Environmental Surveillance of Human and Animal Pathogens in African River Water Using Metagenomics

Alexandra Kossik

A thesis

submitted in partial fulfillment of the

requirements for the degree of:

Master of Public Health

University of Washington

2020

Committee:

J. Scott Meschke, Chair

Alison Cullen

Nicolette Zhou

Program Authorized to Offer Degree:

Environmental and Occupational Health Sciences

School of Public Health

University of Washington

Abstract

Environmental Surveillance of Human and Animal Pathogens in African River Water Using Metagenomics

Alexandra Kossik

Chair of the Supervisory Committee:

Professor J. Scott Meschke

Department of Environmental and Occupational Health Sciences

Diarrheal disease remains a persistent global health threat and is ranked as the second leading cause of death for children under the age of five, with approximately 525,000 child deaths worldwide annually. Children with HIV are eleven times more likely to die from a diarrheal pathogen, making studies that examine disease-causing organisms in HIV burdened areas like Sub-Saharan Africa particularly crucial. In Sub-Saharan African countries, such as Botswana, infectious agents of public health concern have been detected including norovirus, rotavirus, *Salmonella* Typhi, *Mycobacterium*, *Vibrio cholera*, *Cryptosporidium*, *Giardia*, and *E. coli*. As enteric viruses are responsible for a majority of acute waterborne diseases, their surveillance is critical to understand disease prevalence within a community. The specific aims for this project were (1) to analyze Chobe River water samples for both human and animal bacterial, protozoan, and viral pathogens using next-generation sequencing and (2) to determine the relevance and applicability of environmental surveillance and the analyzed Chobe River water samples to nationally reported health data, policy interests, and national priorities in order to determine potential public health and policy impact. A novel environmental surveillance method, originally developed for poliovirus detection called the bag-mediated filtration system (BMFS), enables in-field collection and filtration of large sample volumes. As the BMFS has been shown to be an effective way to sample, process, and analyze large volumes of wastewater and surface water in multiple countries for multiple pathogens, it was an ideal method for this study. For this study BMFS environmental samples were collected at ten transect points along the Chobe River in both Spring and Summer 2017, along with Fall 2019. The sampling locations encompassed areas of wastewater discharge, pit latrine discharge, and animal-impacts. This range of samples was meant to allow for a broad overview of what is impacting the surface water and to determine if there are fluctuations within the same sampling locations between seasons. BMFS samples, 6 liters each, were collected and filtered in the field at the respective sites along the Chobe River and transferred to a field laboratory in Kasane, Botswana for elution, secondary concentration, and nucleic acid extraction. Samples were then shipped to the University of Washington in

Seattle for library preparation using the Nextera DNA Flex Library Prep Kit and were then sequenced on a NextSeq or a NovaSeq, depending on sample batch. Sequencing reads were analyzed for quality, assembled with SPAdes, aligned with Kraken, and pathogens of interest were filtered out using the Rosetta Stone Microbial Database. Pathogens of interest selected using this pipeline were further analyzed through BLAST searches and KMA pipeline analysis to allow for further interrogation and taxonomic annotation. The pathogens with the five highest number of sequence alignments based on the Kraken pipeline were *Bacillus cereus, Clostridium botulinum, Staphylococcus aureus, Pseudomonas aeruginosa, and Bacillus anthracis.* This study demonstrates promising results for possible recovery of a wide range of viruses and bacteria that could be influencing both animals and humans residing in and along the Chobe River in Kasane, Botswana.

Table of Contents

Acknowledgements

I would like to acknowledge the field support and site access that would not have been possible without CARACAL and Dr. Kathleen Alexander. Additionally, the data analysis and bioinformatics pipelines would not have been made possible without the work of Jim Wallace. Finally, I would like to thank the Castner Award and the DEOHS Environmental Health Microbiome Initiative for funding and the entire Environmental and Occupational Health Microbiology Laboratory for their continued support throughout this entire project.

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Diarrheal disease is estimated to kill 2.2 million people each year worldwide, most of which can be attributed to contaminated food, water, and lack of sanitation (WHO, 2020). Diarrheal disease is the second leading cause of death for children under the age of 5 leading to around 525,000 child deaths worldwide annually (WHO, 2018). For children with HIV, they are 11 times more likely to die from a diarrheal pathogen (CDC, 2018). In this respect, studies that look at what pathogens are causing these diseases are very important, especially in high HIV burden areas, such as Botswana, where the HIV prevalence is one of the highest in the world and ranges between 20 and 25 percent, depending on the region and age population of interest (Avert, 2020). The leading cause of diarrhea worldwide are Rotavirus species followed closely by *Escherichia coli* species and then by *Norovirus*, *Shigella* and *Cryptosporidium* and *Giardia* species, as shown in Table 1 (WHO, 2018). This demonstrates the need for both viral and bacterial research in this area, including but not limited to determination of specific causes of diarrhea, environmental and clinical surveillance studies, and prevention studies. Diarrheal disease prevention, through vaccination, increased sanitation, or otherwise is also a cost-effective way to save lives in both adults and children (CDC, 2015).

Table 1: Diarrheal Disease Pathogen Burdens

Pathogen	Cases
Rotavirus species	215,000 deaths per year in children under 5 (WHO, 2018)
E. coli species	111.5 million foodborne illnesses per year worldwide (WHO, 2015)
Norovirus species	124.8 million foodborne illnesses per year worldwide (WHO, 2015)
Shigella species	51 million foodborne illnesses per year worldwide (WHO, 2015)
Cryptosporidium species	8.6 million foodborne illnesses per year worldwide (WHO, 2015)
Giardia species	28.3 million foodborne illnesses per year worldwide (WHO, 2015)

1.2 Kasane, Botswana

Kasane is a small rural town in the Chobe District of Botswana. This town is in the northern most point of the sub-Saharan African country on the Chobe River, which three other countries share a border with, Zimbabwe, Zambia, and Namibia (Figure 1).

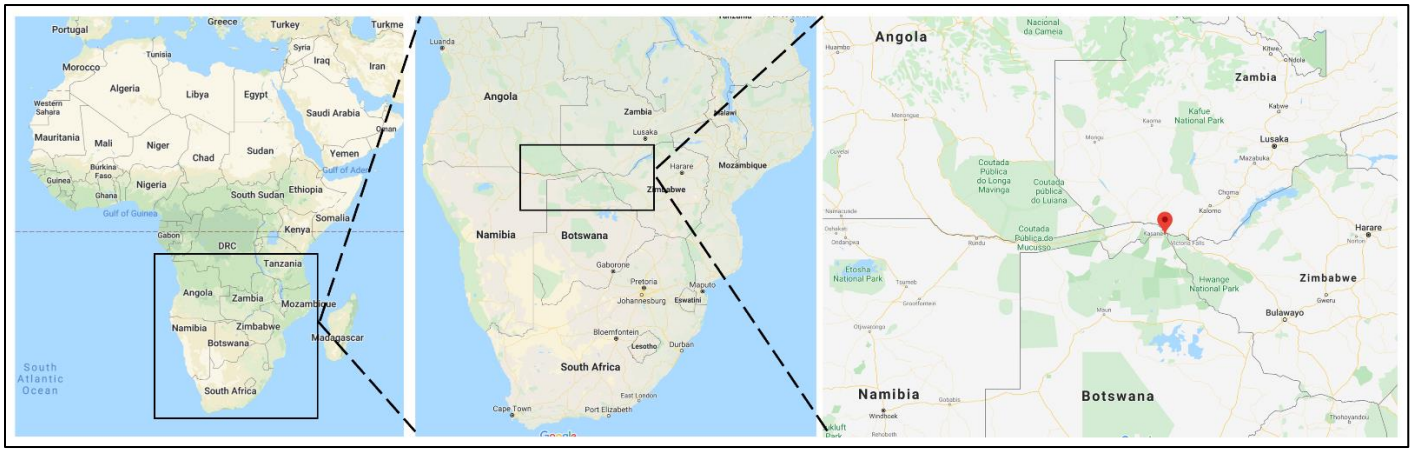


Figure 1: Location of Kasane, Botswana (Images from googlemaps.com)

Botswana has a relatively small population for its land mass with 2.31 million people, around 70% of which live in urban settings either in Francistown or near the capital of Gaborone in the southern part of the country and 19.3% of which live below the poverty line (WHO, 2016; CIA, 2020). With a gross national income per capita of $15,000 and an average life expectancy at birth of 66 years, Botswana is in line with other rural sub-Saharan countries in the region, such as Zambia and Namibia, whose economies are closely associated with the safari tourism industry (WHO, 2016). Botswana was a British colony and gained its independence in 1966, which is still evident in English being the only official language and making the country predominantly Christian (CIA, 2020). However, almost 80% of the population also speaks Setswana and possibly another tribal language, though none of these have official designations (CIA, 2020). The country's fertility rate has been decreasing over recent years, thanks to family planning initiatives, while the HIV/AIDS burden has still been gradually taking lives, even with increased treatment resources, leading to a slower rate of overall population growth (1.48%) with even smaller numbers at the youngest and oldest ages and a median age of 25.7 years (CIA, 2020). While access to an improved drinking water source is 99.2% in urban areas and 92.3% in rural areas, access to improved sanitation facility access is 78.5% in urban areas and only 43.1 % in rural areas (CIA, 2020). Improved drinking water source can be defined by The World Factbook as "use of any of the following sources: piped water into dwelling, yard, or plot; public tap or standpipe; tubewell or borehole; protected dug well; protected spring; or rainwater collection," while improved sanitation facility access can be defined as "use of any of the following facilities: flush or pour-flush to a piped sewer system, septic tank or pit latrine; ventilated improved pit (VIP) latrine; pit latrine with slab; or a composting toilet" (CIA, 2020; CIA, 2020). In the low-income villages in the Chobe Region, around 70% of households use pit latrines and around 15% of households do not have access to improved sanitation (Alexander, 2015).

1.3 Chobe Research Institute and CARACAL

The Chobe Research Institute is located in Kasane, and is comprised of the Alexander WildLab, the Centre for Conservation of African Resources: Animals, Communities, and Land Use (CARACAL), and CARACAL World of Wildlife. This institute is the only research institute, wildlife rehabilitation facility, snake park, and functional research laboratory in Kasane. The center is run by Dr. Kathleen Alexander, DVM, PhD, who is faculty at Virginia Polytechnic Institute and State University as well as Board President and Co-Founder of the center. Dr. Alexander and her team have standing permits with Chobe National Park that allows for collection of samples within park boundaries along with standing relationships with the community to facilitate the collection of samples throughout the river system.

In addition to research, the center is a place for rehabilitation of many species found in Chobe National Park and in Kasane that may have been injured. As with any rehabilitation facility, their primary goal is to return the animals to the wild. However, if the animals cannot recover and be released, they are housed and cared for at the center and used for educational purposes. The center is open daily to the public to encourage education and understanding of the unique species that inhabit the Chobe District. The center also promotes awareness and emphasizes the importance of protecting and preserving this unique habitat.

1.4 Previous Research Efforts in The Chobe District

Previous studies in the Chobe District of Botswana conducted by CARACAL have focused on the detection of antimicrobial resistant *E. coli* and Total Suspended Solids (TSS) in the Chobe River (Jobbins, 2015). A 2013 study found that the wildlife in the region, especially elephants, significantly contributed to the change in water quality across seasons, where the dry seasons showed increased counts of bacteria compared to the wet seasons after initial precipitation run-off events (Alexander, 2013). There is a general lack of sanitation services in the rural areas of Northern Botswana, including the rural towns around Kasane where pit latrines were the dominant form of sanitation (Alexander, 2015).

A study on greywater disposal in the region demonstrated that the use of pit latrines and the dumping of greywater down them could contribute to increased leaching of chemicals and microorganisms into the environment (Alexander, 2015). As the Chobe River is the only body of water in the region, pit latrine waste could reach the river due to these practices. One study in the region looking at child fecal samples to identify a new protoparvovirus also detected Rotavirus, Adenovirius, and Astrovirus, which demonstrates that these viruses, which have the ability to cause diarrhea, are in circulation within the population of Botswana, but have yet to be better quantified (Phan, 2016).

Past research studies in the region have also provided background survey knowledge on the animal pathogens found in areas close in proximity to the Chobe National Park. For instance, the emerging pathogen *Mycobacterium tuberculosis* complex mungi was first identified in this region and its transmission pathway along with its reservoir in the Banded Mongoose has been very well defined (Alexander, 2016, 2016, 2018). Additionally, a study of mongoose fecal samples and swabs confirmed the presence of *Leptospira interrogans*, a zoonotic pathogen that can be spread to humans through urine, soil, or water (Jobbins, 2013; Jobbins, 2014). Though this study did not look at human samples or direct spread through surface water, Leptospirosis infections are a large contributor to disease burdens worldwide and should be researched in more detail (CDC,2013). Other studies have looked at both domestic and wild animals to look at crossover not only between humans and animals, but also between different animal species, which has demonstrated that Rift Valley Fever Virus, another zoonotic pathogen, circulates in both cattle and African Buffalo populations in the region (Jori at al, 2015).

The most recent published studies available from the Chobe region show the importance of both human and animal impact on the water system. For example, one study, which separated sampling sites by land use type (park, town, multi-use) and was conducted using 1L river water grab samples analyzed for *E. coli*, antibiotic resistance, and TSS, showed the highest numbers were reported in the wet season near the most human impacted portion of the river where a majority of households rely on put latrines (Sanderson, 2018). Another study interpreted the interaction between El-Nino Southern Oscillation and childhood diarrheal disease (Heaney, 2019). This study found that, based on clinically collected case data, cooler wetter years were associated with high cases of childhood diarrheal disease, most likely due to increased fecal contamination in the Chobe River due to run-off and animal behavior and interaction with the river. Changes based on season do not seem to be taken into account with the local water quality treatment facility, which combined with only having *E. coli* water quality data, lacking other pathogens, means diarrheal disease-causing pathogens could be missed. In this respect, this

project could contribute to the overall data and understanding in the Chobe region by providing data on other diarrheal disease-causing pathogens in the river water.

1.5 Metagenomics Studies

While no previous studies have combined metagenomic analyses with the Bag Mediated Filtration System, the use of metagenomic methods on environmental samples and water samples in particular has been documented. Next-generation sequencing was combined with metagenomic methods in order to detect viral pathogens in wastewater-impacted river waters in Uganda demonstrating contamination from a nearby treatment plant (O'Brien, 2017). These samples were filtered in field and then transported back to the U.S. for analysis. A small portion of the samples were converted to cDNA and prepped for sequencing using the Illumina TruSeq Nano DNA Library Prep Kit followed by sequencing on an Illumina HiSeq 4000 flow cell. These sequences were then analyzed using BLAST and MEGAN databases. This study showed that it is possible to conduct viral metagenomic analyses on wastewater samples. Other studies have shown similar next-generation methods paired with well-known viral concentration methods, like skim milk flocculation, to isolate bacteria and viruses from wastewater and irrigation waters (Girones, 2020). As with all metagenomics research, the analysis is only as definitive as the databases are robust. Though studies have been done analyzing viral genomes in different water types, more work needs to be done using next-generation sequencing technologies both to help understand how to interpret sequencing findings that are inconclusive and to help characterize water metagenomic profiles in other regions of the world that are less studied (Haramoto, 2018).

The field of metagenomics has also been used to study antimicrobial resistance in various water types showing that resistance profiles are higher closer to regions of human impact (Port, 2014) and that the conclusions are only as good as the constructed databases (Gupta, 2020). The samples require more annotation on the sequence level including processing through MG-RAST, prediction with MetaGeneMark, followed by metagenomic and antimicrobial databases or a ribosomal database followed by the Rosetta database to predict pathogenicity (Port, 2014).

Risk assessment can be used to inform regulation, but only if the data is available to conduct a proper risk assessment, metagenomics can supply this data. For example, genetic information was combined with environmental and social factors, including exposure and disease status, to influence air quality regulation in the U.S. (Kramer, 2006; Cullen, 2008). Though genetic information would not be directly useful or applicable in this project, the understanding of how to incorporate metagenomics approaches with environmental and social aspects could be useful when interpreting the findings of this project.

The field of sequencing, including the databases that are used for alignment of sequences, were originally developed for clinical fields and their corresponding uses, making the adaptation of this technology to environmental media complicated and rather nuanced. Seeing as sequence calls are only as definitive as the databases they are put into, the databases cannot be assumed to account for the many environmental agents that may not be discovered and published yet, so not overstating conclusions is an important point to incorporate into results and discussions based purely on environmental data. Though environmental fields are behind clinical fields when it comes to sequencing databases, progress is being made to give researchers the tools they need. For instance, the FARME database was developed to incorporate the usage of functional

metagenomic gene elements to help researchers get sequencing hits for antibiotic resistance bacteria that are non-culturable and would not be included in other clinically-based databases (Wallace, 2017). Sequence assemblers are also being modified in order to better accommodate environmental sampling data, such as the adjustments that were made to the SPAdes assembler that led to the metaSPAdes assembler option of the software. MetaSPAdes includes updates that help deal with the extensive repetitive sections found in novel environmental sequences along with changes in contig development that help allow the possibility for novel alignments in downstream pipeline steps (Nurk, 2017). In this respect, all of the steps in transforming next-generation sequencing data into results need to be analyzed seeing as the assemblers and databases chosen can greatly influence the outcome (Sutton, 2019).

If there was a simple, reliable method to test for all pathogens present in a water sample from the Chobe region the overall understanding of the water quality in the region would be better and could be applied to future regional issues. For instance, the large elephant die-off currently being experienced in the Chobe region, which is hypothesized to be coming from water contamination, could be analyzed more efficiently (Azeem, 2020).

1.6 Specific Aims

As the rural river town of Kasane faces seasonal fluctuations in diarrheal disease, further examination of pathogens in the Chobe River is needed. Previous studies identifying human and animal pathogens in the region have focused on fecal and swab samples, with very little water sampling conducted. As a result, an in-depth analysis of the water source from which both the humans and animals carrying these pathogens draw is where this research is based. It has been shown that many pathogens that cause diarrheal diseases are spread via the fecal oral route and that these pathogens can be dispersed through common water sources.

Due to the need to understand the characterization and transmission of pathogens that cause diarrheal disease and the interplay between species in this region, there needs to be more extensive sampling and analysis done on human and animal impacted river water samples. To progress towards this goal, the following aims were completed:

Hypothesis I: Viral, bacterial and protozoan pathogens in the concentrated water samples are detectable by sequencing ad have been shown to cause diarrheal disease. We predict zoonotic strains of some of these pathogens will be present, due to the human and animal interaction in the surrounding areas.

Aim 1: Analyze Chobe River water samples for both human and animal bacterial, protozoan, and viral pathogens using next-generation sequencing.

Approach: To better understand what pathogens are influencing disease cases in this region, we collected river water samples at transect points throughout the river system, ranging from the national park to the town. The Bag Mediated Filtration System (BMFS) has been shown to be an effective way to sample, process, and analyze large volumes of wastewater and surface water in multiple countries, making it the method of choice for this study. This system allowed us to detect other pathogens that may be contributing to the diarrheal disease burden in the Chobe region. We collected, filtered, and processed human and animal impacted river water samples across three years and three different

time points, totaling near 30 samples. To complete this process, we utilized next-generation sequencing for bacterial and protozoan targets along with whole transcriptome amplification followed by next-generation sequencing to attempt to identify viral targets. This analysis method, using whole genome shotgun sequencing within metagenomics, allowed us to better determine what pathogens are present within the water system, enabling us to estimate possible contributions to disease burdens.

Hypothesis II: Diarrheal disease cases in humans and the number of diarrhea-causing pathogens in the water are correlated and vary with season.

Aim 2: Determine the relevance and applicability of environmental surveillance and the analyzed Chobe River water samples to nationally reported health data, policy interests, and national priorities to determine potential public health and policy impact.

Approach: Though not all individuals seek treatment when they are ill, especially with diarrheal symptoms, Botswana does track and report notifiable diseases at a national level, including a general category for diarrhea. By comparing sampling results with reported clinical data, we will detect any associations. In this way, we can determine if the water samples are able to predict human disease burden in the region.

This study established new background data on the pathogens present in human and animal impacted African river waters. This provides greater insight as to what pathogens play a role in the increase of diarrheal cases according to seasonal variability and overall disease burden for the region. Applying this new sampling and processing technique will broaden the range of pathogens that can be detected, leading to the ability to examine the utility of environmental surveillance as a public health monitoring tool.

Chapter 2: Environmental Surveillance

2.1 Background and Significance of Environmental Surveillance

Environmental surveillance, is the study of environmental medias, including but not limited to air, surfaces, food, and various types of water, to monitor for pathogens of interest. The purpose is to see if and where pathogens are present in the environment and how humans may be exposed to them, which is an important and extremely relevant field for disease surveillance and emergence. While health systems report clinical cases of reportable diseases, there may be instances where people either do not seek healthcare or are asymptomatic carriers. For instance, a majority of people infected with poliovirus do not present with acute flaccid paralysis, which is when a clinical case would be reported. Similarly, SARS-CoV-2, the virus that causes COVID-19, has been shown to be spread by people who are not symptomatic, leading to increased spread since people are not tested and do not isolate due to their lack of symptoms or knowing that they are infectious and have the ability to spread the virus to others. In these instances, the disease is still spreading within the population without being tracked. This is where environmental surveillance comes into play because it can supply information as to where the disease may be spreading and where containment, treatment, or prevention efforts should be focused. This study was specifically focused around the surveillance of environmental samples collected from a freshwater river that was known to be impacted by both humans and animals leading it to fall within the category of wastewater-impacted river water seeing as both the humans and animals in the region could be contributing pathogens to the river water, including but not limited to fecal contamination or run-off from human activities.

2.2 Typhoid Programs

Studies in Botswana, though it does not monitor for typhoid, environmental or otherwise, have shown that the country is at an increased risk due to the extreme seasonal flooding a large portion of the country experiences. In this respect, the prevalence of typhoid in Botswana is unknown and therefore typhoid is also not part of routine vaccination schedules in Botswana, likely due to lack of funding and a focus on Rotavirus vaccination instead. Multiple papers have shown that typhoid has been isolated from clinical patients who exhibited symptoms for Typhoid Fever in Botswana making typhoid a possible pathogen of interest for future environmental studies (Kim, 2019; Mogeni, 2019; Antillon, 2017).

2.3 Polio Programs

One of the largest, most well-funded, eradication efforts in recent history has been for poliovirus, where the Global Polio Eradication Initiative supplements clinical surveillance for acute flaccid paralysis with environmental surveillance. Use of environmental surveillance is particularly important for poliovirus as a majority of people infected do not show the standard diagnostic symptom of acute flaccid paralysis, but instead are asymptomatic carriers. When the Global Polio Eradication Initiative began, in 1988, poliovirus cases were reported in 125 countries, while today polio is only endemic in two countries (Patel, 2019). Though poliovirus was not eradicated and is likely to see increased case numbers due to complications with both the vaccine and SARS-CoV-2, this initiative was able to accomplish what it did because it used environmental surveillance as a supplement to clinical surveillance to inform vaccination campaigns.

Botswana has demonstrated its investment in immunization campaigns seeing as the government, with assistance from the WHO, funded a door-to-door immunization of all children in 2004 (UNICEF, 2004). The entire campaign was in response to one positive clinical case in Maun, a town in north central Botswana that was likely imported from an ongoing outbreak in Nigeria at the time and was the first diagnosed case in Botswana since the country was declared polio-free in 1991. Since 2004 there have not been any immunization campaigns for poliovirus in Botswana, but the percentage of children who have received a third dose of a vaccine containing poliovirus has been estimated to be 77% as of 2019 (WHO/UNICEF, 2020). However, this coverage percentage is an estimation and "no nationally representative household survey [has been conducted] within the last 5 years [and] WHO and UNICEF recommend a high-quality survey to confirm reported levels of coverage. WHO and UNICEF recommend an assessment of the administrative recording and reporting system" (WHO/UNICEF, 2020).

2.4 Campylobacter Programs

Campylobacter accounts for over 1 million foodborne illness infections a year in the U.S. (CDC, 2019). While numbers are not tracked as closely or at all in low- and middle-income countries, it is clear that Campylobacter species likely account for a large amount of foodborne illness worldwide. A study done completed by CARACAL tested human fecal samples, animal fecal samples, and river water samples for the presence of Campylobacter species and found that they were detectable in the human and wild animal samples (Medley, 2020). The findings of this study allude to the fact that both wild and domestic animals could possibly be reservoirs for Campylobacter species and the transmission pathways should be further studied.

2.5 Antimicrobial Resistance

Antibiotic and antimicrobial resistance (AMR) is a public health concern due to the fact that when bacteria become resistant to antibiotics they become much more difficult to treat. The large factors contributing to this growing threat are human antibiotic use and domestic animal antibiotic use. In this respect, antibiotic resistance is usually higher in human impacted areas or in areas that have a high commercial livestock burden. Though illnesses associated with and deaths due to AMR are not tracked in Botswana, there have been multiple studies in the Chobe Region that look at AMR across land use types and animal species, making this research area one that could be built upon in the future.

2.6 Risk Assessment and Risk Management Importance

Risk assessment and risk management are both important fields that can be influenced by environmental surveillance. If a pathogen of interest is either suspected or found in a region, risk assessment and subsequent risk management techniques would need to be applied, depending on the severity of the situation, to mitigate the effects of the pathogen.

This brings to light the idea and discussion of what a positive result in environmental surveillance correlates to and how that changes by region and possibly by type of positive result. For instance, in the polio environmental surveillance system, if an environmental sample tests positive for wild poliovirus type 1 or circulating vaccine-derived poliovirus type 2 it can trigger human vaccination campaigns, seeing as poliovirus is nearing eradication. However, in the case of other pathogens of interest, the reaction to a positive environmental sample may not be as great. This could be for multiple reasons, one of which could be due to other reservoirs for

pathogens affecting the ability to actually eradicate pathogens, unlike poliovirus that only infects humans. Another reason for less extreme or direct actions to be taken in response to a positive sample could be if a pathogen is very prevalent, not nearing eradication as poliovirus is, so surveillance is more helpful in the understanding of transmission and determining interventions that could be helpful either now or in the future as opposed to direct, immediate action. For example, SARS-CoV-2 is a pathogen where a positive result from a wastewater sample at this time does not trigger an immediate response but is rather being used for fact-finding and transmission understanding. However, in the future, when there is a vaccine for SARS-CoV-2 and the disease is less prevalent, a positive environmental surveillance sample may result in vaccine campaigns or an early warning to trigger shutdowns of certain activities to prevent virus spread.

In the study region of Kasane, Botswana a positive environmental sample could come with a fair number of caveats. Seeing as Kasane relies heavily on the animal tourism industry, the health of the wild animals is of large concern along with the human accessibility to view the animals. If a reaction to a positive sample were to impact this industry's ability to function and continue to make money, it may not be a reaction the community would want to take. For instance, if a disease was found in animals that required distance to be maintained and the national park to be closed, this would negatively impact the town. Additionally, due to a large amount of interaction with the one water source in Kasane, the Chobe River, by both humans and animals, anything that would impact this water source would also be of increased concern. Animals from the entire region travel to the Chobe District during the dry season, which concentrates the animals and humans into a smaller area than throughout the rest of the year. Humans interact with the water for recreational, industrial, and occupational reasons, all of which are important to overall well-being.

Kasane has a limited health infrastructure which consists of one small hospital for the entire region that sometimes relies on researchers for supplies. The 2011 census reported that 9,244 people live in Kasane while an estimated 33,000 tourists visit Kasane each year and there are an estimated 4000 patient visits to a hospital each year in the Chobe Region (Census, 2011; Tourism Report, 2017; Outpatient Stats, 2016). Additionally, there are many underlying health concerns in the region including an increased HIV burden. These points combined with the income disparity between lower income local residents and high-income tourism businesses can lead to different interests when it comes to health and environmental surveillance, all of which would need to factored into determining the reaction to pathogen detection in the region.

In general, Botswana as a country has a reliable economy and a relatively small population in comparison to the physical size of the country. However, the majority of the population is concentrated in and around the capital, Gaborone, which is along the southeast border of the country while Kasane, the study location, is the northern most point of the country. In this respect, a lot of the country-wide decisions are being made in the capital and based on the population center of the country and may not necessarily take into account the smaller more rural regions of the country. Seeing as the decisions are being made so far away from the study region, it may change how seriously a positive sample finding is taken at the national level, seeing as there may be more pressing needs for the majority of the population. Additionally, even if a pathogen is found to be of high concern for the entire country, there may still be a lack of overall

resources to take action. An example of a pathogen that could possibly trigger a response in Kasane would be any that can significantly impact the elephant population, such as *Bacillus anthracis*, where a likely response would be to test the animals that have died and burn their carcasses to prevent potential spread.

Finally, even after the importance of a pathogen of interest is determined, there are still general aspects of execution of environmental surveillance that need to be considered. For instance, how projects are funded could change both the impact they have on the community and the topics of interest. Environmental surveillance studies that are organized around the community interests would have the ability to have the largest impact because both community members and researchers would be interested in the outcomes. While the community was not directly invested in this project, the findings of this study could be applied to future studies to help build an environmental surveillance system that does have community investment and interest.

2.7 Botswana Statistics

The *Botswana Environment Statistics Water and Climate Digest 2017* states that there is a water production facility in Kasane that both treats and supplies water to the surrounding region. The Water Utilities Corporation in Kasane is not large enough to warrant yearly reporting, but the 2017 digest reports a total production of 1,467,904 kiloliters of water across 2016 and 2017. The 2017 digest also states that the total annual rainfall, measured at the Kasane Airport, was 342mm (Botswana Environment Statistics Water and Climate Digest, 2017).

The *National Health Policy: Towards a Healthier Botswana* produced by the Ministry of Health in 2011 states that the life expectancy at birth in Botswana is 54.4 years and that "a total of 25.9% of the population are stunted." This report also states that "digestive system diseases" are one of the "major causes of mortality of public health concern," a "major cause of disease burden in Botswana," and "more than two-thirds of [under-5 mortality rate] deaths are due to communicable disease, with diarrhoea and pneumonia being the two main killers" (National Health Policy Towards a Healthier Botswana, 2011). A majority of the environmental health related policy initiatives covered in this report pair the importance of keeping people healthy with that of maintaining the health of the environment and animal populations, which helps to also maintain the robustness of Botswana's tourism industry.

The 2002 Botswana Central Statistics Office report, *Selected Environmental Indicators,* does not focus on direct human or animal health indicators, but rather places a focus on energy consumption, air pollutants, and protected land usage (Selected Environmental Indicators, 2002). However, this report does denote "Global Warming Potentials" including carbon dioxide and methane. Though these two emissions are not directly impacting surface water, global warming can lead to changes in seasons which can change animal behavior, human behavior, and overall water usage dynamics.

Botswana's *General Outpatient and Preventive Health Statistics Brief 2016* states that diarrhea qualifies as a notifiable disease in three reportable categories, diarrhea with blood, diarrhea (acute) with dehydration, and diarrhea (acute) with no dehydration. The total across these three categories for 2016 was 129,469 cases. The report also states that the Diarrhea Case Fatality Rate fluctuates over different years, being "high in 2006 and 2012 and low in 2015 and 2009"

(General Outpatient and Preventative Health Statistics Brief, 2016). Seeing as most people suffering a diarrheal episode do not seek treatment, these case numbers are likely underestimating the overall health burden of diarrheal disease.

While these nationally published reports provide some insight as to the policy interests and disease tracking abilities of the Botswana health system, a lot of questions are left unanswered, most of which can be attributed to lack of resources. However, it is clear that Botswana values and invests in the health of its population and that diarrheal disease is both a health burden and an important aspect to future plans.

Chapter 3: Methods and Sequencing Results

3.1 Sampling Locations

Ten sampling sites along the Chobe River were chosen by CARACAL colleagues. All of the sites are transect points that CARACAL has routinely sampled at in the past and they cover a broad range of locations, both within the Chobe National Park boundaries and outside of the park closer to the population centers of Kasane. This range of samples allowed for a broad overview of what was impacting the surface water and if there were fluctuations within the same sampling locations between seasons and across years. All samples were taken within 10 feet of the shoreline on the Botswana side of the Chobe River, apart from the samples that were taken within 10 feet of the shoreline of Kazakili Island (an island that sits in the middle of the Chobe River between the Botswana and Namibian banks). This sampling scheme remains the same across all three time points (March 2017, August 2017, October 2017) except for the times when the sites were no longer accessible by water routes, due to extreme drought conditions, these sites were then accessed by land. Table 18 lists GPS coordinates for all sites along with more characteristics and photographs of sampling events that took place in August 2017. Figure 2 is a map of all ten sampling sites to demonstrate distance and spread in comparison to the conservation area and the town. Sampling sites span approximately a 12-mile distance along the Chobe River.

Figure 2: Map of Sampling Sites (Image from googlemaps.com)

3.2 Sample Collection and Processing

A flow chart of sample processing can be seen in Figure 3. All samples consisted of 6L water samples that were filtered by gravity in field using BMFS v.1 or BMFS v.2 and were kept on ice during travel back to the field laboratory, seeing as the ambient temperature (99°F to 105°F) was too warm to leave filter housings out of a cooler due to possible UV inactivation of viruses and bacteria in the samples. The March 2017 and August 2017 sampling events were conducted using the BMFS v.1 while the October 2019 sampling event was completed using BMFS v.2. The BMFS was updated from v.1 to v.2 between the 2017 and 2019 sampling event timelines, making this switch necessary. The BMFS was modified for many reasons including decreased production prices, increased viral recoveries, and user feedback. Figure 3 provides a breakdown of the processing methods used on each of the samples in the field and Table 19 provides an explanation of the processing steps each of the samples went through once they arrived at the University of Washington.

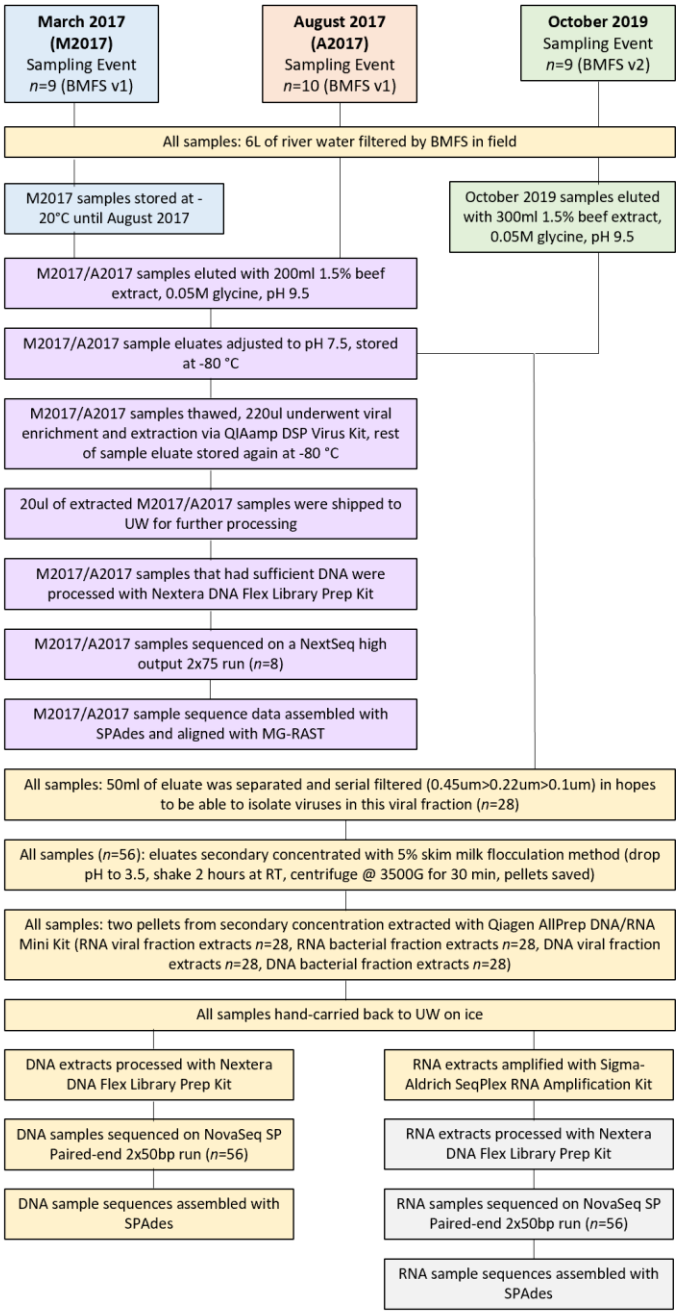


Figure 3: Experimental Field and Laboratory Combined Protocol

3.2.1 Sample Elution

All samples were eluted using a 1.5% beef extract and 0.05M glycine solution that was at pH 9.5 during time of processing. Using a high pH elution buffer with a high protein content has been used efficiently with the BMFS in order to interact with the positively charged filter media to aide in the elution of polioviruses from the filter. Poliovirus is a single stranded RNA virus and is negatively charged, allowed it to be selected by charge with the BMFS filter, the high pH elution buffer changes the isoelectric point of the virus, allowed it to be eluted off of and through the

filter media. While the ability of the BMFS to concentrate other pathogens has been shown in previous studies, the efficiency and mechanisms by which this happens is still to be determined in more detail. Negatively charged single stranded RNA viruses are likely to react similar to polio, while other viruses may not, and pathogens that are larger than the 2-3um pore size of the BMFS filters will likely only have their free DNA concentrated by the BMFS.

March 2017 filters were filtered in field and stored at -20C until thawed and eluted in August 2017. The March 2017 and August 2017 samples were eluted using a double elution technique, where 100ml of the elution buffer was injected into the filter housing, allowed to sit for 15 minutes, and then pumped out using a bilge pump, bringing the final total elution volume to 200ml. Though the elution buffer is used at pH 9.5, the pH is dropped to neutral, 7.0-7.5, prior to storage to prevent potential die-off and inactivation of pathogens.

The October 2019 samples were eluted with the same double elution technique. The elution volume used was increased to 150ml due to the hold-up volume of the BMFS v.2 filter housing, making the total final elution volume 300ml.

3.2.2 Removal and Processing of Viral and free DNA Portion of Samples

Once all samples were eluted, 50ml of each sample eluate was set aside to later become the portion of the samples that is from now on referred to as the viral and free DNA fraction. This 50ml volume of each sample was serial filtered to attempt to remove all other organisms and biological material apart from viruses, seeing as viruses average 0.02-0.25um in size while bacteria average 0.5-5um. The samples were syringe filtered by hand through decreasing pore sized filters: 0.45µm, followed by 0.22µm, and finally 0.1µm (Millipore #SLHP033RRS, #SLGP033RS, #SLVV033RS). The filtrate was saved at 4°C until secondary concentration was performed.

3.2.3 Secondary Concentration

Secondary concentration was completed on all samples, both bacterial and viral portions. March 2017 and August 2017 samples each had a viral and free DNA fraction sample that started as 50ml of serial filtered eluate along with a bacterial fraction sample that started as 150ml of original eluate. October 2019 samples each had a viral and free DNA fraction sample that started as 50ml of serial filtered eluate along with a bacterial fraction sample that started as 250ml of original eluate.

All samples were secondary concentrated using a skimmed milk flocculation method. For every 100ml of sample volume, 1ml of a 5% skimmed milk solution was added. The pH of all samples was acidified to pH 3.5-4.0 using concentrated HCl (pool acid from a local hardware store, Figure 5) and then all samples were shaken by hand (using a modified roller cart, Figure 4) at room temperature for two hours. The samples were then divided into 50ml conicals and centrifuged at 3500G for 20 minutes at room temperature (25-30°C). Finally, the supernatant was poured off of the pellets and discarded while the pellets remained at 4°C until extraction. The March 2017 and August 2017 samples produced 4 pellets for the bacterial fractions and 2 pellets for the viral and free DNA fractions while the October 2019 samples produced 6 pellets for the bacterial fraction and 2 pellets for the viral and free DNA fraction per sample.

Figure 4: Modified shaker constructed from a foldable grocery cart on wheels, rolled by hand for entire shaking period.

Figure 5: Pool acid, purchased from local supply store, used in place of HCl.

3.2.4 Sample Extraction

During the October 2019 trip, two individual pellets from each sample were extracted using the Qiagen AllPrep DNA/RNA Kit. Two pellets were used for this set of extractions because that was the maximum number of pellets produced by any of the viral and free DNA fractions, leading to 100% of all viral and free DNA fractions being processed while only 33% or 50% of bacterial fractions were processed for the 2019 and 2017 sampling events respectively. This set of extractions separates, isolates, and purifies the nucleic acids from the sample into DNA and RNA fractions. This kit ended with an eluate volume of approximately 100µl for the DNA extract output and 60µl for the RNA extract output. Pellets not used in this extraction step were pooled using less than 4ml of sterile DI water and frozen for future use.

A previous set of viral enrichment and extractions was done on a subset of samples (*n*=20) in March 2018 by CARACAL staff. This extraction was done on input volume of 220µl of sample eluate using a QIAamp DSP Virus Kit following filtration on 0.45µm sterile filters and a nuclease treatment. This equated to 0.11% of the original sample being extracted. The final extraction eluent volume was 60µl and was stored at -80°C prior to shipment.

The full protocols for both sets of extractions can be found in the Appendix.

3.2.5 Sample Transport

Samples were transported on two separate occasions. The first round of samples, that were extracted in March 2018, were hand carried to Virginia Tech on ice packs and then a portion was shipped to the University of Washington. The University of Washington received approximately 20ul of each of the samples included in this processing and transport step. Upon arrival at the University of Washington the samples were placed into a -20°C freezer until further processing steps could be completed.

The second transport of samples occurred in November 2019 where all of the samples that were extracted with the AllPrep DNA/RNA Kit were included. These samples were hand carried to the University of Washington on ice packs where they were placed directly into a -20°C freezer after arrival at the airport.

3.2.6 Sequencing Round 1

Samples were quantified using a Qubit 4 Fluorometer, which works through the use of fluorescent dyes attached to DNA or RNA and then read and quantified through UV absorbance. High-sensitivity assays for RNA and DNA were run, though only the DNA assay led to a quantifiable concentration. Due to this finding, it was determined that the best route was to move forward only analyzing these samples for DNA.

The samples with a sufficient amount of DNA (*n*=9 with >1ng in 30µl volume) were processed using the Nextera DNA Flex Library Prep Kit and then sequenced by Seattle Genomics on a NextSeq High Output 2x75bp run in May 2019. Table 3 shows a breakdown of which samples were included on this sequencing run (*n*=9).

3.2.7 Sequencing Round 2 – DNA Sample Processing

All DNA extract samples, both bacterial and viral and free DNA fractions (*n*=56) were quantified using a Qubit 4 Fluorometer. Since all of the samples were determined to be within a usable concentration for the Nextera DNA Flex Library Prep Kit, all samples were prepped using his kit. The starting input volume for the Nextera Flex Kit for all samples was 30µl. All samples concentrations were quantified following the use of the kit and all 56 samples were pooled to create a library pool with a volume of 372.25µl and 18.85nM concentration. Table 20 lists the concentrations of each sample that ended up in the final pool. This pool was sequenced by SeqMatic across two lanes of a NovaSeq Paired-End 2x50bp run in April 2020. A NextSeq run was used for Round 1 samples while a NovaSeq run was used for Round 2 samples in order to attempt to get the same number or reads for each sample in the pools processed on a given run.

3.2.8 Sequencing Round 3 – RNA Sample Processing

RNA extract samples (*n*=56) were analyzed using a Qubit 4 Fluorometer and an RNA High-Sensitivity Kit. Unfortunately, most of the samples registered with very low concentrations of RNA or concentrations that were not even quantifiable at all using this method. Due to this lack of RNA it was decided that the samples would need to be amplified with a whole-transcriptome amplification kit. The SeqPlex RNA Amplification Kit, produced by Sigma-Aldrich, was selected as the kit to use based on previous success with an older version of the kit. Sigma-Aldrich's TransPlex Whole Transcriptome Amplification (WTA) Kit has been used in the past to amplify laboratory prepared samples that mimicked environmental water sources with low concentrations of pathogens (Parker, 2011). The SeqPlex kit offers all of the same technology as the TransPlex kit, but has been improved upon to provide samples in better condition for downstream sequencing analysis. The main difference in the new version of the kit is the addition of a primer removal step that was not necessary in previous kits but is needed in order for next-generation sequencing technologies to work properly. In theory, the final samples after amplification should still be proportionately representative of the RNA that was present in the sample input, everything is just present in higher concentrations.

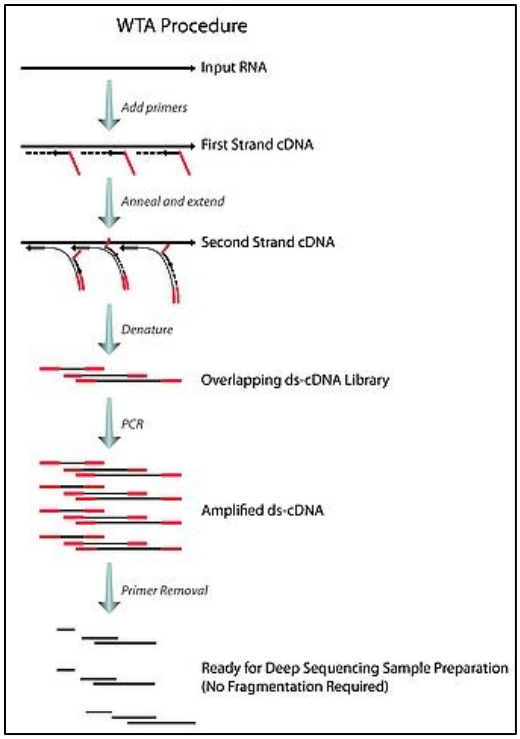


Figure 6: SeqPlex Workflow (Image from SeqPlex RNA Amplification Kit SEQR, 2020)

The protocol for the SeqPlex kit, along with the additional kits used to prepare samples for the kit, are in the Appendix. First the RNA samples were all treated with a RNase-free DNase Kit in order to remove all DNA that may be in the samples. Next the samples were synthesized into libraries followed by an amplification step where all of the RNA still present in the sample was ideally amplified to the same extent. Next the samples were cleaned with a PCR clean-up kit and then a primer removal step was conducted to remove the primer additions that were added to perform the amplification step. The samples were then cleaned a final time with the PCR clean-up kit and stored at -20°C until quantification. The final volume of sample obtained through this process was 30µl. A small portion of the sample (2µl), prior to the final clean-up step, was saved and used to run a primer removal QC reaction. The Ct values for these samples that had gone through the primer removal step were compared to Ct values for portions of the samples that were aliquoted prior to the primer removal reaction. The final samples were quantified using a Qubit 4 Fluorometer and a DNA High Sensitivity kit. The quantification step prior to the kit, the primer removal QC reactions, and the final quantification steps can be seen for each sample in Table 21.

The QC reactions were supposed to shift between 3 to 7 Ct values in order to show proper primer removal. The correct and anticipated shift would be from a higher Ct for samples prior to primer removal and a lower Ct value when the enzyme has been used to remove the primers, meaning that a negative number in Table 21 is correct. However, not all of the samples shifted the correct direction during the QC reactions, 19 of the samples shifted in the wrong direction. Seeing as Sigma-Aldrich, the manufacturer of the SeqPlex kit, had not encountered this specific issue with any other kit users, we tried a few different methods in order to attempt to diagnose the problem and to determine whether or not the samples were ready for further sequencing steps.

The removal of primers from the SeqPlex amplification kit is important for downstream sequencing because primers that are still attached to samples would lead to misalignment of sequences and contigs. Since the QC primer removal reaction was run through RT-qPCR, we tried a traditional PCR using gel electrophoresis in order to determine the lengths of the fragments that were present in the samples. We were hoping to be able to see a difference in samples pre and post primer removal for the samples that passed the QC shift and those that failed the QC shift. Unfortunately, the volume of sample left from each of these steps did not allow the most concentrated aliquot of each sample to be run. However, the gels that were able to be run are shown below.

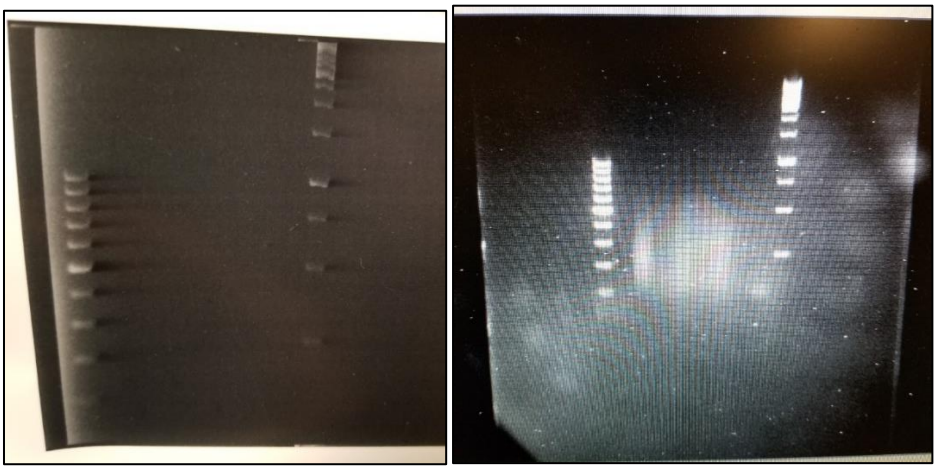


Figure 7: Gels run on RNA samples post-amplification. (left) Gel was run with a 100bp ladder in lane 1 and a 1kb ladder in the last lane, all other non-visible lanes were dilutions of samples processed with the Se Plex Kit. (right) Gel was run with a 100bp ladder in lane 1 and a 1kb ladder in lane 10. Lanes 2 through 9 were samples processed with the SeqPlex Kit and were -1 dilutions of samples prior to primer removal paired with undiluted samples post primer removal. All lanes that are visible in between the ladders are undiluted samples post primer removal.

Through the running of these gels we were able to see that the samples, where there was a concentrated sample left, were present, were not degraded, and seemed to consist of fragments between 200 and 500 base pairs, which is the ideal length for downstream sequencing steps. Based on these findings, and the fact that previous papers siting the use of the SeqPlex kit has not mentioned a similar issue, we decided to proceed with the library prep and sequencing steps.

The Round 3 sequencing, which consisted of the RNA samples after they were processed and amplified using the SeqPlex kit, have still not been sent out. These samples were library prepped using the Nextera DNA Flex Library Prep Kit with an input volume range of 100-200ng. However, when the output of this kit was analyzed by double-stranded DNA high-sensitivity Qubit the DNA concentrations ranged from 0.048 to 0.880 ng/ul, which is too low of a concentration for sequencing.

Due to the low sample output concentration from the Nextera DNA Library Prep Kit, a select number of samples were run on a bioanalyzer. While the electropherograms showed an expected average base pair length range of 300-400, the overwhelming majority of the fragment lengths

for Round 3 samples pre-library prep kit fell in the 40-50bp range (Figure 8). Though more work needs to be done to contact and confirm with the kit manufacturer, the working hypothesis is that the short fragments overwhelmed the beads used in library prep, leading to low output yields.

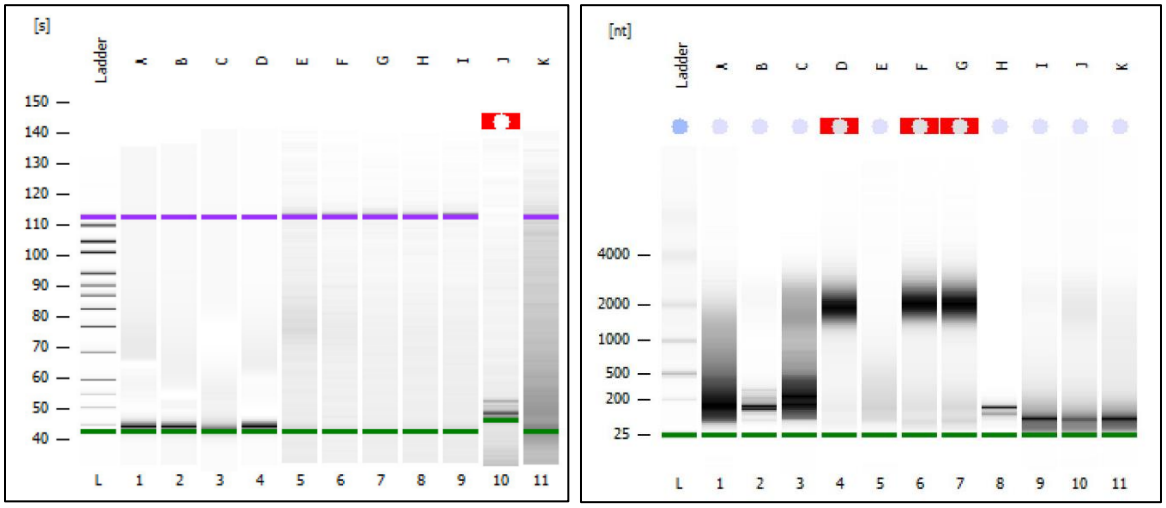


Figure 8: Electrophoresis run summaries for Round 3 samples. (left) Bioanalyzer results where A-D,K are RNA extracted samples pre library prep while E-J are RNA extracted samples post library prep. (right) Bioanalyzer results where A-K are all original RNA extracted samples.

There could be a few explanations for this low output concentration, but more steps need to be taken to determine a path forward. After reaching out to R&D at Illumina to diagnose how fragments bind to the selection beads in the library prep kit and sharing the bioanalyzer results it was determined that the short fragments, which made up a majority of the library sample input, overwhelmed the beads in the Nextera DNA Flex Kit. Still awaiting answers from R&D at Sigma-Aldrich about the inefficiency of the SeqPlex Kit and if that explains the poor fragment lengths. One last effort that was done to analyze the viability of these samples was to Qubit the DNase treated samples and then run those along with the original RNA extracts on a bioanalyzer to address beginning RNA sample quality. None of the DNase treated samples had a high enough concentration to run on a bioanalyzer, but a subset of the original RNA extracts were run as shown in Figure 8. The run showed that while there were some larger fragment lengths in some of the samples, a lot of the fragment lengths were poor or would not be of lengths leading to the RNA viruses of interest during downstream processing steps.

If the samples are ever to be analyzed again, from the beginning, Illumina R&D suggested using the SMARTer Low Input RNA cDNA Kit from Takara BioSciences. Another option would be to start with the secondary concentration pellets and re-extract as opposed to starting with the RNA extracts. This could help by having a more recent extract that is hopefully less degraded or to be able to use an extraction kit that is meant specifically for RNA, increasing efficiency, as opposed to the AllPrep DNA/RNA Extraction Kit that was used (Bibby, 2013). A final suggestion would be to combine those changes listed above with a library prep kit that was designed for lower input sample concentrations, such as the Illumina TruSeq Nano DNA Library Preparation Kit (O'Brien, 2017).

3.2.9 Nextera DNA Flex Library Prep Kit

The Nextera DNA Flex Library Prep Kit, which is sold by Illumina, is designed to prepare DNA samples for next-generation whole-genome sequencing platforms. The samples were sequenced using shotgun-sequencing meaning that each of the samples was broken up into DNA fragments that were then sequenced and were reassembled during the sequencing data analysis step. The Nextera Flex Kit helps to prepare sample libraries for this type of sequencing because it includes steps that assist in fragmentation of DNA sequences (via a bead-linked transposome), the addition of sequencing primers and subsequent indices that essentially add barcodes to every viable fragment of each sample, and assist in purification of the fragment lengths to only include those at are around 350 base pairs, which is ideal for shotgun sequencing.

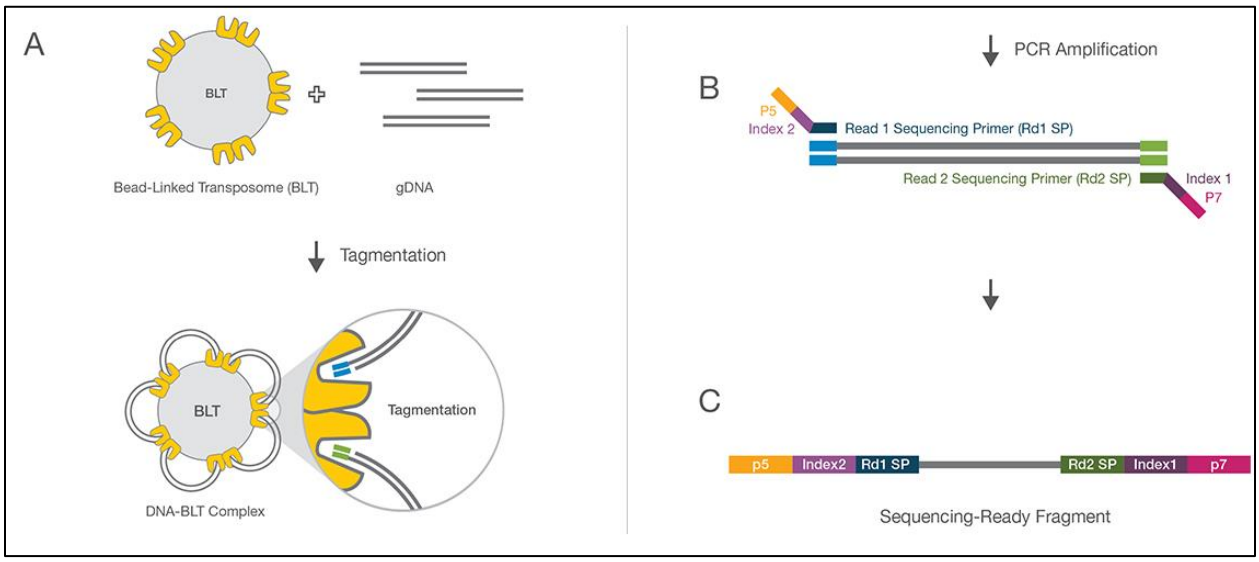


Figure 9: Nextera DNA Flex Library Prep Kit Workflow (Image from Illumina DNA Prep, 2020)

3.2.10 Illumina NextSeq and NovaSeq Sequencers

Different rounds of sequencing have been run on a NextSeq High Output 2x75bp run (Round 1) and on a NovaSeq SP Paired-End 2x50bp run (Round 2, Round 3). Both sequencing platforms can run the prepped pooled DNA libraries that are an output of the Nextera DNA Flex Kit and therefore run multiple samples in the same run, which is more efficient and much more cost-effective. Both machines also use a similar flow cell technology and two-color channel optic system, the main difference being the number of reads each machine can complete per run, as seen in Table 2.

Table 2: NextSeq and NovaSeq Comparisons

	NextSeq High Output 2x75bp	NovaSeq SP Paired End 2x50bp
Output per flow cell	50-60 Gb	65-80 Gb
Paired-end reads	800 million	1.3-1.6 billion
Run Time	18 hours	13-38 hours

Both sequencing platforms work with a similar technology once the prepared sample library has been loaded into the machine. The index adapter regions that were added to the samples during the library preparation steps attach to oligonucleotides that are part of and covering the entire

surface of the flow cells. Next bridge amplification is used to generate clusters of clonal information for each sample. This consists of the DNA strands fragments that are attached to the flow cell oligonucleotides attaching their other adapter end to juxtaposed oligonucleotides and replicating repeatedly; creating clusters of the same DNA sequence fragments near each other on the flow cell. Once this cluster generation is complete sequencing can begin. Illumina platforms use a technology termed "sequencing by synthesis," which can be otherwise described as proprietary. Illumina has designed a method that is able to add a single base at a time to each sample fragment due to their development of a proprietary terminator. By adding one base at a time, per read, the sequencing platforms are able to detect each base that is present and the order in which they appear in each sample fragment. Each nucleotide has a different fluorescent label so the sequencer can image the fluorescence for each sample fragment for each read that is taken as one nucleotide is added at a time for the length of the reads the machine is set to run. The sequencers can then compile this information and can report out the number or reads and the nucleotides read for each sample fragment, based on their unique barcoded indices.

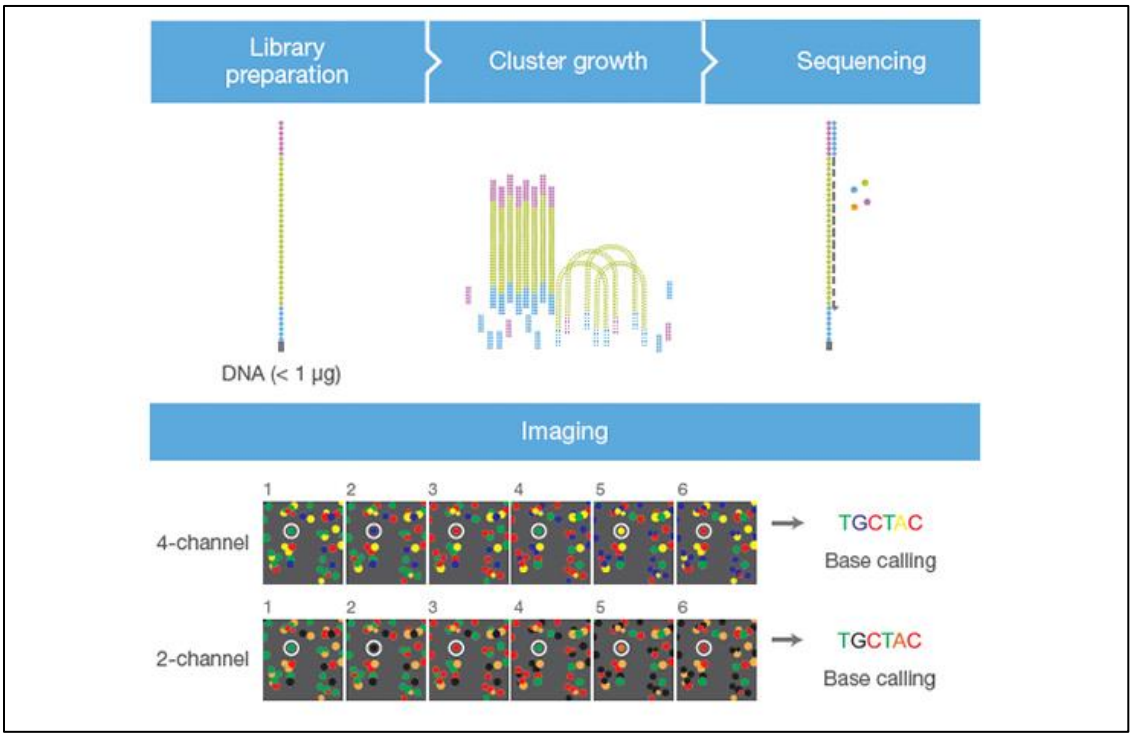


Figure 10: Workflow behind NextSeq and NovaSeq Instruments (Image from 2-Channel SBS Technology, 2020)

3.3 Data Analysis

Though example data tables for only Round 1 or Round 2 can be found in this section, the data tables for the rest of the sequencing can be found in the Appendix. The sequencing assembly pipeline is depicted in Figure 11 and details on these methods are provided below.

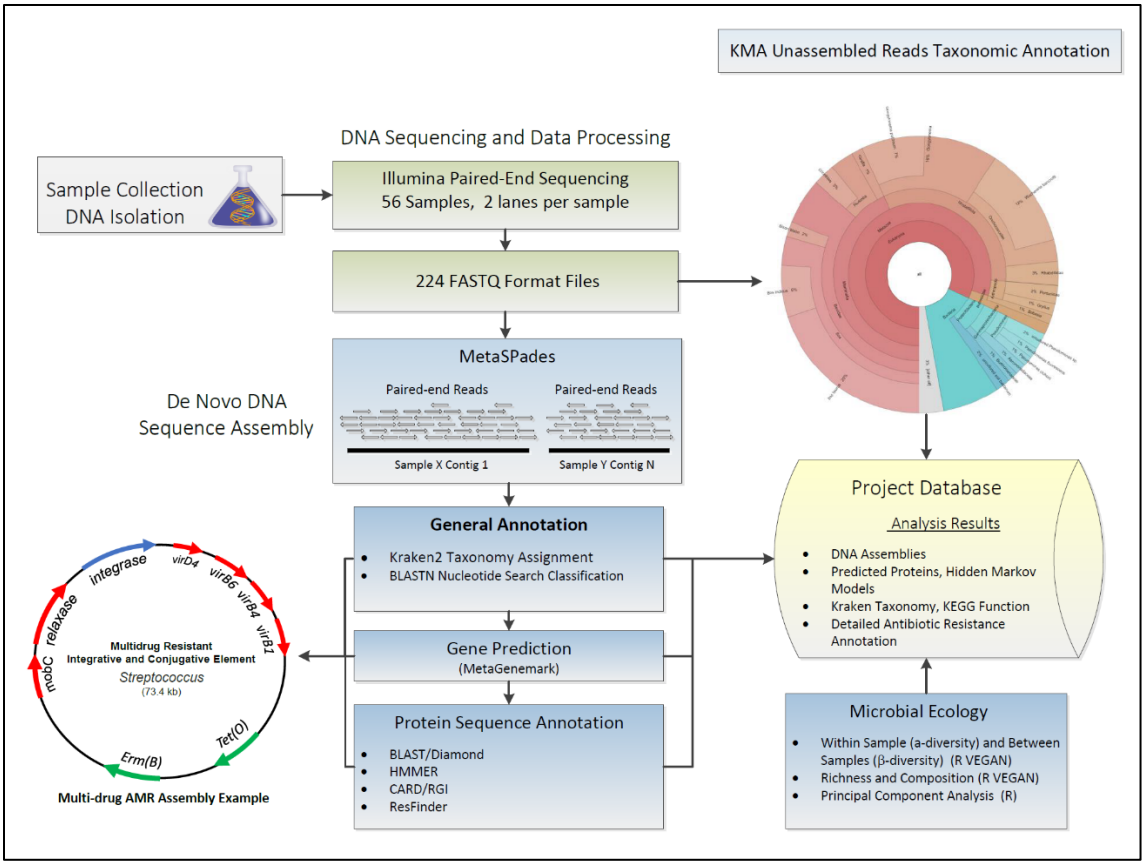


Figure 11: Sequencing Data Analysis Pipeline (Image courtesy of Jim Wallace)

3.3.1 Sequencing Read Data

Following each sequencing run, FASTQ files were created by the sequencer and were sent to UW for analyses. Based on the number of reads per sample, the number of clusters per sample, and the file size, a general quality of the run can be determined. Both Round 1 and Round 2 sequencing data looked promising at this point. Table 3 shows the number of reads per sample from Round 1.

Table 3: Sequencing Round 1 Reads

Sample Name	Date Created	Paired end	Number of Clusters	Number of Reads	Read 1 Length	Read 2 Length	FASTQ file size	Metaspades file size
201708-TR31-pre	5/15/2019	paired	28378347	56756694	76	76	2.55GB	292.36MB
201708-Seep-pre	5/15/2019	paired	25012550	50025100	76	76	2.29GB	319.54MB
201708-TR35-pre	5/15/2019	paired	24365365	48730730	76	76	2.19GB	248.83MB
201708-TR25-pre	1/8/2019	paired	19754876	39509752	76	76	1.92GB	203.46MB
201708-TR27-pre	5/15/2019	paired	22011880	44023760	76	76	1.99GB	235.18MB
201708-TR21-pre	1/8/2019	paired	11478462	22956924	76	76	1.14GB	124.62MB
201708-TR49-pre	5/15/2019	paired	29381562	58763124	76	76	2.63GB	287.66MB
201708-TR1-pre	1/8/2019	paired	72208282	144416565	76	76	6.85GB	too large
201708-TR13-pre	1/8/2019	paired	35182086	70364172	76	76	3.42GB	357.02MB

3.3.2 Read Assembly

Once the reads have been produced and have been assigned back to their original samples, using the barcoded indices attached to each sample read fragment during the library preparation step, the reads need to be assembled. This step takes the samples from just read fragments of lengths of nucleotides to contigs that can eventually be run against different databases. A program called Spades was used and was run through the Illumina Basespace online platform. The setting used for these assembly runs were assembly only, no error correction, paired end, and the Metaspades pipeline within the Spades application was used. This program is able to align different sequences on areas where they have overlapping sequences to attempt to string together longer sequences of the original DNA or RNA that was present in the samples. Table 4 shows the contigs and scaffolds that were created when assembling samples from Round 1. Usually contigs that are less than 500 base pairs in length are not usually relied upon for downstream analysis. GC content relates to how much of the data is in GC repeated regions, which are notoriously difficult to sequence without errors. The N50, N75, L50, and L75 are used to generally analyze completeness of the sequencing assembly. The N50 and N75 are statistics explaining what proportion of base pairs that are in contigs or scaffolds (50% or 75%) are above that base pair length. The L50 and L75 relate to the number of the contig or scaffold, in length order, where the N50 and N75 are found. Generally, a higher N50 and L50 mean better contig or scaffold assembly, but sequencing data quality can still be good with lower N50 and L50 values, these values are used instead to give an idea as to the range of assemblies created.

Table 4: Round 1 Sequencing Contig and Scaffold Assembly

contigs	**# contigs (>= 0 bp)**	**# contigs (>= 1000 bp)**	**Total length (>= 0 bp)**	**Total length (>= 1000 bp)**	**# contigs**	**Largest contig**	**Total length**	**GC (%)**	**N50**	**N75**	**L50**	**L75**	**# N's per 100 kbp**
201708-TR31-pre	244386	25874	142540497	78069793	58820	333003	100796693	45.93	2568	1083	7724	23498	0
201708-Seep-pre	195993	25085	158042915	99132418	56477	176135	120766743	53.44	4546	1357	4547	17696	0
201708-TR35-pre	203795	21944	121346894	65144999	50308	331762	84730094	46.15	2461	1063	6810	20395	0
201708-TR25-pre	158151	17404	99452564	52325289	41477	107697	68883248	46.76	2442	1029	5406	15751	0
201708-TR27-pre	188282	20770	115047704	63282075	47163	331671	81522410	46.04	2602	1088	6129	18716	0
201708-TR21-pre	98303	10493	60834407	31172925	25355	91143	41403224	45.94	2360	1009	3346	10373	0
201708-TR49-pre	244549	25861	145212089	78661839	59567	207249	101896122	45.59	2571	1073	7711	23700	0
201708-TR1-pre	too large												
201708-TR13-pre	296641	31942	173974919	93221117	73465	182813	121899572	45.52	2379	1044	10091	30184	0
scaffolds	**# contigs (>= 0 bp)**	**# contigs (>= 1000 bp)**	**Total length (>= 0 bp)**	**Total length (>= 1000 bp)**	**# contigs**	**Largest contig**	**Total length**	**GC (%)**	**N50**	**N75**	**L50**	**L75**	**# N's per 100 kbp**
201708-TR31-pre	220532	25324	142506586	86140224	55617	333003	107063021	45.86	3293	1223	5978	20032	536.98
201708-Seep-pre	161084	24134	158711704	111150510	52639	184123	130711808	53.33	6878	1668	3256	13828	761.21
201708-TR35-pre	183259	21753	121526156	72222499	48361	331762	90585279	46.09	3087	1187	5430	17816	676.39
201708-TR25-pre	138273	17711	99783849	59283156	40311	121628	74833501	46.7	3146	1159	4253	14772	751.91
201708-TR27-pre	166747	20458	114899560	70476392	44637	331671	87188648	45.98	3389	1247	4753	15892	595.59
201708-TR21-pre	85364	10822	61141440	35678459	25100	102851	45489176	45.88	2973	1129	2786	9351	918.82
201708-TR49-pre	222112	25237	145180089	86437612	56087	209219	107748922	45.53	3329	1212	5952	20121	433.87
201708-TR1-pre	too large												
201708-TR13-pre	266677	32350	174203001	103555656	71542	182813	130689207	45.42	2904	1149	8398	27180	762.72

3.3.3 Assembly Assignment

Once the data was assembled into contigs, Round 1 data was run through MG-RAST in order to see what genomes the contigs aligned to. MG-RAST, which stands for Metagenomic Rapid Annotations using Subsystems Technology, is often used to give a general idea of what a sample of sequencing data consists of in order to help determine next steps. However, MG-RAST is not usually relied upon to make final distinctions of sample contents seeing as there are other alignment options that can be more tailored to datasets. Table 5 lists the class denotations for each sample that was able to be run through MG-RAST (some samples were too large to be assembled with Metaspades and others were too small to be aligned using MG-RAST so those are not included in the data tables).

Table 5: Round 1 Sample MG-RAST Class Assignment

Sample Name	Bacteria %	Viruses %	Eukaryota %	unclassified %	Archaea %	Other %
201708-TR31-pre	80.44	15.04	3.20	1.12	0.30	0.00
201708-Seep-pre	87.32	9.80	1.87	0.56	0.43	0.01
201708-TR35-pre	82.07	13.83	3.01	0.93	0.16	0.00
201708-TR25-pre	85.74	11.46	1.91	0.75	0.13	0.00
201708-TR27-pre	77.93	16.90	3.78	1.22	0.16	0.00
201708-TR21-pre	80.83	15.31	2.64	1.06	0.15	0.01
201708-TR49-pre	80.91	14.48	3.42	0.99	0.20	0.00
201708-TR1-pre	too large					
201708-TR13-pre	83.78	13.01	2.20	0.85	0.16	0.00

3.3.4 Assembly Pipeline

As shown in Figure 11 multiple routes of analysis were attempted, including even running the raw sequences prior to assembly.

The raw unassembled sequencing reads produced in 224 FASTQ files, 56 samples run across two flow cell lanes producing one file for each direction so four files per sample, were run through KMA/CCMetagen for taxonomic annotation. CCMetagen (ConClave Metagenomics) using KMA (k-mer alignment) software is able to us all of the reads from a sample and map them to the database in all possible combinations and then select the mapping assignment that is the most likely so that no short eukaryotic sequences get misread, mis assigned, or left out of analysis, which is a common issue with other databases that use aligned sequences (Clausen, 2018; Marcelino, 2020). This type of output could be useful in both determining what could be leading to low DNA or RNA viral and bacterial concentrations or to compare outputs from the same sample that was run across different lanes in flow cells.

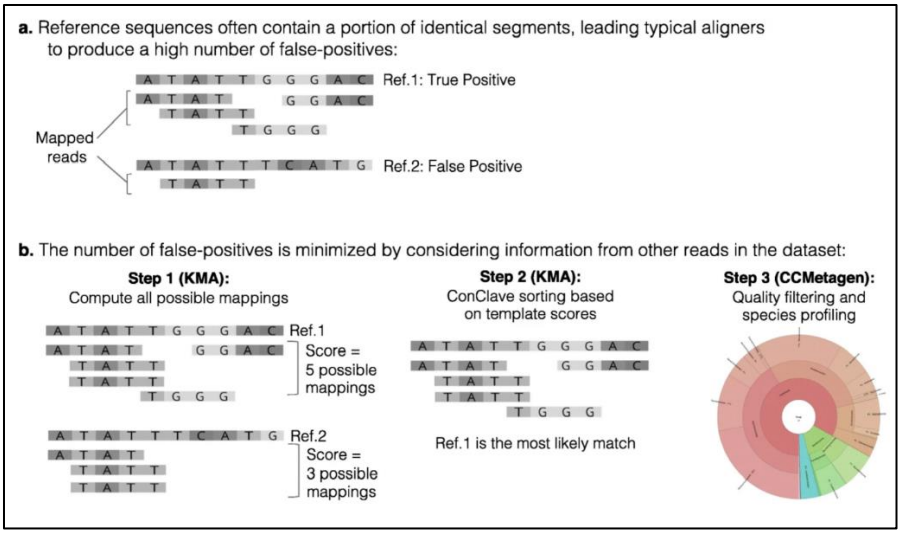


Figure 12: KMA/CCMetagen method explanation (Image from Marcelino, 2020)

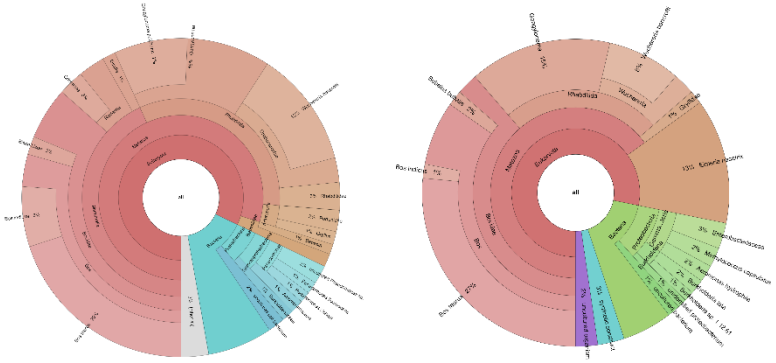


Figure 13: Example of CCMetagen output with Sample 201708_Seep_B

The samples from Round 1 and Round 2 were assembled into contigs/scaffolds using MetaSpades. These de novo DNA sequence assemblies were then aligned using both Kraken2 for taxonomy assignment and BLASTN for nucleotide search classification. Kraken 2 aligns k-mers to its taxonomic database and reports out annotations based on the lowest common ancestor (LCA), which makes for faster processing than previous Kraken programs (Wood, 2019). BLASTN annotates the assembled sequences by comparing them against the NCBI database 11 base pairs at a time (Wood, 2015; BLAST, 2020). Both Kraken 2 and BLASTN perform sequence alignment at similar output speeds, but in slightly different ways. BLASTN aligns against the entire NCBI database so it can be slower but more accurate, while Kraken aligns to a microbial condensed version of the database, allowing for faster alignment though not always as exact. If aligning an entire dataset aligning with Kraken and then annotating further with BLASTN allows the best of both worlds.

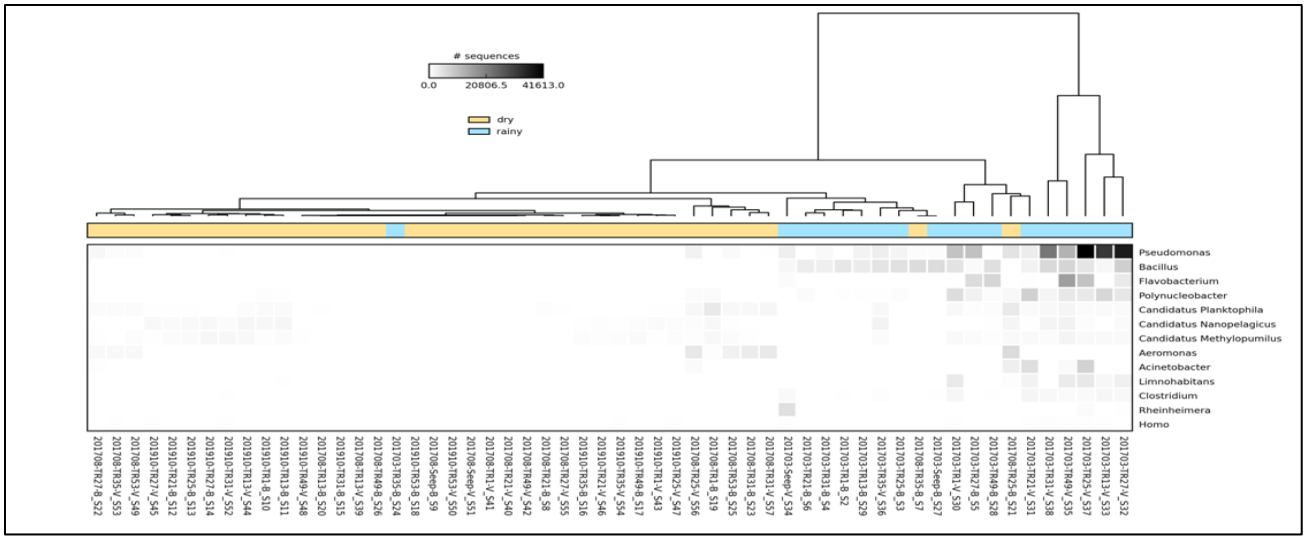


Figure 14: Example of Kraken2 output with all Round 2 DNA Samples. Highest hits for genus
**all heatmaps were made using STAMP Software (STAMP, 2020)*

Classification of probable pathogenic species was done by running the Kraken2 data output through the Microbial Rosetta Stone database. The Rosetta database consists of sequences for "pathogens that impact global public health, emerging infectious organisms, and bioterrorist threat agents" (Ecker, 2005). The sequences that aligned to possible pathogens in the Kraken2 output were further classified through the Rosetta database, adding an extra layer of matching criteria.

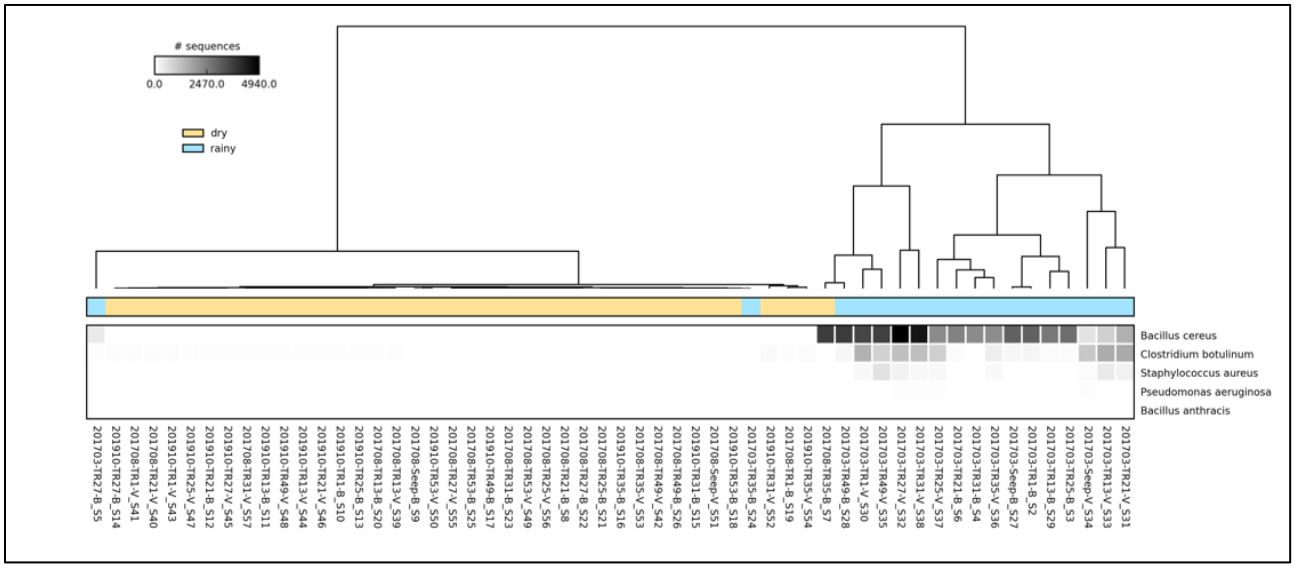


Figure 15: Example of Microbial Rosetta Stone database output with eight top pathogens from all Round 2 DNA samples

Gene prediction was done through MetaGeneMark using the Kraken2 annotations. MetaGeneMark uses not only published genomic sequences for actual gene translation, but also *ab initio* computer algorithms, to increase the ability to predict novel genes, from the short sequences produced through shotgun metagenomics (Zhu, 2010). MetaGeneMark allows one to

gain gene predictions that are novel and would therefore not be aligned using other pipelines or programs seeing as many environmental gene assignments are likely to not yet be entered into common databases. The MetaGeneMark output was then used for protein sequence annotation.

Protein sequence annotation was done through the use of BLAST/Diamond, HMMER, CARD-RGI, and ResFinder. Protein annotation can help to determine if there are functional aspects of the annotated genes present. Diamond is a program that can align protein sequences to a protein database similar to BLASTP, just on a larger and faster scale (Buchfink, 2014). HMMER is a program used to search sequence databases, in this case protein databases using hidden Markov models to help select the most likely sequence alignments (HMMER, 2020). AMR assembly was done through CARD-RGI, Resistance Gene Identifier of the Comprehensive Antibiotic Resistance Database, which has the ability with its algorithms based on homology and single nucleotide polymorphisms, to predict resistomes within the protein annotation (Guitor, 2019; CARD, 2020). ResFinder is a program that is able to identify both acquired antimicrobial resistance genes and chromosomal mutations from input sequences (Bortolaia, 2020). Additional AMR assembly was done following protein annotation by searching all predicted proteins against the entire GenBank NR database, which is the non-redundant protein portion of the GenBank database updated by NCBI (Genbank Overview, 2020). This pipeline can be classified as rather strict matching criteria for identifying AMR genes. Examples of annotated genes and proteins can be found in Table 7 and Table 8.

Table 6: Protein Assembly Statistics

DNA contigs assembled with MetaSPAdes	3,764,994
Base Pairs	1,647,851,336
Predicted protein sequences	3,834,175
Amino acids	469,488,734
GenBank non-redundant proteins searched	290,581,163
Predicted protein sequence BLASTP matches at e-value 0.001	2,729,758
Predicted protein sequence BLASTP matches 90% identity, 90% target alignment	216,832

Table 7: GenBank AMR Protein Matches for Round 2 DNA Samples. Match counts were produced through searching the GenBank NR AMR protein output for “beta-lactamase”, “tetracycline”, “aminoglycoside”, “chloramphenicol”, “fluoroquinolone” "resistance" along with a 90% alignment, 90% matching, and a minimum of 200 amino acid criteria.

Sample	#	Sample	#	Sample	#
201703-TR31-V_S38	39	201703-TR1-V_S30	8	201708-TR31-V_S57	2
201703-TR13-V_S33	31	201703-TR1-B_S2	8	201910-TR13-B_S11	2
201703-TR27-B_S5	18	201708-TR1-B_S19	7	201910-TR1-V_S43	2
201703-TR25-V_S37	15	201703-TR49-B_S28	7	201703-Seep-B_S27	2
201703-TR21-V_S31	14	201703-TR49-V_S35	6	201910-TR21-B_S12	2
201703-TR21-B_S6	13	201708-TR25-B_S21	4	201708-TR53-V_S49	1
201703-TR31-B_S4	13	201910-TR35-V_S54	3	201910-TR13-V_S44	1
201703-Seep-V_S34	11	201910-TR27-V_S45	3	201910-TR27-B_S14	1
201703-TR13-B_S29	10	201910-TR1-B_S10	3	201708-TR53-B_S25	1
201703-TR35-V_S36	9	201910-TR31-V_S52	3	201708-TR31-B_S23	1
201703-TR27-V_S32	9	201910-TR25-V_S47	2	201910-TR25-B_S13	1

Table 8: CARD RGI Strict Matching Output for Round 2 DNA Samples

Predicted Protein ID	Best CARD Hit	% Identity	% aligned	Drug Class	Resistance Mechanism	AMR Gene Family
202403|201703-TR35-V_S36|Pseudomonas_13	FosA	69.85	114.07	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
3684549|201708-TR25-B_S21|Aeromonas_1	OXA-12	94.32	100	cephalosporin; penam	antibiotic inactivation	OXA beta-lactamase
2892593|201703-TR31-V_S38|Pseudomonas_17	FosA	69.85	114.07	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2900577|201703-TR31-V_S38|Proteobacteria_2	tet(A)	99.74	94.1	tetracycline antibiotic	antibiotic efflux	major facilitator superfamily (MFS) antibiotic efflux pump
1295799|201703-TR25-V_S37|Pseudomonas_6	FosA	72.06	102.22	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2754710|201703-TR1-V_S30|Bacillus_1	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2756858|201703-TR1-V_S30|Pseudomonas_2	FosA	73.88	102.22	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
322791|201703-TR21-B_S6|Bacillus_1	BcII	92.16	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
322813|201703-TR21-B_S6|Bacillus_1	FosB	89.52	89.86	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
506933|201703-TR13-B_S29|Bacillus_2	FosB	89.13	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
3377742|201703-TR27-V_S32|Flavobacterium_6	JOHN-1	82.77	97.98	carbapenem; cephalosporin; penam	antibiotic inactivation	JOHN beta-lactamase
2266894|201708-TR35-B_S7|Bacillus_1	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
3133556|201703-TR49-V_S35|Flavobacterium_2	JOHN-1	82.26	100	carbapenem; cephalosporin; penam	antibiotic inactivation	JOHN beta-lactamase
513122|201703-TR13-V_S33|Bacillus_4	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
513122|201703-TR13-V_S33|Bacillus_4	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
515051|201703-TR13-V_S33|Bacillus_2	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2200945|201703-TR31-B_S4|Bacillus_2	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2201033|201703-TR31-B_S4|Bacillus_3	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
2742528|201703-TR1-B_S2|Bacillus_3	FosB	89.13	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
1460707|201703-TR27-B_S5|Bacillus_24	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
1460707|201703-TR27-B_S5|Bacillus_24	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
1460707|201703-TR27-B_S5|Bacillus_24	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
1460806|201703-TR27-B_S5|Bacillus_8	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase

All of the outputs were compiled into a project database that was viewed and interrogated using MySQL Workbench. In this respect, the project database included DNA assemblies, predicted proteins, Hidden Markov Models, Kraken taxonomy, KEGG function, and detailed antibiotic resistance annotations. Overall interpretations of the sequence data processing can be found in Chapter 4.

3.3.5 Future Laboratory Analysis

The Round 3 RNA extract samples will need to be reprocessed and analyzed in hopes of producing viable sequencing data.

Once the sequencing data has been analyzed to a point where pathogens have been identified, PCR could be run on the samples. These pathogen specific PCRs for polio, typhoid, and campylobacter, could be run on the sample extracts, if there is sufficient sample left. Ideally these runs would be used to see if pathogens of interest could be recovered from the samples via PCR as opposed to next-generation sequencing.

3.4 Results

The results discussed in this section are only in reference to Round 2 DNA sequencing samples. Though more results will be able to be obtained by completing further data analysis, the results that have been gathered up to this point are promising. The sequencing data looks to be both substantial and of good quality. Table 28 shows the diversity scores that were calculated by running the Kraken2 output files through the VEGAN program in R. A summary of the data quality scores can be found in Table 10. The average total number of species counts per sample is 10232 and the average richness representing the number of unique species present per sample is 1242. The average Shannon H diversity score, which incorporates both the richness and evenness scores, is 4.4. The average is drawn up by the October 2019 sampling even along with the Seep sampling location, all of which have much higher scores. The Seep is a ritual purging site, meaning that it is a location that is extremely minerally rich and where people visit in order to purge themselves by drinking the water, so it could make sense that the diversity may be less at this sampling location and more closely aligned with human pathogens. The average effective number of species per samples is 191 while the evenness, which is the difference in counts per species present, which ideally is close to one, is 0.6400 for these samples. An evenness score of less than 1 means that not all species are represented with the same counts and that some species are present in higher quantities than others. The quality scores, which are generated by Illumina machines following a sequencing run, demonstrate a respectable average number of DNA reads per sample file at 7314704 and an average quality score of 35.47. Illumina quality scores, or the probability of the wrong base being called in sequencing, above 30 are acceptable. The number of reads per sample are generally well distributed and the contig lengths and other quality statistics are within normal ranges for shotgun metagenomic sequences. However, seeing as each sampling time point was processed differently, though only slightly in some cases, it is difficult to directly compare the samples across seasons or years.

The KMA outputs, where the raw sequencing reads are checked against the entire GenBank database, show that the samples are likely picking up a lot of DNA that would not come up when contigs are assembled and run through microbially focused databases, such as those used with Kraken2. For instance, the samples were able to isolate a lot of mammalian, parasitic, and insect

DNA, which makes sense for a heavily animal and creature impacted region, but is not necessarily important to this analysis, though it does open the door to many future areas of analysis.

The Kraken2 outputs, a summary of which can be found in Appendix #, demonstrate a relatively large range of recovered sequences. The 3,764,994 DNA contig alignments to the Kraken database represent 46 distinct phyla, 86 classes, 193 orders, 440 families, 1554 genera, and 5522 distinct species. A visual representation of a summary of these alignments can be seen in Figure 16.

Table 9: Rosetta Stone Microbial Database Taxonomic Selections

Pathogens by Taxonomic Class	**Count from Microbial Rosetta Stone Database Tables**
Actinobacteria	20
Alphaproteobacteria	44
Bacteroidetes	1
Betaproteobacteria	17
Chlamydia	8
Epsilonproteobacteria	4
Firmicutes	31
Gammaproteobacteria	75
Spirochaetes	3

***this table shows the number of times a unique pathogen was selected from the agency tables that create the Rosetta Stone Microbial Database and does not represent the number of sequences from the Round 2 Samples that aligned to these pathogens.*

Table 10: Diversity Score Average By Sampling Event and Fraction

	Combined	**March**	**August**	**October**	**Viral and free DNA fraction**	**Bacterial Fraction**
Total Counts	10,232	21835.53	4744.42	4420.31	14386.26	5745.76
Richness	1,243	1310.47	940.26	1529.56	1659.33	792.56
Shannon H	4.41	3.10	4.62	5.56	5.19	3.57
$e^{\wedge H}$	191.34	44.31	216.03	318.24	298.67	75.43
Evenness	0.64	0.45	0.70	0.77	0.72	0.56
Quality Scores	35.47	35.60	35.41	35.41	35.44	35.50
Average reads	7,314,705	10,613,574.90	4,970,510.38	6,836,294.09	11,320,997.83	3,196,085.76

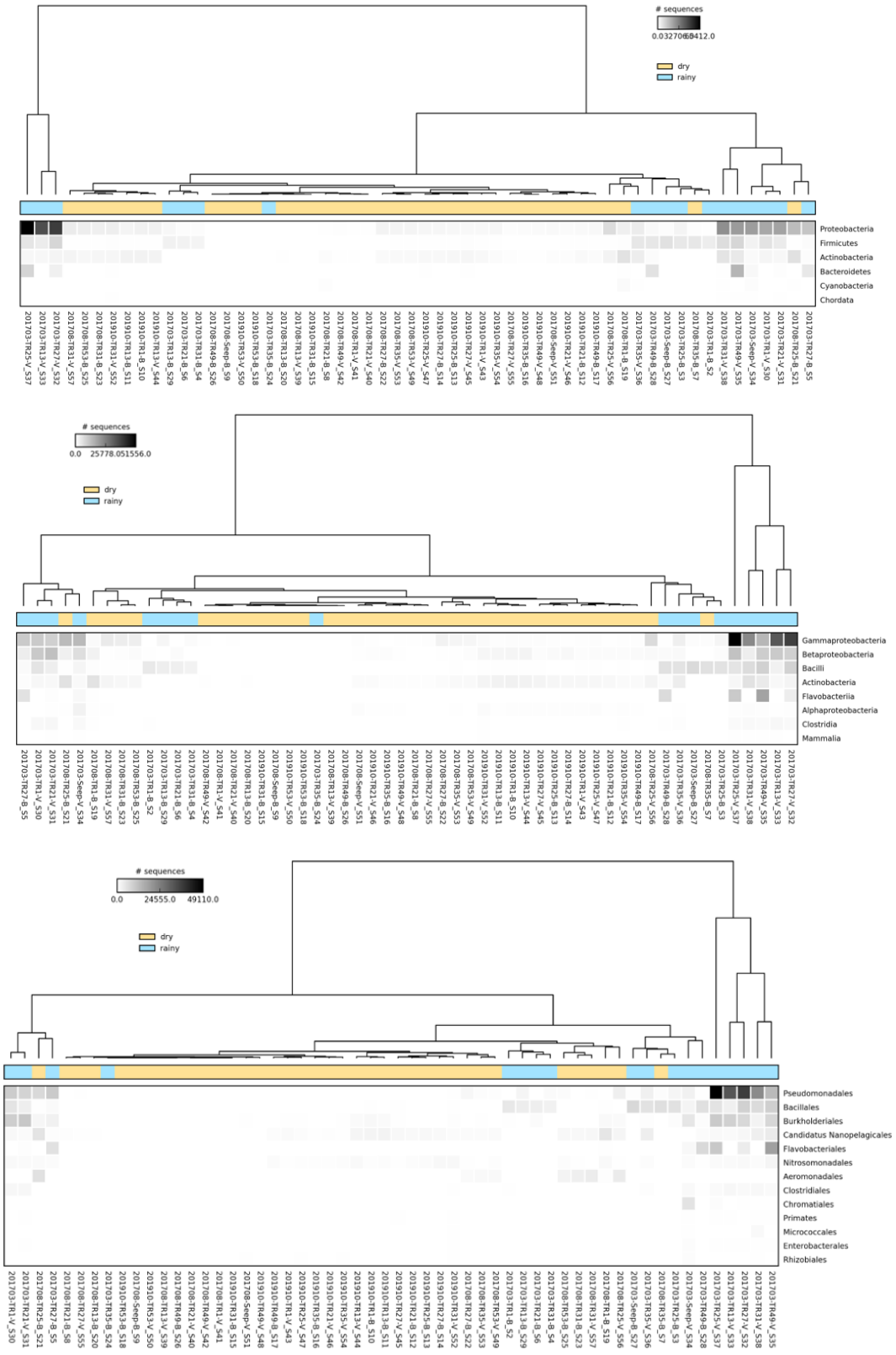


Figure 16: Kraken2 taxonomic classifications summaries for all samples

The Kraken2 alignments that were then sorted by the Rosetta Stone Microbial Database returned 9 different taxonomic classes, 46 genera, and 93 unique species. Figure 17 shows the number of sequences per sample that were sorted and returned for each species identified as a pathogen in the database tables.

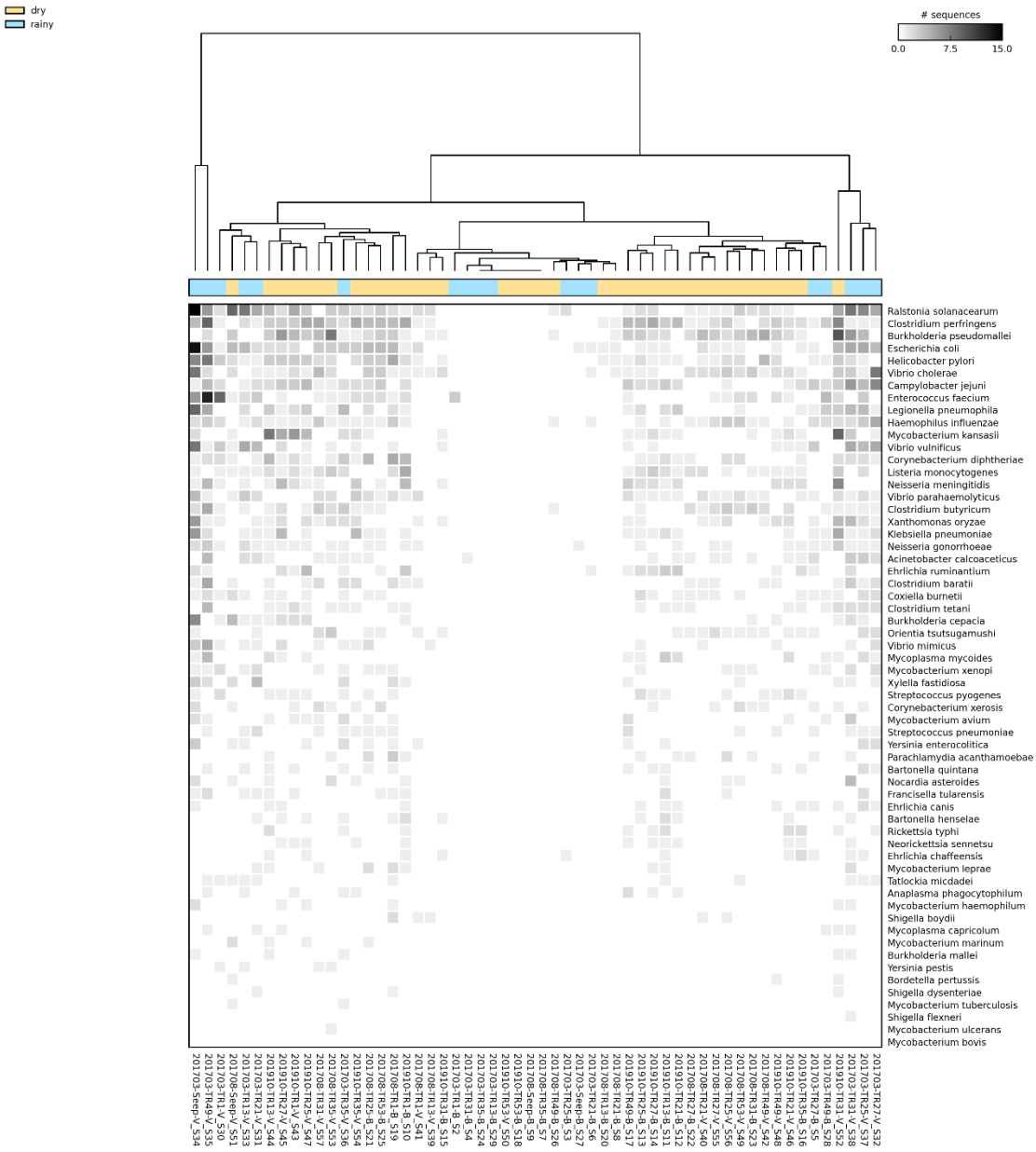


Figure 17: Rosetta Stone Microbial Database pathogens for all sampling events**.** Heatmap depicts the 60 pathogens with the highest number of sequences, but the top 5 highest sequence pathogens have been removed.

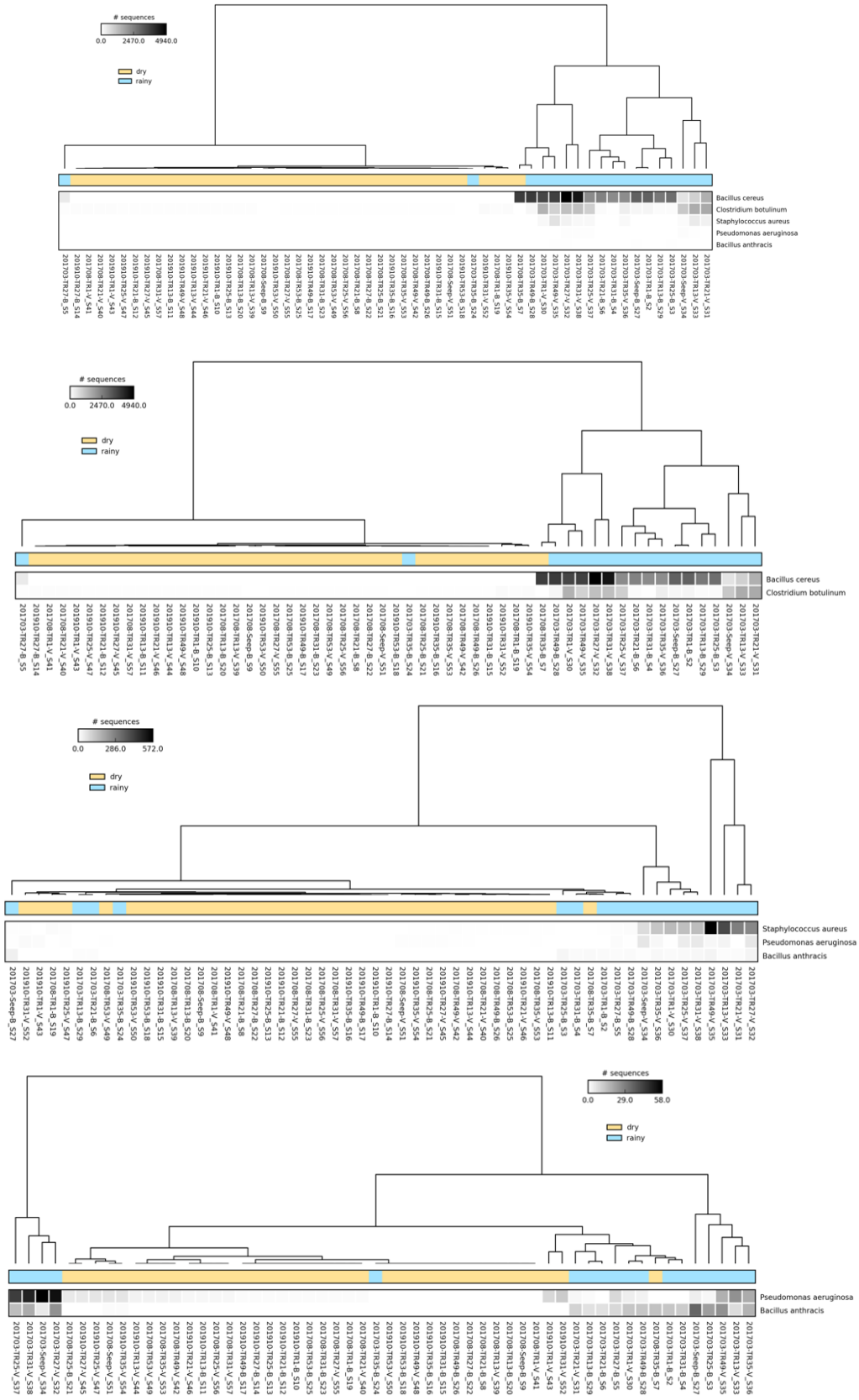


Figure 18: Rosetta Stone Microbial Database pathogens for all sampling events, top 5 pathogens**.** Heatmaps display top 5 pathogens with the highest number of sequences split into heatmap ranges that make the most samples visible in each respective heatmap.

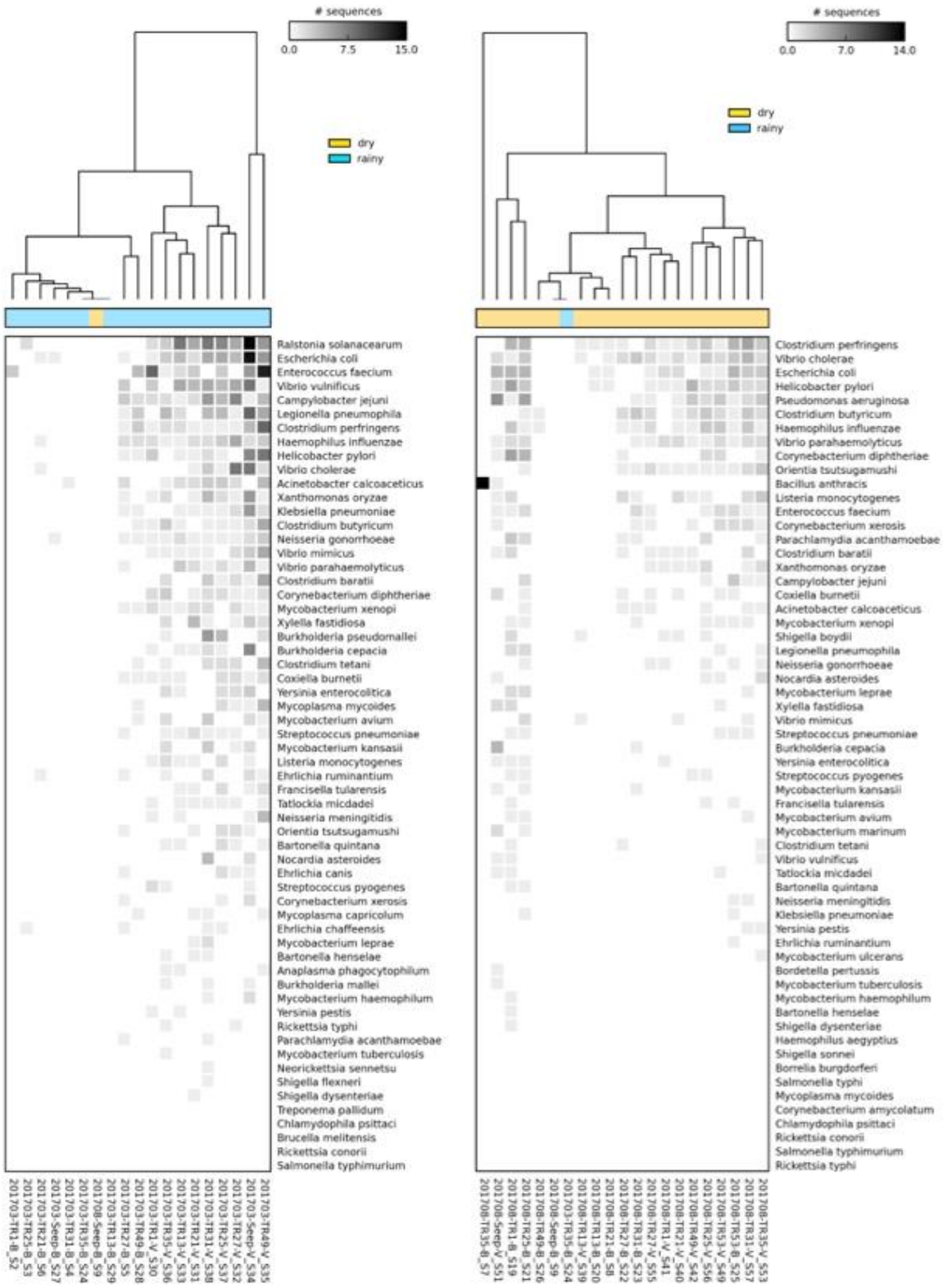


Figure 19: Rosetta Stone Microbial Database top 60 pathogens broken down by sampling events. (left) March sampling, (right) August sampling.

Figure 20: Rosetta Stone Microbial Database top 60 pathogens - October sampling event.

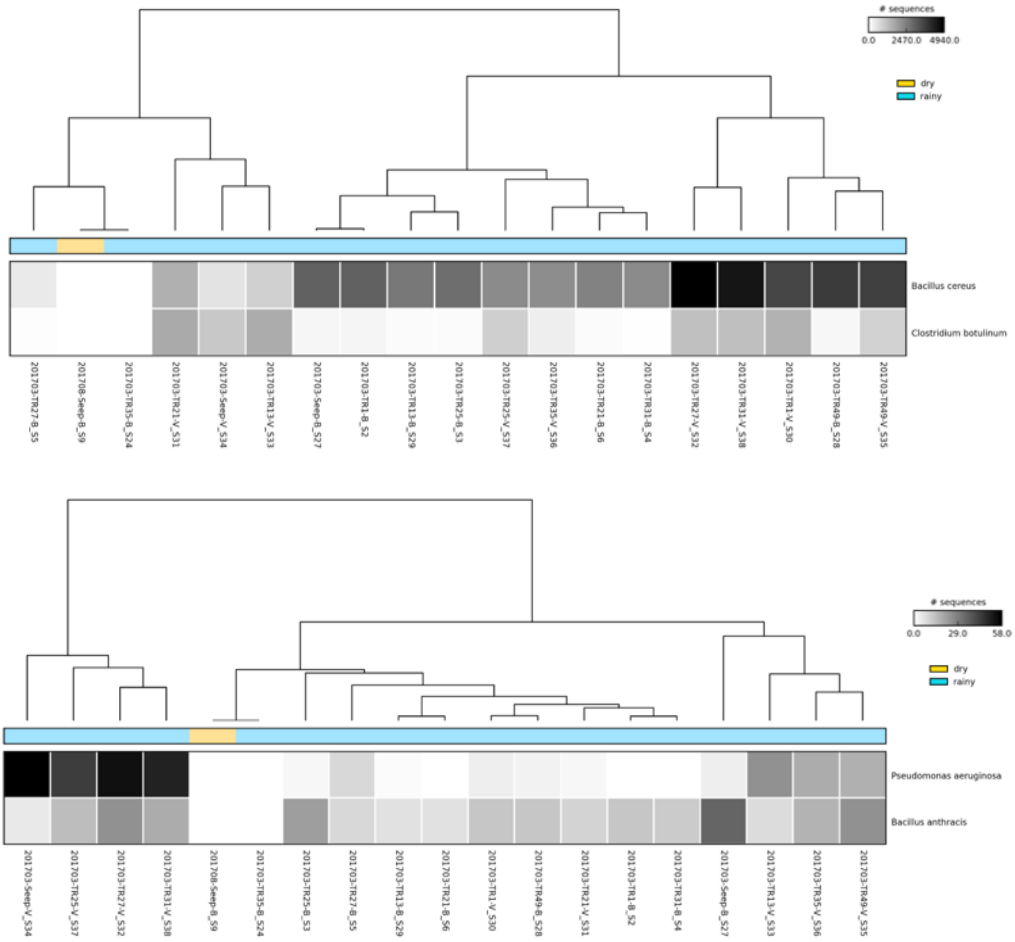


Figure 21: Rosetta Stone Microbial Database top 4 pathogens - March sampling event.

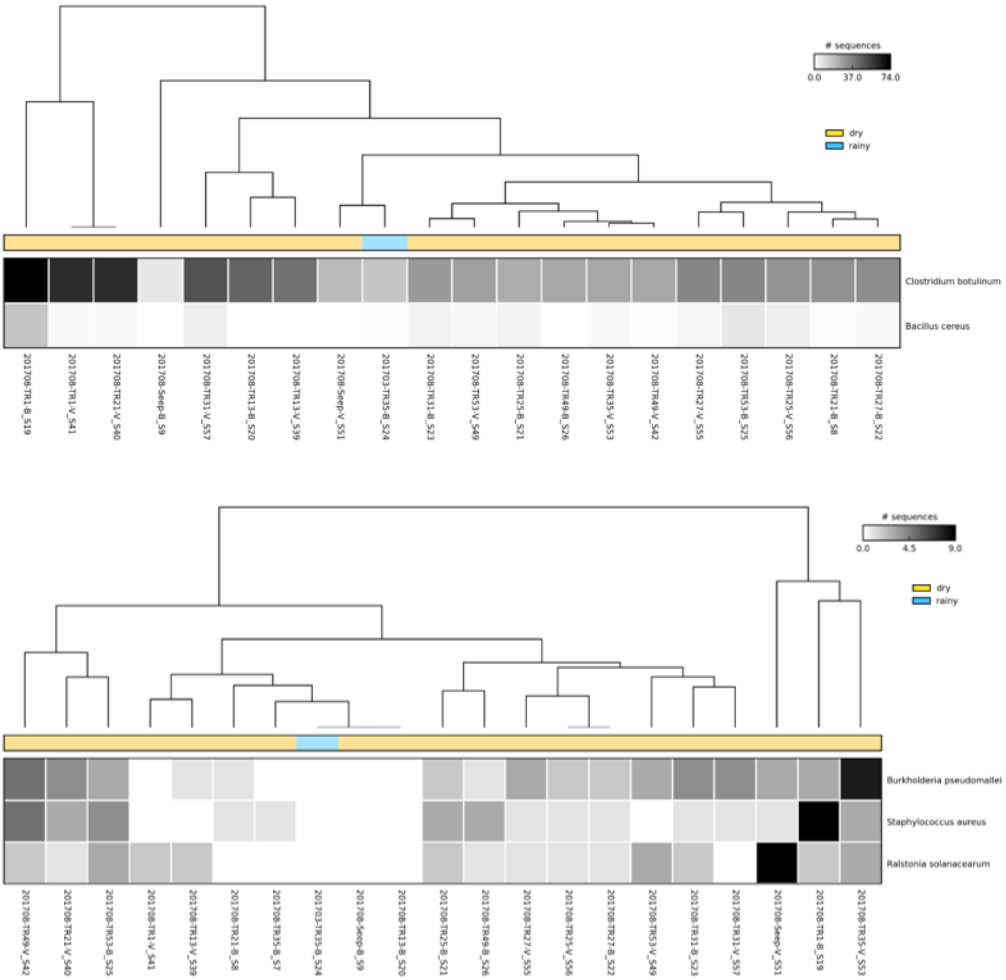


Figure 22: Rosetta Stone Microbial Database top 5 pathogens - August sampling event.

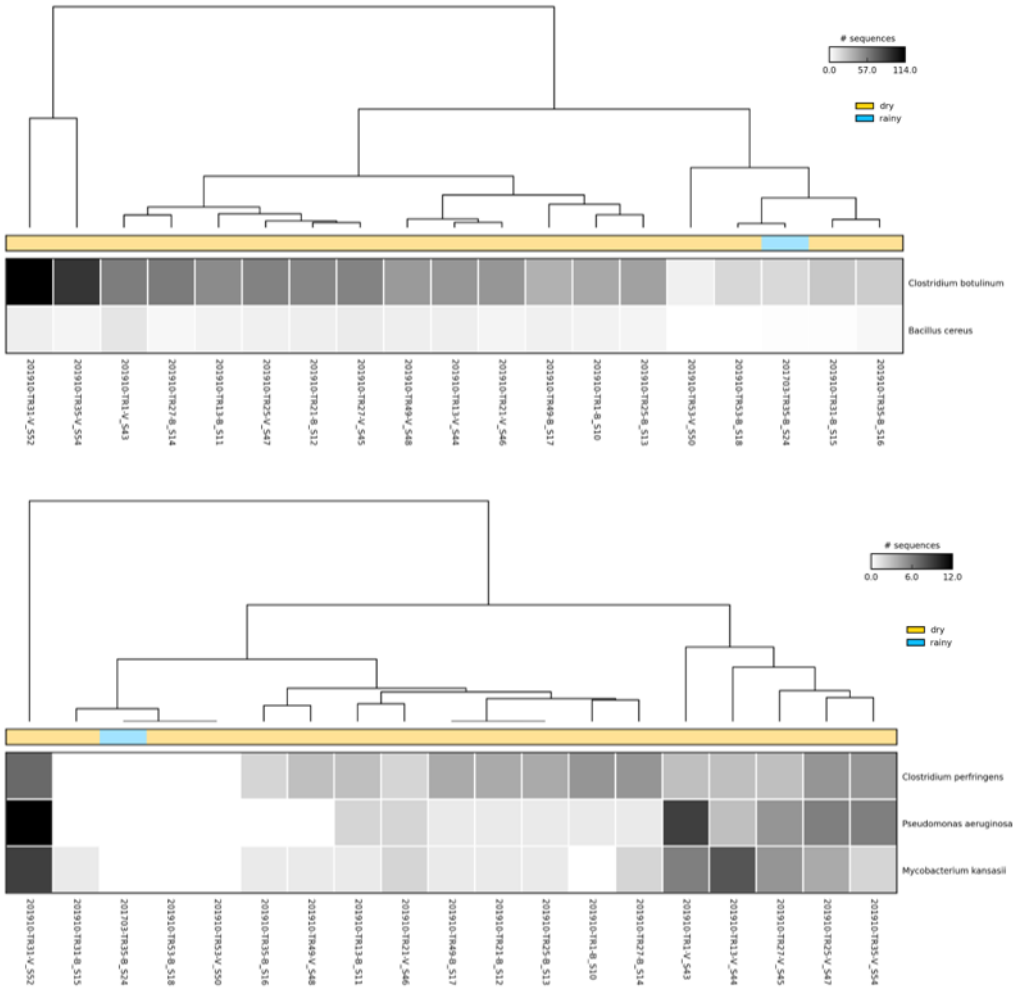


Figure 23: Rosetta Stone Microbial Database top 5 pathogens - October sampling event.

Table 11: 93 Identified Pathogens: Applicability Summaries

Pathogen genus_species	Pathogen Applicability Description
Bacillus cereus	possible human diarrheal pathogen, possibly waterborne
Clostridium botulinum	possible human diarrheal pathogen, possibly waterborne
Staphylococcus aureus	possible human diarrheal pathogen, possibly waterborne
Pseudomonas aeruginosa	possible human diarrheal pathogen, possibly waterborne
Bacillus anthracis	found in region, elephant/animal pathogen, from soil
Ralstonia solanacearum	plant pathogen
Clostridium perfringens	possible human diarrheal pathogen, not waterborne, enterotoxaemia in elephants
Burkholderia pseudomallei	possible human diarrheal pathogen, possibly waterborne
Escherichia coli	possible human diarrheal pathogen, possibly waterborne
Helicobacter pylori	possible human diarrheal pathogen, possibly waterborne
Vibrio cholerae	possible human diarrheal pathogen, possibly waterborne
Campylobacter jejuni	possible human diarrheal pathogen, possibly waterborne
Enterococcus faecium	possible human diarrheal pathogen if immunocompromised, possibly waterborne
Legionella pneumophila	possible human diarrheal pathogen, possibly waterborne, likely in built environments
Haemophilus influenzae	HiB vaccine, meningitis in children
Mycobacterium kansasii	opportunistic human infection
Corynebacterium diphtheriae	diphtheria vaccine available
Vibrio vulnificus	necrotizing fasciitis, waterborne in saltwater
Listeria monocytogenes	foodborne
Clostridium butyricum	possibly pathogenic, used as a probiotic in Asian countries
Neisseria meningitidis	causes meningitis, human to human transmission
Vibrio parahaemolyticus	possible human diarrheal pathogen, waterborne in saltwater
Xanthomonas oryzae	plant pathogen
Klebsiella pneumoniae	opportunistic human infection
Neisseria gonorrhoeae	sexually transmitted infection
Ehrlichia ruminantium	tickborne cattle pathogen
Acinetobacter calcoaceticus	waterborne, but of more concern for AMR in hospital settings
Clostridium baratii	associated with infant deaths and difficulty breathing
Coxiella burnetii	Q Fever, transmitted from animal feces/urine
Clostridium tetani	found in soil
Burkholderia cepacia	often a hospital acquired infection
Orientia tsutsugamushi	scrub typhus disease, mite-borne
Vibrio mimicus	possible human diarrheal pathogen, in freshwater shellfish, turtle eggs, fish
Mycoplasma mycoides	common in ruminants, spread via aerosols
Mycobacterium xenopi	spread via aerosols
Xylella fastidiosa	plant pathogen
Streptococcus pyogenes	Group A Strep, respiratory transmission
Corynebacterium xerosis	endocarditis
Streptococcus pneumoniae	causes pneumonia, possible transmission in recreational waters
Mycobacterium avium	can cause lung disease
Yersinia enterocolitica	possible human diarrheal pathogen, possibly waterborne
Parachlamydia acanthamoebae	resistant to environment, associated with respiratory tract
Francisella tularensis	Causes tularemia
Bartonella quintana	Body louse transmission, can lead to trench fever
Nocardia asteroides	Causes pulmonary infections if immunocompromised
Ehrlichia canis	Canine pathogen
Bartonella henselae	Can cause cat-scratch disease
Rickettsia typhi	Transmitted by flea, can cause murine typhus
Ehrlichia chaffeensis	Transmitted by lone star tick
Neorickettsia sennetsu	Bacteria that can invade human immune system
Mycobacterium leprae	Bacteria that causes leprosy
Tatlockia micdadei	Can cause pneumonia
Anaplasma phagocytophilum	Tick-borne sheep and cattle pathogen
Mycobacterium haemophilum	Can cause skin infections in immunocompromised individuals
Mycoplasma capricolum	Goat, sheep, and cattle pathogen
Shigella boydii	possible human diarrheal pathogen, possibly waterborne
Burkholderia mallei	Human and animal pathogen that can cause formation of glanders

Mycobacterium marinum	waterborne in fish
Yersinia pestis	Flea transmission, plague
Bordetella pertussis	Whooping cough
Shigella dysenteriae	possible human diarrheal pathogen, possibly waterborne
Mycobacterium tuberculosis	TB, respiratory pathogen
Shigella flexneri	possible human diarrheal pathogen, possibly waterborne
Mycobacterium ulcerans	Waterborne human and animal pathogen that can cause ulcers
Brucella melitensis	Sheep pathogen
Rickettsia prowazekii	Transmitted by lice
Rickettsia rickettsii	Tickborne, Rocky Mountain Spotted Fever
Candidatus Liberibacter	Plant pathogen
Ehrlichia ewingii	Tickborne pathogen
Haemophilus ducreyi	Sexually transmitted infection
Salmonella enteritidis	Possible human diarrheal pathogen, likely foodborne
Salmonella typhi	Possible human diarrheal pathogen, possibly waterborne
Salmonella typhimurium	Possible human diarrheal pathogen, likely foodborne
Shigella sonnei	Possible human diarrheal pathogen, possibly waterborne or foodborne
Chlamydophila psittaci	Respiratory pathogen in humans, zoonotic pathogen
Chlamydia trachomatis	Sexually transmitted infection
Borrelia burgdorferi	Lyme Disease, tickborne
Treponema pallidum	Sexually transmitted infection, syphilis
Corynebacterium amycolatum	Opportunistic pathogen, likely hospital acquired
Mycobacterium bovis	cattle respiratory pathogen
Mycobacterium fortuitum	Possibly waterborne, more likely to be a hospital acquired pathogen
Mycobacterium scrofulaceum	Lung infection, cause of scrofula
Mycobacterium abscessus	Possibly spread through water or soil, but more likely airborne
Rhodococcus equi	Horse pathogen, found in dust
Rickettsia conorii	Tickborne pathogen
Chryseobacterium meningosepticum	Pathogen causes meningitis
Ralstonia spp.	Possibly waterborne, opportunistic pathogen
Chlamydophila pneumoniae	Respiratory pathogen, can cause pneumonia
Clostridium difficile	Possible human diarrheal pathogen, possibly waterborne
Acinetobacter baumanii	Opportunistic pathogen, likely hospital acquired
Acinetobacter radioresistans	Opportunistic pathogen, likely hospital acquired
Haemophilus aegyptius	Human pathogen that can cause pink eye

Table 12: Blastn Aligned Contigs

GenBank Division	Name	Contigs	Sum of BLASTN Aligned Nucleotides
BCT	Bacteria	519887	259245808
MAM	Mammal	51838	3768421
ENV	Environmental	12987	2504762
PHG	Phage	3110	1219528
PLN	Plant	1597	142334
INV	Invertebrate	1502	187948
VRT	Vertebrate	649	31370
VRL	Viral	339	106802
PRI	Primate	101	13411

Table 13: Examples of High-Fidelity Matches Using Blastn and Kraken

Sample ID	DNA Length	Percent Alignment	Alignment Length	Bitscore	Blastn	Kraken
201703-TR13-V_S33	107630	99.938	107625	198400	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR13-V_S33	49421	99.913	49422	91005	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR13-V_S33	47465	99.149	42188	75902	Bacillus cereus strain M3, complete sequence	Bacillus cereus FRI-35
201703-TR13-V_S33	47171	99.561	47181	85949	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR13-V_S33	46803	99.907	42785	78788	Bacillus cereus FRI-35, complete genome	Bacillus cereus FRI-35
201703-TR13-V_S33	40250	99.848	40221	73924	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR27-B_S5	165049	99.42	49652	90056	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR27-B_S5	160866	99.869	160852	295800	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR27-B_S5	78450	99.904	78447	144400	Bacillus cereus FRI-35, complete genome	Bacillus cereus FRI-35
201703-TR27-B_S5	72644	99.912	72645	133800	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR27-B_S5	56369	99.89	47372	87190	Bacillus cereus FRI-35, complete genome	Bacillus cereus FRI-35
201703-TR27-B_S5	53954	99.651	53941	98543	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR27-B_S5	52200	99.333	52138	94337	Bacillus cereus FRI-35, complete genome	Bacillus cereus FRI-35
201703-TR27-B_S5	46454	99.763	46455	85174	Bacillus cereus FRI-35, complete genome	Bacillus cereus FRI-35
201703-TR27-B_S5	42756	99.855	42750	78585	Bacillus pacificus strain NCCP 15909 chromosome	Bacillus cereus FRI-35
201703-TR31-B_S4	40658	99.968	40629	74954	Bacillus cereus FRI-35, complete genome	Bacillus cereus FRI-35

In general, all of the sampling events recovered DNA that aligns to sequences that can be classified as pathogens. When looking at the top pathogenic species, based on number of sequences aligned, *Bacillus cereus* and *Clostridium botulinum* are by far the highest in most samples, followed by *Staphylococcus aureus*, *Pseudomonas aeruginosa,* and *Bacillus anthracis*. When the sampling events are broken apart the top five pathogens found change slightly, with the August sampling adding *Burkholderia pseudomallei*. The differences between sampling events could be for a lot of reasons, including but not limited to using BMFS v.1 versus BMFS v.2, number of freeze thaws, and access to necessary equipment.

Table 14: Top CARD AMR Gene Predictions from Round 2 Sequencing Data

Sample ID | Genus	Best CARD Hit	% Identity	% aligned	Drug Class	Resistance Mechanism	AMR Gene Family
202403|201703-TR35-V_S36|Pseudomonas_13	FosA	69.85	114.07	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
3684549|201708-TR25-B_S21|Aeromonas_1	OXA-12	94.32	100	cephalosporin; penam	antibiotic inactivation	OXA beta-lactamase
2892593|201703-TR31-V_S38|Pseudomonas_17	FosA	69.85	114.07	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2900577|201703-TR31-V_S38|Proteobacteria_2	tet(A)	99.74	94.1	tetracycline antibiotic	antibiotic efflux	major facilitator superfamily (MFS) antibiotic efflux pump
1295799|201703-TR25-V_S37|Pseudomonas_6	FosA	72.06	102.22	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2754710|201703-TR1-V_S30|Bacillus_1	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2756858|201703-TR1-V_S30|Pseudomonas_2	FosA	73.88	102.22	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
322791|201703-TR21-B_S6|Bacillus_1	BcII	92.16	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
322813|201703-TR21-B_S6|Bacillus_1	FosB	89.52	89.86	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
506933|201703-TR13-B_S29|Bacillus_2	FosB	89.13	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
3377742|201703-TR27-V_S32|Flavobacterium_6	JOHN-1	82.77	97.98	carbapenem; cephalosporin; penam	antibiotic inactivation	JOHN beta-lactamase
2266894|201708-TR35-B_S7|Bacillus_1	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
3133556|201703-TR49-V_S35|Flavobacterium_2	JOHN-1	82.26	100	carbapenem; cephalosporin; penam	antibiotic inactivation	JOHN beta-lactamase
513122|201703-TR13-V_S33|Bacillus_4	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
513122|201703-TR13-V_S33|Bacillus_4	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
515051|201703-TR13-V_S33|Bacillus_2	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2200945|201703-TR31-B_S4|Bacillus_2	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
2201033|201703-TR31-B_S4|Bacillus_3	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
2742528|201703-TR1-B_S2|Bacillus_3	FosB	89.13	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase
1460707|201703-TR27-B_S5|Bacillus_24	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
1460707|201703-TR27-B_S5|Bacillus_24	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
1460707|201703-TR27-B_S5|Bacillus_24	BcII	91.76	100.39	cephalosporin; penam	antibiotic inactivation	Bc beta-lactamase
1460806|201703-TR27-B_S5|Bacillus_8	FosB	88.41	100	fosfomycin	antibiotic inactivation	fosfomycin thiol transferase

Table 15: High-Coverage Genomes and High-Coverage Plasmids from Round 2 Sequencing Data

SampleDate-Transect-Fraction | GenBank Entry (Template)	KMA Score	Template Length	Template Identity	Template Coverage	Consensus Identity	Consensus Coverage	Coverage Depth
201703-TR35-V_S36.res:198620|CP015852.1 Pseudomonas koreensis strain CRS05-R5 chromosome, complete genome	42607815	5991225	90.2	91.26	98.84	109.58	73.33
201703-TR25-V_S37.res:294|CP010896.1 Pseudomonas fluorescens strain PCL1751, complete genome	7175609	6143950	94.25	94.64	99.59	105.67	11.7
Plasmids							
201703-TR1-V_S30.res:1396|CP016363.1 Bacillus cereus strain M13 plasmid pBCM1303, complete sequence	126534	3088	99.71	100.1	99.61	99.9	48.39
201703-TR1-B_S2.res:334406|AP007214.1 Bacillus cereus NC7401 plasmid pNC4 DNA, complete sequence, strain: NC7401	20167	3091	98.35	98.45	99.9	101.58	12.56
201703-TR21-V_S31.res:405535|CP001286.1 Bacillus cereus AH820 plasmid pAH820_10, complete sequence	9386	10915	94.9	95.8	99.05	104.38	10.17
201703-TR1-V_S30.res:1396|CP023182.1 Bacillus cereus strain CC-1 plasmid p3, complete sequence	33577	5374	91.53	92.98	98.44	107.54	7.37

Table 16: High-Fidelity Matches to AMR Classes Predicted by Protein Blast Search on Round 2 Sequencing Data

AMR Category	Sample ID	% Match	AA Aligned	% Target Aligned	GenBank Description
Beta-Lactamase	201703-TR35-V_S36	98.5	388	1	class C beta-lactamase [Pseudomonas koreensis]
Beta-Lactamase	201703-Seep-V_S34	97.1	385	0.997409	class C beta-lactamase [Pseudomonas sp. R-22-3w-18]
Beta-Lactamase	201703-Seep-V_S34	93.1	217	1	Metallo-beta-lactamase superfamily protein [Candidatus Chazhemtobacterium aquaticus]
Beta-Lactamase	201703-Seep-V_S34	92.3	325	1	CubicO group peptidase (beta-lactamase class C family) [Flavobacterium aquaticum]
Tetracycline	201703-Seep-V_S34	98.5	406	1	tetracycline resistance MFS efflux pump [Flavobacterium sasangense]
Tetracycline	201703-TR31-V_S38	98.3	464	1	Tetracycline resistence protein, metal-tetracycline/H+ antiporter [Bacillus cereus m1293]
Tetracycline	201703-TR31-V_S38	100	399	0.941038	MULTISPECIES: tetracycline efflux MFS transporter Tet(A) [Proteobacteria]
Tetracycline	201703-TR21-B_S6	98.3	464	1	Tetracycline resistence protein, metal-tetracycline/H+ antiporter [Bacillus cereus m1293]
Aminoglycoside	201703-TR35-V_S36	99.4	518	1	bifunctional aminoglycoside phosphotransferase/ATP-binding protein [Pseudomonas koreensis]
Aminoglycoside	201708-TR25-B_S21	92.3	351	1	aminoglycoside phosphotransferase [Actinobacteria bacterium]
Aminoglycoside	201703-TR31-V_S38	99.4	518	1	bifunctional aminoglycoside phosphotransferase/ATP-binding protein [Pseudomonas koreensis]
Aminoglycoside	201703-TR31-V_S38	96.6	268	1	APH(6) family putative aminoglycoside O-phosphotransferase [Psychrobacter sp. FME2]
Chloramphenicol	201703-Seep-V_S34	90.8	207	1	chloramphenicol acetyltransferase [Flavobacterium cucumis]
Chloramphenicol	201703-TR49-V_S35	99	209	1	chloramphenicol acetyltransferase [Flavobacterium anhuiense]
Chloramphenicol	201703-TR49-V_S35	96.7	209	1	chloramphenicol acetyltransferase [Flavobacterium succinicans]

Seeing as shotgun metagenomic sequencing is made up of short reads that are aligned into contigs and then tested against different published DNA sequences to see what fragments align, these hits should be analyzed and not immediately taken as accurate. However, in the case of these samples, where the diversity and quality scores look to be of good quality, and all alignments run through the Rosetta Stone Microbial Database were matches when run with Kraken2, it is likely that these alignments can be trusted. Furthermore, the top five pathogens also make sense in the context of the sampling area. Bacillus anthracis makes sense because anthrax is a known problem in the Chobe Region, leading to many elephant deaths during extreme drought years. *Bacillus cereus* has been shown to be present in many organic matters and in both marine and freshwaters, seeing as the samples are from river water, this is a likely pathogen to find (Bottone, 2010). *Clostridium botulinum* is also known to be found in river waters, so is also a likely and defensible pathogen in these samples (Espelund, 2014). *Staphylococcus aureus* is likely attributable to the human burden in the area, seeing as many people can carry this pathogen (Taylor, 2019). Finally, *Pseudomonas aeruginosa* has been isolated from humans, animals, and the environment, so it definitely makes sense in this context where the sampling sites span a freshwater river that has human, animal, and wastewater impacts (Grosso-Becerra, 2014).

Chapter 4: Discussion

4.1 Limitations

This study has many limitations, some of which can be corrected for in future studies, but other of which are simply inherent to environmental sampling work done on any international scale.

The first major limitation would be that of human error. Having only one person working in country on these samples meant only one set of hands, a fair number of samples, and no other person to be running a double check. This scenario was also compounded by the fact that there was a condensed time on the ground in-country, a lack of lab and boat time due to other experiments and shorter acceptable working days in the region. Additionally, not being a local to the region had its impacts with the extreme heat and drought conditions, dehydration, and accidental concussion due to completing experiments in new working conditions. All of these factors were not ideal for sample management and led to a few of the sample extracts ending up mislabeled. Though these samples were identified once they were brough to the U.W. and were kept separated throughout the entire analysis, relating these back to their original origin samples will be challenging.

A second large limitation is that it will be difficult to make any strong conclusions about what pathogens are present in the river water due to inherent uncertainty withing metagenomic sequencing. For instance, though the sequences may be present in a sample, there is always the possibility of creating chimeras during assembly and then there are additional questions as to whether the pathogens found are infectious and/or waterborne in this case. Additionally, replicates within the same time point do not exist, making comparisons between sampling sites and between time points/seasons almost impossible. However, information that could be gained from this study are pathogens of interest that were found in the samples and could be used in the future either for direct PCR, isolation and growth, or future surveillance study targets.

4.2 Lessons Learned and Applications to Future Environmental Surveillance Studies

Working with collaborators in general comes with challenges and if those collaborators are located on the other side of the world, those challenges can be even more complicated. In this respect, researchers need to put in even more effort on the front-end of international collaborations in order to maximize the chance that everything goes as closely to planned as possible. These planning elements are true for any collaborative effort, and environmental surveillance efforts should not be treated any differently. In actuality, more care should be given to study design and implementation in these instances. Studies of this manner often consist of collection of environmental samples from one region and then analysis in another location by people who are not local to the study region, which can lead to issues of ethics, frustration, and wasted resources. For instance, if locals are not involved both in the sample collection and the sample processing, the value, interpretation, and communication of the data back to the study region can be lost. Additionally, studies organized in this manner are also not sustainable because the skills are not transferred to local staff, making the study management team the only collaborator that is gaining from the interaction. In this respect, locals should be involved from the beginning, including planning steps, and should be trained on the sampling and analysis steps. Though training takes longer, and scientific research benefits from being done quickly, this approach allows the study region to also benefit from the research collaboration. Additional time should be spent discussing with collaborators to design the sampling plan. A non-local researcher

may have an idea as to where samples should be collected based on previous published research, but the environment may have changed since those formal publications or there may be other areas and sampling sites of interest that have come to light. Without a lengthy collaborative process on the front end of an environmental surveillance study to decide the story that both sides want to tell, the samples will be taken and may limit the story that can be told or lack statistical significance. For example, transect points used for sample sites in this study, though they were where collaborators routinely sample and led to a decent coverage of the length of the river, may not have been the best choice since this study was attempting to look at differences between sites. When comparing differences between sites over time and season it is helpful to have duplicates within each sampling event, which was not done in this study. In order to apply these lessons to future sampling efforts, it is important to make sure in-country collaborators and personnel collecting the samples are on the same page and willing to think about the larger picture prior to beginning sampling even if it delays the overall project.

Another important topic to consider if is local collaborators have an interest in the study and if the study is directly applicable and useful to the local region. Often rural and low-income regions may have other more pressing interests that may not be extremely apparent to someone from outside the region. In this respect, consideration should be given to the decision of when to involve the local community. Depending on the project, it may make more sense to involve local community members from the beginning in the decision making process, while in other instances it may be more prudent to develop a plan and then see if buy in can be gained from the local community. Involving the community from the start could lead to a project that is directly applicable and accepted by the community, though it could take much longer to begin and could change from the original project aims, likely adding in more work, time, and need for funding. Developing a project and only involving the local community at the buy in step could lead to a project that is more contained and sticks to the original goals of the researcher, but is likely not as accepted by community members which could lead to difficulties in assistance. For example, in this study access to a boat was challenging during the October 2019 sampling event, which was an unforeseen delay in sampling day and made the actual sampling time more condensed seeing as boat time had to be paid for in cash, which was not the case during the 2017 sampling events. Additionally, access to reagents was a challenge that could have been worked out ahead of time. Hydrochloric acid cannot be carried onto commercial airplanes, so it had to be acquired in country, once borrowed from the local hospital and once bought from a local pool supply store. To ensure that researchers from outside of the region do not blindly impose a project that is not helpful to local citizens, time should be taken to ensure community buy in at the very least. For short term projects, this may add time prior to project execution and data collection but could lead to more assistance and easier field sampling and processing. However, long term environmental surveillance projects should be much more community developed and should incorporate training and involvement of local staff to at least conduct the field sampling.

Study organization is very important and challenges faced in this study could be overcome in future studies by making sure each collaborator has a defined role, piece of the work, and knows what their deliverables are along with what each collaborator wants out of the study in the long term. For example, when working with international collaborators that are based in veterinary medicine and animal rehabilitation as opposed to environmental surveillance and microbiology, projects roles can be confusing. In these situations, special care needs to be given to upfront

discussions about the collaboration to figure out what the trigger points for frustration are for each party. For example, there was confusion over the terms used to refer to samples in this study, which one collaborator thought was just to help identify a sample set, but the other collaborator thought was the other party discrediting their assistance in sample collection. This confusion led to weeks of time that could have been spent moving forward on sample processing, but instead of was spent unpacking the misunderstanding. If the time is taken to work through these potential hurdles prior to beginning the project, it will alleviate emotional distress down the line that is not due to a lack of respect, but rather to misunderstandings of the other parties' interests. While publishable data may be of interest to collaborators in academia, those that are more based in public health practice may place more value in elements that can be used to inform local government and hospital programing. For instance, there may be a lot of publishable datasets determining method development and how to best perform environmental surveillance, whereas public health practice may not be interested until notifiable diseases have been identified. However, these interests could potentially overlap if there is publishable data that is also of extreme current public health practice importance. For instance, environmental surveillance studies that are currently being run for the detection of SARS-CoV-2 in wastewater are also garnering the attention of public health departments due to their current public health necessity and applicability.

Additionally, study design should incorporate planning for seasonal fluctuations when choosing sampling sites. There were two instances in this study where samples either could not be taken or were challenging to collect due directly to their location. One sampling site was not accessible during the extreme drought conditions in October 2019 and another was only accessible through a military base due to low river water levels, making access by boat not possible. In one case the sample was unable to be collected and in the other the collection took much more time, effort, and put samplers in complicated scenarios including talking their way onto and off of a military base, both of which could have been avoided if these sites were not chosen during previous sampling time points. In this respect, sampling site selection and study design need to take into account seasonal impact, especially if a seasonal assessment is part of the overall study design. Finally, if possible, it would be beneficial to design the study to complete as few of the processing steps in the field or all of the processing steps in the field, landing somewhere in between was a challenge faced in this study. A lack of necessary equipment led to issues with processing steps varying between time points.

Table 17: Breakdown of Sample Processing Steps for Ideal Workflows

	Field	Kasane Laboratory	Seattle Laboratory
Processing step breakdown that occurred in this study	• Sample collection & filtration	• Sample elution & secondary concentration • Sample extraction	• Sample library preparation & sequencing • Data analysis
Processing step breakdown that would be easiest and fastest for Seattle collaborators	• Sample collection & filtration		• Sample elution & secondary concentration • Sample extraction • Sample library preparation & sequencing • Data analysis
Processing step breakdown that would be ideal for Kasane collaborators, though not currently possible due to lack of resources and equipment	• Sample collection & filtration	• Sample elution & secondary concentration • Sample extraction • Sample library preparation & sequencing • Data analysis	

Some of these logistical challenges could be overcome in future studies by making sure there is a positive working relationship with the local collaborators along with open lines of communication. Elements of a positive working relationship that are necessary to projects of this manner are open lines of communication, understanding of each collaborator's interests, and respect for each other's differences in research and management styles. Efforts that can be taken to help ensure these positive working relationships are to communicate openly and frequently and to formally assign specific roles, all of which can be challenging when working across time zones and working styles.

Finally, sample shipment is a key topic that should be discussed at length prior to the start of sampling to avoid any major delays, as were experienced in this study. Shipping anything internationally is challenging, and shipping samples is extremely difficult, especially when unknowns about the sample characteristics are factored in. In this study scenario, samples could have been shipped as just environmental river water or as extracts. However, any form of filtration or concentration is not an allowable type of shipment, even with import permitting. In this case, it was not only the CDC import permitting, but rather the exportation permitting out of Botswana that needed to be considered and contributed to the unforeseen challenge of not being able to ship samples. Shipping challenges can be prevented in future surveillance efforts by working with local partners to analyze the shipping situation prior to starting the collection of samples and possibly either modifying the sampling plan to ship during a different experimental step or plan on having personnel always hand-carry the samples when they are traveling between the countries involved in the study.

4.3 Conclusions

Though not all the data analysis has been completed, this study, at the point where it is now, is still an accomplishment. This pilot study has shown that the BMFS is able to be combined with downstream next-generation sequencing technologies to identify bacteria and viruses. Prior to this study being done this was only a hypothesized method and had not been developed or field tested. A lot of the work for this study was in figuring out the best ways to conduct

environmental surveillance internationally in rural areas and combine well-established sampling techniques for wastewater-impacted surface waters with extraction, amplification, library preparation, and sequencing techniques that had not previously been used in that manner. However, moving forward it is apparent that many steps need to be put into place in order for future environmental surveillance efforts to be conducted smoothly.

Sequencing has inherent downfalls and unknowns, especially when it comes to environmental metagenomics, since environmental samples are likely to contain sequences that are not yet published in databases. Though environmental samples are becoming better characterized in sequencing databases every year, there is still a lot left to discover.

The possible policy implications from these findings, if this effort was ever expanded beyond a pilot study could vary according to many factors. It may be challenging to implement findings or adopt environmental surveillance techniques due to the resource limitations in Botswana, the difference in health policy priorities based on the current HIV/AIDS burden, and the lack of population concentration in the study area. Kasane falls within the Chobe region which has a significantly smaller and more rural population compared to the capital city of Gaborone. Seeing as the political offices and health ministries are located in the capital, it may be a challenge to convince policy makers that environmental surveillance efforts are worth the effort and funds. On the other hand, there may be hope for adoption of these techniques or valuing the importance of the data due to the fact that animal conservation is extremely important for the tourism industry, upon which Botswana relies heavily. In this respect, if the data could show impacts of animal pathogens this research could be important to protecting species and understanding future outbreaks, similar to the elephant die off that Botswana is currently facing. Additionally, based on the Botswana national statistics reports, it is clear that diarrheal diseases are of great importance in Botswana due to the large health burden, making adoption of additional tracking methods more likely. Finally, if this research were able to find a pathogen that was of interest to an NGO or the WHO, there would be an ability to acquire outside funding, which would increase the chance of adoption.

References

2011 Population and Housing Census (Results).pdf. (n.d.). Retrieved August 24, 2020, from http://www.statsbots.org.bw/sites/default/files/publications/2011%20Population%20and%20Housing%20Census%20%28Results%29.pdf

2-Channel SBS Technology | Faster sequencing and data acquisition. (n.d.). Retrieved September 1, 2020, from https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html

Abia, A. L. K., Alisoltani, A., Keshri, J., & Ubomba-Jaswa, E. (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, 326–334. https://doi.org/10.1016/j.scitotenv.2017.10.322

Abia, A. L. K., James, C., Ubomba-Jaswa, E., & Benteke Momba, M. N. (2017). Microbial Remobilisation on Riverbed Sediment Disturbance in Experimental Flumes and a Human-Impacted River: Implication for Water Resource Management and Public Health in Developing Sub-Saharan African Countries. International Journal of Environmental Research and Public Health, 14(3). https://doi.org/10.3390/ijerph14030306

Africa: Botswana—The World Factbook—Central Intelligence Agency. (n.d.). Retrieved July 5, 2020, from https://www.cia.gov/library/publications/the-world-factbook/geos/bc.html

Alexander, KA, & Blackburn, J. (2013). Overcoming barriers in evaluating outbreaks of diarrheal disease in resource poor settings: Assessment of recurrent outbreaks in Chobe District, Botswana. BMC Public Health, 13, 775. https://doi.org/10.1186/1471-2458-13-775

Alexander, Kathleen A., Carzolio, M., Goodin, D., & Vance, E. (2013a). Climate change is likely to worsen the public health threat of diarrheal disease in Botswana. International Journal of Environmental Research and Public Health, 10(4), 1202–1230. https://doi.org/10.3390/ijerph10041202

Alexander, Kathleen A., & Godrej, A. (2015a). Greywater Disposal Practices in Northern Botswana—The Silent Spring? International Journal of Environmental Research and Public Health, 12(11), 14529–14540. https://doi.org/10.3390/ijerph121114529

Alexander, Kathleen A., Heaney, A. K., & Shaman, J. (2018). Hydrometeorology and flood pulse dynamics drive diarrheal disease outbreaks and increase vulnerability to climate change in surface-water-dependent populations: A retrospective analysis. PLOS Medicine, 15(11), e1002688. https://doi.org/10.1371/journal.pmed.1002688

Alexander, Kathleen A., Larsen, M. H., Robbe-Austerman, S., Stuber, T. P., & Camp, P. M. (2016). Draft Genome Sequence of the Mycobacterium tuberculosis Complex Pathogen M. mungi, Identified in a Banded Mongoose (Mungos mungo) in Northern Botswana. Genome Announcements, 4(4). https://doi.org/10.1128/genomeA.00471-16

Alexander, Kathleen A., Laver, P. N., Michel, A. L., Williams, M., van Helden, P. D., Warren, R. M., & Gey van Pittius, N. C. (2010). Novel Mycobacterium tuberculosis Complex Pathogen, M. mungi. Emerging Infectious Diseases, 16(8), 1296–1299. https://doi.org/10.3201/eid1608.100314
Alexander, Kathleen A., Laver, P. N., Williams, M. C., Sanderson, C. E., Kanipe, C., & Palmer, M. V. (2018a). Pathology of the Emerging Mycobacterium tuberculosis Complex Pathogen, Mycobacterium mungi, in the Banded Mongoose (Mungos mungo). Veterinary Pathology, 55(2), 303–309. https://doi.org/10.1177/0300985817741730

Alexander, Kathleen A., Sanderson, C. E., Larsen, M. H., Robbe-Austerman, S., Williams, M. C., & Palmer, M. V. (2016). Emerging Tuberculosis Pathogen Hijacks Social Communication Behavior in the Group-Living Banded Mongoose (Mungos mungo). MBio, 7(3). https://doi.org/10.1128/mBio.00281-16

Alexander, Kathleen Anne, Herbein, J., & Zajac, A. (2012). The Occurrence of Cryptosporidium and Giardia Infections Among Patients Reporting Diarrheal Disease in Chobe District, Botswana. Advances in Infectious Diseases, 02(04), 143. https://doi.org/10.4236/aid.2012.24023

Antillón, M., Warren, J. L., Crawford, F. W., Weinberger, D. M., Kürüm, E., Pak, G. D., Marks, F., & Pitzer, V. E. (2017a). The burden of typhoid fever in low- and middle-income countries: A meta-regression approach. PLoS Neglected Tropical Diseases, 11(2). https://doi.org/10.1371/journal.pntd.0005376

Azeem, S., Bengis, R., Aarde, R. V., & Bastos, A. D. S. (2020). Mass Die-Off of African Elephants in Botswana: Pathogen, Poison or a Perfect Storm? African Journal of Wildlife Research, 50(1). https://doi.org/10.3957/056.050.0149

Bačnik, K., Kutnjak, D., Pecman, A., Mehle, N., Tušek Žnidarič, M., Gutiérrez Aguirre, I., & Ravnikar, M. (2020). Viromics and infectivity analysis reveal the release of infective plant viruses from wastewater into the environment. Water Research, 177, 115628. https://doi.org/10.1016/j.watres.2020.115628

Bank-Wolf, B. R., König, M., & Thiel, H.-J. (2010). Zoonotic aspects of infections with noroviruses and sapoviruses. Veterinary Microbiology, 140(3–4), 204–212. https://doi.org/10.1016/j.vetmic.2009.08.021

Bibby, K., Crank, K., Greaves, J., Li, X., Wu, Z., Hamza, I. A., & Stachler, E. (2019). Metagenomics and the development of viral water quality tools. Npj Clean Water, 2(1), 1–13. https://doi.org/10.1038/s41545-019-0032-3

Bibby, K., & Peccia, J. (2013a). Identification of Viral Pathogen Diversity in Sewage Sludge by Metagenome Analysis. Environmental Science & Technology, 47(4), 1945–1951. https://doi.org/10.1021/es305181x

BLAST. (n.d.). Retrieved September 1, 2020, from https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/BLAST/nucleotide_blast.html

Bofill-Mas, S., & Rusiñol, M. (2020). Recent trends on methods for the concentration of viruses from water samples. Current Opinion in Environmental Science & Health, 16, 7–13. https://doi.org/10.1016/j.coesh.2020.01.006

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., Philippon, A., Allesoe, R. L., Rebelo, A. R., Florensa, A. F., Fagelhauer, L., Chakraborty, T., Neumann, B., Werner, G., Bender, J. K., Stingl, K., Nguyen, M., Coppens, J., Xavier, B. B., … Aarestrup, F. M. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy, dkaa345. https://doi.org/10.1093/jac/dkaa345

Botswana emergency polio campaign: Wake-Up Call for the World. (n.d.). UNICEF. Retrieved May 26, 2020, from https://www.unicef.org/media/media_21796.html
Bottone, E. J. (2010). Bacillus cereus, a Volatile Human Pathogen. Clinical Microbiology Reviews, 23(2), 382–398. https://doi.org/10.1128/CMR.00073-09

Botswana Environmental Statistics Water and Climate Digests 2017. (2017). Botswana Central Statistics Office. Retrieved August 24, 2020, from http://www.statsbots.org.bw/sites/default/files/publications/Botswana%20Environment%20Statistics%20Water%20%20climate%20Digest%202017.pdf

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59–60. https://doi.org/10.1038/nmeth.3176

Botswana: WHO and UNICEF estimates of immunization coverage: 2019 revision. Bwa.pdf. (n.d.). Retrieved August 24, 2020, from https://www.who.int/immunization/monitoring_surveillance/data/bwa.pdf?ua=1

Campylobacter (Campylobacteriosis) | Campylobacter | CDC. (2019, December 23). https://www.cdc.gov/campylobacter/index.html

CDC - Schistosomiasis—General Information. (2017, May 2). https://www.cdc.gov/parasites/schistosomiasis/gen_info/faqs.html

Characterisation of the sewage virome: Comparison of NGS tools and occurrence of significant pathogens | Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/j.scitotenv.2020.136604

CARD: The Comprehensive Antibiotic Resistance Database. (n.d.). Retrieved July 20, 2020, from https://card.mcmaster.ca/analyze/rgi

Chen, S.-C., Tan, L.-B., Huang, L.-M., & Chen, K.-T. (2012). Rotavirus infection and the current status of rotavirus vaccines. Journal of the Formosan Medical Association, 111(4), 183–193. https://doi.org/10.1016/j.jfma.2011.09.024

Clausen, P. T. L. C., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics, 19(1), 307. https://doi.org/10.1186/s12859-018-2336-6

Cook, N., Bridger, J., Kendall, K., Gomara, M. I., El-Attar, L., & Gray, J. (2004). The zoonotic potential of rotavirus. The Journal of Infection, 48(4), 289–302. https://doi.org/10.1016/j.jinf.2004.01.018

Diarrhoeal disease. (n.d.). World Health Organization. Retrieved May 31, 2018, from http://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease

Dinsdale, E. A., Edwards, R. A., Bailey, B. A., Tuba, I., Akhter, S., McNair, K., Schmieder, R., Apkarian, N., Creek, M., Guan, E., Hernandez, M., Isaacs, K., Peterson, C., Regh, T., & Ponomarenko, V. (2013). Multivariate Analysis of Functional Metagenomes. Frontiers in Genetics, 4. https://doi.org/10.3389/fgene.2013.00041

Doggett, N. A., Stubben, C. J., Chertkov, O., Bruce, D. C., Detter, J. C., Johnson, S. L., & Han, C. S. (2013). Complete Genome Sequence of Bacillus thuringiensis Serovar Israelensis Strain HD-789. Genome Announcements, 1(6). https://doi.org/10.1128/genomeA.01023-13

Ecker, D. J., Sampath, R., Willett, P., Wyatt, J. R., Samant, V., Massire, C., Hall, T. A., Hari, K., McNeil, J. A., Büchen-Osmond, C., & Budowle, B. (2005b). The Microbial Rosetta Stone Database: A compilation of global and emerging infectious microorganisms and bioterrorist threat agents. BMC Microbiology, 5. https://doi.org/10.1186/1471-2180-5-19

Epizootic Ulcerative Syndrome, Botswana. Retrieved August 16, 2020, from https://www.oie.int/wahis_2/public/wahid.php/Reviewreport/Review?page_refer=MapFullEventReport&reportid=35358

Espelund, M., & Klaveness, D. (2014). Botulism outbreaks in natural environments – an update. Frontiers in Microbiology, 5. https://doi.org/10.3389/fmicb.2014.00287

Fagnant, C. S., Beck, N. K., Yang, M.-F., Barnes, K. S., Boyle, D. S., & Meschke, J. S. (2014). Development of a novel bag-mediated filtration system for environmental recovery of poliovirus. Journal of Water and Health, 12(4), 747–754. https://doi.org/10.2166/wh.2014.032

Fagnant, C. S., Sánchez-Gonzalez, L. M., Zhou, N. A., Falman, J. C., Eisenstein, M., Guelig, D., Ockerman, B., Guan, Y., Kossik, A. L., Linden, Y. S., Beck, N. K., Wilmouth, R., Komen, E., Mwangi, B., Nyangao, J., Shirai, J. H., Novosselov, I., Borus, P., Boyle, D. S., & Meschke, J. S. (2018). Improvement of the Bag-Mediated Filtration System for Sampling Wastewater and Wastewater-Impacted Waters. Food and Environmental Virology, 10(1), 72–82. https://doi.org/10.1007/s12560-017-9311-7

Fagnant, C. S., Toles, M., Zhou, N. A., Powell, J., Adolphsen, J., Guan, Y., Ockerman, B., Shirai, J. H., Boyle, D. S., Novosselov, I., & Meschke, J. S. (2017). Development of an elution

device for ViroCap virus filters. Environmental Monitoring and Assessment, 189(11). https://doi.org/10.1007/s10661-017-6258-y

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., & Raes, J. (2015). Metagenomics meets time series analysis: Unraveling microbial community dynamics. Current Opinion in Microbiology, 25, 56–66. https://doi.org/10.1016/j.mib.2015.04.004

Field Listing: Drinking water source—The World Factbook—Central Intelligence Agency. (n.d.). Retrieved August 24, 2020, from https://www.cia.gov/library/publications/the-world-factbook/fields/361.html

Field Listing: Sanitation facility access—The World Factbook—Central Intelligence Agency. (n.d.). Retrieved August 24, 2020, from https://www.cia.gov/library/publications/the-world-factbook/fields/398.html

Food and Agricultural Organization of the United Nations (FAO). What you need to know about Epizootic ulcerative syndrome (EUS)—An extension brochure. (n.d.). 36. http://www.fao.org/3/i0777e/i0777e.pdf

Fox, J. T., & Alexander, K. A. (2015a). Spatiotemporal Variation and the Role of Wildlife in Seasonal Water Quality Declines in the Chobe River, Botswana. PLoS ONE, 10(10). https://doi.org/10.1371/journal.pone.0139936

Frontiers | The Changing Face of Water: A Dynamic Reflection of Antibiotic Resistance Across Landscapes | Microbiology. (n.d.). Retrieved August 16, 2020, from https://www.frontiersin.org/articles/10.3389/fmicb.2018.01894/full

GenBank Overview. (n.d.). Retrieved September 1, 2020, from https://www.ncbi.nlm.nih.gov/genbank/

General outpatient and Preventive Health Statistics Brief 2016 .pdf. (n.d.). Retrieved August 24, 2020, from http://www.statsbots.org.bw/sites/default/files/publications/General%20outpatient%20and%20Preventive%20Health%20Statistics%20Brief%202016%20.pdf

Genthe, B., Kapwata, T., Le Roux, W., Chamier, J., & Wright, C. Y. (2018). The reach of human health risks associated with metals/metalloids in water and vegetables along a contaminated river catchment: South Africa and Mozambique. Chemosphere, 199, 1–9. https://doi.org/10.1016/j.chemosphere.2018.01.160

Global Diarrhea Burden | Global Water, Sanitation and Hygiene | Healthy Water | CDC. (2018, November 9). https://www.cdc.gov/healthywater/global/diarrhea-burden.html

Guitor, A. K., Raphenya, A. R., Klunk, J., Kuch, M., Alcock, B., Surette, M. G., McArthur, A. G., Poinar, H. N., & Wright, G. D. (2019). Capturing the Resistome: A Targeted Capture Method

To Reveal Antibiotic Resistance Determinants in Metagenomes. Antimicrobial Agents and Chemotherapy, 64(1). https://doi.org/10.1128/AAC.01324-19

Heaney, A. K., Shaman, J., & Alexander, K. A. (2019). El Niño-Southern oscillation and under-5 diarrhea in Botswana. Nature Communications, 10(1), 5798. https://doi.org/10.1038/s41467-019-13584-6

HIV and AIDS in Botswana. (2015, July 21). Avert. https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/botswana

HMMER. (n.d.). Retrieved July 20, 2020, from http://hmmer.org/

Huchzermeyer, K. D. A., & van der Waal, B. C. W. (2012). Epizootic ulcerative syndrome: Exotic fish disease threatens Africa's aquatic ecosystems. Journal of the South African Veterinary Association, 83(1), 204.

Illumina DNA Prep Documentation. (n.d.). Retrieved September 1, 2020, from https://support.illumina.com/sequencing/sequencing_kits/illumina-dna-prep/documentation.html

Illumina DNA Prep | Flexibility for many whole-genome sequencing applications. (n.d.). Retrieved September 1, 2020, from https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-dna-flex.html

Illumina. Understanding Illumina Quality Scores. (2014). https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

Jobbins, S. E., Sanderson, C. E., & Alexander, K. A. (2014). Leptospira interrogans at the Human–Wildlife Interface in Northern Botswana: A Newly Identified Public Health Threat. Zoonoses and Public Health, 61(2), 113–123. https://doi.org/10.1111/zph.12052

Jobbins, Sarah E., & Alexander, K. A. (2015). Evidence of Leptospira sp. infection among a diversity of African wildlife species: Beyond the usual suspects. Transactions of The Royal Society of Tropical Medicine and Hygiene, 109(5), 349–351. https://doi.org/10.1093/trstmh/trv007

Jobbins, Sarah Elizabeth, & Alexander, K. A. (2015). From whence they came—Antibiotic-resistant escherichia coli in african wildlife. Journal of Wildlife Diseases, 51(4), 811–820. https://doi.org/10.7589/2014-11-257

Jori, F., Alexander, K. A., Mokopasetso, M., Munstermann, S., Moagabo, K., & Paweska, J. T. (2015). Serological Evidence of Rift Valley Fever Virus Circulation in Domestic Cattle and African Buffalo in Northern Botswana (2010–2011). Frontiers in Veterinary Science, 2. https://doi.org/10.3389/fvets.2015.00063

Kim, J.-H., Im, J., Parajulee, P., Holm, M., Cruz Espinoza, L. M., Poudyal, N., Mogeni, O. D., & Marks, F. (2019). A Systematic Review of Typhoid Fever Occurrence in Africa. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 69(Suppl 6), S492–S498. https://doi.org/10.1093/cid/ciz525

Leptospirosis Fact Sheet | Leptospirosis | CDC. (2017, October 16). https://www.cdc.gov/leptospirosis/resources/leptospirosis-fact-sheet.html

Lothe, A., Zulu, N., Øyhus, A. O., Kjetland, E. F., & Taylor, M. (2018). Treating schistosomiasis among South African high school pupils in an endemic area, a qualitative study. BMC Infectious Diseases, 18(1), 239. https://doi.org/10.1186/s12879-018-3102-0

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome Biology, 21(1), 103. https://doi.org/10.1186/s13059-020-02014-2

Martínez-Puchol, S., Rusiñol, M., Fernández-Cassi, X., Timoneda, N., Itarte, M., Andrés, C., Antón, A., Abril, J. F., Girones, R., & Bofill-Mas, S. (2020). Characterisation of the sewage virome: Comparison of NGS tools and occurrence of significant pathogens. Science of The Total Environment, 713, 136604. https://doi.org/10.1016/j.scitotenv.2020.136604

Medley, S., Ponder, M., & Alexander, K. A. (2020). Anthropogenic landscapes increase Campylobacter jejuni infections in urbanizing banded mongoose (Mungos mungo): A one health approach. PLoS Neglected Tropical Diseases, 14(3). https://doi.org/10.1371/journal.pntd.0007888

Mladenov, N., Strzepek, K., & Serumola, O. M. (2005). Water quality assessment and modeling of an effluent-dominated stream, the Notwane River, Botswana. Environmental Monitoring and Assessment, 109(1–3), 97–121. https://doi.org/10.1007/s10661-005-5842-8

Mogeni, O. D., Cruz Espinoza, L. M., Im, J., Panzner, U., Toy, T., Pak, G. D., Haselbeck, A., Ramani, E., Schütt-Gerowitt, H., Jacobs, J., Metila, O. L., Adewusi, O. J., Okeke, I. N., Ogunleye, V. I., Owusu-Dabo, E., Rakotozandrindrainy, R., Soura, A. B., Teferi, M., Roy, K. C., … Marks, F. (2019). The Monitoring and Evaluation of a Multicountry Surveillance Study, the Severe Typhoid Fever in Africa Program. Clinical Infectious Diseases, 69(Supplement_6), S510–S518. https://doi.org/10.1093/cid/ciz597

National Health Policy: Towards a Healthier Botswana. (August 2011). Republic of Botswana. Retrieved August 24, 2020, from https://extranet.who.int/nutrition/gina/sites/default/files/BWA%202011%20National%20Health%20Policy.pdf

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017a). metaSPAdes: A new versatile metagenomic assembler. Genome Research, 27(5), 824–834. https://doi.org/10.1101/gr.213959.116

Oakeson, K. F., Wagner, J. M., Mendenhall, M., Rohrwasser, A., & Atkinson-Dunn, R. (2017). Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerging Infectious Diseases, 23(9), 1441. https://doi.org/10.3201/eid2309.170416

O'Brien, E., Nakyazze, J., Wu, H., Kiwanuka, N., Cunningham, W., Kaneene, J. B., & Xagoraraki, I. (2017). Viral diversity and abundance in polluted waters in Kampala, Uganda. Water Research, 127, 41–49. https://doi.org/10.1016/j.watres.2017.09.063

Partridge, S. R., Di Pilato, V., Doi, Y., Feldgarden, M., Haft, D. H., Klimke, W., Kumar-Singh, S., Liu, J.-H., Malhotra-Kumar, S., Prasad, A., Rossolini, G. M., Schwarz, S., Shen, J., Walsh, T., Wang, Y., & Xavier, B. B. (2018). Proposal for assignment of allele numbers for mobile colistin resistance (mcr) genes. Journal of Antimicrobial Chemotherapy, 73(10), 2625–2630. https://doi.org/10.1093/jac/dky262

Patel, J. C. (2019). Surveillance to Track Progress Toward Polio Eradication—Worldwide, 2017–2018. MMWR. Morbidity and Mortality Weekly Report, 68. https://doi.org/10.15585/mmwr.mm6813a4

Phan, T. G., Dreno, B., da Costa, A. C., Li, L., Orlandi, P., Deng, X., Kapusinszky, B., Siqueira, J., Knol, A.-C., Halary, F., Dantal, J., Alexander, K. A., Pesavento, P. A., & Delwart, E. (2016). A new protoparvovirus in human fecal samples and cutaneous T cell lymphomas (mycosis fungoides). Virology, 496, 299–305. https://doi.org/10.1016/j.virol.2016.06.013

Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010 Lisbon, Portugal, April 25-28, 2010 Proceedings. (2010). Retrieved July 26, 2020, from https://link.springer.com/content/pdf/10.1007%2F978-3-642-12683-3.pdf

Rimayi, C., Odusanya, D., Weiss, J. M., de Boer, J., & Chimuka, L. (2018). Contaminants of emerging concern in the Hartbeespoort Dam catchment and the uMngeni River estuary 2016 pollution incident, South Africa. The Science of the Total Environment, 627, 1008–1017. https://doi.org/10.1016/j.scitotenv.2018.01.263

Rodriguez-R, L. M., Tsementzi, D., Luo, C., & Konstantinidis, K. T. (n.d.). Iterative Subtractive Binning of Freshwater Chronoseries Metagenomes Identifies over 400 Novel Species and their Ecologic Preferences. Environmental Microbiology, n/a(n/a). https://doi.org/10.1111/1462-2920.15112

Rotavirus (n.d.). Retrieved May 13, 2018, from https://www.cdc.gov/vaccines/pubs/pinkbook/downloads/rota.pdf

Roychowdhury, T., Vishnoi, A., & Bhattacharya, A. (2013). Next-Generation Anchor Based Phylogeny (NexABP): Constructing phylogeny from Next-generation sequencing data. Scientific Reports, 3. https://doi.org/10.1038/srep02634

Ruggeri, F. M., Bonomo, P., Ianiro, G., Battistone, A., Delogu, R., Germinario, C., Chironna, M., Triassi, M., Campagnuolo, R., Cicala, A., Giammanco, G. M., Castiglia, P., Serra, C.,

Gaggioli, A., & Fiore, L. (2015). Rotavirus Genotypes in Sewage Treatment Plants and in Children Hospitalized with Acute Diarrhea in Italy in 2010 and 2011. Applied and Environmental Microbiology, 81(1), 241–249. https://doi.org/10.1128/AEM.02695-14

Rusiñol, M., Martínez-Puchol, S., Timoneda, N., Fernández-Cassi, X., Pérez-Cataluña, A., Fernández-Bravo, A., Moreno-Mesonero, L., Moreno, Y., Alonso, J. L., Figueras, M. J., Abril, J. F., Bofill-Mas, S., & Girones, R. (2020). Metagenomic analysis of viruses, bacteria and protozoa in irrigation water. International Journal of Hygiene and Environmental Health, 224, 113440. https://doi.org/10.1016/j.ijheh.2019.113440

Schneeberger, P. H. H., Fuhrimann, S., Becker, S. L., Pothier, J. F., Duffy, B., Beuret, C., Frey, J. E., & Utzinger, J. (2019). Qualitative microbiome profiling along a wastewater system in Kampala, Uganda. Scientific Reports, 9(1), 17334. https://doi.org/10.1038/s41598-019-53569-5

Selected Environmental Indicators. (July 2002). Botswana Central Statistics Office. Retrieved August 24, 2020, from http://www.statsbots.org.bw/sites/default/files/publications/Selected%20Environmental%20Indicators.pdf

SeqPlex RNA Amplification Kit SEQR. (n.d.). Sigma-Aldrich. Retrieved September 1, 2020, from https://www.sigmaaldrich.com/catalog/product/sigma/seqr

STAMP - Bioinformatics Software. (n.d.). Retrieved July 18, 2020, from https://beikolab.cs.dal.ca/software/STAMP

Taylor, T. A., & Unakal, C. G. (2020). Staphylococcus Aureus. In StatPearls. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK441868/

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics—A guide from sampling to data analysis. Microbial Informatics and Experimentation, 2, 3. https://doi.org/10.1186/2042-5783-2-3

Tourism Statistics Annual Report 2017.pdf. (n.d.). Retrieved August 24, 2020, from http://www.statsbots.org.bw/sites/default/files/Tourism%20Statistics%20Annual%20Report%202017.pdf

Traoré, A. N., Mulaudzi, K., Chari, G. J. E., Foord, S. H., Mudau, L. S., Barnard, T. G., & Potgieter, N. (2016). The Impact of Human Activities on Microbial Quality of Rivers in the Vhembe District, South Africa. International Journal of Environmental Research and Public Health, 13(8). https://doi.org/10.3390/ijerph13080817

Tubatsi, G., Bonyongo, M. C., & Gondwe, M. (2015). Water use practices, water quality, and households' diarrheal encounters in communities along the Boro-Thamalakane-Boteti river system, Northern Botswana. Journal of Health, Population, and Nutrition, 33, 21. https://doi.org/10.1186/s41043-015-0031-z

Untitled map. (n.d.). Google My Maps. Retrieved May 13, 2018, from https://www.google.com/maps/d/u/0/new?hl=en&at=AAX3J7DEaizkD-IM7QdkKovSX-eVqzNpjA:1526257528849&hl=en

Vezzulli, L., Grande, C., Tassistro, G., Brettar, I., Höfle, M. G., Pereira, R. P. A., Mushi, D., Pallavicini, A., Vassallo, P., & Pruzzo, C. (2017). Whole-Genome Enrichment Provides Deep Insights into Vibrio cholerae Metagenome from an African River. Microbial Ecology, 73(3), 734–738. https://doi.org/10.1007/s00248-016-0902-x

Vierheilig, J., Savio, D., Ley, R. E., Mach, R. L., Farnleitner, A. H., & Reischer, G. H. (2015). Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. Water Science and Technology: A Journal of the International Association on Water Pollution Research, 72(11), 1962–1972. https://doi.org/10.2166/wst.2015.407

WHO | Botswana. (n.d.). WHO; World Health Organization. Retrieved July 5, 2020, from http://www.who.int/countries/bwa/en/

WHO | Rotavirus. (n.d.). WHO; World Health Organization. Retrieved July 5, 2020, from http://www.who.int/immunization/diseases/rotavirus/en/

WHO | Updated WHO/WEDC Technical Notes on WASH in Emergencies. (n.d.). WHO; World Health Organization. Retrieved July 5, 2020, from http://www.who.int/water_sanitation_health/publications/technotes/en/

WHO | Water-related diseases. (n.d.). WHO; World Health Organization. Retrieved July 5, 2020, from http://www.who.int/water_sanitation_health/diseases-risks/diseases/diarrhoea/en/

Wood, D. E., Lu, J., & Langmead, B. (2019a). Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014a). Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46 World Health Organization (Ed.). (2015). WHO estimates of the global burden of foodborne diseases. World Health Organization.

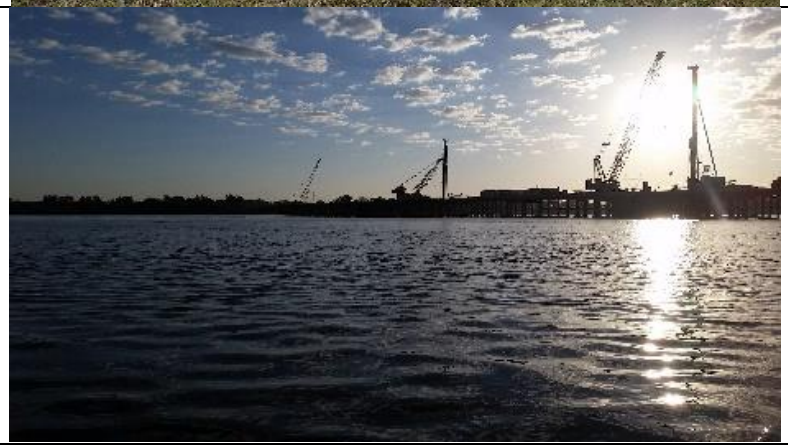
Xiao, L., & Feng, Y. (2008). Zoonotic cryptosporidiosis. FEMS Immunology and Medical Microbiology, 52(3), 309–323. https://doi.org/10.1111/j.1574-695X.2008.00377.x

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy, 67(11), 2640–2644. https://doi.org/10.1093/jac/dks261

Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Research, 38(12), e132–e132. https://doi.org/10.1093/nar/gkq275

Appendices

Table 18: Sampling Sites

Site Name	Latitude	Longitude	Sampling Description	Site Characteristics	Comments	Site Photograph
Seep	-17.79111	25.19809	6L filtered on land	Minerally rich site, used for ritual purging, separated from main river during dry season	Site not accessible for October 2019 sampling	
TR1	-17.792263	25.260647	6L filtered on boat	Site is near the bridge and a hippo habitat, also a bit of a current making bag filling more challenging Mixed Use		

TR13	-17.79149	25.20577	6L filtered on boat	Site is near a shoreline campsite Mixed Use		
TR21	-17.78218	25.17053	6L filtered on boat	“Alligator Cove”. More secluded site, dodging of hippos and crocodiles necessary Town	Site moves slightly due to presence of dangerous wildlife	

TR25	-17.78699	25.15685	6L filtered on boat	Near town, local fishing site Town		
TR27	-17.79441	25.15233	6L filtered on boat	Near town, right outside marina Town		

TR31	-17.81009	25.14426	6L filtered on boat	Near one of the tourist lodges Park		
TR35	-17.82577	25.13466	6L filtered on boat	Site is in park, ~100ft from outfall dock Park		

TR49	-17.84032	25.07253	6L filtered on boat	In park, near hipbone Park		
TR53	-17.83728	25.05486	6L filtered on boat	In park, very close to shore Park	Filter housing broke in freezer storage for March 2017 sampling event	

Table 19: Sample Processing (in field portion) Comparison

Sampling Location	BMFS version used	Filter #	Filter Vol. (L)	Frozen prior to elution	Elution Volume (ml)	Date Eluted	Tissue Kit	Seq. 2019	Viral 50ml	2° Conc Vol (ml)	# pellets	# Used in All Prep Kit	# pellets to UW	# pellets left at field lab
TR1_March 2017	v1	809	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR13_March 2017	v1	812	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR21_March 2017	v1	811	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR25_March 2017	v1	820	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR27_March 2017	v1	813	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR31_March 2017	v1	817	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR35_March 2017	v1	808	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR49_March 2017	v1	801	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR53_March 2017	v1	806	6	yes	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Seep_March 2017	v1	822	6	yes	200	Aug 2017	yes	-	-	150	4	2	2	-
TR1_March 2017 viral fraction	v1	809	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR13_March 2017 viral fraction	v1	812	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR21_March 2017 viral fraction	v1	811	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR25_March 2017 viral fraction	v1	820	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR27_March 2017 viral fraction	v1	813	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR31_March 2017 viral fraction	v1	817	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR35_March 2017 viral fraction	v1	808	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR49_March 2017 viral fraction	v1	801	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-

TR53_March 2017 viral fraction	v1	806	6	yes	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Seep_March 2017 viral fraction	v1	822	6	yes	200	Aug 2017	no	-	yes	50	1	1	0	-
TR1_August 2017	v1	1091	6	-	200	Aug 2017	yes	-	-	150	4	2	2	-
TR13_August 2017	v1	1092	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR21_August 2017	v1	1080	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR25_August 2017	v1	1078	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR27_August 2017	v1	1086	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR31_August 2017	v1	1077	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR35_August 2017	v1	1088	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR49_August 2017	v1	no id	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR53_August 2017	v1	1084	6	-	200	Aug 2017	yes	-	-	150	4	2	2	-
Seep_August 2017	v1	1089	6	-	200	Aug 2017	yes	yes	-	150	4	2	2	-
TR1_August 2017 viral fraction	v1	1091	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR13_August 2017 viral fraction	v1	1092	6	-	200	Aug 2017	no	-	yes	50	1	1	0	
TR21_August 2017 viral fraction	v1	1080	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR25_August 2017 viral fraction	v1	1078	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR27_August 2017 viral fraction	v1	1086	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR31_August 2017 viral fraction	v1	1077	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR35_August 2017 viral fraction (ATR B35V DNA)	v1	1088	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR49_August 2017 viral fraction	v1	no id	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
TR53_August 2017 viral fraction (ATR A35V DNA)	v1	1084	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-
Seep_August 2017 viral fraction	v1	1089	6	-	200	Aug 2017	no	-	yes	50	1	1	0	-

TR1_October 2019	v2	3384	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR13_October 2019	v2	3325	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR21_October 2019	v2	3389	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR25_October 2019	v2	3385	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR27_October 2019	v2	3391	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR31_October 2019	v2	3428	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR35_October 2019	v2	3346	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR49_October 2019	v2	3323	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
TR53_October 2019	v2	3344	6	-	300	Oct 2019	no	-	-	250	6	2	2	2
Seep_October 2019	*not accessible		n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
TR1_October 2019 viral fraction	v2	3384	6	no	300	Oct 2019	no	-	yes	50	1	1	0	-
TR13_October 2019 viral fraction	v2	3325	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR21_October 2019 viral fraction	v2	3389	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR25_October 2019 viral fraction	v2	3385	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR27_October 2019 viral fraction	v2	3391	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR31_October 2019 viral fraction	v2	3428	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR35_October 2019 viral fraction	v2	3346	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR49_October 2019 viral fraction	v2	3323	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
TR53_October 2019 viral fraction	v2	3344	6	-	300	Oct 2019	no	-	yes	50	1	1	0	-
Seep_October 2019 viral fraction	*not accessible		n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Table 20: DNA Extract Sample Laboratory Processing

Sample Name	DNA extract to UW (ul)	ds DNA HS Qubit prior to library prep (ng/ul)	ds DNA HS Qubit after Nextera DNA Flex (ng/ul)	conc. in nM	ul to final pool	nmol to final pool	Seq Date
201703-TR1-B	100	0.898	0.772	3.341991	10	3.34E-05	4/17/20
201703-TR13-B	100	0.888	0.5235	2.266234	10	2.27E-05	4/17/20
201703-TR21-B	100	1.445	0.408	1.766234	10	1.77E-05	4/17/20
201703-TR25-B	100	1.009	0.8795	3.807359	10	3.81E-05	4/17/20
201703-TR27-B	100	1.415	1.63	7.056277	10	7.06E-05	4/17/20
201703-TR31-B	100	1.385	0.4285	1.854978	10	1.85E-05	4/17/20
201703-TR35-B	100	2.84	0.3575	1.547619	10	1.55E-05	4/17/20
201703-TR49-B	100	0.848	1.195	5.17316	10	5.17E-05	4/17/20
201703-TR53-B	n/a	n/a	n/a	n/a	n/a	n/a	n/a
201703-Seep-B	100	0.566	0.4365	1.88961	10	1.89E-05	4/17/20
201703-TR1-V	100	0.0716	3.27	14.15584	10	0.000142	4/17/20
201703-TR13-V	100	0.1545	3.96	17.14286	9.9	0.00017	4/17/20
201703-TR21-V	100	0.0994	2.655	11.49351	10	0.000115	4/17/20
201703-TR25-V	100	0.524	13.8	59.74026	2.8	0.000167	4/17/20
201703-TR27-V	100	0.1495	3.825	16.55844	10	0.000166	4/17/20
201703-TR31-V	100	0.558	16.15	69.91342	2.4	0.000168	4/17/20
201703-TR35-V	100	0.468	15.15	65.58442	2.55	0.000167	4/17/20
201703-TR49-V	100	0.254	7.495	32.44589	5.2	0.000169	4/17/20
201703-TR53-V	n/a	n/a	n/a	n/a	n/a	n/a	n/a
201703-Seep-V	100	0.1605	19.5	84.41558	2	0.000169	4/17/20
201708-TR1-B	100	2.81	3.835	16.60173	10	0.000166	4/17/20
201708-TR13-B	100	2.45	1.02	4.415584	10	4.42E-05	4/17/20
201708-TR21-B	100	1.34	1.7	7.359307	10	7.36E-05	4/17/20
201708-TR25-B	100	9.6	8.145	35.25974	4.8	0.000169	4/17/20
201708-TR27-B	100	9.12	1.765	7.640693	10	7.64E-05	4/17/20
201708-TR31-B	100	4.88	4.235	18.33333	9.1	0.000167	4/17/20
201708-TR35-B	100	1.055	0.301	1.30303	10	1.3E-05	4/17/20
201708-TR49-B	100	0.528	0.487	2.108225	10	2.11E-05	4/17/20
201708-TR53-B	100	4.32	5.92	25.62771	6.5	0.000167	4/17/20
201708-Seep-B	100	1.04	0.1155	0.5	10	0.000005	4/17/20
201708-TR1-V	100	3.135	14.5	62.77056	2.7	0.000169	4/17/20
201708-TR13-V	100	1.75	10.015	43.35498	3.9	0.000169	4/17/20
201708-TR21-V	100	2.275	9.72	42.07792	4	0.000168	4/17/20
201708-TR25-V	100	11.55	5.68	24.58874	6.9	0.00017	4/17/20
201708-TR27-V	100	10.65	7.705	33.35498	5	0.000167	4/17/20
201708-TR31-V	100	12.05	6.87	29.74026	5.6	0.000167	4/17/20
201708-TR35-B	100	7.12	5.605	24.26407	7	0.00017	4/17/20
201708-TR49-V	100	3.335	13.2	57.14286	3	0.000171	4/17/20
201708-TR53-V	100	10.05	9.29	40.21645	4.2	0.000169	4/17/20
201708-Seep-V	100	8.68	13.4	58.00866	2.9	0.000168	4/17/20

201910-TR1-B	100	3.135	7.66	33.16017	5.1	0.000169	4/17/20
201910-TR13-B	100	5.14	9.735	42.14286	4	0.000169	4/17/20
201910-TR21-B	100	2.8	11.8	51.08225	3.3	0.000169	4/17/20
201910-TR25-B	100	4.76	14.8	64.06926	2.6	0.000167	4/17/20
201910-TR27-B	100	5.76	3.78	16.36364	10	0.000164	4/17/20
201910-TR31-B	100	4.98	1.06	4.588745	10	4.59E-05	4/17/20
201910-TR35-B	100	8.42	2.62	11.34199	10	0.000113	4/17/20
201910-TR49-B	100	5.76	6.325	27.38095	6.1	0.000167	4/17/20
201910-TR53-B	100	5.8	0.3805	1.647186	10	1.65E-05	4/17/20
201910-Seep-B	n/a	n/a	n/a	n/a	n/a	n/a	n/a
201910-TR1-V	100	4.94	20.75	89.82684	2	0.00018	4/17/20
201910-TR13-V	100	5.24	21.85	94.58874	2	0.000189	4/17/20
201910-TR21-V	100	5.4	12.05	52.1645	3.2	0.000167	4/17/20
201910-TR25-V	100	5.58	23.3	100.8658	2	0.000202	4/17/20
201910-TR27-V	100	5.26	32.35	140.0433	2	0.00028	4/17/20
201910-TR31-V	100	9.04	9.175	39.71861	4.2	0.000167	4/17/20
201910-TR35-V	100	10.3	19.9	86.14719	2	0.000172	4/17/20
201910-TR49-V	100	6.14	11.6	50.21645	3.4	0.000171	4/17/20
201910-TR53-V	100	7.48	0.0715	0.309524	10	3.1E-06	4/17/20
201910-Seep-V	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Table 21: RNA Extract Sample Laboratory Processing

Sample Name	RNA extract to UW (ul)	hs RNA pre-SeqPlex (ng/ul)	SeqPlex Primer Removal Reaction (Ct)			hs DNA post SeqPlex (ng/ul)
			w/out enzyme	*w/enzyme*	*ΔCt*	
201703-TR1-B	60	-	17.09	18.82	-1.73	13.55
201703-TR13-B	60	-	21.12	20.01	1.11	11.7
201703-TR21-B	60	-	19.63	24.37	-4.74	13.75
201703-TR25-B	60	-	22.42	15.8	6.62	21.75
201703-TR27-B	60	0.3335	17.65	19.99	-2.34	16.8
201703-TR31-B	60	0.523	20.33	16.09	4.24	22.2
201703-TR35-B	60	0.716	17.65	13.13	4.52	high
201703-TR49-B	60	-	19.22	19.85	-0.63	14.1
201703-TR53-B	n/a	n/a	n/a	n/a	n/a	n/a
201703-Seep-B	60	-	16.92	18.55	-1.63	13.55
201703-TR1-V	60	-	13.63	18.25	-4.62	16.6
201703-TR13-V	60	-	14.26	18.7	-4.44	9.3
201703-TR21-V	60	-	12.39	17.91	-5.52	22
201703-TR25-V	60	-	14.28	22.19	-7.91	13.5
201703-TR27-V	60	-	14.84	20.65	-5.81	13.05
201703-TR31-V	60	-	13.24	22.36	-9.12	14.95
201703-TR35-V	60	-	17.74	21.69	-3.95	12.5
201703-TR49-V	60	-	15.27	19.66	-4.39	12.95
201703-TR53-V	n/a	n/a	n/a	n/a	n/a	n/a
201703-Seep-V	60	-	13.24	31.01	-17.77	13.9
201708-TR1-B	60	0.2445	12.51	17.92	-5.41	18.1
201708-TR13-B	60	0.6295	16.46	18.02	-1.56	16.8
201708-TR21-B	60	3.985	14.82	16.55	-1.73	16.7
201708-TR35A-B	60	4.405	19.03	13.75	5.28	23.6
201708-TR27-B	60	7.5	15.83	14.24	1.59	18.85
201708-TR31-B	60	5.45	16.53	13.1	3.43	23.2
201708-TR35B-B	60	3.895	16.4	13.63	2.77	22.8
201708-TR49-B	60	2.695	21.85	13.82	8.03	19.2
201708-TR53-B	60	1.77	15.34	15.61	-0.27	17.2
201708-Seep-B	60	-	15.33	12.95	2.38	17.4
201708-TR1-V	60	-	13.71	21.67	-7.96	13.1
201708-TR13-V	60	0.3605	14.8	22.2	-7.4	12.6
201708-TR21-V	60	0.5525	12.41	14.22	-1.81	12.85
201708-TR25-V	60	1.16	14.19	22.15	-7.96	12.9
201708-TR27-V	60	6.1	12.71	18.56	-5.85	15.5
201708-TR31-V	60	4.125	18.23	18.29	-0.06	20.2
201708-TR35A-V	60	3.38	13.97	16.96	-2.99	21.4
201708-TR49-V	60	0.2635	15.67	19.03	-3.36	18.2
201708-TR35B-V	60	1.16	11.35	15.32	-3.97	21.6
201708-Seep-V	60	3.695	12.03	17.61	-5.58	10.9
201910-TR1-B	60	0.291	18.21	17.82	0.39	7.26
201910-TR13-B	60	0.322	32.77	18.91	13.86	14.5

201910-TR21-B	60	-	17.79	20.16	-2.37	15.45
201910-TR25-B	60	0.3685	16.79	14.86	1.93	21
201910-TR27-B	60	0.3455	21.56	19.05	2.51	15.05
201910-TR31-B	60	0.233	22.58	16.66	5.92	21.6
201910-TR35-B	60	0.2665	18.85	-	-	22.4
201910-TR49-B	60	-	15.85	14.37	1.48	18.15
201910-TR53-B	60	-	16.32	13.36	2.96	19.6
201910-Seep-B	n/a	n/a	n/a	n/a	n/a	n/a
201910-TR1-V	60	0.215	17.19	16.87	0.32	19.5
201910-TR13-V	60	-	15.05	16.36	-1.31	15.5
201910-TR21-V	60	-	15.51	12.47	3.04	18.85
201910-TR25-V	60	-	16.54	32.22	-15.68	13.15
201910-TR27-V	60	-	16.78	26.49	-9.71	17.6
201910-TR31-V	60	-	14.54	18.68	-4.14	15.15
201910-TR35-V	60	-	15.22	18.18	-2.96	14.1
201910-TR49-V	60	-	11.14	20.97	-9.83	17.5
201910-TR53-V	60	-	12.25	17.05	-4.8	22.85
201910-Seep-V	n/a	n/a	n/a	n/a	n/a	n/a
water from 5/19/20	-	-	15.01	16.02	-1.01	9.22
water from 5/21/20	-	-	19.45	-	-	13.15
water from 5/7/20	-	-	26.14	-	-	9.02

Table 22: Round 2 DNA Sequencing File Overview Data

Sample Name	Date Created	Paired end	Number of Clusters	Number of Reads	Read 1 Length	Read 2 Length	FASTQ file size	Metaspades file size
201703-Seep-B	4/17/2020	paired	2,933,020	5,866,040	51	51	187.53MB	10MB
201703-Seep-V	4/17/2020	paired	23,437,571	46,875,144	51	51	1.43GB	134MB
201703-TR13-B	4/17/2020	paired	1,759,475	3,518,950	51	51	113.43MB	11MB
201703-TR13-V	4/17/2020	paired	31,732,240	63,464,480	51	51	1.91GB	116MB
201703-TR1-B	4/17/2020	paired	1,874,977	3,749,954	51	51	122.69MB	11MB
201703-TR1-V	4/17/2020	paired	17,618,083	35,236,166	51	51	1.08GB	71MB
201703-TR21-B	4/17/2020	paired	1,719,877	3,439,754	51	51	114.15MB	12MB
201703-TR21-V	4/17/2020	paired	26,018,522	52,037,044	51	51	1.57GB	91MB
201703-TR25-B	4/17/2020	paired	2,540,343	5,080,686	51	51	165.32MB	9MB
201703-TR25-V	4/17/2020	paired	50,619,394	101,238,788	51	51	3.09GB	Too large
201703-TR27-B	4/17/2020	paired	7,983,681	15,967,363	51	51	506.47MB	46MB
201703-TR27-V	4/17/2020	paired	24,543,749	49,087,498	51	51	1.48GB	117MB
201703-TR31-B	4/17/2020	paired	1,877,127	3,754,254	51	51	122.90MB	11MB
201703-TR31-V	4/17/2020	paired	77,844,214	155,688,428	51	51	4.60GB	Too large
201703-TR35-B	4/17/2020	paired	603,637	1,207,274	51	51	40.31MB	738KB
201703-TR35-V	4/17/2020	paired	42,508,457	85,016,914	51	51	2.54GB	Too large
201703-TR49-B	4/17/2020	paired	5,607,674	11,215,348	51	51	357.07MB	35MB
201703-TR49-V	4/17/2020	paired	44,537,958	89,075,916	51	51	2.64GB	Too large
201708-Seep-B	4/17/2020	paired	48,434	96,868	51	51	3.52MB	75KB
201708-Seep-V	4/17/2020	paired	9,689,796	19,379,592	51	51	632.04MB	73MB
201708-TR13-B	4/17/2020	paired	2,179,625	4,359,250	51	51	145.42MB	7MB
201708-TR13-V	4/17/2020	paired	4,742,153	9,484,306	51	51	315.46MB	21MB
201708-TR1-B	4/17/2020	paired	17,838,318	35,676,636	51	51	1.11GB	101MB
201708-TR1-V	4/17/2020	paired	9,432,555	18,865,110	51	51	622.99MB	38MB
201708-TR21-B	4/17/2020	paired	5,319,562	10,639,124	51	51	344.07MB	19MB
201708-TR21-V	4/17/2020	paired	12,557,193	25,114,386	51	51	816.97MB	57MB
201708-TR25-B	4/17/2020	paired	15,969,804	31,939,608	51	51	1017.05MB	93MB
201708-TR25-V	4/17/2020	paired	12,749,895	25,499,790	51	51	819.94MB	70MB
201708-TR27-B	4/17/2020	paired	5,212,698	10,425,396	51	51	337.37MB	26MB
201708-TR27-V	4/17/2020	paired	11,907,942	23,815,884	51	51	771.44MB	62MB
201708-TR31-B	4/17/2020	paired	12,009,561	24,019,122	51	51	768.73MB	69MB
201708-TR31-V	4/17/2020	paired	16,252,944	32,505,888	51	51	1.01GB	98KB
201708-TR35-B	4/17/2020	paired	894,724	1,789,448	51	51	58.83MB	7MB
201708-TR35-V	4/17/2020	paired	22,862,029	45,862,058	51	51	1.44GB	110MB
201708-TR49-B	4/17/2020	paired	2,127,877	4,255,754	51	51	138.61MB	10MB
201708-TR49-V	4/17/2020	paired	12,528,775	25,057,550	51	51	813.62MB	65MB
201708-TR53-B	4/17/2020	paired	12,252,714	24,505,428	51	51	788.15MB	68MB
201708-TR53-V	4/17/2020	paired	12,174,816	24,349,632	51	51	788.02MB	70MB
201910-TR13-B	4/17/2020	paired	11,736,934	23,473,868	51	51	756.44MB	63MB
201910-TR13-V	4/17/2020	paired	14,889,182	29,778,364	51	51	957.54MB	79MB
201910-TR1-B	4/17/2020	paired	15,258,737	30,517,474	51	51	981.36MB	82MB
201910-TR1-V	4/17/2020	paired	17,528,433	35,056,866	51	51	1.10GB	83MB

201910-TR21-B	4/17/2020	paired	8,207,692	16,415,384	51	51	530.75MB	42MB
201910-TR21-V	4/17/2020	paired	14,566,947	29,133,894	51	51	940.31MB	68MB
201910-TR25-B	4/17/2020	paired	11,894,496	23,788,992	51	51	766.77MB	56MB
201910-TR25-V	4/17/2020	paired	25,648,843	51,297,686	51	51	1.61GB	113MBB
201910-TR27-B	4/17/2020	paired	8,640,177	17,280,354	51	51	561.51MB	41MB
201910-TR27-V	4/17/2020	paired	31,165,492	62,330,984	51	51	1.95GB	131MB
201910-TR31-B	4/17/2020	paired	2,547,112	5,094,224	51	51	167.88MB	8MB
201910-TR31-V	4/17/2020	paired	35,969,782	71,939,564	51	51	2.20GB	197MB
201910-TR35-B	4/17/2020	paired	4,724,897	9,449,794	51	51	310.49MB	22MB
201910-TR35-V	4/17/2020	paired	19,122,637	38,245,274	51	51	1.21GB	111MB
201910-TR49-B	4/17/2020	paired	9,336,424	18,672,848	51	51	604.32MB	54MB
201910-TR49-V	4/17/2020	paired	8,718,304	17,436,608	51	51	574.11MB	61MB
201910-TR53-B	4/17/2020	paired	560,745	1,121,490	51	51	37.86MB	380KB
201910-TR53-V	4/17/2020	paired	27,015	54,030	51	51	1.95MB	76KB

Table 23: Round 2 DNA Sequencing Contig Assembly Overview

	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	Total length (>= 0 bp)	Total length (>= 1000 bp)	# contigs	Largest contig	Total length	GC (%)	N50	N75	L50	L75	# N's per 100 kbp
201703-Seep-B	12618	673	4683636	1169310	2312	39590	2293996	37.04	1018	701	651	1337	0
201703-Seep-V	137024	10925	64773252	31862389	26713	99519	42705589	53.52	2216	986	3576	11094	0
201703-TR13-B	6985	1160	5675016	3863650	2054	41644	4490867	35.53	3700	1540	277	757	0
201703-TR13-V	168436	7910	54313096	26183800	19307	136008	33984216	49.54	2961	1078	2200	7240	0
201703-TR1-B	7382	1313	5315775	3476233	2428	25899	4269253	35.25	2503	1232	443	1064	0
201703-TR1-V	99067	5840	33253964	13037982	14634	47970	19069553	44.31	1569	855	3017	7211	0
201703-TR21-B	6937	1122	5800477	4030004	1991	48494	4637217	35.66	4246	1662	248	689	0
201703-TR21-V	136848	7168	42321629	19463975	15953	51415	25505205	43.6	2247	1038	2545	6839	0
201703-TR25-B	13001	321	4196000	470997	1669	4498	1368364	38.79	806	619	561	1051	0
201703-TR25-V	Too large												
201703-TR27-B	42175	2681	22042363	8698552	9414	185576	13287795	38.88	1527	806	983	4100	0
201703-TR27-V	147343	10140	55271143	20424293	26656	22892	31793953	48.13	1361	804	6238	13968	0
201703-TR31-B	4990	1173	5330497	4142424	1875	89005	4646584	35.44	4313	1757	245	679	0
201703-TR31-V	Too large												
201703-TR35-B	864	42	309109	80293	146	14970	151547	52.12	1150	699	38	82	0
201703-TR35-V	Too large												
201703-TR49-B	36874	3052	17058272	5346345	9385	55227	9691117	34.39	1095	724	2574	5331	0
201703-TR49-V	Too large												
201708-Seep-B	26	1	4489	1433	2	1433	2339	54.25	1433	906	1	2	0
201708-Seep-V	55042	5833	35447653	20766559	12804	148828	25596358	57.07	3609	1260	1279	4432	0
201708-TR13-B	7082	487	3419692	1471550	1329	26991	2045752	45.14	2264	925	179	553	0
201708-TR13-V	19521	1615	10128988	5286748	3694	60188	6722766	46.14	2921	1136	424	1386	0
201708-TR1-B	93773	8063	49101778	23622128	20169	96419	31908728	45.24	2236	971	2722	8377	0
201708-TR1-V	36719	2760	18405159	9361785	6603	65550	11996641	45.77	3028	1117	699	2416	0
201708-TR21-B	19617	1348	8883101	3810066	3217	60532	5078215	44.79	2194	1001	457	1347	0
201708-TR21-V	61222	4271	27367798	13276845	10327	90580	17450760	45.53	2529	1038	1246	4086	0
201708-TR25-B	88556	7994	45127849	20267550	20080	73875	28588908	46.44	1780	904	3426	9232	0
201708-TR25-V	71655	5027	33697218	14396473	12831	78154	19725671	46.68	2150	944	1812	5438	0
201708-TR27-B	25615	1811	12371296	4998981	5034	78698	7205444	44.94	1810	870	737	2246	0
201708-TR27-V	56669	5006	30095072	16421620	10938	83887	20500740	46.98	3092	1201	1247	4050	0
201708-TR31-B	62853	5310	33442233	15959797	13407	94618	21495736	46.53	2296	976	1679	5475	0

201708-TR31-V	86861	7861	47806504	24511236	19225	166468	32317180	46.73	2565	1026	2332	7592	0
201708-TR35-B	8179	550	3423736	900734	1955	12175	1862459	35.25	976	688	581	1154	0
201708-TR35-V	108121	9234	53396801	28241721	20946	94799	36314708	47.04	2605	1091	2623	8272	0
201708-TR49-B	8827	747	4826492	2268969	1934	55605	3068047	42.77	2285	960	232	780	0
201708-TR49-V	63348	5153	31444666	16709510	11742	86311	21271959	47.39	2849	1125	1313	4440	0
201708-TR53-B	61962	4936	32727795	16201724	12184	103465	21152591	44.2	2825	1055	1337	4608	0
201708-TR53-V	65200	5155	33959090	16973522	12667	83094	22116616	46.33	2869	1058	1374	4780	0
201910-TR13-B	63492	5947	30176740	16092916	12734	70123	20798702	44.75	2280	1073	2072	5469	0
201910-TR13-V	73792	6881	38236149	21005446	15048	93859	26665593	46.99	2674	1133	1964	5936	0
201910-TR1-B	80837	7227	39629442	21259565	16002	129061	27324920	45.59	2461	1087	2145	6494	0
201910-TR1-V	76950	6551	40010422	22507070	14435	120532	27942334	47.68	3366	1232	1538	5148	0
201910-TR21-B	44043	3700	20139412	10795087	8024	62455	13785930	43.31	2508	1118	1182	3269	0
201910-TR21-V	63339	5507	33115904	19671331	11413	113877	23762074	45.82	3791	1351	1167	3908	0
201910-TR25-B	65055	4640	26463734	12661872	10924	62671	17004284	46.39	2141	984	1721	4733	0
201910-TR25-V	117488	8317	54588858	30404286	18444	157720	37344957	48.24	3817	1309	1821	6212	0
201910-TR27-B	44116	3635	19694525	8368446	9266	57806	12261533	49.59	1540	860	1811	4530	0
201910-TR27-V	131324	10243	62933573	36232215	21511	126279	43981047	47.53	3726	1339	2322	7431	0
201910-TR31-B	10305	504	3786040	955051	1666	11835	1745053	50.67	1086	714	426	927	0
201910-TR31-V	206538	15887	94795828	52027153	36397	148325	66153233	49.1	2998	1136	4158	13621	0
201910-TR35-B	25480	1659	10236640	3299969	4816	29524	5462095	47.87	1243	759	1146	2578	0
201910-TR35-V	112396	9320	53316648	29856247	20022	101932	37286698	49.28	3033	1196	2460	7585	0
201910-TR49-B	59343	4326	25997808	10311282	11490	82274	15239692	46.85	1580	848	2158	5544	0
201910-TR49-V	59999	4929	29197557	13944605	11697	61718	18603380	49.68	2189	998	1656	4937	0
201910-TR53-B	668	6	135520	7926	53	1771	39814	60.36	759	610	20	35	0
201910-TR53-V	25	0	3470	0	2	918	1448	54.21	918	530	1	2	0

Table 24: Round 2 DNA Sequencing Scaffold Assembly Overview

	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	Total length (>= 0 bp)	Total length (>= 1000 bp)	# contigs	Largest contig	Total length	GC (%)	N50	N75	L50	L75	# N's per 100 kbp
201703-Seep-B	11765	816	4689598	1506889	2339	39590	2554940	37	1173	756	605	1290	308.42
201703-Seep-V	120291	11258	64761145	37560655	24898	130033	46982689	53.3	3097	1194	2725	9131	332.95
201703-TR13-B	5961	1013	5658054	4215577	1660	49353	4666440	35.56	5688	2253	191	527	211.94
201703-TR13-V	165577	7853	54130976	26752100	19260	136008	34589748	49.56	3131	1089	2066	7078	111.68
201703-TR1-B	6161	1186	5328042	3949149	1991	35893	4521827	35.24	3870	1679	302	755	262.99
201703-TR1-V	97449	5835	33232949	13476644	14560	47970	19473574	44.35	1634	869	2868	7047	121.25
201703-TR21-B	5588	917	5769602	4442392	1519	58448	4857060	35.72	8409	2501	135	409	260.38
201703-TR21-V	134763	7012	42232593	19904566	15770	51415	25938875	43.61	2409	1056	2362	6573	123.64
201703-TR25-B	11386	563	4213841	917051	2047	7702	1925291	39.63	961	679	610	1214	784.71
201703-TR25-V													
201703-TR27-B	37235	3303	22090535	10488763	9596	194274	14806451	39.21	1889	897	967	3954	310.28
201703-TR27-V	144584	10227	55218314	21312910	26417	24298	32456130	48.11	1419	819	5932	13581	79.76
201703-TR31-B	3816	874	5309138	4526008	1254	89279	4797799	35.46	8402	3261	136	370	245.65
201703-TR31-V													
201703-TR35-B	726	62	312274	128789	156	19072	196224	51.54	1312	848	36	82	1490.13
201703-TR35-V													
201703-TR49-B	31427	3722	17113413	7441648	9305	61578	11293786	34.58	1378	826	2189	4859	471.3
201703-TR49-V													
201708-Seep-B	26	1	4489	1433	2	1433	2339	54.25	1433	906	1	2	0
201708-Seep-V	41031	6199	35975657	24999753	13663	153267	30022338	57.02	4832	1441	1082	4112	1801.19
201708-TR13-B	5748	544	3454675	1880039	1349	44653	2429113	44.89	3117	1088	135	489	1267.95
201708-TR13-V	15841	1688	10250407	6319326	3991	84318	7879008	46	3682	1238	347	1318	2044.28
201708-TR1-B	80476	8674	49061361	27582286	20210	162226	35546499	45.24	2723	1088	2361	7788	832.61
201708-TR1-V	30050	3076	18620637	11307149	7132	90077	14063455	45.72	3676	1229	609	2389	1613.44
201708-TR21-B	16193	1343	8953485	4727859	3252	83910	6020062	44.89	3291	1134	320	1143	1516.99
201708-TR21-V	50330	4951	27842636	16567867	11947	90580	21297004	45.48	2969	1101	1219	4383	2569.63
201708-TR25-B	70639	8857	45531882	26129022	19786	97494	33690969	46.43	2505	1086	2714	8031	1287.25
201708-TR25-V	57006	5843	34354431	18575139	15159	146653	24818902	46.63	2556	995	1739	5883	2659.99
201708-TR27-B	20150	2107	12553895	6756341	5206	133894	8874529	45	2534	1036	554	2009	1864.3
201708-TR27-V	45313	5311	30595371	19737509	11937	122861	24194542	46.86	3825	1302	1045	3911	2149.4
201708-TR31-B	49945	5824	33719793	19896231	13603	111212	25246332	46.5	3186	1132	1346	4920	1427.15

201708-TR31-V	70806	8480	48223325	29692434	19160	331779	37035821	46.67	3393	1202	1835	6729	1208.71
201708-TR35-B	6972	761	3436039	1393429	1989	12175	2238470	35.4	1246	787	515	1083	507.98
201708-TR35-V	90886	9687	53625384	33525168	20584	94860	41031835	47	3446	1283	2120	7246	1112.93
201708-TR49-B	7401	747	4848912	2719669	1845	109544	3469230	42.88	3175	1123	153	636	661.01
201708-TR49-V	51558	5682	31930805	20198581	13106	93949	25213270	47.26	3489	1216	1224	4512	2289.16
201708-TR53-B	50183	5271	33064294	19709409	12546	115258	24651826	44.26	3840	1210	1084	4160	1494.02
201708-TR53-V	52770	5668	34299861	20636788	13464	151307	25919912	46.19	3629	1185	1163	4565	1430.5
201910-TR13-B	53698	6057	30353254	19277633	11919	83616	23348465	44.93	3163	1329	1587	4523	1144.48
201910-TR13-V	61696	7127	38459577	24541928	15195	109828	30058554	46.99	3401	1296	1655	5364	1357.26
201910-TR1-B	68638	7422	39781974	25092250	15323	167463	30550458	45.81	3410	1309	1740	5512	1191.9
201910-TR1-V	64497	6902	40323342	26125155	15036	146807	31674821	47.74	4135	1356	1306	4853	1477.01
201910-TR21-B	37819	3657	20183739	12727765	7422	62455	15328505	43.44	3590	1401	874	2614	1025.11
201910-TR21-V	53916	5554	33316391	22354287	11691	113877	26529587	45.7	4922	1522	990	3548	1406.77
201910-TR25-B	54825	5127	26697473	15825751	11069	62671	19924879	46.73	2808	1166	1445	4311	1525.24
201910-TR25-V	101230	8732	54882113	34626362	19849	187065	42198947	48.27	4582	1359	1614	6161	1307.08
201910-TR27-B	34297	4482	20044154	11644209	9700	62713	15238660	49.31	2085	1038	1616	4271	2219
201910-TR27-V	112566	10416	63467110	41236221	23027	158221	49777700	47.29	4588	1415	1931	7110	1577.57
201910-TR31-B	8141	715	3860216	1566704	2018	34994	2447710	50.14	1405	792	423	1021	2708.41
201910-TR31-V	181920	16113	94979787	59989005	34300	155592	72615612	48.99	4118	1349	3176	11333	688.72
201910-TR35-B	20037	2229	10410868	5066771	5244	34907	7139007	47.68	1647	908	1050	2532	2274.71
201910-TR35-V	95122	9652	53769179	35240882	19830	104094	42232624	48.96	4062	1411	1976	6641	1273.6
201910-TR49-B	47133	5062	26319660	14036238	12097	93034	18881862	46.8	2142	981	1815	5189	1823.83
201910-TR49-V	46648	5671	29762567	17917107	13226	62712	23017472	49.44	2760	1093	1542	5044	2390.43
201910-TR53-B	604	22	136174	33508	65	3064	63069	58.88	1028	698	21	39	951.34
201910-TR53-V	25	0	3470	0	2	918	1448	54.21	918	530	1	2	0

Table 25: Round 2 DNA Sample MG-RAST Class Assignment

Sample Name	Bacteria %	Viruses %	Eukaryota %	unclassified %	Archaea %	Other %
201703-TR1-B						
201703-TR13-B	98.86	0.02	1.12	0	0	0
201703-TR21-B	99.41	0	0.57	0	0.02	0
201703-TR25-B						
201703-TR27-B	99.46	0.21	0.27	0.04	0.02	0
201703-TR31-B						
201703-TR35-B						
201703-TR49-B						
201703-Seep-B						
201703-TR1-V	98.54	0.08	1.34	0.01	0.02	0
201703-TR13-V	98.56	0.21	1.21	0.01	0.02	0
201703-TR21-V	98.04	0.4	1.51	0.01	0.03	0
201703-TR25-V						
201703-TR27-V	98.73	0.36	0.84	0.04	0.04	0
201703-TR31-V						
201703-TR35-V						
201703-TR49-V						
201703-Seep-V	97.58	0.08	1.12	0.19	1.02	0
201708-TR1-B	95.97	2.91	0.83	0.16	0.13	0
201708-TR13-B	83.41	13.46	1.94	1.13	0.06	0
201708-TR21-B	88	9.33	1.72	0.86	0.09	0
201708-TR25-B						
201708-TR27-B						
201708-TR31-B	91.79	6.25	1.37	0.45	0.15	0
201708-TR35-B	99.73	0	0.27	0	0	0
201708-TR49-B						
201708-TR53-B	91.22	6.78	1.44	0.46	0.09	0
201708-Seep-B						
201708-TR1-V	70.41	23.65	3.62	2.02	0.27	0.03
201708-TR13-V	67.21	25.91	3.77	2.91	0.15	0.05
201708-TR21-V						
201708-TR25-V	93.98	4.7	0.95	0.3	0.08	0
201708-TR27-V						
201708-TR31-V	89.92	7.65	1.66	0.55	0.23	0
201708-TR35-V	81.77	13.93	2.98	1.12	0.19	0
201708-TR49-V	74.12	18.29	5.34	2.03	0.21	0
201708-TR53-V	87.61	9.43	2.05	0.71	0.2	0
201708-Seep-V	89.5	8.37	1.46	0.57	0.1	0
201910-TR1-B	92.42	5.88	0.81	0.75	0.12	0.01
201910-TR13-B	93.13	5.53	0.63	0.6	0.1	0
201910-TR21-B						
201910-TR25-B	88.94	9.15	0.9	0.84	0.16	0.01
201910-TR27-B	88.11	9.93	0.75	1.13	0.08	0.01

201910-TR31-B						
201910-TR35-B	77.62	19.42	1.1	1.77	0.01	0.01
201910-TR49-B	83.3	13.5	1.41	1.65	0.14	0
201910-TR53-B						
201910-TR1-V						
201910-TR13-V						
201910-TR21-V						
201910-TR25-V	85.01	12.31	1.27	1.28	0.12	0
201910-TR27-V	79.85	16.64	1.66	1.66	0.17	0.02
201910-TR31-V	79.2	17.08	1.81	1.63	0.26	0.02
201910-TR35-V	73.42	22.91	1.65	1.84	0.17	0.01
201910-TR49-V	77.09	18.72	1.97	2.04	0.17	0.01
201910-TR53-V						

Table 26: RNA Samples post Nextera DNA Prep Kit

Sample Name	Qubit average (ng/ul)	concentration (nM)
201703-TR1-B	0.370	1.602
201703-TR13-B	0.306	1.323
201703-TR21-B	0.442	1.913
201703-TR27-B	0.356	1.541
201703-TR49-B	0.444	1.922
201703-Seep-B	0.524	2.268
201708-TR1-B	0.420	1.818
201708-TR13-B	0.570	2.468
201708-TR21-B	0.276	1.193
201708-TR27-B	0.784	3.394
201708-TR49-B	0.702	3.039
201708-TR53-B	0.596	2.580
201708-Seep-B	0.602	2.606
201910-TR1-B	0.358	1.550
201910-TR13-B	1.270	5.498
201910-TR21-B	0.736	3.186
201910-TR27-B	0.856	3.706
201910-TR49-B	0.708	3.065
201910-TR53-B	0.880	3.810
201910-TR1-V	0.624	2.701
201910-TR13-V	0.083	0.358
201910-TR21-V	0.440	1.905
201910-TR27-V	0.376	1.628
201910-TR49-V	0.460	1.991
water from 5/19/20	0.026	0.111

Sample Name	Qubit average (ng/ul)	concentration (nM)
201703-TR25-B	0.185	0.799
201703-TR31-B	0.160	0.693
201703-TR35-B	0.280	1.210
201708-TR35A-B	0.071	0.307
201708-TR31-B	0.099	0.429
201708-TR35B-B	0.103	0.446
201910-TR25-B	0.142	0.613
201910-TR31-B	0.066	0.287
201910-TR35-B	0.088	0.382
water from 5/7/20	0.039	0.170

Sample Name	Qubit average (ng/ul)	concentration (nM)
201703-TR1-V	0.106	0.459
201703-TR13-V	0.048	0.210
201703-TR21-V	0.083	0.359
201703-TR25-V	0.037	0.160
201703-TR27-V	0.055	0.240
201703-TR31-V	0.146	0.630
201703-TR35-V	0.108	0.468
201703-TR49-V	0.132	0.571
201703-Seep-V	0.123	0.532
201708-TR1-V	0.144	0.621
201708-TR13-V	0.114	0.494
201708-TR21-V	0.236	1.022
201708-TR25-V	0.182	0.786
201708-TR27-V	0.218	0.942
201708-TR31-V	0.370	1.600
201708-TR35A-V	0.413	1.786
201708-TR49-V	0.664	2.874
201708-TR35B-V	0.728	3.152
201708-Seep-V	0.223	0.963
201910-TR25-V	0.266	1.152
201910-TR31-V	0.610	2.641
201910-TR35-V	0.250	1.082
201910-TR53-V	0.880	3.810
water from 5/21/20	0.083	0.360

Appendix 1: SeqPlex RNA Amplification Protocol

Step 1: Qubit
RNA HS Qubit using 20ul of the samples that have not yet been done (should leave ~30ul of sample to amplify). Will proceed to the next kit no matter what the concentrations are recorded to be via Qubit.

Sample Name	RNA hs Qubit prior to Kit (ng/ul)	DNA hs Qubit after Kit (ng/ul)

Step 2 (Option #1): RNase-free DNase Kit
Run 10 samples through this kit using 30ul of each RNA extract (final sample will be 30ul nuclease free water).

DNase digest eluate using DNA-free DNase treatment (from previous NGS protocol)

a. Add 30 uL RNA, 1 uL rDNase 1, and 3 uL 10X buffer
b. Incubate for 30 minutes at 37ºC
c. Add 6 uL DNase inactivation agent
d. Incubate for 3 minutes at room temperature, mixing every minute
e. Centrifuge for 3 minutes at 12,000 rpm
f. Retain the supernatant for use

Step 3: WTA Kit
Run 10 samples through this kit using 25ul input volume.

Library Synthesis

1. Thaw the Library Synthesis Buffer, Library Synthesis Solution, and water. Mix thoroughly by inversion or brief vortexing. Dissolve any precipitate in the Library Synthesis Solution (L8670) by briefly heating at 37 °C, followed by thorough mixing. Keep on ice.

2. Combine 100 pg to 5 ng of high quality intact total RNA or 1–50 ng damaged RNA (e.g., FFPE or laser capture sample RNA) with Library Synthesis Solution at the following single-reaction scale:

   0.5 µL Library Synthesis Solution (L8670)
   Add Nuclease-free water (W4502) to a total volume of 3.3 µL.

3. Mix by pipeting and incubate reactions in a thermocycler programmed for 70 °C for 5 minutes, then an 18 °C hold. Do not hold at 18 °C for more than 10 minutes. Remove reactions from thermocycler, and place at room temperature or maintain at 18 °C for the next steps, but do not place on ice.
   Note: To avoid RNA renaturation and possible degradation, perform steps 4 and 5, and setup for step 6 as rapidly as possible at 18 °C or room temperature.

4. Prepare the following premix during the incubation in the previous step and immediately combine 1.7 µL of following premix with denatured RNA in Library Synthesis Solution (step 3):

   0.5 µL Library Synthesis Buffer (L9418)
   0.8 µL Water
   0.4 µL Library Synthesis Enzyme (L9543)

5. Mix by pipeting and spin down residue from the sides of reaction tubes.

6. Incubate in a thermocycler using the following conditions:

   18 °C for 10 minutes
   25 °C for 10 minutes
   37 °C for 30 minutes
   42 °C for 10 minutes
   70 °C for 20 minutes
   4 °C hold

7. Spin down any condensation by centrifugation. Samples may be amplified immediately or stored at –20 °C for up to one month.

Amplification

8. Thaw the 5× Amplification Mix. Dissolve any precipitate by briefly heating at 37 °C, followed by thorough mixing. Keep on ice.

9. Transfer the following regents to the library synthesis reaction, using the following single reaction scale*:

   53.50 µL Nuclease-free water
   15.00 µL 5X Amplification Mix (A5112)
   0.75 µL 1:1000 SYBR Green* in 10 mM

Tris-HCl, pH 8.0 (Catalog No. T3038)

0.75 µL Amplification Enzyme (A5237)

75.00 µL Total reaction volume

* Addition of SYBR Green, Catalog No. S9430, not included in the kit, is optional, but strongly recommended for monitoring amplification. Prepare dilution and add immediately to the mix. Discard dilution after each experiment.

10. Mix thoroughly by pipetting or brief vortexing. Spin down residue from top and sides of reaction tubes.

11. Proceed with PCR using the following thermocycler program (real-time qPCR is strongly recommended):

    1 cycle

    94 °C for 2 minutes.

    17-19 cycles*

    94 °C for 30 seconds

    70 °C for 5 minutes (read)

    1 cycle

    70 °C for 30 minutes

    Note: The final 70 °C incubation for 30 minutes is critical for primer removal and downstream sequencing application.

    * – The optimal number of amplification cycles varies with RNA input quantity and quality. Optimal amplification is achieved by proceeding 2–3 cycles into the amplification "plateau", as observed with real-time quantitative PCR. Typically, ~19 cycles are required for 1–5 ng of high-quality RNA or 10–50 ng of FFPE RNA. In this case, if amplifying without real-time monitoring, performing multiple reactions, or otherwise unable to practically monitor individual amplification reactions, set the number of amplification cycles at 20. RNA of lower quality or quantity may require higher input quantities and/or more amplification cycles. If amplifying less than 1 ng of RNA (RIN ≥8.0) or less than 10 ng of damaged RNA (RIN ≤8.0) without real-time monitoring, performing multiple reactions, or otherwise unable to practically monitor individual amplification reactions, set the number of amplification cycles at 25. **For best results, monitor amplification with Sybr Green.**

12. After cycling is complete, maintain the reactions at 4 °C or store at –20 °C until ready for purification.

13. To remove residual primers and nucleotides, purify with a PCR Cleanup Kit (as described in the kit instructions. **Elute with 50 µL nuclease-free water, not kit elution buffer**. Eluate can be concentrated by vacuum centrifugation if necessary, but avoid heating and do not allow the sample to go to dryness. (Because the amplification product is in water alone, without counter-ions present, the sample will denature upon dryness. Denaturation will inhibit primer removal.) The capacity of the GenElute purification column is 10 µg, adequate for purification of a typical amplification reaction.

14. **PCR Cleanup Kit (Monarch Oligonucleotide Cleanup Protocol – NEB #T1030):**

    a. Add 100ul DNA Cleanup Binding Buffer to the 50ul sample. If smaller volume, adjust with nuclease-free water.

    b. Add 300ul ethanol. Mix well by pipetting up and down. Do not vortex.

    c. Insert column into collection tube and load sample onto column. Spin for 1 minute, then discard flow-through.

    d. Re-insert column into collection tube. Add 500ul DNA Wash Buffer and spin for 1 minute.

    e. Repeat step d.

    f. Transfer column to a clean 1.5ml tube. Do not let tip of column contact flow through.

    g. Add 50ul of nuclease free water to the center of the matrix. Wait for 1 minute, then spin for 1 minute to elute DNA.

15. Purified DNA is quantified by measuring absorbance. One A_{260} unit is equivalent to 50 ng/µL DNA.

Primer Removal

Primer removal results in little or no loss of amplification product. However, a 75-µL primer removal reaction input of 2 µg of amplified product is recommended to yield 1 µg of final product for entering the deep sequencing workflow. (see "Considerations for Downstream Sequencing") This allows for loss due to primer removal plus any additional loss during reaction cleanup. A "no-enzyme" control reaction is also performed for each amplified sample. The control reaction is required to test for primer sequence removal (see **Quality Control** below).

Notes: The amplification product cannot be further amplified after primer removal.

Sufficient reagent has been included in the kit for a 75-µL "no-primer-removal-enzyme" control reaction. However, you may wish to

reduce the scale of your "no enzyme control" to save reagent and amplified product, as shown below. A reaction setup described in step 15 is a mix that accounts for both the primer removal reaction and (steps 16 and 17) the no- enzyme control.

15. Combine and mix the following reagents:

    8.50 µL - 10x Primer Removal Buffer (SR401)
    1.70 µL - Primer Removal Solution (SR400)
    X µL - of purified SeqPlex amplification product (**step 14**)
    Y µL - Water (W4502)
    80.75 µL Total reaction volume

16. Transfer 9.5 µL of the mix in **step 15** to a different reaction tube and add 0.5 µL water (W4502). This is the "no-enzyme" reaction. Mix thoroughly by pipeting. Spin down residue from top and sides of reaction tubes by brief centrifugation.

17. To the remaining 71.25 µL, add 3.75 µL Primer Removal Enzyme (SR402). This is the primer removal reaction. Mix thoroughly by pipeting. Spin down residue from top and sides of reaction tubes by brief centrifugation.

18. Incubate both primer removal and no-enzyme control reactions as follows:

    37 °C for 60 minutes
    65 °C for 20 minutes
    4 °C hold

19. Remove samples from thermocycler and centrifuge briefly.

20. Reserve 2 µL each of the primer removal reaction, and the entire no-enzyme control reaction for the **Quality Control** assays below. Keep on ice or store at –20 °C for up to one month.

21. Purify the remaining primer removal reaction products using the GenElute PCR Clean-up Kit as described previously in **step 13**, or store unpurified samples at –20 °C for up to one month.

22. Primer removal reaction product yield and concentration is quantified by measuring absorbance. One A_{260} unit is equivalent to 50 ng/µL DNA.

23. **Check with Qubit at this point instead and check primer removal if time allows (steps below).**

Primer Removal

The efficiency of primer removal can be estimated by qPCR using the 5× Amplification Mix and Amplification Enzyme. Use unpurified primer removal reaction and corresponding control reaction for these assays (**step 20**). Sufficient amplification mix and enzyme are provided in the kit for a 15-µL qPCR test reaction for both the primer removal reaction and "no-primer removal-enzyme" control reaction. A 1/1,000,000 dilution of each primer-removal and control reaction is used for this assay. Expect primer removal to be greater than 90%.

For the Primer Removal QC, combine reagents at following scale per single qPCR reaction:

    3 µL 5× Amplification Mix (A5112)
    0.15 µL Amplification Enzyme (5237)
    1.85 µL 1/10,000 dilution, SYBR Green in 10 mM Tris-HCl, pH 8.0 (Catalog No. T3038)
    10 µL 1/1,000,000 dilution cDNA (from primer removal reaction or no enzyme control)

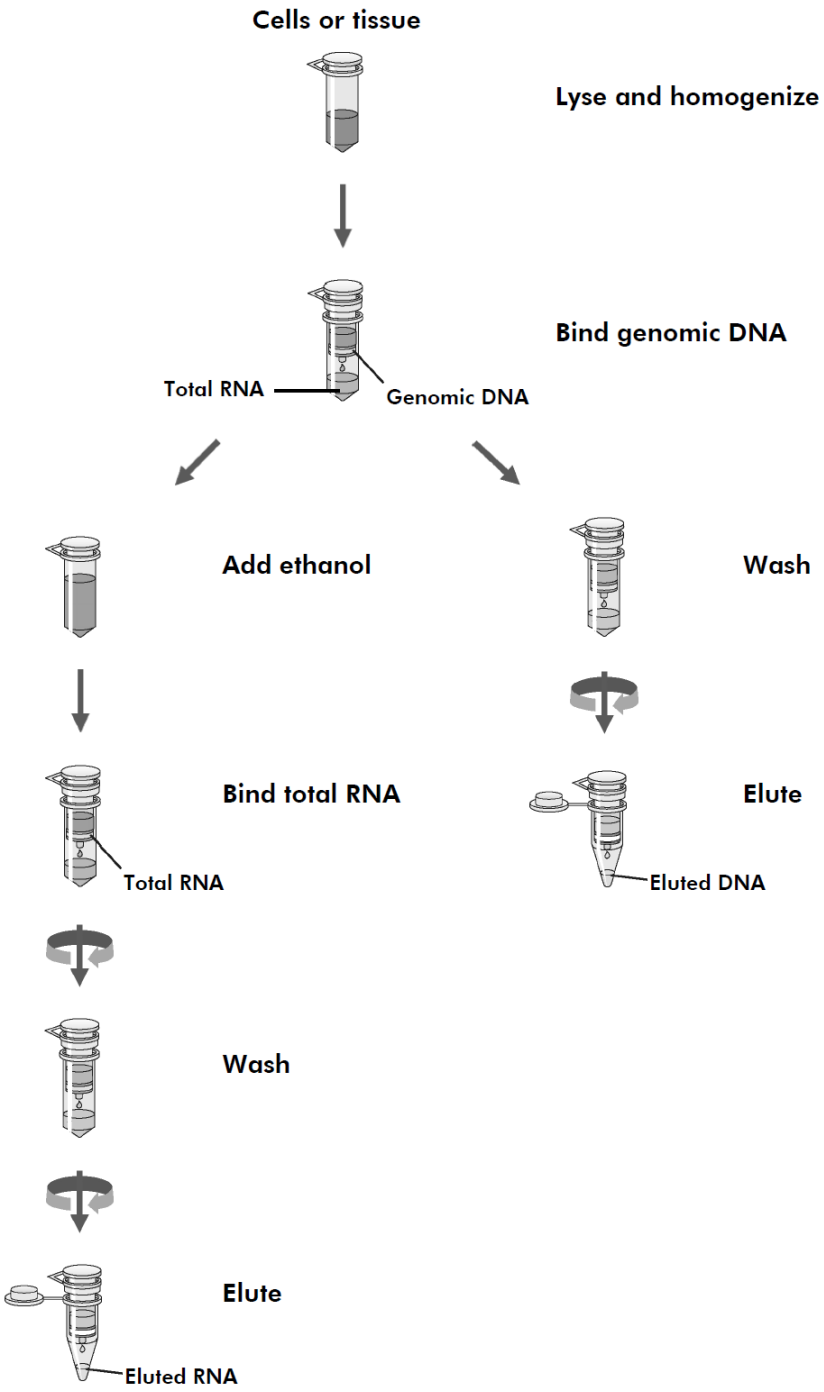
Amplification conditions:

    1 cycle
        94 °C, 2.5 minutes
    40 cycles
        94 °C for 30 seconds
        70 °C for 5 minutes (read)

Expect a ΔCt of 3–7 as an estimate of successful primer removal:

$$(Ct)_{\text{no primer removal enzyme}} - (Ct)_{\text{plus primer removal enzyme}}$$

Appendix 2: AllPrep DNA/RNA Extraction Protocol



1. Add 600µl of Buffer RLT Plus to a 1.5ml tube with 2 pellets from the same sample and homogenize via pipetting up and down.
2. Transfer the homogenized lysate to the AllPrep DNA spin-column placed in a 2ml collection tube. Close the lid and centrifuge for 30s at 8000 x g.
3. Use the flow through for RNA purification, set aside.
4. Place the AllPrep DNA spin column in a new 2ml collection tube. Add 500µl Buffer AW1 to the spin column. Close the lid ad centrifuge for 15s at 8000 x g. Discard the flow through.
5. Add 500µl Buffer AW2 to the AllPrep DNA spin column. Close the lid and centrifuge 2 min at full speed.
6. Place AllPrep DNA spin column in new 1.5ml tube. Add 100µl Buffer EB directly to the spin column membrane and close the lid. Incubate at RT for 1 min. Centrifuge for 1 min at 8000 x g.
7. Repeat step 6 using the eluate from step 6.
8. Add 600µl of 70% ethanol to the flow-through that was set-aside in step 3. Mix by pipetting.
9. Transfer 700µl of the sample to an RNeasy spin column placed in a 2ml collection tube. Centrifuge for 15s at 8000 x g. Discard flow through.
10. Repeat step 9 until entire sample is loaded onto the column.
11. Add 700µl Buffer RW1 to the RNeasy spin column. Close the lid and centrifuge for 15s at 8000 x g. Discard flow through.
12. Add 500µl Buffer RPE to the RNeasy spin column. Close the lid and centrifuge for 15s at 8000 x g. Discard flow through.
13. Add 500µl Buffer RPE to the RNeasy spin column. Close the lid and centrifuge for 2 min at 8000 x g. Discard flow through.
14. Place RNeasy spin column in new 1.5ml tube. Add 30µl RNase-free water directly to spin column membrane. Close the lid and centrifuge for 1 min at 8000 x g. Retain flow through.
15. Repeat step 14 using another 30µl of RNase-free water. Combine both eluates.

Appendix 3: Round 1 Viral Enrichment and Extraction Protocol (courtesy of CARACAL)

Filtration, Nuclease Treatment and Nucleic Acid Extraction from Animal/Human Samples

Preparation

1. Wear gloves for potential hazardous material.
2. Work in the hood.
3. Wipe the hood, pipettes and all surfaces with 10% bleach prior to starting.
4. Avoid cross-contamination of samples by opening one tube at a time.

Homogenizing Tissues

Option A –

a. Cut tissue into small sections and add small amount to clean glass mortar and pestle.
b. Add small amount of 1X PBS and grind solution of a liquid consistency.

Option B –

a. Freeze tissue with liquid nitrogen and break into small pieces.
b. CAREFULLY use small amount of liquid nitrogen in glass mortar and pestle to pulverize small pieces of tissue.
c. Evaporate off liquid nitrogen and suspend in ~150-300 ul of 1X PBS.

Option C-

Use a mini bead-beater or other method for liquefying tissue.

Filtration

1. Place 220 ul of **serum, liquefied tissue, or sample from swab in RNA later** into a ultrafree-MC HV 0.45µm sterile filter (Millipore, UFC30HV0S, red dot)
[If available volume is less than 160 ul, bring to 160 ul with sterile PBS.]
2. Centrifuge at 6,000 rpm for 5 minutes
3. Transfer filtrate to a new 1.5 ml tube. Discard filter.

Nuclease treatment

1. Make the following master mix per sample:

7 µl	Turbo DNase
3 µl	Baseline Zero DNase
3 µl	Benzonase
3 µl	RNase A
14 µl	10×Turbo buffer
30 ul	total/sample

2. Add 30 ul of nuclease mixture to serum filtrate.
3. Incubate at 37°C for 2 hours (mix by gentle tap at 1 hour).

Viral nucleic acid extraction

Extract nucleic acids from nuclease treated sample using the **QIAamp DSP Virus Kit** according to kit instructions with the following modifications:

Do not add carrier RNA to the AL buffer.
If you have 250 ul of sample instead of the 500 ul the kit recommends, add only 250 ul AL to the sample (step 3). (For all other steps use the recommended 1X volume of wash buffers and ethanol.
At Step 10 for elution, after addition of elution buffer, incubate column 5 minutes before centrifugation and add 0.5 µl RNase Inhibitor into the tube before the spin.
Elute in 60 ul AVE or water.

Freeze extracts immediately at -80 °C.

Appendix 4: Nextera DNA Flex Library Prep Kit Protocol
(Protocol Images: Illumina DNA Prep Documentation, 2020)

Tagment Genomic DNA

This step uses the Bead-Linked Transposomes (BLT) to tagment DNA, which is a process that fragments and tags the DNA with adapter sequences.

Consumables

- Bead-Linked Transposomes (BLT)
- Tagmentation Buffer 1 (TB1)
- Nuclease-free water
- 96-well PCR plate
- Microseal 'B' adhesive seal
- 1.7 ml microcentrifuge tubes
- 8-tube strip
- Pipette tips
  - 20 µl multichannel pipettes
  - 200 µl multichannel pipettes

About Reagents

- BLT must be stored at temperatures above 2°C. Do not use BLT that has been stored below 2°C.

Preparation

1 Prepare the following consumables:

Item	Storage	Instructions
BLT	2°C to 8° C	Bring to room temperature. Vortex to mix. Do not centrifuge before pipetting.
TB1	-25°C to -15°C	Bring to room temperature. Vortex to mix.

2 Save the following TAG program on the thermal cycler:
- Choose the preheat lid option and set to 100°C
- Set the reaction volume to 50 µl
- 55°C for 15 minutes
- Hold at 10°C

Procedure

1 Add 2–30 µl DNA to each well of a 96-well PCR plate so that the total input amount is 100–500 ng.

2 If DNA volume < 30 µl, add nuclease-free water to the DNA samples to bring the total volume to 30 µl.

3 Vortex BLT (yellow cap) vigorously for 10 seconds to resuspend. Repeat as necessary.

4 Combine the following volumes to prepare the tagmentation master mix. Multiply each volume by the number of samples being processed.
- BLT (11 µl)
- TB1 (11 µl)

Reagent overage is included in the volume to ensure accurate pipetting.

5 Vortex the tagmentation master mix thoroughly to resuspend.

6 Divide the tagmentation master mix volume equally into an 8-tube strip.

7 Using a 200 µl multichannel pipette, transfer 20 µl tagmentation master mix to each well of the plate containing a sample. Use fresh tips for each sample column.

8 Discard the 8-tube strip after the tagmentation master mix has been dispensed.

9 Pipette each sample 10 times to resuspend. Use fresh tips for each sample column.

10 Seal the plate with Microseal 'B', place on the preprogrammed thermal cycler, and run the TAG program.

Post Tagmentation Cleanup

This step washes the adapter-tagged DNA on the BLT before PCR amplification.

Consumables

- Tagment Stop Buffer (TSB)
- Tagment Wash Buffer (TWB)
- 96-well plate magnet
- Microseal 'B' adhesive seal
- Pipette tips
  - 20 µl multichannel pipettes
  - 200 µl multichannel pipettes

Preparation

1 Prepare the following consumables:

Item	Storage	Instructions
TSB	15°C to 30°C	If precipitates are observed, heat at 37°C for 10 minutes, and then vortex until precipitates are dissolved. Use at room temperature.
TWB	15°C to 30°C	Use at room temperature.

2 Save the following PTC program on the thermal cycler:
- Choose the preheat lid option and set to 100°C
- Set the reaction volume to 60 µl
- 37°C for 15 minutes
- Hold at 10°C

Procedure

1 Add 10 µl TSB to the tagmentation reaction.

2 Slowly pipette each well 10 times to resuspend the beads.

3 Seal the plate with Microseal 'B', place on the preprogrammed thermal cycler, and run the PTC program.

4 Place the plate on the magnetic stand and wait until liquid is clear (~3 minutes).

5 Using a multichannel pipette, remove and discard supernatant.

6 Wash two times as follows:

   a Remove the sample plate from the magnetic stand and use a deliberately slow pipetting technique to add 100 µl TWB directly onto the beads. A deliberately slow pipetting technique minimizes the potential of TWB foaming to avoid incorrect volume aspiration and incomplete mixing.
   b Pipette slowly until beads are fully resuspended.
   c Place the plate on the magnetic stand and wait until the liquid is clear (~3 minutes).
   d Using a multichannel pipette, remove and discard supernatant.

7 Remove the plate from the magnetic stand and use a deliberately slow pipetting technique to add 100 µl TWB directly onto the beads.

8 Pipette each well slowly to resuspend the beads.

9 Seal the plate and place on the magnetic stand until the liquid is clear (~3 minutes). Keep on the magnetic stand until step 4 of the *Procedure* section in *Amplify Tagmented DNA*.
The TWB remains in the wells to prevent overdrying of the beads.

Amplify Tagmented DNA

This step amplifies the tagmented DNA using a limited-cycle PCR program. The PCR step adds Index 1 (i7) adapters, Index 2 (i5) adapters, and sequences required for sequencing cluster generation. To confirm the indexes selected for low plexity pooling have the appropriate color balance, see the *Index Adapters Pooling Guide (document # 1000000041074)*.

Index adapter tubes or plates are ordered separately from the library prep components. For a list of compatible index adapters for use with this protocol, see *Kit Contents* on page 24.

Consumables

- Enhanced PCR Mix (EPM)
- Index adapters (tubes or plate)
- Nuclease-free water
- Microseal 'B' adhesive seal
- 1.7 ml microcentrifuge tubes
- Pipette tips
  - 20 µl multichannel pipettes
  - 200 µl multichannel pipettes

About Reagents

- Index adapter plates
  - A well may contain >10 µl of index adapters.
  - Do not add samples to the index adapter plate.
  - Each well of the index plate is single use only.
- Index adapter tubes
  - Open only one index adapter tube at a time to prevent misplacing caps. Alternatively, use fresh caps after opening each tube.

Preparation

1 Prepare the following consumables:

Item	Storage	Instructions
EPM	-25°C to-15°C	Thaw on ice. Invert to mix, then briefly centrifuge.
Index Adapters	-25°C to-15°C	Thaw at room temperature. [Tubes] Vortex to mix, then centrifuge briefly. [Plates] Spin briefly before use.

2 Save the following BLT PCR program on a thermal cycler using the appropriate number of PCR cycles, indicated in the table below.

- Choose the preheat lid option and set to 100°C
- 68°C for 3 minutes
- 98°C for 3 minutes
- (X) cycles of:
  - 98°C for 45 seconds
  - 62°C for 30 seconds
  - 68°C for 2 minutes
- 68°C for 1 minutes
- Hold at 10°C

Total DNA Input (ng)	Number of PCR Cycles (X)
1–9	12
10–24	8
25–49	6
50–99	5
100–500	5
Blood/Saliva	5

Procedure

1 Combine the following volumes to prepare the PCR master mix. Multiply each volume by the number of samples being processed.

- EPM (22 µl)
- Nuclease-free water (22 µl)

Reagent overage is included in the volume to ensure accurate pipetting.

2 Vortex, and then centrifuge the PCR master mix at 280 × g for 10 seconds.

3 With the plate on the magnetic stand, use a 200 µl multichannel pipette to remove and discard supernatant.

Foam that remains on the well walls does not adversely affect the library.

4 Remove from the magnet.

5 Immediately add 40 µl PCR master mix directly onto the beads in each sample well.

6 Immediately pipette to mix until the beads are fully resuspended. Alternatively, seal the plate and use a plate shaker at 1600 rpm for 1 minute.

7 Seal the sample plate and centrifuge at 280 × g for 3 seconds.

8 Add the appropriate index adapters to each sample.

Index Kit Type	Kit Configuration	Volume of Index Adapter per Sample
24 plex (dual index)	Individual tubes	5 µl i7 adapter 5 µl i5 adapter
96 plex (dual index)	96-well plate	10 µl pre-paired i7 and i5 index adapters

9 Using a pipette set to 40 µl, pipette 10 times to mix. Alternatively, seal the plate and use a plate shaker at 1600 rpm for 1 minute.

10 Seal the plate with Microseal 'B', and then centrifuge at 280 × g for 30 seconds.

11 Place on the thermal cycler and run the BLT PCR program.

SAFE STOPPING POINT

If you are stopping, store at 2°C to 8°C for up to 3 days.

Clean Up Libraries

This step uses double-sided bead purification procedure to purify the amplified libraries.

Consumables

- Sample Purification Beads (SPB)
- Resuspension Buffer (RSB)
- Freshly prepared 80% ethanol (EtOH)
- 96-well 0.8 ml Polypropylene Deepwell Storage Plate (midi plate) (2)
- 96-well PCR plate
- Microseal 'B' adhesive seal
- Microseal 'F' foil seal
- 1.7 ml microcentrifuge tubes
- Nuclease-free water

About Reagents

- Sample Purification Beads
  - Must be at room temperature before use
  - Vortex before each use
  - Vortex frequently to make sure that beads are evenly distributed
  - Aspirate and dispense slowly due to the viscosity of the solution

Preparation

1 Prepare the following consumables:

Item	Storage	Instructions
SPB	2°C to 8°C	Let stand at room temperature for 30 minutes. Vortex and invert to mix.
RSB	-25°C to -15°C	Thaw and bring to room temperature. Vortex to mix.

2 Prepare fresh 80% EtOH from absolute ethanol.

Procedure

1 Centrifuge at 280 × g for 1 minute to collect contents at the bottom of the well.

2 Place the plate on the magnetic stand and wait until the liquid is clear (~5 minutes).

3 Transfer 45 µl supernatant from each well of the PCR plate to the corresponding well of a new midi plate.

4 Vortex and invert SPB multiple times to resuspend.

5 For standard DNA input, perform the following steps.

   a Add 40 µl nuclease-free water to each well containing supernatant.
   b Add 45 µl SPB to each well containing supernatant.
   c Pipette each well 10 times to mix. Alternatively, seal the plate and use a plate shaker at 1600 rpm for 1 minute.
   d Seal the plate and incubate at room temperature for 5 minutes.
   e Place on the magnetic stand and wait until the liquid is clear (~5 minutes).
   f During incubation, thoroughly vortex the SPB (*undiluted* stock tube), and then add 15 µl to each well of a *new* midi plate.
   g Transfer 125 µl supernatant from each well of the first plate into the corresponding well of the second plate (containing 15 µl *undiluted* SPB).
   h Pipette each well in the second plate 10 times to mix. Alternatively, seal the plate and use a plate shaker at 1600 rpm for 1 minute.
   i Discard the first plate.

6 For small PCR amplicon input, perform the following steps.

   a Add 81 µl SPB to each midi plate well containing supernatant.
   b Pipette each well 10 times to mix. Alternatively, seal the plate and use a plate shaker at 1600 rpm for 1 minute.

7 Incubate the sealed midi plate at room temperature for 5 minutes.

8 Place on the magnetic stand and wait until the liquid is clear (~5 minutes).

9 Without disturbing the beads, remove and discard supernatant.

10 Wash two times as follows.

   a With the plate on the magnetic stand, add 200 µl fresh 80% EtOH without mixing.
   b Incubate for 30 seconds.
   c Without disturbing the beads, remove and discard supernatant.

11 Use a 20 µl pipette to remove and discard residual EtOH.

12 Air-dry on the magnetic stand for 5 minutes.

13 Remove from the magnetic stand.

14 Add 32 µl RSB to the beads.

15 Pipette to resuspend.

16 Incubate at room temperature for 2 minutes.

17 Place the plate on the magnetic stand and wait until the liquid is clear (~2 minutes).

18 Transfer 30 µl supernatant to a new 96-well PCR plate.

SAFE STOPPING POINT

If you are stopping, seal the plate with Microseal 'B' adhesive or Microseal 'F' foil seal, and store at -25°C to -15°C for up to 30 days.

Pool Libraries

When the DNA input is 100-500 ng, quantifying and normalizing individual libraries generated in the same experiment is not necessary. However, the final yield of libraries generated in separate experiments can vary slightly.

To achieve optimal cluster density, pool equal library volumes and quantify the pool before sequencing.

DNA Inputs of 100–500 ng

1 Combine 5 µl of each library (up to 384 libraries) in a 1.7 ml microcentrifuge tube.

2 Vortex to mix, and then centrifuge.

3 Quantify the library pool using a dsDNA fluorescent dye method, such as Qubit or PicoGreen.

For DNA Inputs of < 100 ng

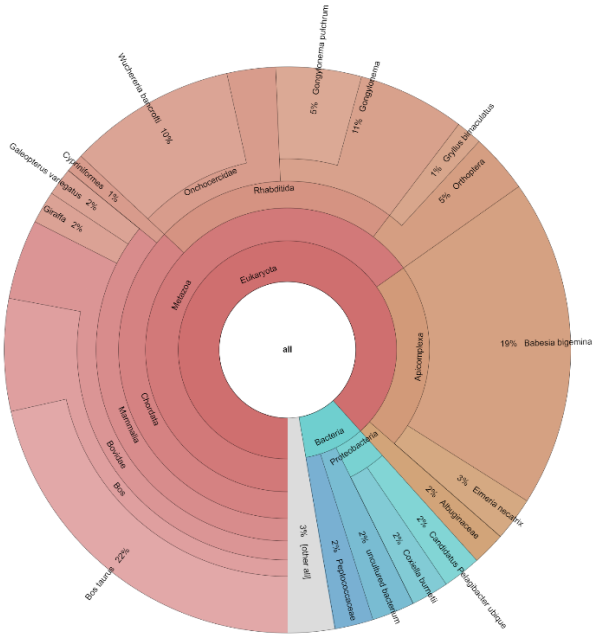
1 Quantify each library individually using Qubit or PicoGreen.

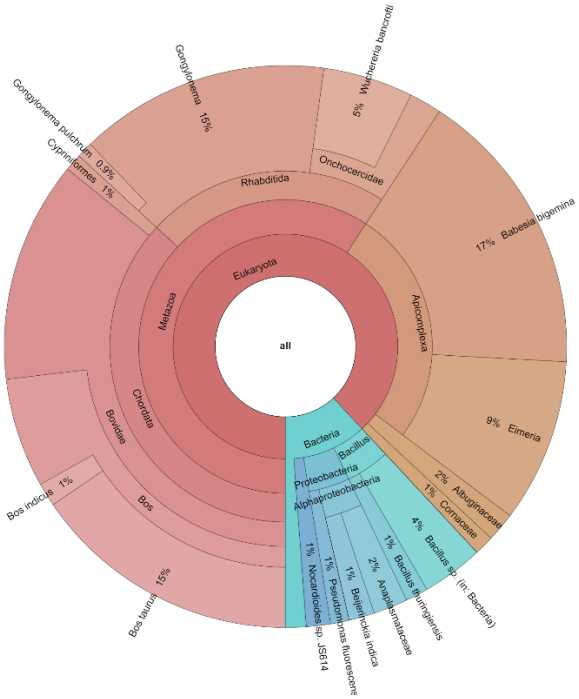
Check Library Quality (Optional)

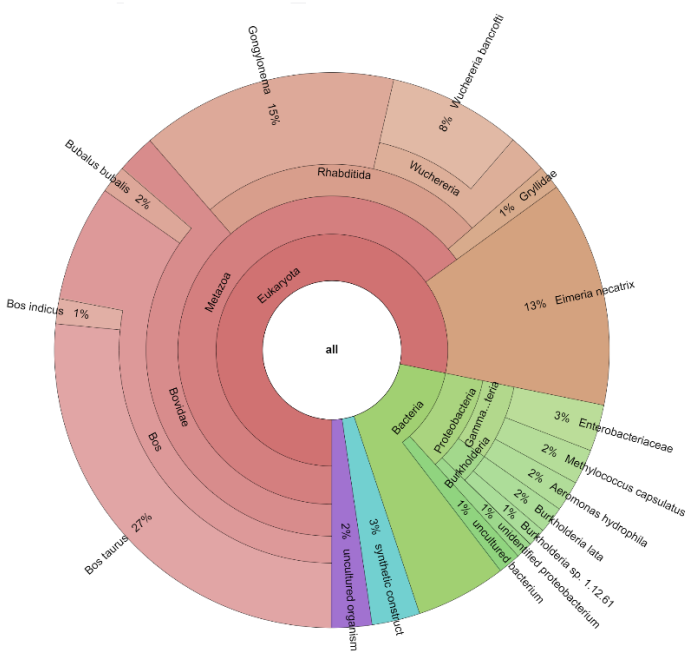
1 Run 1 µl library or pooled libraries on one of the following instruments:
   - Advanced Analytical Fragment Analyzer with the HS-NGS High Sensitivity 474 kit.
     - Add 1 µl RSB to the library to achieve the 2 µl volume required for Fragment Analyzer.
   - Agilent 2100 Bioanalyzer with a High Sensitivity DNA kit.

   The following figures show typical library size profiles with an average fragment size of 600 bp when analyzed with a size range of 150–1500 bp.

Appendix 6: KMA Output Summary from 201703-Seep-B Sample







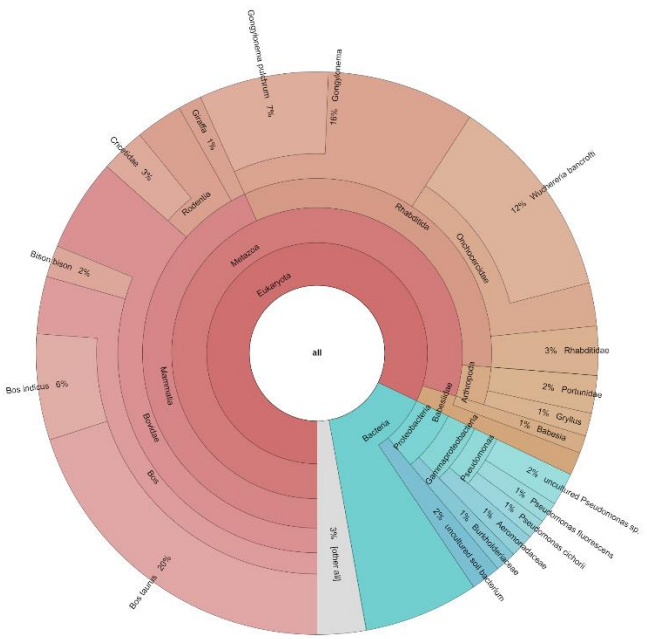


Table 27: Round 2 DNA Sequencing Diversity Scores

Sample	Total Counts	Richness	Shannon H	exp(H)	Evenness
201703-Seep-B_S27	4066	262	1.624295172	5.074840887	0.291701631
201703-Seep-V_S34	24762	3241	5.486468715	241.4032363	0.678712883
201703-TR13-B_S29	4108	161	2.02898843	7.606388073	0.399296786
201703-TR13-V_S33	38569	1753	3.768127032	43.29889139	0.504496547
201703-TR1-B_S2	3875	98	1.155942011	3.17701479	0.25211564
201703-TR1-V_S30	25972	1505	3.468899876	32.10140748	0.474117001
201703-TR21-B_S6	3457	128	1.371793745	3.942416047	0.282725719
201703-TR21-V_S31	24235	1749	3.732083355	41.76603104	0.499823702
201703-TR25-B_S3	4860	344	2.330607804	10.28419041	0.399032836
201703-TR25-V_S37	59986	2382	4.241451259	69.50865371	0.545475465
201703-TR27-B_S5	19066	1049	3.158159752	23.52726008	0.454046108
201703-TR27-V_S32	51451	2120	3.702912272	40.56526961	0.483461212
201703-TR31-B_S4	3248	87	1.115160719	3.050058342	0.249705254
201703-TR31-V_S38	34538	2842	4.206631567	67.1300355	0.528985448
201703-TR35-V_S36	11245	1417	4.126867984	61.98348536	0.568729181
201703-TR49-B_S28	13742	809	2.678535417	14.56374784	0.400032237
201703-TR49-V_S35	44024	2331	4.434687563	84.32577426	0.571918682
201708-Seep-V_S51	2528	1535	7.064065558	1169.188921	0.962894015
201708-TR13-B_S20	451	137	3.746736255	42.38253025	0.761534711
201708-TR13-V_S39	407	311	5.273218328	195.0426647	0.918712296
201708-TR1-B_S19	12115	1910	4.579516093	97.46721768	0.606168346
201708-TR1-V_S41	848	572	5.715795625	303.6256793	0.900247361
201708-TR21-B_S8	1515	387	4.294938664	73.32771667	0.720817814
201708-TR21-V_S40	1073	785	6.150232447	468.8263515	0.922670908
201708-TR25-B_S21	20757	1725	4.098733313	60.26390375	0.549945395
201708-TR25-V_S56	9860	1290	4.092158545	59.8689822	0.571339212
201708-TR27-B_S22	3463	605	4.059667518	57.95503889	0.633805265
201708-TR27-V_S55	1605	932	6.07420069	434.5020618	0.888387454
201708-TR31-B_S23	6519	1113	3.426215165	30.76000064	0.488425637
201708-TR31-V_S57	7917	1539	3.743332638	42.23852158	0.510068088
201708-TR35-B_S7	4536	62	0.771269772	2.162510406	0.186877795
201708-TR35-V_S53	4524	1591	5.278098496	195.996832	0.71595415
201708-TR49-B_S26	213	159	4.550789229	94.70712442	0.897785606
201708-TR49-V_S42	1252	926	6.45007245	632.7481335	0.944252848
201708-TR53-B_S25	6388	1174	3.905335536	49.6667421	0.552524123
201708-TR53-V_S49	4173	1112	4.541903857	93.86934393	0.647556115
201910-TR13-B_S11	5947	1502	4.982883134	145.894409	0.681228675
201910-TR13-V_S44	6457	1698	5.122669308	167.7826359	0.688789453
201910-TR1-B_S10	6764	1980	5.629522333	278.5290417	0.741619286
201910-TR1-V_S43	4370	1753	6.062907963	429.6229498	0.811733816
201910-TR21-B_S12	3494	961	4.550625104	94.6915819	0.662586206
201910-TR21-V_S46	2755	1208	5.756935082	316.3771713	0.811210526
201910-TR25-B_S13	4626	1521	5.243198185	189.2744702	0.715587547
201910-TR25-V_S47	5395	2141	6.457915463	637.7302974	0.842077408
201910-TR27-B_S14	4617	1266	4.573151152	96.84881476	0.640173006
201910-TR27-V_S45	5350	2084	5.975500231	393.6649752	0.781924301
201910-TR31-B_S15	460	271	5.010052549	149.9126137	0.894313867
201910-TR31-V_S52	8265	2703	6.272459384	529.7787064	0.793769436
201910-TR35-B_S16	1921	804	5.119002515	167.1685384	0.765218111
201910-TR35-V_S54	4684	1984	6.429390836	619.7962756	0.846766847
201910-TR49-B_S17	3436	1299	5.211920785	183.4460805	0.726972567
201910-TR49-V_S48	2184	1298	6.538564318	691.2933878	0.912114312

Table 28: Round 2 DNA Sequencing Illumina Quality Scores

Sample ID	FASTQ GZIP File Name	DNA Reads	Average Quality Score
201703-Seep-B_S27	201703-Seep-B_S27_L001_R1_001.fastq.gz	1487873	35.77659921
201703-Seep-B_S27	201703-Seep-B_S27_L001_R2_001.fastq.gz	1487873	35.59471339
201703-Seep-B_S27	201703-Seep-B_S27_L002_R1_001.fastq.gz	1445147	35.70403841
201703-Seep-B_S27	201703-Seep-B_S27_L002_R2_001.fastq.gz	1445147	35.52473001
201703-Seep-V_S34	201703-Seep-V_S34_L001_R1_001.fastq.gz	11826489	35.82458403
201703-Seep-V_S34	201703-Seep-V_S34_L001_R2_001.fastq.gz	11826489	35.4533348
201703-Seep-V_S34	201703-Seep-V_S34_L002_R1_001.fastq.gz	11611083	35.75176786
201703-Seep-V_S34	201703-Seep-V_S34_L002_R2_001.fastq.gz	11611083	35.37196039
201703-TR1-V_S30	201703-TR1-V_S30_L001_R1_001.fastq.gz	8915826	35.65668419
201703-TR1-V_S30	201703-TR1-V_S30_L002_R1_001.fastq.gz	8702257	35.58416673
201703-TR13-B_S29	201703-TR13-B_S29_L001_R2_001.fastq.gz	885992	35.66748007
201703-TR13-B_S29	201703-TR13-B_S29_L002_R1_001.fastq.gz	873483	35.70755699
201703-TR13-B_S29	201703-TR13-B_S29_L002_R2_001.fastq.gz	873483	35.59941521
201703-TR13-V_S33	201703-TR13-V_S33_L001_R1_001.fastq.gz	16032762	35.85338858
201703-TR13-V_S33	201703-TR13-V_S33_L001_R2_001.fastq.gz	16032762	35.47234806
201703-TR13-V_S33	201703-TR13-V_S33_L002_R1_001.fastq.gz	15699478	35.78096699
201703-TR13-V_S33	201703-TR13-V_S33_L002_R2_001.fastq.gz	15699478	35.40291926
201703-TR21-B_S6	201703-TR21-B_S6_L001_R1_001.fastq.gz	866580	35.57112788
201703-TR21-B_S6	201703-TR21-B_S6_L001_R2_001.fastq.gz	866580	35.43428535
201703-TR21-B_S6	201703-TR21-B_S6_L002_R1_001.fastq.gz	853297	35.49129436
201703-TR21-B_S6	201703-TR21-B_S6_L002_R2_001.fastq.gz	853297	35.3336271
201703-TR21-V_S31	201703-TR21-V_S31_L001_R1_001.fastq.gz	13209699	35.76956568
201703-TR21-V_S31	201703-TR21-V_S31_L001_R2_001.fastq.gz	13209699	35.53704009
201703-TR21-V_S31	201703-TR21-V_S31_L002_R1_001.fastq.gz	12808823	35.69321186
201703-TR21-V_S31	201703-TR21-V_S31_L002_R2_001.fastq.gz	12808823	35.47049186
201703-TR25-B_S3	201703-TR25-B_S3_L001_R1_001.fastq.gz	1284541	35.67499597
201703-TR25-B_S3	201703-TR25-B_S3_L001_R2_001.fastq.gz	1284541	35.59535741
201703-TR25-B_S3	201703-TR25-B_S3_L002_R1_001.fastq.gz	1255802	35.600828
201703-TR25-B_S3	201703-TR25-B_S3_L002_R2_001.fastq.gz	1255802	35.51048653
201703-TR25-V_S37	201703-TR25-V_S37_L001_R1_001.fastq.gz	25758854	35.57466877
201703-TR25-V_S37	201703-TR25-V_S37_L001_R2_001.fastq.gz	25758854	35.41011102
201703-TR25-V_S37	201703-TR25-V_S37_L002_R1_001.fastq.gz	24860540	35.49468278
201703-TR25-V_S37	201703-TR25-V_S37_L002_R2_001.fastq.gz	24860540	35.35064484
201703-TR27-B_S5	201703-TR27-B_S5_L001_R1_001.fastq.gz	4024690	35.71761428
201703-TR27-B_S5	201703-TR27-B_S5_L001_R2_001.fastq.gz	4024690	35.69504011
201703-TR27-B_S5	201703-TR27-B_S5_L002_R1_001.fastq.gz	3958991	35.63156067
201703-TR27-B_S5	201703-TR27-B_S5_L002_R2_001.fastq.gz	3958991	35.62069805
201703-TR27-V_S32	201703-TR27-V_S32_L001_R1_001.fastq.gz	12536386	35.77530047
201703-TR27-V_S32	201703-TR27-V_S32_L001_R2_001.fastq.gz	12536386	35.45014831
201703-TR27-V_S32	201703-TR27-V_S32_L002_R1_001.fastq.gz	12007363	35.70781869
201703-TR27-V_S32	201703-TR27-V_S32_L002_R2_001.fastq.gz	12007363	35.39515679
201703-TR31-B_S4	201703-TR31-B_S4_L001_R1_001.fastq.gz	950302	35.75344996

201703-TR31-B_S4	201703-TR31-B_S4_L001_R2_001.fastq.gz	950302	35.3731056
201703-TR31-B_S4	201703-TR31-B_S4_L002_R1_001.fastq.gz	926825	35.68592669
201703-TR31-B_S4	201703-TR31-B_S4_L002_R2_001.fastq.gz	926825	35.28540771
201703-TR31-V_S38	201703-TR31-V_S38_L001_R1_001.fastq.gz	39448019	35.80263998
201703-TR31-V_S38	201703-TR31-V_S38_L001_R2_001.fastq.gz	39448019	35.60649502
201703-TR31-V_S38	201703-TR31-V_S38_L002_R1_001.fastq.gz	38396195	35.72653316
201703-TR31-V_S38	201703-TR31-V_S38_L002_R2_001.fastq.gz	38396195	35.5404871
201703-TR35-B_S24	201703-TR35-B_S24_L001_R1_001.fastq.gz	307012	35.62840215
201703-TR35-B_S24	201703-TR35-B_S24_L001_R2_001.fastq.gz	307012	35.56281188
201703-TR35-B_S24	201703-TR35-B_S24_L002_R1_001.fastq.gz	296625	35.56432196
201703-TR35-B_S24	201703-TR35-B_S24_L002_R2_001.fastq.gz	296625	35.48489844
201703-TR35-V_S36	201703-TR35-V_S36_L001_R1_001.fastq.gz	21645933	35.83832866
201703-TR35-V_S36	201703-TR35-V_S36_L001_R2_001.fastq.gz	21645933	35.49760886
201703-TR35-V_S36	201703-TR35-V_S36_L002_R1_001.fastq.gz	20862524	35.76082252
201703-TR35-V_S36	201703-TR35-V_S36_L002_R2_001.fastq.gz	20862524	35.42783946
201703-TR49-B_S28	201703-TR49-B_S28_L001_R1_001.fastq.gz	2817686	35.78030199
201703-TR49-B_S28	201703-TR49-B_S28_L001_R2_001.fastq.gz	2817686	35.59189562
201703-TR49-B_S28	201703-TR49-B_S28_L002_R1_001.fastq.gz	2789988	35.70844785
201703-TR49-B_S28	201703-TR49-B_S28_L002_R2_001.fastq.gz	2789988	35.51930546
201703-TR49-V_S35	201703-TR49-V_S35_L001_R1_001.fastq.gz	22594410	35.83168863
201703-TR49-V_S35	201703-TR49-V_S35_L001_R2_001.fastq.gz	22594410	35.55388581
201703-TR49-V_S35	201703-TR49-V_S35_L002_R1_001.fastq.gz	21943548	35.76087226
201703-TR49-V_S35	201703-TR49-V_S35_L002_R2_001.fastq.gz	21943548	35.50293271
201708-Seep-B_S9	201708-Seep-B_S9_L001_R1_001.fastq.gz	27053	35.04624256
201708-Seep-B_S9	201708-Seep-B_S9_L001_R2_001.fastq.gz	27053	35.29538314
201708-Seep-B_S9	201708-Seep-B_S9_L002_R1_001.fastq.gz	21381	34.93054581
201708-Seep-B_S9	201708-Seep-B_S9_L002_R2_001.fastq.gz	21381	35.22314204
201708-Seep-V_S51	201708-Seep-V_S51_L001_R1_001.fastq.gz	4953160	35.71960869
201708-Seep-V_S51	201708-Seep-V_S51_L001_R2_001.fastq.gz	4953160	34.84817692
201708-Seep-V_S51	201708-Seep-V_S51_L002_R1_001.fastq.gz	4736636	35.65093645
201708-Seep-V_S51	201708-Seep-V_S51_L002_R2_001.fastq.gz	4736636	34.63027896
201708-TR1-B_S19	201708-TR1-B_S19_L001_R1_001.fastq.gz	8952835	35.75100658
201708-TR1-B_S19	201708-TR1-B_S19_L001_R2_001.fastq.gz	8952835	35.45332557
201708-TR1-B_S19	201708-TR1-B_S19_L002_R1_001.fastq.gz	8885483	35.67055567
201708-TR1-B_S19	201708-TR1-B_S19_L002_R2_001.fastq.gz	8885483	35.33982598
201708-TR1-V_S41	201708-TR1-V_S41_L001_R1_001.fastq.gz	4770007	35.70003839
201708-TR1-V_S41	201708-TR1-V_S41_L001_R2_001.fastq.gz	4770007	34.78699298
201708-TR1-V_S41	201708-TR1-V_S41_L002_R1_001.fastq.gz	4662548	35.6262387
201708-TR1-V_S41	201708-TR1-V_S41_L002_R2_001.fastq.gz	4662548	34.56574903
201708-TR13-B_S20	201708-TR13-B_S20_L001_R1_001.fastq.gz	1100643	35.63192607
201708-TR13-B_S20	201708-TR13-B_S20_L001_R2_001.fastq.gz	1100643	35.21597739
201708-TR13-B_S20	201708-TR13-B_S20_L002_R1_001.fastq.gz	1078982	35.56138008
201708-TR13-B_S20	201708-TR13-B_S20_L002_R2_001.fastq.gz	1078982	35.07211056
201708-TR13-V_S39	201708-TR13-V_S39_L001_R1_001.fastq.gz	2385985	35.47003355
201708-TR13-V_S39	201708-TR13-V_S39_L001_R2_001.fastq.gz	2385985	35.33558761

201708-TR13-V_S39	201708-TR13-V_S39_L002_R1_001.fastq.gz	2356168	35.39427324
201708-TR13-V_S39	201708-TR13-V_S39_L002_R2_001.fastq.gz	2356168	35.17714229
201708-TR21-B_S8	201708-TR21-B_S8_L001_R1_001.fastq.gz	2694603	35.74433525
201708-TR21-B_S8	201708-TR21-B_S8_L001_R2_001.fastq.gz	2694603	35.36334518
201708-TR21-B_S8	201708-TR21-B_S8_L002_R1_001.fastq.gz	2624959	35.67175449
201708-TR21-B_S8	201708-TR21-B_S8_L002_R2_001.fastq.gz	2624959	35.23452862
201708-TR21-V_S40	201708-TR21-V_S40_L001_R1_001.fastq.gz	6312699	35.78198295
201708-TR21-V_S40	201708-TR21-V_S40_L001_R2_001.fastq.gz	6312699	34.94301708
201708-TR21-V_S40	201708-TR21-V_S40_L002_R1_001.fastq.gz	6244494	35.71237141
201708-TR21-V_S40	201708-TR21-V_S40_L002_R2_001.fastq.gz	6244494	34.75025422
201708-TR25-B_S21	201708-TR25-B_S21_L001_R1_001.fastq.gz	8025827	35.76441219
201708-TR25-B_S21	201708-TR25-B_S21_L001_R2_001.fastq.gz	8025827	35.42441408
201708-TR25-B_S21	201708-TR25-B_S21_L002_R1_001.fastq.gz	7943977	35.6892603
201708-TR25-B_S21	201708-TR25-B_S21_L002_R2_001.fastq.gz	7943977	35.3002092
201708-TR25-V_S56	201708-TR25-V_S56_L001_R1_001.fastq.gz	6439093	35.7467047
201708-TR25-V_S56	201708-TR25-V_S56_L001_R2_001.fastq.gz	6439093	35.21189708
201708-TR25-V_S56	201708-TR25-V_S56_L002_R1_001.fastq.gz	6310802	35.6739332
201708-TR25-V_S56	201708-TR25-V_S56_L002_R2_001.fastq.gz	6310802	35.04531674
201708-TR27-B_S22	201708-TR27-B_S22_L001_R1_001.fastq.gz	2623110	35.69428427
201708-TR27-B_S22	201708-TR27-B_S22_L001_R2_001.fastq.gz	2623110	35.36977481
201708-TR27-B_S22	201708-TR27-B_S22_L002_R1_001.fastq.gz	2589588	35.61747506
201708-TR27-B_S22	201708-TR27-B_S22_L002_R2_001.fastq.gz	2589588	35.24014476
201708-TR27-V_S55	201708-TR27-V_S55_L001_R1_001.fastq.gz	6015499	35.7560598
201708-TR27-V_S55	201708-TR27-V_S55_L001_R2_001.fastq.gz	6015499	35.06151061
201708-TR27-V_S55	201708-TR27-V_S55_L002_R1_001.fastq.gz	5892443	35.68369045
201708-TR27-V_S55	201708-TR27-V_S55_L002_R2_001.fastq.gz	5892443	34.88050355
201708-TR31-B_S23	201708-TR31-B_S23_L001_R1_001.fastq.gz	6037928	35.73966268
201708-TR31-B_S23	201708-TR31-B_S23_L001_R2_001.fastq.gz	6037928	35.37158128
201708-TR31-B_S23	201708-TR31-B_S23_L002_R1_001.fastq.gz	5971633	35.6605925
201708-TR31-B_S23	201708-TR31-B_S23_L002_R2_001.fastq.gz	5971633	35.22757644
201708-TR31-V_S57	201708-TR31-V_S57_L001_R1_001.fastq.gz	8169331	35.77193714
201708-TR31-V_S57	201708-TR31-V_S57_L001_R2_001.fastq.gz	8169331	35.39073934
201708-TR31-V_S57	201708-TR31-V_S57_L002_R1_001.fastq.gz	8083613	35.69625735
201708-TR31-V_S57	201708-TR31-V_S57_L002_R2_001.fastq.gz	8083613	35.25807569
201708-TR35-B_S7	201708-TR35-B_S7_L001_R1_001.fastq.gz	453966	35.70362098
201708-TR35-B_S7	201708-TR35-B_S7_L001_R2_001.fastq.gz	453966	35.5321588
201708-TR35-B_S7	201708-TR35-B_S7_L002_R1_001.fastq.gz	440758	35.6292909
201708-TR35-B_S7	201708-TR35-B_S7_L002_R2_001.fastq.gz	440758	35.4612395
201708-TR35-V_S53	201708-TR35-V_S53_L001_R1_001.fastq.gz	11531132	35.52874288
201708-TR35-V_S53	201708-TR35-V_S53_L001_R2_001.fastq.gz	11531132	35.46826669
201708-TR35-V_S53	201708-TR35-V_S53_L002_R1_001.fastq.gz	11399897	35.44496665
201708-TR35-V_S53	201708-TR35-V_S53_L002_R2_001.fastq.gz	11399897	35.3293485
201708-TR49-B_S26	201708-TR49-B_S26_L001_R1_001.fastq.gz	1081118	35.73260828
201708-TR49-B_S26	201708-TR49-B_S26_L001_R2_001.fastq.gz	1081118	35.49669694
201708-TR49-B_S26	201708-TR49-B_S26_L002_R1_001.fastq.gz	1046759	35.65700414

201708-TR49-B_S26	201708-TR49-B_S26_L002_R2_001.fastq.gz	1046759	35.40161202
201708-TR49-V_S42	201708-TR49-V_S42_L001_R1_001.fastq.gz	6357613	35.62939581
201708-TR49-V_S42	201708-TR49-V_S42_L001_R2_001.fastq.gz	6357613	35.25729531
201708-TR49-V_S42	201708-TR49-V_S42_L002_R1_001.fastq.gz	6171162	35.5479336
201708-TR49-V_S42	201708-TR49-V_S42_L002_R2_001.fastq.gz	6171162	35.09205268
201708-TR53-B_S25	201708-TR53-B_S25_L001_R1_001.fastq.gz	6187634	35.74384652
201708-TR53-B_S25	201708-TR53-B_S25_L001_R2_001.fastq.gz	6187634	35.31426633
201708-TR53-B_S25	201708-TR53-B_S25_L002_R1_001.fastq.gz	6065080	35.66213488
201708-TR53-B_S25	201708-TR53-B_S25_L002_R2_001.fastq.gz	6065080	35.16112088
201708-TR53-V_S49	201708-TR53-V_S49_L001_R1_001.fastq.gz	6133336	35.64765146
201708-TR53-V_S49	201708-TR53-V_S49_L001_R2_001.fastq.gz	6133336	35.36677022
201708-TR53-V_S49	201708-TR53-V_S49_L002_R1_001.fastq.gz	6041480	35.57034435
201708-TR53-V_S49	201708-TR53-V_S49_L002_R2_001.fastq.gz	6041480	35.2174828
201910-TR1-B_S10	201910-TR1-B_S10_L001_R1_001.fastq.gz	7658811	35.61862501
201910-TR1-B_S10	201910-TR1-B_S10_L001_R2_001.fastq.gz	7658811	35.49089121
201910-TR1-B_S10	201910-TR1-B_S10_L002_R1_001.fastq.gz	7599926	35.54011657
201910-TR1-B_S10	201910-TR1-B_S10_L002_R2_001.fastq.gz	7599926	35.3639946
201910-TR1-V_S43	201910-TR1-V_S43_L001_R1_001.fastq.gz	8838081	35.77724848
201910-TR1-V_S43	201910-TR1-V_S43_L001_R2_001.fastq.gz	8838081	35.12778159
201910-TR1-V_S43	201910-TR1-V_S43_L002_R1_001.fastq.gz	8690352	35.70301122
201910-TR1-V_S43	201910-TR1-V_S43_L002_R2_001.fastq.gz	8690352	34.96100446
201910-TR13-B_S11	201910-TR13-B_S11_L001_R1_001.fastq.gz	5870741	35.79489114
201910-TR13-B_S11	201910-TR13-B_S11_L001_R2_001.fastq.gz	5870741	35.16190256
201910-TR13-B_S11	201910-TR13-B_S11_L002_R1_001.fastq.gz	5866193	35.72154121
201910-TR13-B_S11	201910-TR13-B_S11_L002_R2_001.fastq.gz	5866193	34.99795114
201910-TR13-V_S44	201910-TR13-V_S44_L001_R1_001.fastq.gz	7518719	35.73399618
201910-TR13-V_S44	201910-TR13-V_S44_L001_R2_001.fastq.gz	7518719	35.22909554
201910-TR13-V_S44	201910-TR13-V_S44_L002_R1_001.fastq.gz	7370463	35.65484828
201910-TR13-V_S44	201910-TR13-V_S44_L002_R2_001.fastq.gz	7370463	35.05570369
201910-TR21-B_S12	201910-TR21-B_S12_L001_R1_001.fastq.gz	4127224	35.69576888
201910-TR21-B_S12	201910-TR21-B_S12_L001_R2_001.fastq.gz	4127224	35.36444908
201910-TR21-B_S12	201910-TR21-B_S12_L002_R1_001.fastq.gz	4080468	35.61283681
201910-TR21-B_S12	201910-TR21-B_S12_L002_R2_001.fastq.gz	4080468	35.22940408
201910-TR21-V_S46	201910-TR21-V_S46_L001_R1_001.fastq.gz	7314521	35.73044592
201910-TR21-V_S46	201910-TR21-V_S46_L001_R2_001.fastq.gz	7314521	35.09921812
201910-TR21-V_S46	201910-TR21-V_S46_L002_R1_001.fastq.gz	7252426	35.65145387
201910-TR21-V_S46	201910-TR21-V_S46_L002_R2_001.fastq.gz	7252426	34.91153608
201910-TR25-B_S13	201910-TR25-B_S13_L001_R1_001.fastq.gz	5982010	35.68895204
201910-TR25-B_S13	201910-TR25-B_S13_L001_R2_001.fastq.gz	5982010	35.3016232
201910-TR25-B_S13	201910-TR25-B_S13_L002_R1_001.fastq.gz	5912486	35.60822284
201910-TR25-B_S13	201910-TR25-B_S13_L002_R2_001.fastq.gz	5912486	35.1461639
201910-TR25-V_S47	201910-TR25-V_S47_L001_R1_001.fastq.gz	12920891	35.75984179
201910-TR25-V_S47	201910-TR25-V_S47_L001_R2_001.fastq.gz	12920891	34.9395044
201910-TR25-V_S47	201910-TR25-V_S47_L002_R1_001.fastq.gz	12727952	35.69277532
201910-TR25-V_S47	201910-TR25-V_S47_L002_R2_001.fastq.gz	12727952	34.74177016

201910-TR27-B_S14	201910-TR27-B_S14_L001_R1_001.fastq.gz	4355847	35.70639327
201910-TR27-B_S14	201910-TR27-B_S14_L001_R2_001.fastq.gz	4355847	35.14488686
201910-TR27-B_S14	201910-TR27-B_S14_L002_R1_001.fastq.gz	4284330	35.62809424
201910-TR27-B_S14	201910-TR27-B_S14_L002_R2_001.fastq.gz	4284330	34.96456272
201910-TR27-V_S45	201910-TR27-V_S45_L001_R1_001.fastq.gz	15734633	35.56078893
201910-TR27-V_S45	201910-TR27-V_S45_L001_R2_001.fastq.gz	15734633	35.29707531
201910-TR27-V_S45	201910-TR27-V_S45_L002_R1_001.fastq.gz	15430859	35.4714444
201910-TR27-V_S45	201910-TR27-V_S45_L002_R2_001.fastq.gz	15430859	35.12686656
201910-TR31-B_S15	201910-TR31-B_S15_L001_R1_001.fastq.gz	1292697	35.71719591
201910-TR31-B_S15	201910-TR31-B_S15_L002_R1_001.fastq.gz	1254415	35.63830949
201910-TR31-V_S52	201910-TR31-V_S52_L001_R1_001.fastq.gz	18064768	35.78719865
201910-TR31-V_S52	201910-TR31-V_S52_L001_R2_001.fastq.gz	18064768	35.40771617
201910-TR31-V_S52	201910-TR31-V_S52_L002_R1_001.fastq.gz	17905014	35.71168785
201910-TR31-V_S52	201910-TR31-V_S52_L002_R2_001.fastq.gz	17905014	35.28599548
201910-TR35-B_S16	201910-TR35-B_S16_L001_R1_001.fastq.gz	2386117	35.49873497
201910-TR35-B_S16	201910-TR35-B_S16_L001_R2_001.fastq.gz	2386117	35.48884527
201910-TR35-B_S16	201910-TR35-B_S16_L002_R1_001.fastq.gz	2338780	35.40244144
201910-TR35-B_S16	201910-TR35-B_S16_L002_R2_001.fastq.gz	2338780	35.33722368
201910-TR35-V_S54	201910-TR35-V_S54_L001_R1_001.fastq.gz	9670349	35.51802184
201910-TR35-V_S54	201910-TR35-V_S54_L001_R2_001.fastq.gz	9670349	35.33779629
201910-TR35-V_S54	201910-TR35-V_S54_L002_R1_001.fastq.gz	9452288	35.42711532
201910-TR35-V_S54	201910-TR35-V_S54_L002_R2_001.fastq.gz	9452288	35.1747214
201910-TR49-B_S17	201910-TR49-B_S17_L001_R1_001.fastq.gz	4704462	35.59944517
201910-TR49-B_S17	201910-TR49-B_S17_L001_R2_001.fastq.gz	4704462	35.39443107
201910-TR49-B_S17	201910-TR49-B_S17_L002_R1_001.fastq.gz	4631962	35.52157854
201910-TR49-B_S17	201910-TR49-B_S17_L002_R2_001.fastq.gz	4631962	35.25147659
201910-TR49-V_S48	201910-TR49-V_S48_L001_R1_001.fastq.gz	4412211	35.62486925
201910-TR49-V_S48	201910-TR49-V_S48_L001_R2_001.fastq.gz	4412211	35.14468007
201910-TR49-V_S48	201910-TR49-V_S48_L002_R1_001.fastq.gz	4306093	35.541378
201910-TR49-V_S48	201910-TR49-V_S48_L002_R2_001.fastq.gz	4306093	34.95725522
201910-TR53-B_S18	201910-TR53-B_S18_L001_R1_001.fastq.gz	289715	35.64902404
201910-TR53-B_S18	201910-TR53-B_S18_L001_R2_001.fastq.gz	289715	35.24079526
201910-TR53-B_S18	201910-TR53-B_S18_L002_R1_001.fastq.gz	271030	35.59774933
201910-TR53-B_S18	201910-TR53-B_S18_L002_R2_001.fastq.gz	271030	35.14861085
201910-TR53-V_S50	201910-TR53-V_S50_L001_R1_001.fastq.gz	15076	35.58914832
201910-TR53-V_S50	201910-TR53-V_S50_L001_R2_001.fastq.gz	15076	35.25331653
201910-TR53-V_S50	201910-TR53-V_S50_L002_R1_001.fastq.gz	11939	35.48412765
201910-TR53-V_S50	201910-TR53-V_S50_L002_R2_001.fastq.gz	11939	35.15110143
201703-TR1-B_S2	201703-TR1-B_S2_L002_R1_001.fastq.gz	921775	35.61949608
201703-TR1-B_S2	201703-TR1-B_S2_L002_R2_001.fastq.gz	921775	35.44960647

Table 29: Round 2 DNA Sequencing Kraken2 Output Count Summaries

sample_id	Count of unique tax_species Kraken hits	Contig counts matching in Kraken
201703-Seep-B_S27	262	12618
201703-Seep-V_S34	3241	137024
201703-TR1-B_S2	98	7382
201703-TR1-V_S30	1505	99067
201703-TR13-B_S29	161	6985
201703-TR13-V_S33	1753	168436
201703-TR21-B_S6	128	6937
201703-TR21-V_S31	1749	136848
201703-TR25-B_S3	344	13001
201703-TR25-V_S37	2382	164898
201703-TR27-B_S5	1049	42175
201703-TR27-V_S32	2120	147343
201703-TR31-B_S4	87	4990
201703-TR31-V_S38	2842	179698
201703-TR35-B_S24	19	864
201703-TR35-V_S36	1417	119934
201703-TR49-B_S28	809	36874
201703-TR49-V_S35	2331	243882
201708-Seep-B_S9	1	26
201708-Seep-V_S51	1535	55042
201708-TR1-B_S19	1910	93773
201708-TR1-V_S41	572	36719
201708-TR13-B_S20	137	7082
201708-TR13-V_S39	311	19521
201708-TR21-B_S8	387	19617
201708-TR21-V_S40	785	61222
201708-TR25-B_S21	1725	88556
201708-TR25-V_S56	1290	71655
201708-TR27-B_S22	605	25615
201708-TR27-V_S55	932	56669
201708-TR31-B_S23	1113	62853
201708-TR31-V_S57	1539	86861
201708-TR35-B_S7	62	8179
201708-TR35-V_S53	1591	108121
201708-TR49-B_S26	159	8827
201708-TR49-V_S42	926	63348
201708-TR53-B_S25	1174	61962
201708-TR53-V_S49	1112	65200
201910-TR1-B_S10	1980	80837
201910-TR1-V_S43	1753	76950
201910-TR13-B_S11	1502	63492
201910-TR13-V_S44	1698	73792
201910-TR21-B_S12	961	44043
201910-TR21-V_S46	1208	63339
201910-TR25-B_S13	1521	65055
201910-TR25-V_S47	2141	117488
201910-TR27-B_S14	1266	44116
201910-TR27-V_S45	2084	131324
201910-TR31-B_S15	271	10305
201910-TR31-V_S52	2703	206538
201910-TR35-B_S16	804	25480
201910-TR35-V_S54	1984	112396
201910-TR49-B_S17	1299	59343
201910-TR49-V_S48	1298	59999
201910-TR53-B_S18	21	668
201910-TR53-V_S50	1	25

Table 30: Round 2 DNA Sequencing Kraken2 Outputs: Domains, Phylum, Class

tax_domain	count(*)	tax_phylum	count(*)	tax_class	count(*)
Archaea	1419	Acidobacteria	197	Acidimicrobiia	194
Bacteria	61539	Actinobacteria	11417	Acidithiobacillia	75
Eukaryota	54	Aquificae	97	Acidobacteriia	162
Viruses	1646	Armatimonadetes	46	Actinobacteria	11065
		Bacteroidetes	5517	Alphaproteobacteria	6992
		Balneolaeota	16	Anaerolineae	34
		Caldiserica	23	Aquificae	98
		Calditrichaeota	8	Archaeoglobi	59
		candidate division Zixibacteria	2	Ardenticatenia	30
		Candidatus Bipolaricaulota	4	Bacilli	5696
		Candidatus Cloacimonetes	25	Bacteroidia	657
		Candidatus Gracilibacteria	3	Betaproteobacteria	6506
		Candidatus Korarchaeota	109	Blastocatellia	19
		Candidatus Microgenomates	5	Caldilineae	16
		Candidatus Saccharibacteria	65	Caldisericia	25
		Chlamydiae	95	Calditrichae	8
		Chlorobi	97	Candidatus Babeliae	26
		Chloroflexi	178	Candidatus Brocadiae	15
		Chordata	59	Chitinophagia	338
		Chrysiogenetes	1	Chlamydiia	101
		Coprothermobacterota	2	Chlorobia	111
		Crenarchaeota	248	Chloroflexia	47
		Cyanobacteria	1530	Chrysiogenetes	1
		Deferribacteres	51	Chthonomonadetes	4
		Deinococcus-Thermus	319	Clostridia	2360
		Dictyoglomi	21	Coprothermobacteria	2
		Elusimicrobia	6	Coriobacteriia	240
		Euryarchaeota	1205	Cytophagia	828
		Fibrobacteres	5	Deferribacteres	50
		Firmicutes	8481	Dehalococcoidia	16
		Fusobacteria	287	Deinococci	329
		Gemmatimonadetes	58	Deltaproteobacteria	1123
		Ignavibacteriae	41	Dictyoglomia	22
		Kiritimatiellaeota	19	Ellioviricetes	32
		Lentisphaerae	11	Elusimicrobia	3
		Negarnaviricota	45	Endomicrobia	3
		Nitrospirae	69	Epsilonproteobacteria	969
		Planctomycetes	1165	Erysipelotrichia	310
		Proteobacteria	29901	Fibrobacteria	51
		Spirochaetes	431	Fimbriimonadia	76
		Synergistetes	53	Flavobacteriia	3132

		Tenericutes	790	Fusobacteriia	292
		Thaumarchaeota	1453	Gammaproteobacteria	14065
		Thermodesulfobacteria	55	Gemmatimonadetes	215
		Thermotogae	137	Gloeobacteria	38
		Verrucomicrobia	311	Halobacteria	512
				Hydrogenophilalia	10
				Ignavibacteria	97
				Kiritimatiellae	11
				Ktedonobacteria	11
				Lentisphaeria	11
				Limnochordia	361
				Mammalia	59
				Methanobacteria	137
				Methanococci	61
				Methanomicrobia	206
				Methanopyri	2
				Methylacidiphilae	95
				Mollicutes	792
				Monjiviricetes	13
				Negativicutes	205
				Nitriliruptoria	213
				Nitrososphaeria	1159
				Nitrospira	72
				Oligoflexia	161
				Opitutae	160
				Phycisphaerae	56
				Planctomycetia	582
				Rubrobacteria	87
				Saprospiria	45
				Spartobacteria	80
				Sphingobacteriia	538
				Spirochaetia	435
				Synergistia	58
				Tepidiformia	19
				Thermococci	127
				Thermodesulfobacteria	55
				Thermoleophilia	231
				Thermomicrobia	19
				Thermoplasmata	117
				Thermoprotei	635
				Thermotogae	153
				Tissierellia	428
				Verrucomicrobiae	160
				Vicinamibacteria	28
				Zetaproteobacteria	22

Table 31: Round 2 DNA Sequencing Kraken2 Outputs: Order, Family

tax_order	count(*)	tax_family	count(*)
Acholeplasmatales	118	Acaryochloridaceae	16
Acidaminococcales	9	Acetobacteraceae	389
Acidiferrobacterales	57	Acholeplasmataceae	118
Acidilobales	18	Acidaminococcaceae	9
Acidimicrobiales	64	Acidiferrobacteraceae	57
Acidithiobacillales	75	Acidilobaceae	6
Acidobacteriales	163	Acidimicrobiaceae	8
Acidothermales	22	Acidithiobacillaceae	75
Actinomycetales	363	Acidobacteriaceae	151
Actinopolysporales	20	Acidothermaceae	22
Aeromonadales	409	Actinomycetaceae	363
Alteromonadales	1591	Actinopolysporaceae	20
Anaerolineales	36	Adenoviridae	11
Aquificales	83	Aerococcaceae	48
Archaeoglobales	51	Aeromonadaceae	349
Ardenticatenales	30	Akkermansiaceae	41
Bacillales	3870	Alcaligenaceae	901
Bacteriovoracales	51	Alcanivoracaceae	78
Bacteroidales	537	Alicyclobacillaceae	56
Bacteroidetes Order II. Incertae sedis	68	Alloherpesviridae	1
Bdellovibrionales	67	Alteromonadaceae	496
Bifidobacteriales	196	Amoebophilaceae	24
Brachyspirales	70	Anaerohalosphaeraceae	13
Bradymonadales	19	Anaerolineaceae	36
Bryobacterales	45	Anaeromyxobacteraceae	49
Bunyavirales	18	Anaplasmataceae	152
Burkholderiales	4736	Anelloviridae	1
Caldilineales	18	Aphanizomenonaceae	86
Caldisericales	26	Aphanothecaceae	88
Calditrichales	8	Aquificaceae	29
Campylobacterales	922	Archaeoglobaceae	51
Candidatus Babeliales	26	Archangiaceae	79
Candidatus Brocadiales	15	Ardenticatenaceae	30
Candidatus Nanopelagicales	379	Atopobiaceae	45
Candidatus Nitrosocaldales	6	Aurantimonadaceae	43
Cardiobacteriales	23	Azonexaceae	99
Catenulisporales	82	Bacillaceae	1980
Caudovirales	1548	Bacteriovoracaceae	14
Caulobacterales	355	Bacteroidaceae	169
Cellvibrionales	369	Baculoviridae	22

Chitinophagales	303	Baekduiaceae	15
Chlamydiales	47	Barnesiellaceae	19
Chlorobiales	111	Bartonellaceae	172
Chloroflexales	50	Bdellovibrionaceae	79
Chromatiales	596	Beijerinckiaceae	78
Chroococcales	229	Bernardetiaceae	17
Chroococcidiopsidales	14	Beutenbergiaceae	29
Chrysiogenales	1	Bifidobacteriaceae	196
Chthonomonadales	4	Blattabacteriaceae	83
Clostridiales	1962	Bogoriellaceae	56
Coprothermobacterales	2	Bornaviridae	1
Coriobacteriales	76	Borreliaceae	108
Corynebacteriales	2699	Brachyspiraceae	70
Cytophagales	823	Bradymonadaceae	13
Deferribacterales	50	Bradyrhizobiaceae	562
Dehalococcoidales	4	Brevibacteriaceae	69
Deinococcales	204	Brucellaceae	81
Desulfarculales	15	Budviciaceae	36
Desulfobacterales	156	Burkholderiaceae	1797
Desulfovibrionales	214	Caedimonadaceae	11
Desulfurellales	24	Caldilineaceae	18
Desulfurobacteriales	15	Caldisericaceae	25
Desulfurococcales	17	Caldisphaeraceae	12
Desulfuromonadales	249	Calditrichaceae	8
Dictyoglomales	22	Calotrichaceae	112
Eggerthellales	80	Campylobacteraceae	707
Egibacterales	12	Candidatus Babeliaceae	26
Egicoccales	15	Candidatus Brocadiaceae	15
Elusimicrobiales	3	Candidatus Deianiraeaceae	17
Emcibacterales	9	Candidatus Methanomethylophilaceae	5
Endomicrobiales	3	Candidatus Midichloriaceae	5
Enterobacterales	2360	Candidatus Nanopelagicaceae	379
Entomoplasmatales	263	Candidatus Nitrosocaldaceae	6
Erysipelotrichales	104	Candidatus Paracaedibacteraceae	8
Euzebyales	23	Cardiobacteriaceae	23
Fibrobacterales	51	Carnobacteriaceae	122
Fimbriimonadales	76	Casimicrobiaceae	21
Flavobacteriales	3132	Catenulisporaceae	83
Frankiales	145	Caulimoviridae	1
Fusobacteriales	292	Caulobacteraceae	355
Gemmatales	154	Cellulomonadaceae	148
Gemmatimonadales	215	Cellvibrionaceae	153
Geodermatophilales	63	Chamaesiphonaceae	12
Gloeobacterales	25	Chelatococcaceae	26
Gloeoemargaritales	4	Chitinophagaceae	303

Glycomycetales	52	Chlamydiaceae	50
Halanaerobiales	134	Chlorobiaceae	111
Halobacteriales	170	Chloroflexaceae	27
Haloferacales	170	Christensenellaceae	10
Herpesvirales	37	Chromatiaceae	205
Holosporales	46	Chromobacteriaceae	266
Hydrogenophilales	10	Chroococcaceae	64
Ignavibacteriales	97	Chroococcidiopsidaceae	14
Immundisolibacterales	11	Chrysiogenaceae	1
Isosphaerales	105	Chthonomonadaceae	4
Jiangellales	79	Closteroviridae	1
Kineosporiales	33	Clostridiaceae	848
Kiritimatiellales	11	Clostridiales Family XIII. Incertae Sedis	34
Kosmotogales	19	Clostridiales Family XVI. Incertae Sedis	2
Ktedonobacterales	11	Clostridiales Family XVII. Incertae Sedis	42
Lactobacillales	1820	Cohaesibacteraceae	31
Legionellales	412	Coleofasciculaceae	8
Leptospirales	95	Colwelliaceae	154
Limnochordales	20	Comamonadaceae	1237
Magnetococcales	11	Conexibacteraceae	33
Marinilabiliales	109	Coprothermobacteraceae	2
Mariprofundales	22	Coriobacteriaceae	31
Mesoaciditogales	15	Coronaviridae	1
Methanobacteriales	131	Corynebacteriaceae	830
Methanocellales	29	Coxiellaceae	145
Methanococcales	61	Crocinitomicaceae	25
Methanomassiliicoccales	20	Cryomorphaceae	10
Methanomicrobiales	41	Cuniculiplasmataceae	3
Methanopyrales	2	Cyanobacteriaceae	20
Methanosarcinales	112	Cyanothecaceae	18
Methylacidiphilales	169	Cyclobacteriaceae	159
Methylococcales	219	Cytophagaceae	273
Micrococcales	2616	Deferribacteraceae	52
Micromonosporales	459	Dehalococcoidaceae	4
Mononegavirales	13	Deinococcaceae	198
Mycoplasmatales	411	Dermabacteraceae	130
Myxococcales	396	Dermacoccaceae	37
Nakamurellales	57	Dermatophilaceae	46
Natranaerobiales	8	Dermocarpellaceae	55
Natrialbales	172	Desulfarculaceae	15
Nautiliales	47	Desulfobacteraceae	134
Neisseriales	662	Desulfobulbaceae	26
Nevskiales	56	Desulfohalobiaceae	14
Nidovirales	11	Desulfomicrobiaceae	11
Nitrosomonadales	711	Desulfovibrionaceae	189

Nitrosopumilales	39	Desulfurellaceae	26
Nitrososphaerales	48	Desulfurobacteriaceae	15
Nitrospirales	72	Desulfurococcaceae	8
Nostocales	538	Desulfuromonadaceae	83
Oceanospirillales	856	Dictyoglomaceae	22
Opitutales	140	Dietziaceae	62
Orbales	28	Dysgonamonadaceae	13
Ortervirales	10	Ectothiorhodospiraceae	235
Oscillatoriales	143	Eggerthellaceae	80
Parachlamydiales	54	Egibacteraceae	12
Parvularculales	12	Egicoccaceae	15
Pasteurellales	447	Elusimicrobiaceae	3
Pelagibacterales	222	Emcibacteraceae	9
Petrotogales	25	Endomicrobiaceae	3
Phycisphaerales	19	Endozoicomonadaceae	27
Picornavirales	8	Enterobacteriaceae	955
Pirellulales	155	Enterococcaceae	233
Planctomycetales	172	Entomoplasmataceae	48
Pleurocapsales	71	Erwiniaceae	367
Primates	55	Erysipelotrichaceae	104
Propionibacteriales	773	Erythrobacteraceae	260
Pseudomonadales	4538	Eubacteriaceae	60
Pseudonocardiales	544	Euzebyaceae	23
Puniceicoccales	21	Ferrimonadaceae	10
Rhizobiales	2746	Ferroplasmaceae	2
Rhodobacterales	1373	Fervidobacteriaceae	56
Rhodocyclales	397	Fibrobacteraceae	51
Rhodospirillales	828	Fimbriimonadaceae	76
Rickettsiales	325	Flammeovirgaceae	113
Rubrobacterales	60	Flavobacteriaceae	3005
Salinisphaerales	13	Francisellaceae	147
Saprospirales	31	Frankiaceae	144
Sedimentisphaerales	34	Fusobacteriaceae	153
Selenomonadales	107	Gallionellaceae	112
Silvanigrellales	43	Gemmataceae	155
Solirubrobacterales	33	Gemmatimonadaceae	217
Sphaerobacterales	19	Geobacteraceae	167
Sphingobacteriales	535	Geodermatophilaceae	63
Sphingomonadales	1065	Gloeobacteraceae	25
Spirochaetales	270	Gloeomargaritaceae	4
Sporichthyales	18	Glycomycetaceae	52
Streptomycetales	2210	Gomontiellaceae	11
Streptosporangiales	244	Gordoniaceae	159
Sulfolobales	147	Gottschalkiaceae	14
Synechococcales	506	Granulosicoccaceae	47

Synergistales	58	Hafniaceae	70
Syntrophobacterales	50	Hahellaceae	29
Tepidiformales	19	Halanaerobiaceae	88
Thermales	126	Halieaceae	58
Thermoanaerobacterales	247	Haliscomenobacteraceae	22
Thermococcales	120	Haloarculaceae	83
Thermodesulfobacteriales	55	Halobacteriaceae	87
Thermomicrobiales	4	Halobacteriovoraceae	38
Thermoplasmatales	30	Halobacteroidaceae	43
Thermoproteales	35	Haloferacaceae	85
Thermotogales	94	Halomonadaceae	399
Thiotrichales	369	Halorubraceae	85
Tissierellales	141	Halothiobacillaceae	28
Veillonellales	59	Hantaviridae	6
Verrucomicrobiales	164	Hapalosiphonaceae	40
Vibrionales	864	Helicobacteraceae	215
Victivallales	11	Heliobacteriaceae	30
Xanthomonadales	847	Hepadnaviridae	1
		Herelleviridae	15
		Herpesviridae	5
		Holosporaceae	28
		Hominidae	55
		Hungateiclostridiaceae	87
		Hydrogenophilaceae	10
		Hydrogenothermaceae	51
		Hyellaceae	16
		Hymenobacteraceae	234
		Hyphomicrobiaceae	210
		Hyphomonadaceae	59
		Iamiaceae	26
		Ichthyobacteriaceae	9
		Idiomarinaceae	62
		Iflaviridae	1
		Ignavibacteriaceae	81
		Ilumatobacteraceae	30
		Immundisolibacteraceae	11
		Inoviridae	2
		Intrasporangiaceae	219
		Iridoviridae	6
		Isosphaeraceae	106
		Jiangellaceae	79
		Jonesiaceae	17
		Kangiellaceae	61
		Kineosporiaceae	33
		Kiritimatiellaceae	17

		Kofleriaceae	20
		Kosmotogaceae	19
		Ktedonosporobacteraceae	11
		Labilitrichaceae	30
		Lachnospiraceae	312
		Lacipirellulaceae	45
		Lactobacillaceae	765
		Lavidaviridae	1
		Legionellaceae	267
		Leptolyngbyaceae	48
		Leptospiraceae	95
		Leptotrichiaceae	140
		Leuconostocaceae	185
		Limnochordaceae	20
		Listeriaceae	169
		Litoricolaceae	17
		Luteoviridae	3
		Magnetococcaceae	12
		Malacoherpesviridae	1
		Marinifilaceae	23
		Marinilabiliaceae	10
		Mariprofundaceae	22
		Marseilleviridae	7
		Melioribacteraceae	16
		Merismopediaceae	8
		Mesoaciditogaceae	15
		Methanobacteriaceae	129
		Methanocaldococcaceae	22
		Methanocellaceae	29
		Methanococcaceae	39
		Methanomassiliicoccaceae	15
		Methanomicrobiaceae	17
		Methanopyraceae	2
		Methanoregulaceae	17
		Methanosarcinaceae	102
		Methanospirillaceae	7
		Methanothermaceae	3
		Methanotrichaceae	9
		Methylacidiphilaceae	169
		Methylobacteriaceae	326
		Methylococcaceae	219
		Methylocystaceae	121
		Methylophilaceae	284
		Microbacteriaceae	1179
		Microbulbiferaceae	102

		Micrococcaceae	572
		Microcoleaceae	55
		Microcystaceae	57
		Micromonosporaceae	459
		Mimiviridae	35
		Mononiviridae	2
		Moraxellaceae	917
		Morganellaceae	388
		Moritellaceae	44
		Muribaculaceae	23
		Mycobacteriaceae	1121
		Mycoplasmataceae	411
		Myoviridae	830
		Myxococcaceae	91
		Nakamurellaceae	57
		Natranaerobiaceae	8
		Natrialbaceae	172
		Nautiliaceae	47
		Neisseriaceae	394
		Nitrosomonadaceae	140
		Nitrosopumilaceae	39
		Nitrososphaeraceae	48
		Nitrospiraceae	72
		Nocardiaceae	479
		Nocardioidaceae	472
		Nocardiopsaceae	73
		Nostocaceae	251
		Nudiviridae	21
		Oceanospirillaceae	174
		Odoribacteraceae	12
		Oleiphilaceae	18
		Opitutaceae	140
		Orbaceae	28
		Oscillatoriaceae	51
		Oscillospiraceae	21
		Oxalobacteraceae	781
		Paenibacillaceae	726
		Paludibacteraceae	13
		Parachlamydiaceae	29
		Parvoviridae	1
		Parvularculaceae	12
		Pasteurellaceae	447
		Pectobacteriaceae	176
		Pelagibacteraceae	222
		Peptococcaceae	197

		Peptoniphilaceae	86
		Peptostreptococcaceae	121
		Peribunyaviridae	2
		Persicobacteraceae	3
		Petrotogaceae	25
		Phycisphaeraceae	19
		Phycodnaviridae	102
		Phyllobacteriaceae	359
		Picrophilaceae	1
		Pirellulaceae	102
		Piscirickettsiaceae	164
		Pithoviridae	2
		Planctomycetaceae	168
		Planococcaceae	302
		Podoviridae	237
		Polyangiaceae	80
		Polydnaviridae	3
		Porphyromonadaceae	59
		Potyviridae	6
		Poxviridae	18
		Prevotellaceae	156
		Prochloraceae	53
		Prochlorotrichaceae	5
		Prolixibacteraceae	67
		Promicromonosporaceae	75
		Propionibacteriaceae	301
		Pseudanabaenaceae	16
		Pseudoalteromonadaceae	333
		Pseudomonadaceae	3621
		Pseudonocardiaceae	544
		Psychromonadaceae	28
		Puniceicoccaceae	21
		Pyrodictiaceae	9
		Reoviridae	5
		Rhabdoviridae	2
		Rhizobiaceae	583
		Rhodanobacteraceae	178
		Rhodobacteraceae	1314
		Rhodobiaceae	31
		Rhodocyclaceae	84
		Rhodospirillaceae	439
		Rhodothermaceae	68
		Rickettsiaceae	151
		Rikenellaceae	57
		Rivulariaceae	20

		Roseiflexaceae	23
		Ruaniaceae	21
		Rubrobacteraceae	45
		Ruminococcaceae	121
		Saccharospirillaceae	53
		Salinisphaeraceae	13
		Salinivirgaceae	9
		Sandaracinaceae	26
		Sanguibacteraceae	18
		Saprospiraceae	9
		Scytonemataceae	29
		Sedimentisphaeraceae	20
		Segniliparaceae	9
		Selenomonadaceae	62
		Shewanellaceae	464
		Silvanigrellaceae	29
		Simkaniaceae	15
		Sinobacteraceae	35
		Siphoviridae	286
		Solibacteraceae	17
		Sphaerobacteraceae	19
		Sphaerolipoviridae	1
		Sphingobacteriaceae	535
		Sphingomonadaceae	806
		Spirochaetaceae	162
		Spiroplasmataceae	215
		Spongiibacteraceae	56
		Sporichthyaceae	18
		Sporolactobacillaceae	38
		Sporomusaceae	45
		Staphylococcaceae	556
		Steroidobacteraceae	21
		Sterolibacteriaceae	118
		Streptococcaceae	467
		Streptomycetaceae	2210
		Streptosporangiaceae	97
		Succinivibrionaceae	60
		Sulfolobaceae	147
		Sutterellaceae	18
		Symbiobacteriaceae	18
		Synechococcaceae	348
		Synergistaceae	58
		Syntrophaceae	34
		Syntrophobacteraceae	16
		Syntrophomonadaceae	48

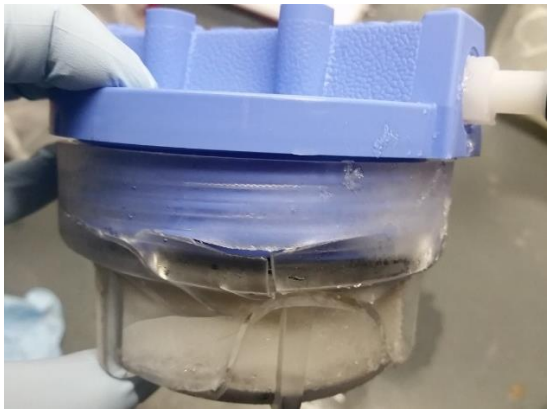
		Tannerellaceae	15
		Tectiviridae	1
		Tepidiformaceae	19
		Thermaceae	126
		Thermoactinomycetaceae	43
		Thermoanaerobacteraceae	73
		Thermoanaerobacterales Family III. Incertae Sedis	120
		Thermoanaerobacterales Family IV. Incertae Sedis	5
		Thermococcaceae	120
		Thermodesulfobacteriaceae	55
		Thermodesulfobiaceae	52
		Thermofilaceae	1
		Thermoguttaceae	8
		Thermomicrobiaceae	4
		Thermomonosporaceae	74
		Thermoplasmataceae	23
		Thermoproteaceae	34
		Thermotogaceae	38
		Thioalkalibacteraceae	11
		Thioalkalispiraceae	29
		Thiobacillaceae	51
		Thiotrichaceae	58
		Tissierellaceae	41
		Totiviridae	1
		Tristromaviridae	4
		Trueperaceae	6
		Tsukamurellaceae	39
		Vallitaleaceae	11
		Veillonellaceae	59
		Verrucomicrobiaceae	124
		Vibrionaceae	864
		Vicinamibacteraceae	41
		Vulgatibacteraceae	19
		Waddliaceae	11
		Wenzhouxiangellaceae	21
		Woeseiaceae	20
		Xanthobacteraceae	123
		Xanthomonadaceae	669
		Yersiniaceae	368
		Zoogloeaceae	204

Appendix 7: Field Photos

1: BMFS v1 during August 2017 sampling event

2: BMFS v2 during October 2019 sampling event

3: CARACAL boat August 2017 sampling, boat was not functional during October 2020 sampling



4: BMFS v1 broken filter housing due to freezing from March 2017 sampling event

5: CARACAL's stand-alone autoclave

6: CARACAL field lab during August 2017 sampling event

7: CARACAL field lab during October 2019 sampling event

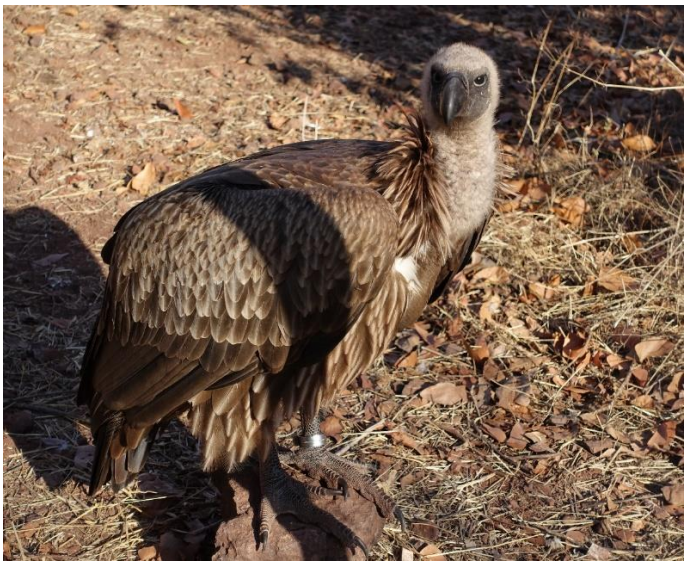
8: Semi-functional hood during August 2017 sampling event with BMFS v1 elution device

9: Non-functional hood during October 2019 sampling event

10: Centrifuge used for AllPrep sample extractions

11: Vortex used for AllPrep sample extractions

12: Pellet example following secondary concentration during October 2019 processing



14: Giles the rescued vulture at CARACAL

13: Badgy the rescued wild Honey Badger at CARACAL