TRACKING THE EVOLUTION OF SARS-COV-2 AND THE EMERGENCE OF OTHER INFECTIOUS DISEASES IN COMMUNITIES USING A WASTEWATER-BASED EPIDEMIOLOGY APPROACH

Report to the
Water Research Commission

by

Rian Pierneef
Agricultural Research Council, Biotechnology Platform, Onderstepoort

WRC Report No. 3065/1/23
ISBN 978-0-6392-0385-0

April 2023

WATER
RESEARCH
COMMISSION

Obtainable from
Water Research Commission
Bloukrans Building, 2nd Floor
Lynnwood Bridge Office Park
4 Daventry Road
Lynnwood Manor
PRETORIA

orders@wrc.org.za or download from www.wrc.org.za

This is the final report of WRC project no. C2020/21-00492.

EXECUTIVE SUMMARY

BACKGROUND

After the first recorded infection of SARS-CoV-2 in China the world changed forever. The COVID-19 pandemic highlighted the vulnerability of populations and the preparedness of the healthcare sector to timeously respond to pandemics. The COVID-19 global pandemic regrettably resulted in large-scale loss of life and economic devastation. By January 2022, South Africa had emerged from a fourth wave of infections and the vaccination programme was underway. As such, SARS-CoV-2 is certain to remain in circulation for the foreseeable time, and the detection of new variants of concern is certain to continue. It is therefore critical that SARS-CoV-2 surveillance is continued and research relating thereto supported in an attempt to curb the infection rate and garner as much information about this virus as possible.

Wastewater-based epidemiology (WBE) is an eloquent alternative in SARS-CoV-2 surveillance and allows for the early detection of SARS-CoV-2. This enables a rapid and consolidated response to curb infection rates and save lives. The use of metagenomic next generation sequencing in wastewater-based epidemiology is well documented. This method has recently demonstrated the ability to recover complete or near complete SARS-CoV-2 genomes from sewage. The recovery of SARS-CoV-2 genomes from wastewater enables evolutionary analysis and the identification of known and novel variants. The added value obtained by using metagenomic sequencing is the ability to detect other pathogens and their functional potential from the same sample in a single sequencing event. As such, investigations into the co-occurrence of other pathogens and the presence of antimicrobial resistance in samples containing SARS-CoV-2 is possible. The great efforts by scientists and researchers have clearly demonstrated the power and application of next-generation sequencing and whole genome sequencing in response to pandemics such as COVID-19.

In this project a next-generation sequencing approach was implemented to assign SARS-CoV-2 lineages in wastewater samples, detect co-occurring pathogens and identify antimicrobial resistance profiles. The next-generation sequencing protocol was divided into an untargeted and targeted approach. The untargeted or metagenomic approach was used to taxonomically categorize wastewater samples and detect the presence and mode of antimicrobial resistance elements. The targeted approach was implemented to amplify the SARS-CoV-2 genome in a wastewater sample and perform whole genome sequencing on the resulting amplicons. This information was then used to assign SARS-CoV-2 lineages per sample. Another targeted approached based on the 16S rRNA gene was further incorporated to provide taxonomic profiles for samples and ascertain the microbial diversity as found in wastewater samples.

AIMS

The following were the aims of the project:

1. Detecting the presence and tracking the evolution of SARS-CoV-2 in freshwater and wastewater samples
2. Identification of pathogens co-occurring with SARS-CoV-2
3. Analysing the Antimicrobial Resistance potential of organisms within freshwater and wastewater.

METHODOLOGY

For this work, different sets of wastewater samples were obtained from collaborators under the South African Collaborative COVID-19 Environmental Surveillance System (SACCESS) network across Gauteng and KwaZulu-Natal. The samples, graciously supplied by collaborators, were in various formats including extracted RNA, RNA extracted after viral concentration, extracted DNA and raw wastewater. Three different next-generation sequencing methods and their application thereof in wastewater-based epidemiology was demonstrated in this study. Targeted sequencing as performed by whole genome sequencing of SARS-CoV-2 in wastewater demonstrates the ability of whole genome sequencing to identify SARS-CoV-2 variants of concern in wastewater samples. Amplicon sequencing such as 16S rRNA was used with great success to provide a taxonomic overview of wastewater samples. Untargeted sequencing obtained by means of metagenomic analysis in wastewater surveillance demonstrates the abilities of metagenomic sequencing to generate taxonomic and antimicrobial resistance profiles. The analyses as performed on each of the samples are described below.

1. Determining the taxonomic composition and the presence of antimicrobial resistance genes metagenomic sequencing

Samples (n=20) from regions across Tshwane were used for RNA metagenomic sequencing. These samples were collected between 17 August 2020 and 6 April 2021 and all tested positive for the presence of SARS-CoV-2. Metagenomic sequencing was done on the RNA extractions from the samples graciously provided by the collaborator. The samples were analysed with regards to taxonomic composition and the presence of antimicrobial resistance. Detection of SARS-Covid-2 in the metagenomic data was further included.

Samples (n=30) were collected from 3 wastewater treatment plants in Tshwane, Gauteng. DNA extractions were done by the ARC Biotechnology including library preparation, amplicon and metagenomic sequencing. Amplicon sequencing produced taxonomic profiles for each sample whereas the metagenomic sequencing was able to detect the presence of antimicrobial resistance within the samples.

Wastewater samples (n=10) were collected from three municipal WWTPs in Pretoria, South Africa, that primarily treat household sewage. Grab samples (influent, activated sludge and secondary settling tank (SST) effluent) were collected from November 2021 to February 2022 at different treatment stages and metagenomic sequencing used to construct metagenome assembled genomes (MAGs). The ability to reconstruct partial to near complete genomes enables the taxonomic classification and detection of antimicrobial resistance. This information is critical as it allows researchers to understand which microorganisms have acquired resistance within a sample and in the community.

Wastewater samples (n=72) were collected from 8 WWTPs located in the East Rand of Gauteng (Mr. W. le Roux). These samples were collected weekly between 26 January 2022 and 22 March 2022 and represent 9 sampling dates. Amplicon and metagenomic sequencing was used to determine the taxonomic and antimicrobial profiles of the samples.

2. Determination of SARS-CoV-2 lineage and variants using whole genome sequencing

Samples (n=73) from across Durban, KwaZulu-Natal, were used for SARS-CoV-2 whole genome sequencing. These samples were collected between 21 July 2020 and 2 November 2021 and all tested positive for the presence of SARS-CoV-2. SARS-CoV-2 whole genome sequencing was done on the RNA extractions from the samples graciously provided by the collaborator. The samples were analysed with regards SARS-CoV-2 lineage and variants detected by means of whole genome sequencing. Currently accepted and published SARS-CoV-2 lineage assignment workflows were implemented and optimised for use in wastewater samples.

3. Detection and characterisation of viruses using viral RNA metagenomic sequencing

Samples (n=17) from across Durban, KwaZulu-Natal, were used for viral RNA metagenomic sequencing. These samples were collected between 25 August 2020 and 3 August 2021 and all tested positive for the presence of SARS-CoV-2. Metagenomic sequencing was done on RNA extracted after viral concentration using ultra (centricon) filtration graciously provided by the collaborator. The samples were analysed with regards to taxonomic composition. Detection of SARS-Covid-2 in the metagenomic data was further included.

RESULTS AND DISCUSSION

1. Taxonomic diversity of microorganisms and the presence of antimicrobial resistance genes in wastewater

In excess of 80 GB of data was produced for the 20 RNA metagenomic sequencing samples from Tshwane. The RNA metagenomic data was able to reveal the presence of SARS-CoV-2 in some of the samples and it was found that there was a correlation between the viral load, measured by means of 7-Day average COVID-19 cases, and the ability to detect SARS-CoV-2 in a sample. The samples displayed a high level of taxonomic diversity and the methodology was able to classify the Archaeal, Bacterial and Viral portions of the wastewater samples. These classifications were further investigated along various taxonomic ranks. The data was further inspected for antimicrobial resistance elements and a high level of diversity and variable between samples was present. Antimicrobial resistance classification was further explored along various resistance classification levels. The 30 samples from Gauteng used for amplicon and metagenomic sequencing produced more than 160 GB of data. The samples had high taxonomic and antimicrobial resistance diversity. This included high levels of Proteobacteria and Tetracycline.

A further 10 samples from Gauteng produced 100 GB of metagenomic data and was used to construct metagenome assembled genomes (MAGs). The ability to extract partial and near complete genomes from wastewater is critical in understanding the acquisition of antimicrobial resistance by certain linages. The data allowed for the reconstruction of 34 medium to high quality MAGs. In this section emphasis was given to *Legionella pneumophila*, *Mycobacterium* spp. and *Aeromonas* spp. and the AMRs and virulence factors encoded within them.

Samples from the East Rand, Gauteng, were used for amplicon and metagenomic sequencing. This part of the project produced more than 600 GB of data. The metagenomic data was used to construct antimicrobial resistance profiles across treatment plants and sampling dates. Varying levels of resistance were found between sampling locations with no significant difference detected between the treatment plants. Clear differences were detected between the sampling dates. An initial increase in the number of AMR genes was followed by a large decrease and then a continuous increase along the sampling dates. Further investigation is required to determine the reason for this and if this would be a reoccurring trend.

2. Determination of SARS-CoV-2 lineage and variants in wastewater

In excess of 9 GB of data was produced for the 73 KwaZulu-Natal samples used for targeted SARS-CoV-2 whole genome sequencing. The NEBNext ARTIC SARS-CoV-2 sequencing protocol was optimised for use in wastewater samples and produced adequate sequencing results. Samples collected between mid-July 2020 and the start of November 2021 displayed varying success with regards to the amount of data generated. It was determined that the length of time between RNA extraction and sequencing is of critical importance, even when stored under optimal conditions. SARS-CoV-2 lineage assignment was possible for more than half of the samples. The SARS-CoV-2 lineage assignments in wastewater samples was in agreement with the prevalent Variant of Concern per sampling period. It was further possible to assemble 3 near complete SARS-CoV-2 genomes from the sequencing results. This report clearly illustrates the application and possibility of SARS-CoV-2 whole genome sequencing in wastewater samples and the contribution thereof to wastewater-based epidemiology.

3. Detection and characterisation of viruses in wastewater

In excess of 80 GB of data was produced for the 17 KwaZulu-Natal samples. These sample were RNA extracted after viral concentration using ultra (centricon) filtration and constitutes the assemblage of viruses or virome. The RNA metagenomic data was able to reveal the presence of SARS-CoV-2 in all but one of the samples. The samples displayed a high level of taxonomic diversity and the methodology was able to classify the virome as found in these wastewater samples. These classifications were further investigated along various taxonomic ranks.

SUMMARY OF FINDINS AND CONCLUSIONS

This Final Technical and Data Report details the work done and results obtained for the amplicon, metagenomic and SARS-CoV-2 whole genome sequencing of wastewater samples under the project titled “Tracking the evolution of SARS-CoV-2 and the emergence of other infectious diseases in communities using a wastewater-based epidemiology approach”. This project aimed to harness the added value afforded by next-generation sequencing in answering various questions related to the presence of SARS-CoV-2, antimicrobial resistance and the microbial content of wastewater samples. The collaborators were all able to accomplish their individual mandates before the samples were passed on to this project. Obtaining samples in this method insured that there was no duplication of results and that the absolute maximum amount of information was extracted per sample in a strategic workflow.

This report highlights the functionality of next-generation sequencing and in particular targeted and untargeted sequencing in wastewater surveillance. The untargeted sequencing or metagenomic methodology was able to provide a holistic view on the taxonomic diversity found in wastewater samples. Furthermore, this methodology allows for the detection of antimicrobial resistance and associated classifications without the need of another data generation event. Although not the most feasible methodology to test for the presence of SARS-CoV-2 in wastewater samples it is still capable of recovering portions of the genome in samples with a high viral load. Data sets such as these contained within this report will greatly assist wastewater surveillance, disease modelling and the prediction of outbreak events.

Targeted sequencing as was used for SARS-CoV-2 whole genome sequencing in these wastewater samples was able to provide SARS-CoV-2 lineage assignments. SARS-CoV-2 whole genome sequencing is generally performed on clinical samples. The application thereof on wastewater samples and the ability to produce lineage assignments and near complete genomes clearly illustrates the functionality of this protocol. This method provides a clear picture on high prevalence SARS-CoV-2 variants as found in a community and has the possibility to detect an upsurge or prevalence of variants of concern.

Continuous monitoring of wastewater samples for the presence of AMR genes is critical in understanding the ebb and flow of these resistance elements in communities. The ability to construct metagenome assembled genomes with metagenomic sequencing data further allows us to classify the recipients of acquired resistance and better understand the spread of AMR in our population.

Metagenomic sequencing and analysis is a powerful tool in wastewater surveillance and epidemiology. The method allows for the taxonomic classification of the organisms present in a sample and furthermore the functional potential of the organisms in a sample. The amount of data generated in a single sequencing event can be used in various research questions and provides a holistic representation of the biological components in a system. The results obtained from metagenomic sequencing analysis will greatly assist in various public health concerns and the associated strategies to be followed in addressing the concerns. Whole genome sequencing and analysis is another powerful tool in wastewater surveillance and epidemiology. The method allows for SARS-CoV-2 lineage assignment and the construction of near complete SARS-CoV-2 genomes. Next-generation sequencing is clearly the future of wastewater-based epidemiological surveillance.

ACKNOWLEDGEMENTS

The project team wishes to thank the following people for their contributions to the project:

Collaborator	Affiliation
Mr Dariah de Villiers	Lumegen
Prof Christo Muller	SAMRC Biomedical Research and Innovation Platform
Dr Awelani Mutshembele	SAMRC Tuberculosis Platform
Prof Martie Van Der Walt	SAMRC Tuberculosis Platform
Dr Shaun Groenink	GreenHill Laboratories
Mr Neil Madgwick	Praecautio
Ms Lisa Schaefer	CSIR Water and Health Microbiology Research Facility
Mr Wouter le Roux	CSIR Water and Health Microbiology Research Facility
Dr Gina Pocock	Waterlab
Dr Annancietar Gomba	NIOH Waterborne Pathogen Laboratory
Dr Said Rachida	NICD Centre for Vaccines and Immunology
Prof Feroz Swalaha	DUT Biotechnology and Food Science
Dr Melinda Suchard	SACCESS network
Prof Faizal Bux	DUT Institute for Water and Wastewater Technology
Prof Sheena Pillai	DUT Institute for Water and Wastewater Technology
Prof Thulani Makhalanyane	University of Pretoria
Dr Oliver Bezuidt	University of Pretoria
Mr Nico van Blerk	ERWAT

Note on samples:

24 Samples for metagenomic sequencing were received from Dr A. Mutshembele of which 20 samples are included in this report. The additional 4 samples will be included in a prospective student project and publication. 73 Samples for whole genome sequencing were received from Prof F. Bux and are included in this report. A subset of 17 samples were used for metagenomic sequencing and the results thereof contained in this report. 39 Samples were received from Dr N. Gomba. These samples were used for whole genome, amplicon and metagenomic sequencing. 72 Samples were received from Mr W. le Roux and were used for amplicon and metagenomic sequencing.

This page was intentionally left blank

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS & ABBREVIATIONS

AMR	Antimicrobial Resistance
ARC	Agricultural Research Council
ARO	Antibiotic Resistance Ontology
BTP	Biotechnology Platform
GB	Gigabyte
mNGS	Metagenomic Next Generation Sequencing
MST	Microbial Source Tracking
NGS	Next-generation sequencing
qRT-PCR	Quantitative Reverse Transcriptase-Polymerase Chain Reaction
SACCESS	South African Collaboration COVID 19 Environmental Surveillance System
WGS	Whole genome sequencing
WWTP	Wastewater Treatment Plant

This page was intentionally left blank

CHAPTER 1: BACKGROUND

1.1 INTRODUCTION

Since the declaration of the pandemic, South Africa has encountered and surpassed a fourth wave of COVID-19 infections. It is clear that the COVID-19 pandemic and the presence of SARS-CoV-2 in our environments will be with us for the foreseeable future. This pandemic and the associated virus require novel, yet reliable technologies and protocols to track the presence thereof and provide timeous reporting of possible outbreaks and resurgence in communities. Wastewater-based epidemiology (WBE) is an eloquent method of quantitatively determining the prevalence of infection in localised areas. This method has been implemented as early as 2011 in the Netherlands to track influenza. During the current COVID-19 pandemic this method has been implemented in numerous countries with great success. Wastewater-based epidemiology will allow for the rapid detection of SARS-CoV-2 in communities and will assist in curbing the spread of COVID-19. This methodology allows for the speedy response to curb the spread of COVID-19 and flattening of the curve in community outbreaks.

As the COVID-19 pandemic progresses the virus is certain to evolve. This has been proven by the emergence of novel and more virulent variants, exemplified by the unfortunate and erroneously named South African SARS-CoV-2 variant. To date, numerous variants across the world have been detected and reported on. The ability to track the evolution of SARS-CoV-2 in wastewater and the variants currently circulating will greatly assist researchers and policy makers with regards to the evolutionary trajectory the virus is on and may assist in fighting the pandemic. The rigorous and frequent analysis of samples will enable a near “real-time” reporting of genomic composition an evolution of SARS-CoV-2 in South African communities. It is further critical to understand the associated pathogens that occur with SARS-CoV-2 in wastewater, including the virulence and antimicrobial properties they may possess. By investigating the co-occurrence of microorganism with SARS-CoV-2 it may be possible to identify indicator or closely associated microorganisms. These may serve as a proxy for the presence of SARS-CoV-2 in wastewater samples and used as a baseline for future studies. This can be tested by assessing the microbial composition of wastewater samples that tested positive and negative for the presence of SARS-CoV-2.

Current technologies further allow for the isolation and extraction of SARS-CoV-2 from samples and the subsequent whole genome sequencing (WGS) and analysis thereof. Due to large collaborative research projects and the communal good will and cause surrounding the COVID-19 pandemic, the genome sequences of more than 600,000 SARS-CoV-2 WGS submissions are publicly available. The serves as an unprecedented database for researchers to identify and track the evolution of SARS-CoV-2. The South African Collaboration COVID 19 Environmental Surveillance System (SACCESS) network is a collection of researchers with an interest in applying WBE with regards to COVID-19 surveillance and includes participants from across South Africa. The SACCESS partners have rigorously collected wastewater samples and conducted COVID-19 diagnostics with great success. These partners have individual sampling schedules and sites which include provincial hotspots across the country and have concluded all documentation required to obtain the samples. This includes retrospective, current and future samples. Based on the collaboration with SACCESS partners samples are easily obtained and redundancy excluded. The SACCES collaborators collect(ed) samples weekly and analysed for the presence of SARS-CoV-2 RNA using the sampling and testing protocols developed in phase one (proof of concept phase) of the WRC’s national programme for monitoring COVID-19 infections in communities using a wastewater-based epidemiology approach.

Urban areas contain comprehensive sewer networks which is fed by various components of the urban population. The collection and analysis of wastewater samples are therefore representative of these urban populations. In rural areas the water resources are based on a freshwater supply. The analysis of samples

both upstream and downstream of these rural communities will give a detailed overview of the presence of SARS-CoV-2 in these communities. Metagenomic analysis of these samples will be done using next generation sequencing (NGS). The NGS strategy will be based on metagenomics. The testing protocols developed in phase one and metagenomic approaches will complement each other with the metagenomic approach providing additional information regarding other viruses and bacteria present in samples. The metagenomic approach further allows for the detection of concurrence of pathogens and SARS-CoV-2 and the identification of antimicrobial resistance (AMR) elements in the samples. This information will be critical in assessing the risk of COVID-19 due to possible co-infection based on the prevalence of other pathogens and AMR in an environment.

The large amount of research that has been concluded with regards to the whole genome sequencing has brought forth the detection of novel SARS-CoV-2 variants with increased virulence and infection rates. The WBE approach is an eloquent solution which will enable the early detection of possible variants and provide retrospective information on the initial occurrence of such variants. The workflow allows for the initial detection of SARS-CoV-2 from wastewater samples, the metagenomic analysis of genomic segments to ascertain the presence of variants and the thereafter WGS of isolated SARS-CoV-2 to classify and inspect the evolutionary track of SARS-CoV-2. This project will compliment other national COVID-19 surveillance projects, in a nonredundant effort, by increasing the number of wastewater samples analysed for the presence of SARS-CoV-2 and reporting on the presence of known and novel variants in retrospective, current and future freshwater and wastewater samples. It will further allow for the detection of co-occurring pathogens in relation to the presence of SARS-CoV-2 and the identification of AMR potential of other organisms in freshwater and wastewater samples. This includes the possible detection of proxies associated with SARS-CoV-2 which may be used in future surveillance strategies.

1.2 RATIONALE FOR THIS PROJECT

This project will aid South Africa's fight against the COVID-19 pandemic. The recent second wave has clearly indicated that we will need to continuously and effectively perform SARS-CoV-2 surveillance in an attempt to timeously warn stakeholders and governing bodies on a possible surge in COVID-19 cases. The ability to rapidly and reliably identify areas of high infection will enable authorities to address and contain localised outbreaks and as such prevent resurgence of the disease. It is critical that pre-emptive community information is gathered after which individual testing would follow. This project will ensure that communities with high levels of SARS-CoV-2 in wastewater or freshwater are identified and the necessary steps are taken and will highlight the South African fight against COVID-19 internationally and may serve as a basis for other studies in other countries, especially in Africa. Current data clearly suggests that the COVID-19 pandemic is far from over. This project will enable government and stakeholders to identify areas of high risk and empower them in the fight against COVID-19. Due to the nature of next generation sequencing, this project will further be able to investigate the co-occurrence of other pathogens and SARS-CoV-2 in South African water samples. This information will greatly assist in determining the risk of co-infection and the relative quality of water. It should be emphasised that COVID-19 is not spread through water but that a high frequency in a sample would indicate high incidence in the community. If a community with a high COVID-19 infection rate is further exposed to other pathogens in their drinking water, this may lead to a high morbidity and mortality rate which would increase the strain on the health sector.

The AMR potential will be examined and further aid in research with regards to the co-occurrence of SARS-CoV-2 and other pathogens. This project will furthermore validate the use of metagenomic next generation sequencing (mNGS) as a robust approach which is unbiased and provides a wealth of information regarding a sample. Using mNGS it is possible to identify all the pathogens and their AMR potential in a single event without the need for prior knowledge and the cumbersome process of isolation and culturing. This project will further develop capacity in the form of a MSc. student and in general will promote capacity building in the water and science sectors. The results of this project will be published in numerous journals and be presented at various conferences. This project will include training workshops and as such further promote the water

research sector and assist in building capacity across the sector. In short, this project will not only assist in fighting the current COVID-19 pandemic but will build capacity, information and skills for any future resurgence or any other pandemic which may arise. The implementation of mNGS in freshwater and wastewater samples to track SARS-CoV-2, other pathogens and AMR will be critical in South Africa's response to the COVID-19 pandemic. The information obtained from this project will allow for the early detection of COVID-19 hotspots and the possible limitation of resurgence in certain areas.

The results will further allow for the possible determination of viral origin and as such potential preventative actions to be taken in the future. The added value afforded by mNGS of water samples include the detection of other pathogens, both viral and bacterial, and the detection of antimicrobial resistance. The possibility of co-infection and the AMR potential of co-occurring pathogens in water resources may be of dire consequence in a COVID-19 pandemic. Tracking SARS-CoV-2 in water samples will allow for the early detection of COVID-19 in communities and areas. This will greatly assist in flattening the curve and allow policy makers and stakeholders to make pre-emptive decisions. The added information with regards to other pathogens and AMR potential in water resources will enable the channelling of resources to areas where critical intervention is required. As this method is based on an environmental sample it negates the required individual testing and increase in numbers to indicate a hotspot or possible resurgence. This data may as such be employed to facilitate strategies in community isolation before the virus is spread to a broader geographic area. This project will be paramount in the early detection of resurgence and the subsequent containment of infection. The mNGS approach will be based on two techniques, a directed primer approach and a metagenomic approach, both of which will be validated by the current qRT-PCR procedure. These techniques and the downstream analysis are easily packaged and made available for commercial use.

Current indications are that the presence of SARS-CoV-2 and the COVID-19 pandemic will be with us for an extended period of time and that resurgence in infection will be seen internationally. Early development of services such as mNGS testing of water samples for SARS-CoV-2 will therefore be economically feasible and viable in the long run. The added information obtained from a shotgun metagenomic approach, e. g. other pathogens, AMR potential, without the need of isolation and culturing will make this an attractive service in the water value chain. This project will include the training of postgraduate students and future water scientists and as such be strategically involved in the development of human capital in the water and science sectors. The knowledge obtained by those involved in this project may be used in future studies, albeit not on SARS-CoV-2, in the water and science sector as the skills are generally transferrable to other pathogens and viruses. This project furthermore allows for the funding of one MSc. student.

1.3 PROJECT AIMS AND OBJECTIVES

1.3.1 Project aims

The following were are the aims of the project:

1. Detecting the presence and tracking the evolution of SARS-CoV-2 in freshwater and wastewater samples
2. Identification of pathogens co-occurring with SARS-CoV-2
3. Analysing the Antimicrobial Resistance potential of organisms within freshwater and wastewater

1.3.2 Project objectives

The objectives of the project were as follows:

1. Establish collaborations within the SACCESS network and other research groups to receive retrospective, current and future samples
2. Obtain samples and validate protocol. If needed, optimize protocols
3. SARS-CoV-2 diagnosis of samples
4. Metagenomic next generation sequencing, analysis and variant detection
5. Identify samples with SARS-CoV-2 variants and enrich for respiratory viruses
6. Phylogenetic and evolutionary analysis of SARS-CoV-2 lineages present in samples
7. Disseminate results of all samples to collaborators and scientific audience
8. Production and dissemination of final report detailing all results

CHAPTER 2: STUDY DESIGN AND METHODS

2.1 INTRODUCTION

For this work, different sets of wastewater samples were obtained from collaborators under the South African Collaborative COVID-19 Environmental Surveillance System (SACCESS) network across Gauteng and KwaZulu-Natal. The samples, graciously supplied by collaborators, were in various formats including extracted RNA, RNA extracted after viral concentration, extracted DNA and raw wastewater. Three different next-generation sequencing methods and their application thereof in wastewater-based epidemiology was demonstrated in this study. Targeted sequencing as performed by whole genome sequencing of SARS-CoV-2 in wastewater demonstrates the ability of whole genome sequencing to identify SARS-CoV-2 variants of concern in wastewater samples. Amplicon sequencing such as 16S rRNA was used with great success to provide a taxonomic overview of wastewater samples. Untargeted sequencing obtained by means of metagenomic analysis in wastewater surveillance demonstrates the abilities of metagenomic sequencing to generate taxonomic and antimicrobial resistance profiles. The methods used by collaborators for SARS-CoV-2 detection activities have been documented under the WRC publication; *"A compendium of emerging South African testing methodologies for detecting of SARS-CoV-2 RNA in wastewater surveillance"* (WRC, 2020).

2.2 SAMPLE INFORMATION

Wastewater samples used for this were selected from provincial hotspots by collaborators to ensure non-redundancy of sampling and a concerted effort. Samples considered for analysis included retrospective, current and future sampling activities from the selected wastewater sites and adjacent freshwater sources over a period of 12 months. The sampling frequency was based on the collaborator's sampling schedule but a weekly frequency was preferred as this has been recommended for generating timely information on SARS-CoV-2 circulation in a community. The sub-sections below provide information on the samples selected for analysis as means of achieving the objectives of the project.

2.2.1 Samples for determining the taxonomic composition and the presence of antimicrobial resistance genes

Samples (n=20) from regions across Tshwane were used for RNA metagenomic sequencing. These samples were collected between 17 August 2020 and 6 April 2021 and all tested positive for the presence of SARS-CoV-2. Metagenomic sequencing was done on the RNA extractions from the samples graciously provided by the collaborator. The samples were analysed with regards to taxonomic composition and the presence of antimicrobial resistance. Detection of SARS-Covid-2 in the metagenomic data was further included.

Samples (n=30) were collected from 3 wastewater treatment plants in Tshwane, Gauteng. DNA extractions were done by the ARC Biotechnology including library preparation, amplicon and metagenomic sequencing. Amplicon sequencing produced taxonomic profiles for each sample whereas the metagenomic sequencing was able to detect the presence of antimicrobial resistance within the samples.

Wastewater samples (n=10) were collected from three municipal WWTPs in Pretoria, South Africa, that primarily treat household sewage. Grab samples (influent, activated sludge and secondary settling tank (SST) effluent) were collected from November 2021 to February 2022 at different treatment stages and metagenomic sequencing used to construct metagenome assembled genomes (MAGs). The ability to reconstruct partial to near complete genomes enables the taxonomic classification and detection of antimicrobial resistance. This

information is critical as it allows researchers to understand which microorganisms have acquired resistance within a sample and in the community.

Wastewater samples (n=72) were collected from 8 WWTPs located in the East Rand of Gauteng (Mr. W. le Roux). These samples were collected weekly between 26 January 2022 and 22 March 2022 and represent 9 sampling dates. Amplicon and metagenomic sequencing was used to determine the taxonomic and antimicrobial profiles of the samples.

2.2.2 Samples for determining SARS-CoV-2 lineage and variants

Samples (n=73) from across Durban, KwaZulu-Natal, were used for SARS-CoV-2 whole genome sequencing. These samples were collected between 21 July 2020 and 2 November 2021 and all tested positive for the presence of SARS-CoV-2. SARS-CoV-2 whole genome sequencing was done on the RNA extractions from the samples graciously provided by the collaborator. The samples were analysed with regards SARS-CoV-2 lineage and variants detected by means of whole genome sequencing. Currently accepted and published SARS-CoV-2 lineage assignment workflows were implemented and optimised for use in wastewater samples.

2.2.3 Samples selected for the detection and characterisation of viruses

Samples (n=17) from across Durban, KwaZulu-Natal, were used for viral RNA metagenomic sequencing. These samples were collected between 25 August 2020 and 3 August 2021 and all tested positive for the presence of SARS-CoV-2. Metagenomic sequencing was done on RNA extracted after viral concentration using ultra (centricon) filtration graciously provided by the collaborator. The samples were analysed with regards to taxonomic composition. Detection of SARS-Covid-2 in the metagenomic data was further included.

The analyses as performed on each of the sample sets are described below.

2.3 METHODS FOR SAMPLES ANALYSES

2.3.1 General

Samples, including detailed sample collection information, were sent to the ARC-BTP for SARS-CoV-2 RNA detection using the sampling and testing protocols described in the WRC publication; *"A compendium of emerging South African testing methodologies for detecting of SARS-CoV-2 RNA in wastewater surveillance"* (WRC, 2020). Similarly, protocols described in the compendium were used to recover/concentrate the virus. to ensure comparable results to other testing facilities. From consultations with various collaborators, it was apparent that sample extractions needed to facilitate the diagnostic testing for SARS-CoV-2 were already conducted in the samples selected for this study. This was an advantage to this project as only an aliquot of the extracted sample is needed to achieve the aims as set out in this project. Furthermore, this decreases the costing for each sample to be analysed.

2.3.2 Sample collection method

The protocol for grab and composite samples is as follows: The samples were either obtained manually or by means of automated samplers. One (1) litre of wastewater sample was used for testing. However, the volume of sample to be collect varied, depending on the viscosity of the initial sample. CDC protocols were used for guidance (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/wastewater-surveillance/developing-a-wastewater-surveillance-sampling-strategy.html).

2.3.3 Sample processing and analysis

Collected samples were stored at 4°C immediately after collection and, where possible, processed within 24 hours to reduce SARS-CoV-2 RNA degradation and increase surveillance utility. In circumstances where sample processing was not possible within 24 hours after collection, the samples were frozen at -20°C or -70°C. Samples were mixed by inverting samples several times (liquid samples) or by vortexing. The sample was then concentrated by filtration through a membrane whereafter nucleic acid extraction using the CDC approved wastewater surveillance testing method. Thereafter, the method displayed in Figure 2-1 was followed.

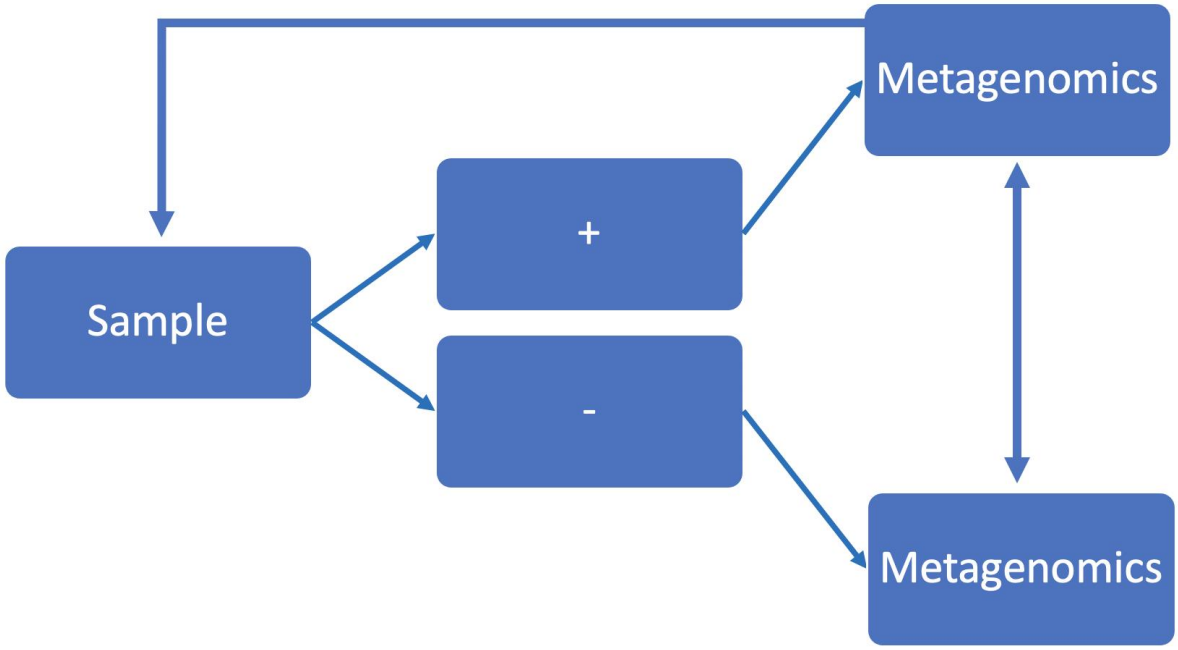


Figure 2-1: Per sample workflow. Samples obtained from collaborators (extracted and not extracted) were firstly subjected to SARS-CoV-2 diagnostics whereafter metagenomic analysis followed. SARS-CoV-2 negative samples were included in this process to serve as a baseline. Within the SARS-CoV-2 samples variants were detected and thereafter resequencing with enrichment of the SARS-CoV-2 samples was conducted. This approach allowed for all the aims as detailed in the project to be achieved in an optimised workflow.

2.3.4 SARS-CoV-2 detection

Initial SARS-CoV-2 diagnosis was performed by the collaborators or by ourselves in an aligned effort to avoid duplication. After the initial diagnosis samples were selected in consultation with collaborators, in such a manner to ensure robust and significant results. These samples were chosen based on location, date of sampling, COVID-19 infection rate, to name a few. As the associated metadata is critical to the significance of the results, a detailed discussion was held with the collaborator(s) in this regard. Samples which were diagnosed as SARS-CoV-2 negative were also included in this analysis to ensure meaningful comparisons and investigations into the co-occurrence of other pathogens. This methodology enabled further detection of possible proxies in SARS-CoV-2 surveillance programs. Initial analysis was based on publicly available datasets and includes the SARS-CoV-2 genome database as hosted by GISAID (Shu and McCauley, 2017). The results obtained from this step allowed for the detection of SARS-CoV-2 variants and in cases where they were found, the initial samples were then enriched for respiratory virus cDNA.

2.3.5 SARS-CoV-2 genomic sequencing

After the initial SARS-CoV-2 diagnosis, another data generation event was conducted to extract complete or near complete SARS-CoV-2 genomes which was used for phylogenetic and evolutionary analysis. The metagenomic approach was based on currently accepted standards and protocols used by BTP for samples such as dung, water, microbiome and other diverse environments and include the application of commercially available extraction kits. BTP has been involved in numerous metagenomic projects which required optimization of extraction and library preparation and as such houses the required capacity to adequately fulfil this requirement inhouse. This initial round of data generation included retrospective samples from collaborators extracted using two different kits and a set of current samples with an inhouse extraction kit. This step enabled us to determine the best extraction procedure based on the data produced and the protocol to be implemented for future samples. The data generation and analysis for both approaches was done at the ARC-BTP, Onderstepoort, South Africa. All sequence data was housed and analysed on the High-Performance Cluster (HPC) located at the ARC-BTP, Onderstepoort, South Africa. Data can be shared with other research groups if an official request is made to the WRC pertaining to development of tools or additional research data. Sequence data was analysed using established and published protocols. This included raw and filtered sequence quality inspection with FASTQC (Andrews, S., 2010). Quality control, adapter removal, decontamination and error correction of the raw sequence data was done using the BBDuk software suite (Bushnell, B.). Filtered reads were aligned to know SARS-CoV-2 genomes using BBMap (Bushnell, B., 2014). This allowed for the identification of mutations and variations in SARS-CoV-2 genomes found in freshwater and wastewater samples.

2.3.6 Determining the taxonomic classification and the presence of other pathogens and antimicrobial resistance genes in samples

Taxonomic classification of the filtered reads was done using Kaiju (Menzel et al., 2016) and Kraken 2 (Wood et al., 2019). This data was used to indicate the general taxonomic composition of a sample and the presence of other pathogens in a sample. This was followed by assembly with metaSPAdes (Nurk et al., 2017), gene prediction with Prodigal (Hyatt et al., 2010) and gene annotation by means of DIAMOND (Buchfink et al., 2015) against the NCBI nr database (Coordinators, N.R., 2018.). Further functional annotation was done using MG-RAST (Meyer et al., 2008) and InterProScan (Jones et al., 2014). Detection and annotation of AMRs in the samples was done using RGI from the CARD database (Alcock et al., 2020) and AMRFinder (Feldgarden et al., 2019). Statistical analysis and visualisation was done using R version 3.6.0 (Team, R.C., 2019).

2.4 SUMMARY OF DATA GENERATED

2.4.1 Collaborations

Numerous meetings with potential collaborators have been held. Successful collaborations to date include Dr A. Mutshembele (SAMRC Tuberculosis Platform), Prof F. Bux (DUT Institute for Water and Wastewater Technology), Dr Noncy Gomba (NIOH), Mr Wouter le Roux (CSIR), Prof Thulani Makhalanyane (UP) and Dr Oliver Bezuidt (UP).

2.4.2 Sample collection

24 Samples for metagenomic sequencing were received from Dr A. Mutshembele of which 20 samples are included in this report. The additional 4 samples will be included in a prospective student project and publication. 73 Samples for whole genome sequencing were received from Prof F. Bux and are included in this report. A subset of 17 samples were used for metagenomic sequencing and the results thereof contained in this report. 39 Samples were received from Dr N. Gomba. These samples were used for whole genome,

amplicon and metagenomic sequencing. 72 Samples were received from Mr W. le Roux and were used for amplicon and metagenomic sequencing.

2.4.3 Data generation

The methodology as described for metagenomics has been implemented and executed by the project team with great success on various environments. In particular, the metatranscriptomic approach has recently been successfully applied to classify viruses as found in the samples. Figure 2-2 below is one example of the visual representation of some of the results obtained using a metagenomic approach as proposed in this project. The ability to identify SARS-CoV-2, co-occurring pathogens and other functional elements such as AMR will ensure that the results produced are significant and accepted by the research community. It will further highlight the power and application of mNGS in water research.

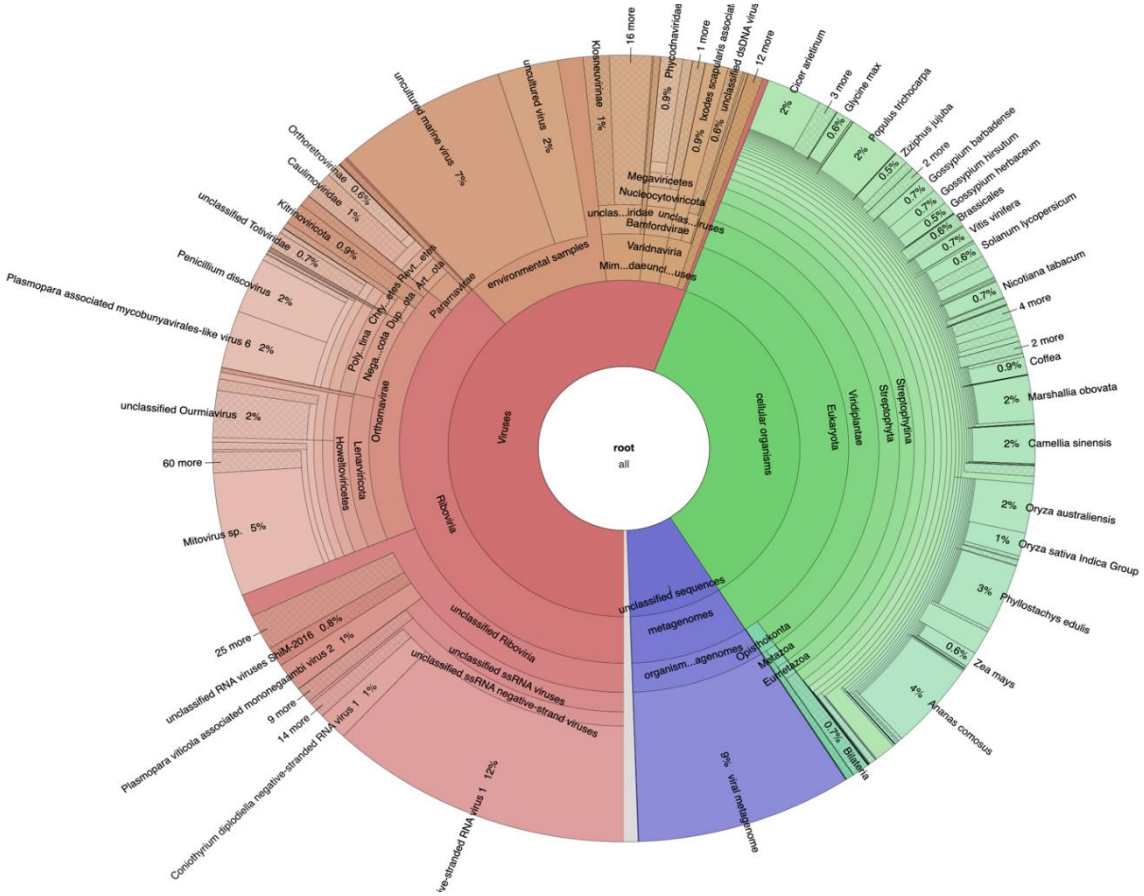


Figure 2-2: Metagenomic analysis of plant viruses using methodology as proposed for this project. This is a visual representation of the frequency and classification of viruses from an environmental sample.

CHAPTER 3: METAGENOMIC SEQUENCING OF WASTEWATER SAMPLES POSITIVE FOR THE PRESENCE OF SARS-COV-2 FROM THE TSHWANE DISTRICT

3.1 INTRODUCTION

Metagenomics is defined as “the application of modern genomics technique without the need for isolation and lab cultivation of individual species” (Chen and Pachter, 2005). This means that genetic material sampled directly from an environment is study and as such negates various of the time consuming and laborious processes associated with the isolation and cultivation of single species. This method therefor allows for the classification of a copious number of organisms as present in a sample. An added benefit is the detection of the functional potential available in a sample.

Metagenomic analysis of wastewater samples provides insights to human health related factors which includes the distribution of pathogens and antibiotic resistance genes (Yang et al., 2014). The contents of a wastewater sample provide researchers and stakeholders a glimpse as to what is circulating in the host associated environment and as such host health. Wastewater samples may be regarded as a pooled version of the human gut microbiome. Pathogens and antimicrobial resistance which are present in a wastewater sample may be presumed to have been present in the population gut microbiome prior to the sampling. These sewage water accurately reflect a population’s gut microbial composition which therefor allows metagenomics to assist in obtaining information regarding the infection dynamics in a given population (Fresia et al., 2019).

The COVID-19 pandemic has increased awareness regarding the power and resolution of next-generation sequencing and genomics. This has been evident in the detection of SARS-CoV-2 variants and the tracking of COVID-19 infection. Wastewater-based epidemiology is a critical component in the detection and tracking of SARS-CoV-2 and it has been shown that sequencing of viral concentrations and RNA extracted directly from wastewater can identify multiple SARS-CoV-2 genotypes, including variants not yet observed in clinical sequencing programmes (Crits-Christoph et al., 2021).

Genomics and in particular metagenomics are therefore an eloquent application in wastewater surveillance and epidemiology. This method enables the detection and classification of a multitude of organisms in a sample in a single data generation event or sequencing run. Additionally, it allows for the detection of antimicrobial resistance and other functionalities. The results obtained from a metagenomic sequencing event can further be stored for long term use and be used as a baseline for future research endeavours.

Although not the preferential method in SARS-CoV-2 detection, metagenomics still has the potential to screen samples for possible fragments or portions of the SARS-CoV-2 genome on a large scale. The frequent metagenomic analysis of wastewater samples will alert stakeholders, government and other interested bodies to the detection of SARS-CoV-2 and allow for the rapid implementation of target testing. The data generated in these metagenomic sequencing events will further be available for various other research endeavours and surveillance projects.

In the sections below, we clearly outline the methodology used and results obtained in the metagenomic analysis of wastewater samples obtained from the Tshwane region during the period 17 August 2020 and 6 April 2021. The results illustrate the functionality, benefits and potential of metagenomic sequencing of wastewater samples.

3.2 MATERIALS AND METHODS

Grab samples (n=20) were collected from various wastewater treatment sites across the Tshwane district by the SAMRC (Dr Awelani Mutshembele). The sampling sites included Baviaanspoort (n=3), Daspoort (n=6) and Rietgat (n=11) wastewater treatment plans. These sampling sites cover Tshwane east (Baviaanspoort), central (Daspoort) and west (Rietgat). The samples were collected between 17 August 2020 and 6 April 2021 (Table 3-1 and Figure 3-1). These samples all tested positive for the presence of SARS-CoV-2. RNA extractions were done by the SAMRC and the resulting extractions delivered to the ARC Biotechnology for library preparation and sequencing (Supplementary Sequencing Quotation).

Table 3-1: Samples received for metagenomic sequencing.

Sample ID	Type of sample	Concentration	A260/280	A260/230	Collection Date	Collection Site
BSW1_1A	Grab	1294,58	2,197	2,347	2020/08/17	Baviaanspoort
RTW1_1A	Grab	2617,13	2,219	2,394	2020/08/17	Rietgat
BSW2_1A	Grab	3686,30	2,228	2,378	2020/08/31	Baviaanspoort
RTW2_1A	Grab	1907,83	2,236	2,327	2020/08/31	Rietgat
BSW6_1A	Grab	818,19	2,204	2,348	2020/10/26	Baviaanspoort
DW6_1A	Grab	1621,13	2,231	2,348	2020/10/26	Daspoort
RTW6_1A	Grab	1373,90	2,21	2,42	2020/10/26	Rietgat
DW7_1A	Grab	1429,46	2,202	2,332	2020/11/09	Daspoort
RTW7_1A	Grab	1502,49	2,228	2,392	2020/11/09	Rietgat
DW8_1A	Grab	1343,98	2,197	2,276	2020/11/23	Daspoort
RTW8_1A	Grab	583,27	2,19	2,288	2020/11/23	Rietgat
DW10_1A	Grab	1224,639	2,151	2,338	2020/12/21	Daspoort
RTW10_1A	Grab	685,23	2,169	2,202	2020/12/21	Rietgat
RTW11_1A	Grab	2057,402	2,197	2,216	2021/01/04	Rietgat
DW12_1A	Grab	1802,222	2,244	2,407	2021/01/18	Daspoort
RTW12_1A	Grab	4644,048	2,255	2,327	2021/01/18	Rietgat
RTW13_1A	Grab	886,559	2,165	2,267	2021/03/23	Rietgat
RTW14_1A	Grab	693,557	2,055	2,134	2021/03/29	Rietgat
DW15_1A	Grab	192,5	1,99	1,5	2021/04/06	Daspoort
RTW15_1A	Grab	1126,2	2,29	2,14	2021/04/06	Rietgat

RNA samples were processed using a ribodepletion step to remove abundant RNAs such as rRNAs and globin RNAs which then enables us to focus on the high-value, informative portions contained within the mRNA, and therefor also lower the sequencing cost. The resulting libraries were sequenced on a HiSeq 2500 with roughly 4.1 GB of data per sample requested.

Initial sequence data quality and filtered data quality was inspected using FastQC version 0.11.8 (Andrews, S., 2010). Sequence data was quality trimmed and filtered, including adapter removal and decontamination, using BBDuk version 38.91 available from the BBTools suite of tools (Bushnell, B., 2014). Human contamination in the quality filtered sequencing data was removed by aligning the sequence data against the latest reference human genome (GRCh38.p13) using BBMap version 38.91, available from BBTools. To identify the presence of SARS-CoV-2 in these samples, filtered and decontaminated paired-end reads were aligned to the SARS-CoV-2 reference genome (NC_045512.2) with BBMap and coverage statistics calculated. Possible correlations between the amount of sequence data or 7-Day average COVID-19 cases and the detection of SARS-CoV-2 in the metagenomic sequencing data was tested using Spearman's rank correlation coefficient as available from the ggstatsplot library (Patil, I., 2021) implemented in RStudio version 1.4.1717 (Team, RStudio, 2021) and R version 4.0.2 (Team, R Core, 2020).

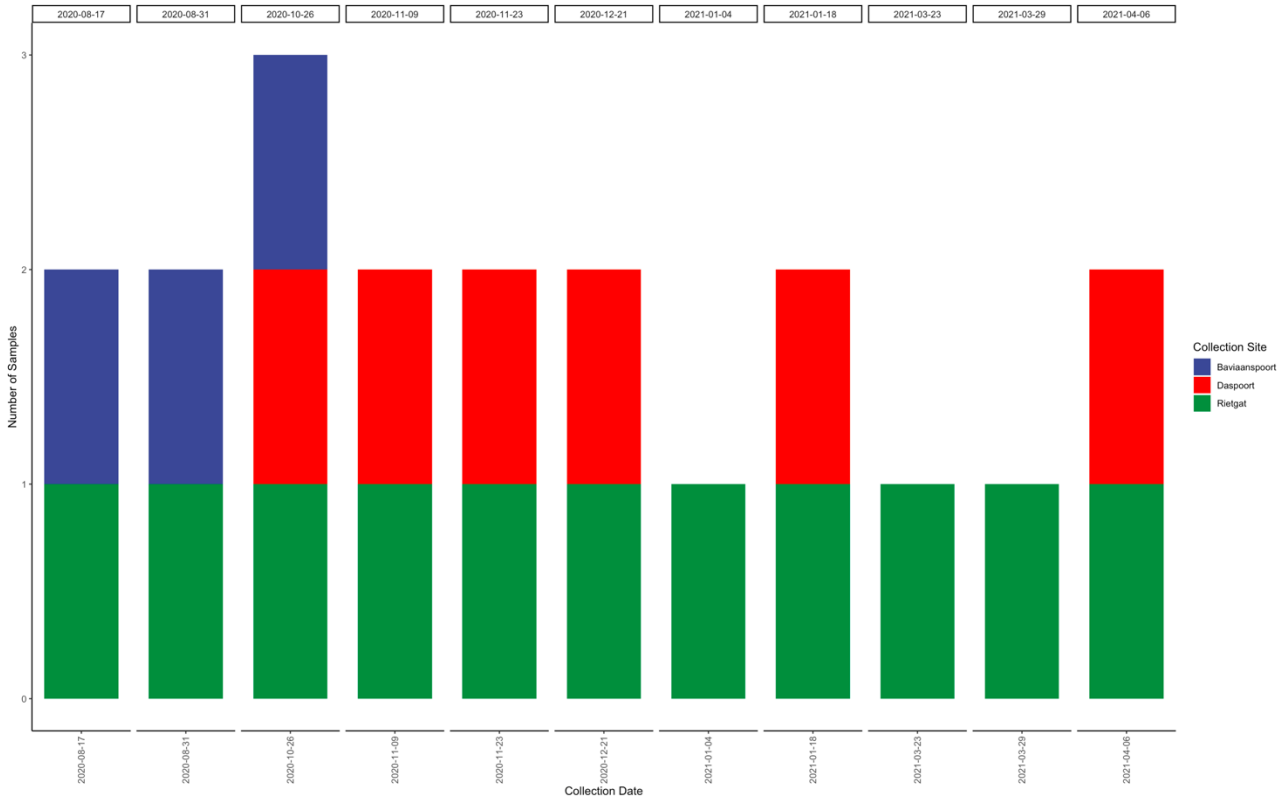


Figure 3-1: Samples received for metagenomic sequencing. The colours indicate the sampling location, x-axis the date of sampling and y-axis the number of samples. The subtitles on the top of each bar indicate the sampling date.

Taxonomic classification of the filtered and decontaminated sequencing data was done using Kaiju version 1.8.0 (Menzel et al., 2016) and the Kaiju formatted refseq database as available on 2021/02/26. The Kaiju formatted refseq database contains complete assembled and annotated reference genomes of Archaea, Bacteria, and viruses from the NCBI RefSeq database (O'Leary et al., 2016).

The quality filtered sequence data was used to identify the presence of antimicrobial resistance (AMR) genes in the samples. The Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2020) contains 3,339 reference sequences including the associated annotations and phenotypes. The database was accessed on 2021/09/01 and the “nucleotide_fasta_protein_homolog_model.fasta” file used as suggested by Alcock et al (2020). The sequence data was aligned against the antibiotic resistance genes using BBMap and coverage statistics calculated. Results were filtered to only include antibiotic resistance genes covered by at least 80% by the sequencing data. These would represent high confidence alignments. Each antibiotic resistance gene is annotated with an Antibiotic Resistance Ontology (ARO) accession which is then further categorized by gene family, drug class and resistance mechanism.

The quality filtered sequencing data was assembled into transcripts using SPAdes version 3.15.3 (Bushmanova et al., 2019) with the “--rna” option enabled. The “hard_filtered_transcripts.fasta” was used for further analysis. Variations in the number of transcripts between samples was investigated by testing correlations between the amount of sequence data or 7-Day average COVID-19 cases and the number of transcripts per sample. This was done using Spearman’s rank correlation coefficient as available from the ggstatsplot library implemented in RStudio version 1.4.1717 and R version 4.0.2.

The filtered transcripts for all samples were samples were aligned against the NCBI nt database (Sayers et al., 2021) using blastn version 2.12.0+ with the “megablast” option invoked and an e-value cut-off set to 1e-5. The results were filtered for the top hit with at least 80% identity with an alignment length of at least 80% of the query transcript to identify taxonomic classification.

This methodology was further used to detect AMRs in the filtered transcripts with the CARD as reference. The results were filtered for the top hit with at least 80% identity with an alignment length of at least 80% of the subject reference.

3.3 RESULTS

3.3.1 Data quality filtering and decontamination

Approximately 82 GB worth of raw sequencing data was produced for the 20 samples. The raw sequencing data was quality filtered and the resulting sequence quality of the filtered reads were again inspected using FastQC. Sequencing data which mapped to the human genome was removed and the quality of the remaining sequence data again quality checked with FastQC. The number of paired-end sequences for each sample is presented in Table 3-2 and Figure 3-2. Data loss due to quality and contamination was as expected and more than enough paired-end reads remained for further analysis. The low levels of data loss after decontamination, i.e. human, clearly illustrates the application of the ribodepletion step. This step greatly assists in removing unwanted human ribosomal RNA and as such allows for focused sequencing on the archaeal, bacterial and viral portions of samples. If this step was not included a large portion of the data sequenced would have been of human origin and not usable in this project.

Table 3-2: Number of paired-end reads at each stage of quality control and decontamination.

Sample ID	Raw Reads	QC Reads	No Human QC Reads
BSW1_1A	16,401,593	15,366,564	15,354,175
RTW1_1A	20,558,275	19,030,402	19,029,720
BSW2_1A	25,492,099	22,701,059	22,699,962
RTW2_1A	20,014,123	18,403,174	18,402,192
BSW6_1A	19,141,747	17,879,637	17,879,263
DW6_1A	22,209,262	20,774,336	20,773,285
RTW6_1A	20,361,919	19,193,278	19,191,816
DW7_1A	25,232,401	23,709,787	23,707,775
RTW7_1A	19,473,521	18,037,201	18,036,076
DW8_1A	28,563,786	26,571,417	26,562,715
RTW8_1A	17,541,916	16,345,364	16,326,850
DW10_1A	20,695,652	19,525,114	19,523,460
RTW10_1A	19,764,354	18,306,700	18,300,525
RTW11_1A	17,895,659	16,794,703	16,789,746
DW12_1A	10,576,422	9,603,211	9,540,452
RTW12_1A	16,120,513	14,881,971	14,877,332
RTW13_1A	19,272,681	17,745,505	17,743,870
RTW14_1A	18,062,438	17,014,991	17,013,800
DW15_1A	23,166,309	21,886,226	21,884,681
RTW15_1A	18,373,620	17,414,457	17,410,347

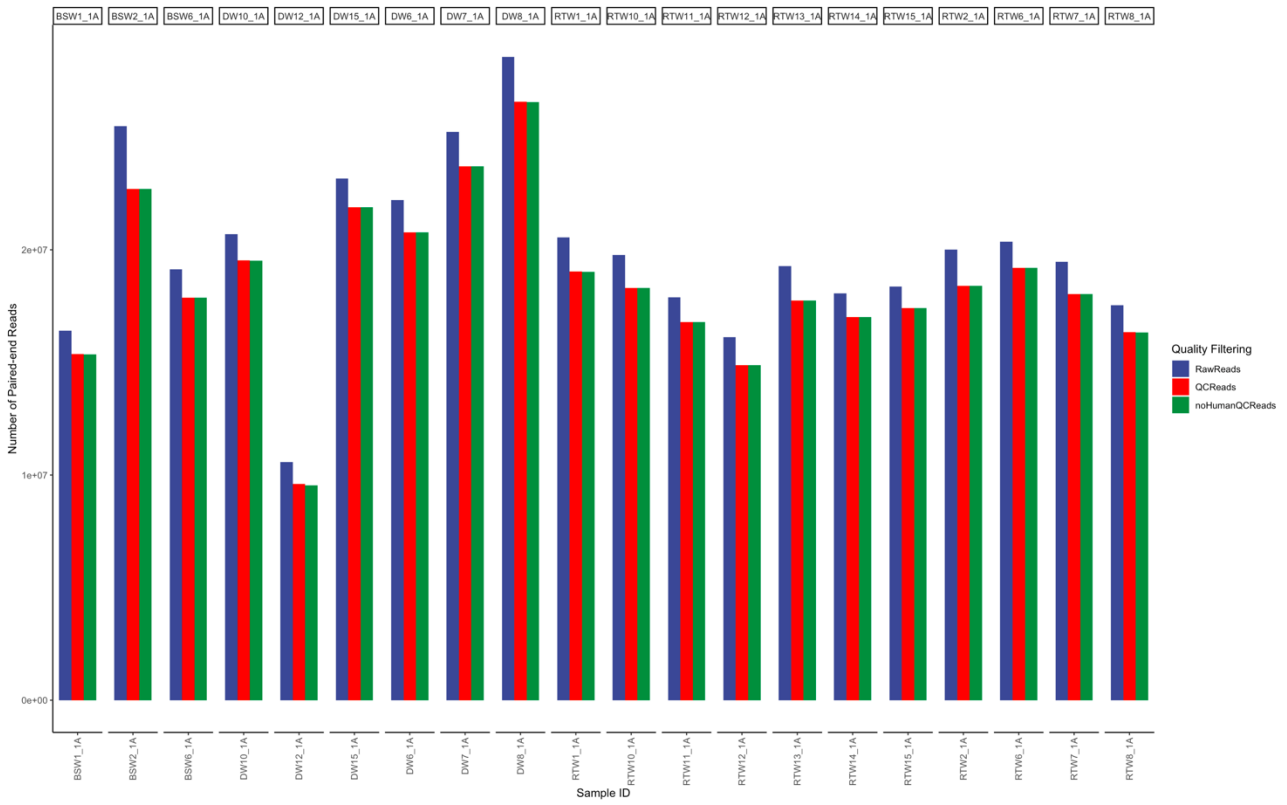


Figure 3-2: Number of paired-end reads at each stage of quality control and decontamination. The colours indicate the quality control step, x-axis the sample and y-axis the number of paired-end reads. The subtitles on the top of each bar indicate the sample ID. Low levels of data loss was seen and the number of paired-end reads surviving quality filtering and human decontamination was more than adequate for the project.

It should be emphasized that the computational removal of human sequencing data is still required as non-ribosomal RNA will still be present in a sample after the ribodepletion protocol. This portion of the data may influence results and workflows and it is therefore recommended to still filter the data for any human contamination. The lowest amount of paired-end reads remaining after quality filtering and decontamination was in excess of 9 million paired-end reads. This is more than enough data for adequate inferences and exploratory analysis. It should be noted that these are RNA samples and not DNA samples. This data therefore represents the actively expressed portions of archaeal, bacterial and viral genomes and will further include the genomic content of RNA viruses.

3.3.2 Detection of SARS-CoV-2

The presence of SARS-CoV-2 fragments were detected in 5 samples using RNA metagenomic sequencing (Table 3-3 and Figure 3-3). This was as expected as the RNA metagenomic sequencing protocol is not target or directed against the SARS-CoV-2 genome. As all 20 samples were positive for the presence of SARS-CoV-2 using conventional diagnostics, possible reasons for the detection of SARS-CoV-2 in only 5 samples was the amount of sequence data generated per sample or the viral load present in a sample. Sequencing is measured by the amount of data generated. The higher the amount of data or reads per sample the greater the possibility of detecting all microorganisms present in a sample. In RNA metagenomics one pays for the amount of data generated and as such there is a trade-off between cost and detection. In essence, the more you sequence the greater the possibility of detecting SARS-CoV-2 in a sample using this approach. Another possibility for the detection of SARS-CoV-2 is the amount of virus present in a sample. Higher quantities of the virus present in a sample will increase the probability of viral segments being sequenced and as such detected.

Table 3-3: Number of paired-end reads at each stage of quality control and decontamination.

Sample ID	Reference Covered Percent	Collection Date	Collection Site
BSW1_1A	0.1906	2020/08/17	Baviaanspoort
DW6_1A	0.8360	2020/10/26	Daspoort
DW12_1A	0.6220	2021/01/18	Daspoort
RTW10_1A	1.9095	2020/12/21	Rietgat
RTW11_1A	0.5685	2021/01/04	Rietgat

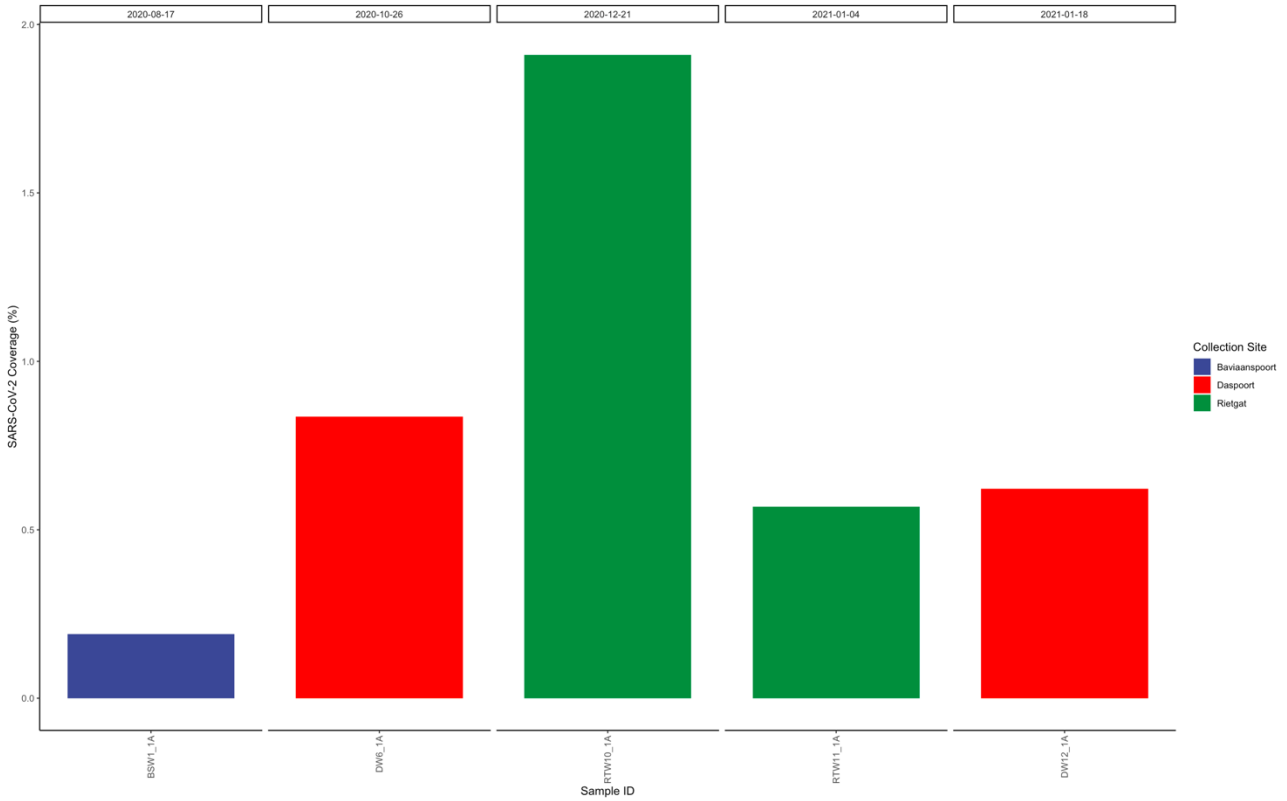


Figure 3-3: Percentage of the SARS-CoV-2 reference genome (NC_045512.2) covered per sample. The colours indicate the sample collection site, x-axis the sample and y-axis the percentage SARS-CoV-2 reference genome coverage. The subtitles on the top of each bar indicate the sampling date.

As all 20 samples were positive for the presence of SARS-CoV-2 using conventional diagnostics, possible reasons for the detection of SARS-CoV-2 in only 5 samples was the amount of sequence data generated per sample or the viral load present in a sample. Sequencing is measured by the amount of data generated. The higher the amount of data or reads per sample the greater the possibility of detecting all microorganisms present in a sample. In RNA metagenomics one pays for the amount of data generated and as such there is a trade-off between cost and detection. In essence, the more you sequence the greater the possibility of detecting SARS-CoV-2 in a sample using this approach. Another possibility for the detection of SARS-CoV-2 is the amount of virus present in a sample. Higher quantities of the virus present in a sample will increase the probability of viral segments being sequenced and as such detected.

The number of paired-end reads and 7-Day average COVID-19 cases were tested as contributors to the amount of SARS-CoV-2 detected (Figure 3-4). The Spearman's rank correlation coefficient test results indicated that there was no correlation between SARS-CoV-2 coverage (%) and amount of data generated, i.e. number of paired-end reads (p-value = 0.2447) (Figure 3-4.a). The 7-Day average COVID-19 cases indicated a significant positive correlation with SARS-CoV-2 coverage (%) (p-value = 0.0321) (Figure 3-4.b).

This result clearly indicates that the detection of SARS-CoV-2 using RNA metagenomics is influenced by the amount of virus present in a sample. In essence, the higher the amount of COVID-19 cases reported leads to a higher viral load in wastewater samples. This then increases the probability of recovering SARS-CoV-2 genome fragments from a sample by means of RNA metagenomic sequencing. Bearing in mind the high levels of diversity or microbial content the detection of portions of the SARS-CoV-2 genome within some of these samples clearly illustrates the power of RNA metagenomic sequencing in pathogen surveillance and wastewater analysis.

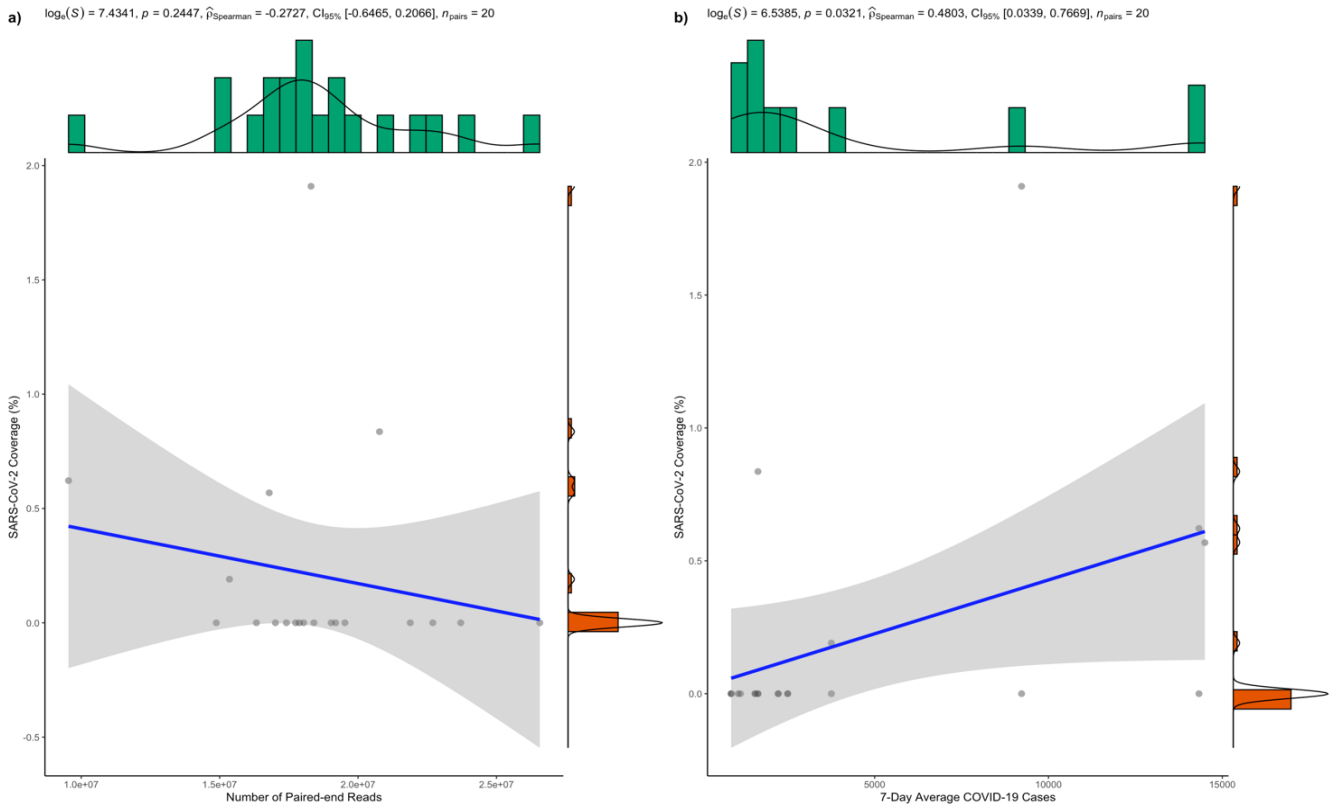


Figure 3-4: Spearman's rank correlation coefficient test results for a) correlation between SARS-CoV-2 coverage (%) and amount of data generated and b) correlation between SARS-CoV-2 coverage (%) and 7-Day average COVID-19 cases. The results from the statistical test are reported in the subtitles on the top of each graph. The marginal distributions for the x and y variables are overlaid on the axes of each graph.

3.3.3 Taxonomic profile of samples based on unassembled sequencing data

Taxonomic classification as produced by Kaiju using the quality filtered, decontaminated reads indicated a high proportion of Bacterial paired-end reads in the samples, as was expected. The RNA metagenomic protocol was further able to detect various Archaea and Viruses. Certain samples did indicate a higher relative abundance of Archaeal and Viral paired-end reads and is of interest. Deviations such as these clearly indicate that the microbial composition or diversity in wastewater samples fluctuates and is not constant. These fluctuations may be linked to various factors such as circulating pathogens in a community and should be further investigated. The relative abundance, as indicated by the percentage of reads which mapped to a taxonomic kingdom are presented in Figure 3-5 and Table 3-4. The large red portions of each bar represent Bacteria and it is evident from the figure below that the largest portion of paired-end reads per sample were classified as bacterial of origin.

The blue segments indicate Archaeal paired-end reads and the green segments those of Viral origin. The relative abundance of these fluctuate across samples and are not bound to a certain sampling area or location.

The ability of RNA metagenomic sequencing to identify various Kingdoms in a single sample is emphasised by the graph above and illustrates the applicability of this protocol in wastewater analysis and testing. This method does not discriminate or isolate and provides a greater understanding of the current content of a wastewater sample.

Each of the taxonomic kingdoms were further inspected at different taxonomic levels.

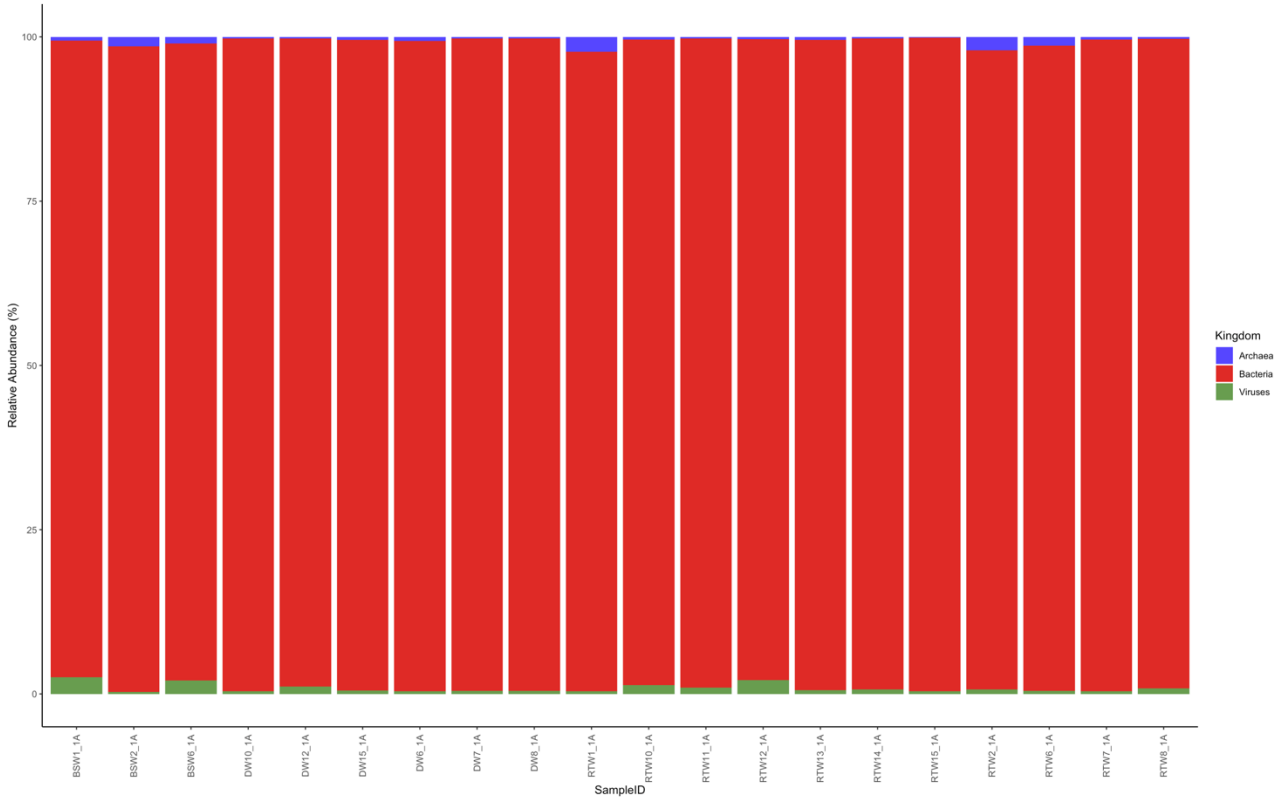


Figure 3-5: Relative abundance, as indicated by the percentage of reads, for Archaea, Bacteria and Viruses for each sample. The colours represent the different taxonomic kingdoms. The samples are on the x-axis and the relative abundance of each kingdom is displayed on the y-axis.

Table 3-4: Relative abundance, as indicated by the percentage of reads, for Archaea, Bacteria and Viruses for each sample.

Sample ID	Archaea	Bacteria	Viruses
BSW2_1A	1.41861212158	98.2827326949	0.298655183491
RTW11_1A	0.25119851404	98.7641715917	0.984629894277
RTW12_1A	0.358577033904	97.5485868432	2.09283612286
RTW15_1A	0.155482007212	99.4353548159	0.409163176873
DW6_1A	0.603934214132	98.995582836	0.400482949846
DW7_1A	0.220161058384	99.3104687949	0.469370146678
RTW1_1A	2.28276948959	97.3111469926	0.406083517844
RTW6_1A	1.32463538414	98.1930587846	0.48230583123
RTW7_1A	0.392123066316	99.1744262115	0.433450722157
RTW10_1A	0.415395031064	98.264809171	1.31979579793
BSW6_1A	0.983129484161	96.9828933386	2.03397717723
RTW8_1A	0.276712678491	98.8697353152	0.853552006335
DW12_1A	0.224453679585	98.6602771868	1.11526913364
DW15_1A	0.443900272486	99.0545812477	0.50151847984
RTW13_1A	0.439440018375	98.965600259	0.59495972261
BSW1_1A	0.555811129797	96.9016231897	2.54256568049
DW10_1A	0.2516505397	99.3504034653	0.39794599504
DW8_1A	0.234004547432	99.3183479927	0.447647459886
RTW14_1A	0.241099942208	99.0452920899	0.713607967904
RTW2_1A	2.05606539291	97.2640718692	0.679862737934

The Archaeal portion indicated the presence of 7 different phyla of which 4 were classified as *Candidatus* (Figure 3-6). This term *Candidatus* indicates that the phylum is well characterized but yet-uncultured. This is of interest and clearly illustrates the power of metagenomic sequencing and the ability to classify unculturable or yet-uncultured phyla in a sample. The *Candidatus* phyla observed at phylum level were *Candidatus* Korarchaeota, *Candidatus* Lokiarchaeota, *Candidatus* Micrarchaeota and *Candidatus* Thermoplasmatoa.

The Archaeal diversity and differences between the various samples became evident when moving down to the lower taxonomic ranks of class (Figure 3-7), order (Figure 3-8), family (Figure 3-9) and genus (Figure 3-10). Per sample Archaeal taxonomic classifications are further available in the Supplementary Material. Various *Candidatus* classifications were found for Archaeal orders, families and genera. The Archaeal order classifications included *Candidatus* Nitrosocaldales, the families *Candidatus* Nitrosocaldaceae and *Candidatus* Methanomethylophilaceae with *Candidatus* Halobonum, *Candidatus* Korarchaeum, *Candidatus* Mancarchaeum, *Candidatus* Methanomethylophilus, *Candidatus* Methanoplasma, *Candidatus* Micrarchaeum, *Candidatus* Nitrosocaldus, *Candidatus* Nitrosocosmicus, *Candidatus* Nitrosomarinus, *Candidatus* Nitrosopelagicus, *Candidatus* Nitrosotenuis and *Candidatus* Prometheoarchaeum found in the Archaeal genera classification.

The detection of Archaeal communities, including various *Candidatus* classifications, in wastewater samples highlights the application of metagenomic sequencing in wastewater surveillance. The ability to extract the taxonomic information for the Archaeal portion of a sample negates the laborious, costly and time consuming efforts normally associated with the isolation and cultivation of Archaeal individuals. Although Archaea are not currently regarded as pathogenetic the occurrence and diversity of the Archaeal population in wastewater should be monitored.

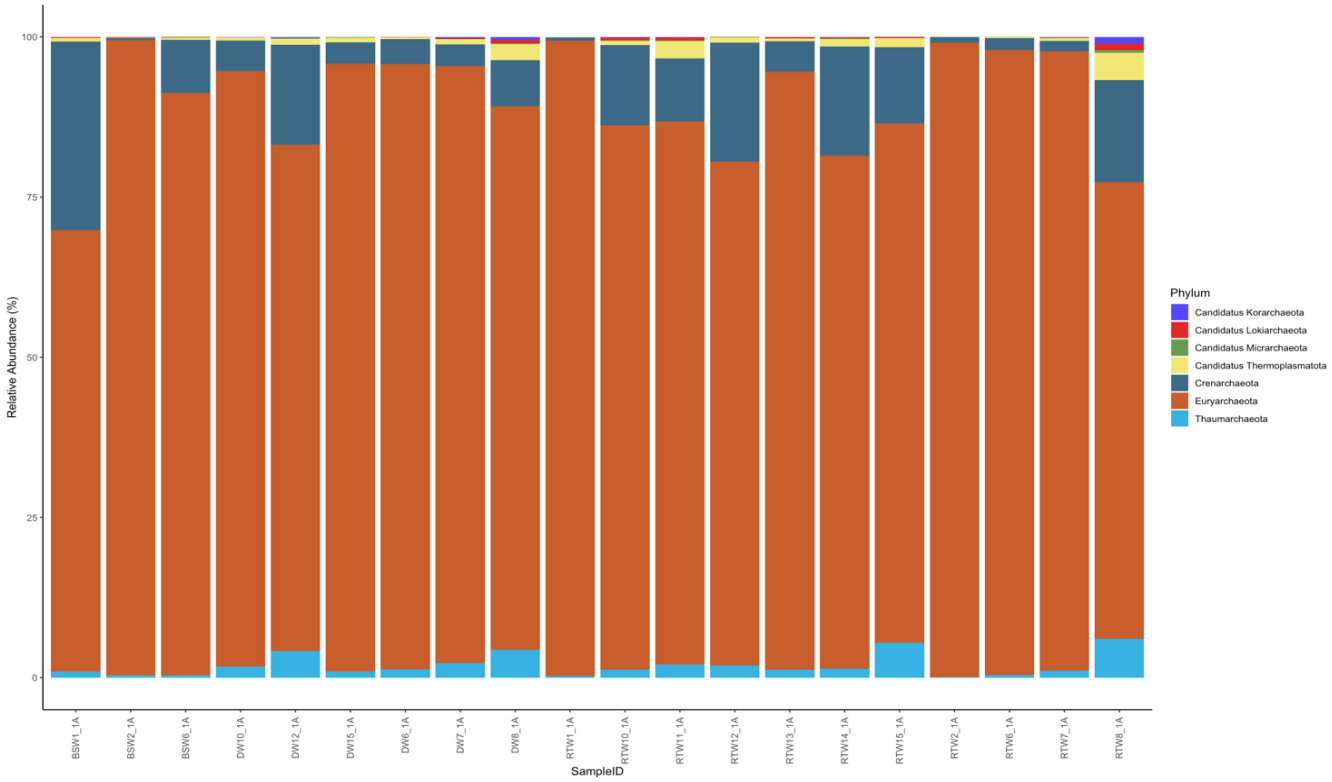


Figure 3-6: Relative abundance, as indicated by the percentage of reads, for Archaeal phyla. A total of 7 phyla were detected with 4 of these classified as *Candidatus.* Each colour is representative of a phylum.

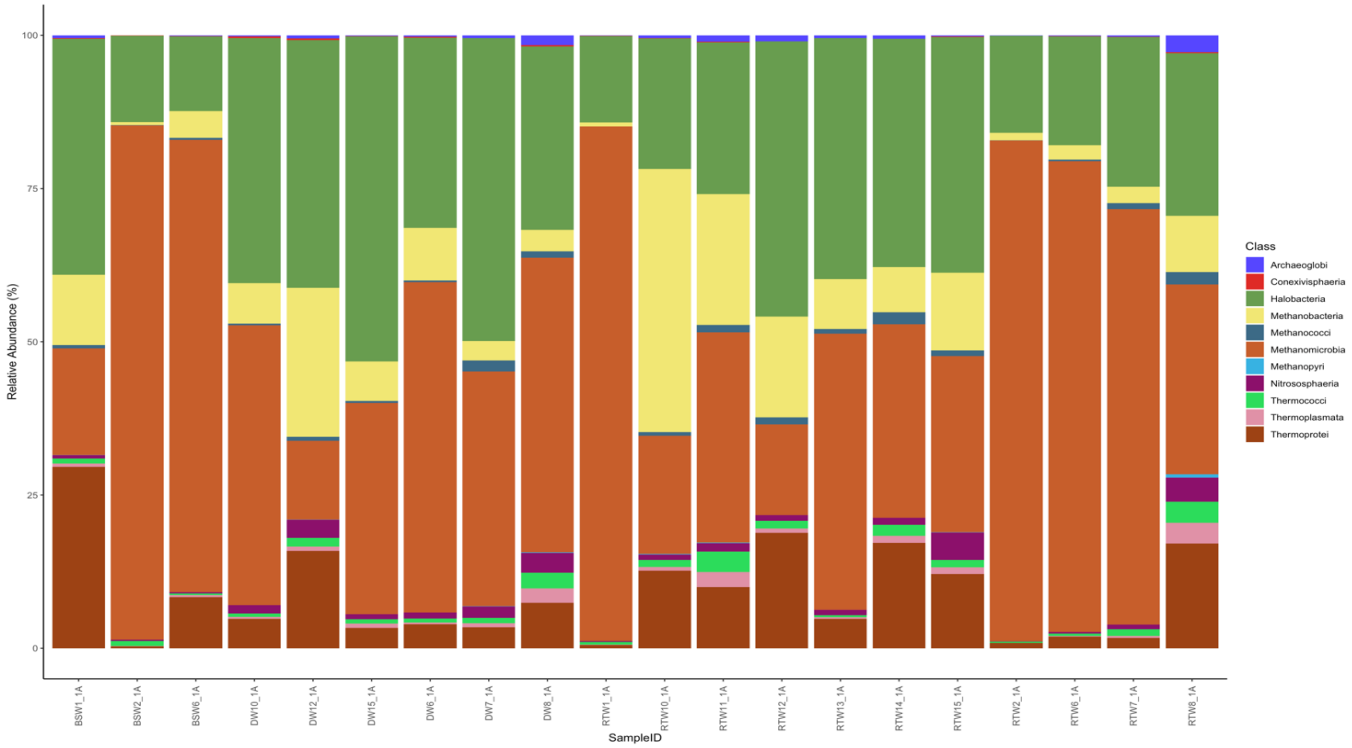


Figure 3-7: Relative abundance, as indicated by the percentage of reads, for Archaeal classes. A total of 11 different classes were detected and are each represented by a different colour. Visually, differences between the samples based on Archaeal classes are evident.

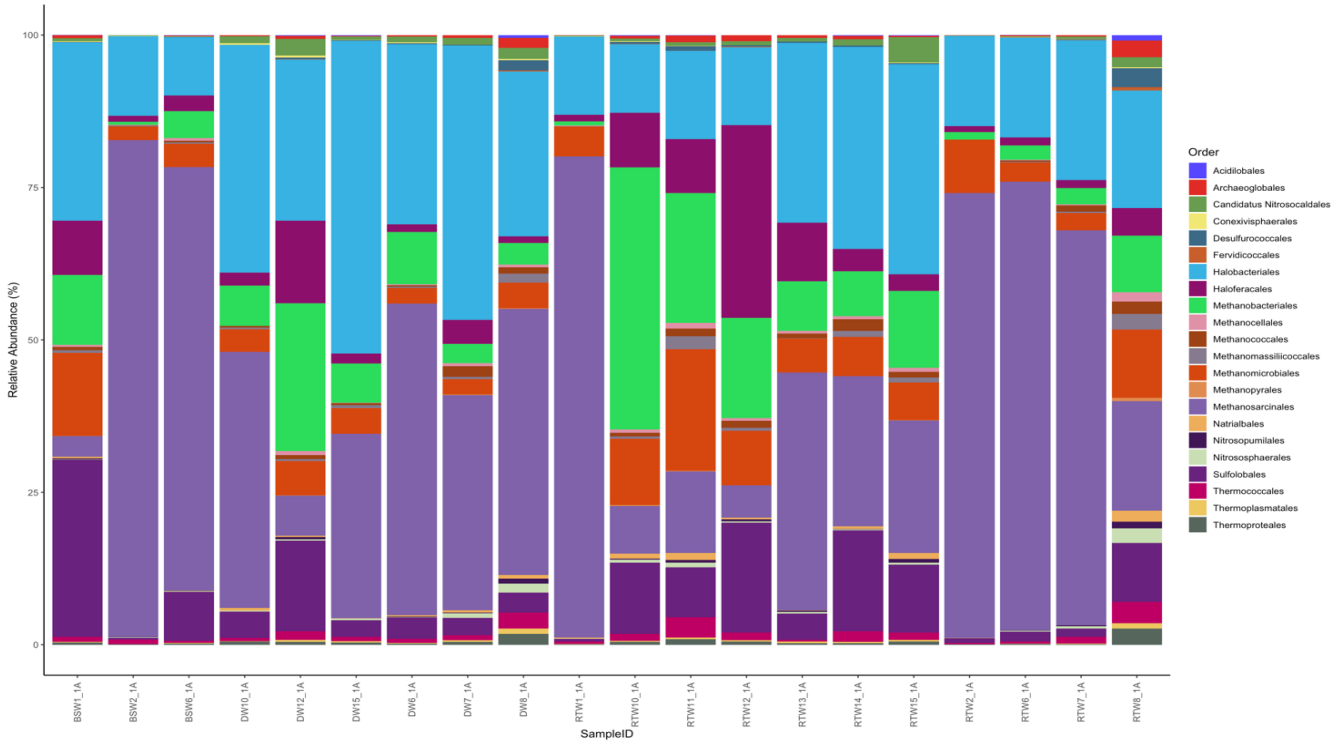


Figure 3-8: Relative abundance, as indicated by the percentage of reads, for Archaeal orders. A total of 22 different orders were detected and are each represented by a different colour. Visually, differences between the samples based on Archaeal orders are evident.

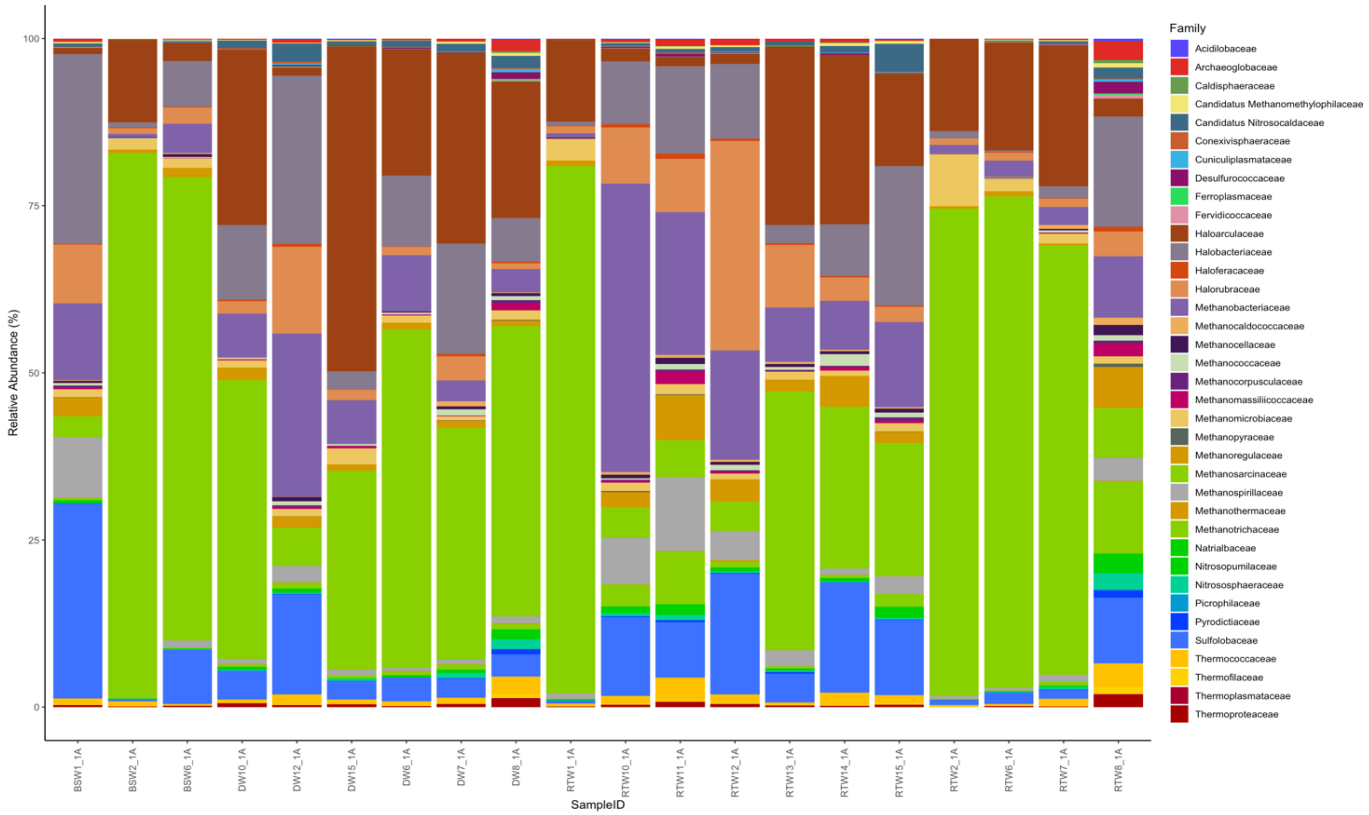


Figure 3-9: Relative abundance, as indicated by the percentage of reads, for Archaeal families. A total of 37 different families were detected and are each represented by a different colour. Visually, differences between the samples based on Archaeal families are evident. Of particular interest is the detection of *Candidatus* Nitrosocaldaceae and *Candidatus* Methanomethylophilaceae.

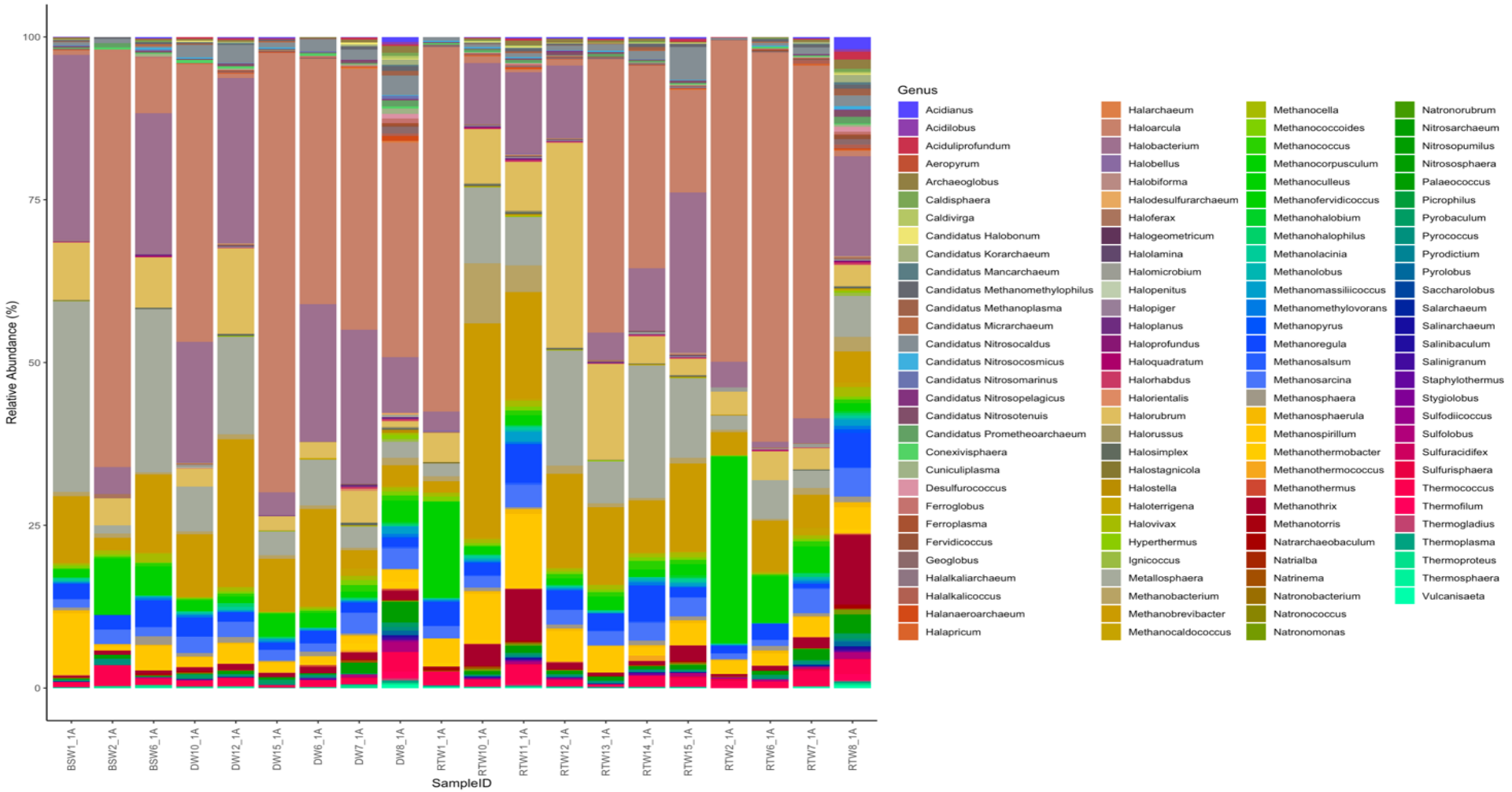


Figure 3-10: Relative abundance, as indicated by the percentage of reads, for Archaeal genera. A total of 118 different genera were detected and are each represented by a different colour. Visually, differences between the samples based on Archaeal genera are evident. Of particular interest is the detection of *Candidatus* Halobonum, *Candidatus* Korarchaeum, *Candidatus* Mancarchaeum, *Candidatus* Methanomethylophilus, *Candidatus* Methanoplasma, *Candidatus* Micrarchaeum, *Candidatus* Nitrosocaldus, *Candidatus* Nitrosocosmicus, *Candidatus* Nitrosomarinus, *Candidatus* Nitrosopelagicus, *Candidatus* Nitrosotenuis and *Candidatus* Prometheoarchaeum.

The Bacterial portion represented the highest relative abundance for each of the samples, as was expected. In particular, the phylum Proteobacteria was highly represented and was followed by Bacteriodetes (Figure 3-11). The samples further included 4 *Candidatus* phyla, i.e. *Candidatus* Bipolaricaulota, *Candidatus* Cloacimonetes, *Candidatus* Omnitrophica and *Candidatus* Saccharibacteria. Numerous Bacterial classes (n=74) and orders (n=168) were detected across the samples (Figure 3-12 and Figure 3-13). Bacterial classes included *Candidatus* Babeliae, *Candidatus* Brocadiae and *Candidatus* Saccharimonia and the orders included *Candidatus* Babeliales, *Candidatus* Brocadiales, *Candidatus* Nanopelagicales and *Candidatus* Nanosynbacterales. Due to the high diversity per sample, the top 10 bacterial families and genera, based on relative abundance, per sample were inspected and are presented in Figure 3-14 and Figure 3-15. The top 10 genera per sample resulted in a combined set of 23 genera across all 20 samples. These are further described in Table 3-5. Per sample taxonomic classification is further available in the Supplementary Material.

Of interest was the high relative abundance of genera generally associated with disease or pathogenicity. These included the genera Aeromonas, Arcobacter, Coxiella, Klebsiella, Listeria, Moraxella, and Pseudomonas to name a few. The ability to detect a multitude of possible bacterial pathogens from a single sequencing event illustrates the power of metagenomic and in particular RNA metagenomic sequencing of wastewater samples for surveillance. This methods negates the individual isolation and cultivation events needed to cover a broad range of bacterial pathogens and provides a holistic overview of a sample.

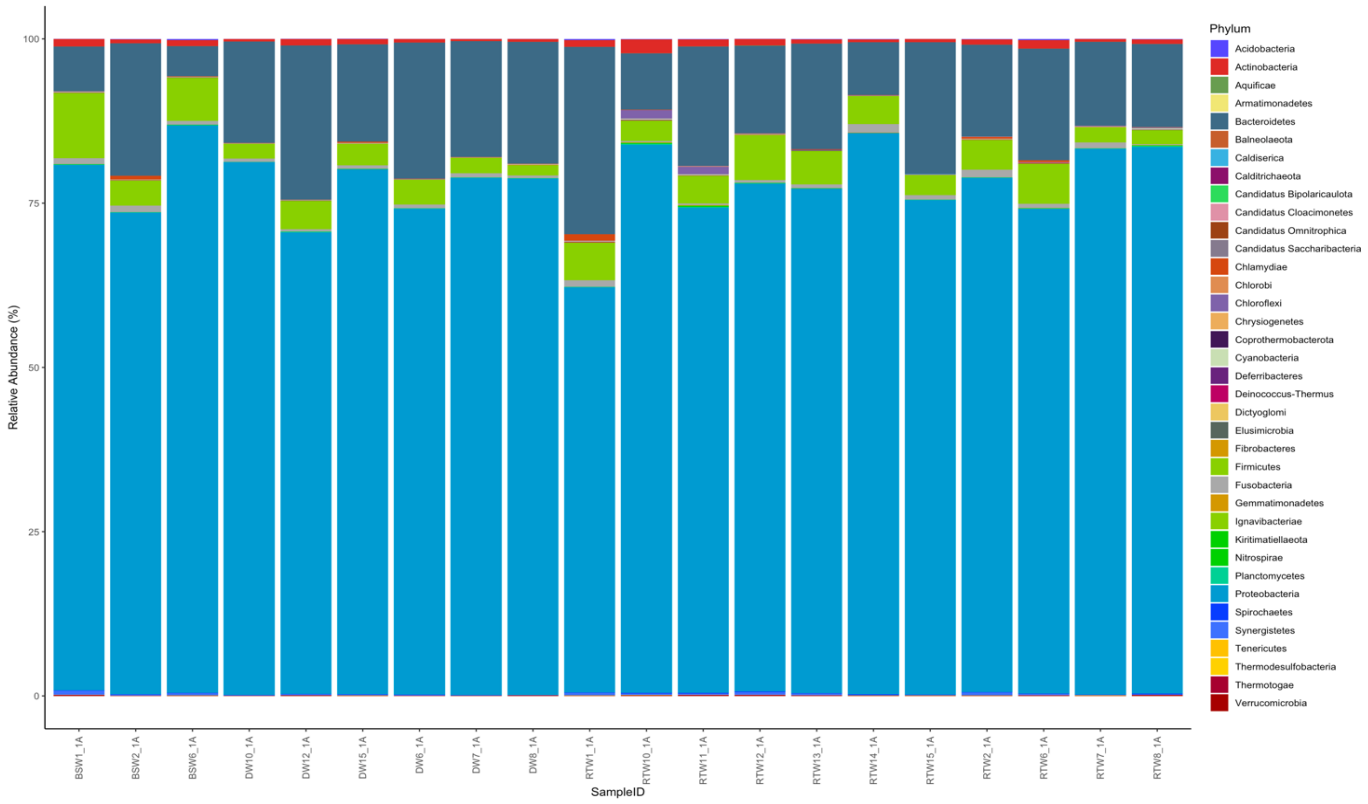


Figure 3-11: Relative abundance, as indicated by the percentage of reads, for Bacterial phyla. A total of 37 phyla were detected with 4 of these classified as *Candidatus.* Each colour is representative of a phylum.

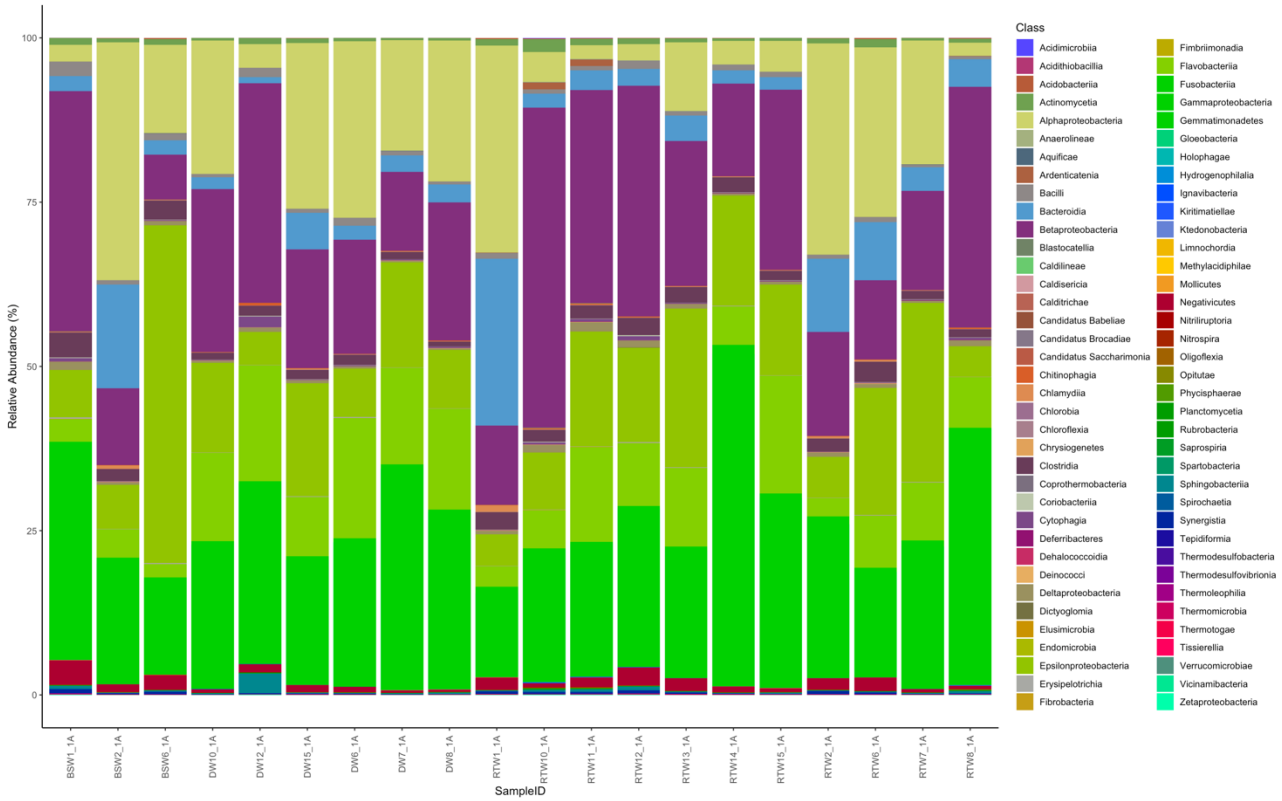


Figure 3-12: Relative abundance, as indicated by the percentage of reads, for Bacterial classes. A total of 74 classes were detected with 3 of these classified as *Candidatus.* Each colour is representative of a class. Differences in bacterial composition across samples are clear.

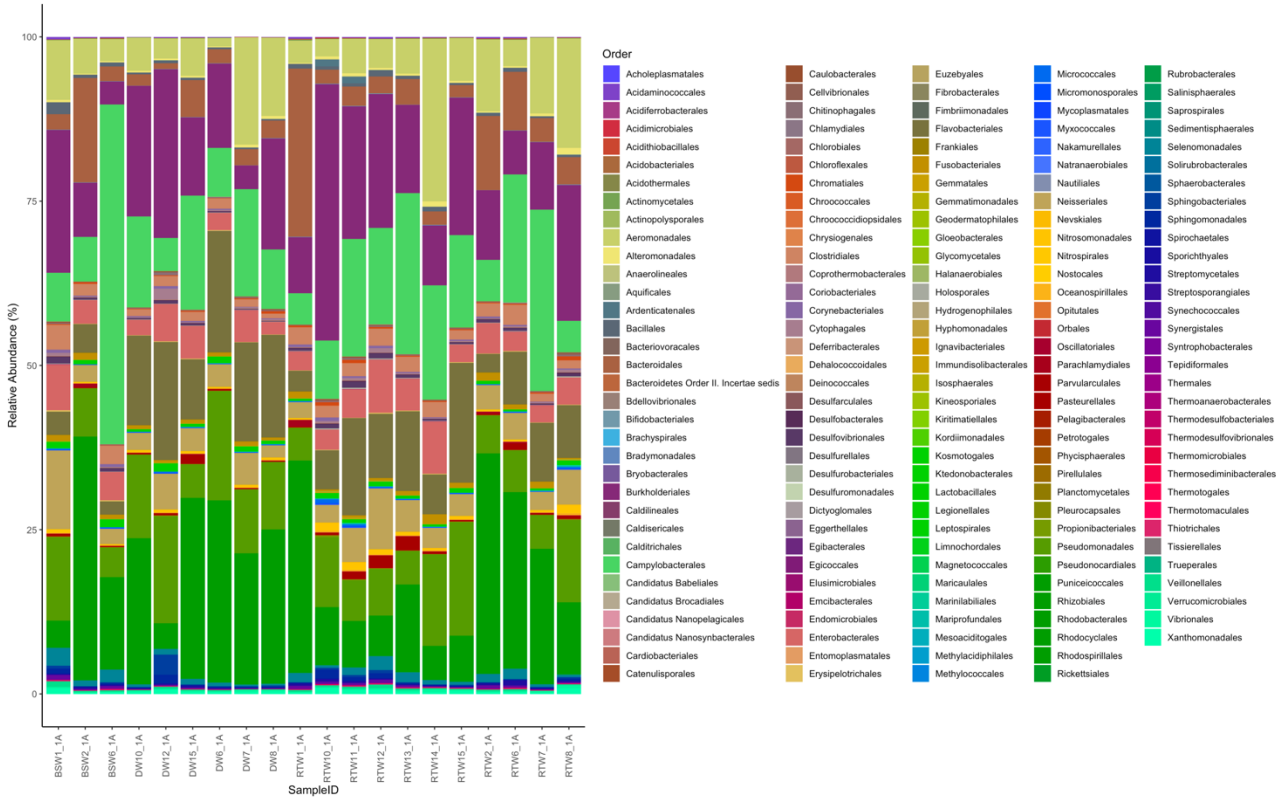


Figure 3-13: Relative abundance, as indicated by the percentage of reads, for Bacterial orders. A total of 74 orders were detected with 4 of these classified as *Candidatus.* Each colour is representative of an order. Differences in bacterial composition across samples are clear.

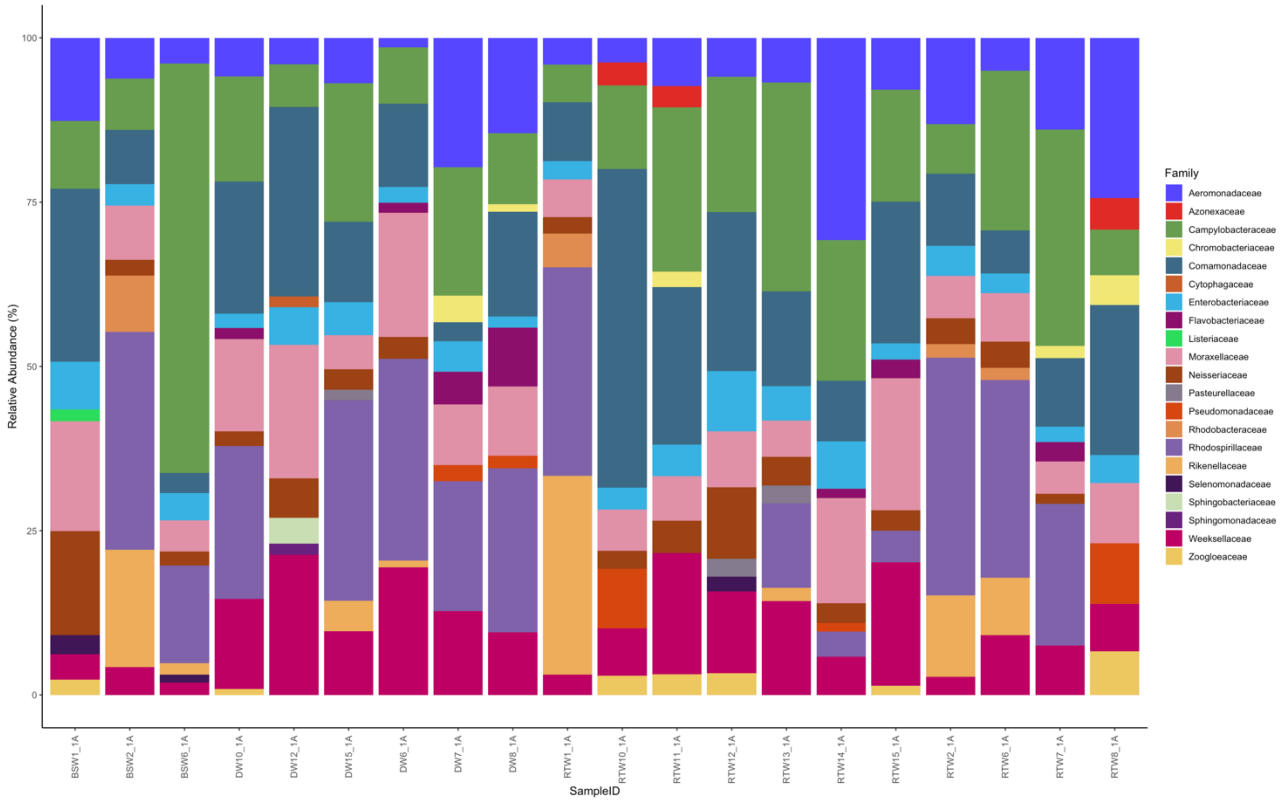


Figure 3-14: Relative abundance, as indicated by the percentage of reads, for the top 10 Bacterial families. Each colour is representative of a top 10 family. The high levels of diversity with regards to the Bacterial families are clearly evident.

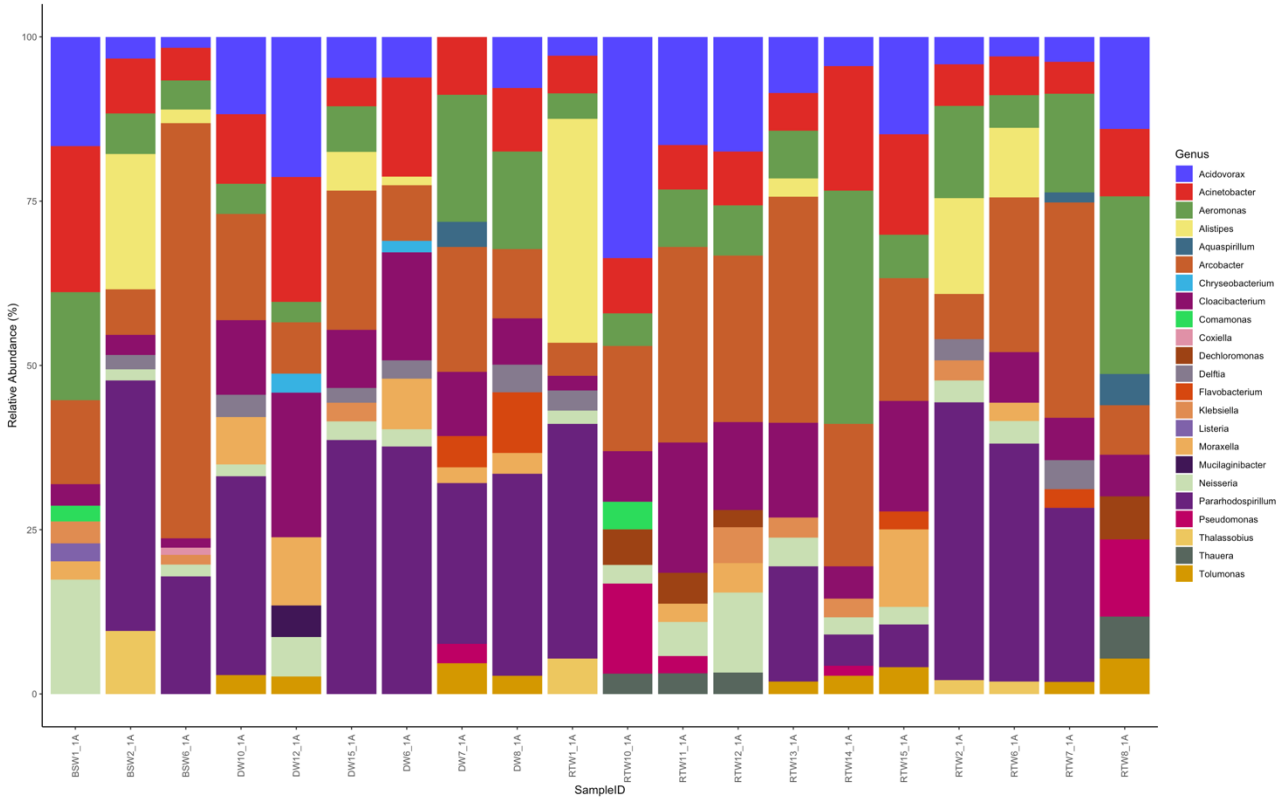


Figure 3-15: Relative abundance, as indicated by the percentage of reads, for the top 10 Bacterial genera. Each colour is representative of a top 10 genus. The high levels of diversity with regards to the Bacterial genera are clearly evident.

Table 3-5: Combined set of top 10 Bacterial genera per sample detected across all the samples.

Phylum	Class	Order	Family	Genus
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas
Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Aquaspirillum
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Arcobacter
Bacteroidetes	Flavobacteriia	Flavobacteriales	Weeksellaceae	Chryseobacterium
Bacteroidetes	Flavobacteriia	Flavobacteriales	Weeksellaceae	Cloacibacterium
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonas
Proteobacteria	Gammaproteobacteria	Legionellales	Coxiellaceae	Coxiella
Proteobacteria	Betaproteobacteria	Rhodocyclales	Azonexaceae	Dechloromonas
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia
Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Flavobacterium
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Klebsiella
Firmicutes	Bacilli	Bacillales	Listeriaceae	Listeria
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Moraxella
Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	Mucilaginibacter
Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria
Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Pararhodospirillum
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Thalassobius
Proteobacteria	Betaproteobacteria	Rhodocyclales	Zoogloeaceae	Thauera
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Tolumonas

The International Committee on Taxonomy of Viruses (ICTV) classification system was used for Viruses. Viral classification for Realm, Kingdom and Phylum was inspected and high levels of viral diversity for the different taxonomic ranks detected (Figure 3-16, Figure 3-17 and Figure 3-18). A total of 11 realms, 38 kingdoms and 53 phyla were detected and including various “unclassified”. The most abundant Viral classes were found to be *Caudoviricetes* and *Allassoviricetes* with the Viral orders being *Caudovirales* and *Levivirales*. Per sample Viral taxonomic classification is further available from the Supplementary Material.

The high resolution offered by metagenomics and in particularly RNA metagenomics in viral identification is clearly illustrated here. It is generally very difficult to isolate viruses from a sample and in particularly from samples with high viral diversity. By using metagenomics approaches a large portion of the virome is accurately identified without the need for per virus tests. This is critical in monitoring human health and possible outbreaks of infection. Early detection of viral infection as presented in wastewater will greatly aid relevant parties and allow for the rapid intervention.

Viruses are not only important in human health but have recently been proposed as a human-specific microbial source tracking (MST) marker. This form of marker detection is used to identify specific sources of faecal contamination and has the ability to differentiate between human, animal and bird origin. Metagenomic sequencing may therefore have an alternative application in wastewater samples and possibly enable the detection of sources of contamination.

As illustrated here and in other sections, metagenomic sequencing affords a wealth of information in a single data generation event. The data generated is then available to be scrutinised for various questions of interest and research endeavours.

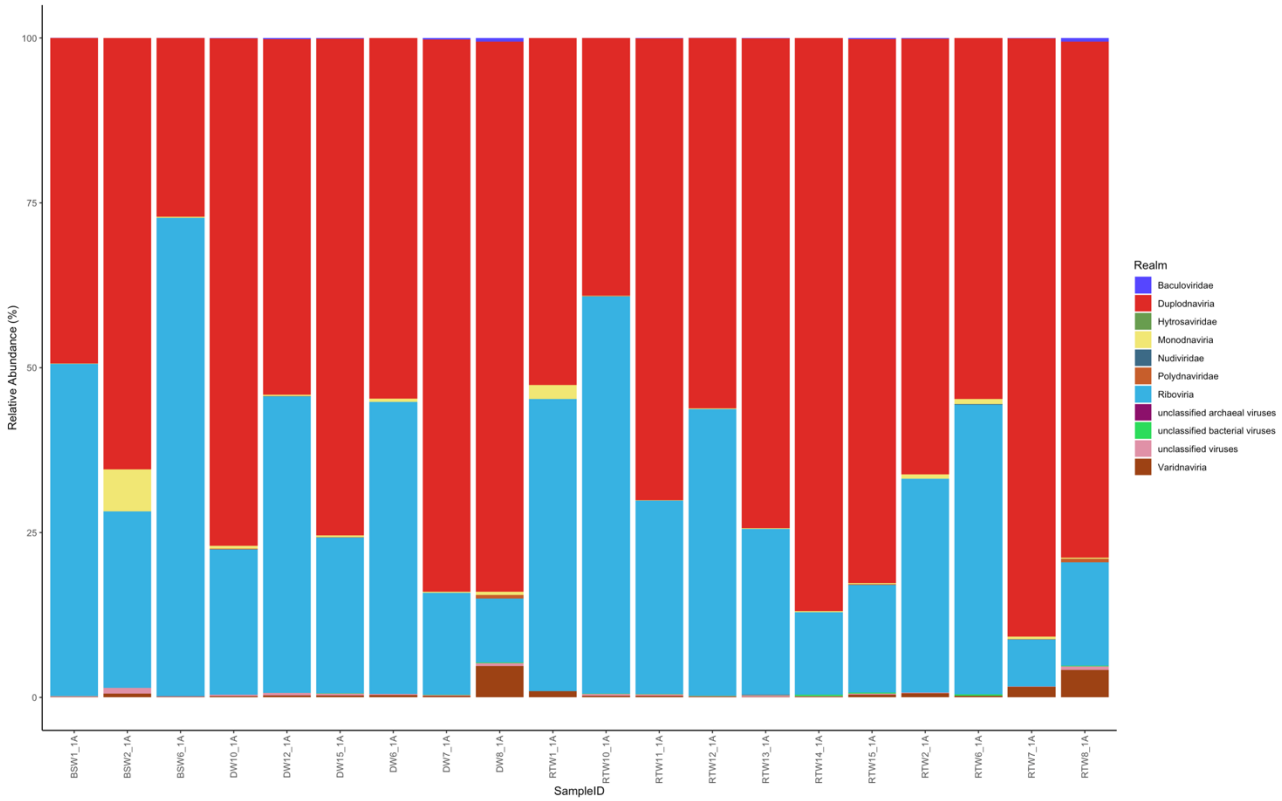


Figure 3-16: Relative abundance, as indicated by the percentage of reads, for the Viral realm classification. Each colour is representative of a viral realm. *Duplodnaviria* and *Riboviria* were found to be in high abundance. Certain samples further indicated a high abundance of *Varidnaviria* and *Monodnaviria*.

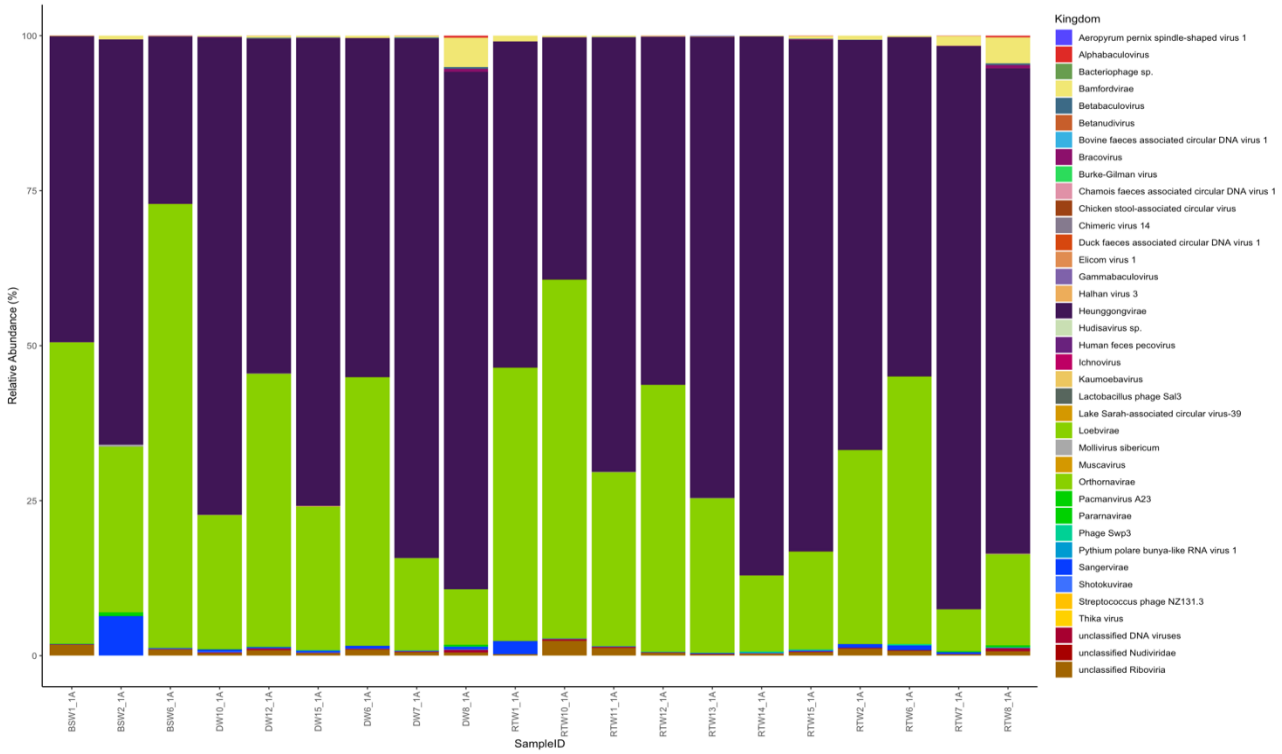


Figure 3-17: Relative abundance, as indicated by the percentage of reads, for the Viral kingdom classification. Each colour is representative of a viral kingdom. *Heunggongvirae* and *Orthornavirae* were found to be in high abundance.

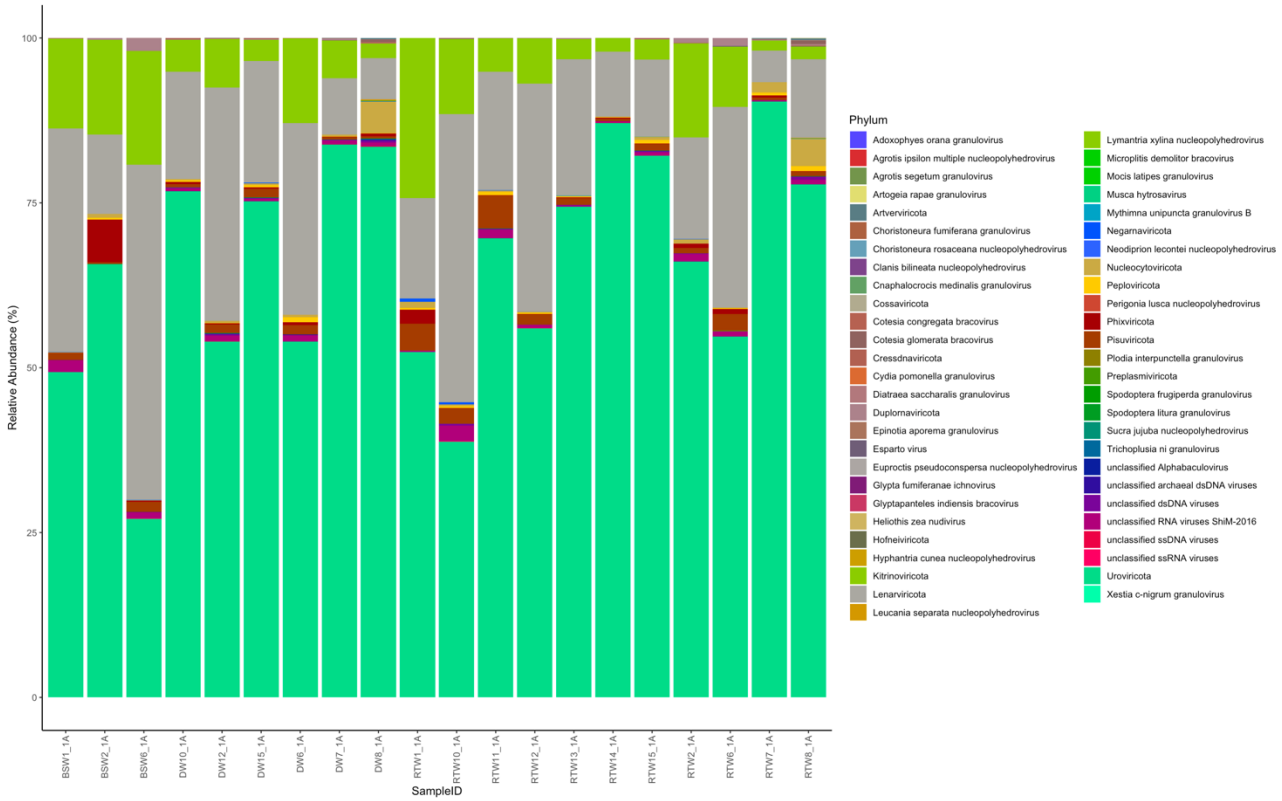


Figure 3-18: Relative abundance, as indicated by the percentage of reads, for the Viral phylum classification. Each colour is representative of a viral phylum. *Uroviricota* and *Lenarviricota* were found to be in high abundance.

3.3.4 AMR profile of samples based on unassembled sequencing data

The AMR profile of each sample was determined by aligning the quality filtered, decontaminated reads against CARD and filtering for results with at least 80% coverage of the reference AMR sequence. Based on the filtering criteria, 3 samples (BSW2_1A, RTW1_1A and RTW2_1A) were found to be void of any AMRs.

A total of 103 unique Antibiotic Resistance Ontology (AROs) were detected across 17 samples and are displayed in Figure 3-19. The number of unique and shared AROs per sample are presented in Figure 3-20. In general, the Rietgat (RTW) samples displayed higher levels of ARO frequency. Sample RTW8_1A contains 27 unique AROs and sample RTW11_1A 5 unique AROs. These samples further share 12 AROs not found in any of the other samples.

The AROs were further classified into AMR Gene Families and Drug Classes. There was a total of 39 AMR Gene Families (Figure 3-21 and Figure 3-22) and 39 AMR Drug Classes (Figure 3-23 and Figure 3-24) detected across 17 samples. Certain RTW samples, RTW8_1A and RTW11_1A, dominated the AMR Gene Family and Drug Class frequencies. RTW8_1A further had various unique AMR Gene Families and Drug Classes not found in the other samples and together RTW11_1A shared various AMR Gene Families and Drug Classes not found in any of the other samples. Each ARO further has an AMR Resistance Mechanism associated with it. In total, 7 AMR Resistance Mechanisms were found across the 17 samples (Figure 3-25 and Figure 3-26). The AMR Resistance Mechanisms were also variable across the samples and further emphasizes the high AMR diversity contained within the wastewater samples. The added benefit and ability of metagenomic sequencing to detect AMR potential within samples, in addition to the taxonomic classification, is clearly highlighted by the figures and information presented below.

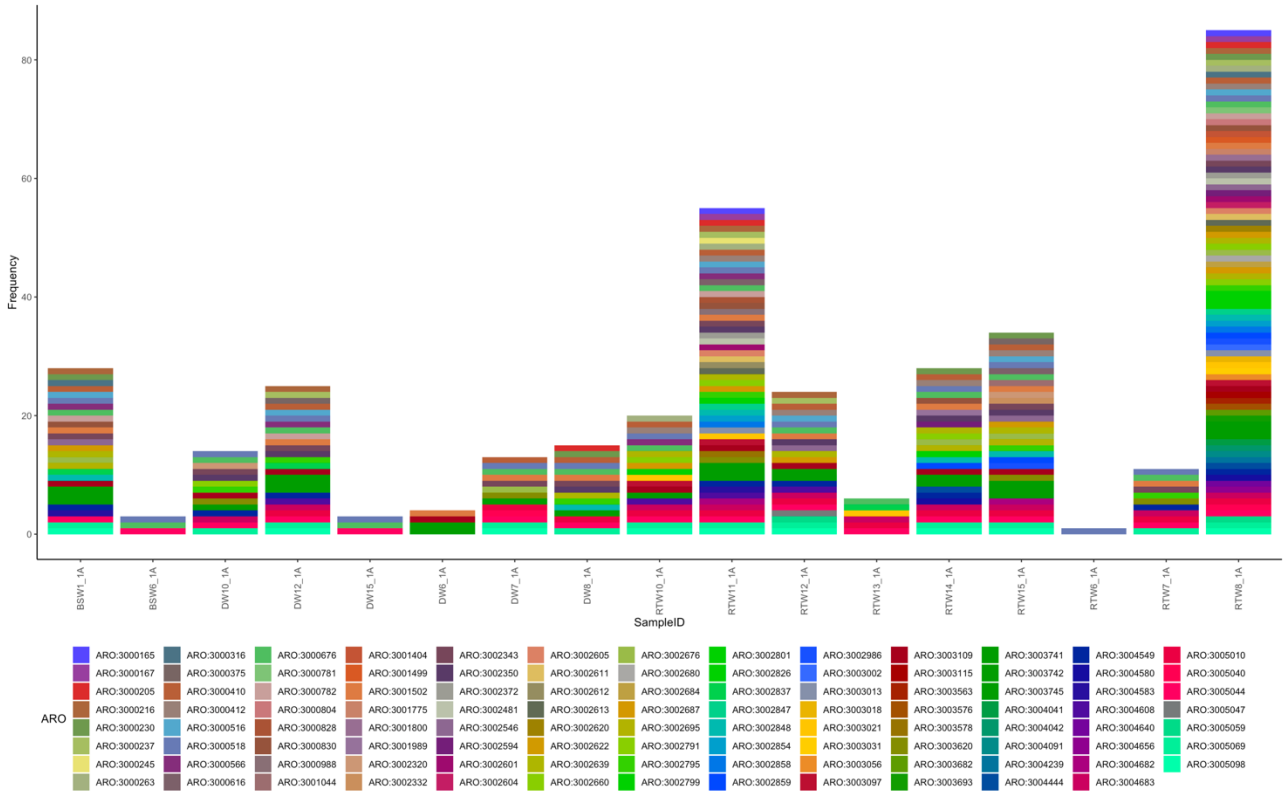


Figure 3-19: Distribution of AROs across each sample. Each colour is representative of an ARO. The samples all displayed different ARO diversity and quantity.

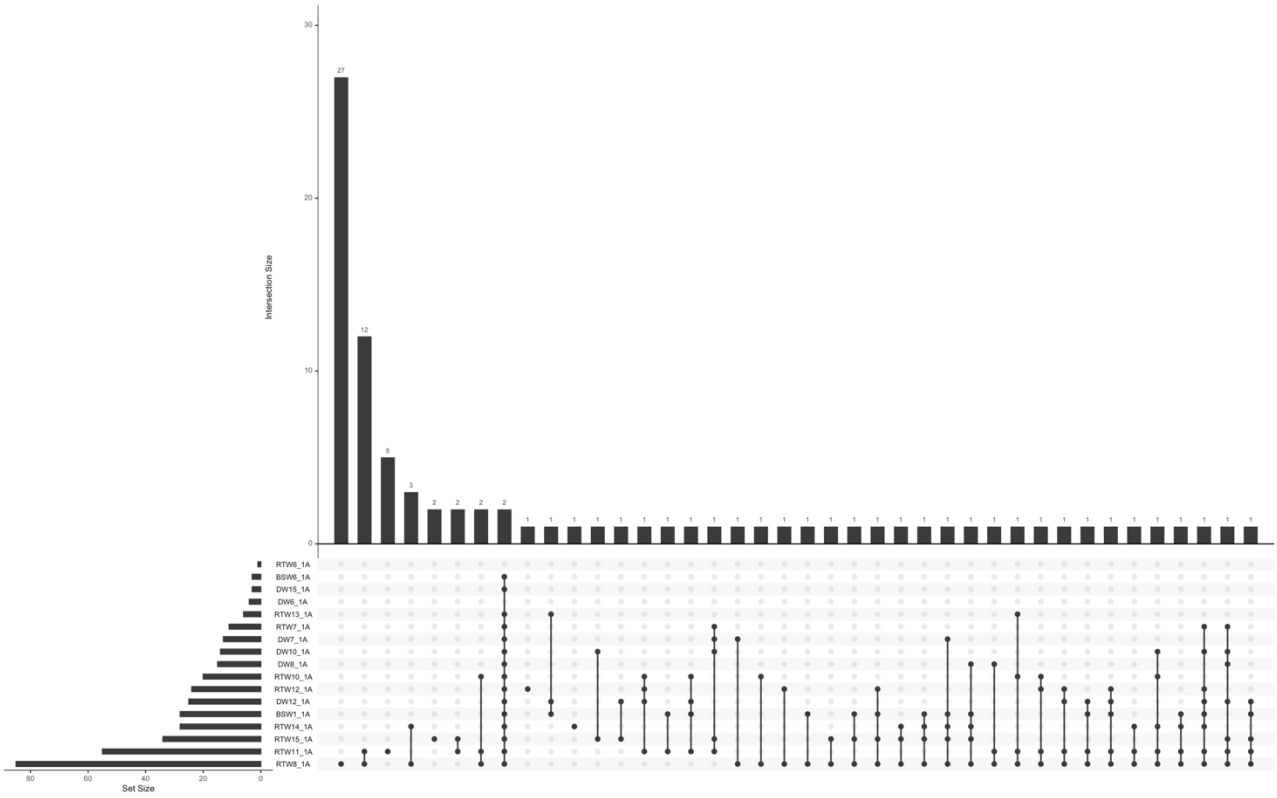


Figure 3-20: Unique and shared AROs across each sample. The bars on the left indicate the number of AROs per sample. The grid in the middle indicates which samples share a set of AROs and the bars on the top represent the size of the shared set.

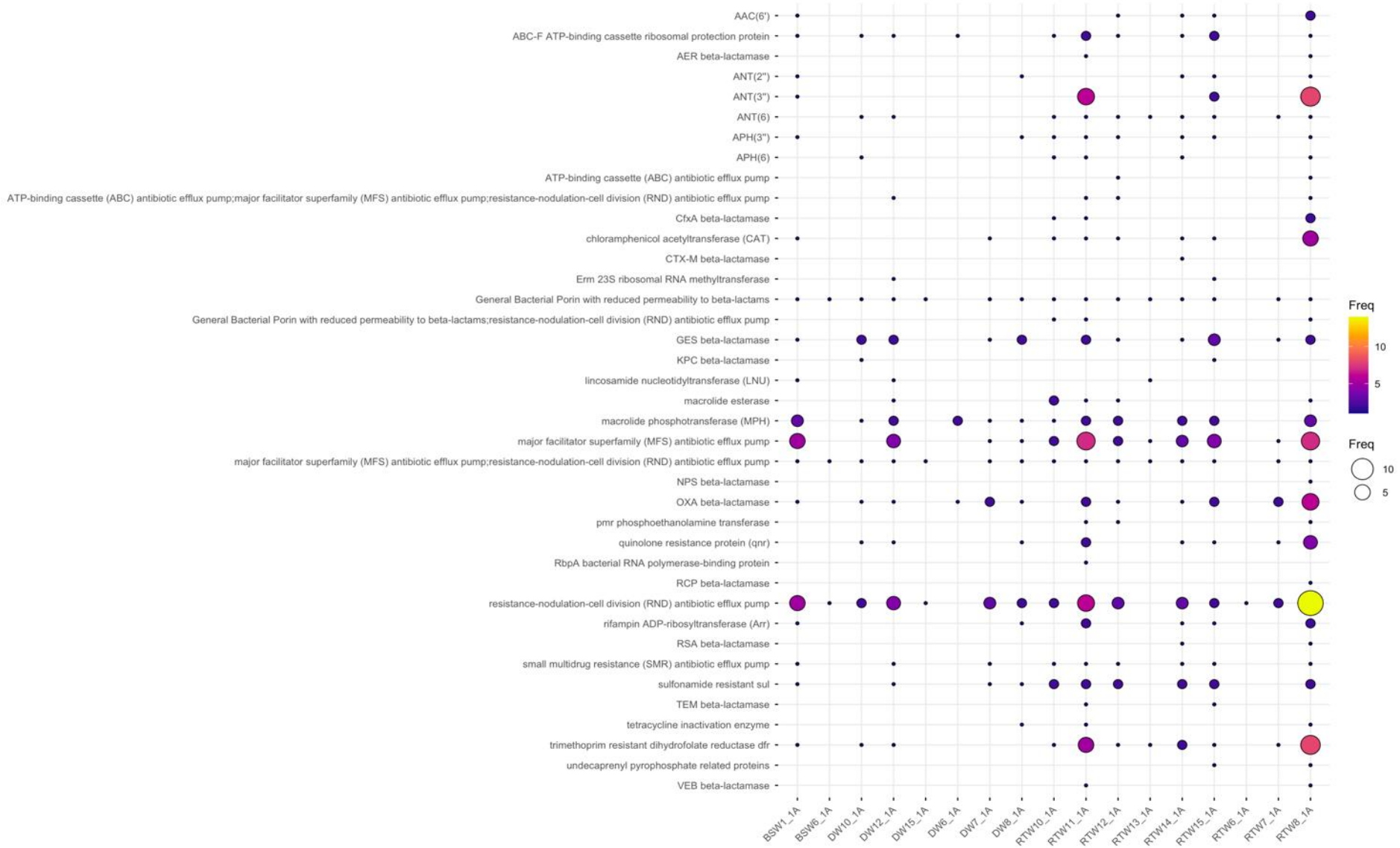


Figure 3-21: Distribution of AMR Gene Families across each sample. The colour and size of each circle represent the frequency of the particular AMR Gene Family which is indicated on the y-axis. The samples are on the x-axis.

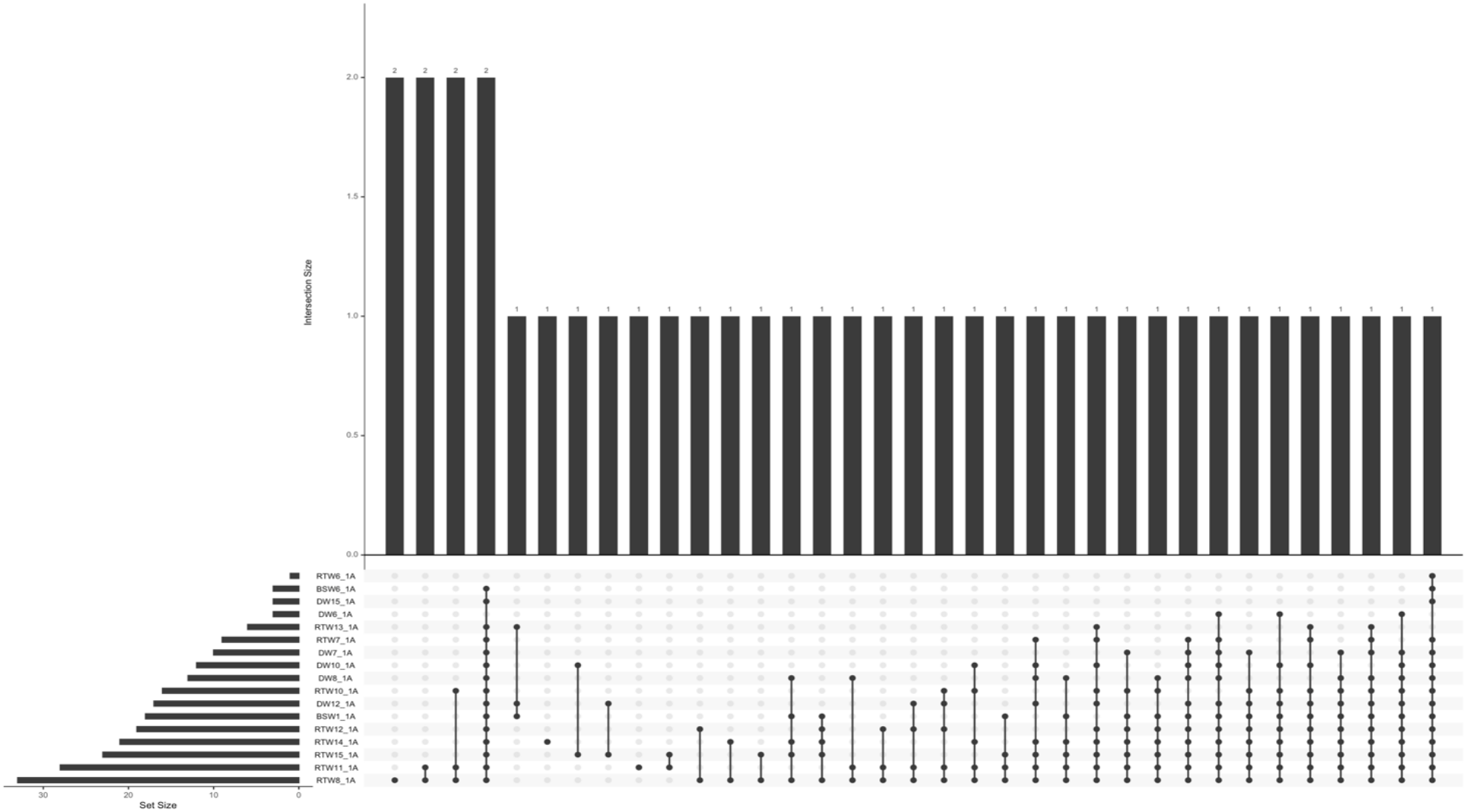


Figure 3-22: Unique and shared AMR Gene Families across each sample. The bars on the left indicate the number of AMR Gene Families per sample. The grid in the middle indicates which samples share a set of AMR Gene Families and the bars on the top represent the size of the shared set.

Figure 3-23: Distribution of AMR Drug Classes across each sample. The colour and size of each circle represent the frequency of the particular AMR Drug Classes which is indicated on the y-axis. The samples are on the x-axis.

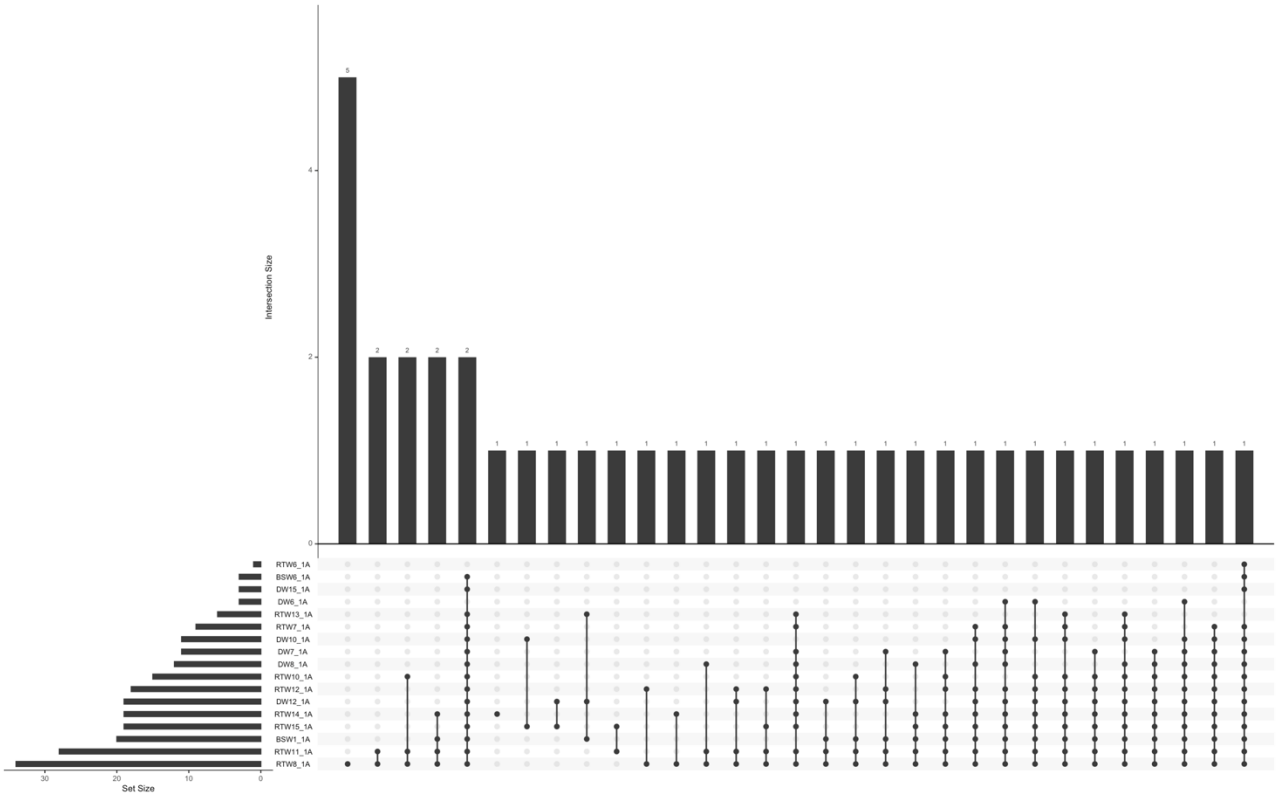


Figure 3-24: Unique and shared AMR Drug Classes across each sample. The bars on the left indicate the number of AMR Drug Classes per sample. The grid in the middle indicates which samples share a set of AMR Drug Classes and the bars on the top represent the size of the shared set.

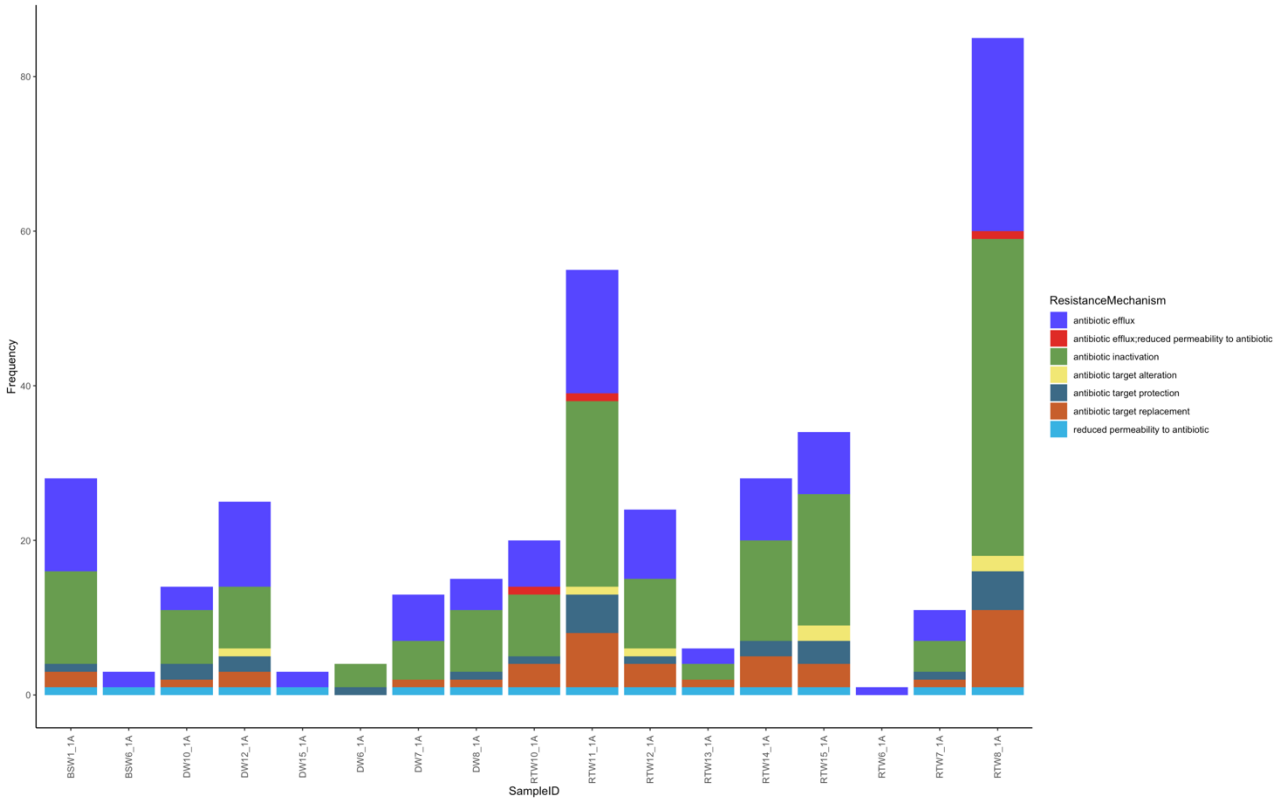


Figure 3-25: Distribution of AMR Resistance Mechanisms across each sample. The colours represent a particular AMR Resistance Mechanisms and the frequency of the AMR Resistance Mechanisms is indicated on the y-axis. The samples are on the x-axis.

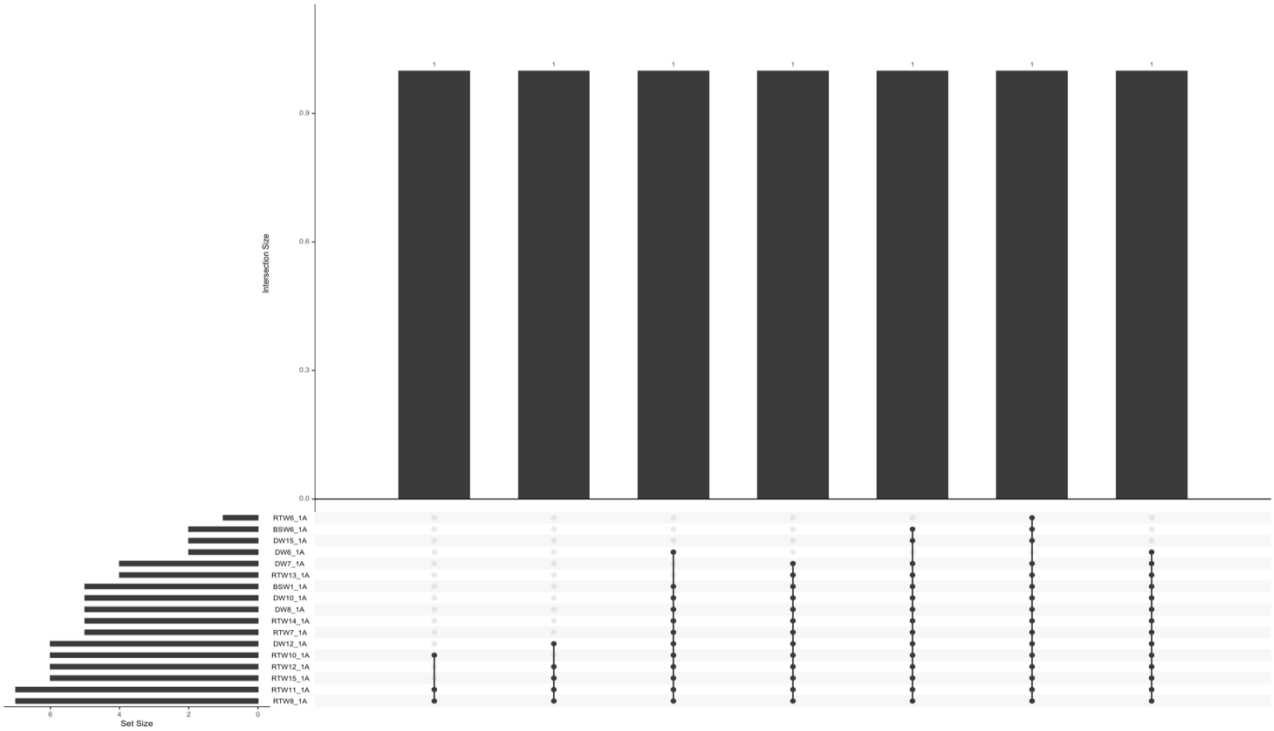


Figure 3-26: Unique and shared AMR Resistance Mechanisms across each sample. The bars on the left indicate the number of AMR Resistance Mechanisms per sample. The grid in the middle indicates which samples share a set of AMR Resistance Mechanisms and the bars on the top represent the size of the shared set.

3.3.5 Taxonomic profile of samples based on *de novo* assembled transcripts

Quality filtered, decontaminated reads were *de novo* assembled and filtered for long reliable transcripts with high expression. A total of 715,307 reliable transcripts were identified across all 20 samples (Figure 3-27). These transcripts were analysed for taxonomic and AMR profiles using a similar approach as above. Some discrepancies may exist between the approaches and this can be attributed to various factors such as database used, format of the query, i.e. paired-end read or transcript, and software application used. Furthermore, the *de novo* assembled transcripts were subjected to another round of filtering and quality control. That being said, there is a large overlap between the results of the methodologies and the trends found within the results. The *de novo* assembled transcripts represent the actively expressed portions of genomes as found within the wastewater samples. Relatively high numbers of transcripts were again found among the Rietgat (RTW) samples. This may be due to various factors and should be investigated further. One speculation may be that there is an increase or higher levels of activity within the RTW samples.

The data available allowed us to investigate possible reasons for the variation in transcript numbers. One possible explanation for this may be the amount of data generated per samples. Testing of Spearman correlation between the number of paired-end reads and number of transcripts indicate a significant negative correlation (p-value = 0.0114) (Figure 3-28 a)). This result indicates that the amount of sequencing data produced was not an influencing factor with regards to the variability in number of transcripts for each sample. The 7-Day average COVID-19 cases was thereafter used to test the variation in number of transcripts. There was a positive correlation, albeit not significant, between the number of transcripts and the 7-Day average COVID-19 cases (Figure 3-28 b)). This may indicate an increase in the activity found within a sample due to the presence of SARS-CoV-2. This is purely speculative and will be investigated further.

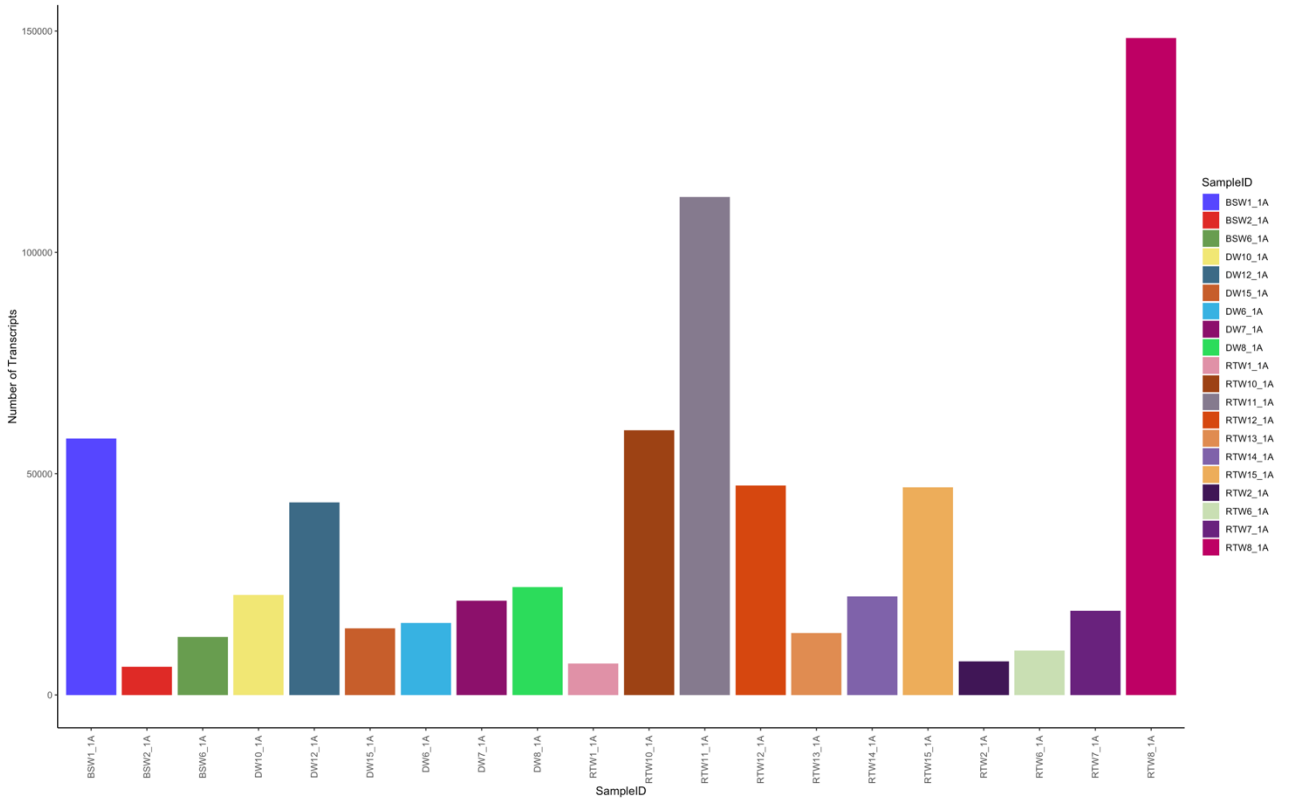


Figure 3-27: Number of transcripts per sample. The samples are represented on the x-axis and coloured according to sample. The number of transcripts are indicated on the y-axis.

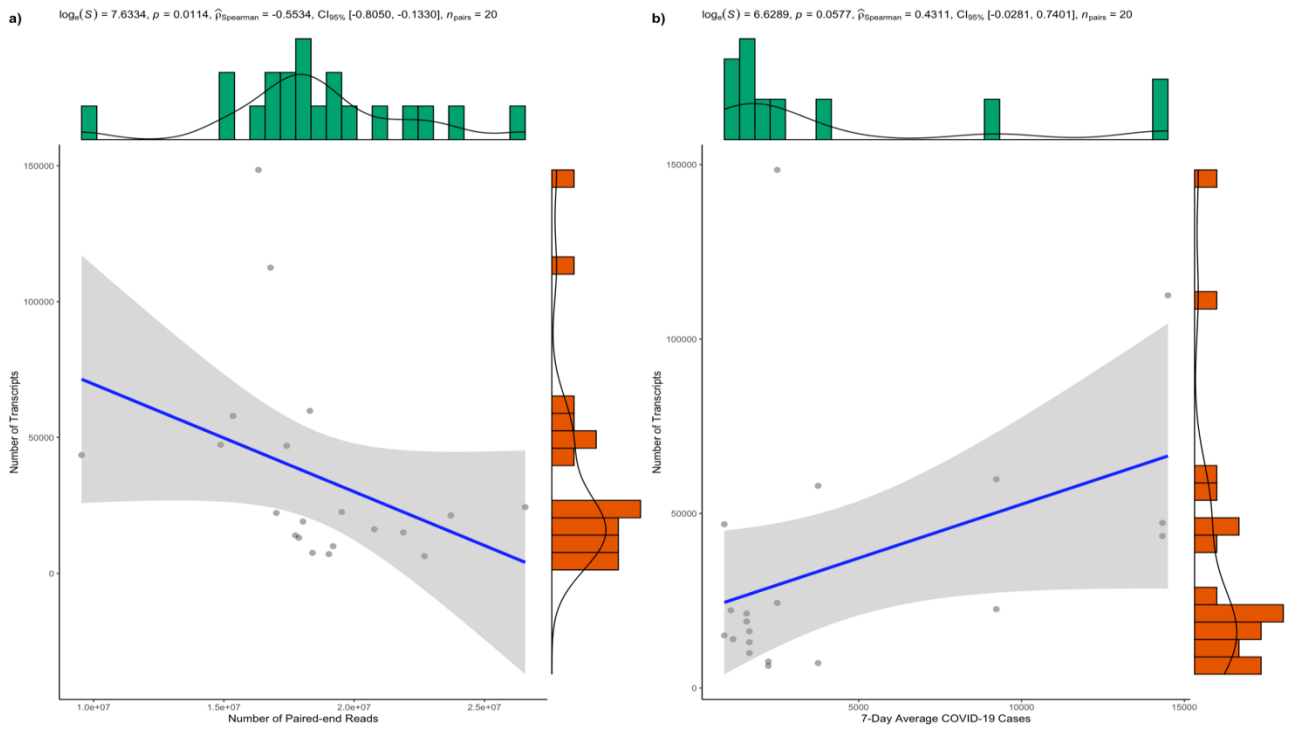


Figure 3-28: Spearman's rank correlation coefficient test results for a) correlation between the number of transcripts and amount of data generated and b) correlation between the number of transcripts and 7-Day average COVID-19 cases. The results from the statistical test are reported in the subtitles on the top of each graph. The marginal distributions for the x and y variables are overlaid on the axes of each graph.

Transcripts were aligned against the NCBI nt database. This nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. The genome, gene and transcript sequence data provided within are a foundation for research and discovery. After filtering, 390,615 alignment hits against the NCBI nt database were found. This was divided into 385,953 Bacterial, 1,601 Archaeal and 3,061 Viral hits of origin. The Archaeal classifications most frequently detected across all samples are described in Table 3-6. A total of 125 different Archaeal classifications were found. The Archaeal portion of transcripts indicated relatively high levels in some of the Rietgat (RTW) samples (Figure 3-29**)**.

Table 3-6: Most frequently detected Archaeal classifications across all samples.

Classification	Frequency
Methanobrevibacter smithii	268
uncultured archaeon	233
Methanospirillum hungatei	118
Methanospirillum sp. J.3.6.1-F.2.7.3	86
Methanospirillum hungatei JF-1	70
Methanothrix soehngenii GP6	60
Methanobrevibacter smithii ATCC 35061	49
Methanobacterium formicicum	44
Methanoregula formicica	44
Methanoregula boonei	31
Methanomassiliicoccaceae archaeon DOK	29
Methanothrix soehngenii	27
uncultured euryarchaeote	25
Methanobrevibacter arboriphilus	22
Methanomassiliicoccales archaeon	20
Methanoregula formicica SMSP	20
Methanobacterium sp. BAmetb5	19
Methanoregula boonei 6A8	18
Methanosphaera stadtmanae	18
Methanomethylovorans hollandica DSM 15978	17
Candidatus Diapherotrites archaeon	16

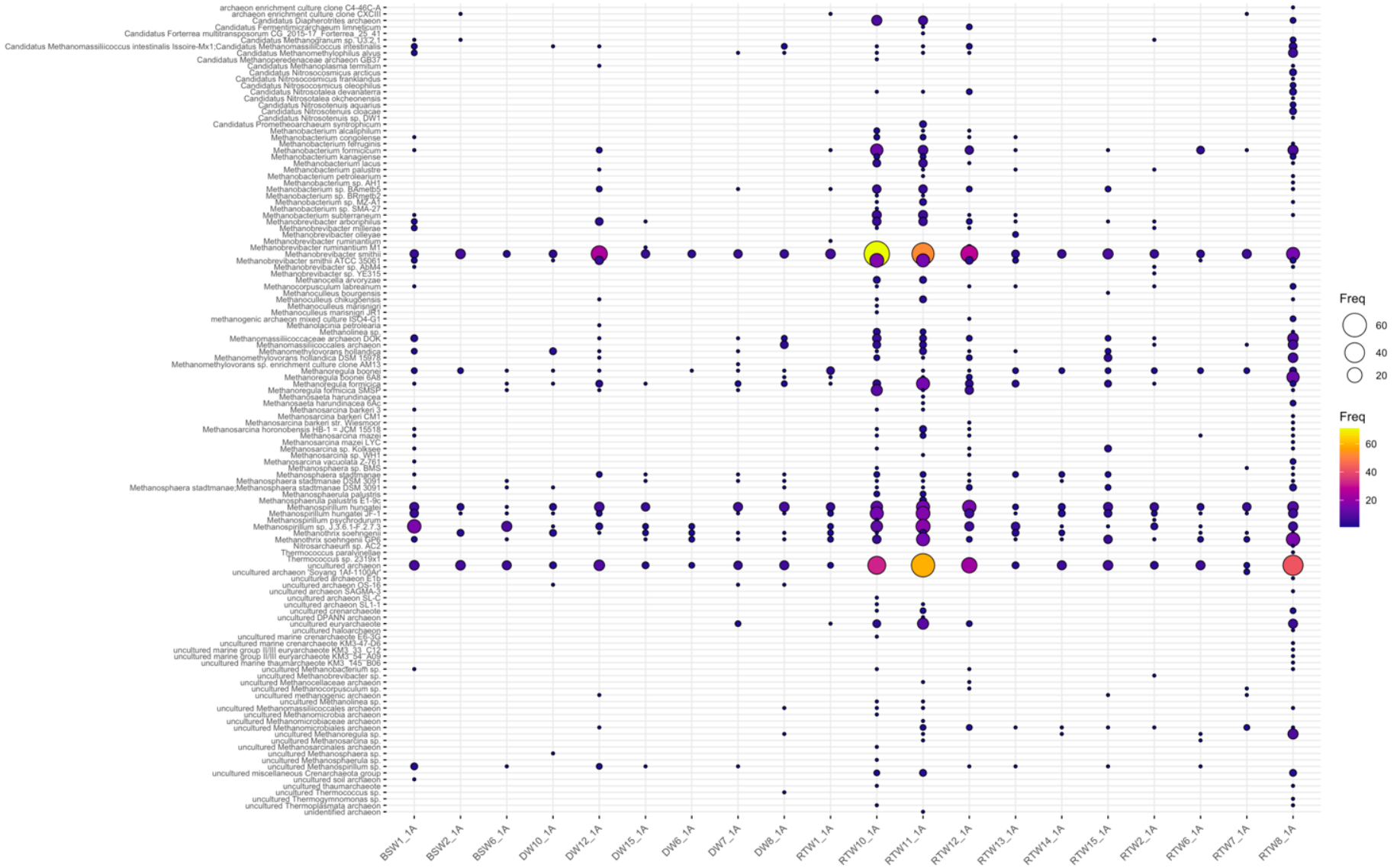


Figure 3-29: Distribution of Archaeal classifications across each sample. The colour and size of each circle represent the frequency of the particular Archaea which is indicated on the y-axis. The samples are on the x-axis.

A total of 7,355 different Bacterial classifications were detected and the most frequently detected across all samples are described in Table 3-7. Samples from Rietgat again displayed a high frequency of detected Bacterial classifications (Figure 3-30). The bacterial classifications included more than 25,000 “uncultivated bacteria”. These are of interest as they have not yet been cultivated but may be crucial in human health. Further investigation and phylogenetic analysis will be required to adequately classify these transcripts.

Table 3-7: Most frequently detected Bacterial classifications across all samples.

Classification	Frequency
Aliarcobacter cryaerophilus	17,854
Cloacibacterium caeni	16,051
Acidovorax sp. 1608163	12,620
Aeromonas caviae	12,424
Cloacibacterium normanense	10,268
Moraxella osloensis	8,561
Thauera sp. MZ1T	6,237
Tolumonas auensis DSM 9187	6,073
Acinetobacter johnsonii	5,551
Aliarcobacter cryaerophilus D2610	4,706
Prevotella copri	4,520
Acidovorax sp. HDW3	4,200
Acidovorax sp. KKS102	4,103
Aquaspirillum sp. LM1	3,960
Acinetobacter baumannii	3,853
Alicycliphilus denitrificans	3,769
Acinetobacter towneri	3,524
Dechloromonas sp.	3,393
Pseudomonas alcaligenes	3,121
Acinetobacter sp. NEB 394	3,102
Sphaerotilus natans subsp. sulfidivorans	3,009
Acidovorax carolinensis	2,759
Klebsiella pneumoniae	2,313

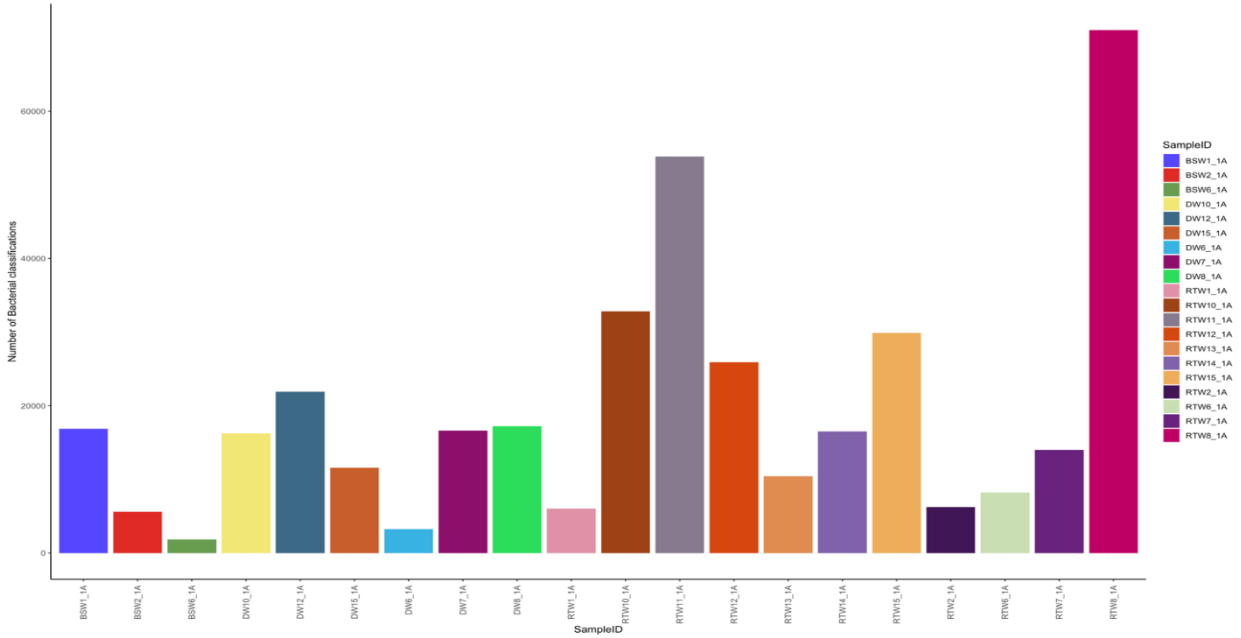


Figure 3-30: Number of Bacterial classifications per sample.

The highest occurring Viral classifications are described in Table 3-8 and the viral portion included 180 different classifications are shown in Figure 3-31. The viral abundance was heavy and included diverse annotations. One of these was a CrAss-like virus sp. These have been only recently been identified and are dominant viruses in the human gut virome. CrAssphages have been proposed as a human-specific MST marker and the application is currently under development. Numerous “uncultured human fecal virus” classifications were also present and demonstrates the long road ahead to fully characterize the human gut microbiome and virome.

Table 3-8: Most frequently detected Viral classifications across all samples.

Classification	Frequency
Siphoviridae sp.	774
Myoviridae sp.	452
uncultured human fecal virus	318
Bacteriophage sp.	120
Podoviridae sp.	116
ssRNA phage SRR7976325_7	112
ssRNA phage SRR5466369_1	100
Escherichia virus Qbeta	56
Pepper mild mottle virus	56
Tobacco mosaic virus	53
ssRNA phage SRR6960507_10	51
ssRNA phage SRR5466365_2	44
ssRNA phage SRR5466727_4	41
Tobacco mild green mosaic virus	36
Escherichia virus BZ13	35
Leviviridae sp.	30
ssRNA phage SRR6960799_15	29
Tomato mosaic virus	25
Herelleviridae sp.	23
CrAss-like virus sp.	20
ssRNA phage SRR7976326_4	20
Microviridae sp. ctOX110	18

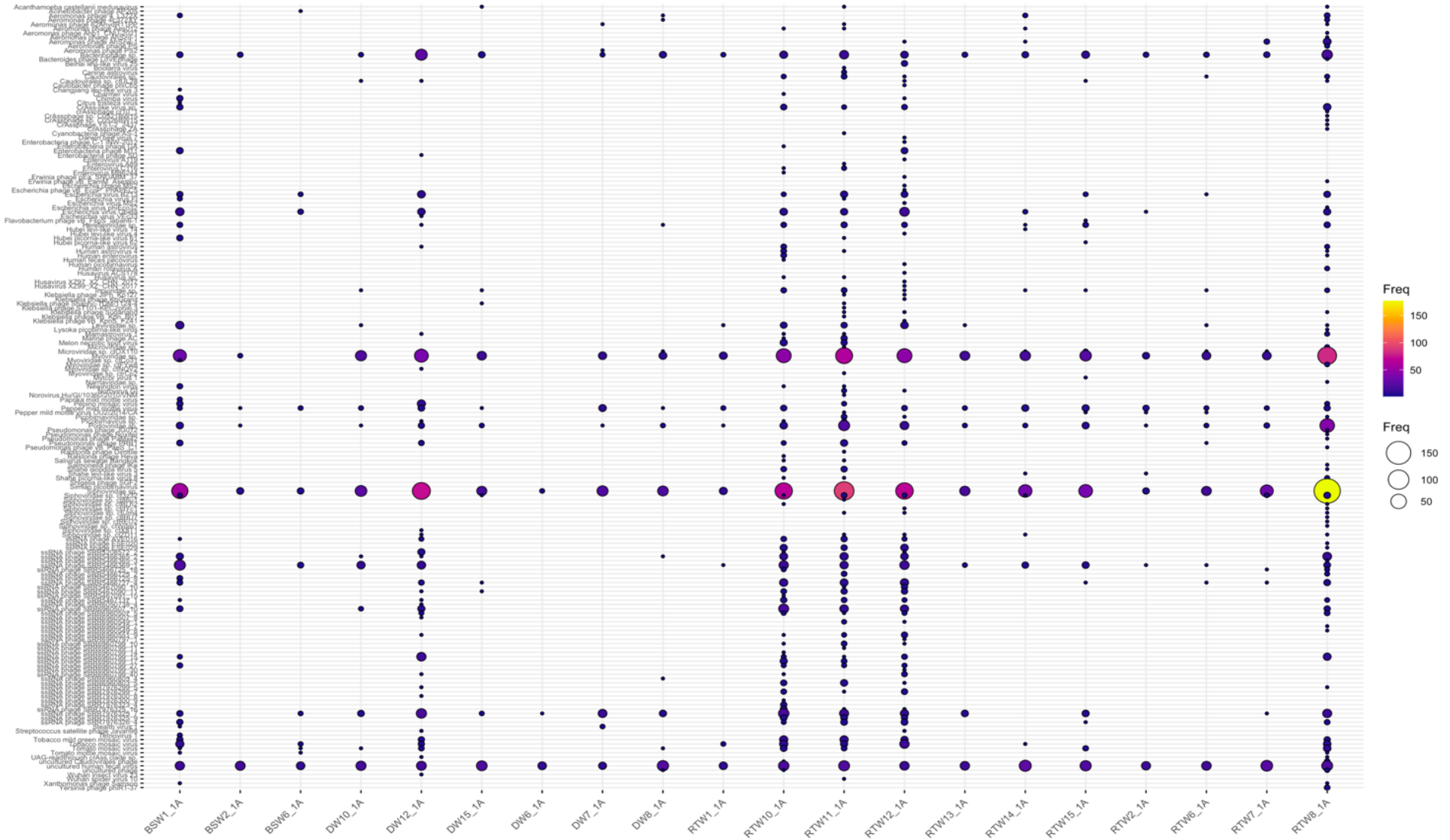


Figure 3-31: Distribution of Viral classifications across each sample.

3.3.6 AMR profile of samples based on *de novo* assembled transcripts

The AMR profile of each sample was determined by aligning the quality filtered transcripts against CARD and filtering for results with at least 80% identity over at least 80% of the reference AMR sequence. Based on the filtering criteria, 7 samples were found to be void of any AMRs. The reason for the discrepancy with the previous AMR results where the reads were aligned may be due to the added filtering and *de novo* assembly of transcripts. A total of 52 unique Antibiotic Resistance Ontology (AROs) were detected across 13 samples and are displayed in Figure 3-32.

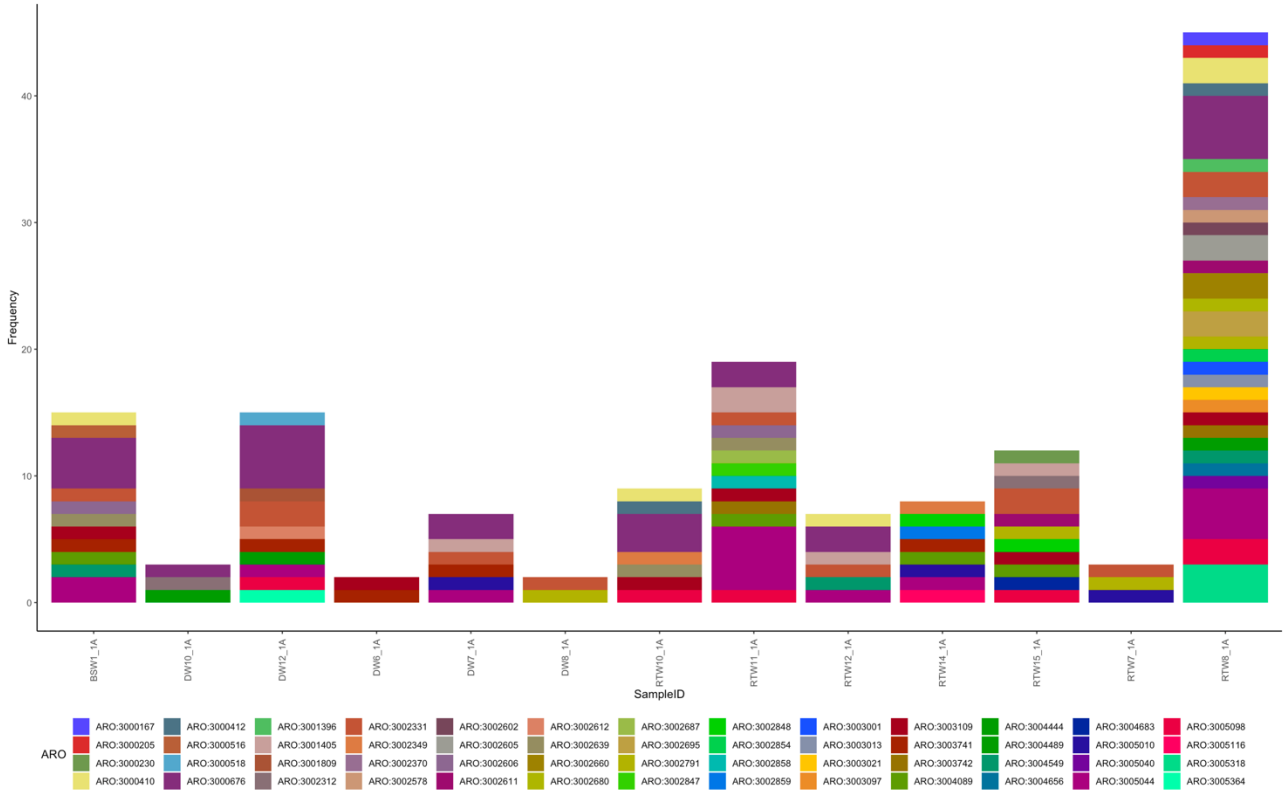


Figure 3-32: Distribution of AROs across each sample. Each colour is representative of an ARO. The samples all displayed different ARO diversity and quantity. In general, the Rietgat (RTW) samples displayed higher levels of ARO frequency.

The number of unique and shared AROs per sample are presented in Figure 3-33. The high levels of AROs in the Rietgat samples are again evident. Sample RTW8_1A contains 19 unique AROs and sample RTW11_1A 3 unique AROs. The AROs were further classified into AMR Gene Families and Drug Classes. There was a total of 26 AMR Gene Families (Figure 3-34 and Figure 3-35) and 22 AMR Drug Classes (Figure 3-36 and Figure 3-37) detected across 13 samples. Each ARO further has an AMR Resistance Mechanism associated with it. In total, 5 AMR Resistance Mechanisms were found across the 13 samples (Figure 3-38 and Figure 3-39). The results presented below corroborate the methodology used earlier and again indicates a high occurrence of AMR potential in RTW samples relative to the other sampling locations.

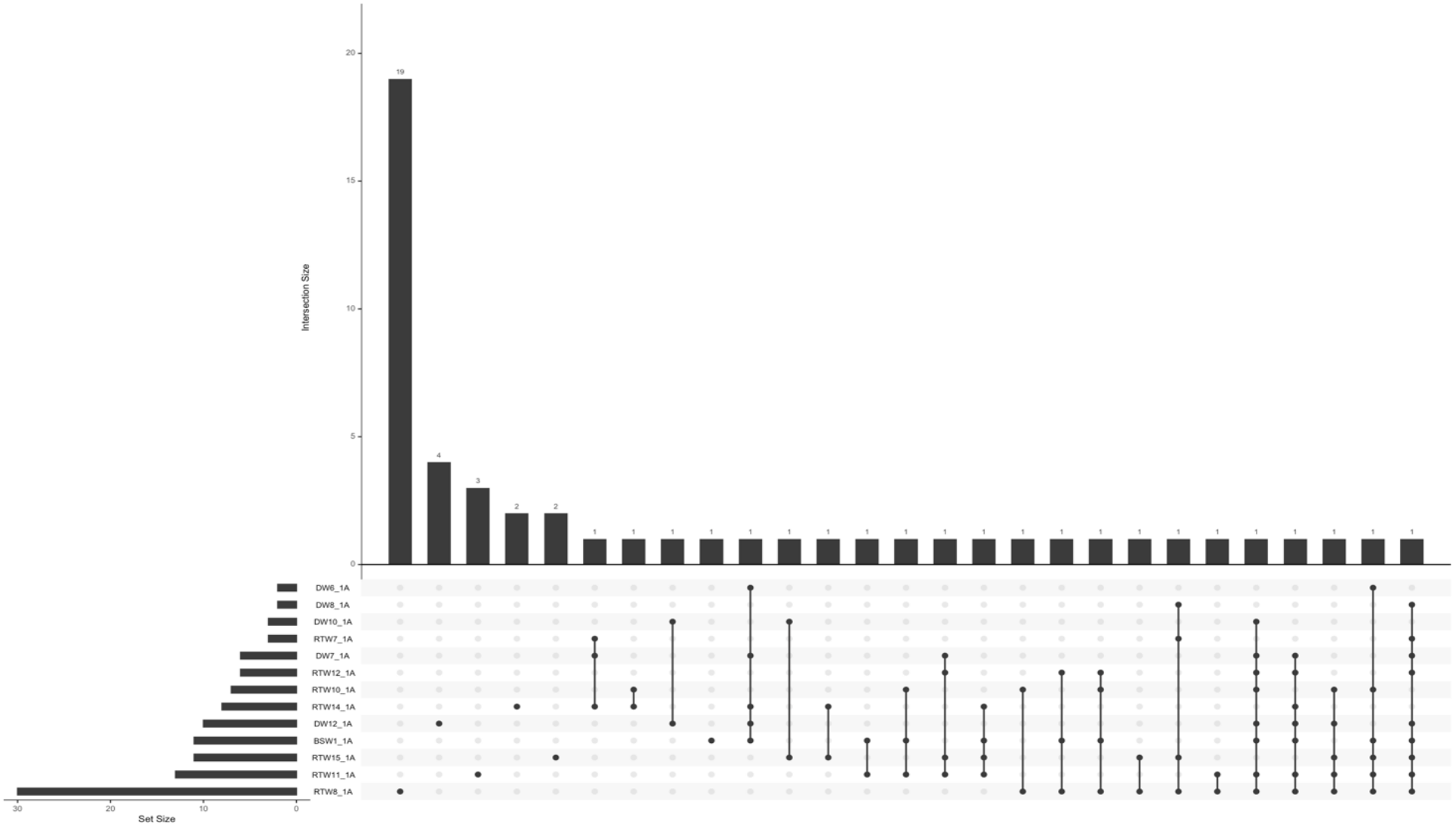


Figure 3-33: Unique and shared AROs across each sample. The bars on the left indicate the number of AROs per sample. The grid in the middle indicates which samples share a set of AROs and the bars on the top represent the size of the shared set.

Figure 3-34: Distribution of AMR Gene Families across each sample. The colour and size of each circle represent the frequency of the particular AMR Gene Family which is indicated on the y-axis. The samples are on the x-axis.

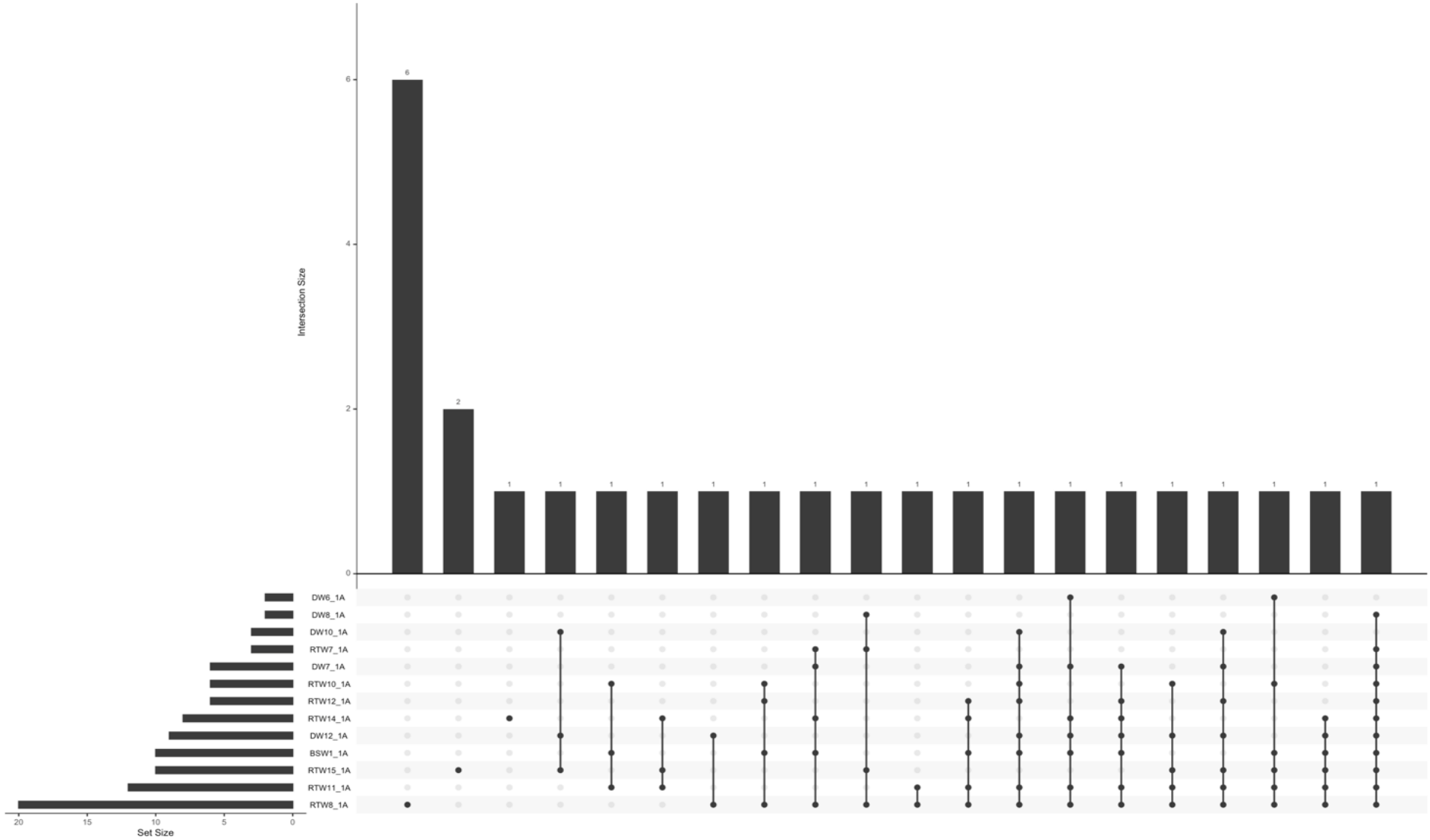


Figure 3-35: Unique and shared AMR Gene Families across each sample.

Figure 3-36: Distribution of AMR Drug Classes across each sample. The colour and size of each circle represent the frequency of the particular AMR Drug Classes which is indicated on the y-axis. The samples are on the x-axis.

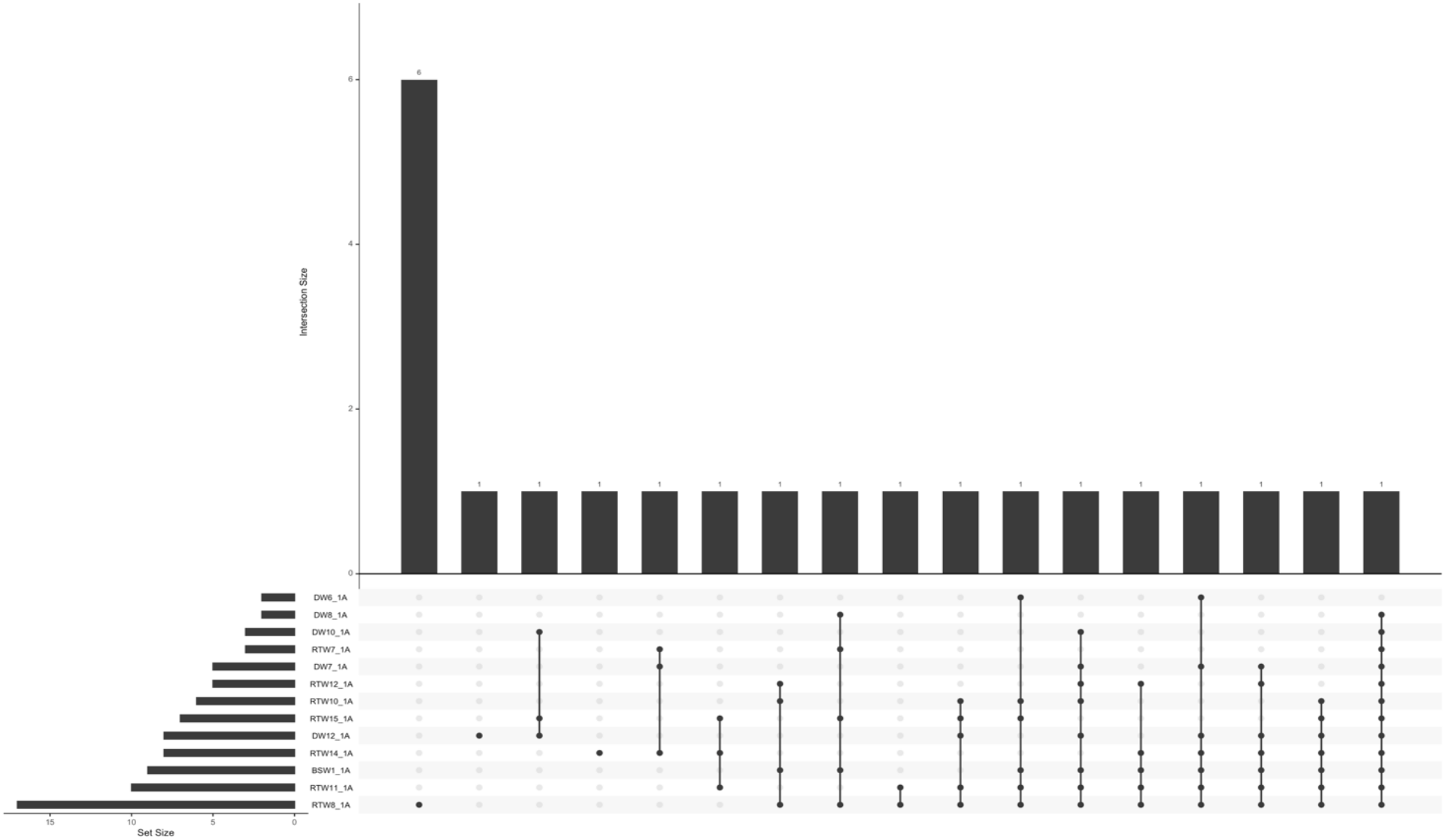


Figure 3-37: Unique and shared AMR Drug Classes across each sample. The bars on the left indicate the number of AMR Drug Classes per sample.

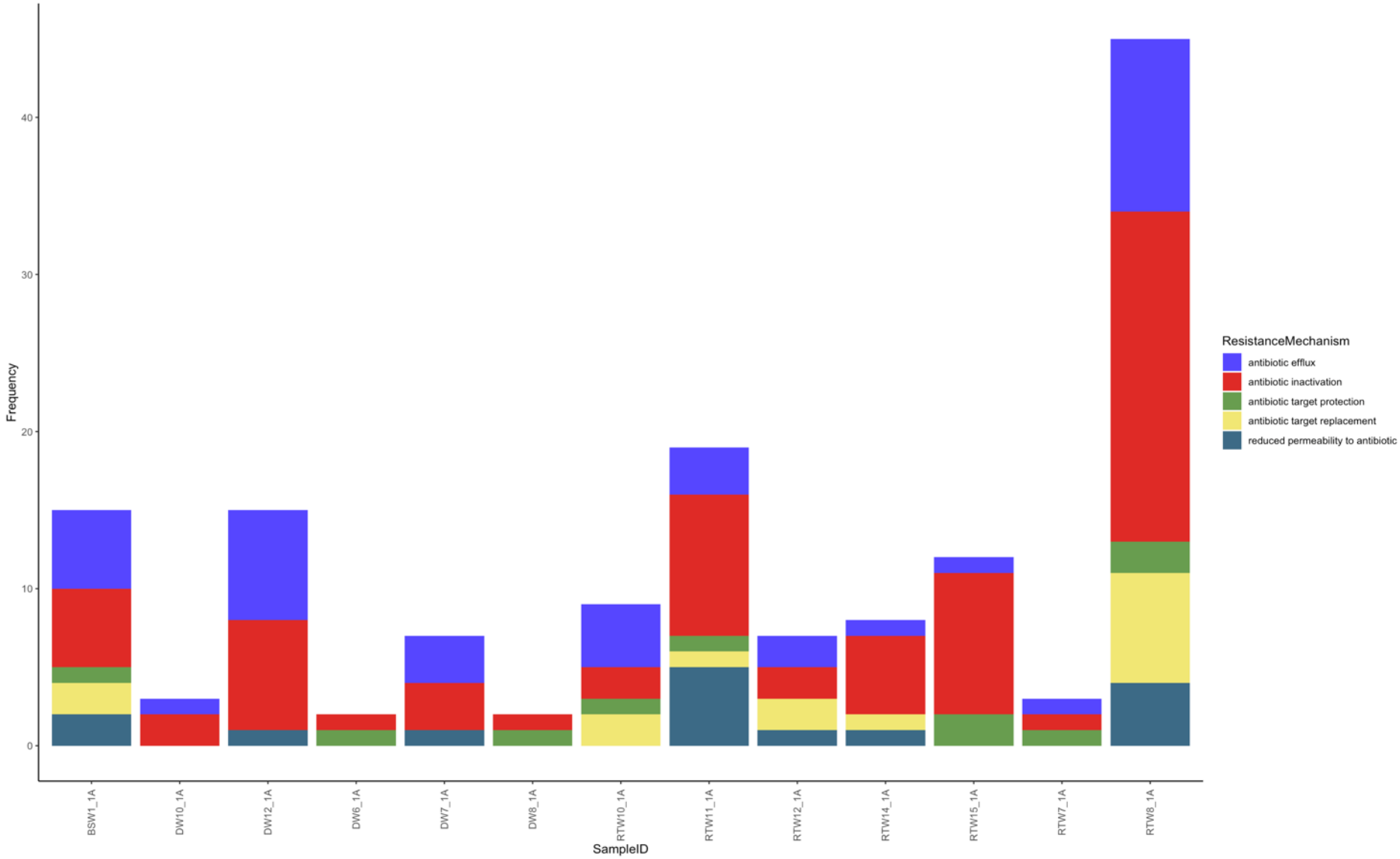


Figure 3-38: Distribution of AMR Resistance Mechanisms across each sample. The colours represent a particular AMR Resistance Mechanisms and the frequency of the AMR Resistance Mechanisms is indicated on the y-axis. The samples are on the x-axis.

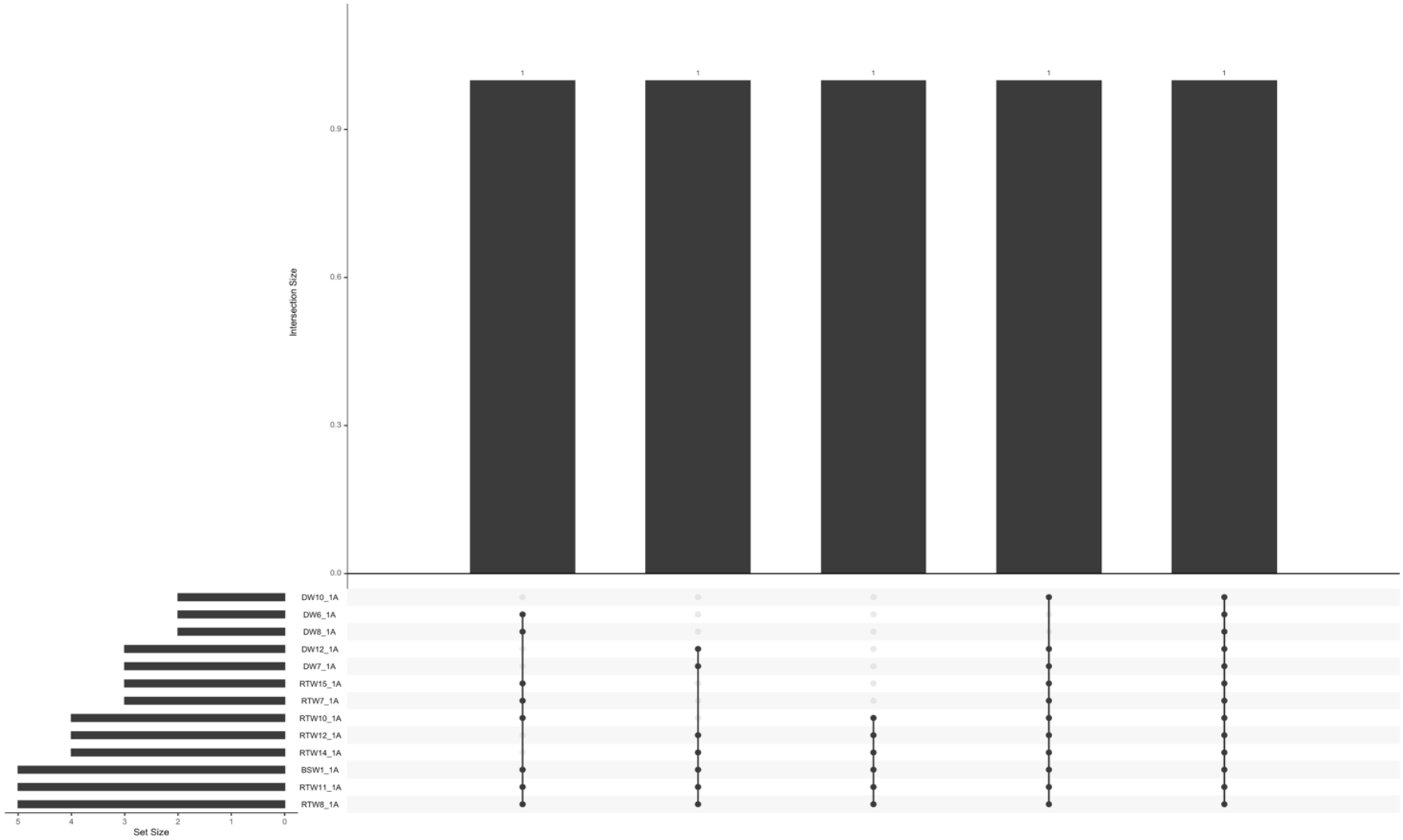


Figure 3-39: Unique and shared AMR Resistance Mechanisms across each sample. The bars on the left indicate the number of AMR Resistance Mechanisms per sample. The grid in the middle indicates which samples share a set of AMR Resistance Mechanisms and the bars on the top represent the size of the shared set.

3.4 DISCUSSION

The RNA metagenomic sequencing of 20 samples collected from various wastewater treatment sites across the Tshwane district during the period 17 August 2020 and 6 April 2021 resulted in the generation of approximately 80 GB of raw data. Standard quality filtering and decontamination protocols did not produce any significant data loss and the number of paired-end reads used in downstream analysis was in the range of 9-26 million. This is more than adequate for RNA metagenomic studies and inference.

The sequencing results were inspected for the presence of SARS-CoV-2. The samples have previously been confirmed to be positive for the presence of SARS-CoV-2 and the metagenomic sequencing results were analysed to provide some insights with regards to how deep needs to be sequenced, i.e. how much data needs to be produced, to detect portions of the SARS-CoV-2 genome using an RNA metagenomic sequencing approach. Segments of the SARS-CoV-2 genome was detected in 5 samples. Possible reasons for the detection of SARS-CoV-2 in RNA metagenomes is the amount of data produced or the viral load. These two possibilities were investigated by testing correlations. The results indicated that the 7-Day average COVID-19 cases were positively correlated with the percentage of the SARS-CoV-2 genome covered. The 7-Day average COVID-19 cases was used as a proxy to indicate viral load. As such it became clear that the detection of SARS-CoV-2 in RNA metagenomic sequence data was not dependent on the amount of data produced but rather the viral load, i.e. the amount of virus, present in a sample.

The ability to determine taxonomic profiles for each of the samples was clearly highlighted. Using both paired-end reads and *de novo* assembled transcripts, the high levels of diversity in the wastewater samples could be catalogued. Taxonomic classification was available for the Archaeal, Bacterial and Viral portions of each sample. This included the detection of various *Candidatus* classifications. The term *Candidatus* indicates that an organism is well characterized but not yet-uncultured. These are of great interest and should be investigated further. The metagenomic data further contained various “CrAss” annotations. These phages have only recently been identified and are dominant viruses in the human gut virome. CrAssphages have been proposed as a human-specific MST marker and are under investigation by various groups.

The added benefit of metagenomic sequencing is the ability to elucidate other potential functionalities in a sample. In this report we focused on the presence of AMRs within the samples. The samples were variable with regards to the presence and frequency of AMRs. High levels of AMRs were detected in various Rietgat samples and the reason for this should be further investigated. The detection of AMRs in wastewater samples will greatly in the modelling or prediction of AMR outbreaks.

The potential of metagenomic analysis in wastewater surveillance was clearly demonstrated. The generation of metagenomic data only requires extraction of DNA or RNA from samples. Thereafter a single sequencing event is able to produce data with various applications. These applications include taxonomic and functional characterisation. This methodology further circumvents various laborious and time-consuming activities and is able to detect organism which are not currently well documented or understood. Albeit at relatively low levels, RNA metagenomic sequencing was still able to detect portions of the SARS-CoV-2 genome in wastewater samples. This is remarkable as the genome of SARS-CoV-2 is only 30,000 base pairs and the diversity found within the wastewater samples was large.

In summary, the metagenomic approach as detailed in this report is a valuable alternative to the current protocols used in wastewater analysis. The results from this dataset have been submitted to the South African Journal of Science and is currently under review. The submitted manuscript abstract is in **Supplementary SAJS manuscript**.

CHAPTER 4: WHOLE GENOME SEQUENCING OF SARS-COV-2 AS FOUND IN WASTEWATER SAMPLES OBTAINED FROM DURBAN, KWAZULU-NATAL

4.1 INTRODUCTION

The recent classification of Omicron as a variant of concern has again demonstrated the abilities of SARS-CoV-2 to change and evolve. South Africa has been critical in identifying and alerting the rest of the world with regards to novel variants and possible variants of concern. This has clearly demonstrated our scientific capabilities as a nation. The growing list of high impact publications clearly illustrates these capabilities. This prowess can be ascribed to our expertise in genomics and bioinformatics and in particular whole genome sequencing and analysis of SARS-CoV-2.

Next-generation sequencing analysis of wastewater samples provides insights to human health related factors which includes the distribution of pathogens and antibiotic resistance genes (Yang et al., 2014). The contents of a wastewater sample provide researchers and stakeholders a glimpse as to what is circulating in the host associated environment and as such host health. Wastewater samples may be regarded as a pooled version of the human gut microbiome. Pathogens and antimicrobial resistance which are present in a wastewater sample may be presumed to have been present in the population gut microbiome prior to the sampling. These sewage water accurately reflect a population's gut microbial composition which therefor allows metagenomics and targeted whole genome sequencing to assist in obtaining information regarding the infection dynamics in a given population (Fresia et al., 2019).

The COVID-19 pandemic has increased awareness regarding the power and resolution of next-generation sequencing and genomics. This has been evident in the detection of SARS-CoV-2 variants and the tracking of COVID-19 infection. Wastewater-based epidemiology is a critical component in the detection and tracking of SARS-CoV-2 and it has been shown that sequencing of viral concentrations and RNA extracted directly from wastewater can identify multiple SARS-CoV-2 genotypes, including variants not yet observed in clinical sequencing programmes (Crits-Christoph et al., 2021).

Genomics and in particular targeted whole genome sequencing are therefore an eloquent application in wastewater surveillance and epidemiology. This method enables the detection and classification of SARS-CoV-2 in numerous samples based on a single data generation event or sequencing run. The results obtained from these targeted whole genome sequencing events can further be stored for long term use and be used as a baseline for future research endeavours.

Although whole genome sequencing and variant analysis of SARS-CoV-2 is generally done on clinical samples, wastewater samples have the potential to screen samples for SARS-CoV-2 and variants of concern on a large scale. The frequent analysis of wastewater samples will alert stakeholders, government and other interested bodies to the detection of SARS-CoV-2 and variants of concern which will allow for the rapid implementation of target testing. The data generated in these whole genome sequencing events will further be available for various other research endeavours and surveillance projects.

In the sections below, we clearly outline the methodology used and results obtained in the whole genome sequencing of SARS-CoV-2 obtained from wastewater samples collected during the period 21 July 2020 to 2 November 2021 from the Durban region in KwaZulu-Natal. The results illustrate the functionality, benefits and potential of SARS-CoV-2 whole genome sequencing of wastewater samples.

4.2 MATERIALS AND METHODS

Samples (n=73) were collected from various wastewater treatment sites across Durban, KwaZulu-Natal by DUT IWWT (Prof F. Bux). The sampling sites included Central (n=37), Isipingo (n=13), KwaMashu (n=11) and Phoenix (n=12) wastewater treatment plants. The samples were collected between 21 July 2020 and 2 November 2021 (Table 4-1 and Figure 4-1). These samples all tested positive for the presence of SARS-CoV-2. RNA extractions were done by the DUT IWWT and the resulting extractions delivered to the ARC Biotechnology for library preparation and SARS-CoV-2 whole genome sequencing (Supplementary Sequencing Quotation).

Table 4-1: Samples received for SARS-CoV-2 whole genome sequencing.

Sample ID	Date	Location
RP_21_07_2020_S53	2020/07/21	Central
RP_25_08_2020_S54	2020/08/25	Central
RP_29_09_2020_S55	2020/09/29	Central
RP_27_10_2020_S56	2020/10/27	Central
RP_17_11_2020_S57	2020/11/17	Central
RP_24_11_2020_S58	2020/11/24	Central
RP_15_12_2020_S12	2020/12/15	Central
RP_22_12_2020_S59	2020/12/22	Central
RP_29_12_2020_S60	2020/12/29	Central
RP_19_01_2021_S61	2021/01/19	Central
RP_26_01_2021_S13	2021/01/26	Central
RP_02_02_2021_S62	2021/02/02	Central
RP_23_02_2021_S63	2021/02/23	Central
RP_09_03_2021_S14	2021/03/09	Central
RP_30_03_2021_S64	2021/03/30	Central
RP_08_04_2021_S65	2021/04/08	Central
RP_13_04_2021_S15	2021/04/13	Central
RP_18_05_2021_S16	2021/05/18	Central
RP_27_05_2021_S17	2021/05/27	Central
RP_24_06_2021_S18	2021/06/24	Central
RP_30_06_2021_S19	2021/06/30	Central
RP_01_07_2021_S20	2021/07/01	Central
RP_27_07_2021_S2	2021/07/27	Central
RP_03_08_2021_S3	2021/08/03	Central
IWWT_1_S1	2021/08/10	Phoenix
IWWT_2_S2	2021/08/10	Isipingo
IWWT_3_S3	2021/08/10	KwaMashu
IWWT_4_S4	2021/08/10	Central
IWWT_5_S5	2021/08/17	Phoenix
IWWT_6_S6	2021/08/17	Isipingo
IWWT_7_S7	2021/08/17	KwaMashu
IWWT_8_S8	2021/08/17	Central
IWWT_10_S10	2021/08/24	Isipingo
IWWT_11_S11	2021/08/24	KwaMashu
IWWT_12_S12	2021/08/24	Central
IWWT_9_S9	2021/08/24	Phoenix
IWWT_13_S13	2021/08/31	Phoenix
IWWT_14_S14	2021/08/31	Isipingo
IWWT_15_S15	2021/08/31	KwaMashu
IWWT_16_S16	2021/08/31	Central

Sample ID	Date	Location
IWWT_17_S17	2021/09/08	Phoenix
IWWT_18_S18	2021/09/08	Isipingo
IWWT_19_S19	2021/09/08	KwaMashu
IWWT_20_S20	2021/09/08	Central
IWWT_21_S21	2021/09/14	Phoenix
IWWT_22_S22	2021/09/14	Isipingo
IWWT_23_S23	2021/09/14	KwaMashu
IWWT_24_S24	2021/09/14	Central
IWWT_26_S26	2021/09/21	Isipingo
IWWT_27_S27	2021/09/21	KwaMashu
IWWT_28_S28	2021/09/21	Central
IWWT_29_S29	2021/09/21	Phoenix
IWWT_30_S30	2021/09/28	Isipingo
IWWT_31_S31	2021/09/28	KwaMashu
IWWT_32_S32	2021/09/28	Central
IWWT_33_S33	2021/09/28	Phoenix
IWWT_34_S34	2021/10/07	Isipingo
IWWT_35_S35	2021/10/07	KwaMashu
IWWT_36_S36	2021/10/07	Central
IWWT_37_S37	2021/10/12	Phoenix
IWWT_38_S38	2021/10/12	Isipingo
IWWT_40_S40	2021/10/12	Central
IWWT_41_S41	2021/10/19	Phoenix
IWWT_42_S42	2021/10/19	Isipingo
IWWT_43_S43	2021/10/19	KwaMashu
IWWT_44_S44	2021/10/19	Central
IWWT_45_S45	2021/10/26	Phoenix
IWWT_46_S46	2021/10/26	Isipingo
IWWT_47_S47	2021/10/26	KwaMashu
IWWT_48_S48	2021/10/26	Central
IWWT_49_S49	2021/11/02	Phoenix
IWWT_50_S50	2021/11/02	Isipingo
IWWT_52_S52	2021/11/02	Central

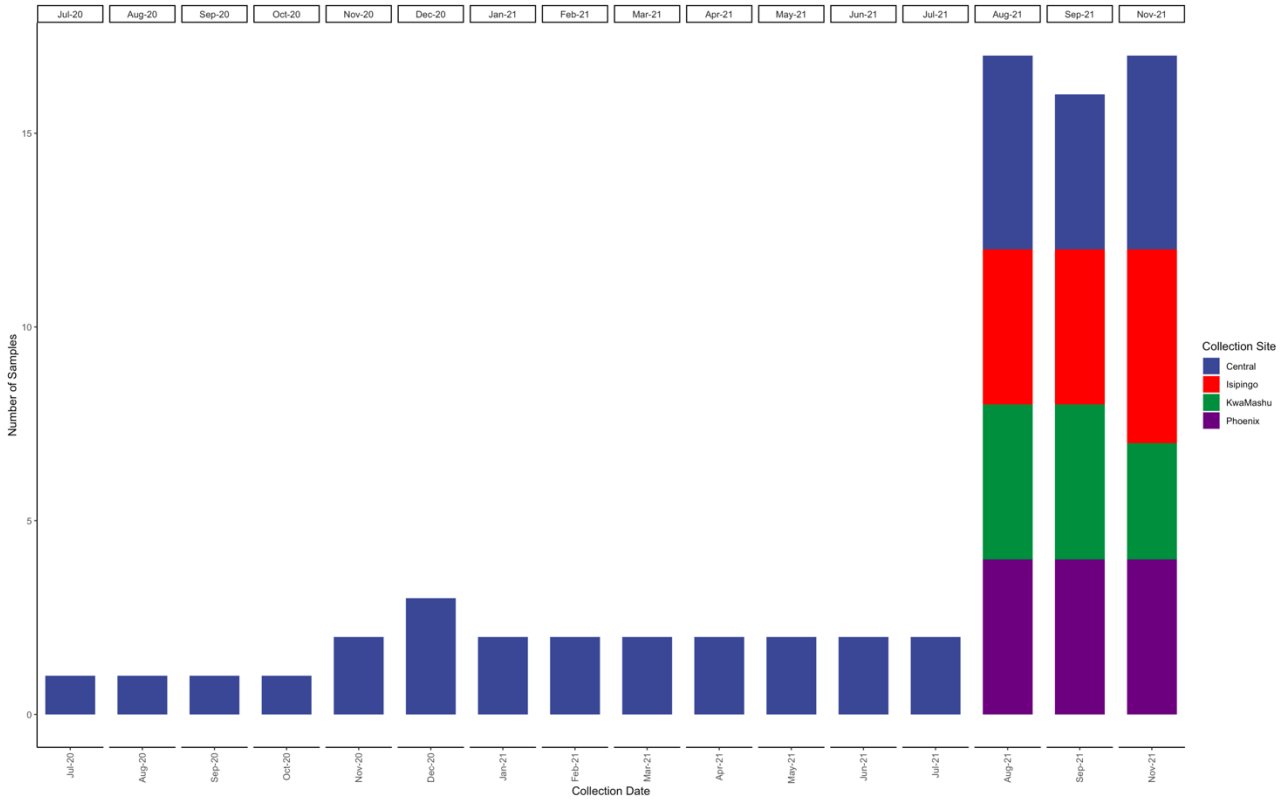


Figure 4-1: Samples received for SARS-CoV-2 whole genome sequencing. The colours indicate the sampling location, x-axis the date of sampling and y-axis the number of samples. The subtitles on the top of each bar indicate the sampling date.

RNA samples were processed using the NEBNext® ARTIC SARS-CoV-2 Library Prep Kit (Illumina®) and sequenced on an Illumina® MiSeq platform at the ARC-Biotechnology Platform. Initial sequence data quality and filtered data quality was inspected using FastQC version 0.11.8 (Andrews, S., 2010). Sequence data was quality trimmed and filtered, including adapter removal and decontamination, using BBDuk version 38.91 available from the BBTools suite of tools (Bushnell, B., 2014). Human contamination in the quality filtered sequencing data was removed by aligning the sequence data against the latest reference human genome (GRCh38.p13) using BBMap version 38.91, available from BBTools. To identify the portion of the SARS-CoV-2 genome sequenced, filtered and decontaminated paired-end reads were aligned to the SARS-CoV-2 reference genome (MN908947.3) with BBMap and coverage statistics calculated. The quality filtered sequencing data was *de novo* assembled into scaffolds using SPAdes version 3.15.3 (Bushmanova et al., 2019) with the "coronaSPAdes" option specified (Meleshko et al., 2021). This is a special mode of "rnaviralSPAdes" specifically aimed for SARS-CoV-2 *de novo* assembly. The quality of the *de novo* assemblies were assessed with QUAST version 5. 0. 2 (Gurevich et al., 2013).

The Utah DoH ARTIC/Illumina Bioinformatic Workflow (https://github.com/CDCgov/SARS-CoV-2_Sequencing/tree/master/protocols/BFX-UT_ARTIC_Illumina) was followed to construct a consensus sequence for each sample based on the V3 ARTIC primer scheme and ARTIC default reference (MN908947.3). Quality filtered, decontaminated reads are mapped to the SARS-CoV-2 reference genome using BWA-MEM version 0.7.17-r1188 (Li, H and Durbin, R., 2009) and thereafter sorted and unmapped reads removed with samtools version 1.10 (Li et al., 2009). The ARTIC primers are removed from the resulting "bam" files with iVar version 1.3.1 (Grubaugh et al., 2019). The "bam" files obtained after primer removal are again sorted using samtools and a consensus sequence generated using samtools (-aa -A -d 0 -B -Q 0) and iVar (-t 0.9 -m 20 -n N).

The consensus sequence for each sample was used to identify possible SARS-CoV-2 lineages with Pangolin (O'Toole et al., 2021). Pangolin version 3.1.17 was implemented in both the default and "UShEr" mode (Turakhia et al., 2021). The version used by Pangolin were pangolearn: 2021-11-25, constellations: v0.0.28, scorpio: 0.3.15, pango-designation used by pangoLEARN/Usher: v1.2.101 and pango-designation aliases: 1.2.107An additional program, hedgehog version 1.0.6 (https://github.com/cov-lineages/hedgehog), was implemented to determine SARS-CoV-2 lineage assignment. Consensus sequences which passed the internal Pangolin quality control settings were aligned using the "--alignment" flag. This alignment was then used in combination with IQ-TREE version 2.1.1 (Nguyen et al., 2015) to construct a maximum likelihood phylogeny. Additional data analysis and visualization was done in R version 4.0.2 (Team, R Core, 2020) implemented in RStudio version 1.4.1717 (Team, RStudio, 2021) with added libraries ggstatsplot library (Patil, I., 2021) and ggtree (Yu et al., 2017).

4.3 RESULTS

4.3.1 Data quality filtering and decontamination

Approximately 9 GB worth of raw sequencing data was produced for the 73 samples. The raw sequencing data was quality filtered and the resulting sequence quality of the filtered reads were again inspected using FastQC. Sequencing data which mapped to the human genome was removed and the quality of the remaining sequence data again quality checked with FastQC. The number of reads for each sample is presented in Table 4-2 and Figure 4-2.

Table 4-2: Number of reads at each stage of quality control and decontamination.

Sample ID	Raw Reads	QC Reads	No Human QC Reads
IWWT_10_S10	329,308	302,486	302,484
IWWT_11_S11	497,203	457,804	457,801
IWWT_12_S12	388,431	366,546	366,546
IWWT_13_S13	443,471	410,964	410,964
IWWT_14_S14	461,888	431,768	431,711
IWWT_15_S15	276,354	252,143	252,136
IWWT_16_S16	444,047	416,137	416,134
IWWT_17_S17	370,196	337,728	337,713
IWWT_18_S18	330,781	297,674	297,673
IWWT_19_S19	347,507	319,192	319,188
IWWT_1_S1	389,923	359,938	359,937
IWWT_20_S20	386,265	353,498	353,305
IWWT_21_S21	353,809	324,464	324,462
IWWT_22_S22	250,731	231,462	231,461
IWWT_23_S23	274,802	245,460	245,455
IWWT_24_S24	402,170	376,706	376,705
IWWT_26_S26	254,354	219,912	219,910
IWWT_27_S27	285,410	264,555	264,553
IWWT_28_S28	145,962	131,441	131,437
IWWT_29_S29	473,220	430,798	430,794
IWWT_2_S2	307,071	220,069	218,957
IWWT_30_S30	242,173	188,418	188,383
IWWT_31_S31	332,929	219,212	219,205
IWWT_32_S32	372,624	332,079	332,079
IWWT_33_S33	335,993	300,067	300,058
IWWT_34_S34	347,705	275,593	275,570
IWWT_35_S35	429,358	273,250	273,227
IWWT_36_S36	293,750	269,209	269,204

Sample ID	Raw Reads	QC Reads	No Human QC Reads
IWWT_37_S37	310,604	233,731	233,679
IWWT_38_S38	135,606	120,069	120,068
IWWT_3_S3	298,319	266,878	266,874
IWWT_40_S40	249,494	230,056	230,045
IWWT_41_S41	421,040	325,805	325,735
IWWT_42_S42	400,726	359,541	359,507
IWWT_43_S43	372,621	334,828	334,698
IWWT_44_S44	191,418	175,536	175,524
IWWT_45_S45	202,531	169,959	169,845
IWWT_46_S46	205,381	180,073	180,031
IWWT_47_S47	284,970	226,819	226,787
IWWT_48_S48	444,694	338,111	338,104
IWWT_49_S49	341,113	276,960	276,940
IWWT_4_S4	282,085	223,969	223,963
IWWT_50_S50	407,584	292,163	292,073
IWWT_52_S52	340,815	317,228	317,227
IWWT_5_S5	442,156	417,503	417,500
IWWT_6_S6	404,413	383,989	383,988
IWWT_7_S7	284,235	269,760	269,759
IWWT_8_S8	249,482	215,868	215,868
IWWT_9_S9	492,517	460,645	460,643
RP_01_07_2021_S20	949,371	802,360	802,330
RP_02_02_2021_S62	177,226	91,272	91,248
RP_03_08_2021_S3	516,260	465,224	465,224
RP_08_04_2021_S65	140,713	90,715	90,623
RP_09_03_2021_S14	742,135	626,343	625,650
RP_13_04_2021_S15	367,168	197,292	197,258
RP_15_12_2020_S12	221,287	158,140	158,136
RP_17_11_2020_S57	52,215	27,175	27,175
RP_18_05_2021_S16	242,477	158,046	157,981
RP_19_01_2021_S61	133,991	48,497	48,312
RP_21_07_2020_S53	53,423	25,319	25,278
RP_22_12_2020_S59	94,459	36,951	28,344
RP_23_02_2021_S63	168,037	77,269	70,707
RP_24_06_2021_S18	811,074	718,295	718,283
RP_24_11_2020_S58	38,867	19,252	19,240
RP_25_08_2020_S54	55,134	24,001	21,948
RP_26_01_2021_S13	304,356	172,441	172,433
RP_27_05_2021_S17	435,953	299,226	299,091
RP_27_07_2021_S2	492,571	445,811	445,811
RP_27_10_2020_S56	28,192	8,812	8,812
RP_29_09_2020_S55	57,062	26,385	26,346
RP_29_12_2020_S60	167,962	37,294	37,290
RP_30_03_2021_S64	144,370	104,151	104,108
RP_30_06_2021_S19	800,541	654,291	654,264

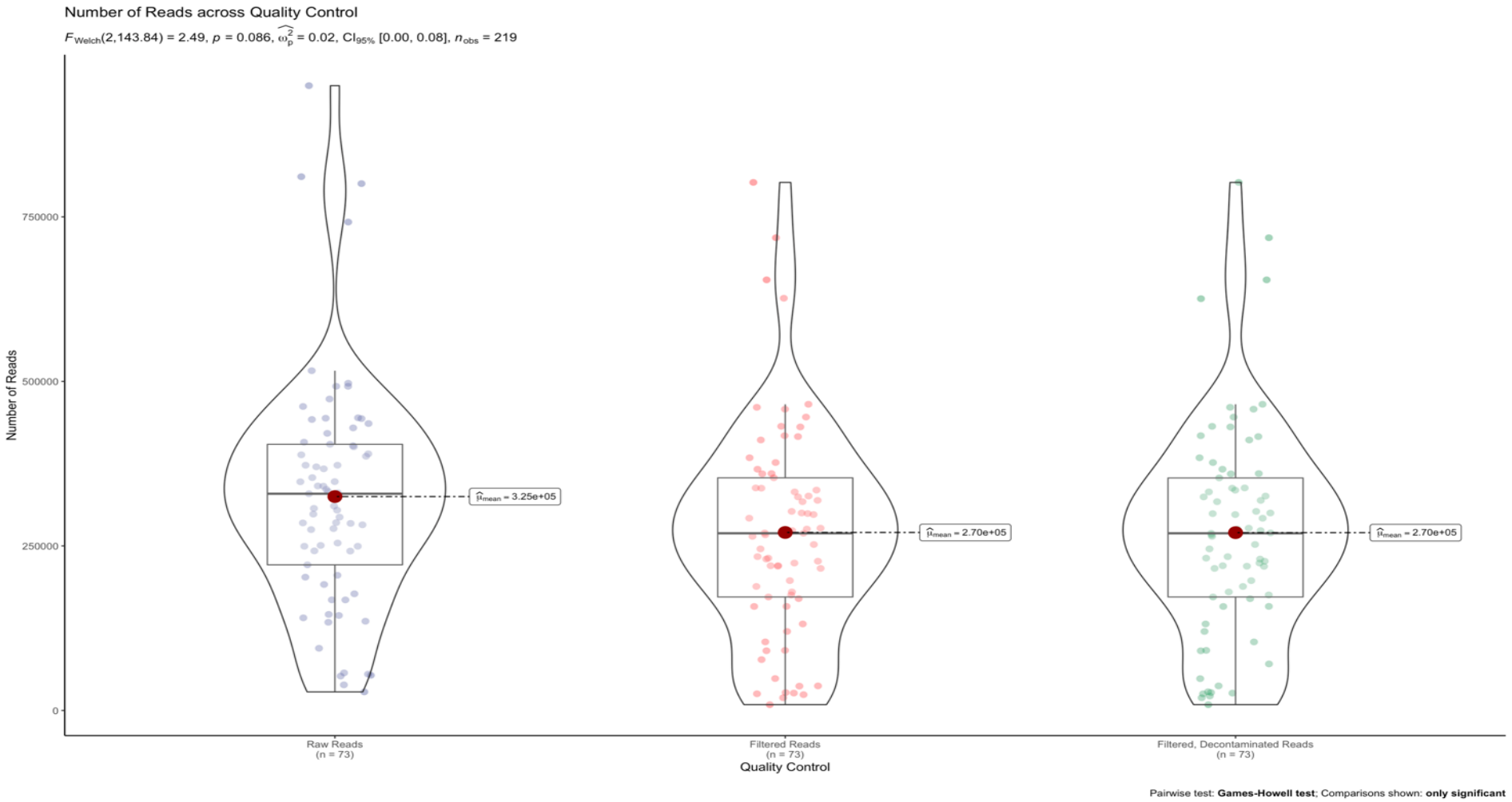


Figure 4-2: Number of reads at each stage of quality control and decontamination. The colours indicate the quality control step, x-axis the sample and y-axis the number of reads. Low levels of data loss were seen and the number of reads surviving quality filtering and human decontamination was more than adequate for the project. No significant differences in the number of reads between any of the processing steps were observed. The results from the statistical test are reported in the subtitles on the top of each graph.

Data loss due to quality and contamination was as expected and more than enough reads remained for further analysis. The low levels of data loss after decontamination, i.e. human, clearly illustrates the application of the NEBNext ARTIC SARS-CoV-2 protocol for targeted sequencing. On average the raw dataset contained 324,987 reads, the quality filtered 270,447 reads and the quality filtered decontaminated set 270,162 reads. No significant differences in the number of reads between of the processing steps were observed (p-value = 0.086). In general, the older samples performed worse, i.e. produced less quality controlled, decontaminated reads than the more recent samples (Figure 4-3). This may be due to the stability of RNA samples which generally decrease over a period of time even under optimal conditions.

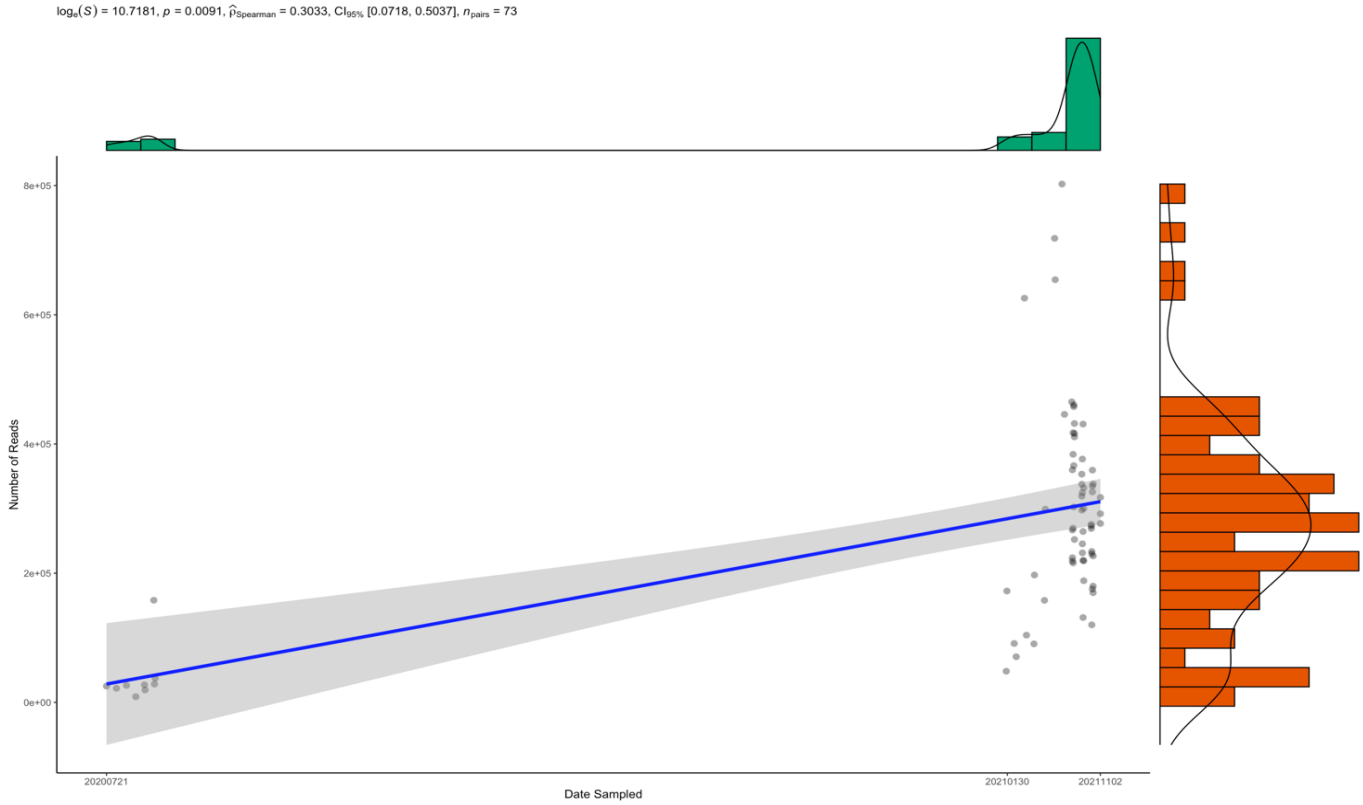


Figure 4-3: Spearman's rank correlation coefficient test results for the correlation between the number of quality controlled, decontaminated reads and the date sampled. The results from the statistical test are reported in the subtitles on the top of each graph. The marginal distributions for the x and y variables are overlaid on the axes of each graph. A significant positive correlation was evident and as such the "age" of a sample or duration from date sampled to the date sequenced is crucial and influences the number of reads generated.

4.3.2 Classification of SARS-CoV-2 lineages

The different SARS-CoV-2 lineage assignment protocols, i.e. Pangolin, Pangolin+Usher and Hedgehog, produced varying results with some overlap. For both the Pangolin and Pangolin+Usher protocol 38 samples passed the internal quality control parameters and for the Hedgehog protocol 33 samples were deemed adequate (Table 4-3).

Table 4-3: SARS-CoV-2 lineage assignment.

Sample ID	Date	Location	Pangolin	Usher	Hedgehog	Reference Coverage (%)
RP_21_07_2020_S53	2020/07/21	Central	None	None	None	17.9280
RP_25_08_2020_S54	2020/08/25	Central	None	None	None	54.5999
RP_29_09_2020_S55	2020/09/29	Central	None	None	None	0.0000
RP_27_10_2020_S56	2020/10/27	Central	None	None	None	78.5975
RP_17_11_2020_S57	2020/11/17	Central	None	None	None	64.1909
RP_24_11_2020_S58	2020/11/24	Central	None	None	None	82.6238
RP_15_12_2020_S12	2020/12/15	Central	None	None	None	98.8061
RP_22_12_2020_S59	2020/12/22	Central	None	None	None	94.2447
RP_29_12_2020_S60	2020/12/29	Central	None	None	None	89.7301
RP_19_01_2021_S61	2021/01/19	Central	None	None	None	80.3498
RP_26_01_2021_S13	2021/01/26	Central	None	None	None	100.0000
RP_02_02_2021_S62	2021/02/02	Central	None	None	None	2.9830
RP_23_02_2021_S63	2021/02/23	Central	None	None	None	98.5286
RP_09_03_2021_S14	2021/03/09	Central	P.1	C.6	P.1	100.0000
RP_30_03_2021_S64	2021/03/30	Central	None	None	None	1.3811
RP_08_04_2021_S65	2021/04/08	Central	None	None	None	98.8061
RP_13_04_2021_S15	2021/04/13	Central	None	None	None	94.6159
RP_18_05_2021_S16	2021/05/18	Central	None	None	None	97.9534
RP_27_05_2021_S17	2021/05/27	Central	None	None	None	91.2952
RP_24_06_2021_S18	2021/06/24	Central	C.36.3.1	None	B.1.616	100.0000
RP_30_06_2021_S19	2021/06/30	Central	AY.45	AY.45	B.1.617.2	100.0000
RP_01_07_2021_S20	2021/07/01	Central	B.1.617.2	B.1.617.2	B.1.617.2	100.0000
RP_27_07_2021_S2	2021/07/27	Central	AY.43	B.1.617.2	B.1.617.2	100.0000
RP_03_08_2021_S3	2021/08/03	Central	AY.44	AY.45	B.1.617.2	100.0000
IWWT_1_S1	2021/08/10	Phoenix	AY.45	AY.45	B.1.617.2	99.8763
IWWT_2_S2	2021/08/10	Isipingo	None	None	None	100.0000
IWWT_3_S3	2021/08/10	KwaMashu	AY.45	AY.45	B.1.617.2	100.0000
IWWT_4_S4	2021/08/10	Central	B.1.2	None	None	100.0000
IWWT_5_S5	2021/08/17	Phoenix	AY.44	B.1.617.2	B.1.617.2	100.0000
IWWT_6_S6	2021/08/17	Isipingo	AY.44	B.1.617.2	B.1.617.2	100.0000
IWWT_7_S7	2021/08/17	KwaMashu	AY.44	AY.45	B.1.617.2	99.9532
IWWT_8_S8	2021/08/17	Central	B.1.629	None	A	100.0000
IWWT_10_S10	2021/08/24	Isipingo	None	None	None	99.8763
IWWT_11_S11	2021/08/24	KwaMashu	AY.45	AY.45	B.1.617.2	100.0000
IWWT_12_S12	2021/08/24	Central	B.1.2	None	B.1.616	100.0000
IWWT_9_S9	2021/08/24	Phoenix	AY.45	AY.45	B.1.617.2	100.0000
IWWT_13_S13	2021/08/31	Phoenix	AY.45	AY.45	B.1.617.2	99.8763
IWWT_14_S14	2021/08/31	Isipingo	B.1.2	None	B.1.616	99.9465
IWWT_15_S15	2021/08/31	KwaMashu	AY.45	AY.45	B.1.617.2	100.0000
IWWT_16_S16	2021/08/31	Central	AY.45	AY.45	B.1.617.2	100.0000
IWWT_17_S17	2021/09/08	Phoenix	AY.45	AY.45	B.1.617.2	99.9398
IWWT_18_S18	2021/09/08	Isipingo	B.1.617.2	AY.45	None	100.0000
IWWT_19_S19	2021/09/08	KwaMashu	AY.45	AY.45	B.1.617.2	100.0000
IWWT_20_S20	2021/09/08	Central	B.1	None	A	100.0000
IWWT_21_S21	2021/09/14	Phoenix	AY.43	AY.45	B.1.617.2	99.8763
IWWT_22_S22	2021/09/14	Isipingo	AY.45	AY.45	B.1.617.2	100.0000
IWWT_23_S23	2021/09/14	KwaMashu	AY.45	AY.45	None	99.9030
IWWT_24_S24	2021/09/14	Central	B.1.2	None	B.1.616	100.0000

Sample ID	Date	Location	Pangolin	Usher	Hedgehog	Reference Coverage (%)
IWWT_26_S26	2021/09/21	Isipingo	AY.45	AY.45	B.1.617.2	100.0000
IWWT_27_S27	2021/09/21	KwaMashu	AY.45	AY.45	B.1.617.2	99.8796
IWWT_28_S28	2021/09/21	Central	None	None	B.1.616	100.0000
IWWT_29_S29	2021/09/21	Phoenix	AY.45	AY.45	B.1.617.2	99.9064
IWWT_30_S30	2021/09/28	Isipingo	None	None	None	98.0771
IWWT_31_S31	2021/09/28	KwaMashu	None	None	None	98.2276
IWWT_32_S32	2021/09/28	Central	None	None	None	99.8796
IWWT_33_S33	2021/09/28	Phoenix	None	None	None	99.2977
IWWT_34_S34	2021/10/07	Isipingo	None	None	None	98.2443
IWWT_35_S35	2021/10/07	KwaMashu	None	None	None	96.8465
IWWT_36_S36	2021/10/07	Central	B.1	B.1	A	100.0000
IWWT_37_S37	2021/10/12	Phoenix	None	None	None	92.4957
IWWT_38_S38	2021/10/12	Isipingo	C.1.2	C.1.2	None	99.8763
IWWT_40_S40	2021/10/12	Central	B.1	B.1	A	99.8863
IWWT_41_S41	2021/10/19	Phoenix	None	None	None	94.6159
IWWT_42_S42	2021/10/19	Isipingo	AY.45	AY.45	None	100.0000
IWWT_43_S43	2021/10/19	KwaMashu	None	None	None	99.2977
IWWT_44_S44	2021/10/19	Central	B.1	None	A	100.0000
IWWT_45_S45	2021/10/26	Phoenix	None	None	None	82.9883
IWWT_46_S46	2021/10/26	Isipingo	None	None	None	99.8796
IWWT_47_S47	2021/10/26	KwaMashu	None	None	None	88.5028
IWWT_48_S48	2021/10/26	Central	None	None	None	95.1443
IWWT_49_S49	2021/11/02	Phoenix	None	None	None	0.0000
IWWT_50_S50	2021/11/02	Isipingo	None	None	None	80.5906
IWWT_52_S52	2021/11/02	Central	A.2.5	B.1.140	P.1	100.0000

The Pangolin analysis identified 38 samples to be adequate for lineage assignment and thereafter failed to assign a lineage to 1 of the 38 samples. There were 11 lineages detected and frequency found is in Table 4-4. The Pangolin analysis indicated a high proportion of lineage AY.45 (Delta variant). The lineage assignments per sample are displayed in Figure 4-4 and Figure 4-5. These figures included the portion of the SARS-CoV-2 reference genome covered, date sampled, location and active COVID-19 cases in KwaZulu-Natal.

Table 4-4: Pangolin SARS-CoV-2 lineage assignment.

SARS-CoV-2 Lineage	Number of Samples
A.2.5	1
AY.43	2
AY.44	4
AY.45	16
B.1	4
B.1.2	4
B.1.617.2	2
B.1.629	1
C.1.2	1
C.36.3.1	1
P.1	1

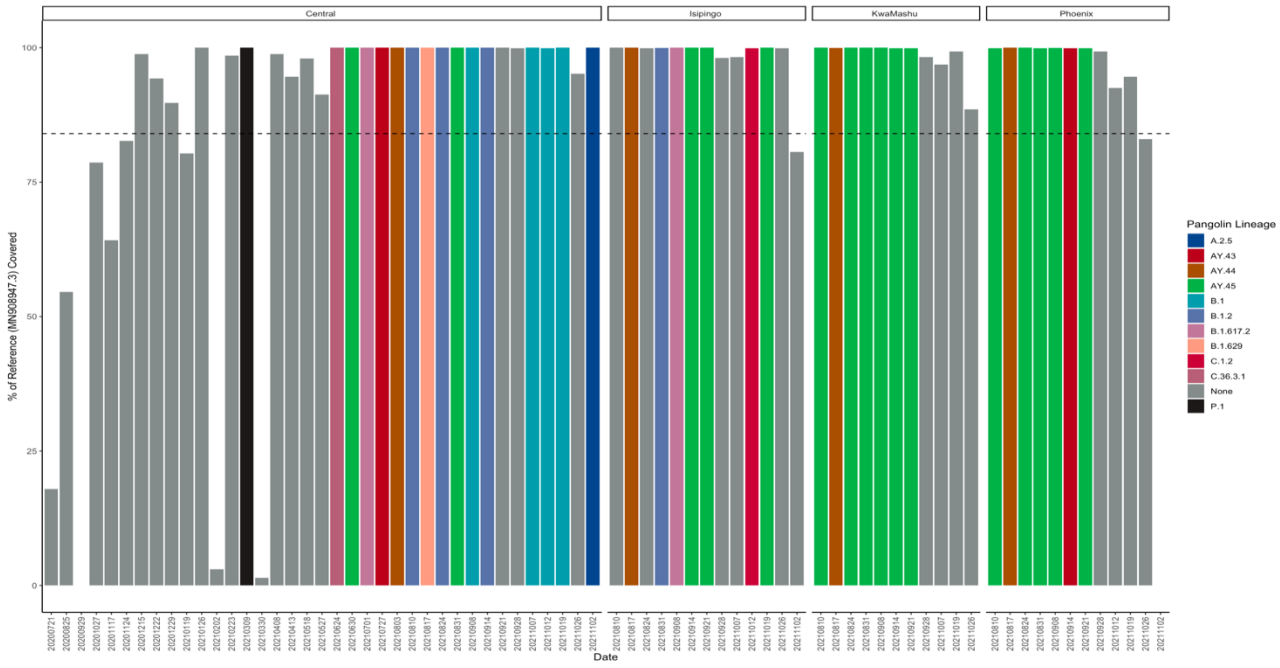


Figure 4-4: Pangolin lineage assignment for all samples. The x-axis indicates the date sampled and the y-axis the percentage SARS-CoV-2 reference genome covered. Each bar represents a sample and the colours indicate the assigned lineage. The samples are further grouped according to location. The grey bars represent samples for which lineage assignment was not possible. The dashed horizontal line indicates a cut-off for the percentage coverage. Samples below this threshold would not cover enough of the reference to produce results.

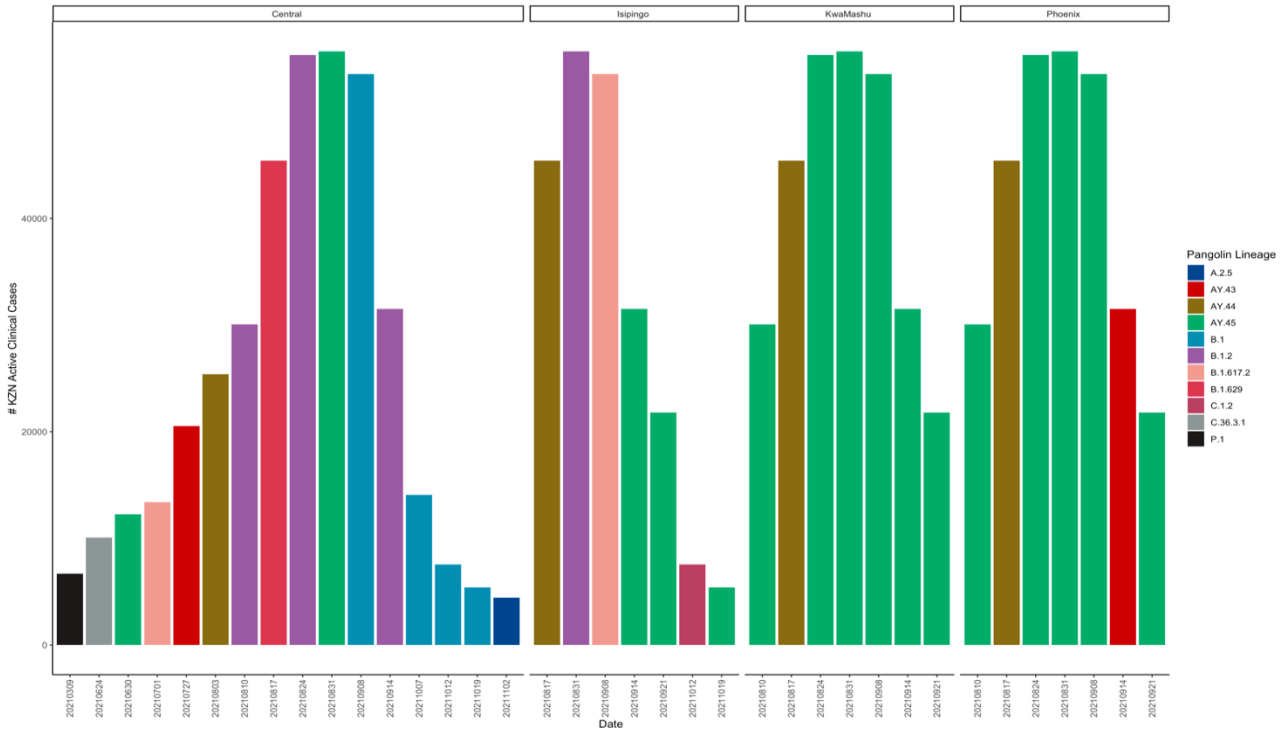


Figure 4-5: Pangolin lineage assignment with "None" assigned removed. The x-axis indicates the date sampled and the y-axis the number of active clinical COVID-19 cases. Each bar represents a sample and the colours indicate the assigned lineage. Differences in variant diversity can be seen based on the colours. Central and Isipingo appear more diverse with regards to lineages assigned than KwaMashu and Phoenix.

In Figure 4-4 possible reasons for the failure to assign a lineage are seen. These are coloured in grey. Samples not covering a large enough portion of the reference genome will not be assigned a lineage as there is not

enough certainty with regards to the possible variations. These would be represented by samples with bars not extending past the dashed horizontal line and would explain a large portion of the samples not assigned a lineage. The samples with an adequate amount of reference coverage yet not assigned a lineage would possibly indicate a high diversity of variants in a sample. Due to various mutations present for numerous variants in some of samples the lineage assignment protocol would not be able to distinguish between the possible variants with high confidence. The number of active clinical COVID-19 cases in KwaZulu-Natal are presented in Figure 4-5, coloured by the assigned lineage, samples with “None assignments removed”, and grouped according to the sampling location. The locations displayed varying variant diversity which may be due to the demographic or other social factors such as movement.

The Pangolin+ UShEr analysis identified 38 samples to be adequate for lineage assignment and thereafter failed to assign a lineage to 9 of the 38 samples. There were 6 lineages detected and frequency found is in Table 4-5. The Pangolin+UShEr analysis indicated a high abundance of lineage AY.45 (Delta variant). The lineage assignments per sample are displayed in **Figure 3.6** and **Figure 3.7**. These figures included the portion of the SARS-CoV-2 reference genome covered, date sampled, location and active COVID-19 cases in KwaZulu-Natal.

Table 4-5: Pangolin+UShEr SARS-CoV-2 lineage assignment.

SARS-CoV-2 Lineage	Number of Samples
AY.45	20
B.1	2
B.1.140	1
B.1.617.2	4
C.1.2	1
C.6	1

In Figure 4-6 possible reasons for the failure to assign a lineage are seen. These are coloured in dark red. Samples not covering a large enough portion of the reference genome will not be assigned a lineage as there is not enough certainty with regards to the possible variations. These would be represented by samples with bars not extending past the dashed horizontal line and would explain a large portion of the samples not assigned a lineage. The samples with an adequate amount of reference coverage yet not assigned a lineage would possibly indicate a high diversity of variants in a sample. Due to various mutations present for numerous variants in some of samples the lineage assignment protocol would not be able to distinguish between the possible variants with high confidence.

The number of active clinical COVID-19 cases in KwaZulu-Natal are presented in Figure 4-7, coloured by the assigned lineage, samples with “None assignments removed”, and grouped according to the sampling location. The locations displayed varying variant diversity which may be due to the demographic or other social factors such as movement.

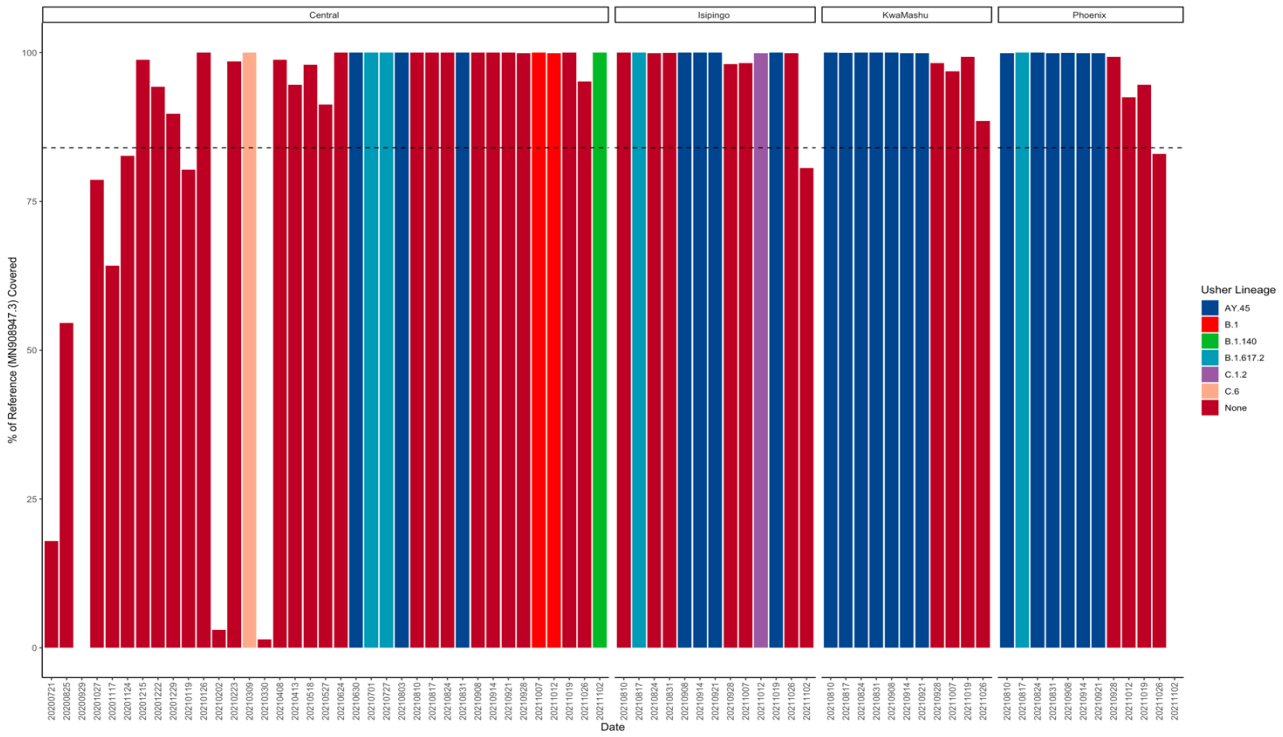


Figure 4-6: Pangolin+UShEr lineage assignment for all samples. The x-axis indicates the date sampled and the y-axis the percentage SARS-CoV-2 reference genome covered. Each bar represents a sample and the colours indicate the assigned lineage. The samples are further grouped according to location. The dark red bars represent samples for which lineage assignment was not possible. The dashed horizontal line indicates a cut-off for the percentage coverage. Samples below this threshold would not cover enough of the reference to produce results.

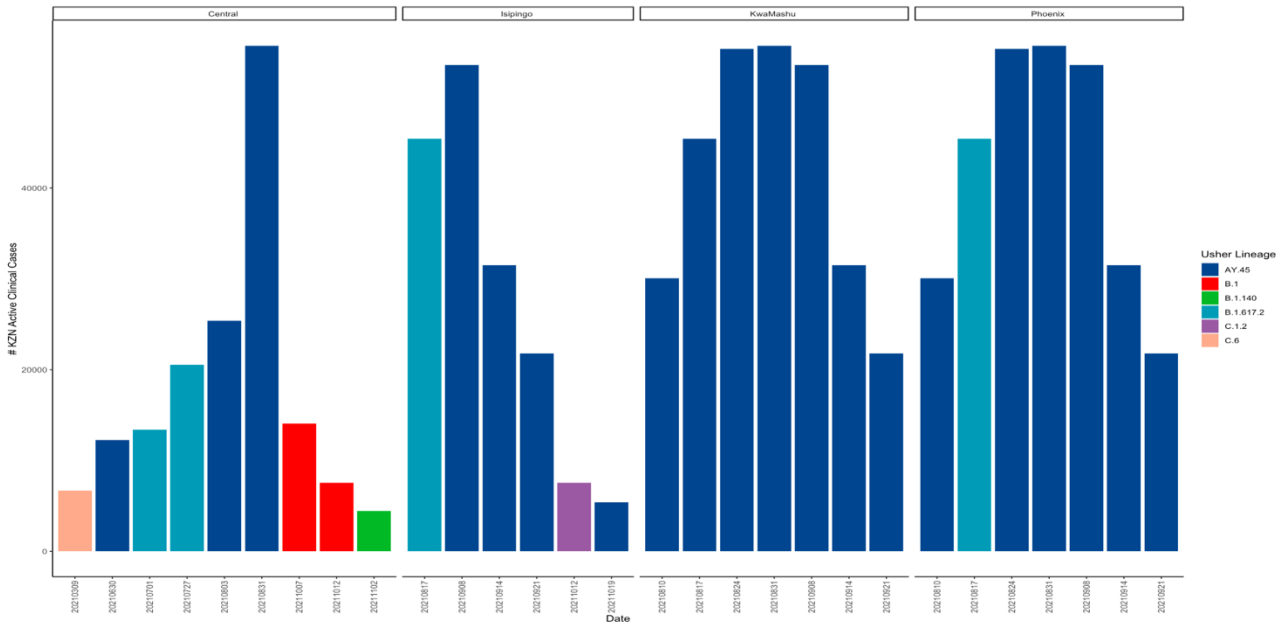


Figure 4-7: Pangolin+UShEr lineage assignment with "None" assigned removed. The x-axis indicates the date sampled and the y-axis the number of active clinical COVID-19 cases. Each bar represents a sample and the colours indicate the assigned lineage. Differences in variant diversity can be seen based on the colours. Central and Isipingo appear more diverse with regards to lineages assigned than KwaMashu and Phoenix.

The Hedgehog analysis identified 33 samples to be adequate for lineage assignment and thereafter none failed lineage assignment. There were 6 lineages detected and frequency found is in Table 4-6.

Table 4-6: Hedgehog SARS-CoV-2 lineage assignment.

SARS-CoV-2 Lineage	Number of Samples
A	5
B.1.616	5
B.1.617.2	21
P.1	2

The Hedgehog analysis indicated a high abundance of lineage B.1.617.2 (Delta variant). The lineage assignments per sample are displayed in Figure 4-8 and Figure 4-9. These figures included the portion of the SARS-CoV-2 reference genome covered, date sampled, location and active COVID-19 cases in KwaZulu-Natal.

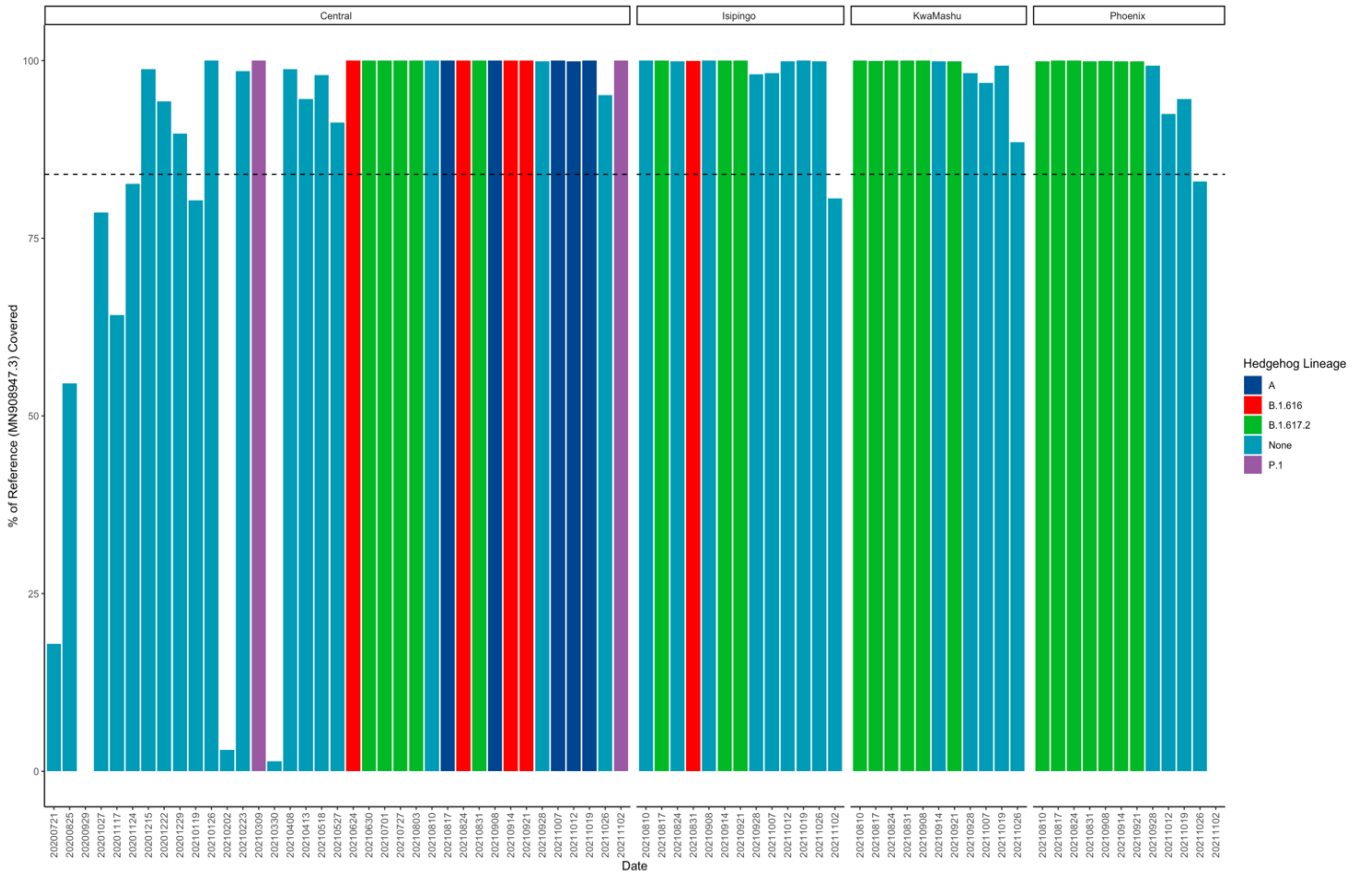


Figure 4-8: Hedgehog lineage assignment for all samples. The x-axis indicates the date sampled and the y-axis the percentage SARS-CoV-2 reference genome covered. Each bar represents a sample and the colours indicate the assigned lineage. The samples are further grouped according to location. The turquoise bars represent samples for which lineage assignment was not possible. The dashed horizontal line indicates a cut-off for the percentage coverage. Samples below this threshold would not cover enough of the reference to produce results.

In Figure 4-8 possible reasons for the failure to assign a lineage are seen. These are coloured in dark turquoise. Samples not covering a large enough portion of the reference genome will not be assigned a lineage as there is not enough certainty with regards to the possible variations. These would be represented by samples with bars not extending past the dashed horizontal line and would explain a large portion of the samples not

assigned a lineage. The samples with an adequate amount of reference coverage yet not assigned a lineage would possibly indicate a high diversity of variants in a sample. Due to various mutations present for numerous variants in some of samples the lineage assignment protocol would not be able to distinguish between the possible variants with high confidence.

The number of active clinical COVID-19 cases in KwaZulu-Natal are presented in Figure 4-9, coloured by the assigned lineage, samples with “None assignments removed”, and grouped according to the sampling location. The locations displayed varying variant diversity which may be due to the demographic or other social factors such as movement.

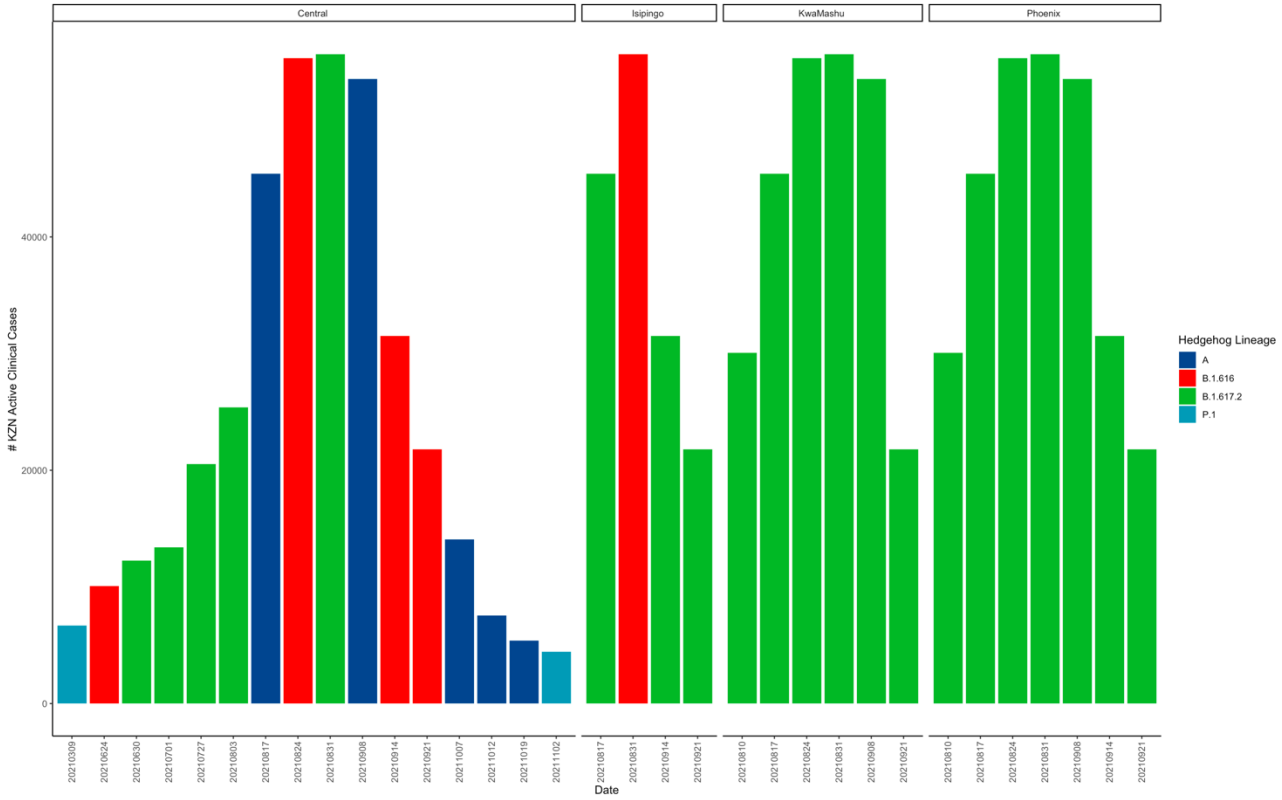


Figure 4-9: Hedgehog lineage assignment with “None” assigned removed. The x-axis indicates the date sampled and the y-axis the number of active clinical COVID-19 cases. Each bar represents a sample and the colours indicate the assigned lineage. Differences in variant diversity can be seen based on the colours. Central and Isipingo appear more diverse with regards to lineages assigned than KwaMashu and Phoenix.

The results of Pangolin, Pangolin+UShEr and Hedgehog all indicated a high representation of Delta variants (Figure 4-10). The high proportion of Delta variants are as expected. They would be the most dominant variant in wastewater samples collected recently and as indicated the more recent samples produced better sequencing results. The ability to detect other variants, e. g. Beta variant (sampled 2021/03/09), in wastewater sample is of critical importance in wastewater-based epidemiology.

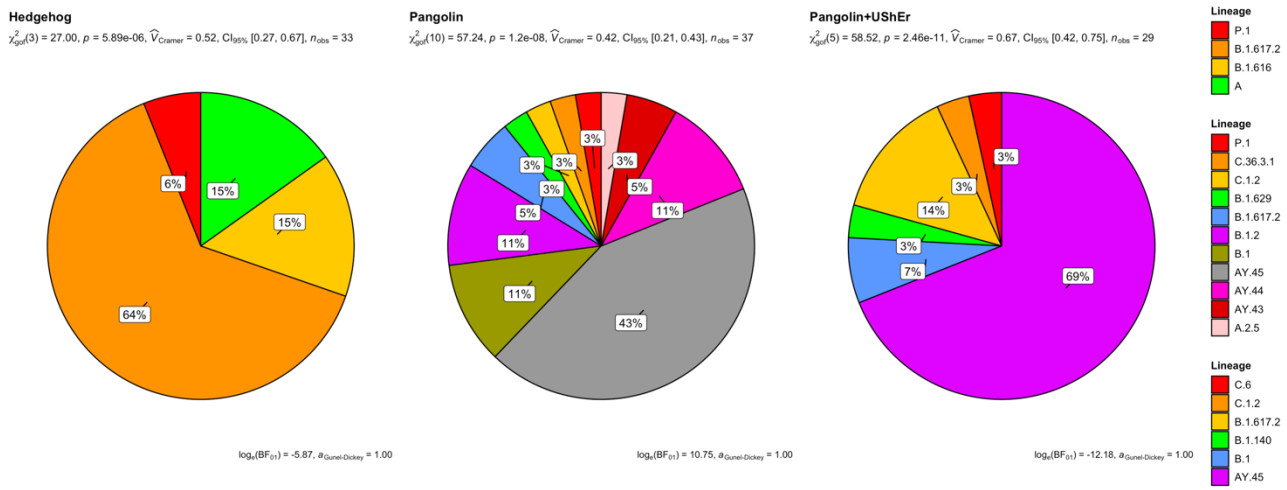


Figure 4-10: Lineage assignment results for Hedgehog, Pangolin and Pangolin+UShEr. The legends follow the pie chart order. High prevalence of Delta variants is clearly evident and supported by the proportion test results in the subtitles.

4.3.3 Phylogenetic analysis of SARS-CoV-2 lineages

SARS-CoV-2 consensus sequences obtained from the samples passing the Pangolin filtering standards were aligned in combination with the SARS-CoV-2 reference (MN908947.3) and representative sequences of the Alpha, Beta and Delta variants. The multiple sequence alignment included 42 sequences of which 4 were references and the other 38 consensus sequences passing the Pangolin quality filtering as detailed in the section above. The inferred maximum likelihood phylogeny tree was annotated using the lineage assignment of Pangolin (Figure 4-11), Pangolin+UShEr (Figure 4-12) and Hedgehog (Figure 4-13) and location sampled. The tree produced is used in all three of the figures with the annotation based on different lineage assignment results. The tree is rooted at the SARS-CoV-2 reference genome (MN908947.3). The results are in agreement with the date sampled and expected variants. Samples sampled from mid-June 2021 until November 2021 grouped with the Delta variant reference. These dates align with the high incidence of the Delta variant in South Africa. A sample collected mid-March 2021 clustered with the Beta variant and this is again in agreement with the incidence of the Beta variant during this period. It should be emphasized that a different may be obtained with the inclusion of additional SARS-CoV-2 variants and genomes.

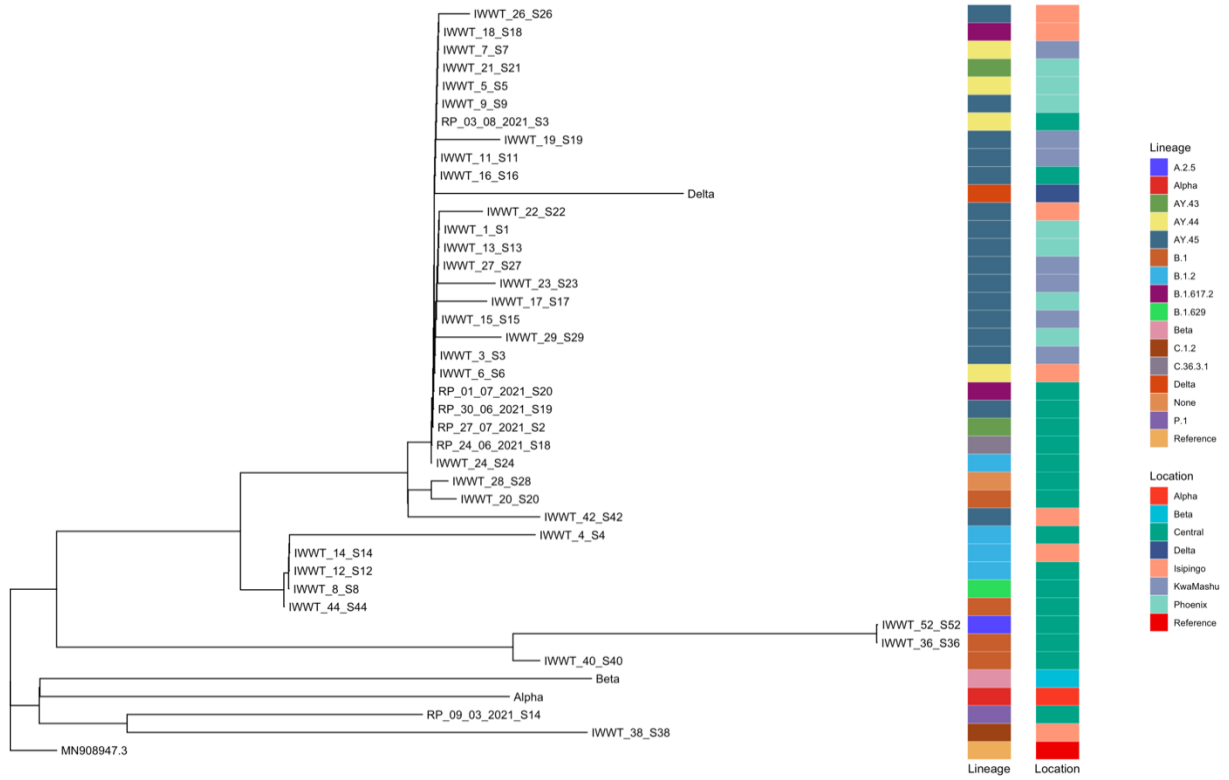


Figure 4-11: Maximum likelihood phylogenetic tree annotated with Pangolin lineage assignment.

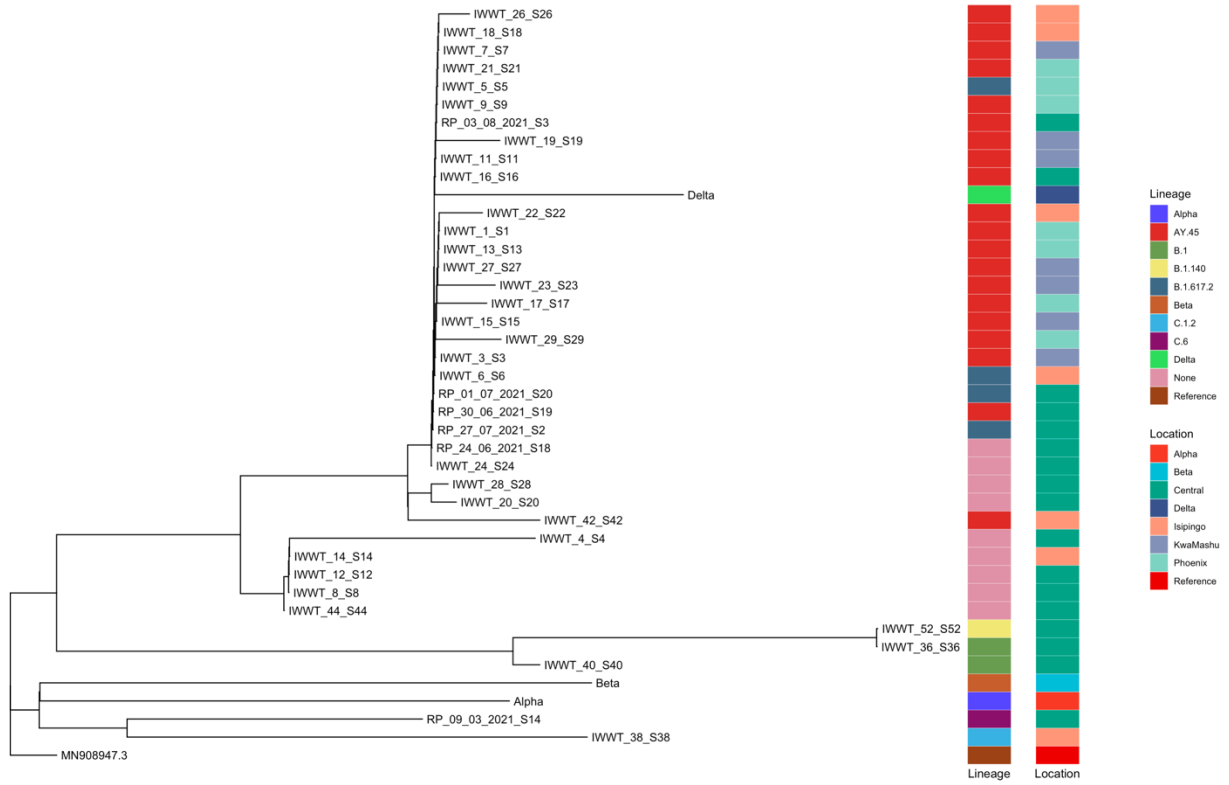


Figure 4-12: Maximum likelihood phylogenetic tree annotated with Pangolin+UShEr lineage assignment.

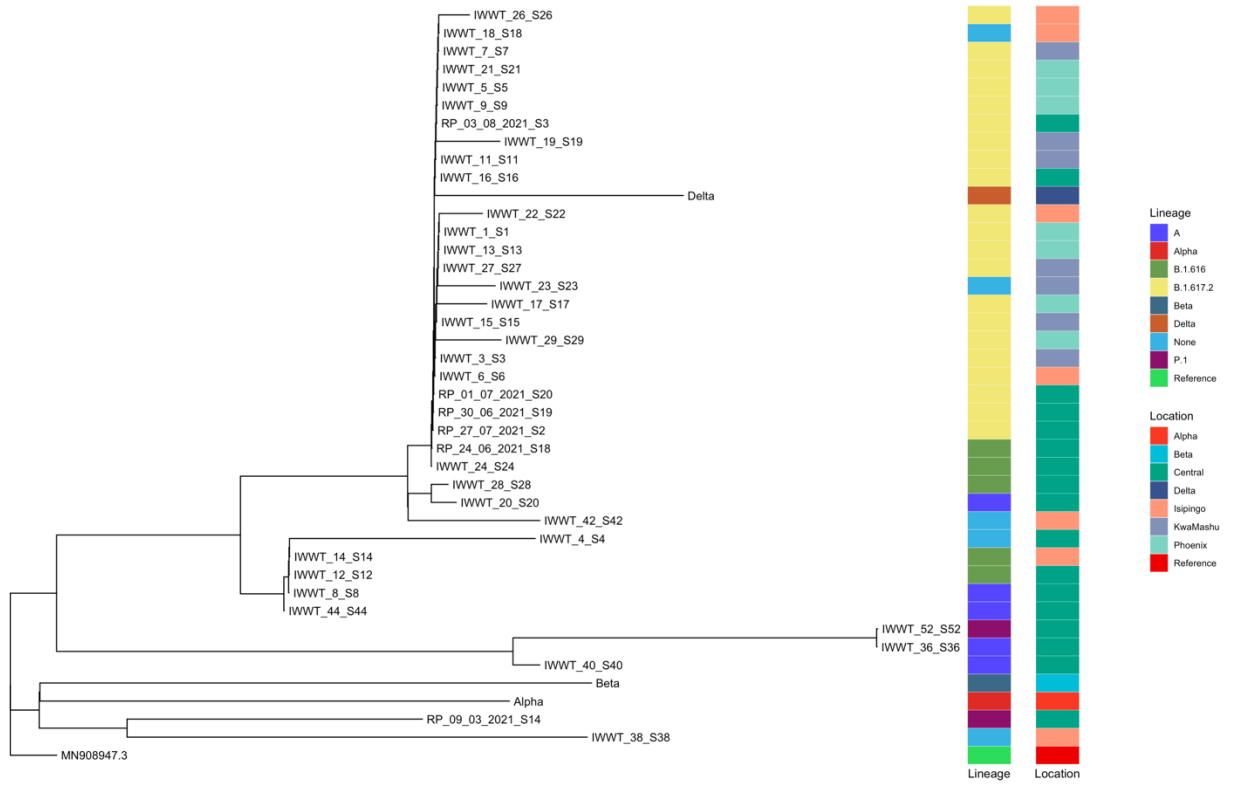


Figure 4-13: Maximum likelihood phylogenetic tree annotated with Hedgehog lineage assignment.

4.3.4 SARS-CoV-2 *de novo* assembled genomes from wastewater samples

All SARS-CoV-2 WGS samples were *de novo* assembled with varying results. Near complete SARS-CoV-2 genomes were based on *de novo* assemblies producing a singular scaffold with a length of larger than 29,000 bp as the length of the reference SARS-CoV-2 genome is 29,903 bp and consists of a single scaffold. The *de novo* assemblies of 3 samples adhered to this criterion. Samples RP_27_07_2021_S2, IWWT_24_S24 and IWWT_36_S36 passed the above-mentioned filter (Table 4-7). The quality assessment reports for RP_27_07_2021_S2 (Figure 4-14), IWWT_24_S24 (Figure 4-15) and IWWT_36_S36 (Figure 4-16) detail the near completeness of the *de novo* assemblies. The ability to construct near complete genomes is critical in lineage assignment, phylogenetic analysis and the identification of current or novel variants which include variants of concern. Given the highly diverse sample conditions from which SARS-CoV-2 WGS data was produced near complete *de novo* assemblies were not expected. In essence, each sample consists of numerous SARS-CoV-2 entities which will drastically complicate the construct of near complete genomes.

Table 4-7: Near complete *de novo* assembled SARS-CoV-2 genome from wastewater.

Sample ID	Date	Location	Pangolin	Usher	Hedgehog	Reference Coverage (%)	Assembly Length
RP_27_07_2021_S2	2021/07/27	Central	AY.43	B.1.617.2	B.1.617.2	100.0000	29,920 bp
IWWT_24_S24	2021/09/14	Central	B.1.2	None	B.1.616	100.0000	30,044 bp
IWWT_36_S36	2021/10/07	Central	B.1	B.1	A	100.0000	29,953 bp

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	RP_27_07_2021_S2
# contigs	1
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
# contigs (>= 5000 bp)	1
# contigs (>= 10000 bp)	1
# contigs (>= 25000 bp)	1
# contigs (>= 50000 bp)	0
Largest contig	29 920
Total length	29 920
Total length (>= 0 bp)	29 920
Total length (>= 1000 bp)	29 920
Total length (>= 5000 bp)	29 920
Total length (>= 10000 bp)	29 920
Total length (>= 25000 bp)	29 920
Total length (>= 50000 bp)	0
N50	29 920
N75	29 920
L50	1
L75	1
GC (%)	37.97
Mismatches	
# N's	0
# N's per 100 kbp	0

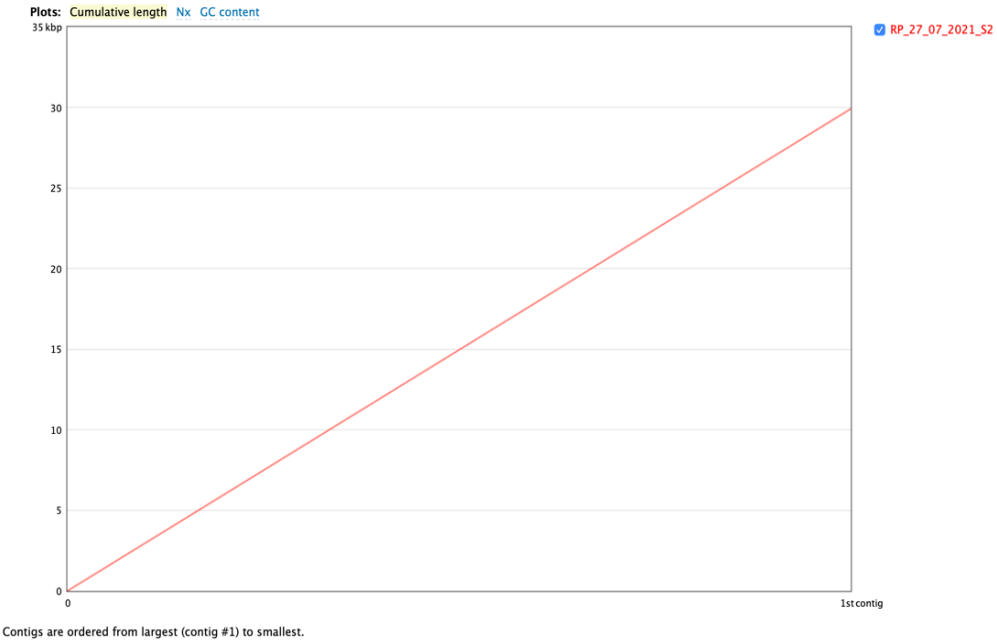


Figure 4-14: Quality assessment of *de novo* assembly for sample RP_27_07_2021_S2.

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	IWWT_24_S24
# contigs	1
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
# contigs (>= 5000 bp)	1
# contigs (>= 10000 bp)	1
# contigs (>= 25000 bp)	1
# contigs (>= 50000 bp)	0
Largest contig	30 044
Total length	30 044
Total length (>= 0 bp)	30 044
Total length (>= 1000 bp)	30 044
Total length (>= 5000 bp)	30 044
Total length (>= 10000 bp)	30 044
Total length (>= 25000 bp)	30 044
Total length (>= 50000 bp)	0
N50	30 044
N75	30 044
L50	1
L75	1
GC (%)	38.03
Mismatches	
# N's	0
# N's per 100 kbp	0

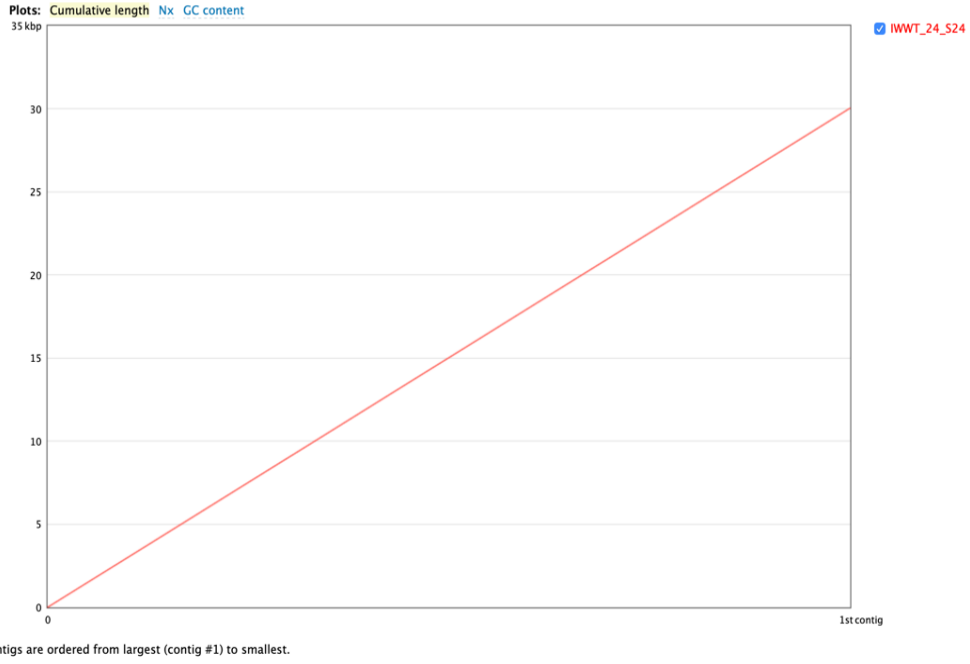


Figure 4-15: Quality assessment of *de novo* assembly for sample IWWT_24_S24.

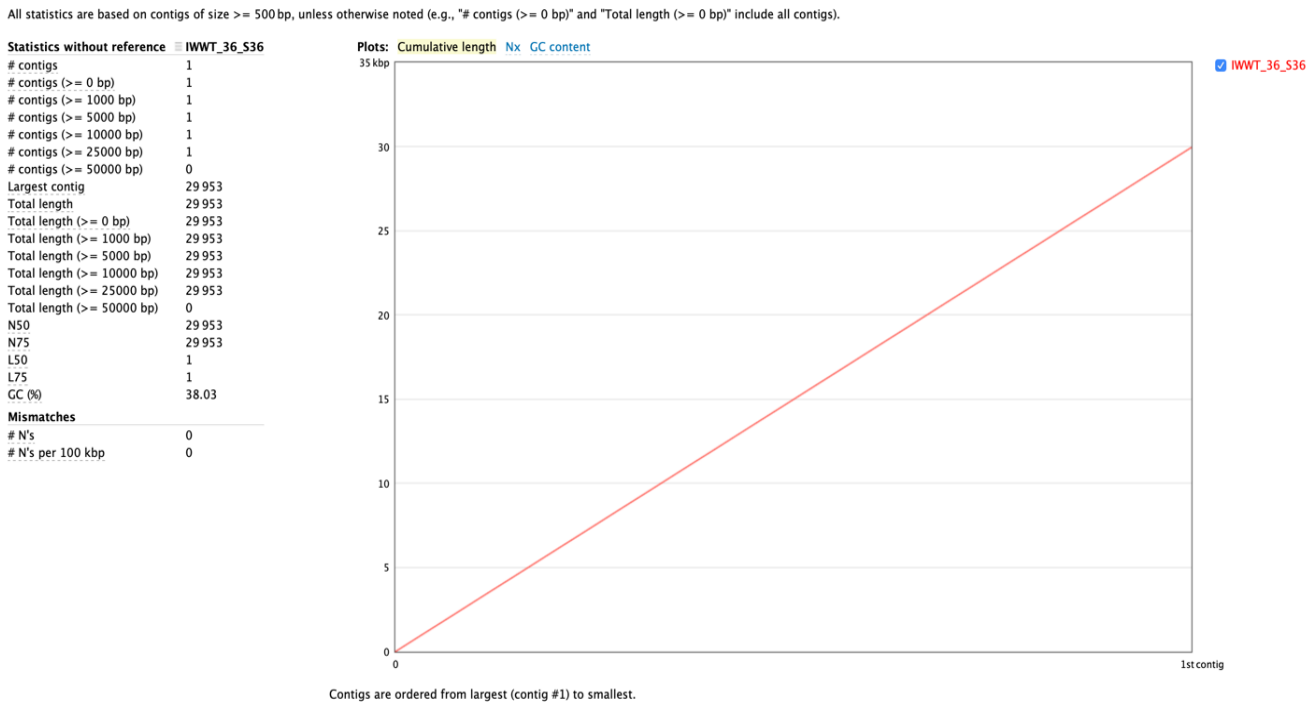


Figure 4-16: Quality assessment of *de novo* assembly for sample IWWT_36_S36.

4.4 DISCUSSION

The COVID-19 pandemic and associated detection of Variants of Concern highlighted the power and application of next-generation sequencing in epidemiology. The unprecedented rate at which SARS-CoV-2 genomes was sequenced allowed researchers to rapidly detect novel variants and identify the functional changes brought forth by the genomic variations. Protocols to rapidly yet adequately produce SARS-CoV-2 genomic information by means of whole genome sequencing have developed in leaps and bounds in a very short period of time. One positive outcome of the COVID-19 pandemic would be the demonstration and global acceptance of whole genome sequencing as the gold standard in SARS-CoV-2 research.

In this report a total of 73 wastewater samples from 4 different areas in Durban, KwaZulu-Natal, were subjected to SARS-CoV-2 whole genome sequencing. After RNA extraction NEBNext ARTIC kits were used produce whole genome sequencing data. The amount of raw data produced was more than ample although certain samples performed poorly. The reason for this poor performance was investigated and concluded that the age of a sample, period between RNA extraction and sequencing, was a vital component in the production of high coverage sequencing data. It is therefore critical to regularly sample and rapidly sequence in an attempt to circumvent the degradation of RNA as was seen in this report. Data loss between the various quality control and decontamination steps was as expected. It was found that the protocol used produced good quality data and that the NEBNext ARTIC kit was able to target and amplify SARS-CoV-2 genomes even from highly diverse and contaminated samples such as wastewater.

The ARCTIC analysis pipeline and subsequent lineage assignment protocol was able to assign variant annotation to more than half of the samples. The inability to assign lineages was predominantly due to low coverage of the SARS-CoV-2 genome as not enough data was produced. This was found to be caused by the age of the sample where the period between RNA extraction and sequencing was prolonged. For some of the samples with an adequate amount of sequencing data no lineage assignment was possible. This was more than likely due to a highly diverse sample which contained various SARS-CoV-2 variants. This leads to the

lineage assignment algorithm not being able to ascertain the presence of specific mutations and due to this unreliability not able to assign a SARS-CoV-2 lineage.

A variety of different variants were detected across all the samples. These were found to be in accordance to the expected Variant of Concern as was prevailing at the period sampled in South Africa. The detection of the Beta variant in a sample from early March 2021 and that of numerous Delta variants in samples collected after July 2021 is in accordance with the clinical data. Of particular interest was the difference in variant diversity detected between sampling locations. Samples from Central Durban and Isipingo displayed a number of different variants whereas KwaMashu and Phoenix were found to be more homogenous with regards to variants. This may be due to geographical or social factors which could influence the diversity of the variants present in a sample. The possible reasons for this still need further investigation.

The phylogenetic analysis was in general agreement with the lineage classifications and expected variants as per date sampled. It should be stated that phylogenetic analysis such as included here are continuous evolving as more sequences are added and included in the maximum likelihood phylogenetic analyses.

If particular interest was the ability to construct near complete SARS-CoV-2 genomes from the wastewater samples. The *de novo* assembly results included 3 near complete genomes consisting of a single scaffold with adequate length. This is of importance as these samples are highly diverse and contaminated. The ability to construct near complete genomes is of critical importance in the detection of current and novel variants and greatly contributes to the global understanding of SARS-CoV-2.

This report clearly illustrates the ability and applicability of SARS-CoV-2 whole genome sequencing in wastewater-based epidemiology. By means of whole genome sequencing we were able to adequately assign SARS-CoV-2 lineages to more than half of the samples. Furthermore, it was possible to construct near complete SARS-CoV-2 genomes from the data produced for 3 of the samples. It was found that the time span between RNA extraction and sequencing is of critical importance and greatly influences the sequencing performance.

The SARS-CoV-2 virus is continuously evolving as is evident with the continuous characterization of new variants including Variants of Concern. The implementation of wastewater-based epidemiology and in particular SARS-CoV-2 whole genome sequencing of wastewater samples is an eloquent method to firstly detect the presence and secondly characterize the variants present in communities.

CHAPTER 5: VIRAL CONCENTRATED RNA METAGENOMIC SEQUENCING OF WASTEWATER SAMPLES POSITIVE FOR SARS-COV-2 FROM DURBAN, KWAZULU-NATAL

5.1 INTRODUCTION

SARS-CoV-2 and the COVID-19 pandemic has clearly indicated the need for routine viral surveillance in an attempt to prevent any future outbreaks. Wastewater-based epidemiology is an eloquent and robust method to do broad-scale surveillance with a quick turnaround time. The COVID-19 pandemic has increased awareness regarding the power and resolution of next-generation sequencing and genomics. This has been evident in the detection of SARS-CoV-2 variants and the tracking of COVID-19 infection. Wastewater-based epidemiology is a critical component in the detection and tracking of SARS-CoV-2 and it has been shown that sequencing of viral concentrations and RNA extracted directly from wastewater can identify multiple SARS-CoV-2 genotypes, including variants not yet observed in clinical sequencing programmes (Crits-Christoph et al., 2021).

Next-generation sequencing analysis of wastewater samples provides insights to human health related factors which includes the distribution of pathogens and antibiotic resistance genes (Yang et al., 2014). The contents of a wastewater sample provide researchers and stakeholders a glimpse as to what is circulating in the host associated environment and as such host health. Wastewater samples may be regarded as a pooled version of the human gut microbiome. Pathogens and antimicrobial resistance which are present in a wastewater sample may be presumed to have been present in the population gut microbiome prior to the sampling. The sewage water accurately reflects a population's gut microbial composition which therefor allows metagenomics and targeted whole genome sequencing to assist in obtaining information regarding the infection dynamics in a given population (Fresia et al., 2019).

The virome is defined as the collection of all viruses found within a particular environment and is described using metagenomic sequencing and the appropriate bioinformatic tools. Viruses are the most abundant and diverse entities on earth and comprise of viruses that infect bacteria, other cellular organisms and eukaryotes (Liang and Bushman, 2021). Metagenomic sequencing has the potential to detect any viral genomic material in a sample without prejudice (Niewenhuijse et al., 2020). This enables rapid and robust detection of any potentially harmful viruses in the community in a single data generation event. It is possible to detect numerous viral entities without the need to purify and isolate individually. This method is able to identify and functionally profile viruses in wastewater and further allows for viral discovery and the study of viral dynamics (Gulino et al., 2020).

In the sections below, we clearly outline the methodology used and results obtained in the metagenomic sequencing of the virome from wastewater samples collected during the period 25 August 2020 to 3 August 2021 from the Durban region in KwaZulu-Natal. The results illustrate the functionality, benefits and potential of metagenomic sequencing of RNA extracted after viral concentration using ultra (centricon) filtration. The wealth of information obtained per sample will greatly assist in creating a baseline wastewater virome. This information will be crucial in future studies and aid in detecting fluctuations in a system which is indicative of human health.

5.2 MATERIALS AND METHODS

Samples (n=17) were collected from the Central wastewater treatment site in Durban, KwaZulu-Natal by DUT IWWT (Prof F. Bux). The samples were collected between 25 August 2020 and 3 August 2021 (Table 5-1). These samples all tested positive for the presence of SARS-CoV-2. RNA extractions were done by the DUT IWWT after viral concentration using ultra (centricon) filtration and the resulting extractions delivered to the ARC Biotechnology for library preparation and metagenomic sequencing (Supplementary Sequencing Quotation). The resulting libraries were sequenced on a HiSeq 2500 with roughly 5 GB of data per sample requested.

Table 5-1: Samples received for concentrated viral RNA metagenomic sequencing.

Sample ID	Date	Location
RP25_08_2020	2020/08/25	Central
RP29_09_2020	2020/09/29	Central
RP15_12_2020	2020/12/15	Central
RP29_12_2020	2020/12/29	Central
RP19_01_2021	2021/01/19	Central
RP26_01_2021	2021/01/26	Central
RP02_02_2021	2021/02/02	Central
RP23_02_2021	2021/02/23	Central
RP09_03_2021	2021/03/09	Central
RP30_03_2021	2021/03/30	Central
RP08_04_2021	2021/04/08	Central
RP13_04_2021	2021/04/13	Central
RP24_06_2021	2021/06/24	Central
RP30_06_2021	2021/06/30	Central
RP01_07_2021	2021/07/01	Central
RP27_07_2021	2021/07/27	Central
RP03_08_2021	2021/08/03	Central

Initial sequence data quality and filtered data quality was inspected using FastQC version 0.11.8 (Andrews, S., 2010). Sequence data was quality trimmed and filtered, including adapter removal and decontamination, using BBDuk version 38.91 available from the BBTools suite of tools (Bushnell, B., 2014). Human contamination in the quality filtered sequencing data was removed by aligning the sequence data against the latest reference human genome (GRCh38.p13) using BBMap version 38.91, available from BBTools.

To identify the portion of the SARS-CoV-2 genome sequenced, filtered and decontaminated paired-end reads were aligned to the SARS-CoV-2 reference genome (MN908947.3) with BBMap and coverage statistics calculated.

Taxonomic classification of the filtered and decontaminated sequencing data was done using Kaiju version 1.8.0 (Menzel et al., 2016) and the Kaiju formatted refseq database as available on 2021/02/26. The Kaiju formatted refseq database contains complete assembled and annotated reference genomes of Archaea, Bacteria, and viruses from the NCBI RefSeq database (O'Leary et al., 2016).

Additional data analysis and visualization was done in R version 4.1.2 (Team, R Core, 2020) implemented in RStudio version 1.4.1717 (Team, RStudio, 2021) with added library ggstatsplot library (Patil, I., 2021).

5.3 RESULTS

5.3.1 Data quality filtering and decontamination

Approximately 80 GB worth of raw sequencing data was produced for the 17 samples. The raw sequencing data was quality filtered and the resulting sequence quality of the filtered reads were again inspected using FastQC. Sequencing data which mapped to the human genome was removed and the quality of the remaining sequence data again quality checked with FastQC. The number of reads for each sample is presented in Table 5-2 and Figure 5-1.

Table 5-2: Number of reads at each stage of quality control and decontamination.

Sample ID	Raw Reads	QC Reads	No Human QC Reads
RP25_08_2020	22,835,770	21,771,515	21,642,781
RP29_09_2020	20,606,005	19,603,927	19,536,310
RP15_12_2020	23,169,154	20,726,855	20,513,751
RP29_12_2020	1,110	908	907
RP19_01_2021	16,018,105	13,927,558	13,770,107
RP26_01_2021	13,726,764	11,481,766	11,422,988
RP02_02_2021	19,659,049	18,113,815	18,071,149
RP23_02_2021	11,897,432	9,409,646	9,335,900
RP09_03_2021	34,303,692	32,550,232	32,539,331
RP30_03_2021	37,005,595	35,121,766	35,081,227
RP08_04_2021	35,962,856	33,848,418	33,807,610
RP13_04_2021	20,295,971	17,712,417	17,707,345
RP24_06_2021	32,857,585	30,889,860	30,869,611
RP30_06_2021	30,761,843	28,937,351	28,921,671
RP01_07_2021	32,785,293	31,303,985	31,283,229
RP27_07_2021	33,818,850	32,680,271	32,656,154
RP03_08_2021	33,331,403	32,312,085	32,300,863

Data loss due to quality and contamination was as expected and more than enough reads remained for further analysis. Unfortunately, sample RP29_12_2022 did not produce the expected number of reads and was removed in further analysis. This may be due to various factors and will be investigated. The low levels of data loss after decontamination, i.e. human, clearly illustrates the application of the viral concentration using ultra (centricon) filtration in removing unwanted contamination. This process enables focused sequencing and the viral portion of a sample and little to no data is wastefully expended. On average the raw dataset contained 26,189,710 reads, the quality filtered 24,399,467 reads and the quality filtered decontaminated set 24,341,252 reads. No significant differences in the number of reads between of the processing steps were observed (p-value = 0.79).

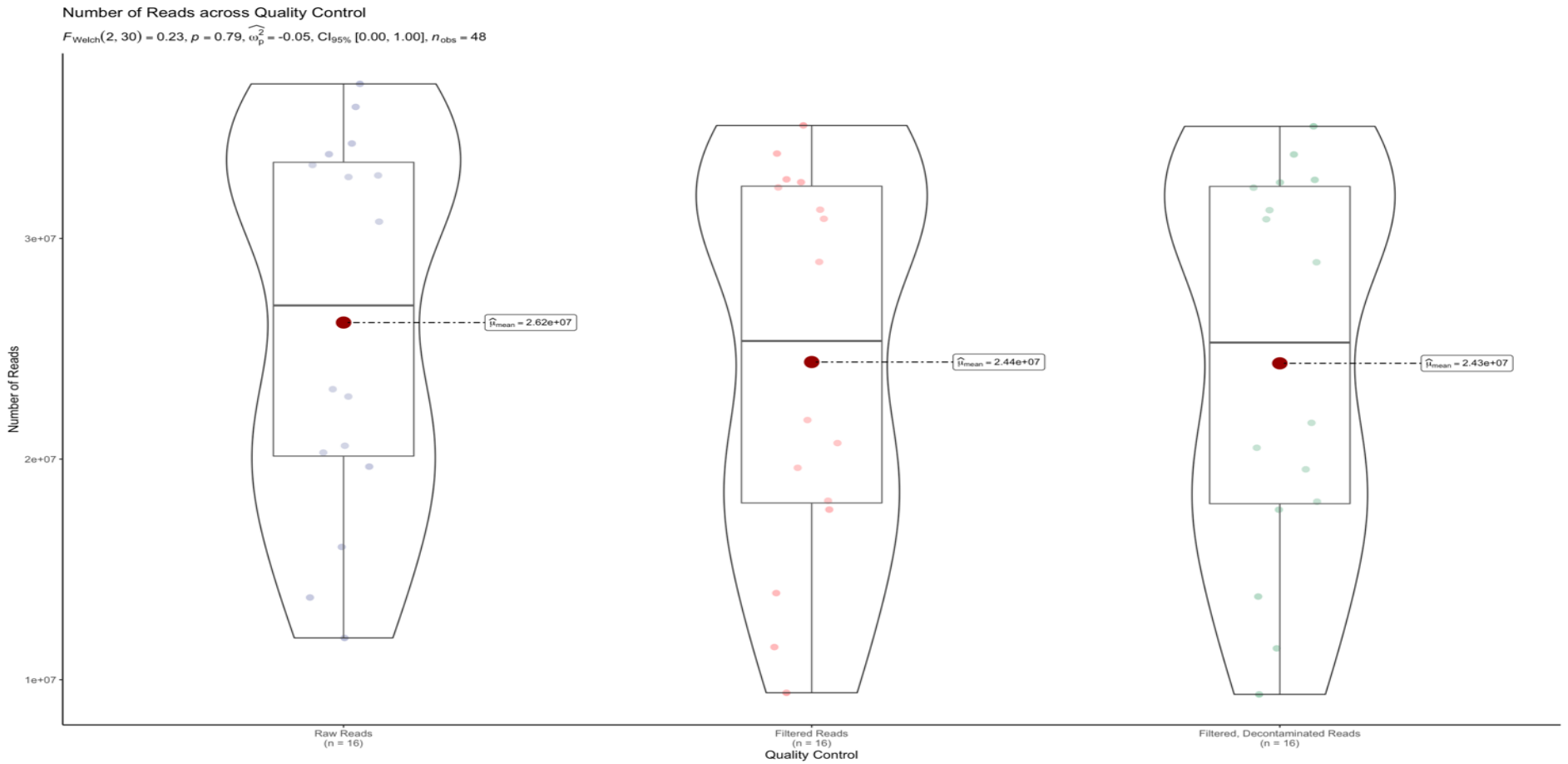


Figure 5-1: Number of reads at each stage of quality control and decontamination. The colours indicate the quality control step, x-axis the sample and y-axis the number of reads. Low levels of data loss were seen and the number of reads surviving quality filtering and human decontamination was more than adequate for the project. No significant differences in the number of reads between any of the processing steps were observed. The results from the statistical test are reported in the subtitles on the top of each graph.

5.3.2 Detection of SARS-CoV-2

The presence of SARS-CoV-2 fragments were detected in all 16 samples, sample RP29_12_2022 excluded due to low read count, using metagenomic sequencing on RNA extracted after viral concentration using ultra (centricon) filtration (Table 5-3 and Figure 5-2). The ability to detect SARS-CoV-2 in all of the samples clearly illustrates the importance of viral concentration when doing metagenomic sequencing for virome detection. As all 16 samples were positive for the presence of SARS-CoV-2 using conventional diagnostics the detection of SARS-CoV-2 genomic segments by means of metagenomic sequencing is clearly illustrated in the table above. Of particular interest was the high percentage SARS-CoV-2 genome coverage in sample RP27_07_2021, sampled 2021/07/27. This sample was collected during the peak of the third wave in South Africa and the high percentage coverage may be indicative of the high viral load in the sample.

Table 5-3: Number of paired-end reads at each stage of quality control and decontamination.

Sample ID	Collection Date	Collection Site	Reference Covered Percent
RP25_08_2020	2020/08/25	Central	0.5819
RP29_09_2020	2020/09/29	Central	1.3109
RP15_12_2020	2020/12/15	Central	3.7521
RP19_01_2021	2021/01/19	Central	2.5616
RP26_01_2021	2021/01/26	Central	4.2203
RP02_02_2021	2021/02/02	Central	0.9196
RP23_02_2021	2021/02/23	Central	1.6587
RP09_03_2021	2021/03/09	Central	45.2028
RP30_03_2021	2021/03/30	Central	0.1538
RP08_04_2021	2021/04/08	Central	1.6754
RP13_04_2021	2021/04/13	Central	0.1505
RP24_06_2021	2021/06/24	Central	18.2657
RP30_06_2021	2021/06/30	Central	4.5046
RP01_07_2021	2021/07/01	Central	26.0342
RP27_07_2021	2021/07/27	Central	98.9733
RP03_08_2021	2021/08/03	Central	52.2958

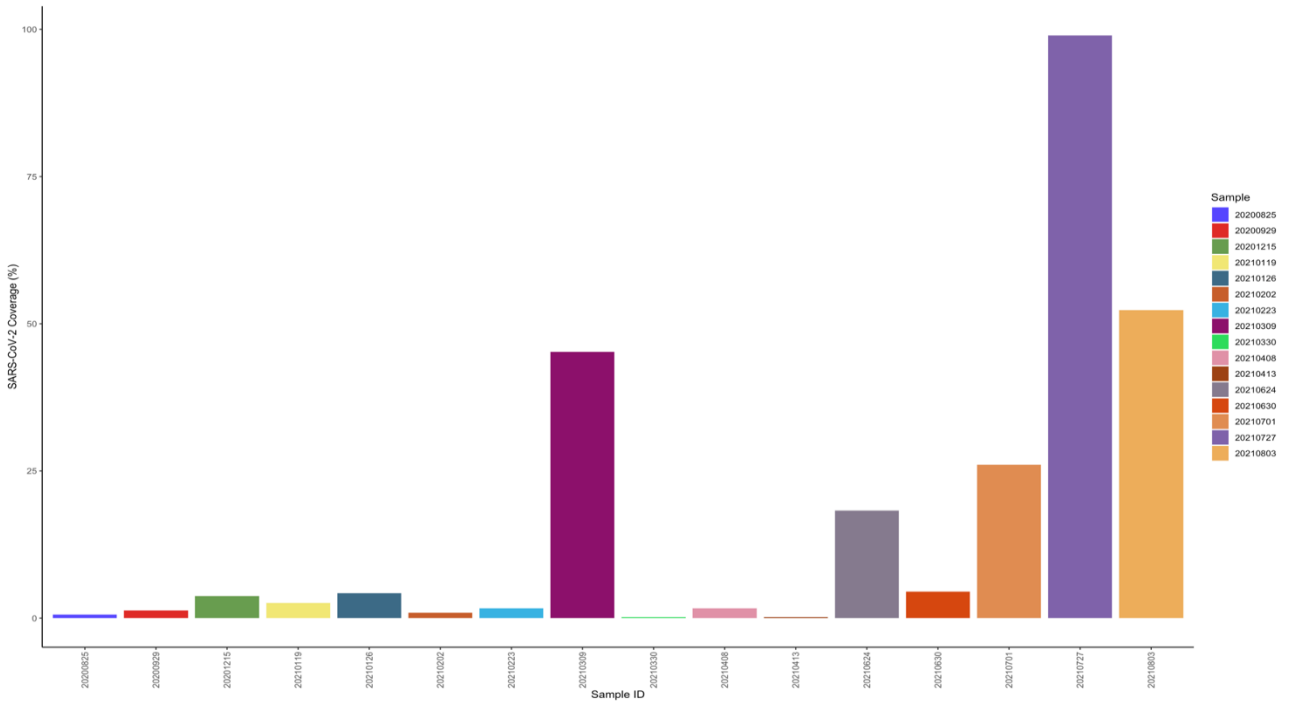


Figure 5-2: Percentage of the SARS-CoV-2 reference genome (NC_045512.2) covered per sample. The colours indicate the sample as represented by sampling date, x-axis the sample as represented by sampling date and y-axis the percentage SARS-CoV-2 reference genome coverage.

5.3.3 Virome taxonomic profile of samples based on unassembled sequencing data

Taxonomic classification as produced by Kaiju using the quality filtered, decontaminated reads indicated a high proportion of Viral paired-end reads in the samples, as was expected.

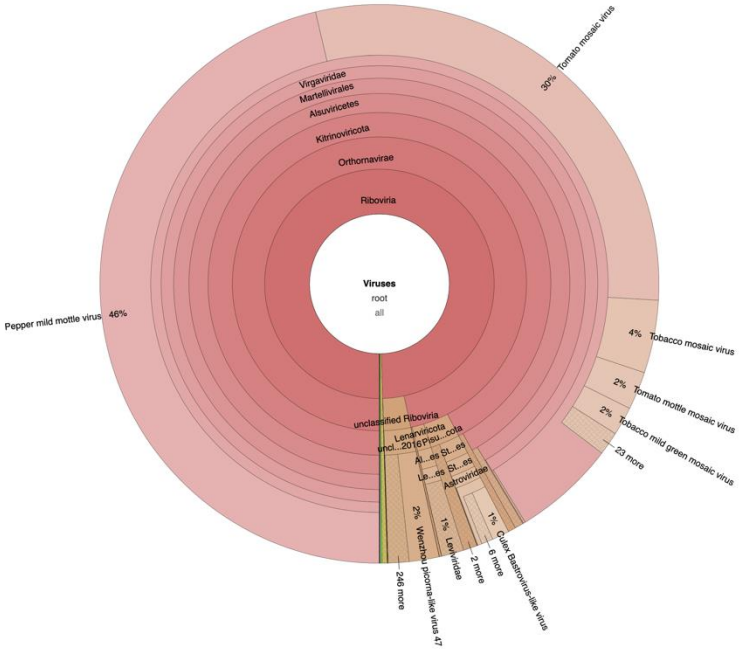


Figure 5-3: Viral taxonomic profile of sample RP25_08_2020, collected 2020/08/25.

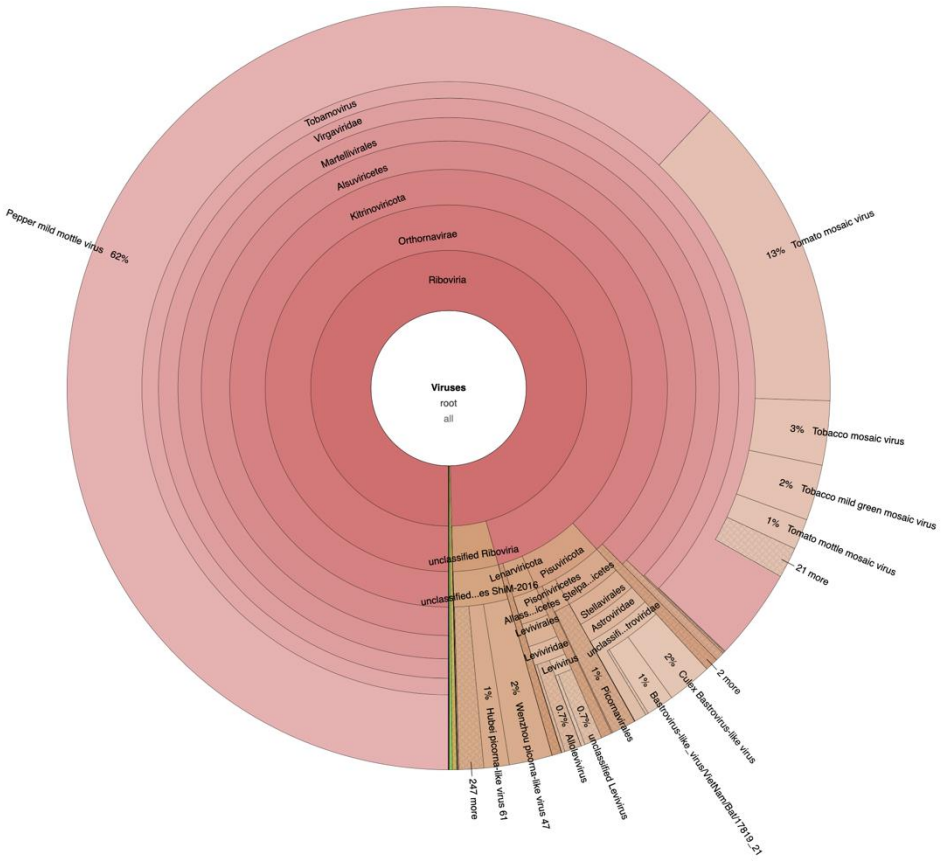


Figure 5-4: Viral taxonomic profile of sample RP29_09_2020, collected 2020/09/29.

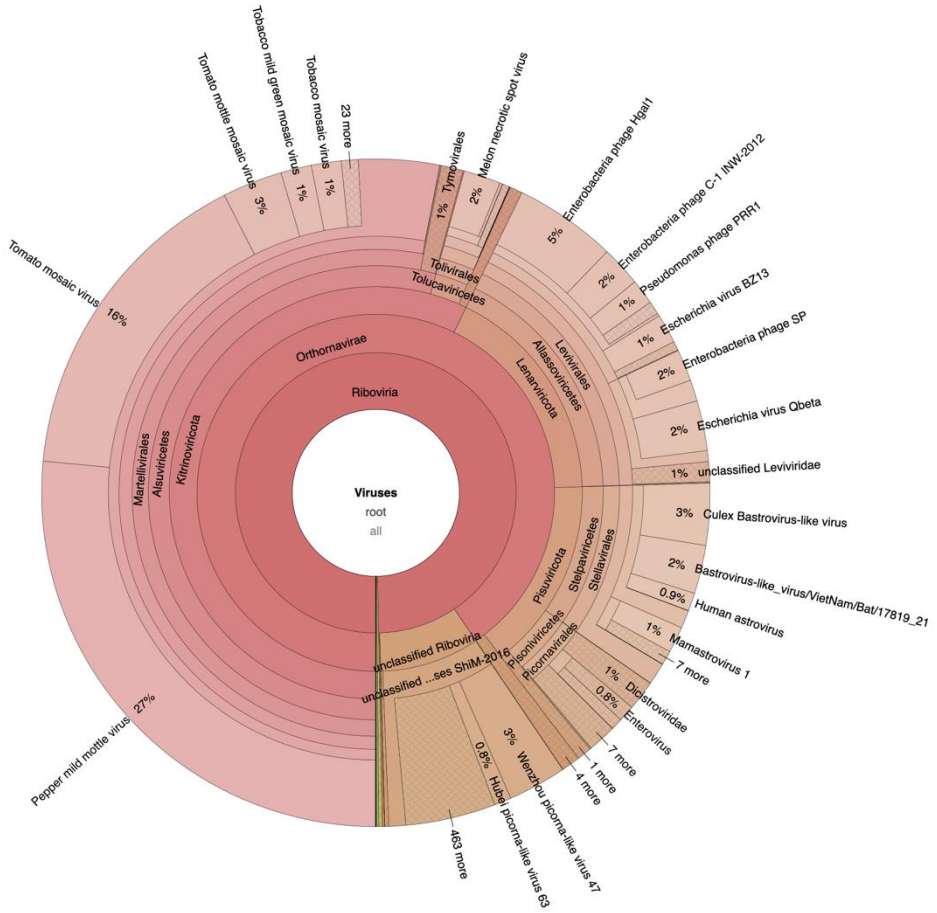


Figure 5-5: Viral taxonomic profile of sample RP15_12_2020, collected 2020/12/15.

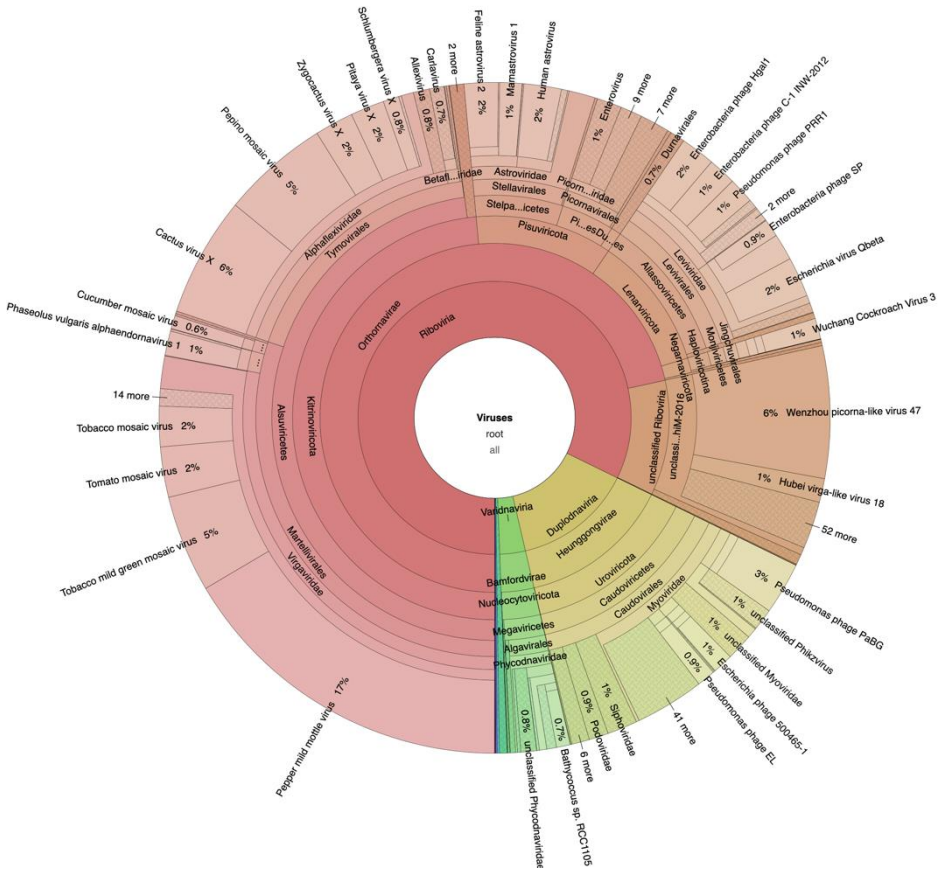


Figure 5-6: Viral taxonomic profile of sample RP19_01_2021, collected 2021/01/19.

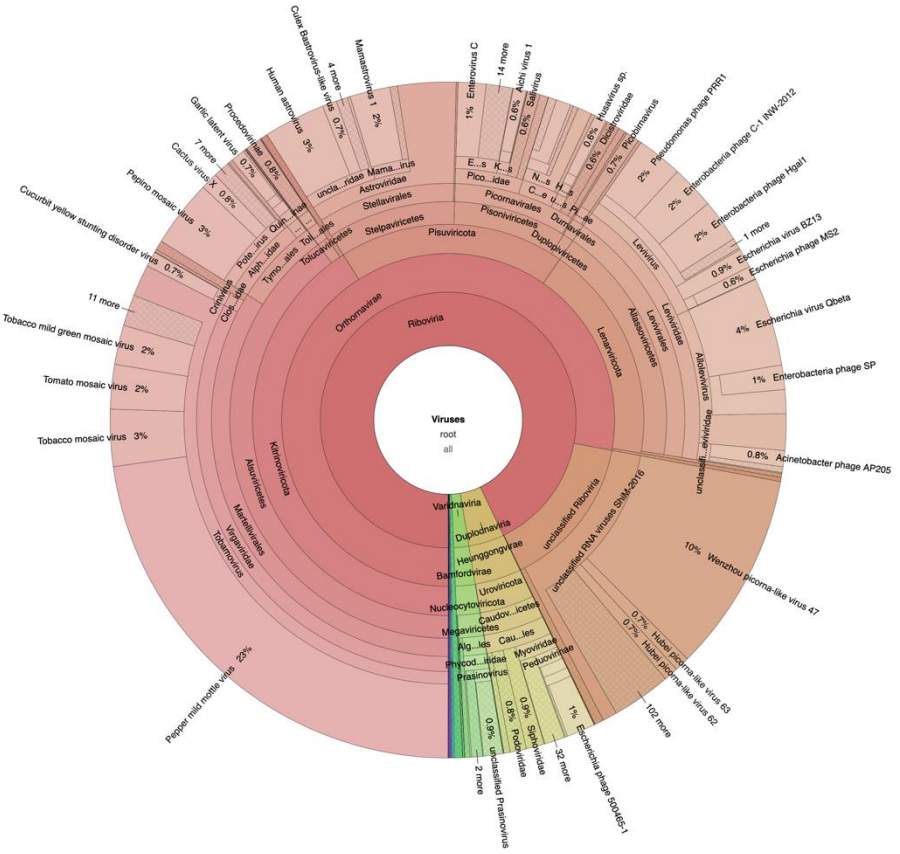


Figure 5-7: Viral taxonomic profile of sample RP26_01_2021, collected 2021/01/26.

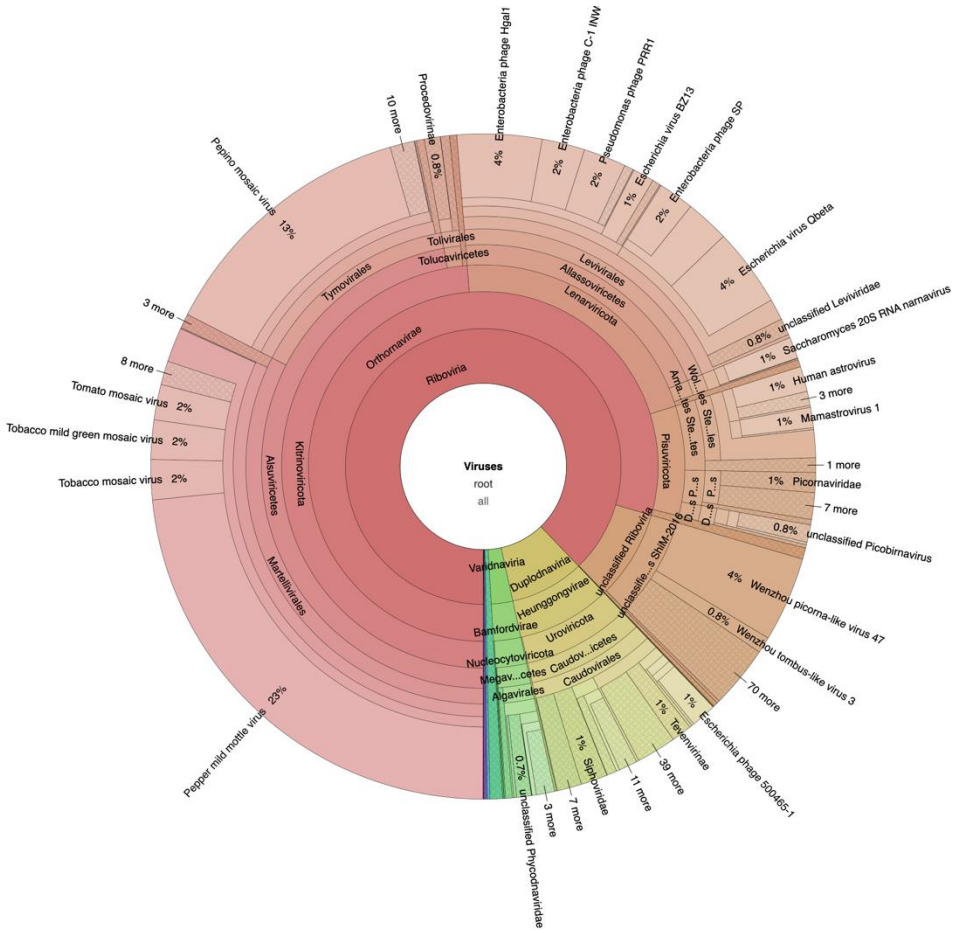


Figure 5-8: Viral taxonomic profile of sample RP02_02_2021, collected 2021/02/02.

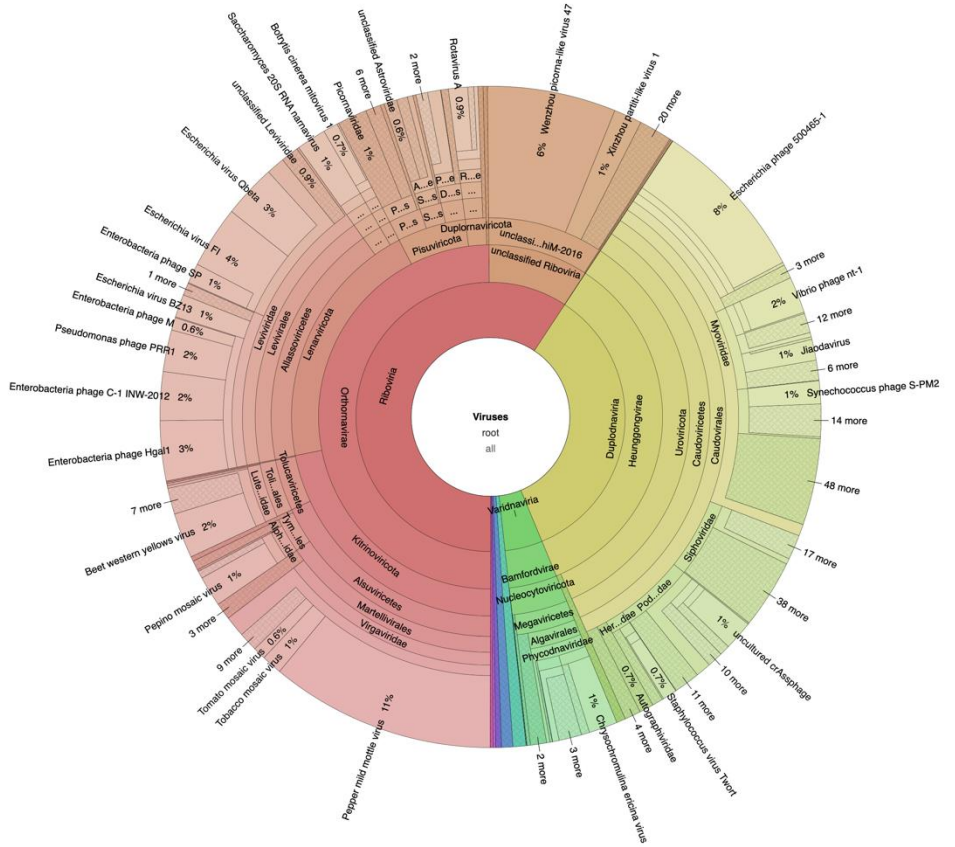


Figure 5-9: Viral taxonomic profile of sample RP23_02_2021, collected 2021/02/23.

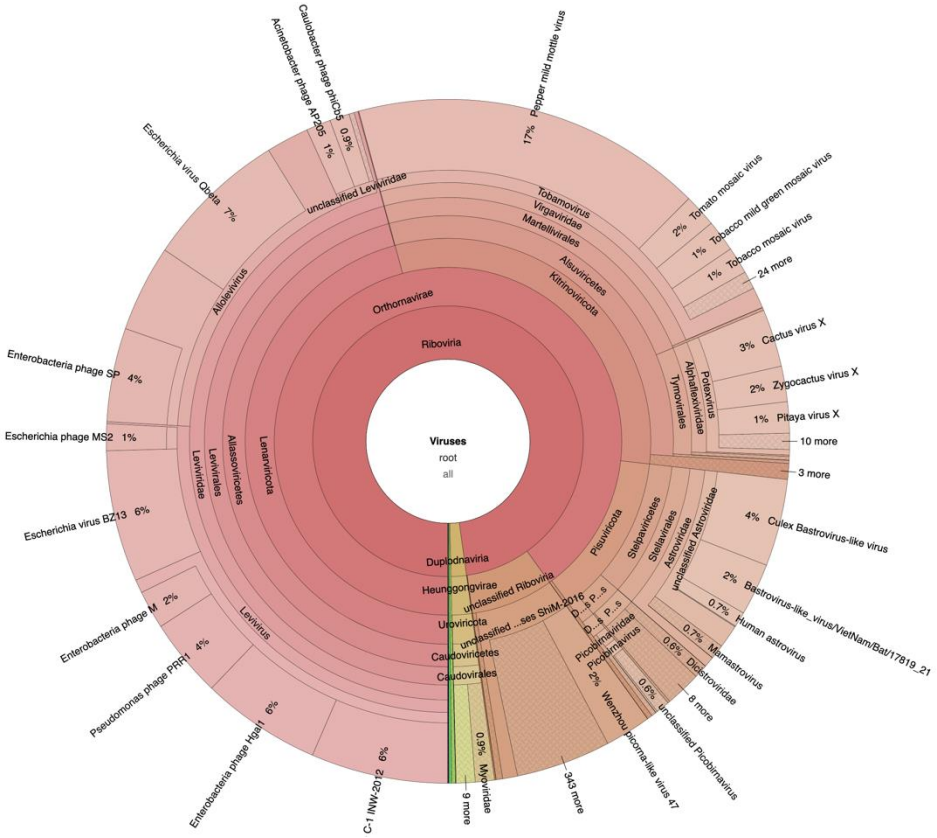


Figure 5-10: Viral taxonomic profile of sample RP09_03_2021, collected 2021/03/09.

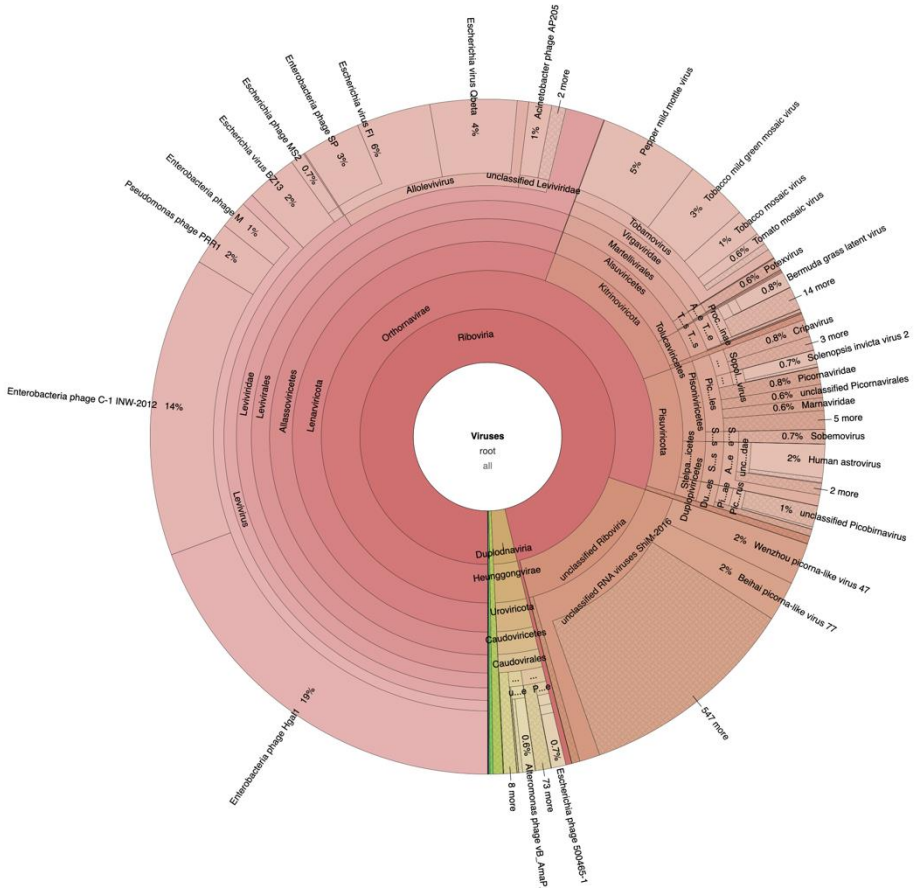


Figure 5-11: Viral taxonomic profile of sample RP30_03_2021, collected 2021/03/30.

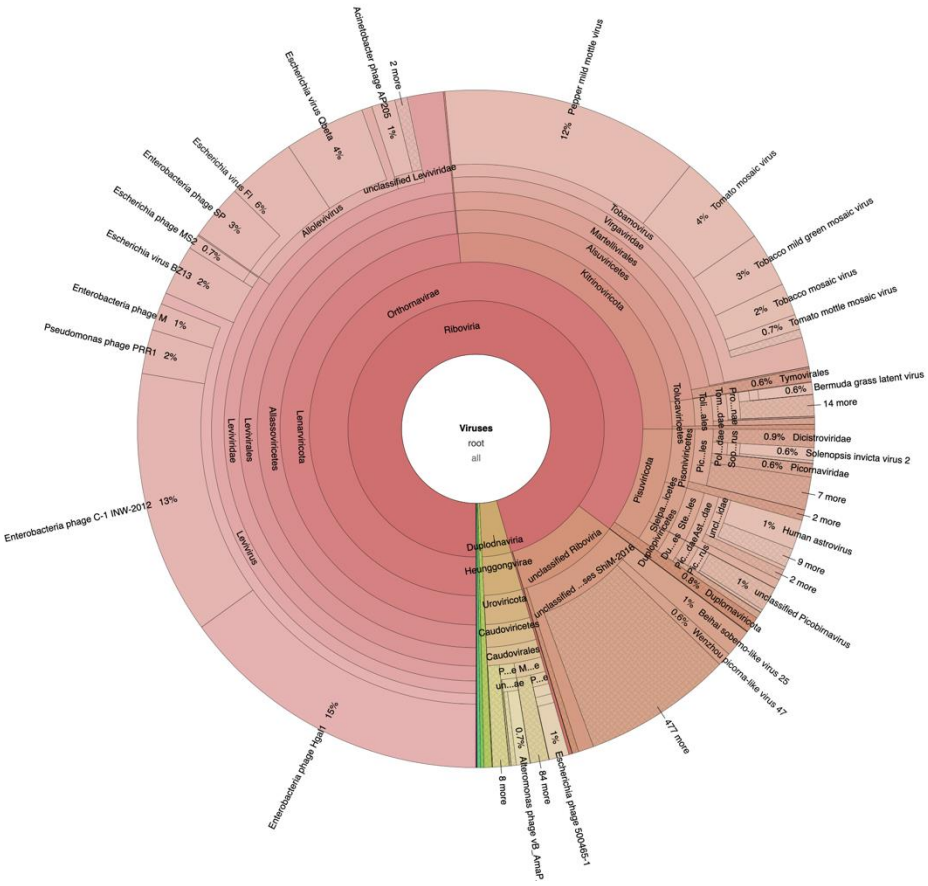


Figure 5-12: Viral taxonomic profile of sample RP08_04_2021, collected 2021/04/08.

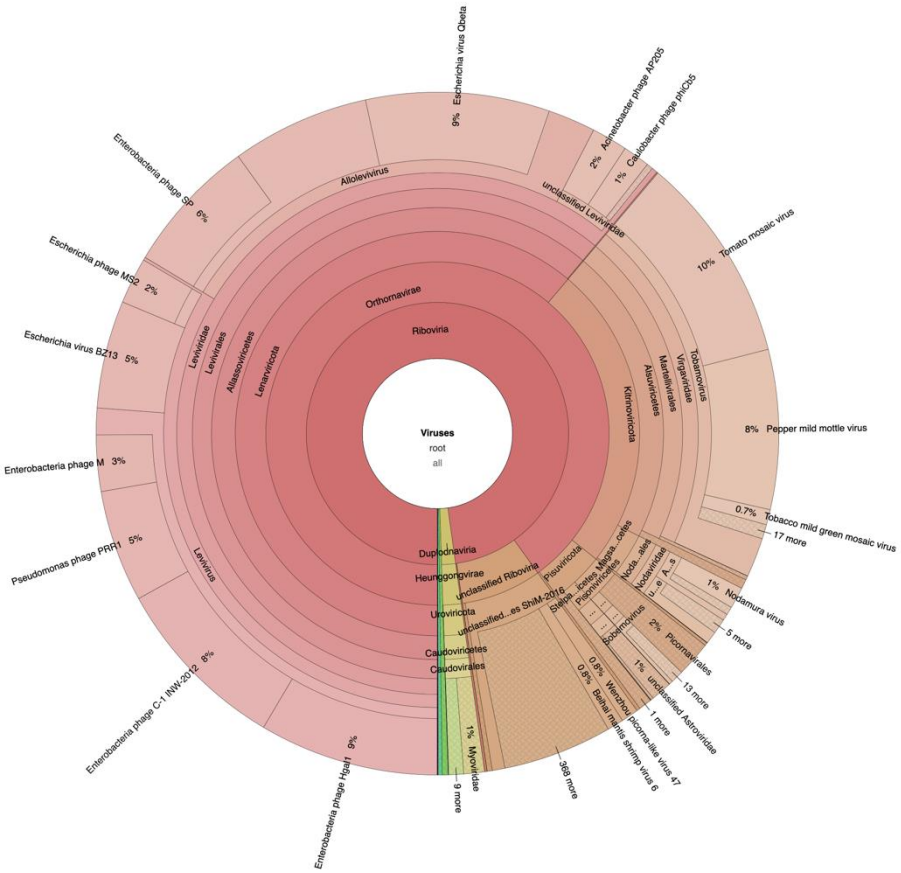


Figure 5-13: Viral taxonomic profile of sample RP13_04_2021, collected 2021/04/13.

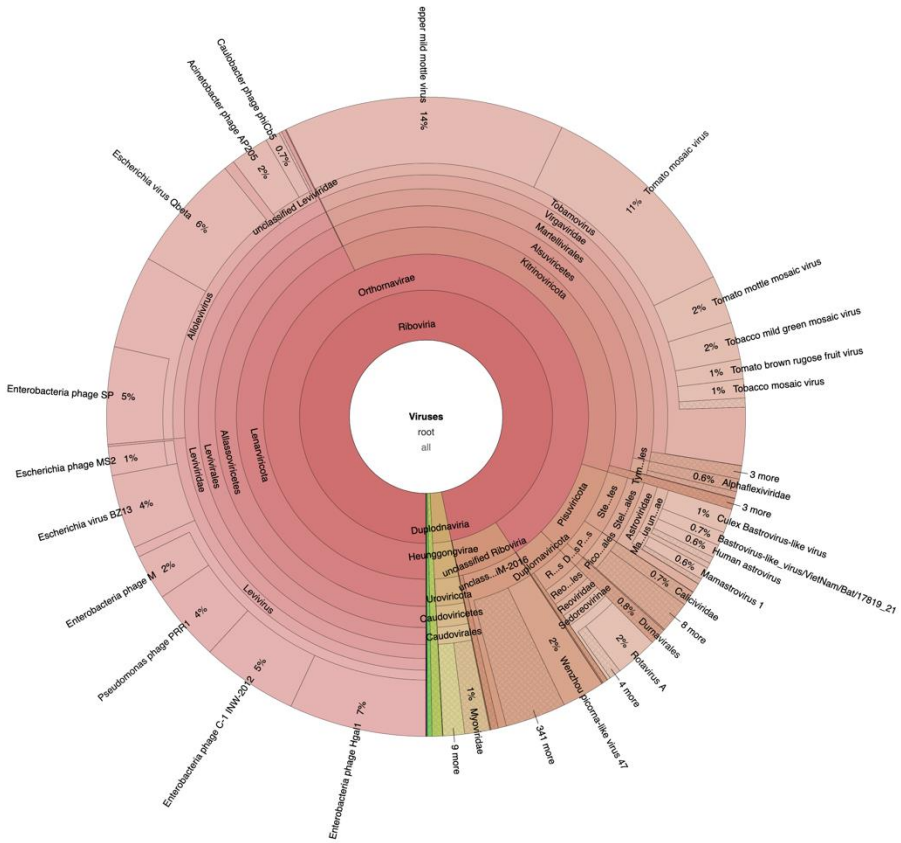


Figure 5-14: Viral taxonomic profile of sample RP24_06_2021, collected 2021/06/24.

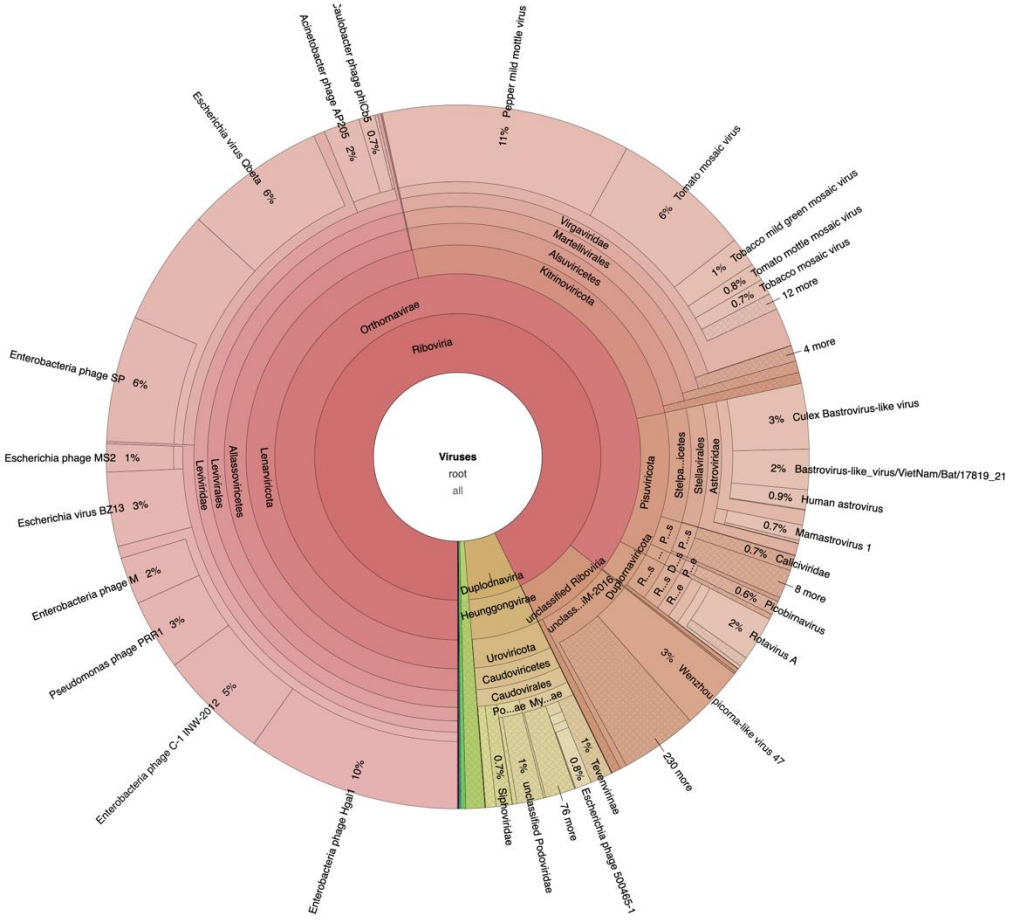


Figure 5-15: Viral taxonomic profile of sample RP30_06_2021, collected 2021/06/30.

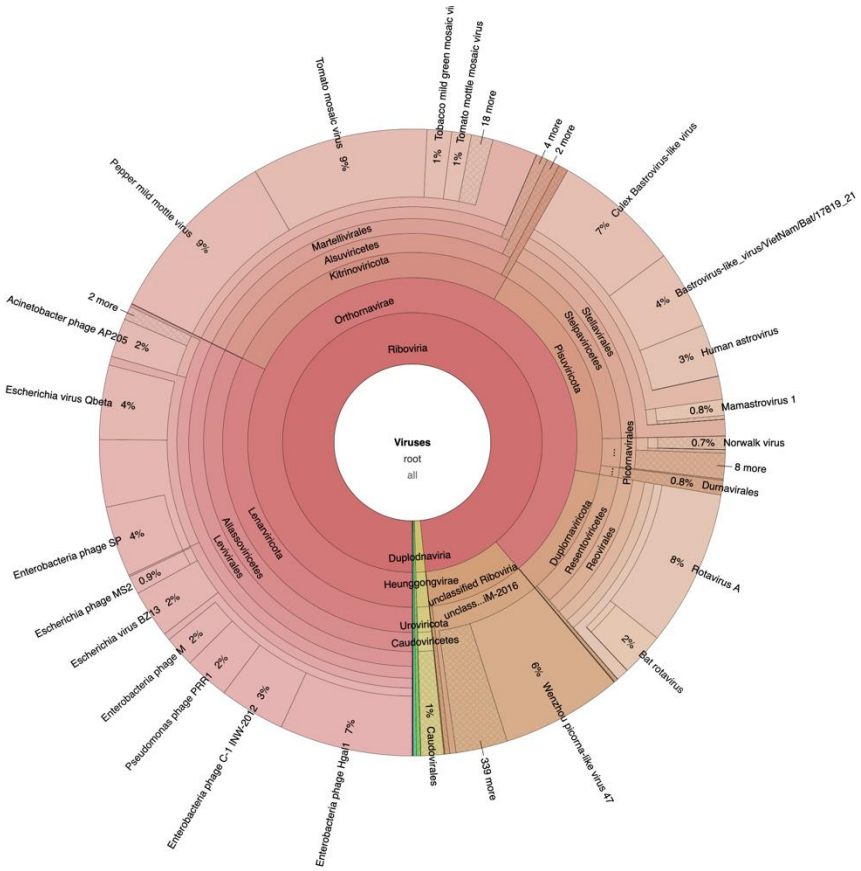


Figure 5-16: Viral taxonomic profile of sample RP01_07_2021, collected 2021/07/01.

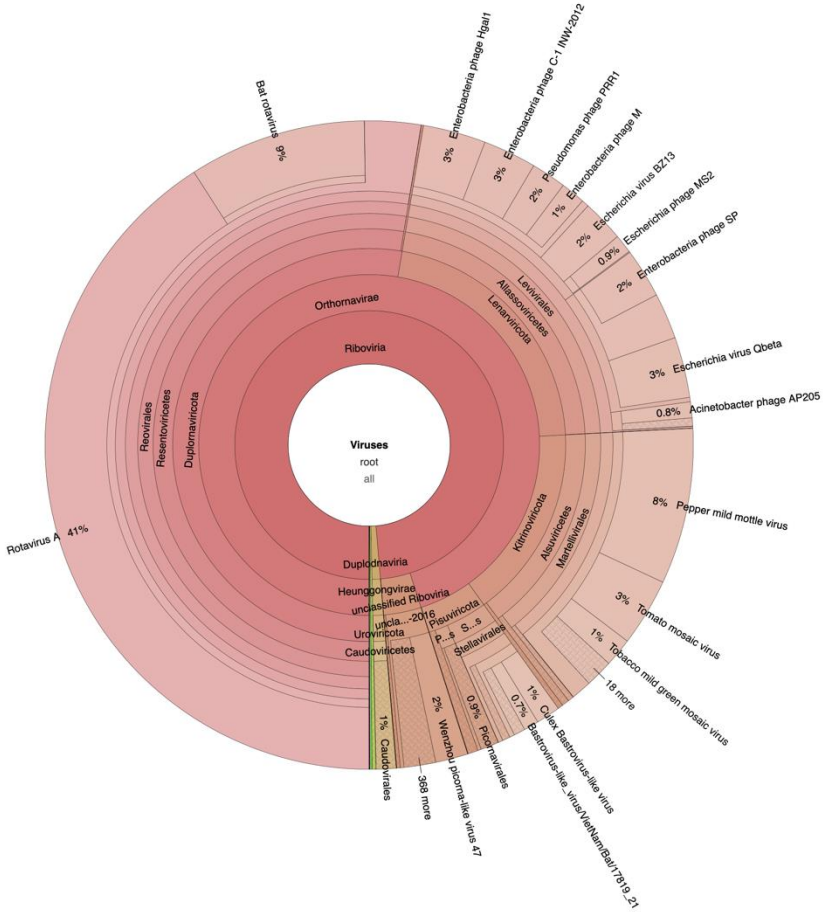


Figure 5-17: Viral taxonomic profile of sample RP27_07_2021, collected 2021/07/27.

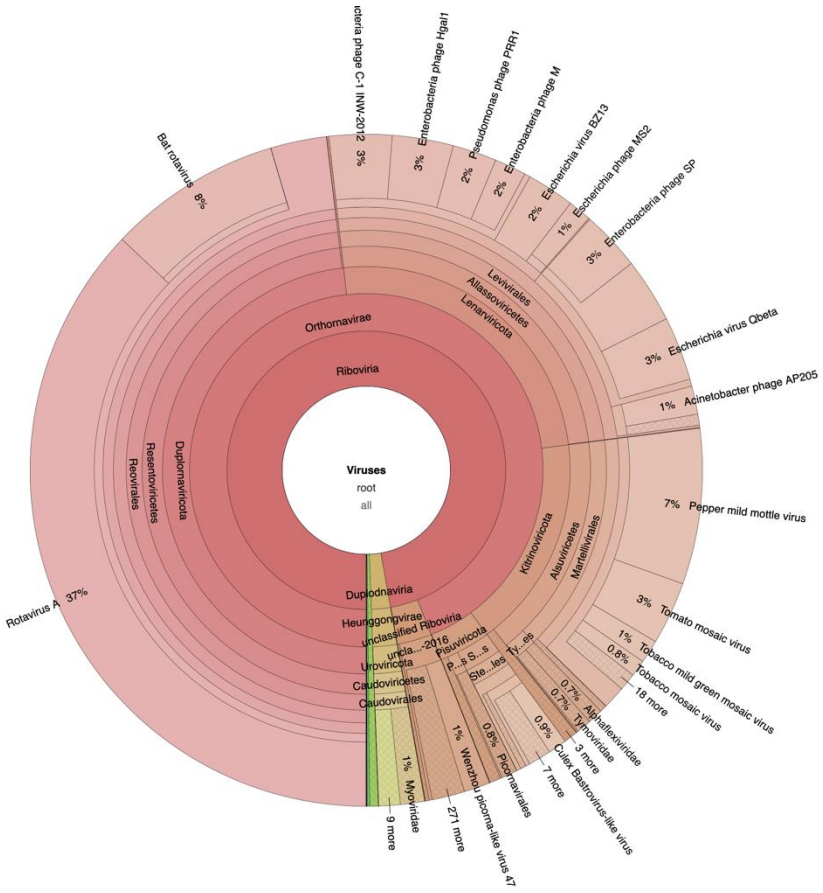


Figure 5-18: Viral taxonomic profile of sample RP03_08_2021, collected 2021/08/03.

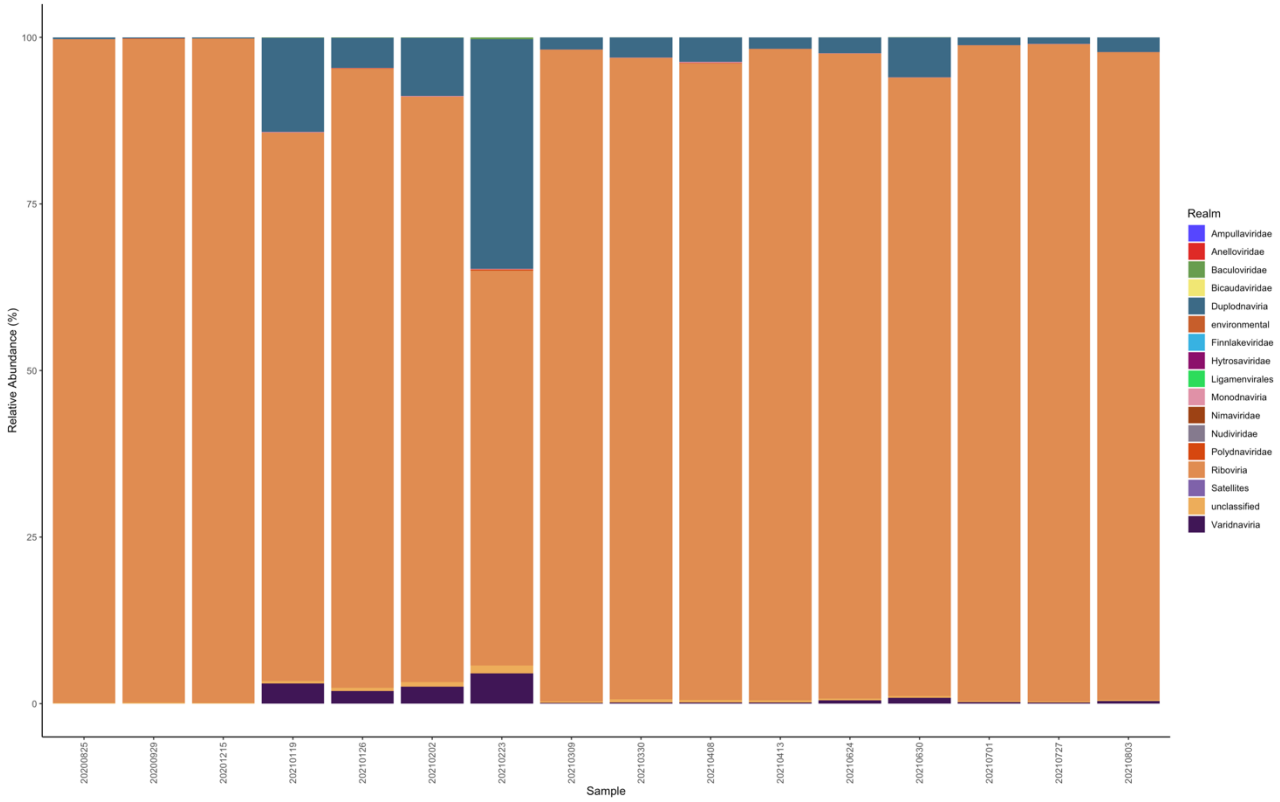


Figure 5-19: Relative abundance, as indicated by the percentage of reads, for the Viral realm classification. Each colour is representative of a viral realm. *Riboviria* was found to be in high abundance across all samples with *Duplodnaviria* and *Varidnaviria* high in certain samples.

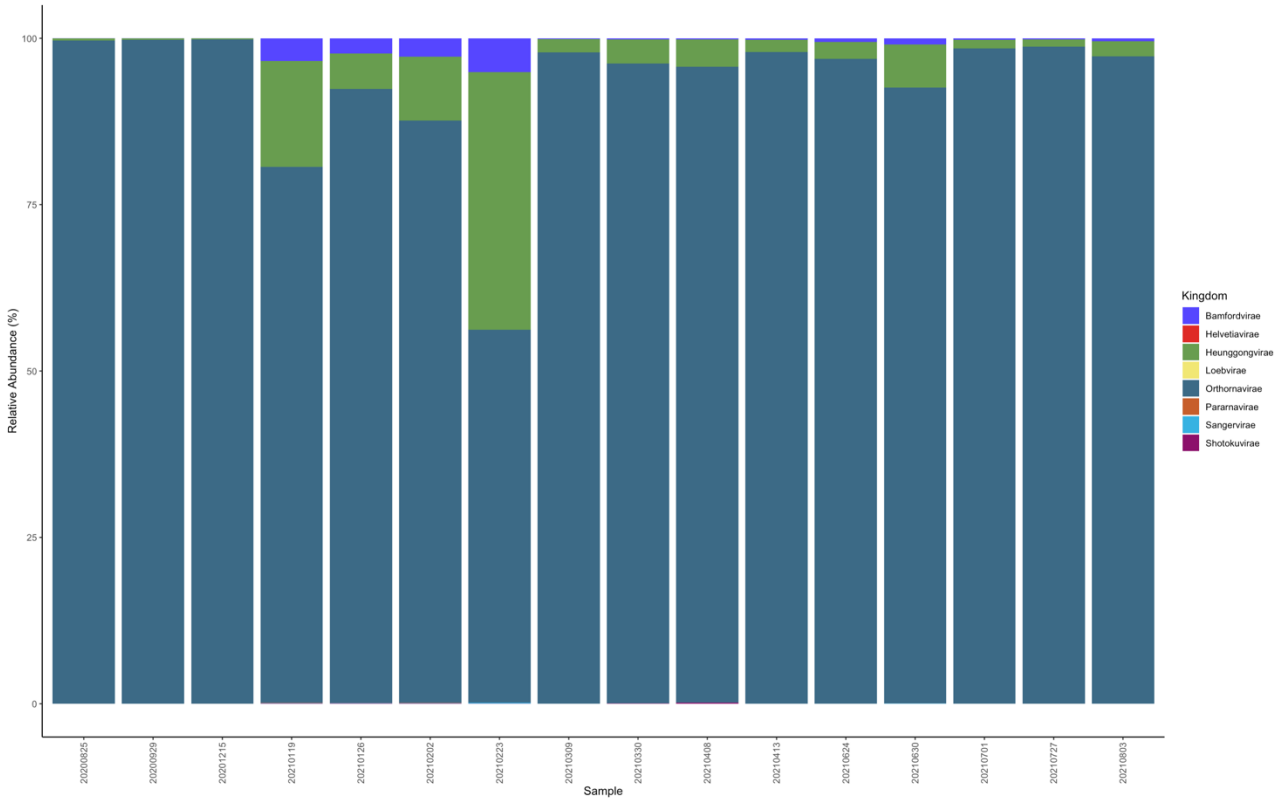


Figure 5-20: Relative abundance, as indicated by the percentage of reads, for the Viral kingdom classification. Each colour is representative of a viral kingdom. *Orthornavirae* was found to be in high abundance across all samples with *Heunggongvirae* and *Bamfordvirae* high in some of the samples.

5.4 DISCUSSION

The virome is defined as the assemblage of viruses in a particular environment, determined and classified by means of metagenomic sequencing. This is a focused method of metagenomic sequencing where various techniques are used to focus on the viral portion of a sample. One of these techniques is based on viral concentration using ultra (centricon) filtration whereafter RNA is extracted. This method worked particularly well when used with SARS-CoV-2 whole genome sequencing as described in the previous chapter.

Here, the RNA extracted after viral concentration (n=17) was subject to metagenomic sequencing and resulted in a high-quality data set per sample, with the exception of one sample (RP29_12_2020). All the samples were initially tested for the presence of SARS-CoV-2 by the collaborator and this was confirmed with the metagenomic sequencing. The samples all contained segments of the SARS-CoV-2 genome in varying quantities. This result highlights the applicability of untargeted metagenomic sequencing in wastewater epidemiology and the ability thereof to detect pathogens of interest.

Using metagenomic sequencing it was possible to taxonomically classify the virome as found in wastewater samples. As expected, the samples displayed a high prevalence of *Riboviria*. This realm of viruses includes all viruses using a homologous RNA-dependent polymerase for replication. A large portion of known viral diseases are caused by viruses in the *Riboviria* realm. As the COVID-19 pandemic is slowly but surely winding down, one has to be cognisant of other potential viral threats to human health. Metagenomic sequencing allows for broad epidemiological surveillance based on a single data generation event and should regularly be used as an early detection tool.

CHAPTER 6: AMPLICON AND METAGENOMIC SEQUENCING OF WASTEWATER SAMPLES FROM TSHWANE, GAUTENG

6.1 INTRODUCTION

Microorganisms are an essential part of our lives and are ubiquitous. These microbes often survive by interacting with other living organisms within the same environment through the establishment of mutual/symbiotic relationships. Bacteria in our bodies help us digest complex carbohydrates by using enzymes that we lack and compete with pathogenic species to prevent their overgrowth. In other environments like wastewater treatment plants, bacteria and fungi help break down organic nutrients and remove these from the water before it is released into various water bodies. In the early days before advancements in molecular biology techniques, scientists relied on the use of culture-dependent methods such as microscopy to study the morphology and behaviour of microorganisms. This method was however not efficient in studying microbes from complex environments where the exact community composition was often not known. As sequencing (especially third-generation) technologies became more accessible and much more affordable to use, the field of metagenomics also expanded, and more environments were explored.

The most common method used to determine the taxonomic composition of bacteria and fungi is the amplification and sequencing of marker genes (Escobar-Zepeda et al., 2015; Breitwieser et al., 2019). The 16S rRNA, 18S rRNA genes and ITS regions are examples of marker genes used by most of researchers to determine which bacterial and fungal groups form part of the community systems in an environment. This method has some disadvantages such as low resolution, especially at species level, amplification bias depending on the type of primer used (e.g. in bacterial studies, one can use the V1-V3, V3-V4, or just V4 region to make primers for PCR amplification) (Matsuo et al., 2021). However, it is a fast and cost-effective method of identifying bacteria and other eukaryotes (Breitwieser et al., 2019).

The functions that the microbiota perform in the human GI tract remain conserved from one individual to the other to individual, but often the composition will not be identical between two people (Coman & Vodnar, 2020). The different types of microbial communities that are present on or in the human body are affected by multiple factors that lead to variation between individuals (Moschen et al., 2012). Despite these differences, we often share a group of microbiota that remains the same because of the important functions they perform that cannot be traded off. These conserved groups are commonly known as the ‘core taxa’ (Moschen et al., 2012). The core taxa are spread throughout the gastrointestinal tract because the sections of the tract have different conditions that are not optimum for all the taxa. For example, the small intestine is known to support more fast-growing anaerobes compared to the colon due to differences in oxygen levels and tolerance to antimicrobials and bile acids (Donaldson et al., 2016).

The predominant bacterial species that exist in the human gut have been found to belong to the Firmicutes, Bacteroidetes, Proteobacteria phyla (Kho & Lal, 2018; Coman & Vodnar, 2020; Ghosh & Pramanik, 2021). There appears to be a shift in the bacterial composition from infancy to adulthood, with less representation of the above-mentioned phyla as one gets older. The Bacteroidetes and Firmicutes represent over 90% of the gut microbiota (Klement & Pazienza, 2019; Sakkas et al., 2020). The Firmicutes phylum is composed of genera such as Lactobacillus, Bacillus, Clostridium, Enterococcus, and Ruminicoccus. Clostridium being the most abundant human gut genus in this phylum (Rinninella et al., 2019; Beam et al., 2021). The ratio of these phyla to each other can be used as biomarkers for certain diseases (Ghosh & Pramanik, 2021), e.g. the ratio of Bacteroidetes to Firmicutes correlates with obesity. A study was done using sewage water from 71 United States cities and it revealed that there was similar human faecal oligotypes between the different communities. An estimated 27 abundant oligotypes were identified and labelled as the core taxa of the U.S population. These were either Prevotellaceae, Bacteroidaceae, or Ruminococcaceae oligotypes (Newton et al., 2015).

In this chapter, both amplicon and shotgun sequencing were used to investigate the taxonomic and antimicrobial resistance composition of wastewater samples found in Tshwane, Gauteng. These methods are used as an alternative to the culture-based methods and are able to produce much more information in a fraction of the time when compared to the culture-based methods.

6.2 MATERIALS AND METHODS

Samples (n=30) were collected from 3 wastewater treatment plants in Tshwane, Gauteng (Dr A. Gomba) (Table 6-1 and Figure 6-1). DNA extractions were done by the ARC Biotechnology including library preparation, amplicon and metagenomic sequencing (Supplementary Sequencing Quotation). The resulting libraries were sequenced on a MiSeq (amplicon) and a MGI DNBSEQ-G400 (metagenomics). Initially, a SARS-CoV-2 whole genome sequencing approach was attempted on the samples as discussed in Chapter 4. This method was unfortunately not successful on these samples as they may have been too old. In Chapter 4 difficulties regarding the SARS-CoV-2 whole genome sequencing approach is discussed and it is clear that older samples do perform worse and in many cases fail. It was therefor decided to proceed with an amplicon based alternative which would enable taxonomic profiling from these samples with high microbial diversity.

Table 6-1: Samples received for amplicon and metagenomic sequencing.

Sample ID	Sampling Site	Sample Type	Date Collected
1RR0721	Rooiwal WWTP	Influent	21/07/2020
2PR0728	Rooiwal WWTP	Primary Sludge	28/07/2020
3PR0804	Rooiwal WWTP	Primary Sludge	04/08/2020
4PR0811	Rooiwal WWTP	Primary Sludge	11/08/2020
5PR0818	Rooiwal WWTP	Primary Sludge	18/08/2020
6PR0826	Rooiwal WWTP	Primary Sludge	26/08/2020
7PR0901	Rooiwal WWTP	Primary Sludge	01/09/2020
8PR0908	Rooiwal WWTP	Primary Sludge	08/09/2020
9PR0915	Rooiwal WWTP	Primary Sludge	15/09/2020
10PR0929	Rooiwal WWTP	Primary Sludge	29/09/2020
1AD0721	Daspoort	Activated Sludge	21/07/2020
2PD0728	Daspoort	Primary Sludge	28/07/2020
3PD0804	Daspoort	Primary Sludge	04/08/2020
4PD0811	Daspoort	Primary Sludge	11/08/2020
5PD0818	Daspoort	Primary Sludge	18/08/2020
6PD0826	Daspoort	Primary Sludge	26/08/2020
7PD0901	Daspoort	Primary Sludge	01/09/2020
8PD0908	Daspoort	Primary Sludge	08/09/2020
9PD0915	Daspoort	Primary Sludge	15/09/2020
10PD0929	Daspoort	Primary Sludge	29/09/2020
2PS0728	Sunderland Ridge	Primary Sludge	28/07/2020
3PS0804	Sunderland Ridge	Primary Sludge	04/08/2020
4PS0811	Sunderland Ridge	Primary Sludge	11/08/2020
5PS0818	Sunderland Ridge	Primary Sludge	18/08/2020
6PS0826	Sunderland Ridge	Primary Sludge	26/08/2020
7PS0901	Sunderland Ridge	Primary Sludge	01/09/2020
8PS0908	Sunderland Ridge	Primary Sludge	08/09/2020
9PS0915	Sunderland Ridge	Primary Sludge	15/09/2020

Sample ID	Sampling Site	Sample Type	Date Collected
10PS0929	Sunderland Ridge	Primary Sludge	29/09/2020
11AS1014	Sunderland Ridge	Activated Sludge	14/10/2020

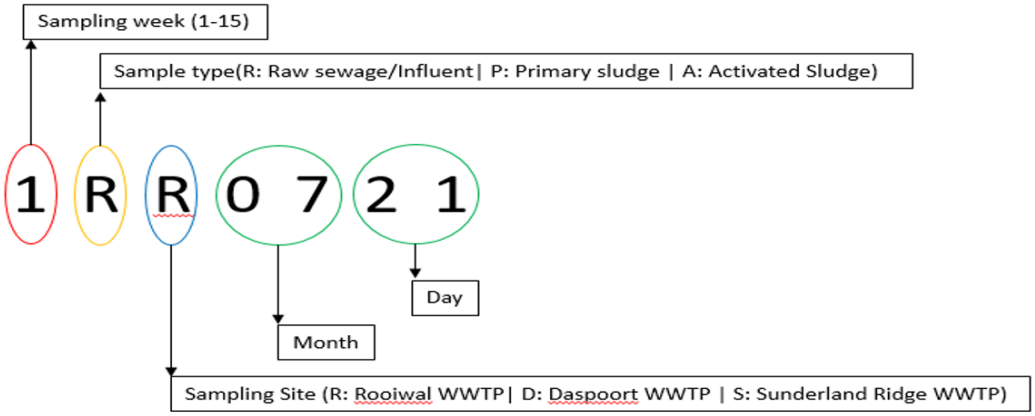


Figure 6-1: Sample ID description and identification.

The amplicon sequencing followed the following procedure. Microbial DNA was extracted using the Macherey-Nagel™ NucleoSpin™ DNA Stool kit, and 16S ribosomal RNA (16S rRNA) amplification and sequencing were performed according to the Illumina 16S protocol (16S Metagenomic Sequencing Library Preparation Guide). Briefly, the variable V3 and V4 regions of the 16S rRNA gene were amplified primers from Klindworth et al. (2013) from the samples, followed by library amplification and sequencing on the Illumina MiSeq instrument using V3 chemistry. The primer sequence was as follows: 16S forward primer = 5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG and 16S Reverse primer = 5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC. The PCR program was as follows: 95 °C for 3 min, 25 cycles of; 95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s and a final extension at 72 °C for 5 min, held at 4 °C. Generated data were evaluated for quality and used for downstream bioinformatic pipelines. Low-quality sequencing reads were filtered and trimmed to a consistent length with a maximum of 2 expected errors per-read enforced (Edga and Flyvbjerg, 2015). This is done on paired reads jointly, after which amplicon sequence variants are inferred and downstream analysis is done using the DADA2 method (Callahan et al., 2016).

This method combines identical sequencing reads into “unique sequences” with a corresponding abundance value followed by the identification of sequencing errors. Thereafter the forward and reverse reads are merged, and paired sequences that do not perfectly overlap are discarded. The resulting sequence table was inspected for chimeras which were removed. Taxonomy was assigned to the final, filtered sequence table using the SILVA ribosomal RNA gene reference database (Quast et al., 2012). The R package, phyloseq (McMurdie et al., 2013), was used to further analyze and graphically display the sequencing data which was clustered into amplicon sequence variants (ASVs) with the protocol described above. The ASV table was agglomerated onto operational taxonomic units (OTUs) according to taxonomic classification and inspected at “phylum” level to remove any unclassified OTUs. The OTU table was normalized using the ‘normalize function’ and the ‘median ratio” method implemented in MetalonDA R package (Metwally et al., 2018) which uses the DESeq2 “estimate size factors” function (Love et al., 2014). For this analysis, we added a pseudo count of 1 to the initial OTU table, running the normalization prior to rounding off the normalized table to the largest integer not exceeding the normalized value. Floor rounding was applied to negate the effect of the pseudo count addition. All bioinformatic analysis was done using a RStudio environment with R version 4.1.2.
The DNA metagenomic sequencing was done on the same extraction used for the above amplicon method and sequenced on a MGI DNBSEQ-G400. Initial sequence data quality and filtered data quality was inspected using FastQC version 0.11.8 (Andrews, S., 2010). Sequence data was quality trimmed and filtered, including adapter removal and decontamination, using BBDuk version 38.91 available from the BBTools suite of tools

(Bushnell, B., 2014). Filtered reads were assembled using SPAdes v.3.15.3 (Nurk et al., 2017) and only contigs with length exceeding 1,500 bp used for further analyses. ABRicate (https://github.com/tseemann/abricate) was used to detect antimicrobial resistance genes. Abricate allows for the mass screening of contigs for AMR genes. This program only detects acquired resistance and is not suitable for the detection of point mutations. Abricate was run with default parameters and the “ncbi” database selected. This database was locally updated 2023/01/05 and at time of usage included 6,334 sequences. The output from Abricate includes AMR gene name and putative antibiotic resistance phenotype.

6.3 RESULTS

6.3.1 Amplicon approach

Approximately 12 GB worth of raw sequencing data was produced for 29 samples. One sample failed extraction and library preparation and was excluded from further analyses. The initial results included 3,494 ASVs which were agglomerated into 750 Operational Taxonomic Units (OTU). The taxonomy profiles (Phylum, Class, Order) are for each sample is presented in Figure 6-2 – 6-4.

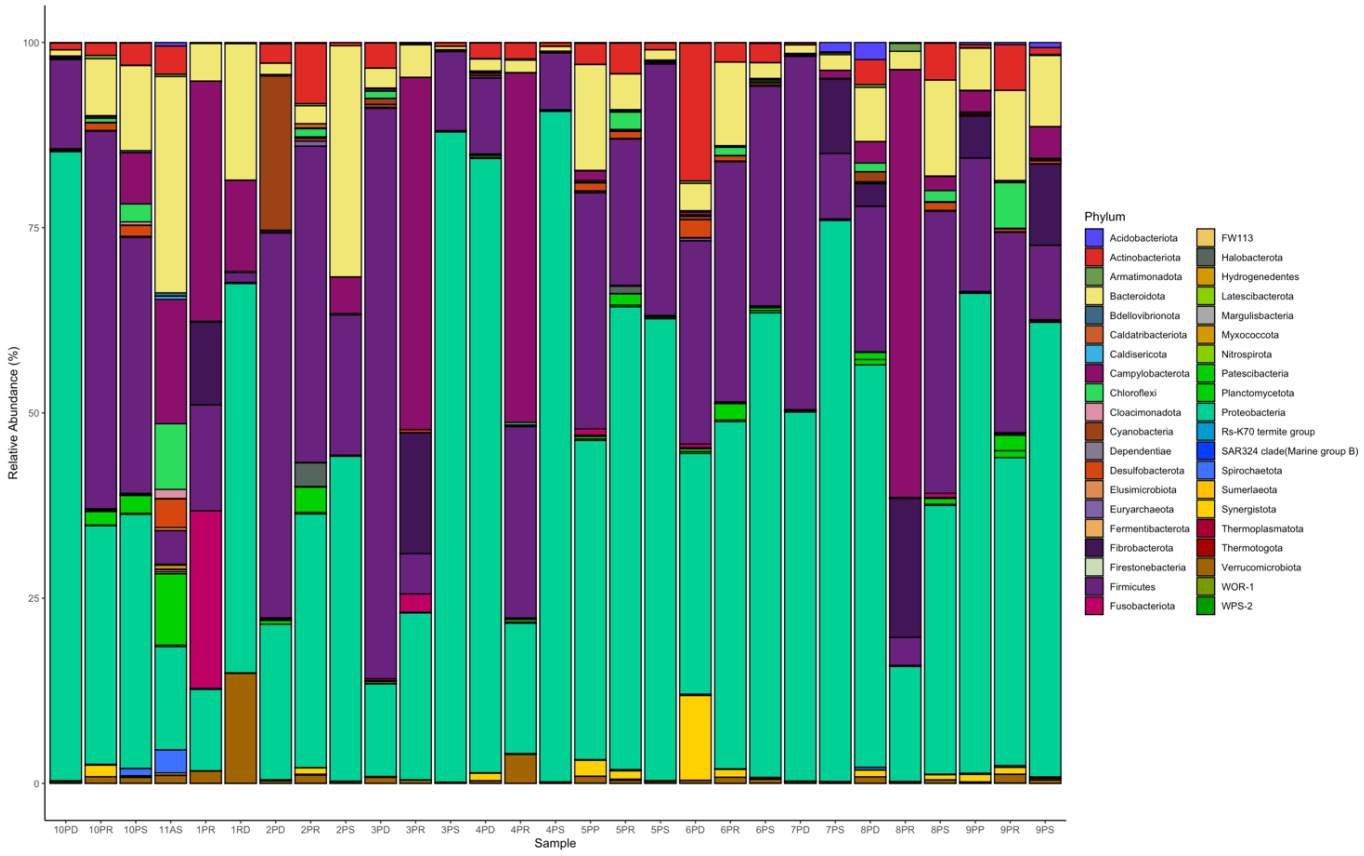


Figure 6-2: Relative abundance of the different Phyla for each of the 29 samples. This figure indicates high levels of diversity within each sample and large differences between samples. A high representation of Proteobacteria is evident, as is expected.

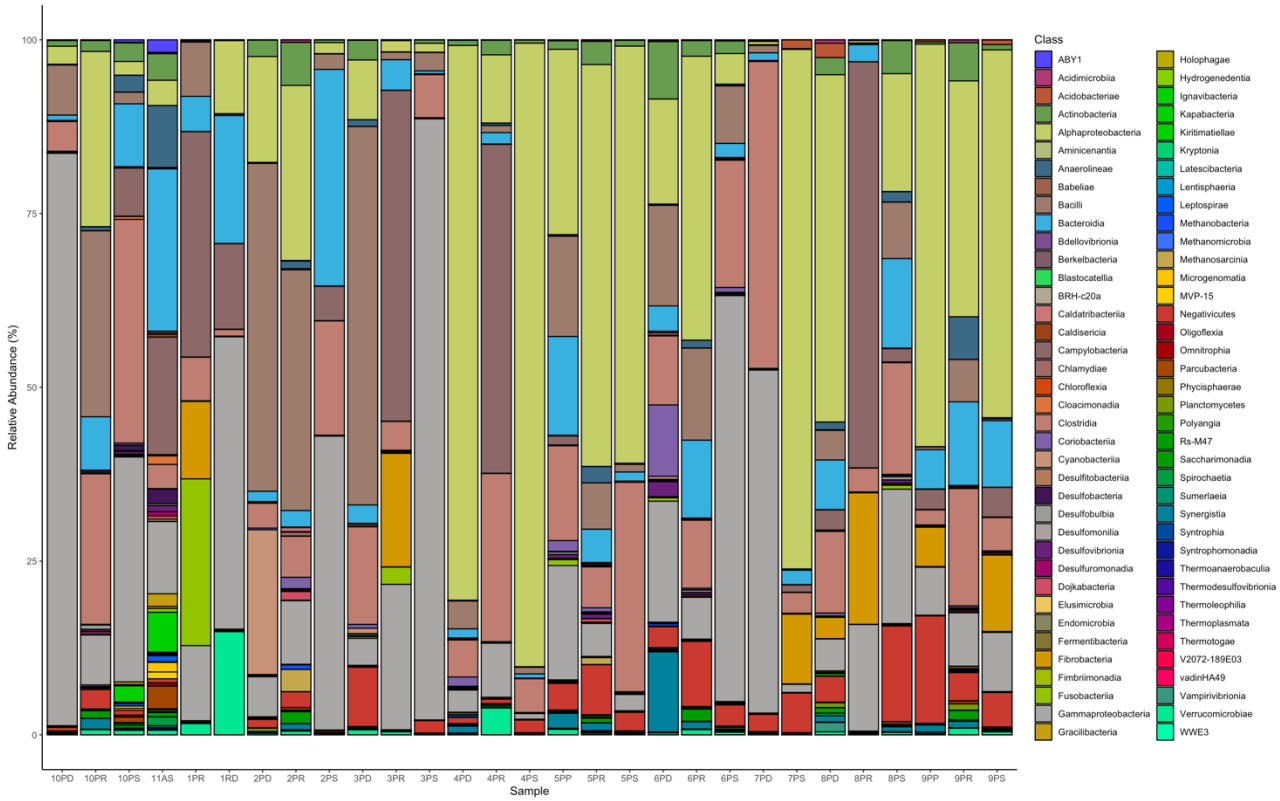


Figure 6-3: Relative abundance of the different Classes for each of the 29 samples. This figure indicates high levels of diversity within each sample and large differences between samples.

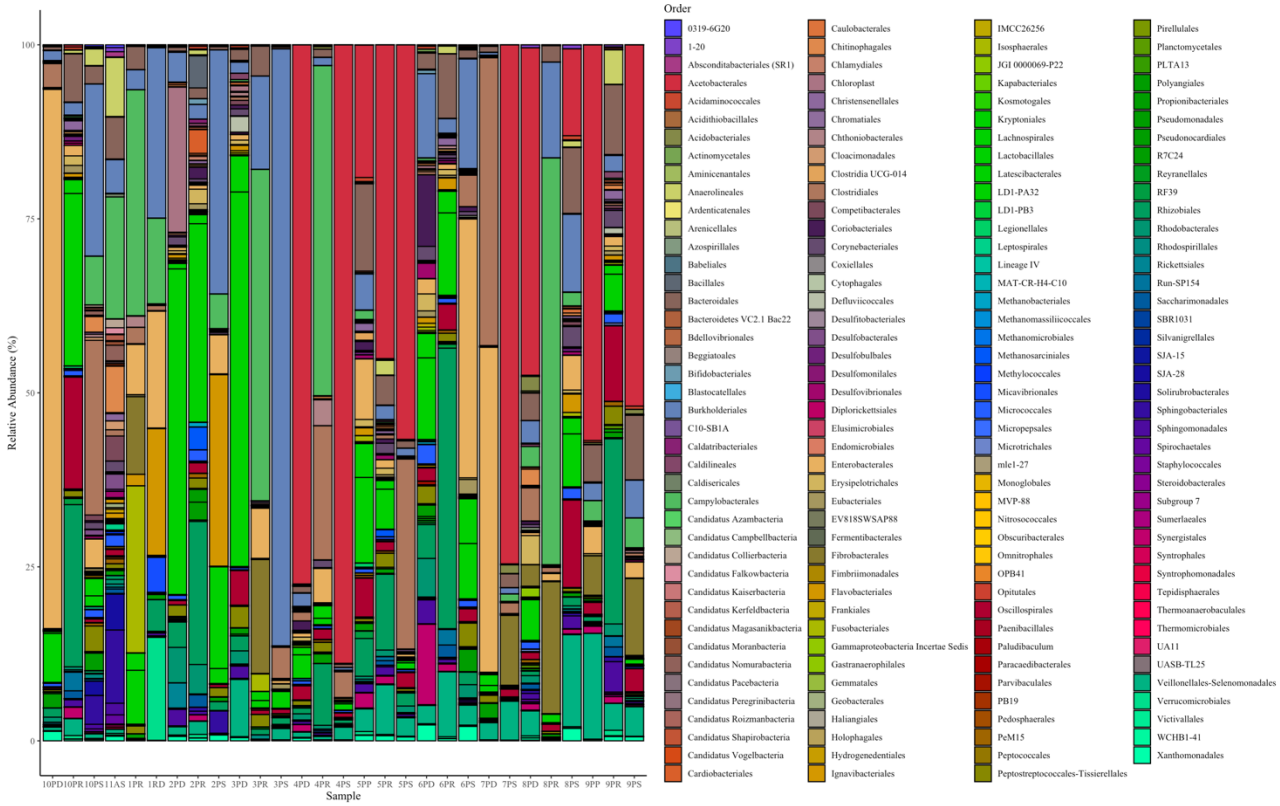


Figure 6-4: Relative abundance of the different Orders for each of the 29 samples. This figure indicates high levels of diversity within each sample and large differences between samples.

6.3.2 Detection of antimicrobial resistance

DNA extraction and metagenomic sequencing was successfully performed for 28 samples. High AMR levels were present in all the samples, except 7PD (Daspoort, Primary sludge, 2020/09/01) (Figure 6-5). The reason for this needs to be investigated as all other samples had in excess of 10 AMR elements whereas 7PD only displayed 2. A total of 136 different AMRs were detected across the 28 samples. The highest occurrence of AMRs was found in sample 6PR (Rooiwal, Primary sludge, 2020/08/26). The samples displayed the presence of 28 different resistance phenotypes.

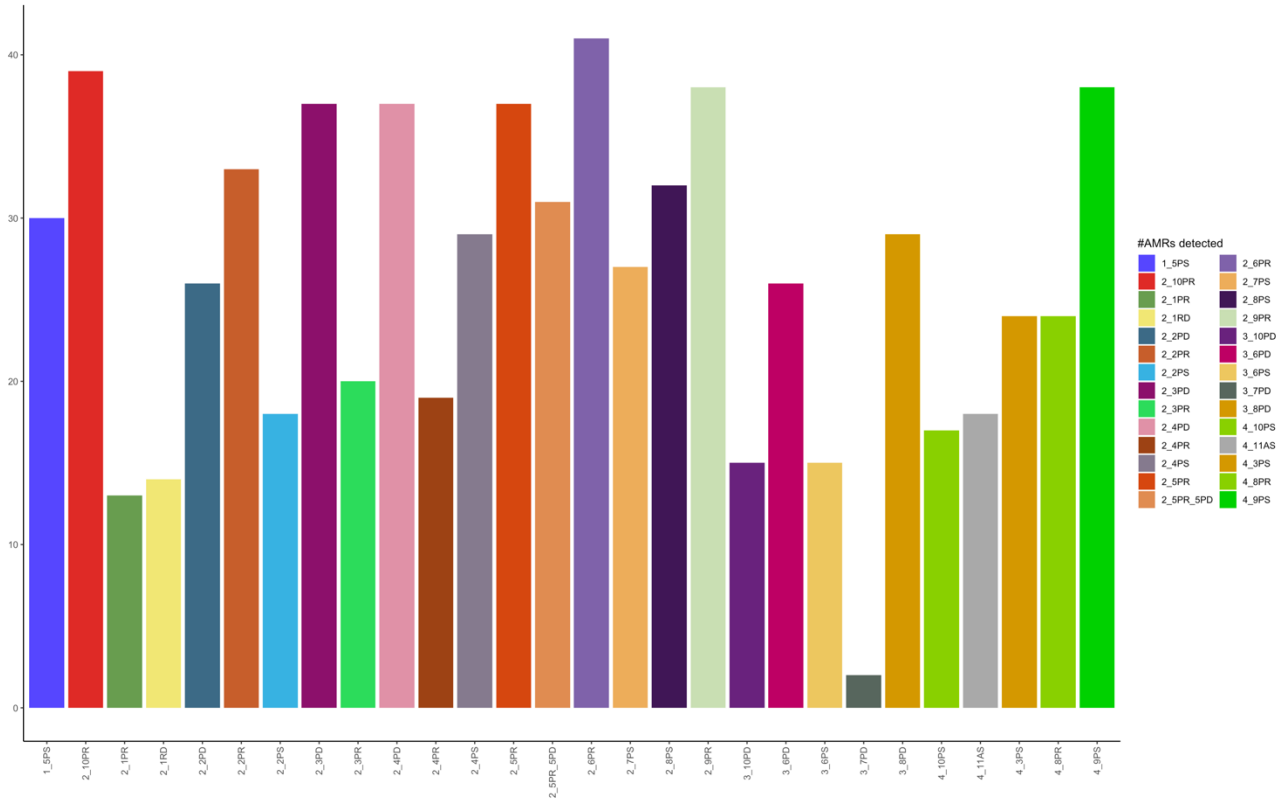


Figure 6-5: Number of AMR elements detected per sample. High levels of AMR were observed for all samples except 7PD.

High levels of tetracycline resistance phenotype were found across all samples and followed closely by Macrolide resistance phenotype and then Beta-lactam resistance phenotype (Figure 6-6 and Table 6-2) with a high diversity of AMR genes distributed within each sample (Figure 6-7 and Table 6-2).

The extreme incidence and diversity of AMR is clearly portrayed in Table 6-2. Wastewater samples from Tshwane treatment plants. Each sample displays a multitude of AMR genes and resistance phenotypes and could possibly indicate extreme levels of AMR in these Tshwane communities.

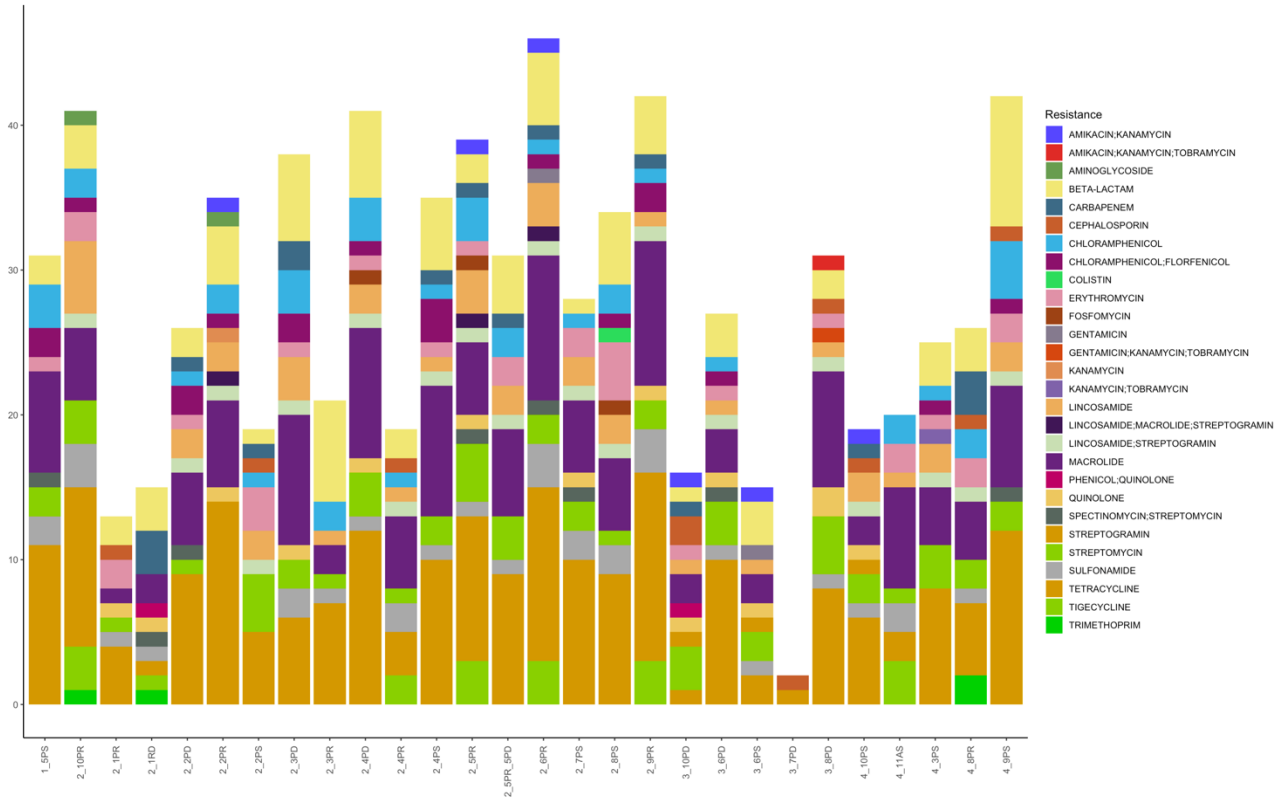


Figure 6-6: Resistance phenotype profile for each sample. High levels of Tetracycline were found in all the samples. A total of 28 different resistance phenotypes were found.

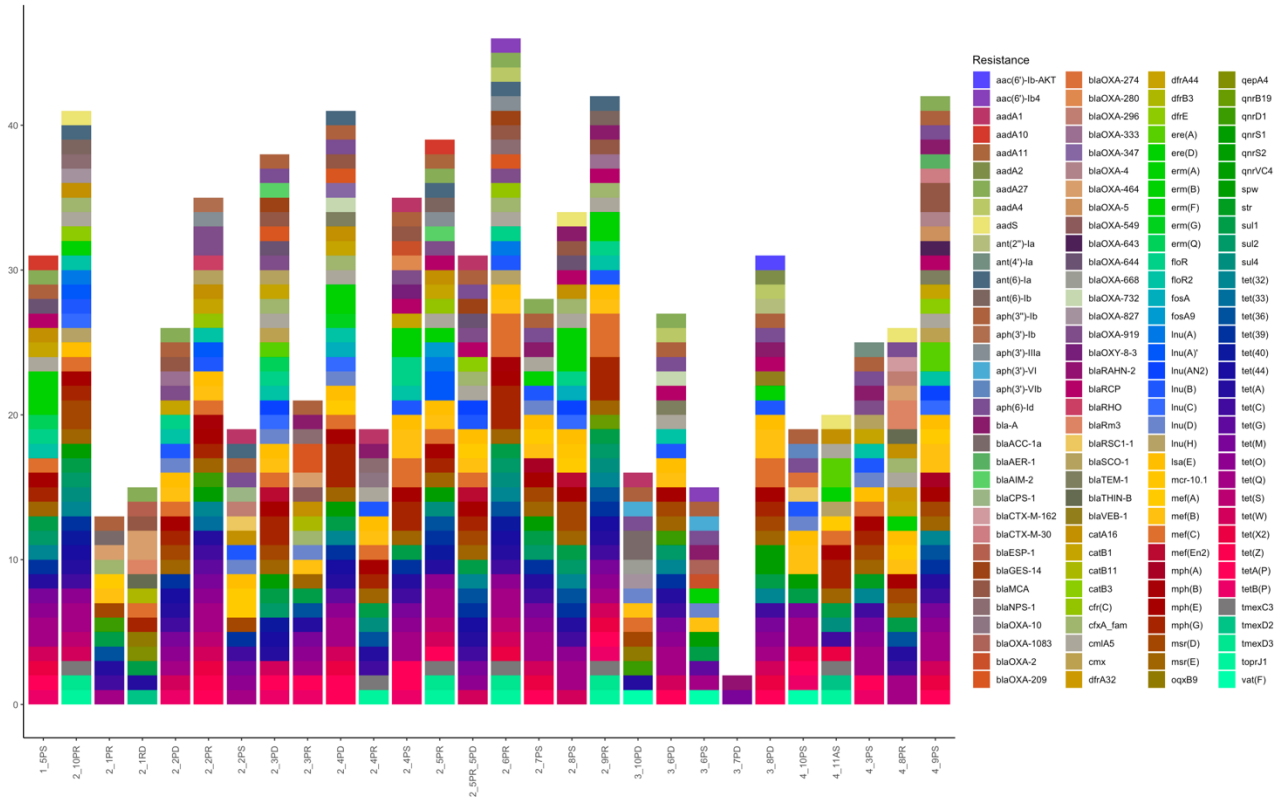


Figure 6-7: Resistance gene profile for each sample. A total of 136 different resistance genes were detected.

Table 6-2: AMR profiles for each sample.

Sample ID	Resistance Phenotype	Resistance Gene
3_7PD	CEPHALOSPORIN; TETRACYCLINE	tet(M); blaRAHN-2
1_5PS	CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; SULFONAMIDE; CHLORAMPHENICOL; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN	tetA(P); tetB(P); tet(Q); msr(E); floR2; erm(F); ere(D); tet(M); blaOXA-644; tet(39); mef(C); tet(O); aadA27; catB1; tet(W); erm(B); floR; catA16; tet(32); sul1; sul2; mph(G); cmlA5; mph(E); tet(X2); aph(3'')-Ib; erm(Q); tet(A); aadA10; blaRCP
2_1RD	TIGECYCLINE; SULFONAMIDE; QUINOLONE; TRIMETHOPRIM; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; CARBAPENEM; BETA-LACTAM; MACROLIDE; PHENICOL;QUINOLONE	oqxB9; dfrB3; tmexD2; blaOXA-464; sul1; blaTHIN-B; mph(G); blaESP-1; qepA4; blaMCA; aadA27; blaRm3; mef(C); tet(A)
2_9PR	CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; SULFONAMIDE; QUINOLONE; CHLORAMPHENICOL; TETRACYCLINE; CARBAPENEM; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; TIGECYCLINE; STREPTOMYCIN; LINCOSAMIDE	tetA(P); tetB(P); toprJ1; tet(Q); msr(E); ant(6)-Ib; floR2; tmexC3; erm(F); ant(6)-Ia; tet(M); tet(39); mef(C); tet(O); bla-A; tet(W); erm(B); mef(B); sul4; floR; cfxA_fam; tet(32); sul1; sul2; mph(G); cmlA5; tet(X2); tet(C); lnu(B); tmexD3; tet(44); tet(A); tet(36); lsa(E); blaOXA-333; blaRCP; qnrB19; blaMCA
3_10PD	CEPHALOSPORIN; STREPTOGRAMIN; AMIKACIN;KANAMYCIN; QUINOLONE; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; BETA-LACTAM; MACROLIDE; PHENICOL;QUINOLONE; STREPTOMYCIN; LINCOSAMIDE	aadA1; oqxB9; aph(3'')-Ib; mef(B); msr(D); aph(6)-Id; blaACC-1a; lnu(D); aph(3')-VI; blaOXA-827; mef(C); blaOXA-668; qnrD1; vat(F); tet(A)
2_8PS	FOSFOMYCIN; COLISTIN; CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; SULFONAMIDE; LINCOSAMIDE;STREPTOGRAMIN; CHLORAMPHENICOL; TETRACYCLINE;	mef(A); tet(Q); msr(E); mef(En2); lsa(E); ere(D); blaOXA-644; tet(39); erm(B); blaMCA; lnu(C); fosA; bla-A; tet(W); tet(O); floR; catA16; cfxA_fam; tet(32); sul1; sul2; cmlA5; erm(A); mph(E); mcr-10.1; aadS; tet(A); lnu(AN2); tet(36); tet(C); blaRCP; msr(D)

Sample ID	Resistance Phenotype	Resistance Gene
	ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	
3_6PS	GENTAMICIN; STREPTOGRAMIN; AMIKACIN;KANAMYCIN; SULFONAMIDE; QUINOLONE; TETRACYCLINE; BETA-LACTAM; MACROLIDE; STREPTOMYCIN; LINCOSAMIDE	qnrS1; mef(B); aph(3'')-Ib; blaOXA-1083; tet(G); tet(Q); aph(6)-Id; sul1; aph(3')-VI; lnu(D); bla-A; blaOXA-2; aac(6')-Ib4; vat(F); erm(B)
3_6PD	CHLORAMPHENICOL;FLORFENICOL; CHLORAMPHENICOL; MACROLIDE; SULFONAMIDE; QUINOLONE; ERYTHROMYCIN; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tetA(P); tet(Q); msr(E); aph(6)-Id; floR2; lsa(E); qnrD1; tet(39); tet(M); tet(O); aadA27; tet(W); mef(B); aadA4; blaOXA-732; msr(D); tet(32); sul2; cmlA5; mph(E); aph(3'')-Ib; blaTEM-1; lnu(B); tet(A); tet(C); blaRCP
4_3PS	CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; LINCOSAMIDE;STREPTOGRAMIN; KANAMYCIN;TOBRAMYCIN; CHLORAMPHENICOL; TETRACYCLINE; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tetA(P); tetB(P); tet(Q); msr(E); blaSCO-1; aph(6)-Id; floR2; lsa(E); tet(M); mef(C); tet(O); blaOXA-919; catB1; bla-A; ant(4')-Ia; msr(D); tet(32); mph(G); mph(E); lnu(B); aph(3'')-Ib; lnu(D); str; tet(C)
4_11AS	TIGECYCLINE; SULFONAMIDE; CHLORAMPHENICOL; TETRACYCLINE; ERYTHROMYCIN; MACROLIDE; STREPTOMYCIN; LINCOSAMIDE	tet(C); mef(A); msr(E); mph(G); sul4; tmexD2; sul1; aadS; tmexC3; cmlA5; mef(C); lnu(H); mph(E); toprJ1; tet(X2); catA16; ere(A); ere(D)
2_2PR	AMINOGLYCOSIDE; LINCOSAMIDE;MACROLIDE;STREPTOGRAMIN; CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; QUINOLONE; CHLORAMPHENICOL; TETRACYCLINE; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; AMIKACIN;KANAMYCIN; KANAMYCIN; LINCOSAMIDE	tetA(P); tet(Q); msr(E); blaSCO-1; tet(S); spw; floR2; lsa(E); lnu(A)'; tet(M); mef(C); tet(O); blaOXA-919; blaRHO; catB1; tet(W); mef(B); tet(33); catA16; tet(32); mph(G); mph(E); tet(X2); tet(Z); mph(B); cfr(C); lnu(B); tet(G); aph(3')-IIIa; tet(A); aph(3')-Ib; tet(C); qnrD1

Sample ID	Resistance Phenotype	Resistance Gene
2_2PD	CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; CHLORAMPHENICOL; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tetB(P); tet(Q); msr(E); floR2; lsa(E); tet(M); tet(39); mef(C); tet(O); blaOXA-919; aadA27; catB1; msr(D); mef(B); floR; tet(W); mph(G); mph(E); aph(3'')-Ib; tet(C); lnu(B); tet(44); lnu(D); tet(A); blaOXA-333; blaMCA
2_2PS	CEPHALOSPORIN; LINCOSAMIDE;STREPTOGRAMIN; CHLORAMPHENICOL; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tet(O); aadA1; mef(A); tet(Q); blaCPS-1; msr(D); catA16; aph(6)-Id; blaOXA-296; blaRSC1-1; lnu(D); ant(6)-Ia; lsa(E); tet(M); tet(39); aph(3'')-Ib; tet(C); lnu(B)
2_6PR	CHLORAMPHENICOL;FLORFENICOL; LINCOSAMIDE;MACROLIDE;STREPTOGRAMIN; TIGECYCLINE; SULFONAMIDE; GENTAMICIN; CHLORAMPHENICOL; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; CARBAPENEM; AMIKACIN;KANAMYCIN; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; MACROLIDE; STREPTOMYCIN; LINCOSAMIDE	aph(3')-IIIa; toprJ1; tet(Q); msr(E); tet(S); tmexC3; lsa(E); lnu(H); blaGES-14; mef(C); tet(O); lnu(A); aadA27; tet(W); mef(B); aadA4; blaOXA-209; blaOXA-919; sul4; floR; cfxA_fam; tet(32); sul1; sul2; mph(G); cmlA5; mph(E); aac(6')-Ib4; tet(X2); mph(B); cfr(C); lnu(B); blaNPS-1; tmexD3; tet(44); tet(A); tet(36); ant(6)-Ia; tet(C); tet(40); blaMCA
3_8PD	CEPHALOSPORIN; AMIKACIN;KANAMYCIN;TOBRAMYCIN; MACROLIDE; SULFONAMIDE; QUINOLONE; ERYTHROMYCIN; GENTAMICIN;KANAMYCIN;TOBRAMYCIN; TETRACYCLINE; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tetA(P); tet(Q); msr(E); aph(6)-Id; erm(F); lsa(E); tet(M); mef(C); tet(O); qnrS2; qnrVC4; blaVEB-1; bla-A; tet(W); mef(B); aadA2; aadA4; msr(D); tet(32); sul1; mph(G); mph(E); aac(6')-Ib-AKT; aph(3'')-Ib; tet(X2); lnu(B); ant(2'')-Ia; tet(C); blaRCP
4_10PS	CEPHALOSPORIN; STREPTOGRAMIN; MACROLIDE; SULFONAMIDE; QUINOLONE; TETRACYCLINE; CARBAPENEM;	mef(B); tetA(P); qnrS2; lnu(B); tet(Q); blaOXA-274; aph(6)-Id; sul1; blaRSC1-1; lnu(D); tetB(P); lsa(E); tet(M); aph(3'')-Ib; tet(X2); vat(F); aph(3')-VIb

Sample ID	Resistance Phenotype	Resistance Gene
	LINCOSAMIDE;STREPTOGRAMIN; AMIKACIN;KANAMYCIN; STREPTOMYCIN; LINCOSAMIDE	
2_4PR	CEPHALOSPORIN; MACROLIDE; SULFONAMIDE; CHLORAMPHENICOL; TETRACYCLINE; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; TIGECYCLINE; STREPTOMYCIN; LINCOSAMIDE	mef(B); lnu(B); msr(E); blaNPS-1; bla-A; sul4; blaOXA-10; aadA1; sul1; mph(G); tmexC3; cmlA5; lsa(E); tet(A); mph(E); tet(36); toprJ1; tet(C); mef(C)
2_4PS	CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; SULFONAMIDE; ERYTHROMYCIN; CHLORAMPHENICOL; TETRACYCLINE; CARBAPENEM; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tetA(P); tetB(P); tet(Q); msr(E); mef(B); floR2; lsa(E); ere(D); tet(39); tet(M); mef(C); erm(B); blaOXA-919; catB1; blaOXA-280; tet(O); tet(A); aadA1; floR; sul1; mph(G); mph(E); lnu(B); aph(3'')-Ib; blaMCA; blaOXY-8-3; tet(C); blaRCP; blaOXA-2
2_4PD	CHLORAMPHENICOL;FLORFENICOL; FOSFOMYCIN; MACROLIDE; SULFONAMIDE; QUINOLONE; CHLORAMPHENICOL; TETRACYCLINE; ERYTHROMYCIN; BETA-LACTAM; LINCOSAMIDE;STREPTOGRAMIN; STREPTOMYCIN; LINCOSAMIDE	tetA(P); mef(A); tet(Q); msr(E); aph(6)-Id; floR2; erm(F); ant(6)-Ia; tet(39); mef(C); tet(O); qnrS2; catB1; lnu(C); fosA; tet(W); erm(B); blaOXA-209; blaOXA-732; catA16; cfxA_fam; tet(32); sul1; blaOXA-347; erm(G); mph(G); cmlA5; mph(E); tet(X2); blaTEM-1; aph(3'')-Ib; tet(G); tet(44); lnu(D); tet(A); lsa(E); blaMCA
4_9PS	CEPHALOSPORIN; CHLORAMPHENICOL;FLORFENICOL; MACROLIDE; CHLORAMPHENICOL; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; ERYTHROMYCIN; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	tetA(P); mef(A); tet(Q); msr(E); aph(6)-Id; floR2; lsa(E); tet(M); tet(39); tet(O); aadA27; catB1; lnu(C); mph(A); blaAER-1; bla-A; tet(W); blaOXA-643; mef(B); msr(D); cmx; tet(32); blaCTX-M-30; cmlA5; blaOXA-4; mph(E); catB3; tet(X2); ere(A); blaTEM-1; aph(3'')-Ib; tet(A); lnu(AN2); tet(36); blaOXA-5; tet(C); blaRCP; blaMCA
2_3PR	MACROLIDE; SULFONAMIDE; CHLORAMPHENICOL; TETRACYCLINE; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	mef(B); tetA(P); tetB(P); aph(3'')-Ib; blaOXA-209; blaNPS-1; tet(Q); catA16; cfxA_fam; blaOXA-464; sul1; lnu(D); bla-A; tet(A); msr(E); tet(36); tet(M); blaOXA-549; catB11; tet(O)
2_3PD	CHLORAMPHENICOL;FLORFENICOL; QUINOLONE; MACROLIDE; SULFONAMIDE;	tetA(P); tetB(P); msr(E); mef(En2); aph(6)-Id; floR2; lsa(E); blaOXA-644; blaGES-14; mef(C); blaOXA-919; qnrS2; catB1; lnu(C); blaAIM-2;

Sample ID	Resistance Phenotype	Resistance Gene
	LINCOSAMIDE;STREPTOGRAMIN; CHLORAMPHENICOL; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	msr(D); blaSCO-1; mef(B); blaOXA-209; floR; cfxA_fam; sul1; sul2; mph(G); cmlA5; mph(E); tet(W); ere(A); aph(3'')-Ib; erm(Q); tet(44); lnu(D); tet(A); lnu(AN2); cmx; tet(40); blaMCA
2_5PR_5PD	MACROLIDE; SULFONAMIDE; LINCOSAMIDE;STREPTOGRAMIN; CHLORAMPHENICOL; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE	mef(A); msr(E); mef(B); mef(En2); aph(6)-Id; lsa(E); tet(M); blaOXA-644; tet(39); blaGES-14; mef(C); tet(O); blaOXA-919; catB3; tet(W); aadA1; msr(D); cfxA_fam; tet(32); sul1; mph(G); cmlA5; mph(E); lnu(B); aph(3'')-Ib; tet(44); tet(A); lnu(AN2); tet(36); tet(C); blaRCP
2_1PR	CEPHALOSPORIN; MACROLIDE; SULFONAMIDE; QUINOLONE; TETRACYCLINE; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN	mef(B); mef(A); tet(Q); msr(D); cfxA_fam; blaOXA-464; sul1; qnrD1; tet(36); aph(3'')-Ib; tet(C); blaACC-1a; tet(A)
2_10PR	CHLORAMPHENICOL;FLORFENICOL; AMINOGLYCOSIDE; TIGECYCLINE; SULFONAMIDE; TRIMETHOPRIM; ERYTHROMYCIN; CHLORAMPHENICOL; TETRACYCLINE; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; MACROLIDE; STREPTOMYCIN; LINCOSAMIDE	toprJ1; tet(Q); msr(E); tet(S); spw; floR2; tmexC3; erm(F); lsa(E); lnu(A)'; lnu(H); ant(6)-Ib; tet(39); mef(C); tet(O); lnu(A); lnu(C); tet(W); dfrE; sul4; msr(D); catA16; cfxA_fam; tet(32); sul1; sul2; mph(G); cmlA5; mph(E); tet(C); lnu(B); blaNPS-1; tmexD3; blaOXA-827; tet(44); tet(A); ant(6)-Ia; aadS; tet(40)
2_5PR	TIGECYCLINE; LINCOSAMIDE;MACROLIDE;STREPTOGRAMIN; FOSFOMYCIN; MACROLIDE; SULFONAMIDE; QUINOLONE; CHLORAMPHENICOL; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; LINCOSAMIDE;STREPTOGRAMIN; BETA-LACTAM; AMIKACIN;KANAMYCIN; STREPTOMYCIN; LINCOSAMIDE	tetA(P); tmexD3; toprJ1; tet(Q); msr(E); tet(S); fosA9; tmexC3; lsa(E); ere(D); ant(6)-Ib; tet(39); mef(C); tet(O); qnrS1; lnu(A); aadA27; catB1; blaAIM-2; tet(W); mef(B); blaOXA-919; sul4; catA16; mph(G); cmlA5; mph(E); aadA11; aph(3')-IIIa; tet(A); tet(36); ant(6)-Ia; aadA10; cfr(C); blaRCP; tet(40); lnu(A)'
4_8PR	CEPHALOSPORIN; MACROLIDE; SULFONAMIDE; TRIMETHOPRIM;	dfrA32; mef(A); tet(Q); msr(E); aph(6)-Id; blaOXA-464; erm(F); lsa(E); blaCTX-M-162; mef(B); dfrA44; msr(D); catA16; cfxA_fam; sul1;

Sample ID	Resistance Phenotype	Resistance Gene
	CHLORAMPHENICOL; TETRACYCLINE; CARBAPENEM; ERYTHROMYCIN; BETA-LACTAM; STREPTOMYCIN; LINCOSAMIDE;STREPTOGRAMIN	blaTHIN-B; cmlA5; blaRm3; mph(E); blaOXA-296; aadS; tet(A); tet(36); tet(C)
2_7PS	MACROLIDE; SULFONAMIDE; QUINOLONE; CHLORAMPHENICOL; SPECTINOMYCIN;STREPTOMYCIN; TETRACYCLINE; ERYTHROMYCIN; BETA-LACTAM; LINCOSAMIDE;STREPTOGRAMIN; STREPTOMYCIN; LINCOSAMIDE	tetA(P); mef(A); tet(Q); msr(E); aph(6)-Id; erm(F); lsa(E); tet(39); tet(M); tet(O); qnrS2; aadA27; mph(A); bla-A; tet(W); mef(B); msr(D); sul1; sul2; cmlA5; mph(E); aph(3'')-Ib; tet(X2); lnu(B); lnu(D); tet(A); tet(C)

6.4 DISCUSSION

The wastewater samples obtained from the 3 treatment plants in Tshwane, Gauteng, displayed high taxonomic diversity based on the 16S rRNA analyses. The 3,494 amplicon sequencing variants produced by amplicon sequencing and subsequent analyses were agglomerated into 750 operational taxonomic units (OTUs). A total of 40 different phyla were detected with members of Proteobacteria and Campylobacterota present in high abundance. There were 76 different taxonomic Classes and 167 Orders present across the samples. This methodology allows for the rapid taxonomic profiling of wastewater samples and provides researchers with an alternative solution to classic isolation and cultivation strategies. The wealth of diversity present, as detected by amplicon sequencing, further promotes this method as a viable alternative to currently used protocols.

Metagenomic sequencing of the samples enable the construction of AMR profiles across the samples. All samples with the exception of one had high AMG gene levels. A total of 136 different AMR genes were detected which related to 28 different resistance phenotypes. High levels of tetracycline resistance phenotype were found across all samples and followed closely by Macrolide resistance phenotype and then Beta-lactam resistance phenotype. The high levels of AMR found in these wastewater plants are of concern and will be compared to other treatment plants within the same municipality and treatment plants from other municipalities.

The data generated during this part of the project is assisting Mr Don Jambo in his MSc Microbiology (NWU) degree. He is currently busy with further analyses which will be included in his thesis and publications.

CHAPTER 7: THE RECONSTRUCTION OF METAGENOME ASSEMBLED GENOMES FROM WASTEWATER SAMPLES

7.1 INTRODUCTION

Activated sludge is the most common treatment form in wastewater systems, where microbes are used to remove carbon, nitrogen, phosphorus, pathogens, and other pollutants such as pharmaceutical products and pesticides from agricultural waste (Wu et al., 2019). Bacteria is considered the most important microorganism in wastewater systems because they are responsible for most of the waste removal and chemical transformation in the entire process (Silva-Bedoya et al., 2016). In early days when researchers were relying on culture dependent methods to perform such studies, the full extent of the diversity of these communities was not known. That has since changed since metagenomics became more popular because of the accessibility of sequencing and its reduced cost. Most of the studies performed revealed a similar pattern in the bacterial composition of wastewater systems, despite the different geographical locations.

The human body is home to trillions of microorganisms living on and within it through. This is made possible by the symbiotic relationship that these microbial cells have with the different cell types in various parts of our bodies (Clemente et al., 2012; Lagier et al., 2012; Moschen et al., 2012). The different microbial groups are collectively known as the human microbiota and consist of bacteria, eukaryotes, viruses (Lagier et al., 2012; Cani, 2018), and some archaeal cells. These microbial cells function to bring about some homeostatic balance in the body through energy storage (Moschen et al., 2012), metabolic assistance, and even form an integral part of the immune system (Clemente et al., 2012; Bull & Plummer, 2014; Almeida et al., 2019; Shahi, S. K. et al., 2019).

The human gut is commonly referred to as the “second brain” in the body and is connected to the central nervous system through the gut-brain axis. This connection allows the linkage of cognitive functions to the intestinal functions (Bull & Plummer, 2014), making the gut a critical component in understanding most disorders in the body that are linked to the central nervous system (Oluwagbemigun et al., 2022). As such, the gut and its microbiome has been the most studied than any other body site. The studies around this area of research also tend to focus on the bacterial component (Oliphant & Allen-Vercoe, 2019; Shahi, Shailesh K et al., 2019), just like in wastewater metagenomics studies.

The bacterium in the gastrointestinal region is responsible for a range of functions that mainly assist with digestion and retaining nutrients. These are in the form of carbohydrate and protein metabolism, into products that can be used easily by the host because human cells often lack the ability to produce enzymes that can easily and quickly break down complex macromolecules into simple products that can be absorbed into the blood stream (Oliphant & Allen-Vercoe, 2019; Ghosh & Pramanik, 2021). The degradation of proteins is relatively less understood compared to complex carbohydrates but is important for the normal functioning of the GI tract.

16S rRNA amplification is the most common and standard method being used today for taxonomic and phylogenetic identification of microorganisms from environmental samples. This technique is PCR amplification reaction that uses the V3-V4 region of the 16S rRNA gene to construct primers, this is because the V3-V4 region is highly conserved between different types of bacteria (Wang et al., 2022). The PCR amplicons are then sequenced and the differences in the less conserved regions allow for the identification of specific taxa. There are however limitations when using 16S rRNA sequencing in metagenomic studies. The biggest disadvantage being that it is less precise at identifying microorganisms at species level and cannot identify other specific genes that are associated with the microbiota (Ranjan et al., 2016), this limits our understanding of the microbiome. The technique is however cheap, and the results do not require extensive data analysis. The 16S rRNA genes

have been used for over 40 years as phylogenetic markers, hence there is a wide representation of this marker in many databases (Escobar-Zepeda et al., 2015).

Shotgun metagenomics on the other hand consists of sequencing the entire DNA of the bacteria isolated from the environment. The DNA is prepared to construct whole shotgun libraries, The information from shotgun sequencing can be used to identify the different genes that are present in that particular sample (not just the composition of the microbiota), as well as the metabolic potential (Escobar-Zepeda et al., 2015). This is the preferred method in analysing the genomes of microorganisms directly from the environment.

The ability afforded by shotgun sequencing includes the assembly of partial and near complete genomes directly form the environment. The construction of metagenome assembled genomes (MAGs) enables detailed investigation into the taxonomic classification and functional potential of microorganisms as found in wastewater samples. As certain microorganisms are extremely difficult to isolate and cultivate, shotgun metagenomics offers an alternative to culture-based methods. This method is demonstrated in this chapter and illustrates the functionality thereof focusing on *Legionella pneumophila*, *Mycobacterium* spp. and *Aeromonas* spp.

7.2 MATERIALS AND METHODS

Wastewater samples (n=10) were collected from three municipal WWTPs in Pretoria, South Africa, that primarily treat household sewage (Dr A. Gomba). Table 7-1 summarizes the characteristics and treatment processes used at each sampling site. Grab samples (influent, activated sludge and secondary settling tank (SST) effluent) were collected from November 2021 to February 2022 at different treatment stages. Sterile one-liter bottles were used to collect samples, which were then transported to the laboratory on ice and stored at 4°C until processing.

Table 7-1: Characteristics of WWTP sampling sites.

Site	Treatment capacity (ML/day*)	Aeration technology	Source of wastewater (%)	Population size served
WWTP1	35	surface aeration	domestic (90) industrial (10)	366 709
WWTP2	60	surface aeration	domestic (100)	600 000
WWTP3	93	surface aeration	domestic (80) industrial (20)	236 580

* ML/day – mega litres per day

After DNA extraction the samples were sequenced on a MGI DNBSEQ-G400 with 10 GB or roughly 25,000,000 reads requested per sample (Supplementary Sequencing Quotation). Initial sequence data quality and filtered data quality was inspected using FastQC version 0.11.8 (Andrews, S., 2010). Sequence data was quality trimmed and filtered, including adapter removal and decontamination, using BBDuk version 38.91 available from the BBTools suite of tools (Bushnell, B., 2014). Filtered reads were assembled using SPAdes v.3.15.3 (Nurk et al., 2017) and only contigs with length exceeding 1,500 bp used for further analyses. For each sample the contigs were binned using MetaBAT v.2.15 (Kang et al., 2019) and genome quality of each bin assessed with CheckM v.1.1.3 (Parks et al., 2015). A bin was assigned as being a MAG of medium quality if the completeness was larger or equal to 50% and contamination less than 10% (Bowers et al., 2017). Each medium quality MAG was then assigned a taxonomic classification using GTDB-Tk v.1.7.0 (Chaumeil et al., 2020). The multiple sequence alignment for 120 bacterial markers as produced by the GTDB-Tk workflow was used to produce a phylogenetic tree with FastTree v.2.1.11 (Price, et al., 2010) and visualized with ggtree (Yu et al., 2017). ABRicate (https://github.com/tseemann/abricate) was used to detect antimicrobial resistance genes and virulence factors in species of interest. The species of interest for AMR and virulence factor detection were *Legionella pneumophila*, *Mycobacterium* spp. and *Aeromonas* spp. Abricate was run with default parameters

and the “ncbi” database selected for AMR detection. This database was locally updated 2023/01/05 and at time of usage included 6,334 sequences. For virulence factors the “vfdb” database was used, updated on 2022/11/02, containing 4,332 sequences.

7.3 RESULTS

7.3.1 Metagenome Assembled Genomes

Metagenomic binning resulted in the construction of 34 medium quality MAGs (Table 7-2). The MAG statistics indicated the presence of high-quality MAGs with completeness larger than 90% and contamination less than 5% with some MAGs found to be very near complete and of draft genome quality. Species level classification was further possible for the majority of the MAGs and included representatives of the species of interest. The medium quality mags included 4 Actinobacteriota, 7 Firmicutes and 23 Proteobacteria at the phylum level (Figure 7-1). The numerous taxa obtained from the MAGs is shown in Table 7-3.

Table 7-2: MAG statistics.

Bin Id	Completeness	Contamination	Genome size (bp)	# contigs	GC	# predicted genes
WP1_INF.metabat.10	50.88	0.00	2796765	15	50.5	2596
WP1_INF.metabat.11	96.07	1.08	5479669	127	49.9	4862
WP1_INF.metabat.2	88.19	1.08	4207953	50	40.3	3923
WP1_INF.metabat.4	94.66	0.19	3177878	60	38.2	2838
WP1_INF.metabat.6	94.45	3.19	5337866	1012	67.2	5937
WP1_INF.metabat.9	95.29	0.00	4245992	87	61.9	3895
WP1_SST.metabat.1	57.99	0.54	2271464	21	40.3	1990
WP1_SST.metabat.3	62.85	1.08	2332096	25	42.1	2119
WP2_A5.metabat.12	94.98	2.70	4550476	57	40.0	4137
WP2_A5.metabat.15	99.75	1.32	5829570	102	67.1	5610
WP2_A5.metabat.2	82.34	4.73	3810755	84	41.8	3449
WP2_A5.metabat.3	97.58	0.79	5496814	131	49.9	4906
WP2_A5.metabat.4	89.65	0.00	4200368	53	40.5	3828
WP2_A5.metabat.8	99.24	0.19	3435984	35	38.2	3069
WP2_INF.metabat.5	97.92	0.73	5397266	469	67.2	5514
WP2_SST.metabat.1	99.24	0.19	3335074	18	38.2	3005
WP3_A5.metabat.1	58.62	0.00	3046587	315	62.2	2962
WP3_A5.metabat.11	87.27	0.27	3475702	305	51.5	3395
WP3_A5.metabat.14	68.09	0.00	3442818	597	51.4	3581
WP3_A5.metabat.18	50.88	0.00	1730833	325	41.6	1735
WP3_A5.metabat.3	85.06	3.33	3746732	242	41.3	3521
WP3_A5.metabat.5	99.73	0.54	5672061	127	49.9	5016
WP3_INF.metabat.2	68.97	0.00	3782452	125	62.3	3512
WP3_INF.metabat.7	99.24	0.19	3295813	19	38.2	2961
WP3_INF.metabat.8	87.65	1.86	6868624	145	67.1	6736
WP4_A5.metabat.14	98.58	0.00	5031109	51	51.7	4649
WP4_A5.metabat.2	56.76	0.38	1406442	264	41.4	1548
WP4_A5.metabat.8	55.77	0.00	3817422	100	41.7	3499
WP5_A5.metabat.13	92.67	2.34	6110892	90	64.9	5626
WP5_A5.metabat.6	98.64	0.50	5340117	114	51.6	4785
WP5_INF.metabat.5	71.14	1.74	3316617	912	51.1	3452
WP5_INF.metabat.6	57.85	5.26	5324284	175	40.4	5003

Bin Id	Completeness	Contamination	Genome size (bp)	# contigs	GC	# predicted genes
WP5_INF.metabat.8	88.53	4.14	5088418	278	39.9	4780
WP5_INF.metabat.9	97.87	2.97	4795601	164	41.3	4523

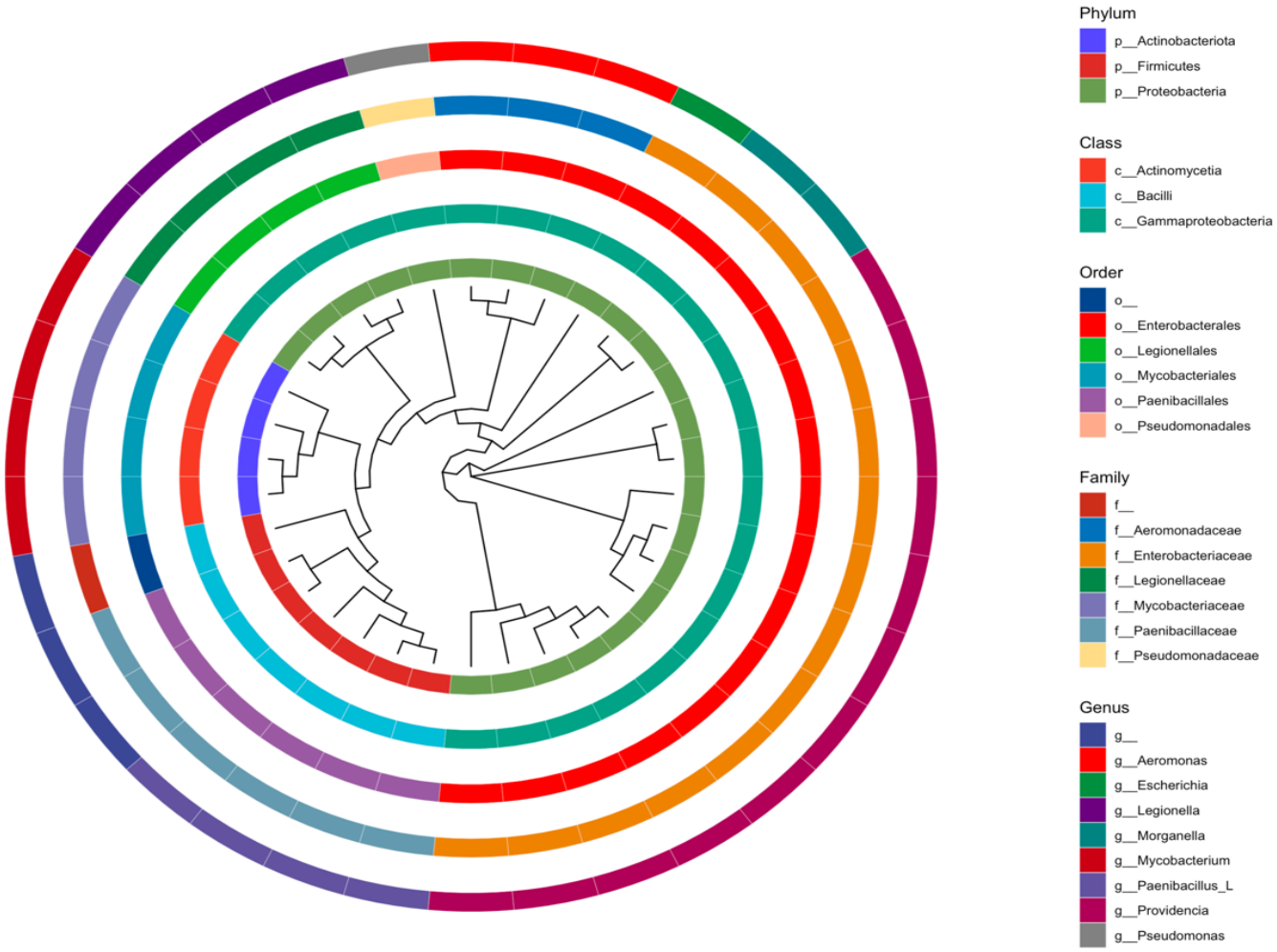


Figure 7-1: Phylogenetic tree of the 34 medium quality MAGs based on 120 universal bacterial markers. The coloured rings represent taxonomic classification.

Table 7-3: MAG taxonomic classification.

Bin Id	Phylum	Class	Order	Family	Genus	Species
WP1_INF.metabat.10	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Morganella	s__Morganella morganii_A
WP1_INF.metabat.11	p__Firmicutes	c__Bacilli	o__Paenibacillales	f__Paenibacillaceae	g__Paenibacillus_L	s__Paenibacillus_L sp007679495
WP1_INF.metabat.2	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia rettgeri
WP1_INF.metabat.4	p__Proteobacteria	c__Gammaproteobacteria	o__Legionellales	f__Legionellaceae	g__Legionella	s__Legionella pneumophila
WP1_INF.metabat.6	p__Actinobacteriota	c__Actinomycetia	o__Mycobacteriales	f__Mycobacteriaceae	g__Mycobacterium	s__Mycobacterium phocaicum
WP1_INF.metabat.9	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Aeromonadaceae	g__Aeromonas	s__Aeromonas caviae
WP1_SST.metabat.1	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia stuartii_A
WP1_SST.metabat.3	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__
WP2_A5.metabat.12	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia stuartii_A
WP2_A5.metabat.15	p__Actinobacteriota	c__Actinomycetia	o__Mycobacteriales	f__Mycobacteriaceae	g__Mycobacterium	s__Mycobacterium phocaicum
WP2_A5.metabat.2	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia alcalifaciens
WP2_A5.metabat.3	p__Firmicutes	c__Bacilli	o__Paenibacillales	f__Paenibacillaceae	g__Paenibacillus_L	s__Paenibacillus_L sp007679495
WP2_A5.metabat.4	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__
WP2_A5.metabat.8	p__Proteobacteria	c__Gammaproteobacteria	o__Legionellales	f__Legionellaceae	g__Legionella	s__Legionella pneumophila
WP2_INF.metabat.5	p__Actinobacteriota	c__Actinomycetia	o__Mycobacteriales	f__Mycobacteriaceae	g__Mycobacterium	s__Mycobacterium phocaicum
WP2_SST.metabat.1	p__Proteobacteria	c__Gammaproteobacteria	o__Legionellales	f__Legionellaceae	g__Legionella	s__Legionella pneumophila
WP3_A5.metabat.1	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Aeromonadaceae	g__Aeromonas	s__Aeromonas hydrophila
WP3_A5.metabat.11	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Morganella	s__Morganella morganii
WP3_A5.metabat.14	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Escherichia	s__Escherichia flexneri
WP3_A5.metabat.18	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia stuartii_A
WP3_A5.metabat.3	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia stuartii
WP3_A5.metabat.5	p__Firmicutes	c__Bacilli	o__Paenibacillales	f__Paenibacillaceae	g__Paenibacillus_L	s__Paenibacillus_L sp007679495
WP3_INF.metabat.2	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Aeromonadaceae	g__Aeromonas	s__Aeromonas caviae
WP3_INF.metabat.7	p__Proteobacteria	c__Gammaproteobacteria	o__Legionellales	f__Legionellaceae	g__Legionella	s__Legionella pneumophila
WP3_INF.metabat.8	p__Actinobacteriota	c__Actinomycetia	o__Mycobacteriales	f__Mycobacteriaceae	g__Mycobacterium	s__Mycobacterium mageritense
WP4_A5.metabat.14	p__Firmicutes	c__Bacilli	o__Paenibacillales	f__Paenibacillaceae	g__	s__
WP4_A5.metabat.2	p__Firmicutes	c__Bacilli	o__	f__	g__	s__

Bin Id	Phylum	Class	Order	Family	Genus	Species
WP4_A5.metabat.8	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__
WP5_A5.metabat.13	p__Proteobacteria	c__Gammaproteobacteria	o__Pseudomonadales	f__Pseudomonadaceae	g__Pseudomonas	s__Pseudomonas nitroreducens
WP5_A5.metabat.6	p__Firmicutes	c__Bacilli	o__Paenibacillales	f__Paenibacillaceae	g__	s__
WP5_INF.metabat.5	p__Firmicutes	c__Bacilli	o__Paenibacillales	f__Paenibacillaceae	g__Paenibacillus_L	s__Paenibacillus_L sp007679495
WP5_INF.metabat.6	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__
WP5_INF.metabat.8	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__Providencia stuartii_A
WP5_INF.metabat.9	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacterales	f__Enterobacteriaceae	g__Providencia	s__

On a taxonomic class level this related to 4 Actinomycetia, 7 Bacilli and 23 Gammaproteobacteria. Only one MAG could not be assigned at an order level with the others distributed across Enterobacterales (n=18), Legionellales (n=4), Mycobacteriales (n=4), Paenibacillales (n=6) and Pseudomonadales (n=1). There were 6 different family classifications with only one MAG not classified at a family level. The family taxonomic distribution was as follows: 3 *Aeromonadaceae*, 15 *Enterobacteriaceae*, 4 *Legionellaceae*, 4 *Mycobacteriaceae, 6 Paenibacillaceae* and 1 *Pseudomonadaceae.* At the level of genus, 3 MAGs could not be assigned and 8 at a species level. The genera identified consisted of *Aeromonas* (n=3), *Escherichia* (n=1), *Legionella* (n=4), *Morganella* (n=2), *Mycobacterium* (n=4), *Paenibacillus*_L (n=4), *Providencia* (n=12) and *Pseudomonas* (n=1). A total of 14 different species found with *Legionella pneumophila*, *Paenibacillus*_L sp007679495 and *Providencia stuartii*_A the highest occurring species detect with 4 representatives each. The other species detected were *Aeromonas caviae* (n=2), *Aeromonas hydrophila* (n=1), *Escherichia flexneri* (n=1), *Morganella morganii* (n=2), *Mycobacterium mageritense* (n=1), *Mycobacterium phocaicum* (n=3), *Providencia alcalifaciens* (n=1), *Providencia rettgeri* (n=1), *Providencia stuartii* (n=1) and *Pseudomonas nitroreducens* (n=1).

7.3.2 *Legionella pneumophila* MAGs

The MAGs were inspected for the presence of *L*. *pneumophila*. A total of 4 *L*. *pneumophila* medium quality MAGs were found originating from samples across all sample types (influent, activated sludge and secondary settling tank (SST) effluent). Resistance to Spectinomycin by means of aminoglycoside O-phosphotransferase APH(9)-Ia was found in all 4 the MAGs. The 4 *L*. *pneumophila* MAGs were found to have a high incidence of virulence factors with WP1_INF.metabat.4 containing 397, 418 in WP2_A5.metabat.8., 438 in WP2_SST.metabat.1 and 434 in WP3_INF.metabat.7.

7.3.3 *Aeromonas* spp. MAGs

Aeromonas spp. included *Aeromonas caviae* (n=2) and *Aeromonas hydrophila* (n=1). Table 7-4 details the AMRs found in the *Aeromonas* spp. MAGs. Virulence factors in the 3 *Aeromonas* spp. were found to be the highest in the *Aeromonas caviae* MAGs. There were 81 virulence factors in WP1_INF.metabat.9 and 79 in WP3_INF.metabat.2. The Aeromonas hydrophila (MAG WP3_A5.metabat.1) contained 54 virulence factors.

Table 7-4: AMRs detected in Aeromonas spp.

MAG	Gene	Coverage (%)	Identity (%)	Product	Resistance
WP1_INF.metabat.9	aph(6)-Id	100.00	100.00	aminoglycoside O-phosphotransferase APH(6)-Id	STREPTOMYCIN
WP1_INF.metabat.9	aph(3")-Ib	100.00	99.88	aminoglycoside O-phosphotransferase APH(3")-Ib	STREPTOMYCIN
WP1_INF.metabat.9	blaMOX-4	100.00	95.65	CMY-1/MOX family class C beta-lactamase MOX-4	CEPHALOSPORIN
WP1_INF.metabat.9	blaOXA-1143	100.00	96.48	class D beta-lactamase OXA-1143	BETA-LACTAM
WP3_A5.metabat.1	blaOXA-724	100.00	99.75	OXA-12 family class D beta-lactamase AmpH/OXA-724	BETA-LACTAM
WP3_A5.metabat.1	cepH	99.91	99.65	cephalosporin-hydrolyzing class C beta-lactamase CepH	CEPHALOSPORIN

MAG	Gene	Coverage (%)	Identity (%)	Product	Resistance
WP3_A5.metabat.1	imiH	88.89	97.21	ChpA family subclass B2 metallo-beta-lactamase ImiH	CARBAPENEM
WP3_INF.metabat.2	aph(6)-Id	100.00	100.00	aminoglycoside O-phosphotransferase APH(6)-Id	STREPTOMYCIN
WP3_INF.metabat.2	aph(3")-Ib	100.00	99.88	aminoglycoside O-phosphotransferase APH(3")-Ib	STREPTOMYCIN
WP3_INF.metabat.2	blaMOX-4	100.00	95.65	CMY-1/MOX family class C beta-lactamase MOX-4	CEPHALOSPORIN
WP3_INF.metabat.2	blaOXA-1143	100.00	96.48	class D beta-lactamase OXA-1143	BETA-LACTAM

7.3.4 *Mycobacterium* spp. MAGs

The 4 *Mycobacterium* spp. MAGs were identified as *Mycobacterium mageritense* (n=1) and *Mycobacterium phocaicum* (n=3). Sample WP3_INF (influent), MAG metabat.8, classified as *Mycobacterium mageritense*, contained 4 different AMRs (Table 7-5). In comparison with the 2 other species above, the *Mycobacterium* spp. MAGs contained much less virulence factors. The highest number of virulence factors, 10, were found in the *Mycobacterium mageritense* MAG (WP3_INF.metabat.8.). The other three *Mycobacterium phocaicum* MAGs contained between 5 and 6 virulence factors.

Table 7-5: AMRs detected in *Mycobacterium mageritense.*

MAG	Gene	Coverage (%)	Identity (%)	Product	Resistance
WP3_INF.metabat.8	erm(40)	100.00	97.35	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(40)	MACROLIDE
WP3_INF.metabat.8	tet(V)	97.30	81.24	tetracycline efflux MFS transporter Tet(V)	TETRACYCLINE
WP3_INF.metabat.8	aac(2')-Ib	90.31	84.53	aminoglycoside N-acetyltransferase AAC(2')-Ib	GENTAMICIN;TOBRAMCYIN
WP3_INF.metabat.8	vanR-O	94.87	84.38	vancomycin resistance response regulator transcription factor VanR-O	VANCOMYCIN

7.4 DISCUSSION

Metagenomic sequencing data allows researcher to reconstruct metagenome assembled genomes (MAGs). These MAGs are obtained based on the most current bioinformatic workflows and provides the ability to taxonomically classify the partial to near complete genomes as found in wastewater samples. This process is especially useful when interested in microorganism which are difficult to isolate and cultivate. The process of isolation and cultivation is needed to perform whole genome sequencing. When this option is not feasible or the microorganisms are viable but not culturable a metagenomic pipeline can be followed. Using metagenome sequencing it was possible to reconstruct 34 MAGs with high to medium quality genomes. These MAGs were assigned at a phylum level as Proteobacteria, Firmicutes and Actinobacteriota which is congruent with literature.

Of particular interest were the species *Legionella pneumophila*, *Mycobacterium* spp. and *Aeromonas* spp. which were further analysed for antimicrobial resistance and virulence factors. A total of 4 *Legionella pneumophila* MAGs were found in all the sample types, i.e. nfluent, activated sludge and secondary settling tank (SST) effluent) and all the MAGs included the presence of O-phosphotransferase APH(9)-Ia which confers resistance to Spectinomycin. The *L*. *pneumophila* MAGs included a wide range of virulence factors. The *Aeromonas* spp. MAGs included *Aeromonas caviae* (n=2) and *Aeromonas hydrophila* (n=1). These MAGs had a much larger range of AMRs than the *L*. *pneumophila* MAGs but in comparison had much less virulence factors. The 4 *Mycobacterium* spp. MAGs were identified as *Mycobacterium mageritense* (n=1) and *Mycobacterium phocaicum* (n=3) with the *M. mageritense* MAG containing 4 different AMR genes with different resistance phenotypes.

It is evident that metagenomic sequencing and the construction of metagenome assembled genomes provides researchers with high resolution results. Numerous microorganisms, including pathogens, may be difficult to isolate and sequence individually. Using metagenomics, it is possible to reconstruct these notoriously difficult microorganisms on a high to medium quality genome level. Thereafter the MAGs may be analysed for the presence of AMRs and virulence factors. This method negates the time-consuming and laborious alternative protocols and provides researchers with a wealth of information per sample or sequencing event.

The data generated during this part of the project is supporting Ms E. Poopedi in pursuit of her PhD degree at the University of the Witwatersrand. She is currently busy with additional analyses and the results will form part of her thesis and publications.

CHAPTER 8: AMPLICON AND METAGENOMIC SEQUENCING OF WASTEWATER SAMPLES FROM THE EAST RAND OF GAUTENG

8.1 INTRODUCTION

Wastewater-Based Epidemiology (WBE), commonly known as wastewater analysis or sewage epidemiology is a popular method used to monitor the wastewater composition of a particular region to detect specific chemical compounds or determine the microbial composition (Mackuľak et al., 2021). This technique was initially used to determine the use of illicit drugs (Gao et al., 2015) in different municipal regions by detecting specific metabolites in the water. It has however evolved as a tool to monitor various metabolites and chemicals for a particular population to answer questions about the livelihood of the people that form part of it (Erickson et al., 2021). An important extension of this method is the use of metagenomics to also assess the microbiome that represent a population from a particular area.

The quantitative measure of specific biomarkers in wastewater can be used to evaluate the lifestyles of people from different regions, such as the type of diet the majority follows and how this could be influencing their health and the incidence of diseases (Picó & Barceló, 2021). The wastewater from hospitals can be used to detect the type of antibiotic resistance genes that exist in the area and for the surveillance of pathogens such as SARS-CoV-2 (Erickson et al., 2021; Mackuľak et al., 2021; Picó & Barceló, 2021). Environmental contamination of pesticides and mycotoxins can also be determined by analysing the wastewater of that location. This method of monitoring the lifestyles of populations and the state of the environment is not always welcomed because it can reveal the negligence of the city when it comes to taking care of the environment. Despite the disadvantage, WBE is beneficial as it makes it possible to monitor individual communities, combating the expenses that come with individual sampling and sequencing.

The advancements of sequencing technologies and the increase in the use of culture independent methods to study various environments (Escobar-Zepeda et al., 2015), has led to an increased curiosity to study the microbial communities present in diverse environments like the activated sludge of wastewater treatment plants (WWTPs) (Wu et al., 2019; Ji et al., 2020). Previous studies have reported that WWTPs contain a diverse community of microorganisms ranging from archaea, fungi, bacteria, and viruses. Majority of the previous and current studies in this environment are focused on understanding the bacterial community compositions and functions (Silva-Bedoya et al., 2016; Ji et al., 2020). This is because it has been established that bacteria play the major role in the removal of organic waste from the water. Most of these studies have however been carried out in first world countries, leaving a research gap in developing countries of major continents like Africa and Asia, where majority of the world population resides (Abdill et al., 2022; Dueholm et al., 2022). Treated wastewater effluent is often released into water bodies such as rivers and dams where it ends up being used for human activities. It is therefore important to know the microbial composition of influent and effluent of a wastewater treatment centre. This will allow us to determine how effective the treatment process is at removing microbes from the influent, because some of these microorganisms can be pathogenic to plants, animals and human beings once released into the environment (Abia et al., 2018).

Many countries, especially in the global north have conducted studies with the aim of determining the core microbial taxa in wastewater systems. The results generated show similar trends in microbial composition especially for bacterial groups. Studies that sampled from municipal wastewater report that the most dominant bacterial phylum is Proteobacteria. Cydzik-Kwiatkowska & Zielińska (2016) reported 21-65%, Bedoya et al. (2020) a 9-23%, and a 17-31% dominance in the Zhang et al. (2019) study. Numerous studies report similar results, and this suggests that this phylum is very important in wastewater ecology. For the second most

dominant phylum in wastewater systems, most studies report different results, but the option is always almost between Chloroflexi, Bacteroidetes, Acidobacteria, and Firmicutes (Yang et al., 2014; Abia et al., 2018; Zhang et al., 2018; Zhang et al., 2019). In class ranking the dominant groups belonging to the Proteobacteria phylum and are categorised as Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, and Epsilonproteobacteria (Huo et al., 2017).

The human microbiota is the combination of all the microorganisms (bacteria, fungi, archaea, and viruses) that exist within and on the human body. This includes the skin, the gut and reproductive organs. The gut microbiota has been the subject of most studies in the past decade or more in this area of study. Most of these studies have reported similar and sometimes identical results with regards to the microbial composition of the human gut, focusing specifically on the bacterial kingdom. The Bacteroidetes and Firmicutes are the most dominant phyla in the gut and forms part of the core taxa of all healthy individuals. Different factors are known to affect the composition of the gut microbiota, these include diet, age, sex, disease, pre and probiotic use and many others. The advancements in sequencing technology have allowed researchers to study the gut microbiota without relying on culture dependent methods. They are still met with difficulties due to the intensive work and time consumption that comes with sampling individuals' waste in a population to study the gut. It was however revealed in recent studies that the wastewater of a region can be a good reflection of the human gut through the faecal-derived waste that enter the wastewater systems. Which is not a surprise because wastewater systems have been used in the past and even in the present to detect illicit drugs such as cocaine and make deductions on the drug use of a specific region. This technique is known as waste-water based epidemiology/wastewater surveillance. With this surveillance system we can detect human pathogens, antibiotic resistance genes and make inferences about the human gut of a specific population without manual individual sampling and the high cost of sequencing that comes with it.

In this chapter, a large cohort of samples is analysed using amplicon and shotgun metagenomic sequencing to investigate the taxonomic and antimicrobial resistance profiles of samples obtained from the East Rand, Gauteng. This method enables comparisons between treatment plants and date of collection, further serving as a proxy for the community gut microbiome in a certain region at a specific time.

8.2 MATERIALS AND METHODS

Wastewater samples (n=72) were collected from 8 WWTPs located in the East Rand of Gauteng (Mr. W. le Roux). These samples were collected weekly between 26 January 2022 and 22 March 2022 and represent 9 sampling dates (Table 8-1).

Table 8-1: Samples collected from the East Rand of Gauteng.

SampleID	Location	Date
B1A	Daveyton WCW	2022/01/26
B1B	Olifantsfontein WCW	2022/01/26
B1C	Vlakplaats WCW	2022/01/26
B1D	Carl Grundlingh WCW	2022/01/26
B1E	Herbert Bickley WCW	2022/01/26
B1F	Jan Smuts WCW	2022/01/26
B1G	JP Marais WCW	2022/01/26
B1H	Rynfield WCW	2022/01/26
B2A	Daveyton WCW	2022/03/01
B2B	Olifantsfontein WCW	2022/03/01
B2C	Vlakplaats WCW	2022/03/01
B2D	Carl Grundlingh WCW	2022/03/01
B2E	Herbert Bickley WCW	2022/03/01
B2F	Jan Smuts WCW	2022/03/01

SampleID	Location	Date
B2G	JP Marais WCW	2022/03/01
B2H	Rynfield WCW	2022/03/01
B3A	Daveyton WCW	2022/03/08
B3B	Olifantsfontein WCW	2022/03/08
B3C	Vlakplaats WCW	2022/03/08
B3D	Carl Grundlingh WCW	2022/03/08
B3E	Herbert Bickley WCW	2022/03/08
B3F	Jan Smuts WCW	2022/03/08
B3G	JP Marais WCW	2022/03/08
B3H	Rynfield WCW	2022/03/08
B4A	Daveyton WCW	2022/02/08
B4B	Olifantsfontein WCW	2022/02/08
B4C	Vlakplaats WCW	2022/02/08
B4D	Carl Grundlingh WCW	2022/02/08
B4E	Herbert Bickley WCW	2022/02/08
B4F	Jan Smuts WCW	2022/02/08
B4G	JP Marais WCW	2022/02/08
B4H	Rynfield WCW	2022/02/08
B5A	Daveyton WCW	2022/03/15
B5B	Olifantsfontein WCW	2022/03/15
B5C	Vlakplaats WCW	2022/03/15
B5D	Carl Grundlingh WCW	2022/03/15
B5E	Herbert Bickley WCW	2022/03/15
B5F	Jan Smuts WCW	2022/03/15
B5G	JP Marais WCW	2022/03/15
B5H	Rynfield WCW	2022/03/15
B6A	Daveyton WCW	2022/03/22
B6B	Olifantsfontein WCW	2022/03/22
B6C	Vlakplaats WCW	2022/03/22
B6D	Carl Grundlingh WCW	2022/03/22
B6E	Herbert Bickley WCW	2022/03/22
B6F	Jan Smuts WCW	2022/03/22
B6G	JP Marais WCW	2022/03/22
B6H	Rynfield WCW	2022/03/22
B7A	Daveyton WCW	2022/02/01
B7B	Olifantsfontein WCW	2022/02/01
B7C	Vlakplaats WCW	2022/02/01
B7D	Carl Grundlingh WCW	2022/02/01
B7E	Herbert Bickley WCW	2022/02/01
B7F	Jan Smuts WCW	2022/02/01
B7G	JP Marais WCW	2022/02/01
B7H	Rynfield WCW	2022/02/01
B8A	Daveyton WCW	2022/02/23
B8B	Olifantsfontein WCW	2022/02/23
B8C	Vlakplaats WCW	2022/02/23
B8D	Carl Grundlingh WCW	2022/02/23
B8E	Herbert Bickley WCW	2022/02/23
B8F	Jan Smuts WCW	2022/02/23
B8G	JP Marais WCW	2022/02/23
B8H	Rynfield WCW	2022/02/23
B9A	Daveyton WCW	2022/02/15

SampleID	Location	Date
B9B	Olifantsfontein WCW	2022/02/15
B9C	Vlakplaats WCW	2022/02/15
B9D	Carl Grundlingh WCW	2022/02/15
B9E	Herbert Bickley WCW	2022/02/15
B9F	Jan Smuts WCW	2022/02/15
B9G	JP Marais WCW	2022/02/15
B9H	Rynfield WCW	2022/02/15

Raw wastewater samples were collected at the inlet works of each relevant treatment plant. Samples were collected after coarse (grid size 6mm) and fine (grid size 4mm) screening using a stationary composite sampler set to collect samples every hour over a 24-hour period (composite samples). If a composite sample could not be collected, grab sampling was performed. One litre samples were collected in polyethylene terephthalate (PET) bottles and used for DNA extraction. Preparation and extraction protocol for the wastewater samples were as follows:

- Sample preparation: 200 ml of each sample was filtered through a Macherey-Nagel Glass Fiber Filter 45 mm (EO- treated). Filter was dissolved in 3 ml deionized water.
- Extraction protocol: Macherey-Nagel Genomic DNA from stool samples protocol (https://www.mn-net.com/media/pdf/e3/88/69/Instruction-NucleoSpin-DNA-Stool.pdf) was used, with the following changes:
    i. 700 µl ST1 buffer used to dissolve 300 µl of sample.
    ii. 30 µl SE buffer used for DNA elution.

The extracted DNA was for both amplicon and metagenomic sequencing (Supplementary Sequencing Quotation). 16S ribosomal RNA (16S rRNA) amplification and sequencing were performed according to the Illumina 16S protocol (16S Metagenomic Sequencing Library Preparation Guide). Briefly, the variable V3 and V4 regions of the 16S rRNA gene were amplified primers from Klindworth et al. (2013) from the samples, followed by library amplification and sequencing on the Illumina MiSeq instrument using V3 chemistry. The primer sequence was as follows: 16S forward primer = 5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG and 16S Reverse primer = 5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC. The PCR program was as follows: 95 °C for 3 min, 25 cycles of; 95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s and a final extension at 72 °C for 5 min, held at 4 °C. Generated data were evaluated for quality and used for downstream bioinformatic pipelines. Low-quality sequencing reads were filtered and trimmed to a consistent length with a maximum of 2 expected errors per-read enforced (Edga and Flyvbjerg, 2015). This is done on paired reads jointly, after which amplicon sequence variants are inferred and downstream analysis is done using the DADA2 method (Callahan et al., 2016).

This method combines identical sequencing reads into “unique sequences” with a corresponding abundance value followed by the identification of sequencing errors. Thereafter the forward and reverse reads are merged, and paired sequences that do not perfectly overlap are discarded. The resulting sequence table was inspected for chimeras which were removed. Taxonomy was assigned to the final, filtered sequence table using the SILVA ribosomal RNA gene reference database (Quast et al., 2012). The R package, phyloseq (McMurdie et al., 2013), was used to further analyze and graphically display the sequencing data which was clustered into amplicon sequence variants (ASVs) with the protocol described above. Detailed analyses methodology is further available in Supplementary Dineo Raphela BScHons(Genetics). The DNA metagenomic sequencing was done on the same extraction used for the above amplicon method and sequenced on a HiSeq 2500. For each sample 8 GB of data or roughly 20,000,000 reads were requested. Initial sequence data quality and filtered data quality was inspected using FastQC version 0.11.8 (Andrews, S., 2010). Sequence data was quality trimmed and filtered, including adapter removal and decontamination, using BBDuk version 38.91 available from the BBTools suite of tools (Bushnell, B., 2014). Filtered reads were assembled using SPAdes v.3.15.3 (Nurk et al., 2017) and only contigs with length exceeding 1,500 bp used for further analyses. ABRicate (https://github.com/tseemann/abricate) was used to detect antimicrobial resistance genes. Abricate allows for

the mass screening of contigs for AMR genes. This program only detects acquired resistance and is not suitable for the detection of point mutations. Abricate was run with default parameters and the “ncbi” database selected. This database was locally updated 2023/01/05 and at time of usage included 6,334 sequences. The output from Abricate includes AMR gene name and putative antibiotic resistance phenotype.

8.3 RESULTS

All samples were successfully sequenced with the exception of B9A, collected on 2022/02/15 from Daveyton WCW. This sample has been resubmitted for sequencing and will be included in future analyses and publications. Genes conferring resistance were found in all samples, the lowest being 1 AMR gene and the highest 58 (Jan Smuts WCW, 2022/02/01) (Figure 8-1). On average, there were 19.61 AMR genes present across all the samples. A total of 221 different AMR genes were found across all the samples. No significant differences in the number of AMR genes were found between the treatment plants (Figure 8-2) when all sampling dates grouped per sampling location. A minimum spanning tree based on the presence/absence of each AMR gene displayed clustering but this could not be associated with a particular sampling location (Figure 8-3).

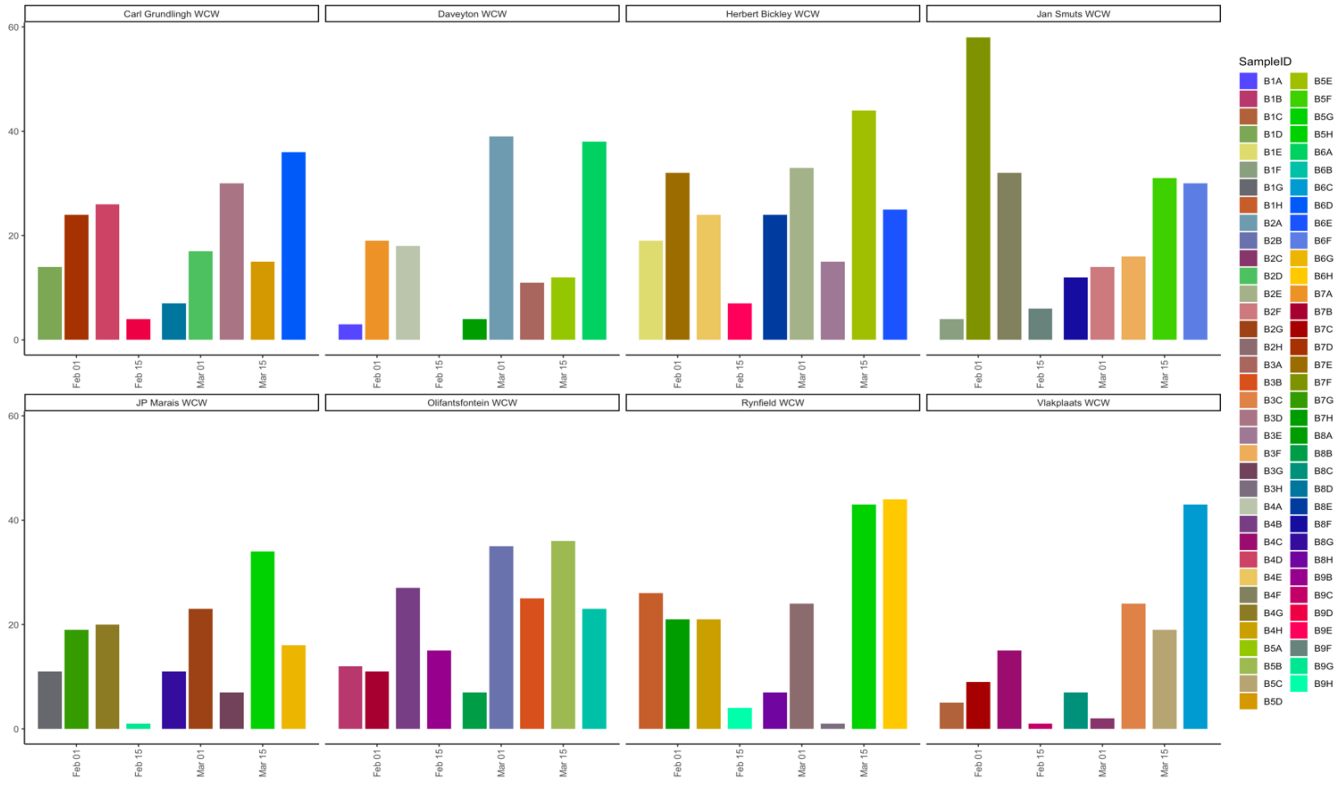


Figure 8-1: Number of AMR genes found per sample. The figure is grouped according to sampling location with date of collection on the x-axis. The y-axis for each sub-graph represents the number of AMR genes and each sample has a different colour.

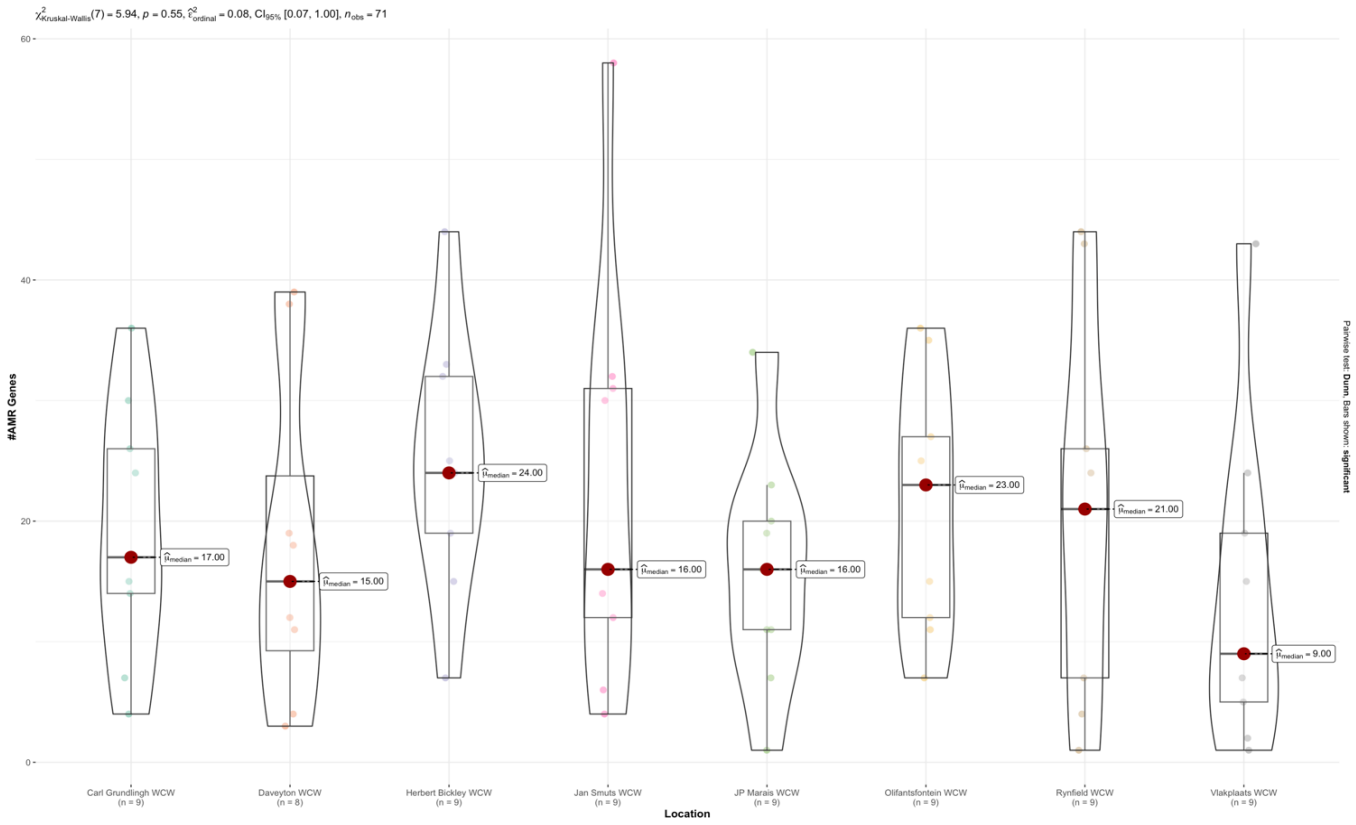


Figure 8-2: Number of AMR genes per sampling location. No significant difference between the treatment plants were found.

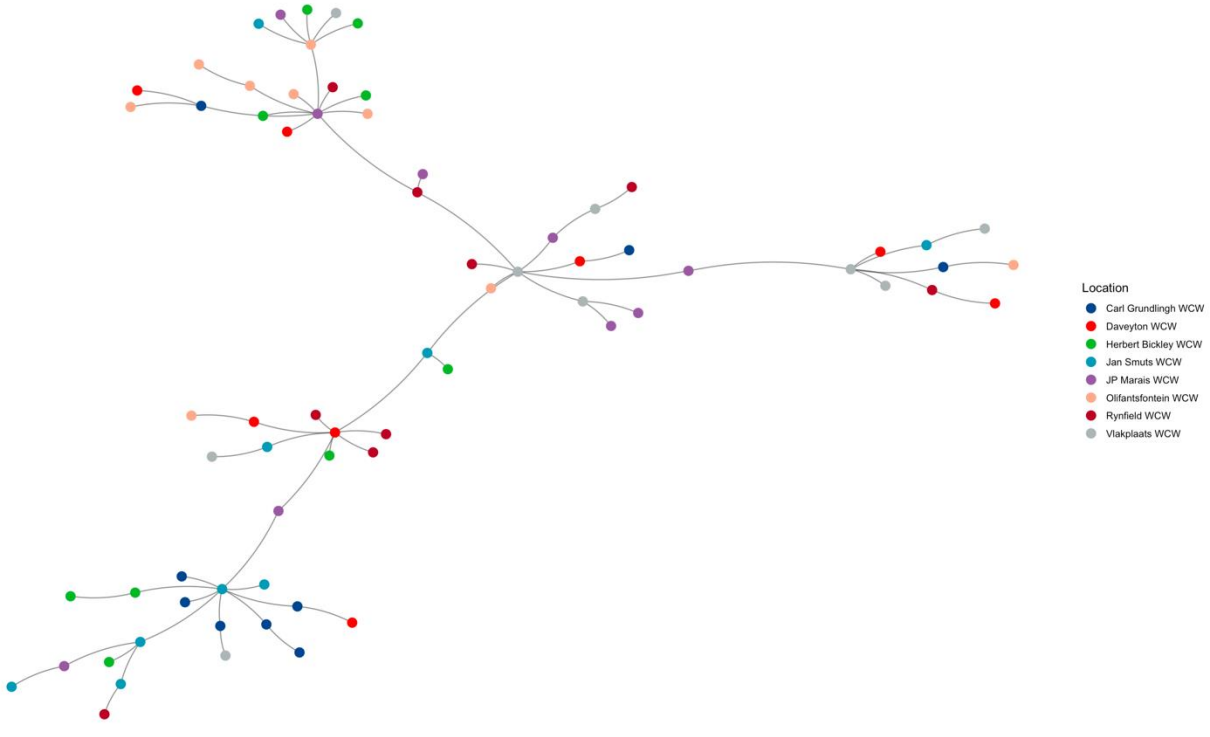


Figure 8-3: Minimum spanning tree based on the presence/absence of AMR genes. Each dot represents a sample and is coloured by the sampling location.

Different AMR phenotypes were found across all samples (Figure 8-4). The highest occurring AMR phenotypes were Macrolide and Tetracycline and in total 32 different AMR phenotypes were detected. In Figures 8-1 – 8-4 samples are pooled according to treatment plant. Each treatment plant thus had a sample included for the

specific sampling week and as such there were 9 samples per treatment plant. From Figure 8-1 – 8-4 there is no clear trend with regards to the treatment plants. They all have relatively the same amount of AMR genes and the incidence of these fluctuate during the 9 sampling weeks. In Figure 8-1 and Figure 8-4 all treatment plants start with relatively high levels of AMR genes during the first sampling week which generally remains constant.

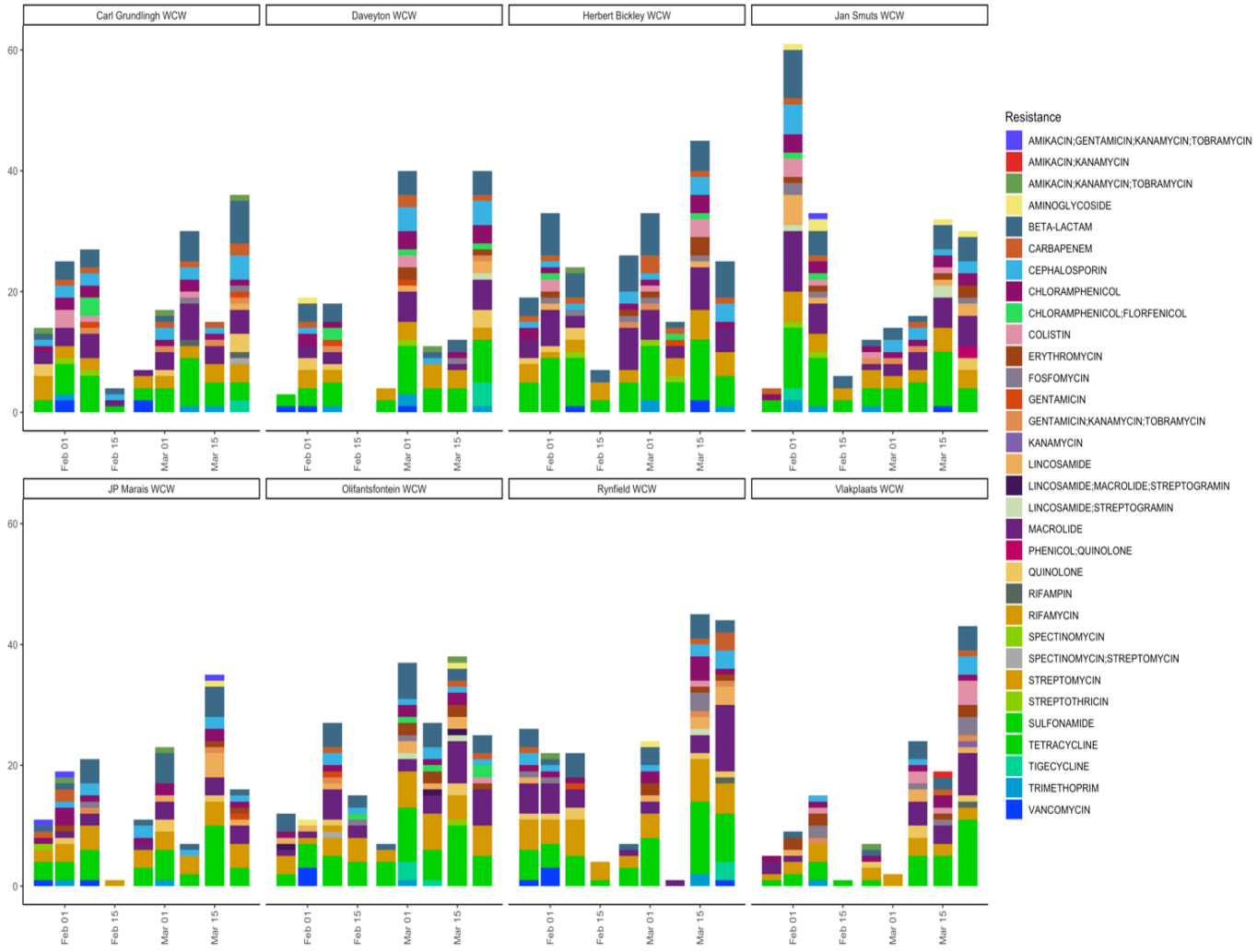


Figure 8-4: Number of AMR phenotypes found per sample. The figure is grouped according to sampling location with date of collection on the x-axis. The y-axis for each sub-graph represents the number of AMR phenotypes with different colours for each phenotype.

A drop in AMR gene levels is seen across all samples during the fourth sampling week (15 February 2022) which is then followed by a steady increase in the number of AMR genes for the consequent sampling weeks (Figure 8-5). This occurrence needs to be further analysed in conjunction with the amplicon data to obtain a clear picture of what is happening. It should be mentioned that the sample which failed metagenomic sequencing was part of sampling week 4 and should be included in future analyses to negate any bias. In Figures 8-5 – 8-9 the samples were pooled according to sampling date. Each sampling date therefor had a representative from each one of the sampling locations. A sampling date consists of 8 treatment plants.

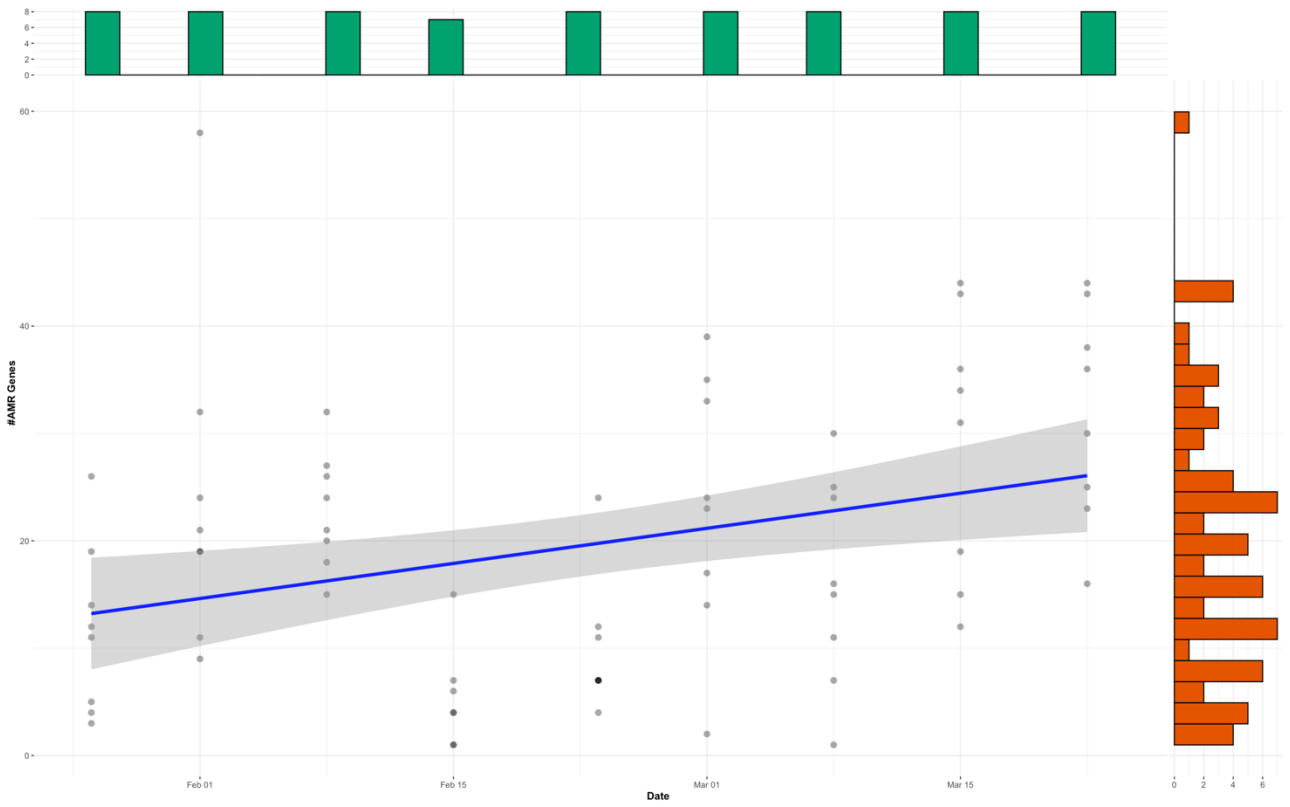


Figure 8-5: Scatter plot of the number of AMR genes per sampling period.

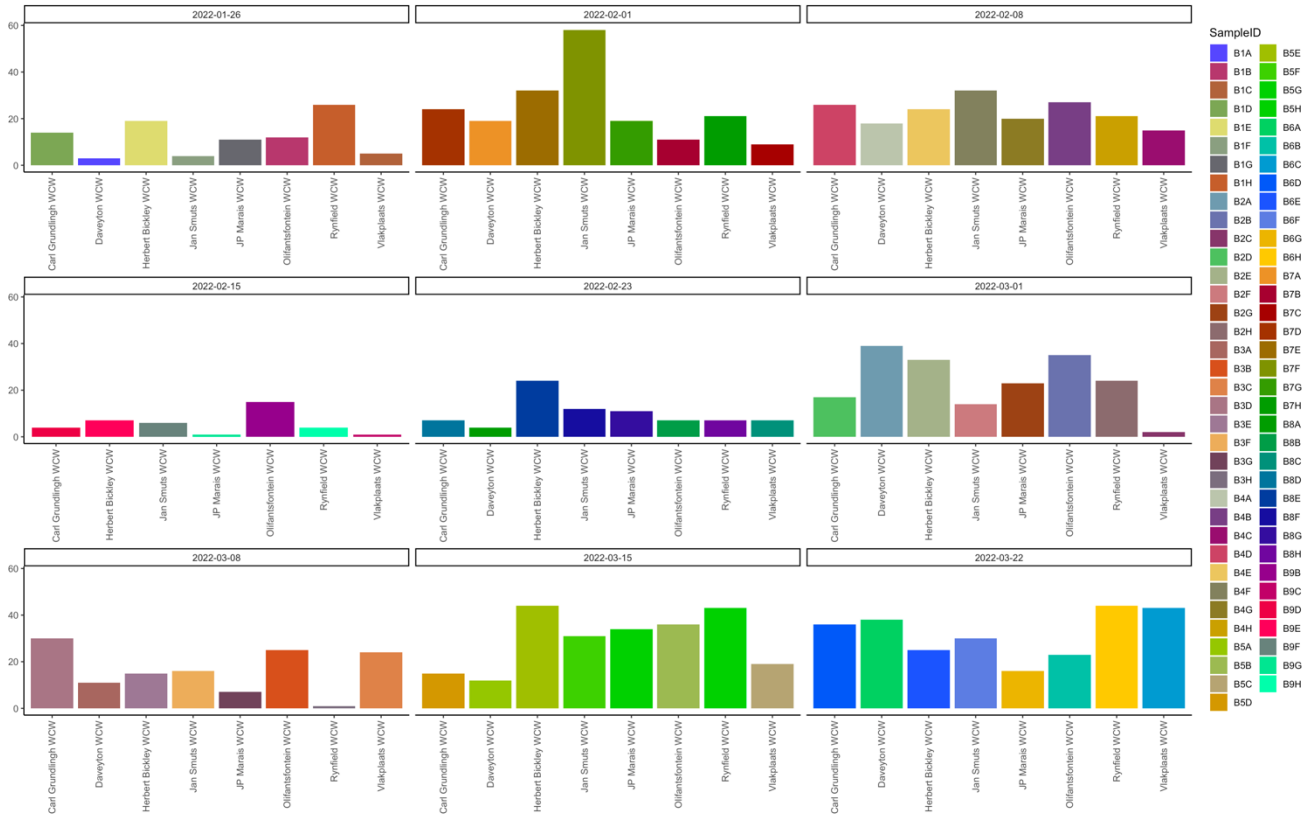


Figure 8-6: Number of AMR genes found per sample. The figure is grouped according to sampling date with location of collection on the x-axis. The y-axis for each sub-graph represents the number of AMR genes and each sample has a different colour.

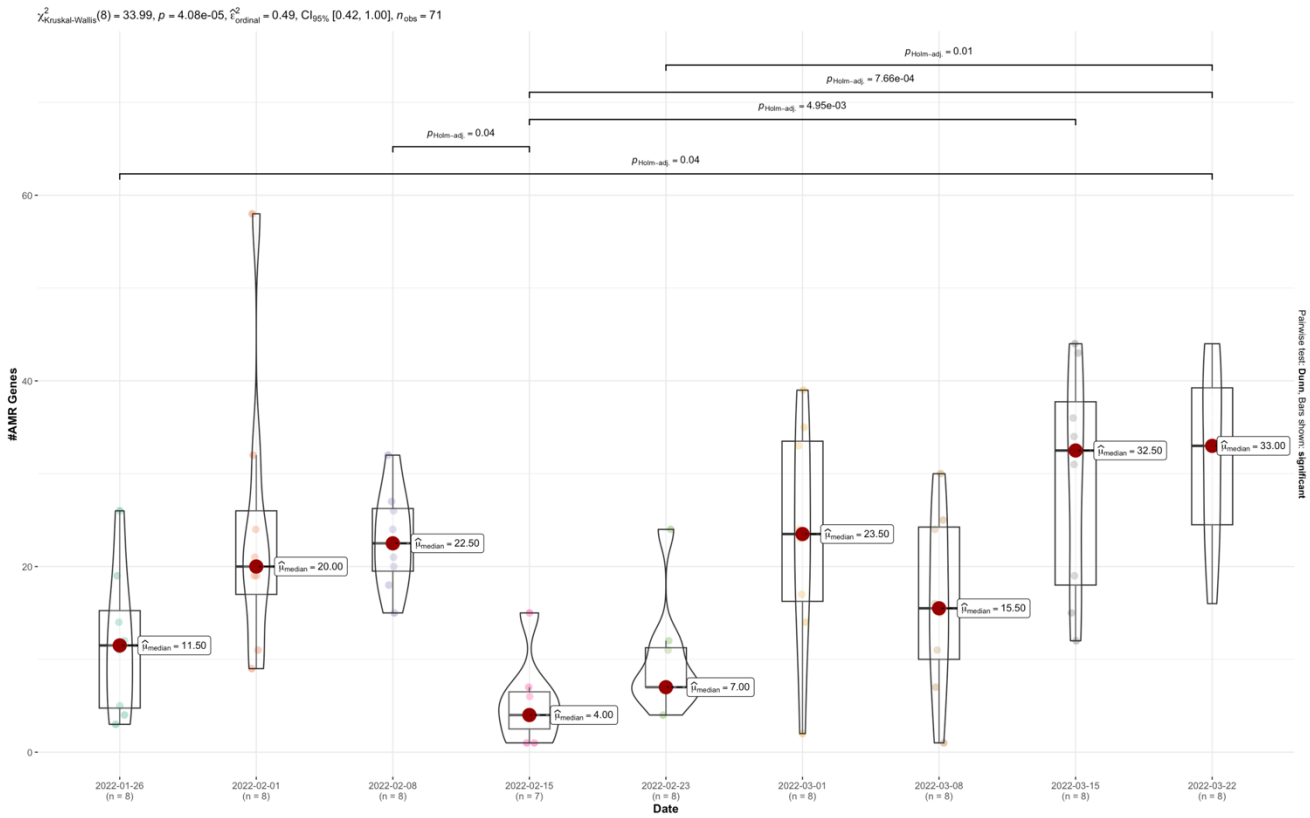


Figure 8-7: Number of AMR genes per sampling date. Significant differences (after p-value adjustment) are indicated by lines on top of the graph.

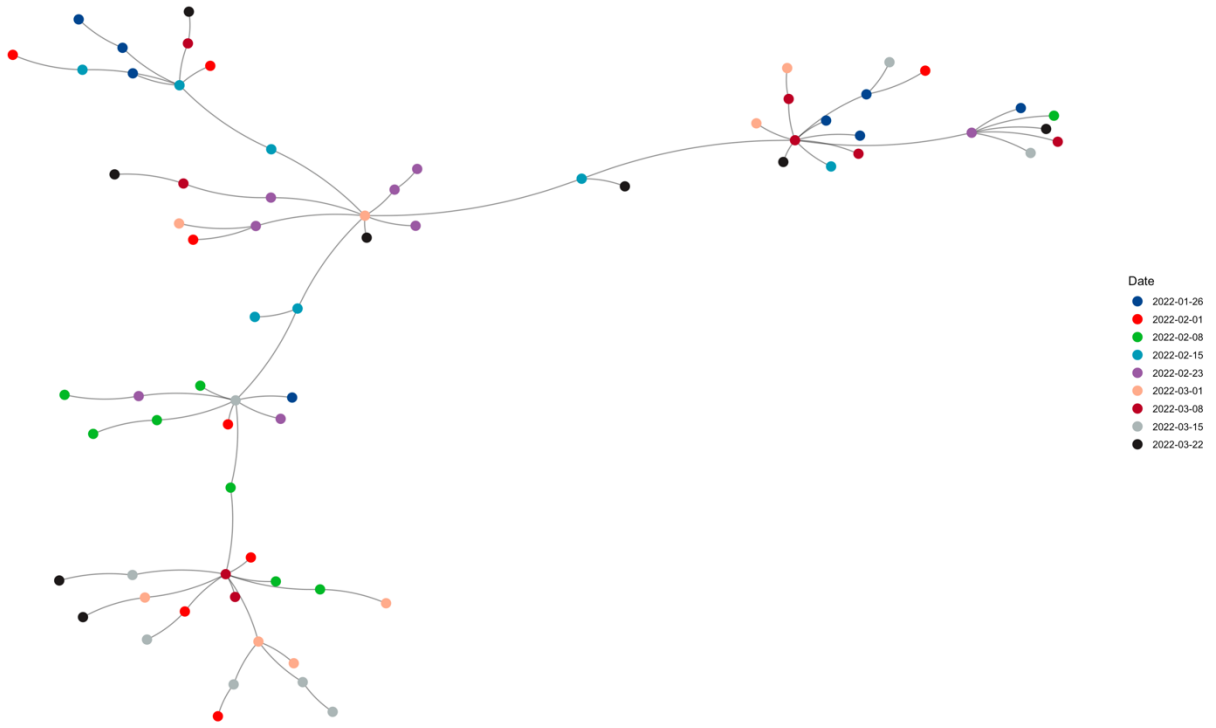


Figure 8-8: Minimum spanning tree based on the presence/absence of AMR genes. Each dot represents a sample and is coloured by the sampling date.

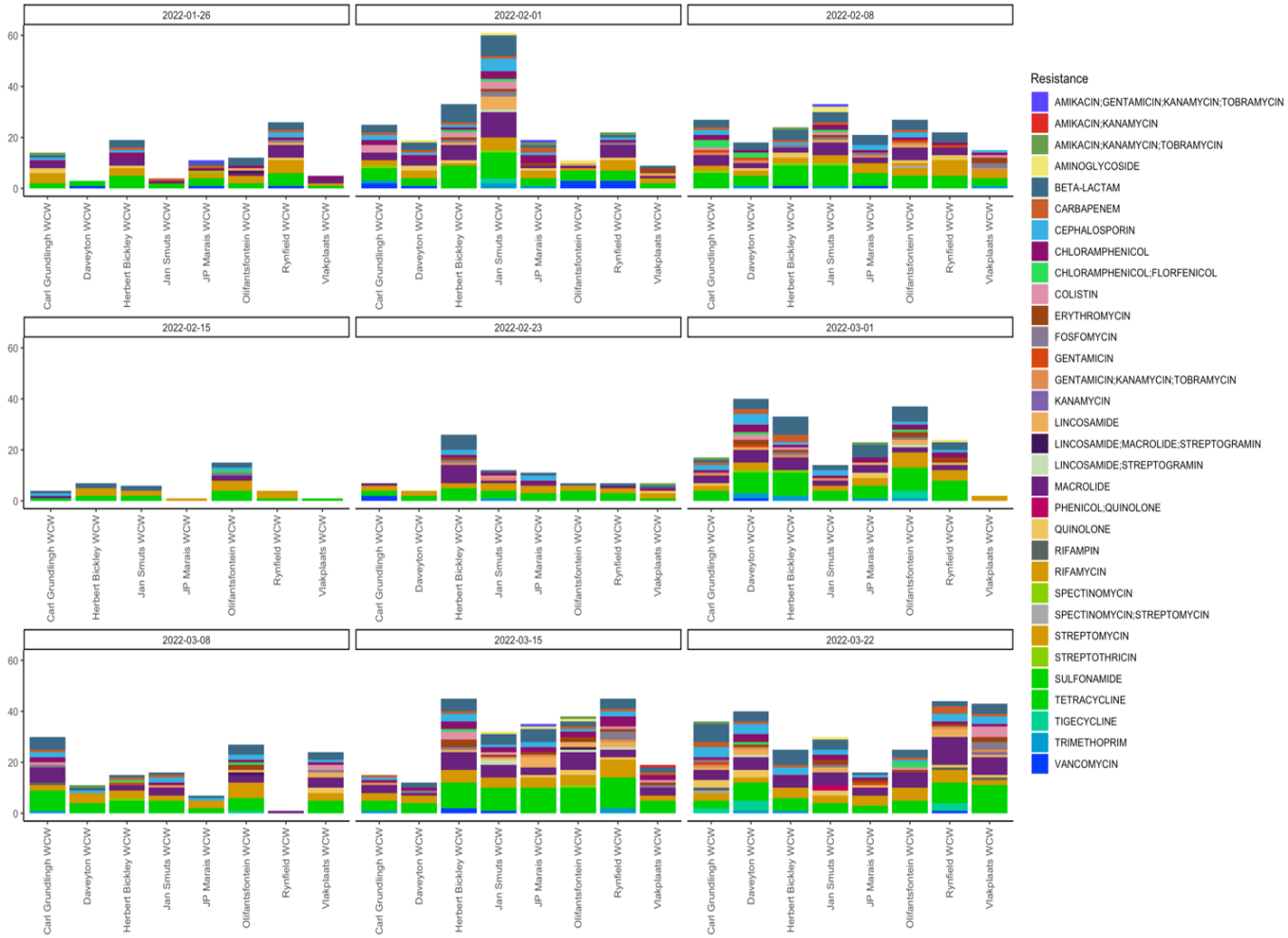


Figure 8-9: Number of AMR phenotypes found per sample. The figure is grouped according to sampling date with location of collection on the x-axis. The y-axis for each sub-graph represents the number of AMR phenotypes with different colours for each phenotype.

In Figure 8-6 and Figure 8-8 which is based on grouping the sampling locations per sample date it can be seen that all the treatment plants, with one or two exceptions have similar levels of AMR genes per date. The low AMR gene presence during the fourth sampling week is also evident. The significant differences between sampling dates are clearly evident in Figure 8-7. In general, the differences are significant between the earlier and later sampling dates. Samples from the last sampling week contained significantly more AMR genes than those collected during the middle sampling weeks. There is further a significant decrease in the number of AMR genes between sampling week 3 and 4. The significant increase in AMR gene numbers between the first and last sampling period is also of interest.

8.4 DISCUSSION

8.4.1 Amplicon Sequencing

The findings in this study regarding the bacterial community composition in wastewater systems are consistent with the findings of other studies performed in the past, especially in the first world countries. These findings support the fact that Proteobacteria, Firmicutes, Bacteroidetes, Acidiobacteria and Chloroflexi are the most abundant phyla found in wastewater (Yang et al., 2014; Silva-Bedoya et al., 2016; Liu et al., 2017; Zhang et al., 2019; Bedoya et al., 2020; Ji et al., 2020; Xiang et al., 2021; Zhang et al., 2022). The 16S rRNA gene is quite variable and different primers (V1-V3, V3-V4, V4, etc.) can be used during amplification in PCR. As such it is important to note that the bacterial community composition will be affected (Albertsen et al., 2015) if a different primer is to be used for the study. The core dominant phyla observed will most likely not be affected by primer changes. The microbial diversity between the WWTPs was measured using different indices such as the Shannon, Simpson, and Chao1 (Table 2 in Supplementary information). The results suggest that the microbial diversity between all the WWTPs is similar, with very little differences. There were two WWTPs (Herbert Bickley and Jan Smuts), that had a slightly higher alpha diversity measure (Figure 2, Figure 7 in supplementary information). It is possible to hypothesize that the higher diversity observed in Jan Smuts is because of where it is located. This WWTP receives influent from the biggest and busiest airport in South Africa (O.R. Tambo International Airport) where passengers between all six inhabited continents land. Very few variables were considered in this study, and that affected the extent in which we could explore the different factors that may or may not affect microbial diversity between the different WWTPs. The time variable was considered, and the findings while preliminary suggest that the microbial diversity does not change within a shorter time-scale (Figure 5). This suggests that sampling from these areas can be done once in a while. This will however have to be tested again in a future study, by taking into consideration other factors like physiochemical properties and their changes, as well as the microbial diversity changes from other microbial communities such as fungi and archaea (Liu et al., 2017).

Microorganisms are an important component of our lives and the different ecosystems we have on earth. How these microorganisms interact with one another, and other organisms can help us determine their importance in different environments. An important example of these environments is the complex wastewater systems that harbour an abundance of microorganisms from bacteria and archaea to fungi and viruses. These microbial communities are responsible for the removal of waste and pathogens from sewage, bacteria being the most abundant and important in this system. Most studies about microbial ecology of wastewater systems that have been made available to the public have been carried out in highly developed countries, leaving a big research gap in areas of the world where majority of the global population resides. As such, this study set out to characterise the microbial community structure of wastewater systems in Gauteng, South Africa. The results from this metagenomics-based study are consistent with studies performed in other parts of the world, that reveal that the most dominant bacterial phyla in wastewater systems are Proteobacteria, Firmicutes, Bacteroidetes and Acidiobacteria.

8.4.2 Metagenomic Sequencing

A large cohort of different AMR genes was detected across the samples with a high diversity of AMR phenotypes or putative resistance. The samples (n=71) could be grouped by treatment plant and date of sampling. No significant differences were detected between the sampling locations based on AMR gene numbers. This was of interest as the assumption was that certain treatment plants would have higher AMR gene presence based on the community and location it serves. Significant differences were found based on sampling date. For the first 3 weeks of sampling there was a gradual increase followed by a decrease in the number of AMR genes in the fourth sampling week. Thereafter a gradual increase was again detected. Significant differences (after p-value adjustment) were found between the first and last sampling week. This suggests that during the sampling period there was a large increase in the number of AMR genes between the first and last sampling date. A

significant decrease in the number of AMR genes was found between the third and fourth sampling week. These differences were further found between the middle sampling dates and the last sampling date. This data suggests that there is an increase in the incidence of AMR genes based on the date. This trend will be further investigated and external data included to identify the possible reason for this. An extended sampling period is further proposed to clearly evaluate any cyclical patterns.

CHAPTER 9: CONCLUSIONS AND RECOMMENDATIONS

This Final Technical and Data Report details the work done and results obtained for the amplicon, metagenomic and SARS-CoV-2 whole genome sequencing of wastewater samples under the project titled “Tracking the evolution of SARS-CoV-2 and the emergence of other infectious diseases in communities using a wastewater-based epidemiology approach”. This project aimed to harness the added value afforded by next-generation sequencing in answering various questions related to the presence of SARS-CoV-2, antimicrobial resistance and the microbial content of wastewater samples. The collaborators were all able to accomplish their individual mandates before the samples were passed on to this project. Obtaining samples in this method insured that there was no duplication of results and that the absolute maximum amount of information was extracted per sample in a strategic workflow.

This report highlights the functionality of next-generation sequencing and in particular targeted and untargeted sequencing in wastewater surveillance. The untargeted sequencing or metagenomic methodology was able to provide a holistic view on the taxonomic diversity found in wastewater samples. Furthermore, this methodology allows for the detection of antimicrobial resistance and associated classifications without the need of another data generation event. Although not the most feasible methodology to test for the presence of SARS-CoV-2 in wastewater samples it is still capable of recovering portions of the genome in samples with a high viral load. Data sets such as these contained within this report will greatly assist wastewater surveillance, disease modelling and the prediction of outbreak events.

Targeted sequencing as was used for SARS-CoV-2 whole genome sequencing in these wastewater samples was able to provide SARS-CoV-2 lineage assignments. SARS-CoV-2 whole genome sequencing is generally performed on clinical samples. The application thereof on wastewater samples and the ability to produce lineage assignments and near complete genomes clearly illustrates the functionality of this protocol. This method provides a clear picture on high prevalence SARS-CoV-2 variants as found in a community and has the possibility to detect an upsurge or prevalence of variants of concern.

Continuous monitoring of wastewater samples for the presence of AMR genes is critical in understanding the ebb and flow of these resistance elements in communities. The ability to construct metagenome assembled genomes with metagenomic sequencing data further allows us to classify the recipients of acquired resistance and better understand the spread of AMR in our population.

Metagenomic sequencing and analysis is a powerful tool in wastewater surveillance and epidemiology. The method allows for the taxonomic classification of the organisms present in a sample and furthermore the functional potential of the organisms in a sample. The amount of data generated in a single sequencing event can be used in various research questions and provides a holistic representation of the biological components in a system. The results obtained from metagenomic sequencing analysis will greatly assist in various public health concerns and the associated strategies to be followed in addressing the concerns. Whole genome sequencing and analysis is another powerful tool in wastewater surveillance and epidemiology. The method allows for SARS-CoV-2 lineage assignment and the construction of near complete SARS-CoV-2 genomes. Next-generation sequencing is clearly the future of wastewater-based epidemiological surveillance.

REFERENCES

Abdill, R.J., Adamowicz, E.M. and Blekhman, R., 2022. Public human microbiome data are dominated by highly developed countries. PLoS biology, 20(2), p.e3001536.

Abia, A.L.K., Alisoltani, A., Keshri, J. and 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. Science of the Total Environment, 616, pp.326-334.

Alcock, B.P., Raphenya, A.R., Lau, T.T., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.L.V., Cheng, A.A., Liu, S. and Min, S.Y., 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research, 48(D1), pp.D517-D525.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D. and Finn, R.D., 2019. A new genomic blueprint of the human gut microbiota. Nature, 568(7753), pp.499-504.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.
Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. and Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nature medicine, 26(4), pp.450-452.

Beam, A., Clinger, E. and Hao, L., 2021. Effect of diet and dietary components on the composition of the gut microbiota. Nutrients, 13(8), p.2795.

Bedoya, K., Hoyos, O., Zurek, E., Cabarcas, F. and Alzate, J.F., 2020. Annual microbial community dynamics in a full-scale anaerobic sludge digester from a wastewater treatment plant in Colombia. *Science of The Total Environment*, *726*, p.138479.

Breitwieser, F.P., Lu, J. and Salzberg, S.L., 2019. A review of methods and databases for metagenomic classification and assembly. Briefings in bioinformatics, 20(4), pp.1125-1136.

Bibby, K. and Peccia, J., 2013. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. Environmental science & technology, 47(4), pp.1945-1951.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A. and Tringe, S.G., 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology*, *35*(8), pp.725-731.

Buchfink, B., Xie, C. and Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nature methods, 12(1), pp.59-60.

Bull MJ, Plummer NT. 2014. "Part 1: The human gut microbiome in health and disease." (“RACGP – The gut microbiome – Australian Family Physician”) Integrative Medicine: A Clinician's Journal 13: 17-22.

Bushmanova, E., Antipov, D., Lapidus, A. and Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience, 8(9), p.giz100.

Bushnell, B., BBDuk: Adapter. Quality Trimming and Filtering. httpssourceforgenetprojectsbbmap.

Bushnell, B., 2014. BBMap: a fast, accurate, splice-aware aligner (No. LBNL-7065E). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581.

Cani, P.D., 2018. Human gut microbiome: hopes, threats and promises. Gut, 67(9), pp.1716-1725.

Chaumeil, P.A., Mussig, A.J., Hugenholtz, P. and Parks, D.H., 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.

Chen, K. and Pachter, L., 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS computational biology*, *1*(2), p.e24.

Chen, Y., Chen, L., Deng, Q., Zhang, G., Wu, K., Ni, L., Yang, Y., Liu, B., Wang, W., Wei, C. and Yang, J., 2020. The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients. Journal of medical virology.

Clemente, J.C., Ursell, L.K., Parfrey, L.W. and Knight, R., 2012. The impact of the gut microbiota on human health: an integrative view. Cell, 148(6), pp.1258-1270.

Crits-Christoph, A., Kantor, R.S., Olm, M.R., Whitney, O.N., Al-Shayeb, B., Lou, Y.C., Flamholz, A., Kennedy, L.C., Greenwald, H., Hinkle, A. and Hetzel, J., 2021. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *Mbio*, *12*(1).

Coordinators, N.R., 2018. Database Resources of the National Center for Biotechnology Information.

Coman, V. and Vodnar, D.C., 2020. Gut microbiota and old age: Modulating factors and interventions for healthy longevity. Experimental gerontology, 141, p.111095.

Cydzik-Kwiatkowska, A. and Zielińska, M., 2016. Bacterial communities in full-scale wastewater treatment systems. World Journal of Microbiology and Biotechnology, 32, pp.1-8.

Daughton, C., 2020. The international imperative to rapidly and inexpensively monitor community-wide Covid-19 infection status and trends. The Science of the Total Environment, 726, p.138149.

Devi, S., 2020. COVID-19 resurgence in Iran. Lancet (London, England), 395(10241), p.1896.

Donaldson, G.P., Lee, S.M. and Mazmanian, S.K., 2016. Gut biogeography of the bacterial microbiota. Nature Reviews Microbiology, 14(1), pp.20-32.

Dueholm, M.K.D., Nierychlo, M., Andersen, K.S., Rudkjøbing, V., Knutsson, S., Albertsen, M. and Nielsen, P.H., 2022. MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. Nature communications, 13(1), p.1908.

Edgar, R.C.; Flyvbjerg, H. Error Filtering, Pair Assembly and Error Correction for Next-Generation Sequencing Reads. *Bioinformatics* **2015**, *31*, 3476-3482.

Erickson, T.B., Endo, N., Duvallet, C., Ghaeli, N., Hess, K., Alm, E.J., Matus, M. and Chai, P.R., 2021. “Waste not, want not” – leveraging sewer systems and wastewater-based epidemiology for drug use trends and pharmaceutical monitoring. Journal of Medical Toxicology, 17(4), pp.397-410.

Escobar-Zepeda, A., Vera-Ponce de León, A. and Sanchez-Flores, A., 2015. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in genetics, 6, p.348.

Eurosurveillance editorial team, 2020. Rapid risk assessment from ECDC: Resurgence of reported cases of COVID-19 in the EU/EEA, the UK and EU candidate and potential candidate countries. Eurosurveillance, 25(26), p.2007021.

Feldgarden, M., Brover, V., Haft, D.H., Prasad, A.B., Slotta, D.J., Tolstoy, I., Tyson, G.H., Zhao, S., Hsu, C.H., McDermott, P.F. and Tadesse, D.A., 2019. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. Antimicrobial agents and chemotherapy, 63(11), pp.e00483-19.

Force, T.T. and Atlantic, R.S.C., Wastewater-Based Epidemiology for SARS-CoV-2.

Fresia, P., Antelo, V., Salazar, C., Giménez, M., D'Alessandro, B., Afshinnekoo, E., Mason, C., Gonnet, G.H. and Iraola, G., 2019. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. Microbiome, 7(1), pp.1-9.

Gao, J., O'Brien, J., Lai, F.Y., van Nuijs, A.L., He, J., Mueller, J.F., Xu, J. and Thai, P.K., 2015. Could wastewater analysis be a useful tool for China? – A review. Journal of Environmental Sciences, 27, pp.70-79.

Giandhari, J., Pillay, S., Wilkinson, E., Tegally, H., Sinayskiy, I., Schuld, M., Lourenço, J., Chimukangara, B., Lessells, R., Moosa, Y. and Gazy, I., 2021. Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. International Journal of Infectious Diseases, 103, pp.234-241.

Ghosh, S. and Pramanik, S., 2021. Structural diversity, functional aspects and future therapeutic applications of human gut microbiome. Archives of microbiology, 203(9), pp.5281-5308.

Grubaugh, N.D., Gangavarapu, K., Quick, J., Matteson, N.L., De Jesus, J.G., Main, B.J., Tan, A.L., Paul, L.M., Brackney, D.E., Grewal, S. and Gurfield, N., 2019. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome biology, 20(1), pp.1-19.

Gulino, K., Rahman, J., Badri, M., Morton, J., Bonneau, R. and Ghedin, E., 2020. Initial Mapping of the New York City Wastewater Virome. Msystems, 5(3), pp.e00876-19.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), pp.1072-1075.

Hart, O.E. and Halden, R.U., 2020. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. Science of The Total Environment, p.138875.

Heijnen, L. and Medema, G., 2011. Surveillance of influenza A and the pandemic influenza A (H1N1) 2009 in sewage and surface water in the Netherlands. Journal of water and health, 9(3), pp.434-442.

Holshue, M.L., DeBolt, C., Lindquist, S., Lofy, K.H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A. and Diaz, G., 2020. First case of 2019 novel coronavirus in the United States. New England Journal of Medicine.

Huo, Y., Bai, Y. and Qu, J., 2017. Unravelling riverine microbial communities under wastewater treatment plant effluent discharge in large urban areas. Applied Microbiology and Biotechnology, 101, pp.6755-6764.
Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), p.119.

Ji, B., Wang, S., Guo, D. and Pang, H., 2020. Comparative and comprehensive analysis on bacterial communities of two full-scale wastewater treatment plants by second and third-generation sequencing. Bioresource Technology Reports, 11, p.100450.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. and Pesseat, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), pp.1236-1240.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H. and Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, p.e7359.

Klement, R.J. and Pazienza, V., 2019. Impact of different types of diet on gut microbiota profiles and cancer prevention and treatment. Medicina, 55(4), p.84.

Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glöckner, F.O. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. **2013**, *41*, e1. https://doi.org/10.1093/nar/gks808.

Kho, Z.Y. and Lal, S.K., 2018. The human gut microbiome – a potential controller of wellness and disease. Frontiers in microbiology, p.1835.

Lagier, J.C., Million, M., Hugon, P., Armougom, F. and Raoult, D., 2012. Human gut microbiota: repertoire and variations. *Frontiers in cellular and infection microbiology*, *2*, p.136.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *bioinformatics*, *25*(14), pp.1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), pp.2078-2079.

Li, Y., Deng, X., Hu, F., Wang, J., Liu, Y., Huang, H., Ma, J., Zhang, J., Zhang, F. and Zhang, C., 2018. Metagenomic analysis identified co-infection with human rhinovirus C and bocavirus 1 in an adult suffering from severe pneumonia. The Journal of infection, 76(3), p.311.

Liang, G. and Bushman, F.D., 2021. The human virome: assembly, composition and host interactions. Nature Reviews Microbiology, 19(8), pp.514-527.

Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biol*. **2014**, *15*, 1-21.

Mackuľak, T., Gál, M., Špalková, V., Fehér, M., Briestenská, K., Mikušová, M., Tomčíková, K., Tamáš, M. and Butor Škulcová, A., 2021. Wastewater-based epidemiology as an early warning system for the spreading of SARS-CoV-2 and its mutations in the population. International Journal of Environmental Research and Public Health, 18(11), p.5629.

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., Kryukov, K., Fukuda, A., Morimoto, Y., Naito, Y. and Okada, H., 2021. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. BMC microbiology, 21, pp.1-13.

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. and Brouwer, A., 2020. Presence of SARS-Coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands. *Environmental Science & Technology Letters*, *7*(7), pp.511-516.

Meleshko, D., Hajirasouliha, I. and Korobeynikov, A., 2021. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. *bioRxiv*, pp.2020-07.

Menzel, P., Ng, K.L. and Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications, 7(1), pp.1-9.

Metwally, A.A.; Yang, J.; Ascoli, C.; Dai, Y.; Finn, P.W.; Perkins, D.L. MetaLonDA: A flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome* **2018**, *6*, 1-12.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A. and Wilkening, J., 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics, 9(1), pp.1-8.

McMurdie, P.J.; Holmes, S. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **2013**, *8*, e61217. https://doi.org/10.1371/journal.pone.0061217.

Moschen, A.R., Wieser, V. and Tilg, H., 2012. Dietary factors: major regulators of the gut's microbiota. Gut and liver, 6(4), p.411.

Newton, R.J., McLellan, S.L., Dila, D.K., Vineis, J.H., Morrison, H.G., Eren, A.M. and Sogin, M.L., 2015. Sewage reflects the microbiomes of human populations. MBio, 6(2), pp.e02574-14.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), pp.268-274.

Nieuwenhuijse, D.F., Oude Munnink, B.B., Phan, M.V., Munk, P., Venkatakrishnan, S., Aarestrup, F.M., Cotten, M. and Koopmans, M.P., 2020. Setting a baseline for global urban virome surveillance in sewage. Scientific reports, 10(1), pp.1-13.

Nurk, S., Meleshko, D., Korobeynikov, A. and Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome research, 27(5), pp.824-834.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. and Astashyn, A., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), pp.D733-D745.

Oliphant, K. and Allen-Vercoe, E., 2019. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. Microbiome, 7(1), pp.1-15.

Oluwagbemigun, K., Schnermann, M.E., Schmid, M., Cryan, J.F. and Nöthlings, U., 2022. A prospective investigation into the association between the gut microbiome composition and cognitive performance among healthy young adults. Gut Pathogens, 14(1), p.15.

Orive, G., Lertxundi, U. and Barcelo, D., 2020. Early SARS-CoV-2 outbreak detection by sewage-based epidemiology. Science of The Total Environment, p.139298.

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B. and Yeats, C., 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, *7*(2), p.veab064.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. and Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, *25*(7), pp.1043-1055.

Patil, I., 2021. Visualizations with statistical details: The'ggstatsplot'approach. Journal of Open Source Software, 6(61).

Peddu, V., Shean, R.C., Xie, H., Shrestha, L., Perchetti, G.A., Minot, S.S., Roychoudhury, P., Huang, M.L., Nalla, A., Reddy, S.B. and Phung, Q., 2020. Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. Clinical Chemistry.

Picó, Y. and Barceló, D., 2021. Mass spectrometry in wastewater-based epidemiology for the determination of small and large molecules as biomarkers of exposure: toward a global view of environment and human health under the COVID-19 outbreak. ACS omega, 6(46), pp.30865-30872.

Price, M.N., Dehal, P.S. and Arkin, A.P., 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), p.e9490.

Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res*. **2012**, *41(D1)*, D590-D596.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L. and Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology, 5(11), pp.1403-1407.

Ranjan, R., Rani, A., Metwally, A., McGee, H.S. and Perkins, D.L., 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochemical and biophysical research communications, 469(4), pp.967-977.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation forStatistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G.A.D., Gasbarrini, A. and Mele, M.C., 2019. What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. Microorganisms, 7(1), p.14.

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.

Sakkas, H., Bozidis, P., Touzios, C., Kolios, D., Athanasiou, G., Athanasopoulou, E., Gerou, I. and Gartzonika, C., 2020. Nutritional status and the influence of the vegan diet on the gut microbiota and human health. Medicina, 56(2), p.88.

Sayers, E.W., Beck, J., Bolton, E.E., Bourexis, D., Brister, J.R., Canese, K., Comeau, D.C., Funk, K., Kim, S., Klimke, W. and Marchler-Bauer, A., 2021. Database resources of the national center for biotechnology information. Nucleic acids research, 49(D1), p.D10.

Shahi, S.K., Zarei, K., Guseva, N.V. and Mangalam, A.K., 2019. Microbiota analysis using two-step PCR and next-generation 16S rRNA gene sequencing. JoVE (Journal of Visualized Experiments), (152), p.e59980.

Shu, Y. and McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, *22*(13), p.30494.

Silva-Bedoya, L.M., Sánchez-Pinzón, M.S., Cadavid-Restrepo, G.E. and Moreno-Herrera, C.X., 2016. Bacterial community analysis of an industrial wastewater treatment plant in Colombia with screening for lipid-degrading microorganisms. Microbiological Research, 192, pp.313-325.

Tan, B., Ng, C.M., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.H. and Thompson, J.R., 2015. Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. Frontiers in microbiology, 6, p.1027.

Team, R.C., 2019. R: A language and environment for statistical computing. Vienna, Austria: Foundation for Statistical Computing.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N. and Mlisana, K., 2020. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*.
Tegally, H., Wilkinson, E., Lessells, R.J., Giandhari, J., Pillay, S., Msomi, N., Mlisana, K., Bhiman, J.N., von Gottberg, A., Walaza, S. and Fonseca, V., 2021. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nature medicine*, pp.1-7.

Turakhia, Y., Thornlow, B., Hinrichs, A.S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D. and Corbett-Detig, R., 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, *53*(6), pp.809-816.

Wang, S., Song, F., Gu, H., Shu, Z., Wei, X., Zhang, K., Zhou, Y., Jiang, L., Wang, Z., Li, J. and Luo, H., 2022. Assess the diversity of gut microbiota among healthy adults for forensic application. Microbial Cell Factories, 21(1), p.46.

Wang, X.W., Li, J.S., Guo, T.K., Zhen, B., Kong, Q.X., Yi, B., Li, Z., Song, N., Jin, M., Xiao, W.J. and Zhu, X.M., 2005. Concentration and detection of SARS coronavirus in sewage from Xiao Tang Shan Hospital and the 309th Hospital. Journal of virological methods, 128(1-2), pp.156-161.

WRC 2020. A compendium of emerging South African testing methodologies for detecting of SARS-CoV-2 RNA in wastewater surveillance. WRC Report No. SP 143/20. Water Research Commission. Pretoria.

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., Zhang, Q., Brown, M.R., Li, Z., Van Nostrand, J.D. and Ling, F., 2019. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nature microbiology, 4(7), pp.1183-1195.

Xiao, F., Sun, J., Xu, Y., Li, F., Huang, X., Li, H., Zhao, J., Huang, J. and Zhao, J., 2020. Infectious SARS-CoV-2 in feces of patient with severe COVID-19. Emerg Infect Dis, 26(8), pp.10-3201.

Yang, Y., Li, B., Zou, S., Fang, H.H. and Zhang, T., 2014. Fate of antibiotic resistance genes in sewage treatment plant revealed by metagenomic approach. Water research, 62, pp.97-106.
Yu, G., Smith, D.K., Zhu, H., Guan, Y. and Lam, T.T.Y., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8(1), pp.28-36.

Zhang, M., Liu, Y.S., Zhao, J.L., Liu, W.R., He, L.Y., Zhang, J.N., Chen, J., He, L.K., Zhang, Q.Q. and Ying, G.G., 2018. Occurrence, fate and mass loadings of antibiotics in two swine wastewater treatment systems. Science of the Total Environment, 639, pp.1421-1431.

Zhang, M., Wang, S., Ji, B. and Liu, Y., 2019. Towards mainstream deammonification of municipal wastewater: Partial nitrification-anammox versus partial denitrification-anammox. *Science of the Total Environment*, *692*, pp.393-401.

Zhang, Z.B., Li, P., Liu, H.J., Zhong, J.Y., Zheng, Y., Li, K.B., Qin, P.Z., Zeng, Q., Li, J.H., Li, L.Z. and Cao, L., 2020. Genomic surveillance of a resurgence of COVID-19 in Guangzhou, China.

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
Agricultural Research Council

PierneefR@arc.agric.za

Date 10 March 2021
Quotation No. **Rian Pierneef_10 March 2021**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Sample prep	20	R 3 240,34	R 64 806,84
Sequencing (GB)	82	R 961,35	R 78 831,04
			R 143 637,88
		15% VAT	R -
		TOTAL	R 143 637,88

An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until the full amount has been transferred to our account
2. If quoted "per gigabase" for sequencing, a 20% - up or down deviation, will constitute the
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange
4. Quote will only be valid for 30 days from the date of the quotation
4. **ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide* . If samples do not comply to these requirements, ALL EXTRA EXPENSES will be for your expense at a**
5. **No data/information** will be released until we've been fully reimbursed, **including all extra expenses**

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 28/09/2021
Quotation No. **Rian Pierneef_28/09/2021_24Covid**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Amplicon based whole genome sequening of SARS-CoV-2 from RNA samples	24	R 1 805,00	R 43 320,00
			R 43 320,00
		15% VAT	R -
		TOTAL	R 43 320,00

Dear valued client

Please note that the ARC-BTP HiSeq2500 has reached end-of-life (EOL). We are currently in the final stages of procuring a new high-throughput sequencing platform and in the interim the work that you have been quoted on will be done using an external service provider.

An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until a Purchase Order (PO) or Internal Approval form has been received
2. If quoted "per gigabase" for sequencing, a 20% - up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will
4. Quote will only be valid for 30 days from the date of the quotation
5. ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply
6. No data/information will be released until we've been fully reimbursed, including all extra expenses

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 28/09/2021
Quotation No. **Rian Pierneef_28/09/2021_24shotgun**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Sample prep	24	R 2 477,46	R 59 459,02
Sequencing (GB)	120	R 500,00	R 60 000,00
			R 119 459,02
		15% VAT	R -
		TOTAL	R 119 459,02

Dear valued client

Please note that the ARC-BTP HiSeq2500 has reached end-of-life (EOL). We are currently in the final stages of procuring a new high-throughput sequencing platform and in the interim the work that you have been quoted on will be done using an external service provider.

An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until a Purchase Order (PO) or Internal Approval form has been received
2. If quoted "per gigabase" for sequencing, a 20% - up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will
4. Quote will only be valid for 30 days from the date of the quotation
5. ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply
6. No data/information will be released until we've been fully reimbursed, including all extra expenses

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 18/11/2021
Quotation No. **Rian Pierneef_18/11/2021**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Amplicon based whole genome sequening of SARS-CoV-2 from RNA samples	48	R 1 805,00	R 86 640,00
			R 86 640,00
		15% VAT	R -
		TOTAL	R 86 640,00

Dear valued client

Please note that the ARC-BTP HiSeq2500 has reached end-of-life (EOL). We are currently in the final stages of procuring a new high-throughput sequencing platform and in the interim the work that you have been quoted on will be done using an external service provider.

An order number is required prior to the start of any work done by the ARC Sequencing facility.

This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.

Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until a Purchase Order (PO) or Internal Approval form has been received
2. If quoted "per gigabase" for sequencing, a 20% - up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will
4. Quote will only be valid for 30 days from the date of the quotation
5. ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply
6. No data/information will be released until we've been fully reimbursed, including all extra expenses

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 13/01/2022
Quotation No. **Rian Pierneef_13/01/2022_shotgun**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
DNA extraction	30	R 175,00	R 5 250,00
Sample prep	30	R 396,88	R 11 906,49
Sequencing (GB)	150	R 500,00	R 75 000,00
			R 92 156,49
		15% VAT	R
		TOTAL	R 92 156,49

Dear valued client
Please note that the ARC-BTP HiSeq2500 has reached end-of-life (EOL). We are currently in the final stages of procuring a new high-throughput sequencing platform and in the interim the work that you have been quoted on will be done using an external service provider.
An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until a Purchase Order (PO) or Internal Approval form has been received
2. If quoted "per gigabase" for sequencing, a 20% up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will
4. Quote will only be valid for 30 days from the date of the quotation
5. ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply
6. No data/information will be released until we've been fully reimbursed, including all extra expenses

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 13/01/2022
Quotation No. **Rian Pierneef_13/01/2022**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Amplicon based whole genome sequening of SARS-CoV-2 from RNA samples	30	R 805,00	R 24 150,00
			R 24 150,00
		15% VAT	R
		TOTAL	R 24 150,00

Dear valued client

Please note that the ARC-BTP HiSeq2500 has reached end-of-life (EOL). We are currently in the final stages of procuring a new high-throughput sequencing platform and in the interim the work that you have been quoted on will be done using an external service provider.

An order number is required prior to the start of any work done by the ARC Sequencing facility.

This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.

Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until a Purchase Order (PO) or Internal Approval form has been received
2. If quoted "per gigabase" for sequencing, a 20% up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will
4. Quote will only be valid for 30 days from the date of the quotation
5. ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply
6. No data/information will be released until we've been fully reimbursed, including all extra expenses

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

2.2.1

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 03/03/2022
Quotation No. **Rian Pierneef_03/03/2022**
(P11000078)

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
DNA extraction	75	R 175,00	R 13 125,00
Sample prep	75	R 1 240,63	R 93 047,48
Sequencing (GB)	375	R 500,00	R 187 500,00
			R 293 672,48
		15% VAT	R -
		TOTAL	R 293 672,48

Dear valued client

Please note that the ARC-BTP HiSeq2500 has reached end-of-life (EOL). We are currently in the final stages of procuring a new high-throughput sequencing platform and in the interim the work that you have been quoted on will be done using an external service provider.

An order number is required prior to the start of any work done by the ARC Sequencing facility.

This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.

Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until a Purchase Order (PO) or Internal Approval form has been received
2. If quoted "per gigabase" for sequencing, a 20% - up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will
4. Quote will only be valid for 30 days from the date of the quotation
5. ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply
6. No data/information will be released until we've been fully reimbursed, including all extra expenses

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 04/07/2022
Quotation No. **Rian Pierneef_04/07/2022**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Sample prep	72	R 425,09	R 30 606,54
Sequencing (GB)	12	R 3 367,34	R 40 408,10
			R 71 014,64
		15% VAT	R
		TOTAL	R 71 014,64

An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until the full amount has been transferred to our account
2. If quoted "per gigabase" for sequencing, a 20% up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will only be valid for our current kits in SA
4. Quote will only be valid for 30 days from the date of the quotation
4. **ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply to these requirements, ALL EXTRA EXPENSES will be for your expense at a minimum cost of R1000 per sample.**
5. **No data/information** will be released until we've been fully reimbursed, **including all extra expenses**

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
ARC-BTP
Onderstepoort
PierneefR@arc.agric.za

Date 04/07/2022
Quotation No. **Rian Pierneef_04/07/2022**

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Sequencing (GB)	216	R 500,00	R 108 000,00
			R 108 000,00
		15% VAT	R
		TOTAL	R 108 000,00

An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

1. No sequencing preparation will commence until the full amount has been transferred to our account
2. If quoted "per gigabase" for sequencing, a 20% up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will only be valid for our current kits in SA
4. Quote will only be valid for 30 days from the date of the quotation
4. **ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply to these requirements, ALL EXTRA EXPENSES will be for your expense at a minimum cost of R1000 per sample.**
5. **No data/information** will be released until we've been fully reimbursed, **including all extra expenses**

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

Supplementary: Sequencing Quotation

AGRICULTURAL RESEARCH COUNCIL (ARC)
BIOTECHNOLOGY PLATFORM
Private Bag X05, Onderstepoort 0110, South Africa
Tel: (012) 529 9121 (Int: +27 12) E-Mail: BTP-Core@arc.agric.za
Web site: www.arc.agric.za

Quotation

Genomics Core Facility

Dr
Rian Pierneef
pierneefr@arc.agric.za

Date 26/09/2022
Quotation No. Rian_Pierneef_26_09_2022

Dear Dr Rian Pierneef

Please find attached the quote for the sequencing of your sample(s). Please note that our financial department requires payment (or proof thereof) before the results may be released.

Description	Qty	Unit Price	Total (excl VAT)
Sample prep	10	R 1,216.83	R 12,168.30
Sequencing (GB)	100	R 234.83	R 23,482.80
			R 35,651.10
		15% VAT	R -
		TOTAL	R 35,651.10

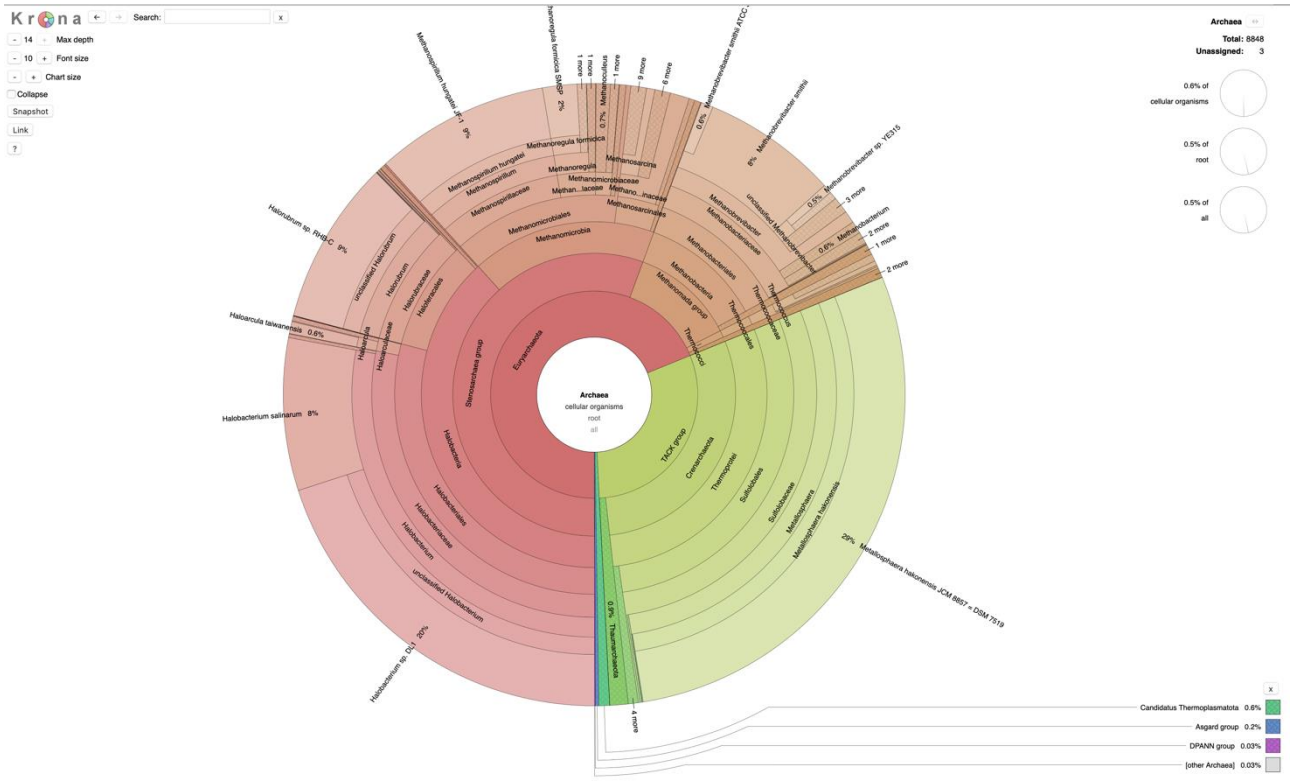
An order number is required prior to the start of any work done by the ARC Sequencing facility.
This quotation represents an estimate of the cost for the requested sequencing work only. This does not constitute a contract and does not imply any warranties.
Please see attached conditions for payment and sample requirements.

Conditions:

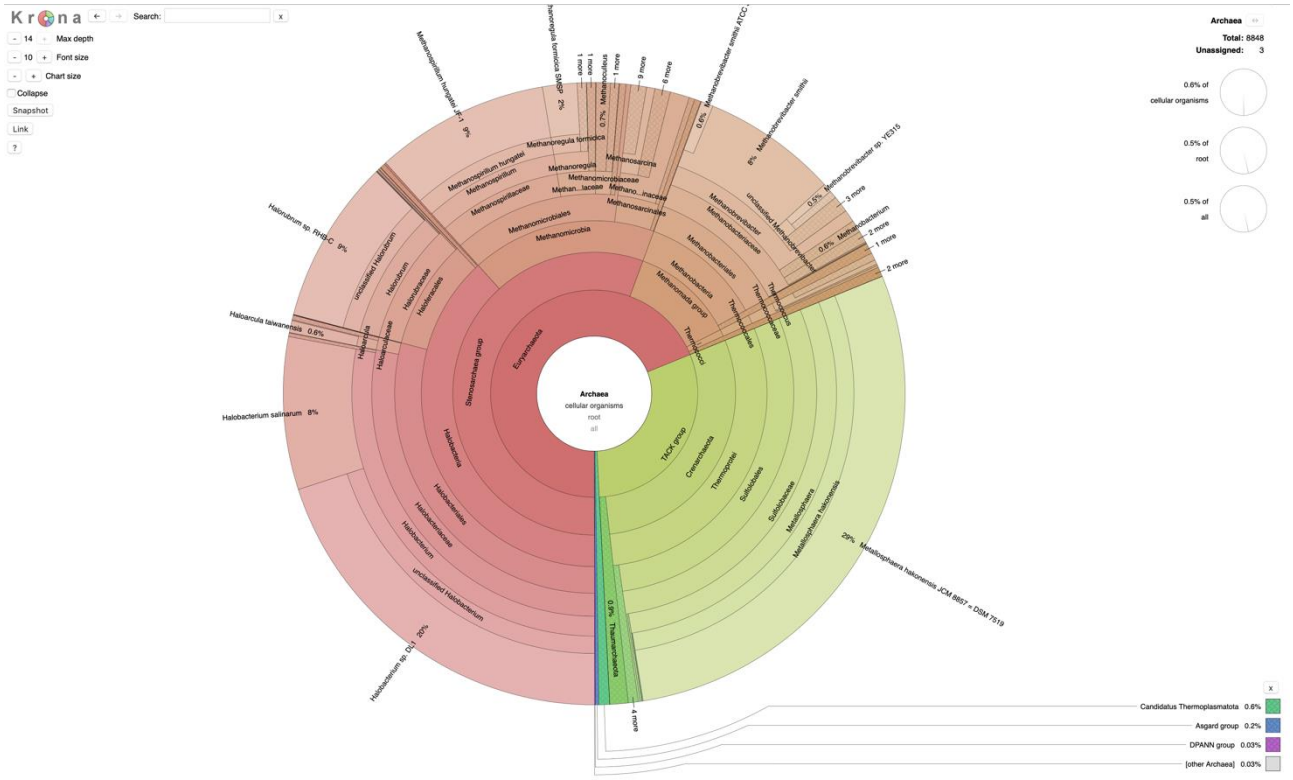
1. No sequencing preparation will commence until the full amount has been transferred to our account
2. If quoted "per gigabase" for sequencing, a 20% - up or down deviation, will constitute the completion of the order
3. Kits are imported on existing projects and prices are based on our current stock. Due to exchange rates, this quote will only be valid for our current kits in SA
4. Quote will only be valid for 30 days from the date of the quotation
4. **ALL samples must conform to the requirements as stipulated in the *Sample Preparation Guide*. If samples do not comply to these requirements, ALL EXTRA EXPENSES will be for your expense at a minimum cost of R1000 per sample.**
5. **No data/information** will be released until we've been fully reimbursed, **including all extra expenses**

Kindly email your order number to BTP-Core@arc.agric.za before work can commence

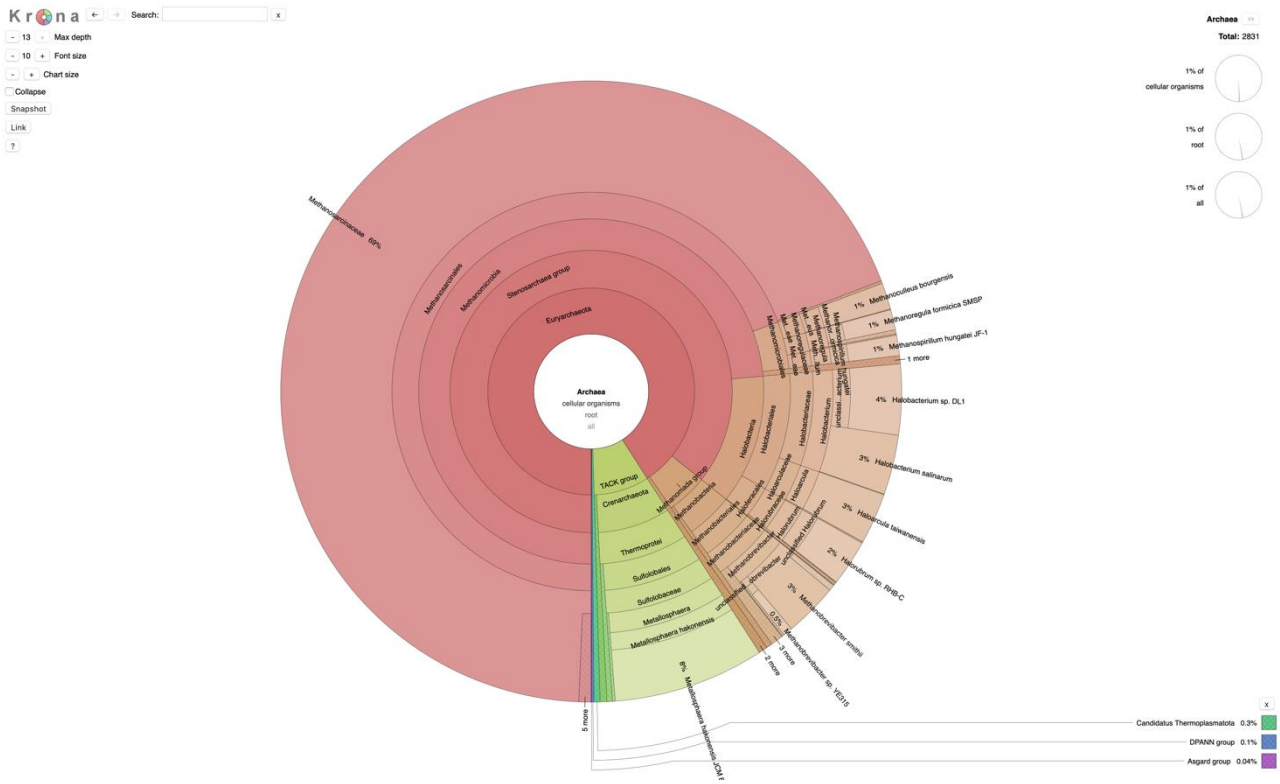
Supplementary: Per Sample Taxonomy Archaea BSW1_1A



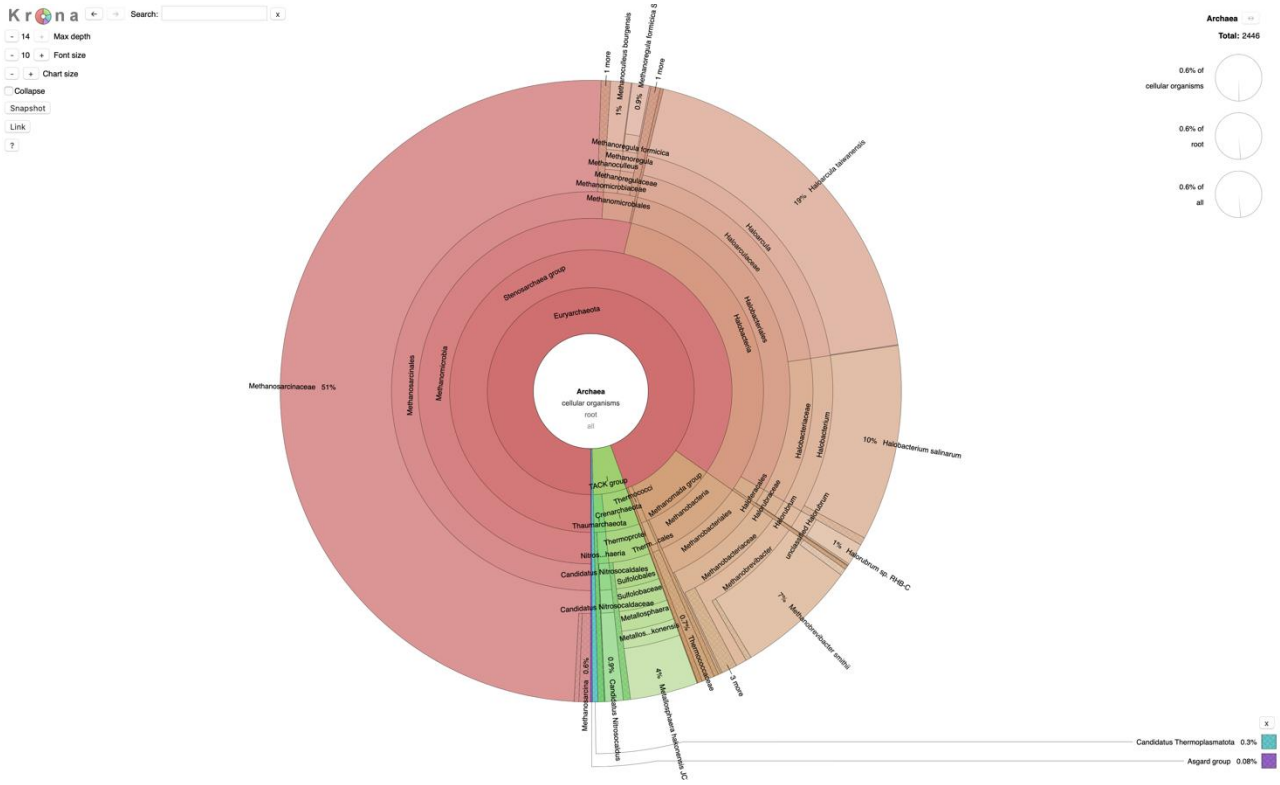
Supplementary: Per Sample Taxonomy Archaea BSW2_1A



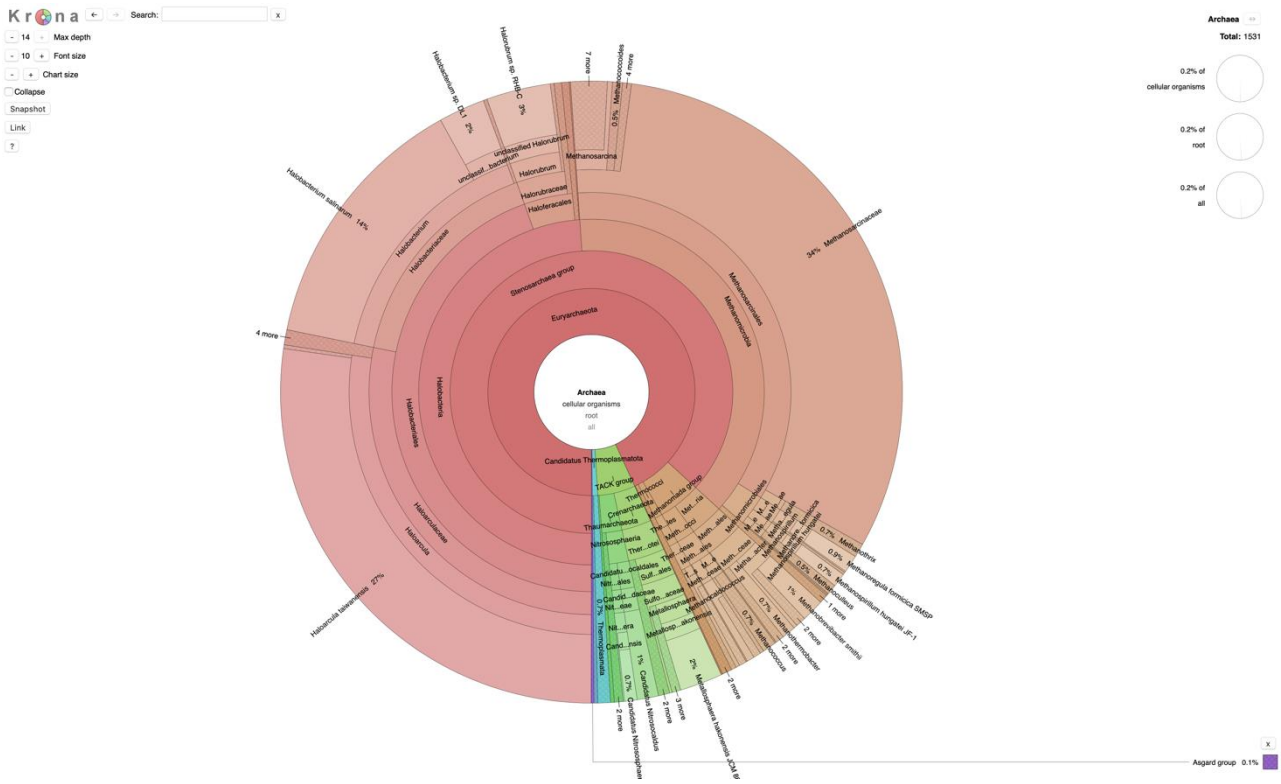
Supplementary: Per Sample Taxonomy Archaea BSW6_1A



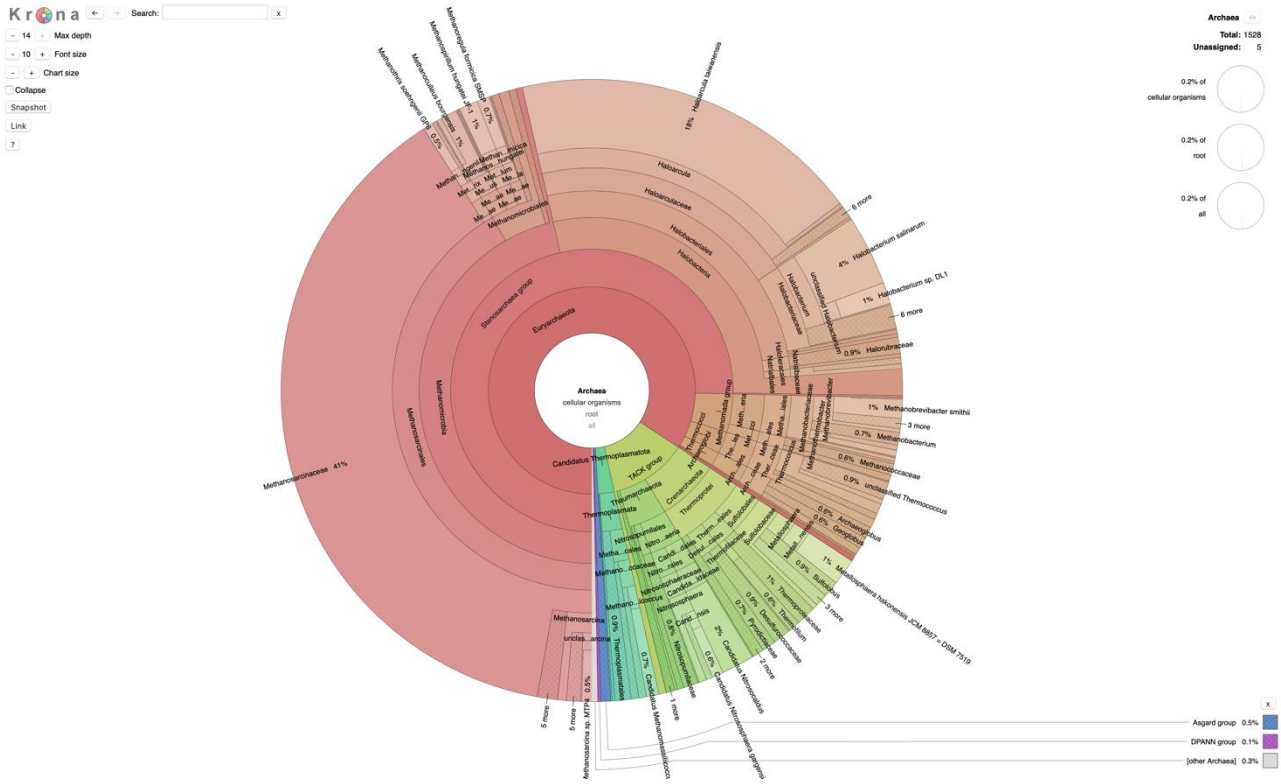
Supplementary: Per Sample Taxonomy Archaea DW6_1A



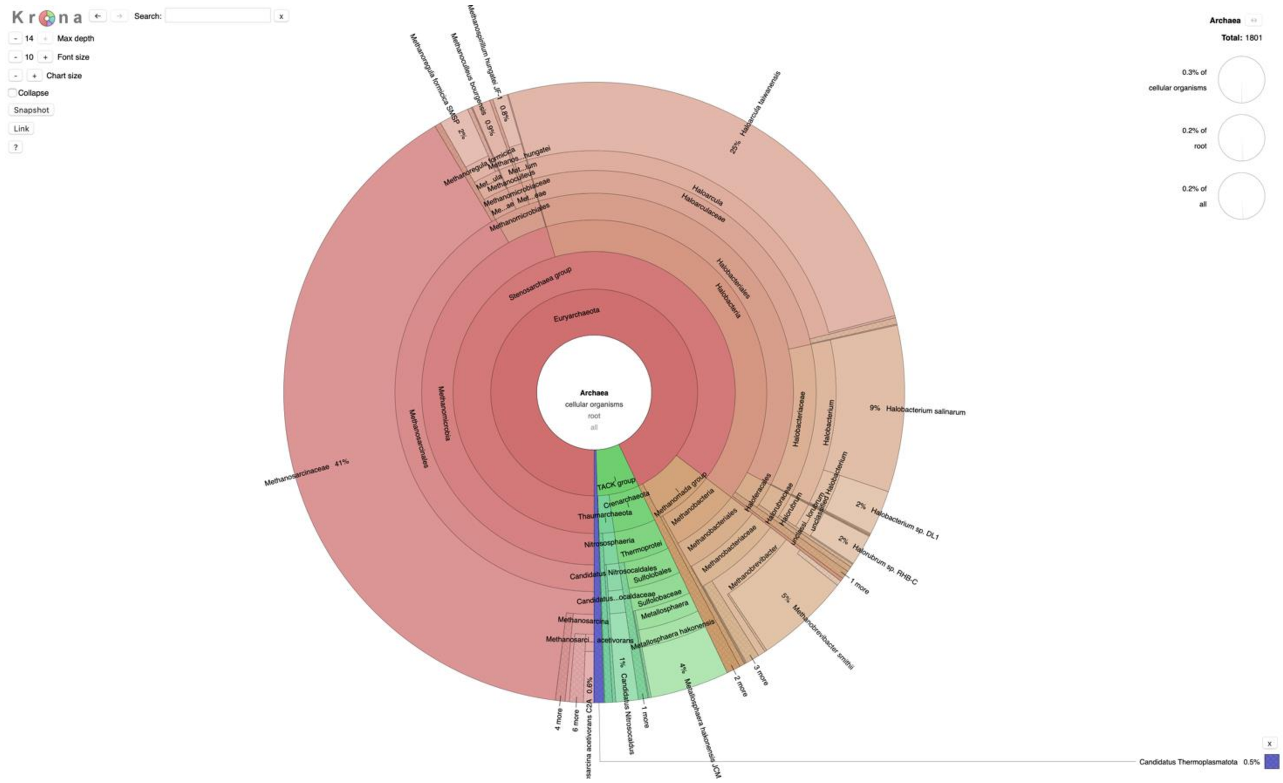
Supplementary: Per Sample Taxonomy Archaea DW7_1A



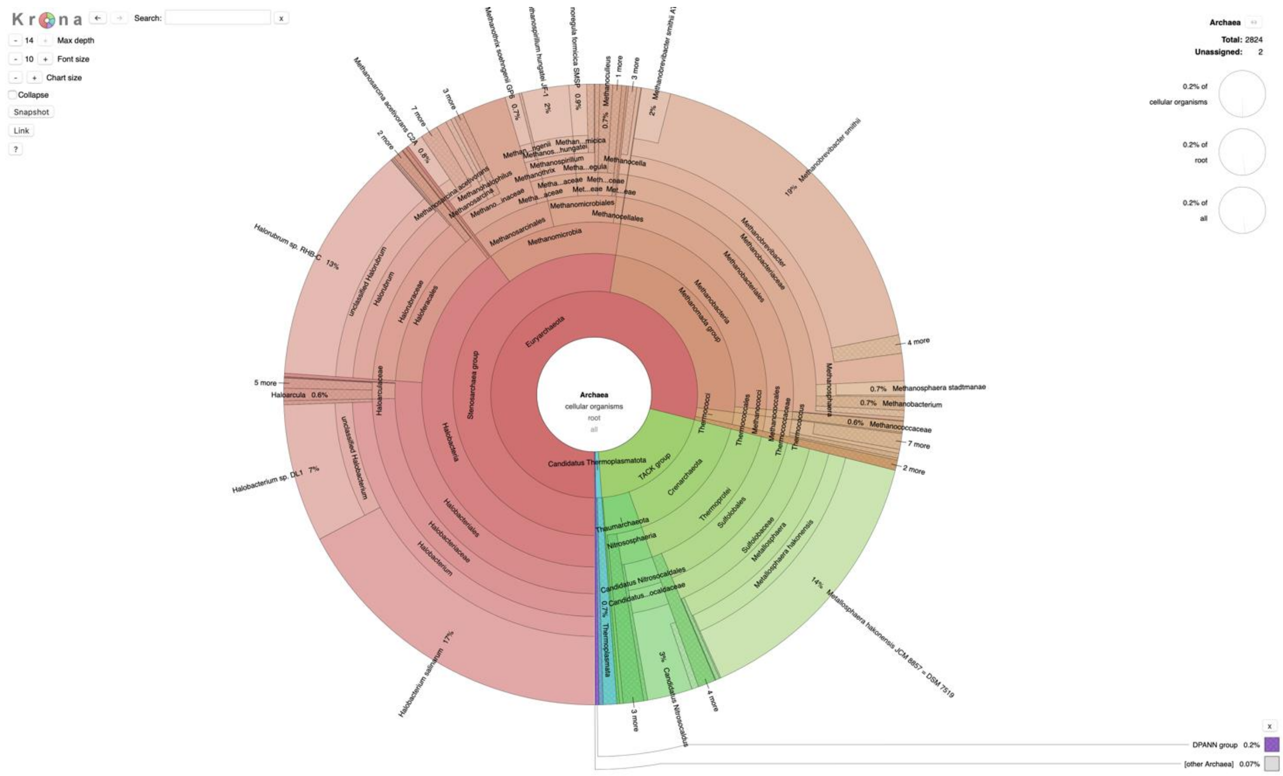
Supplementary: Per Sample Taxonomy Archaea DW8_1A



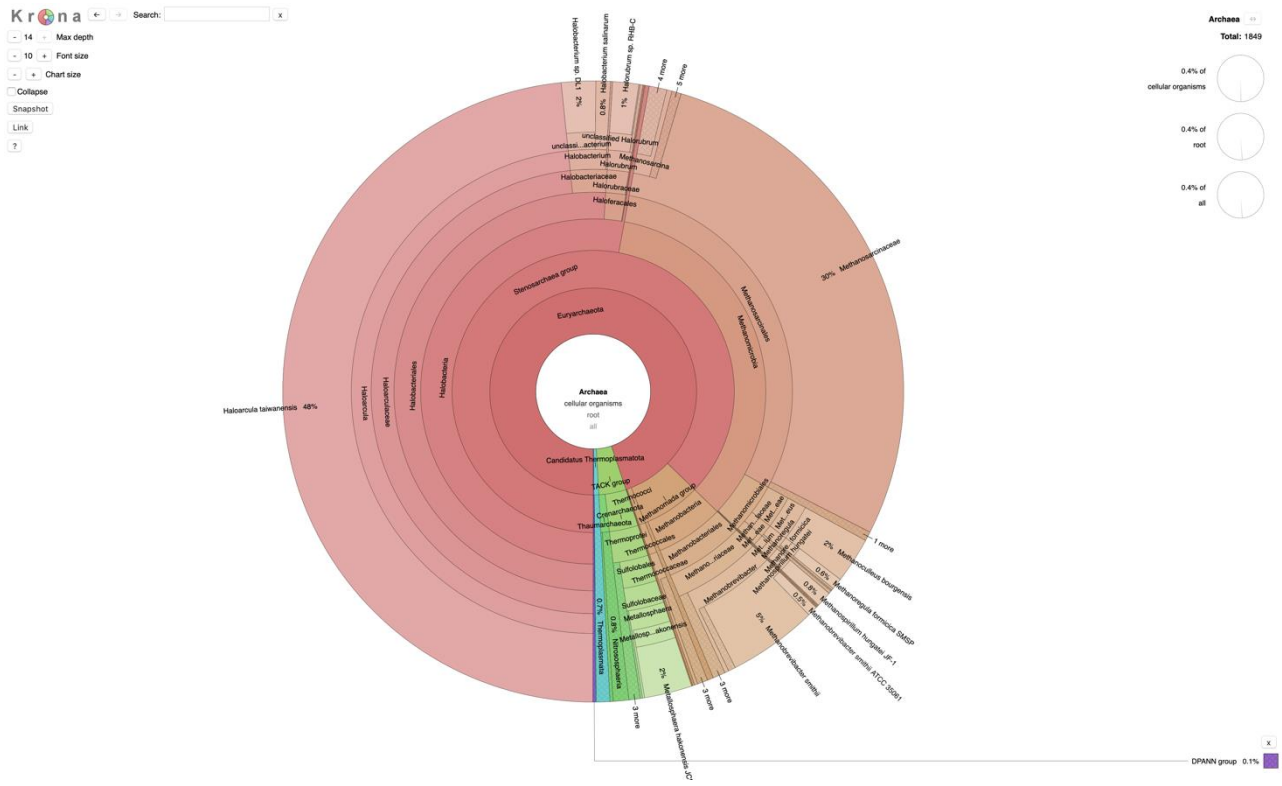
Supplementary: Per Sample Taxonomy Archaea DW10_1A



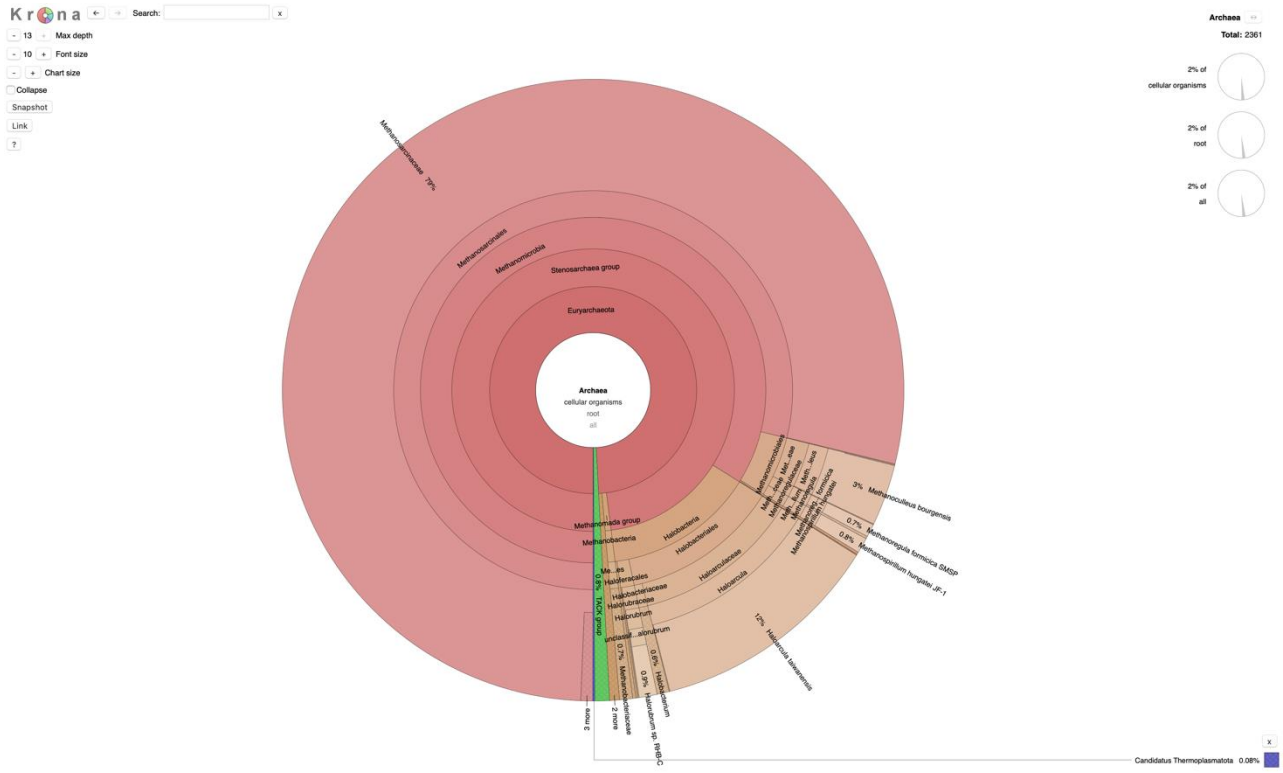
Supplementary: Per Sample Taxonomy Archaea DW12_1A



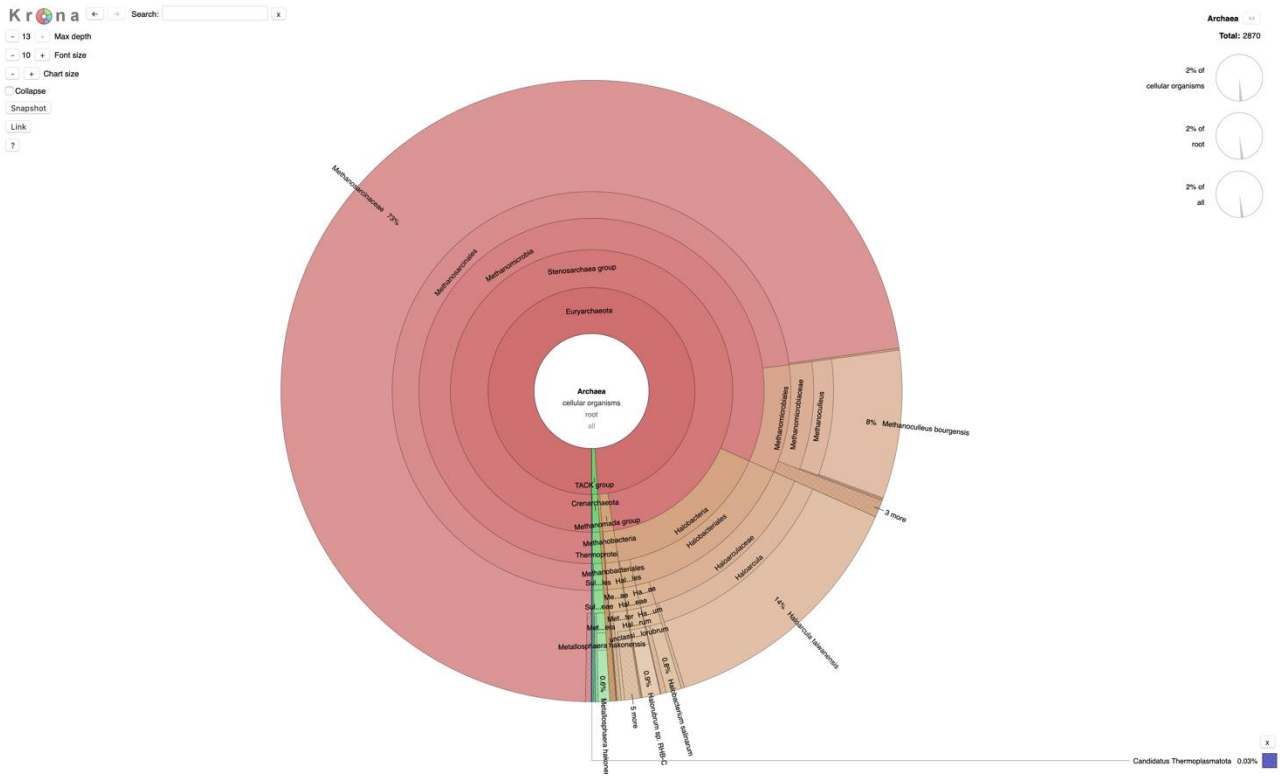
Supplementary: Per Sample Taxonomy Archaea DW15_1A



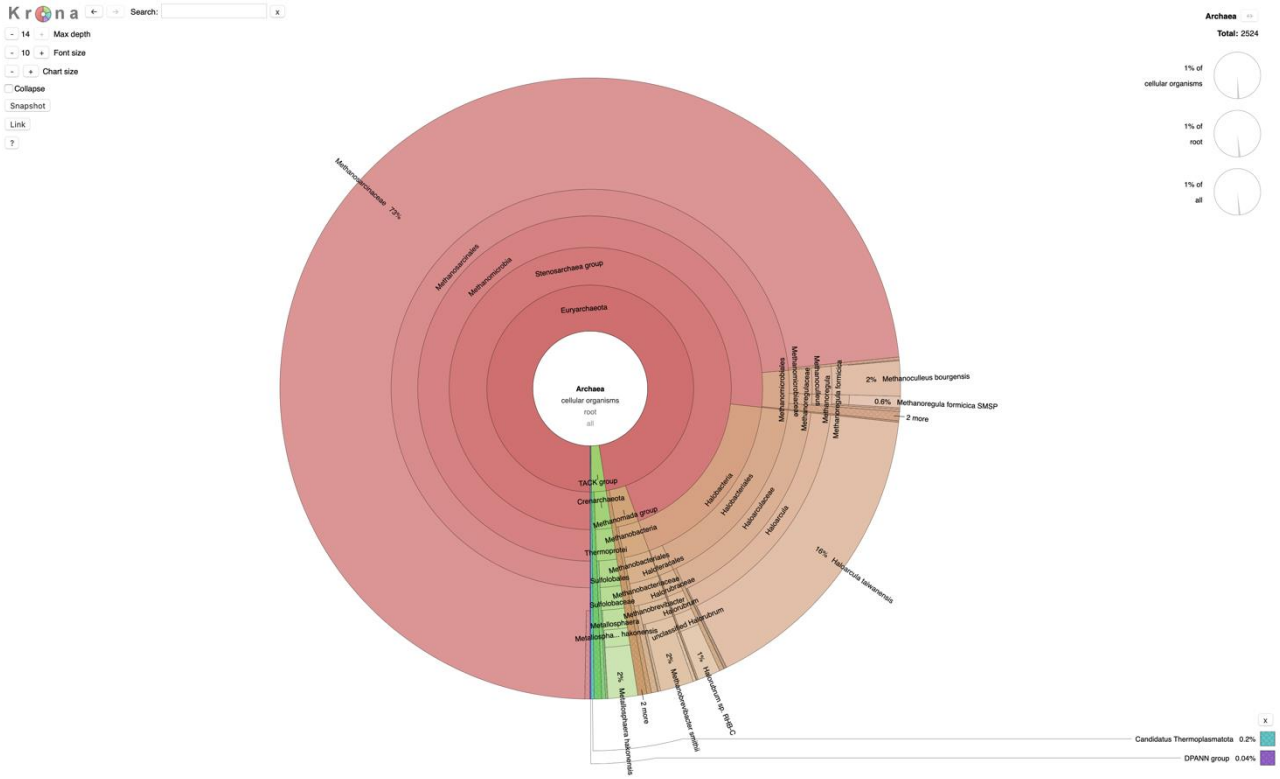
Supplementary: Per Sample Taxonomy Archaea RTW1_1A



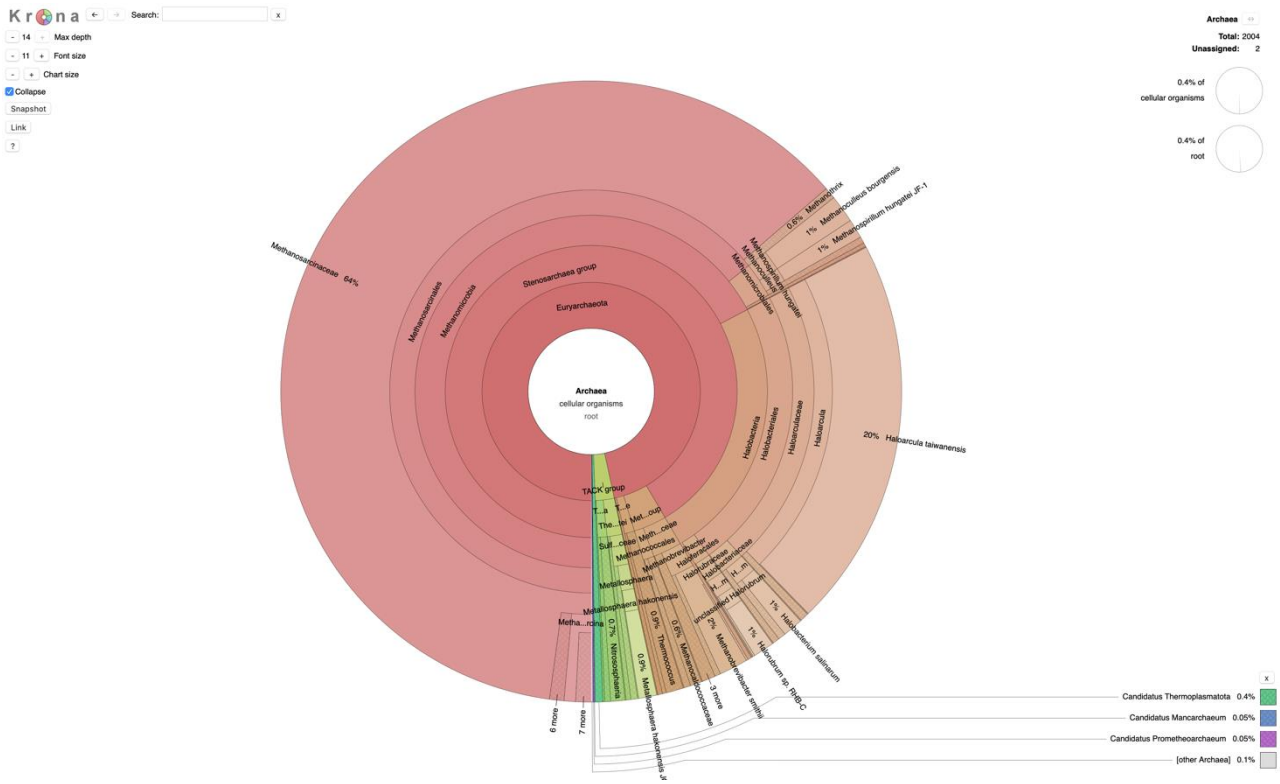
Supplementary: Per Sample Taxonomy Archaea RTW2_1A



Supplementary: Per Sample Taxonomy Archaea RTW6_1A

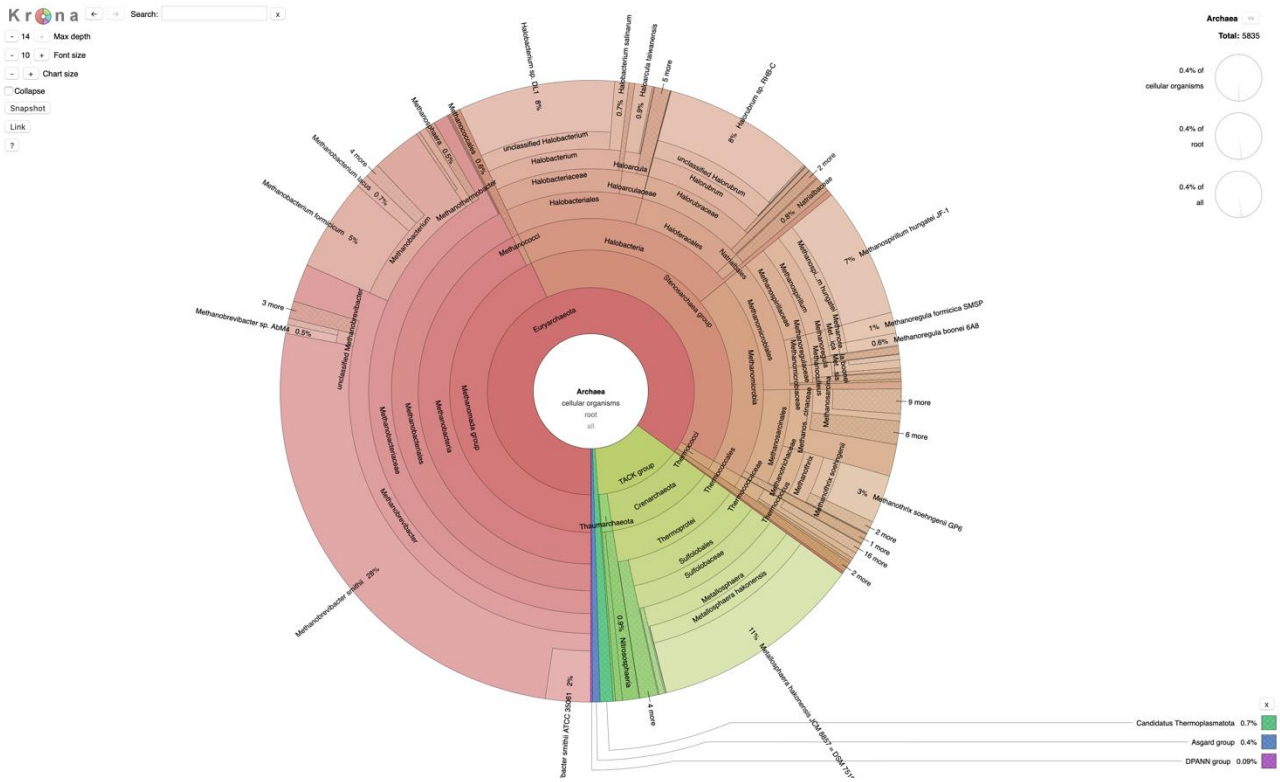


Supplementary: Per Sample Taxonomy Archaea RTW7_1A

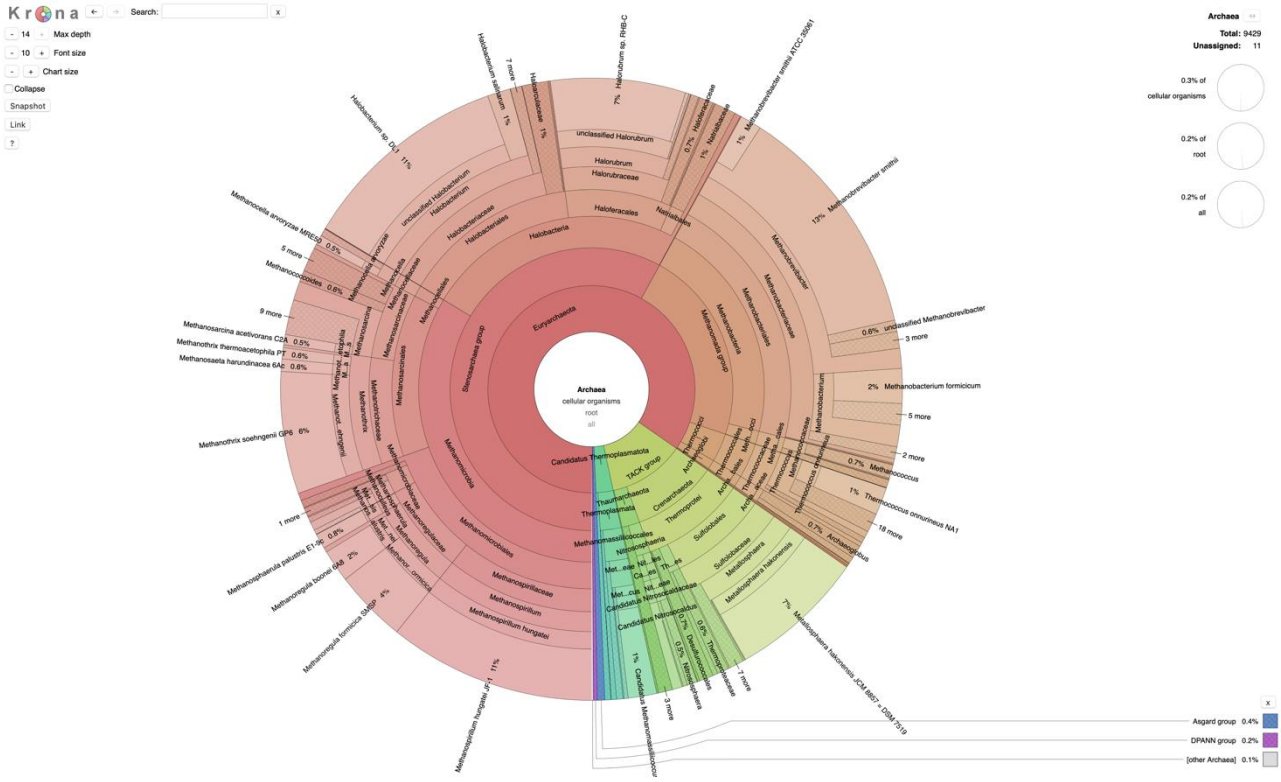


Supplementary: Per Sample Taxonomy Archaea RTW8_1A

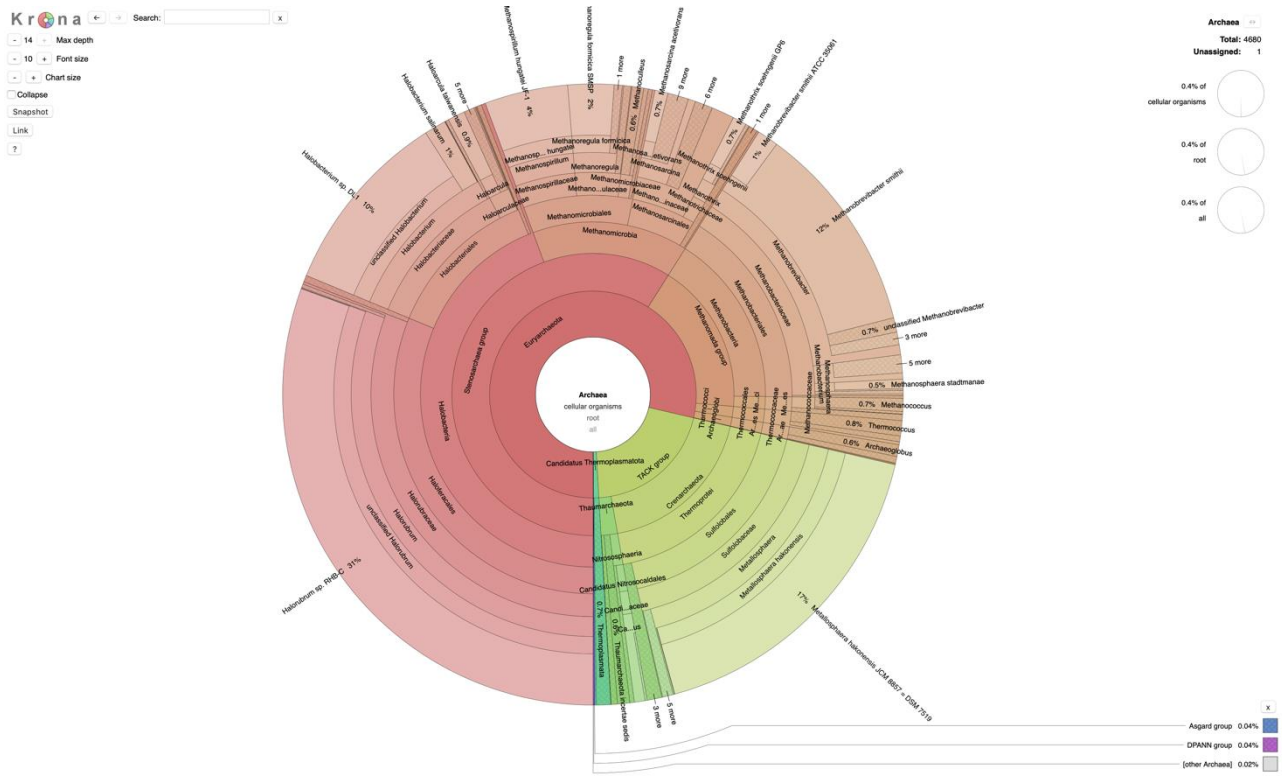
Supplementary: Per Sample Taxonomy Archaea RTW10_1A



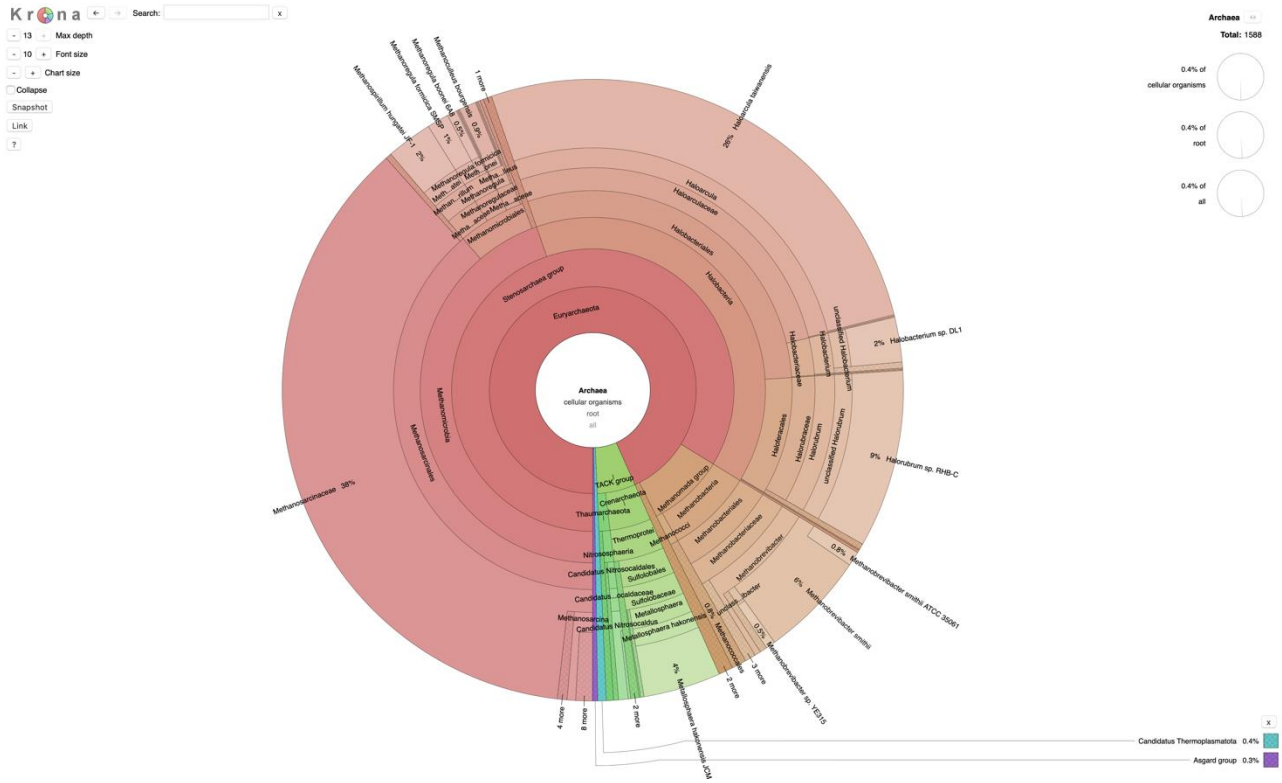
Supplementary: Per Sample Taxonomy Archaea RTW11_1A



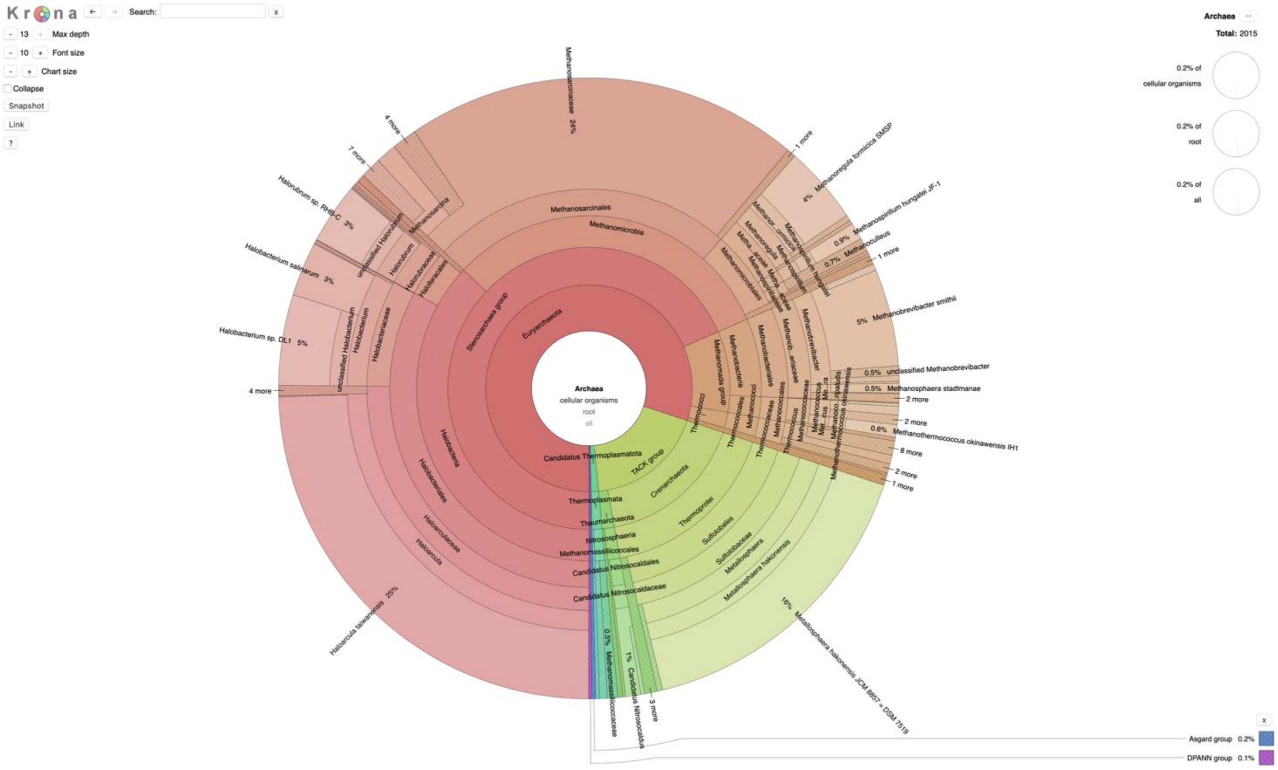
Supplementary: Per Sample Taxonomy Archaea RTW12_1A



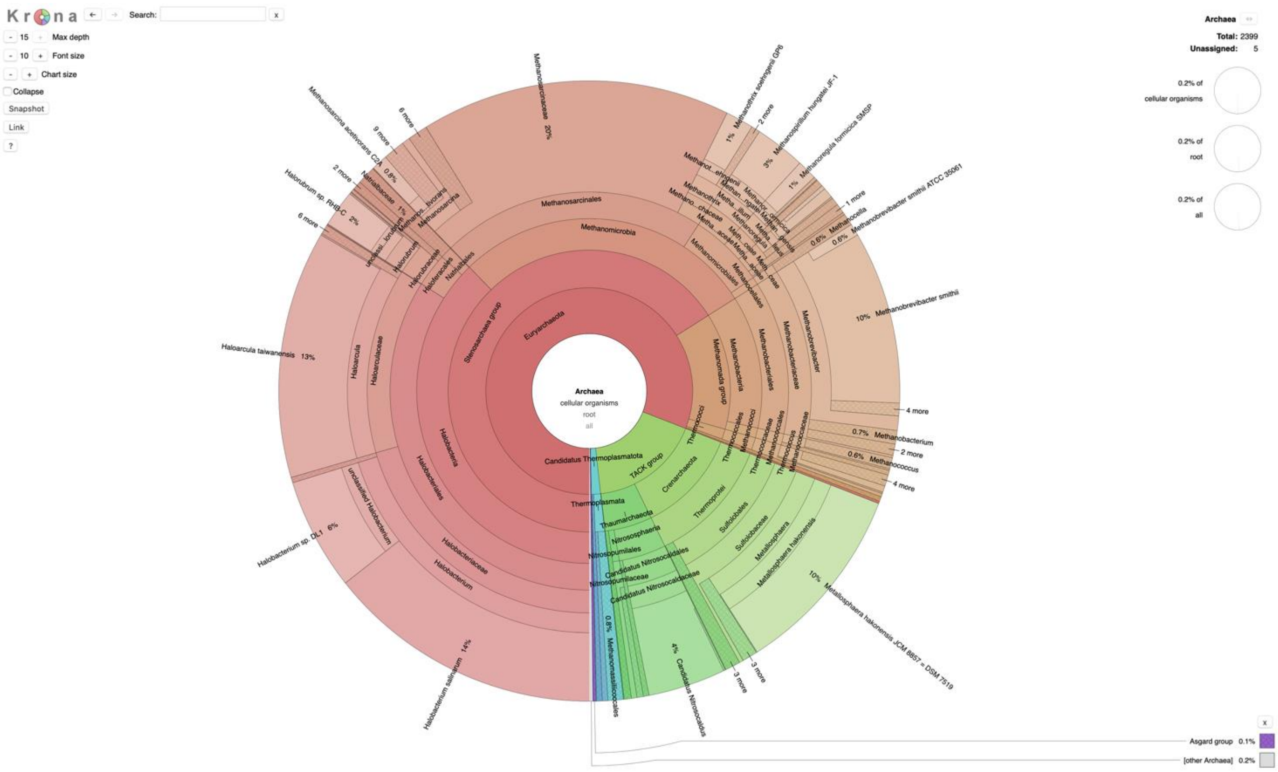
Supplementary: Per Sample Taxonomy Archaea RTW13_1A



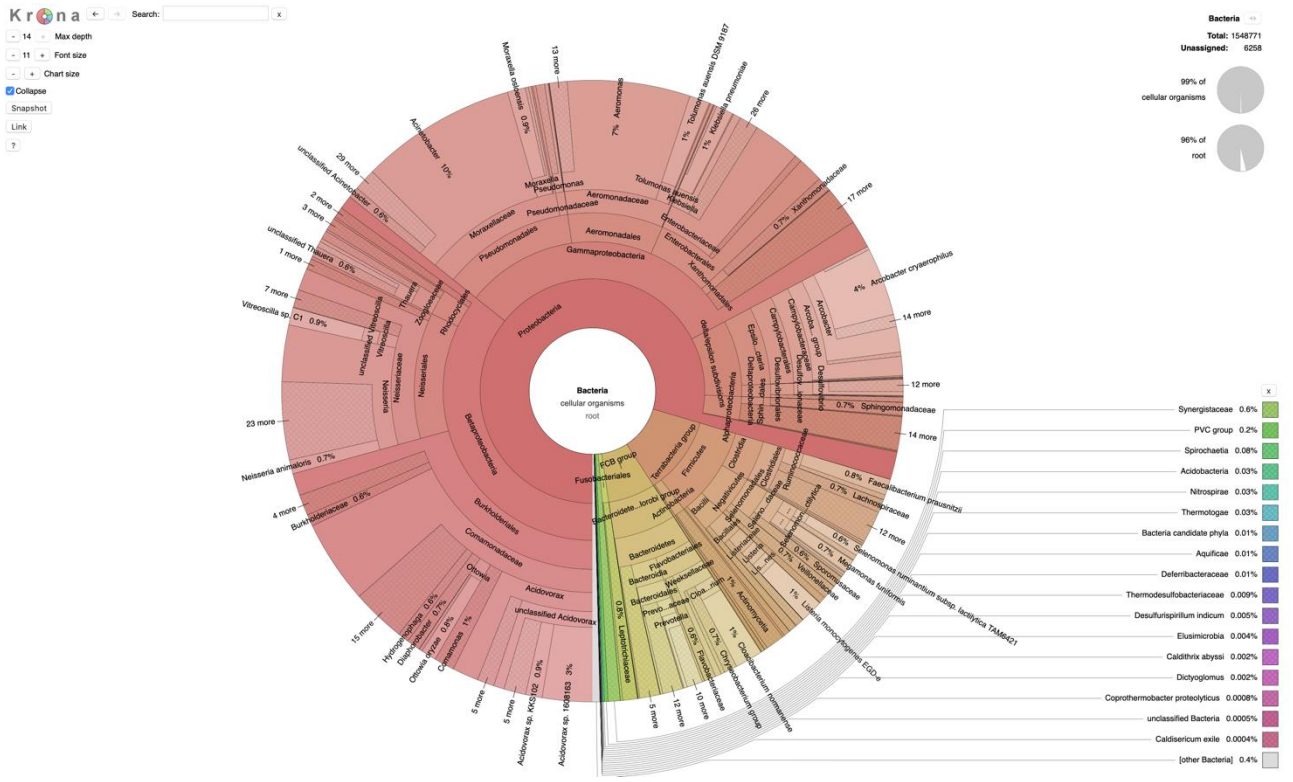
Supplementary: Per Sample Taxonomy Archaea RTW14_1A



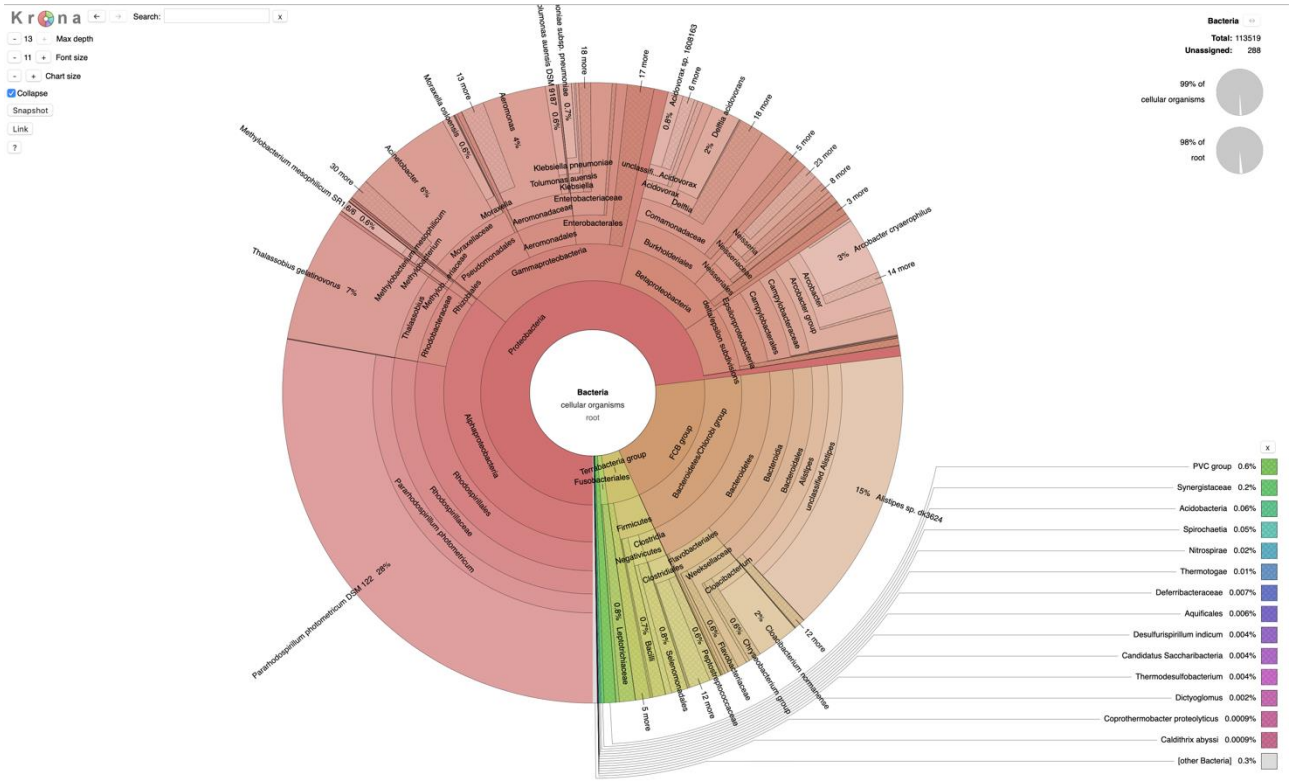
Supplementary: Per Sample Taxonomy Archaea RTW15_1A



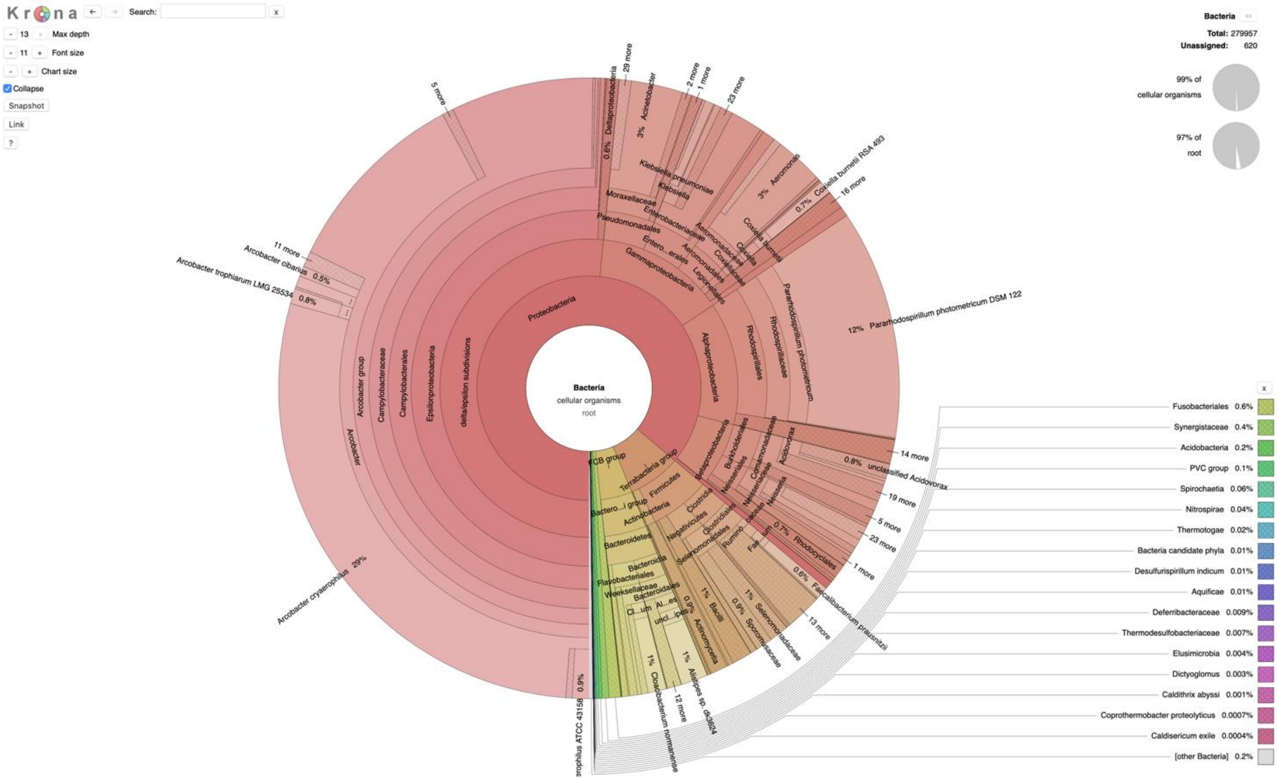
Supplementary: Per Sample Taxonomy Bacteria BSW1_1A



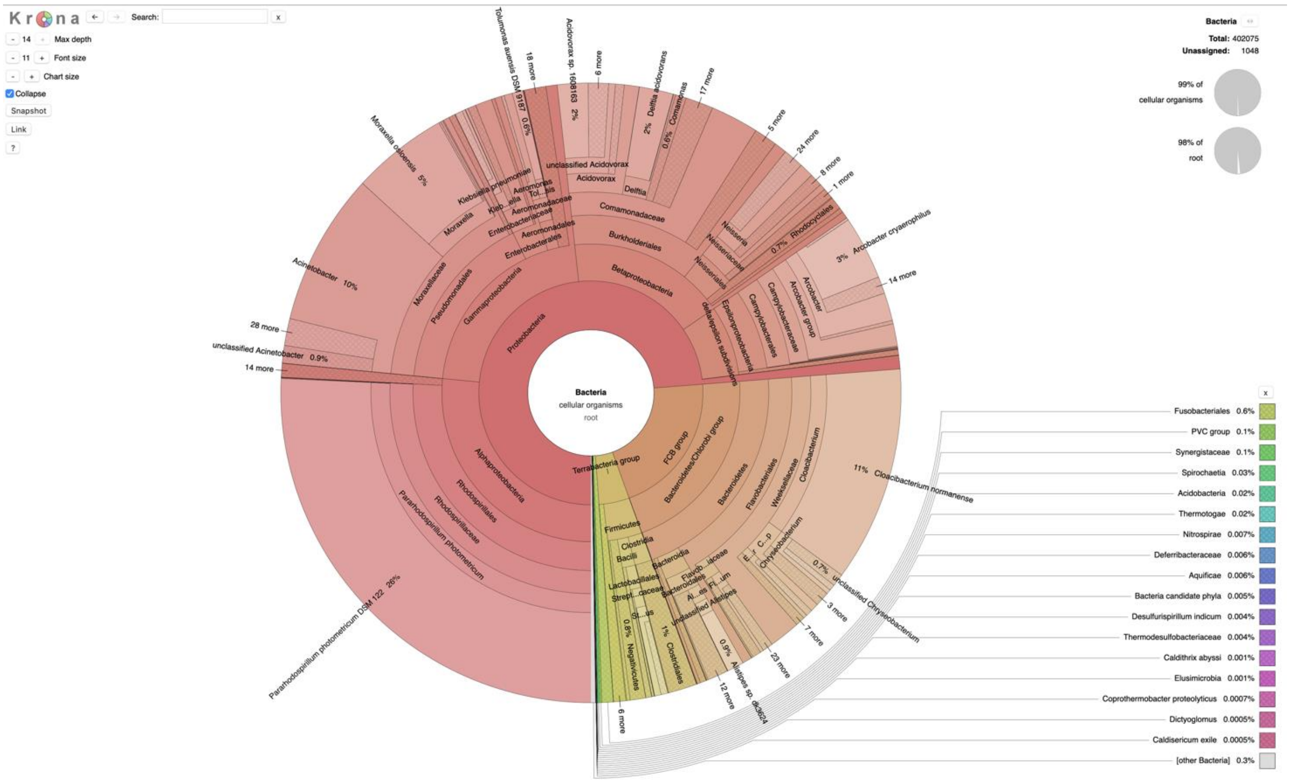
Supplementary: Per Sample Taxonomy Bacteria BSW2_1A



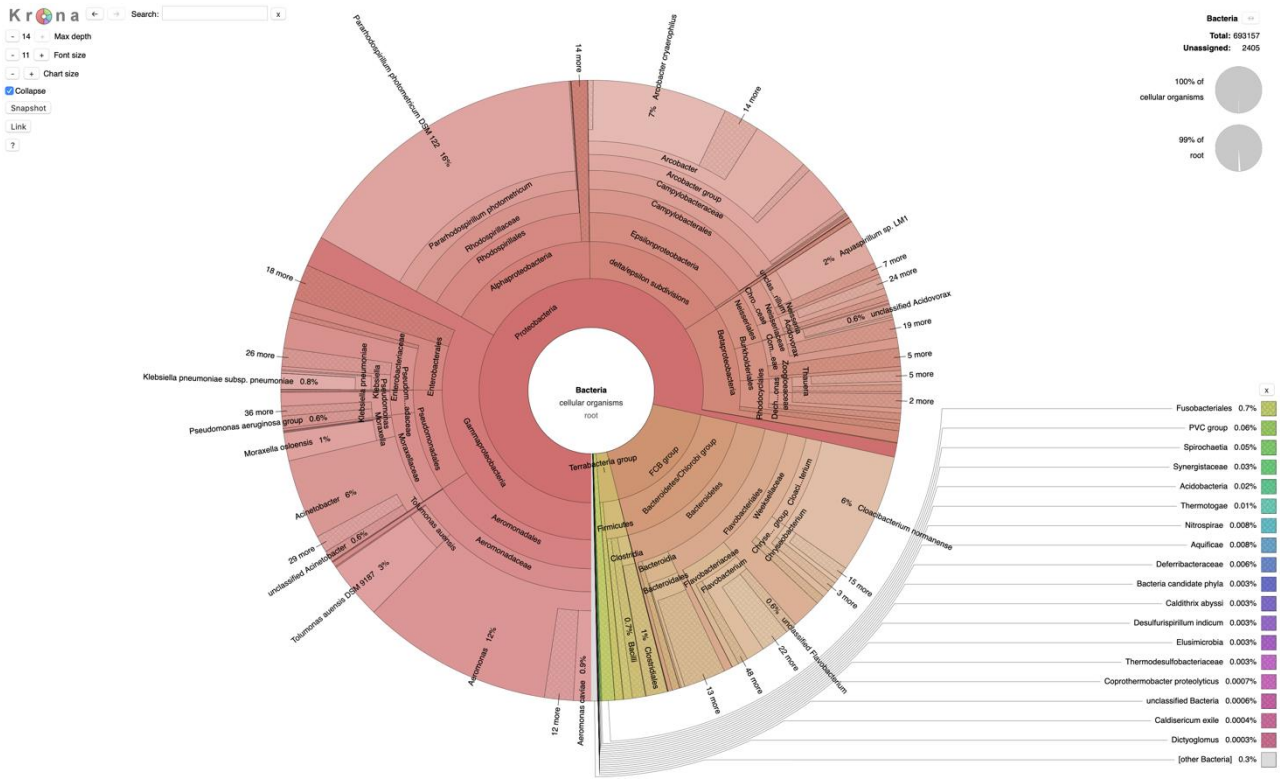
Supplementary: Per Sample Taxonomy Bacteria BSW6_1A



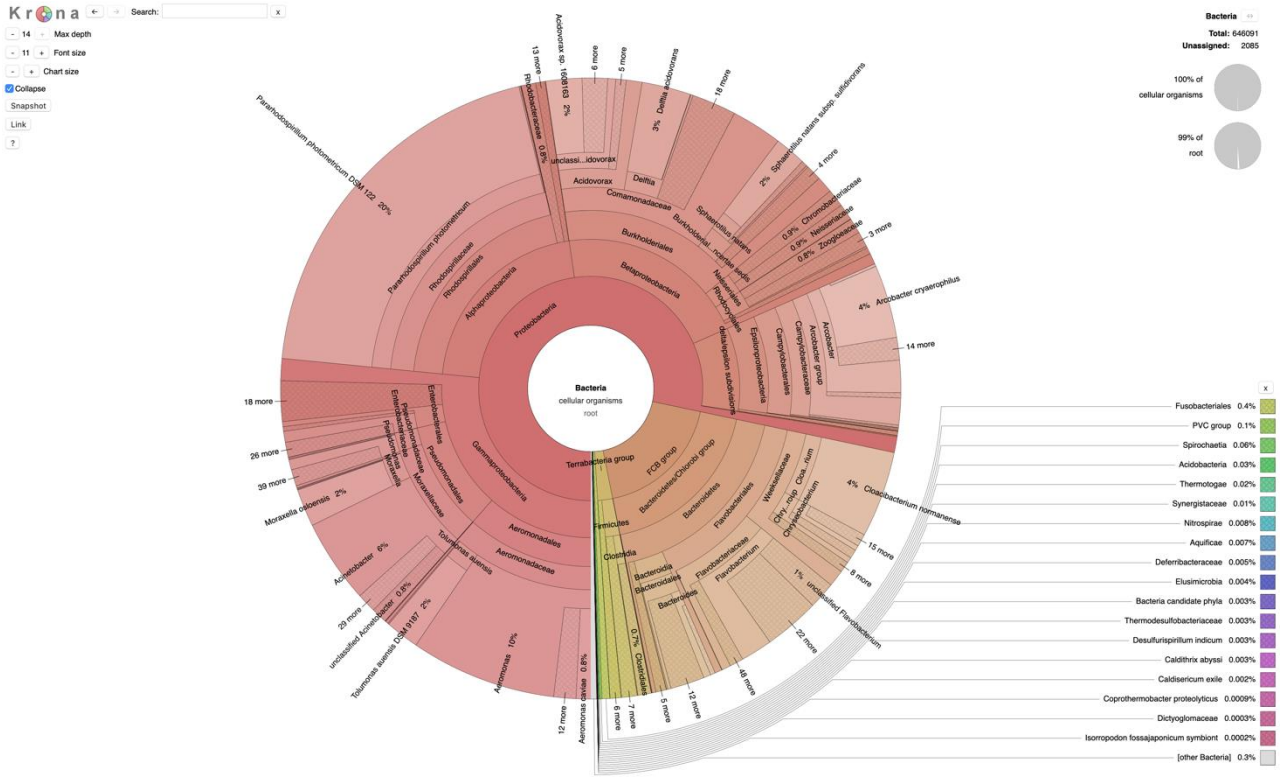
Supplementary: Per Sample Taxonomy Bacteria DW6_1A



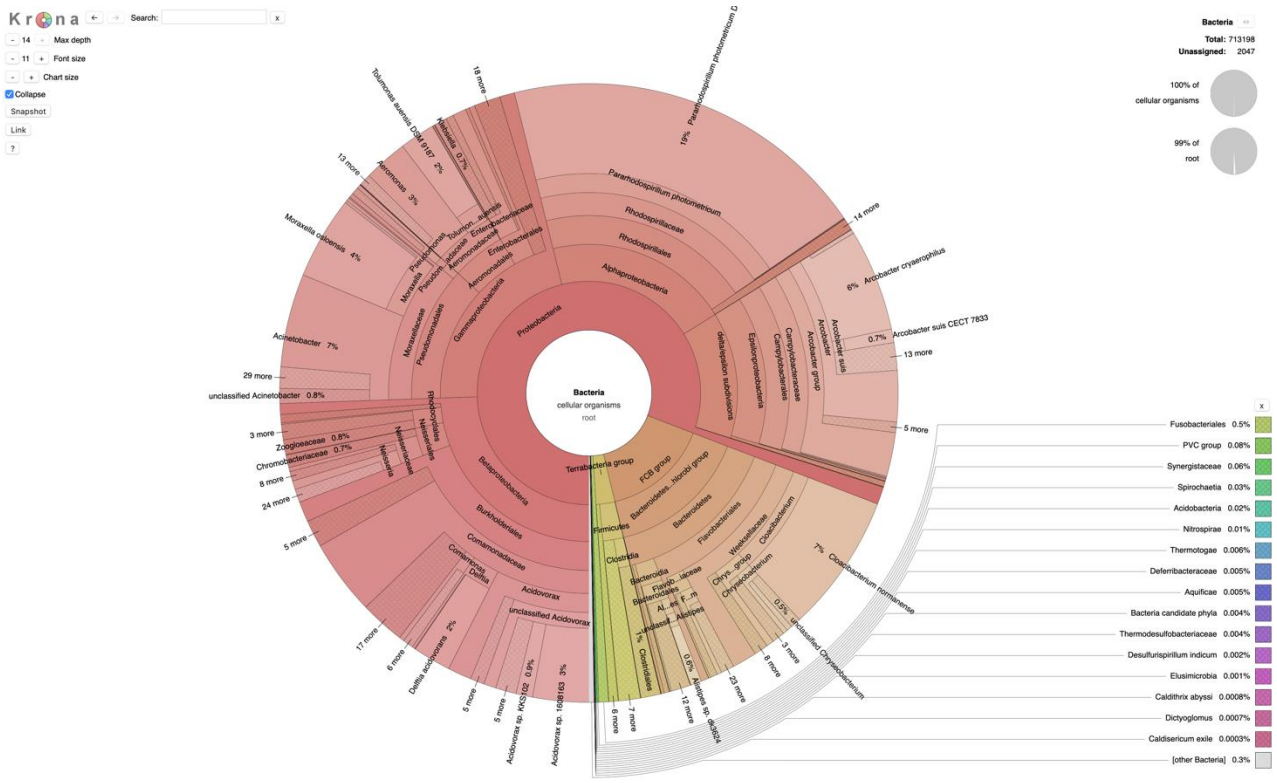
Supplementary: Per Sample Taxonomy Bacteria DW7_1A



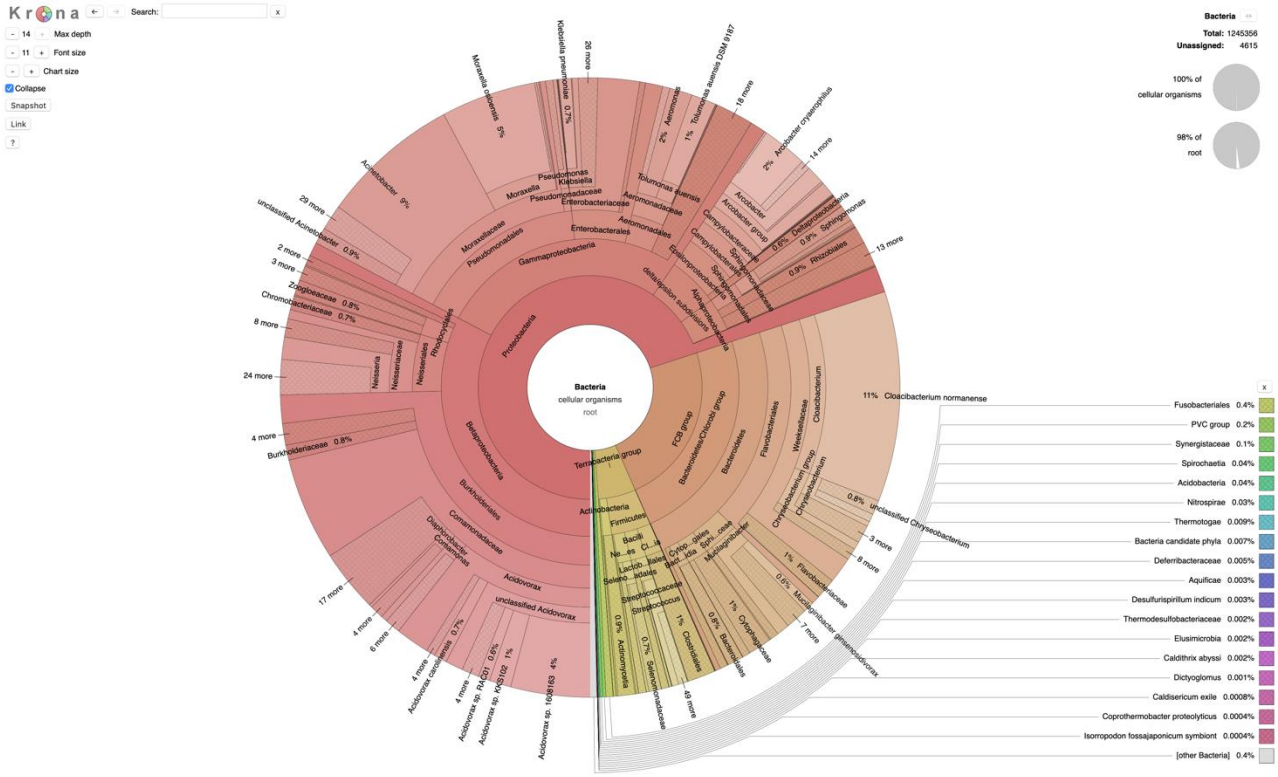
Supplementary: Per Sample Taxonomy Bacteria DW8_1A



Supplementary: Per Sample Taxonomy Bacteria DW10_1A

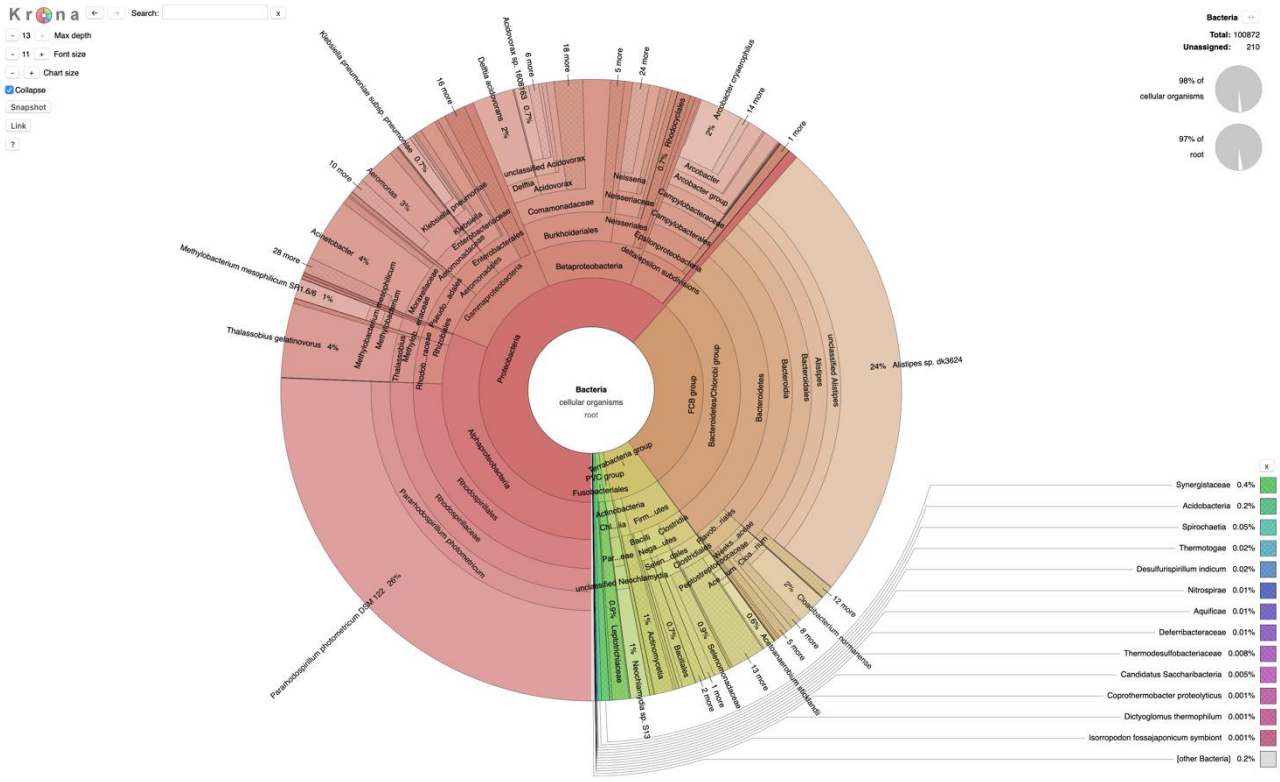


Supplementary: Per Sample Taxonomy Bacteria DW12_1A



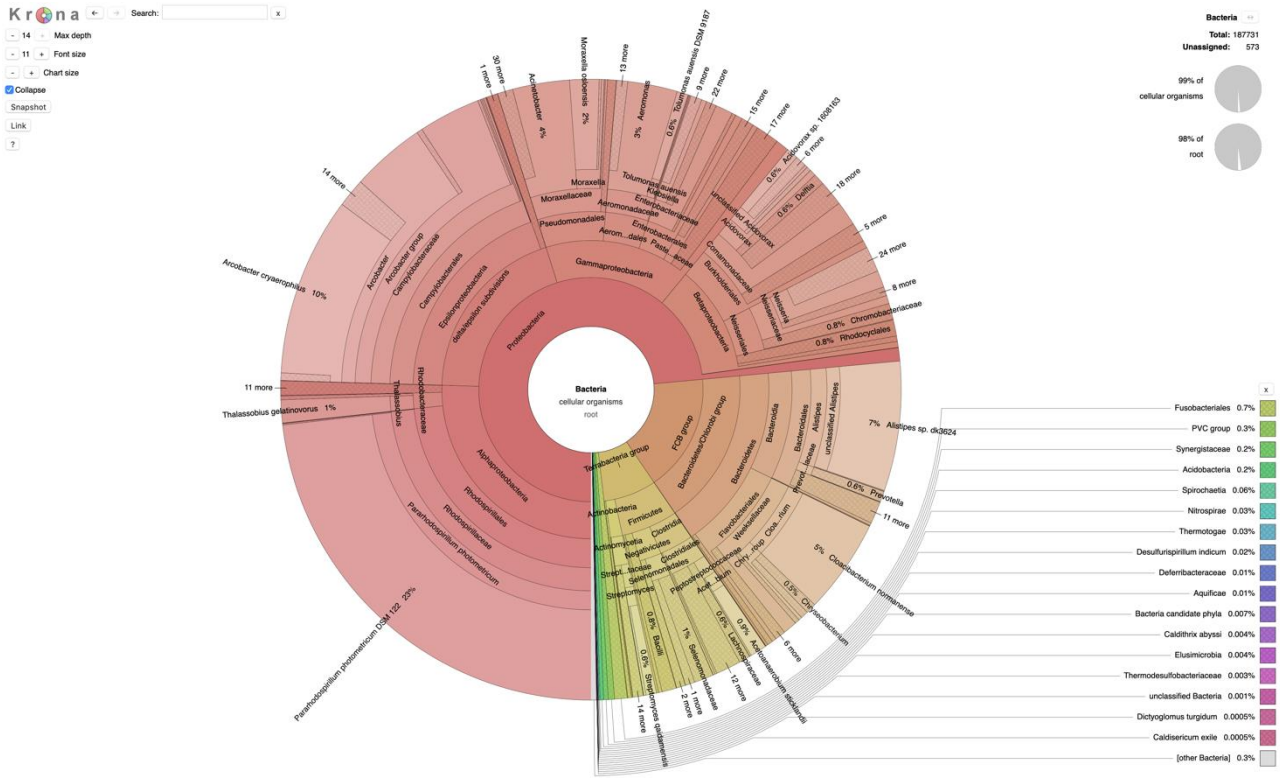
Supplementary: Per Sample Taxonomy Bacteria DW15_1A

Supplementary: Per Sample Taxonomy Bacteria RTW1_1A

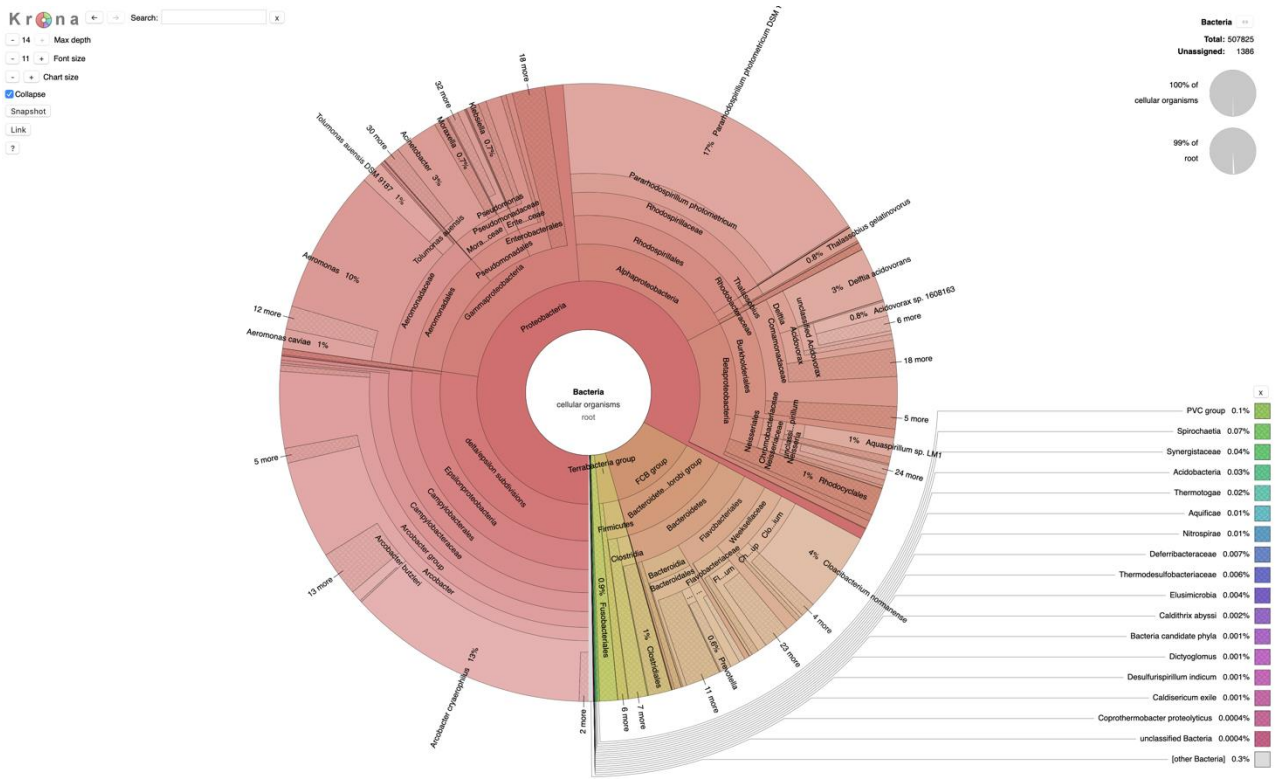


Supplementary: Per Sample Taxonomy Bacteria RTW2_1A

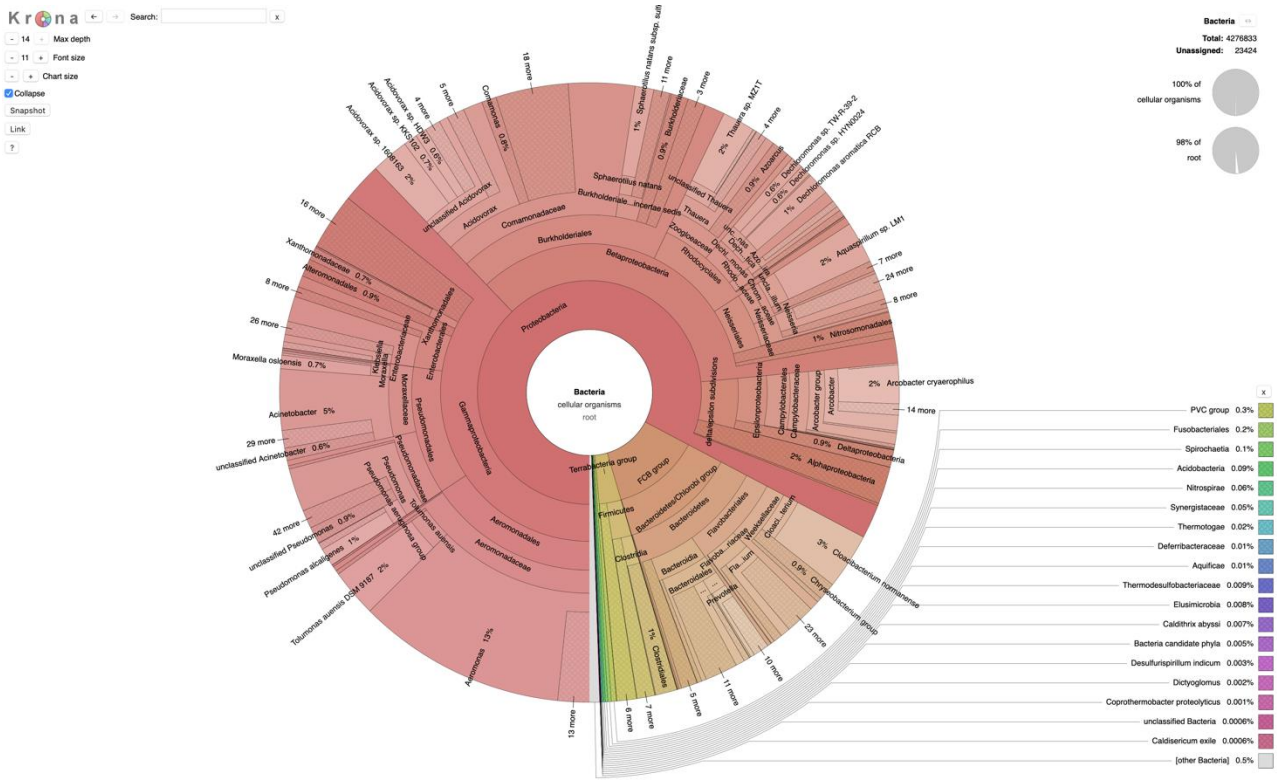
Supplementary: Per Sample Taxonomy Bacteria RTW6_1A



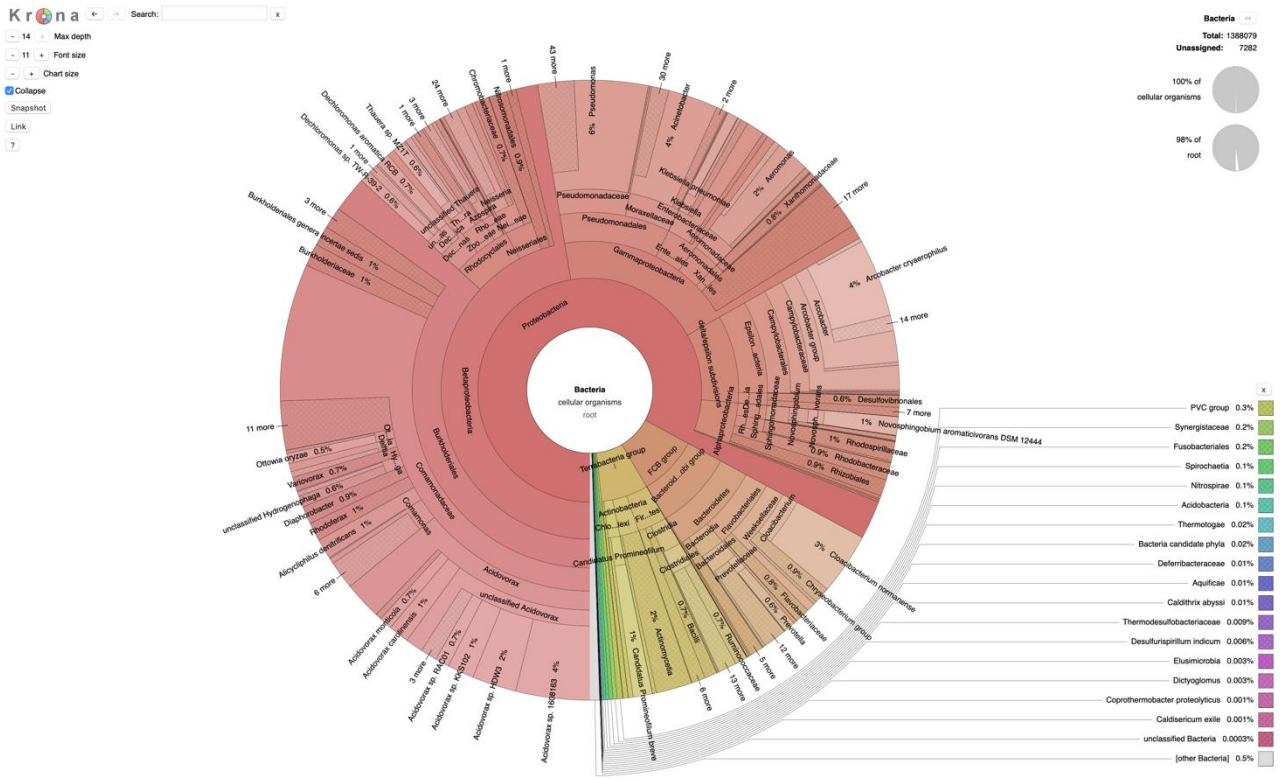
Supplementary: Per Sample Taxonomy Bacteria RTW7_1A



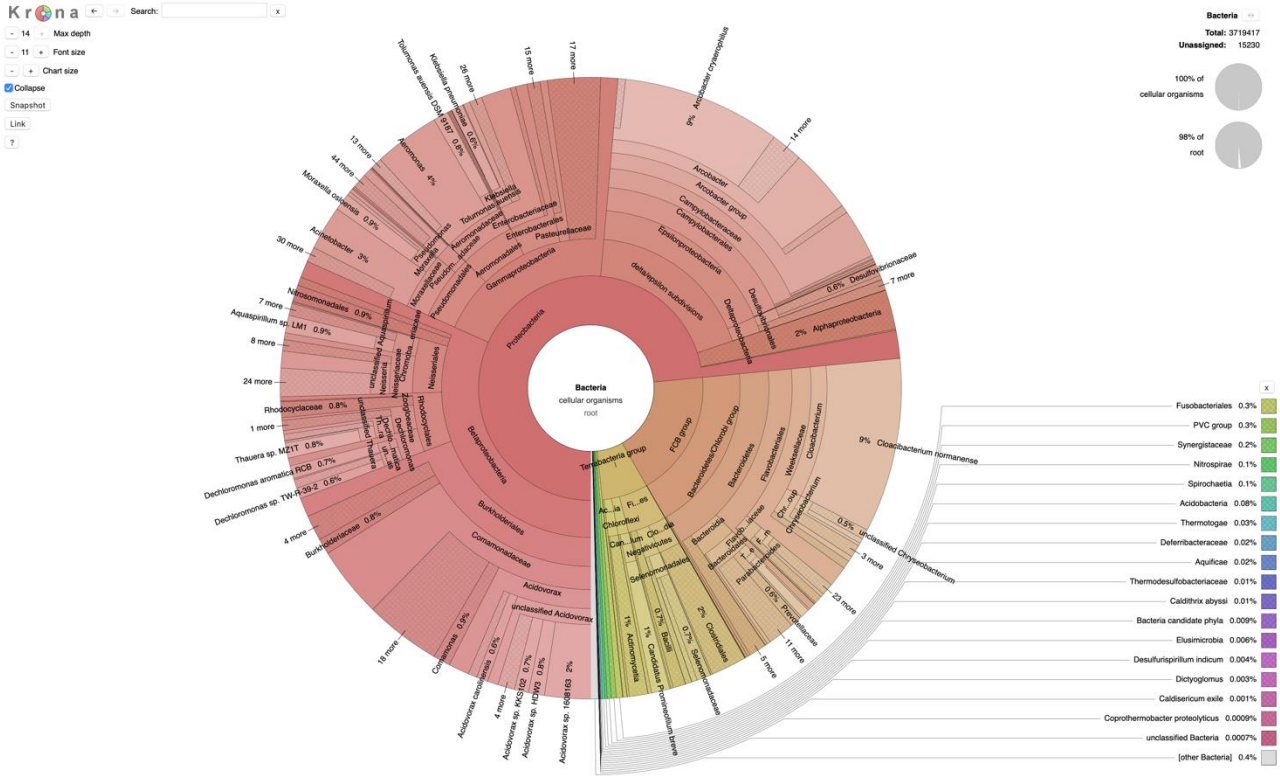
Supplementary: Per Sample Taxonomy Bacteria RTW8_1A



Supplementary: Per Sample Taxonomy Bacteria RTW10_1A

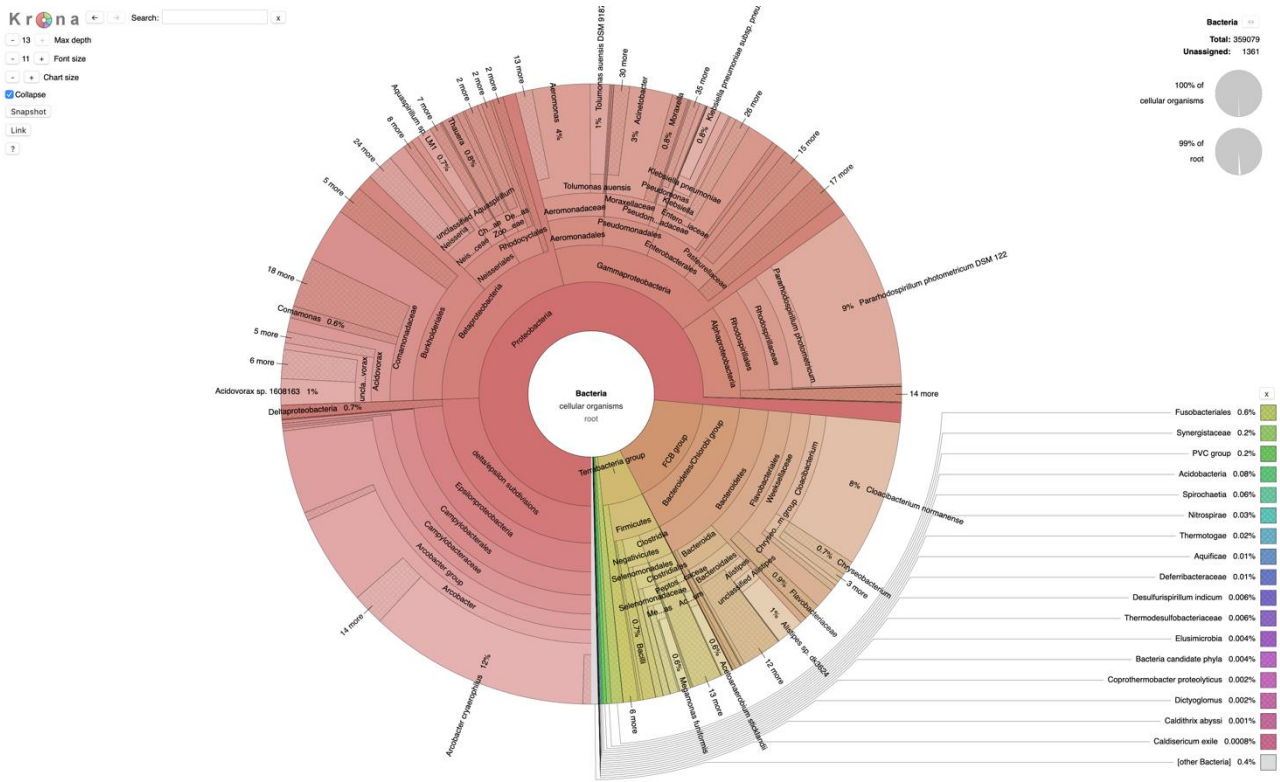


Supplementary: Per Sample Taxonomy Bacteria RTW11_1A

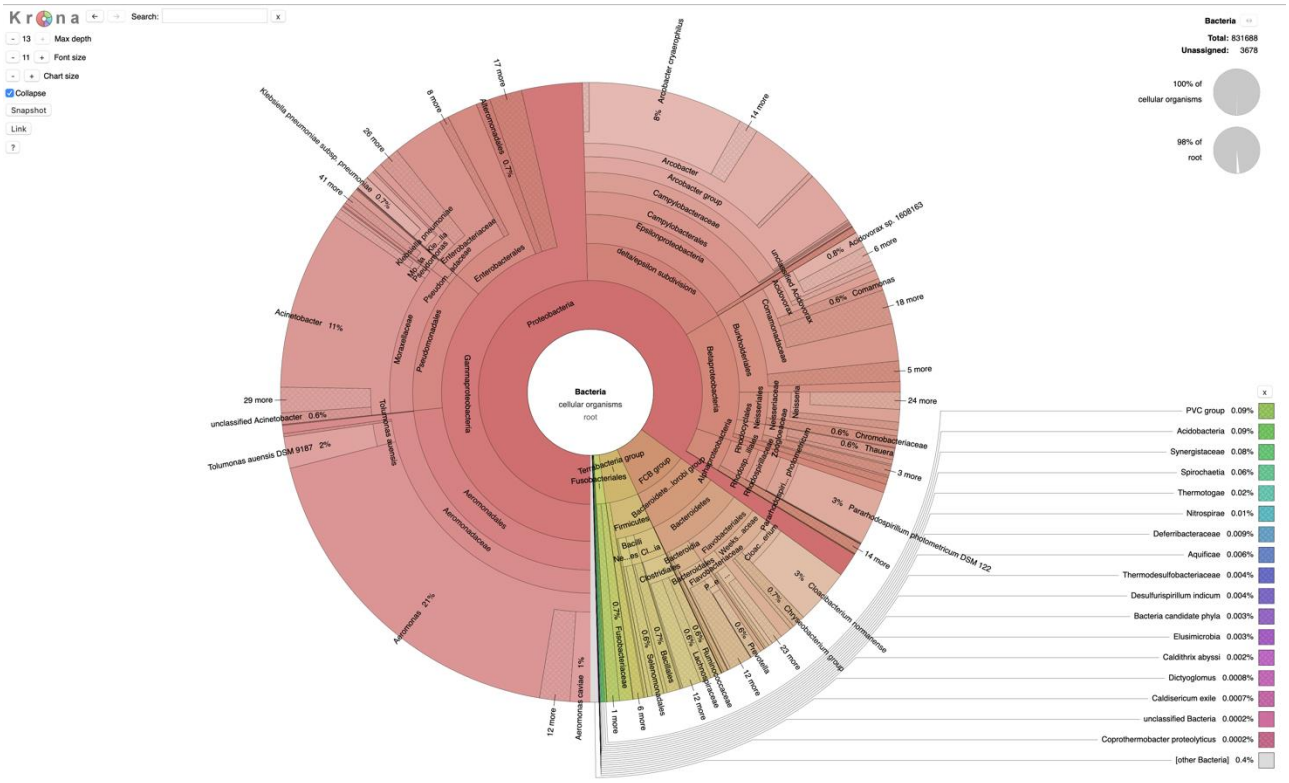


Supplementary: Per Sample Taxonomy Bacteria RTW12_1A

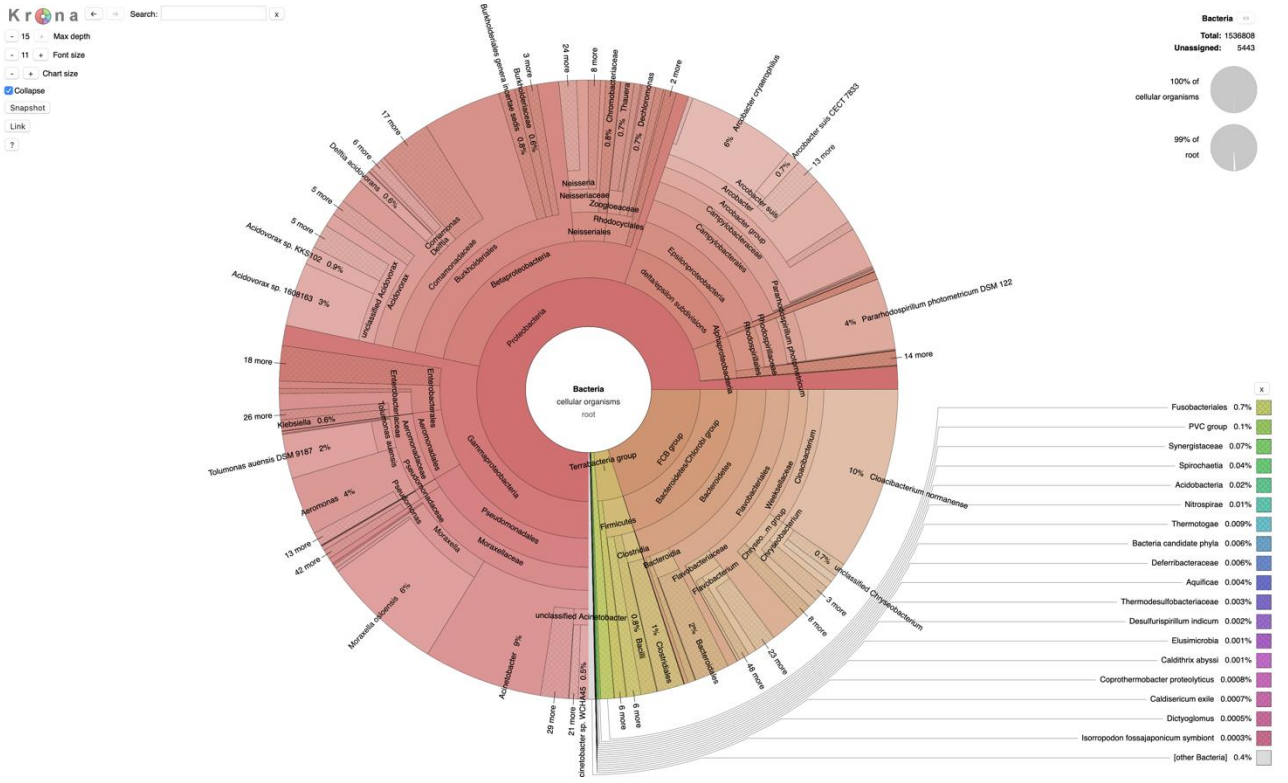
Supplementary: Per Sample Taxonomy Bacteria RTW13_1A



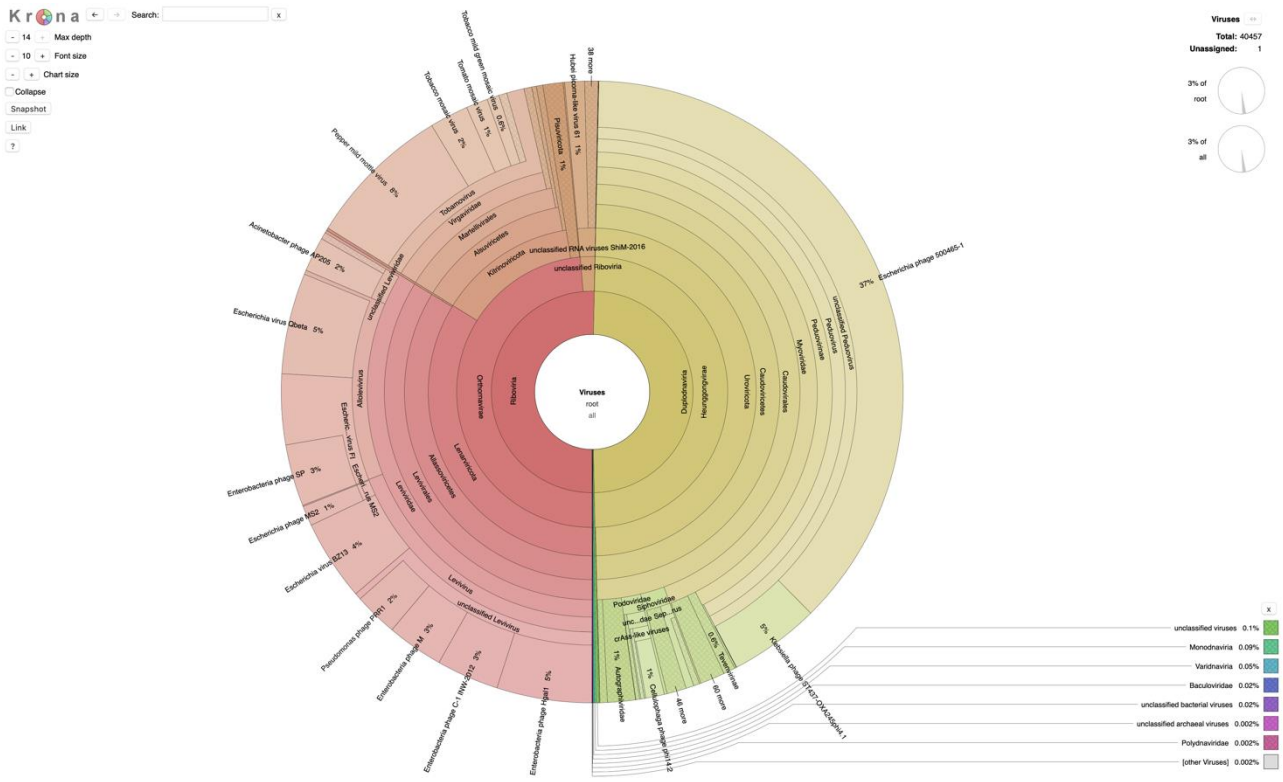
Supplementary: Per Sample Taxonomy Bacteria RTW14_1A



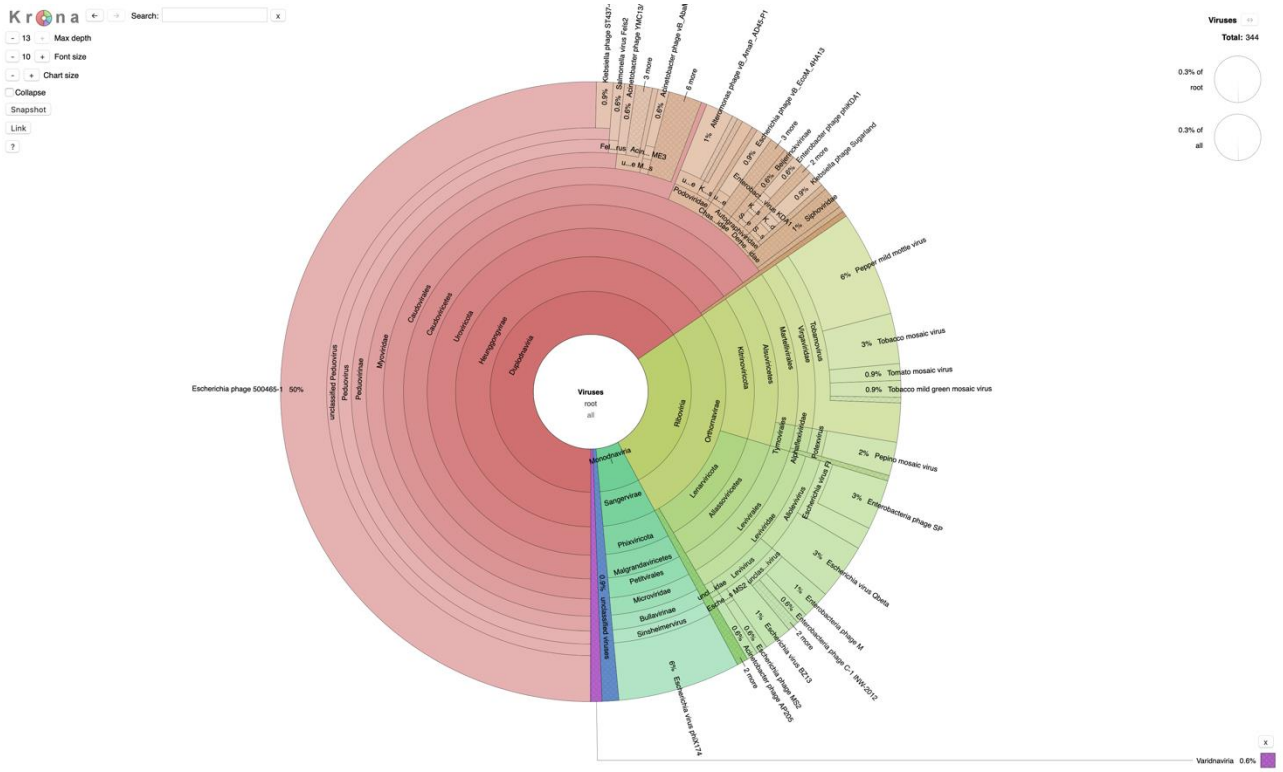
Supplementary: Per Sample Taxonomy Bacteria RTW15_1A



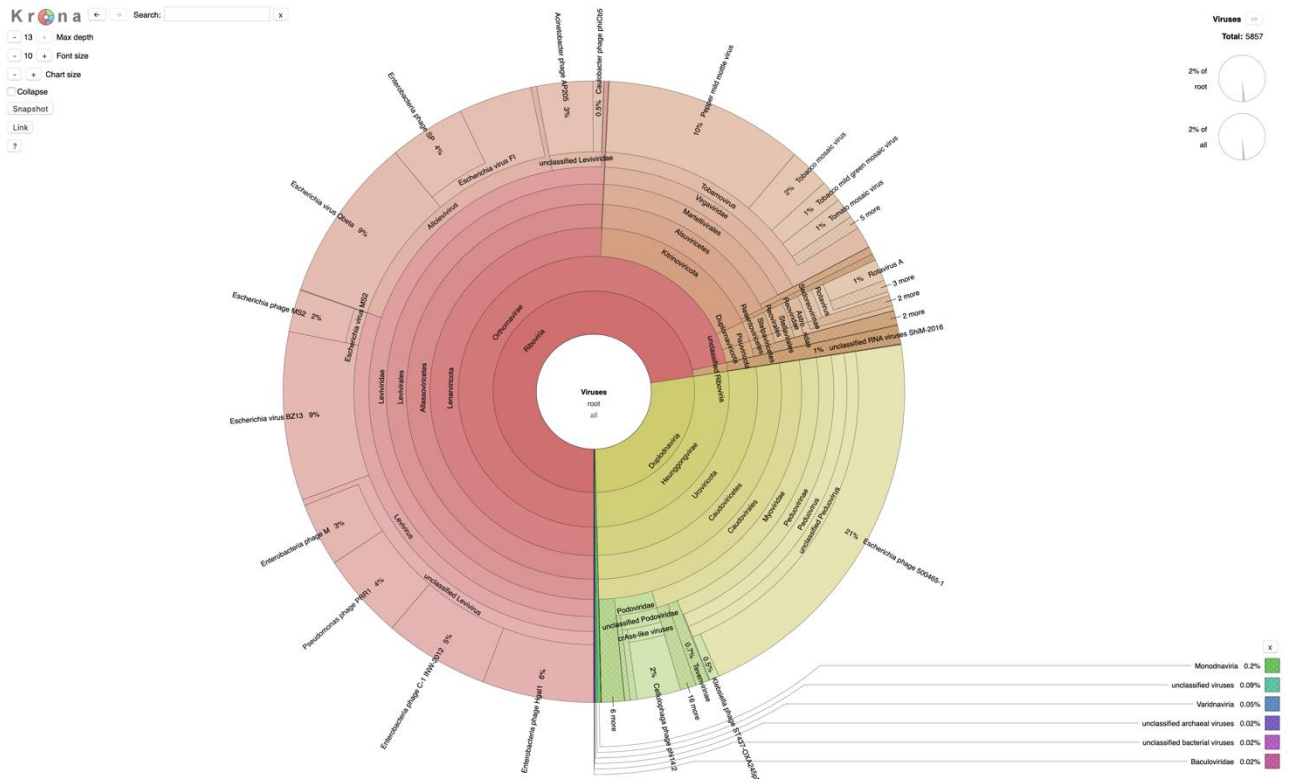
Supplementary: Per Sample Taxonomy Virus BSW1_1A



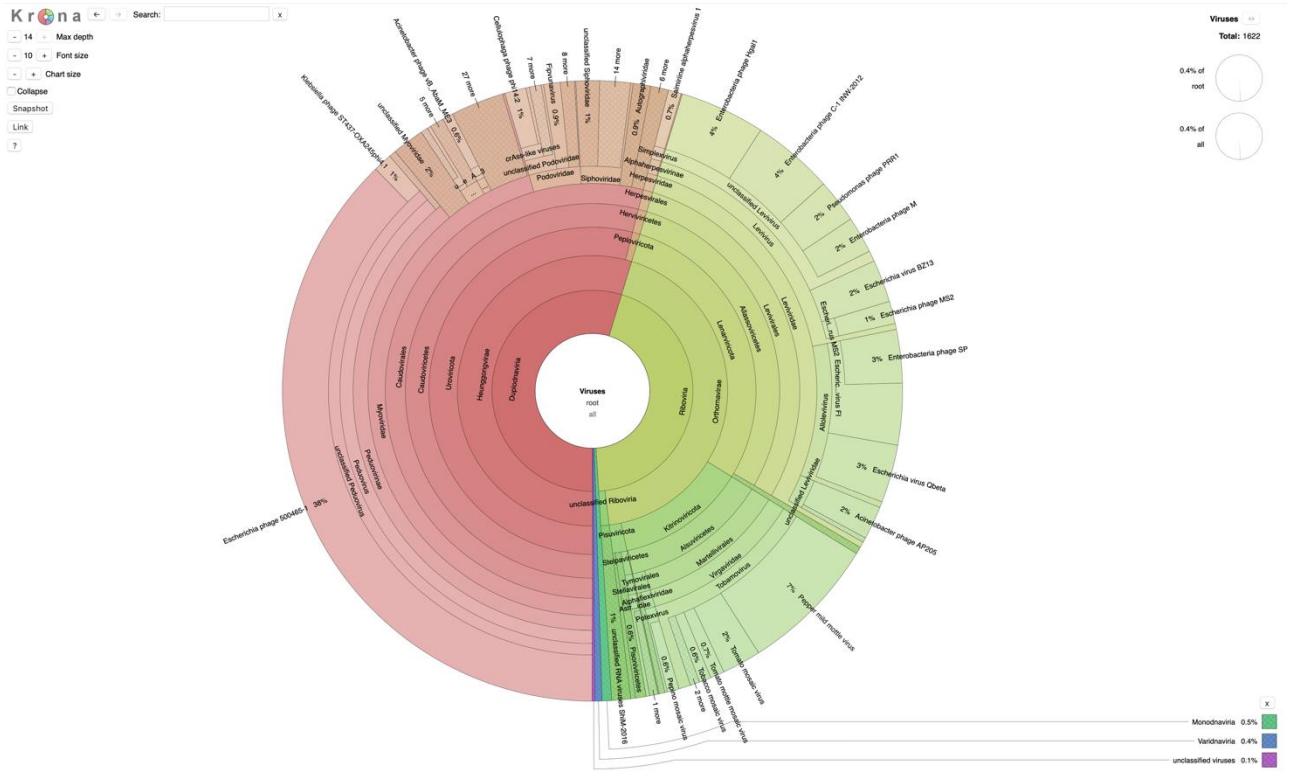
Supplementary: Per Sample Taxonomy Virus BSW2_1A



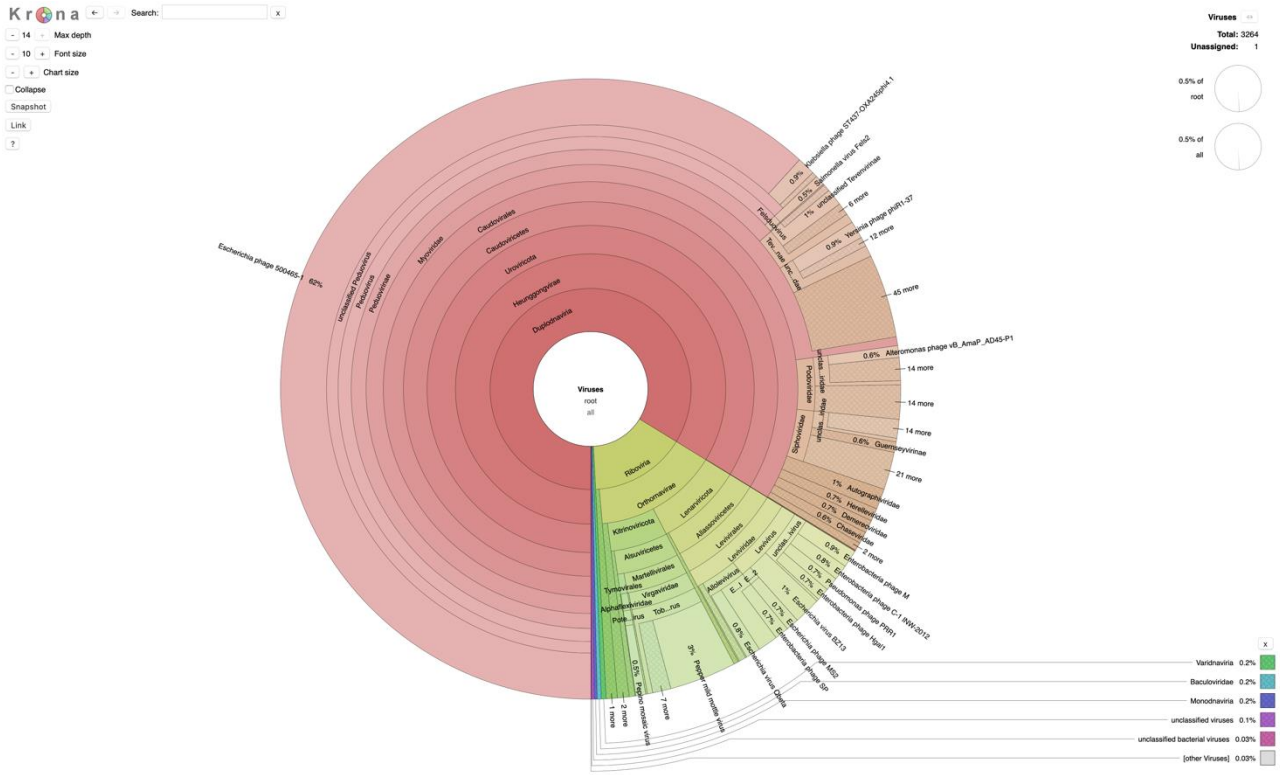
Supplementary: Per Sample Taxonomy Virus BSW6_1A



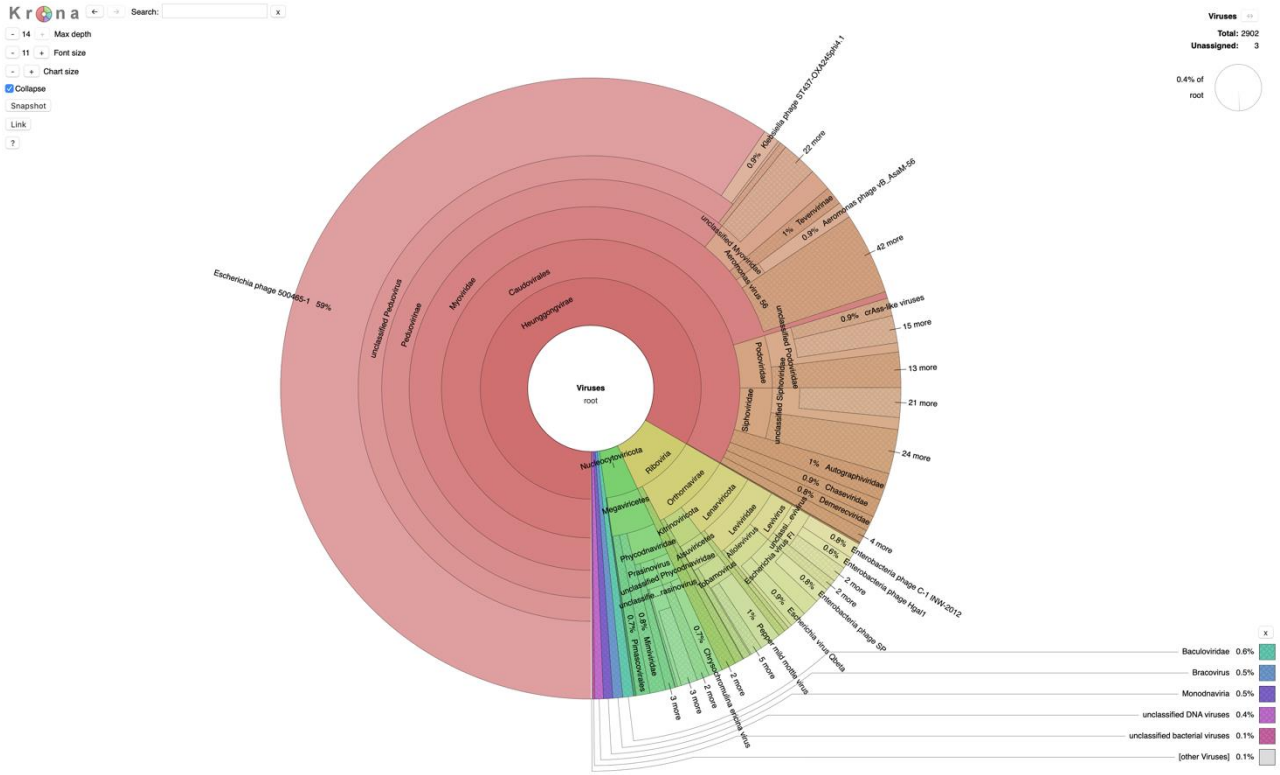
Supplementary: Per Sample Taxonomy Virus DW6_1A



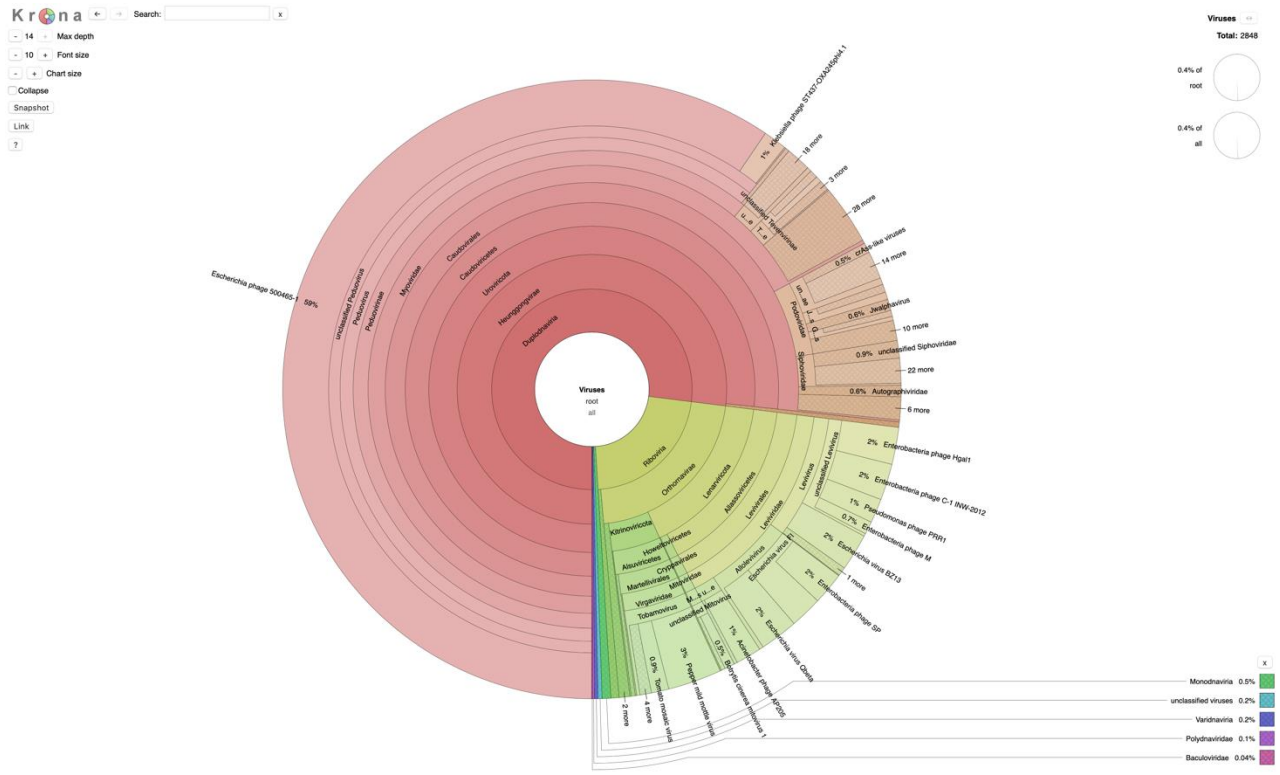
Supplementary: Per Sample Taxonomy Virus DW7_1A



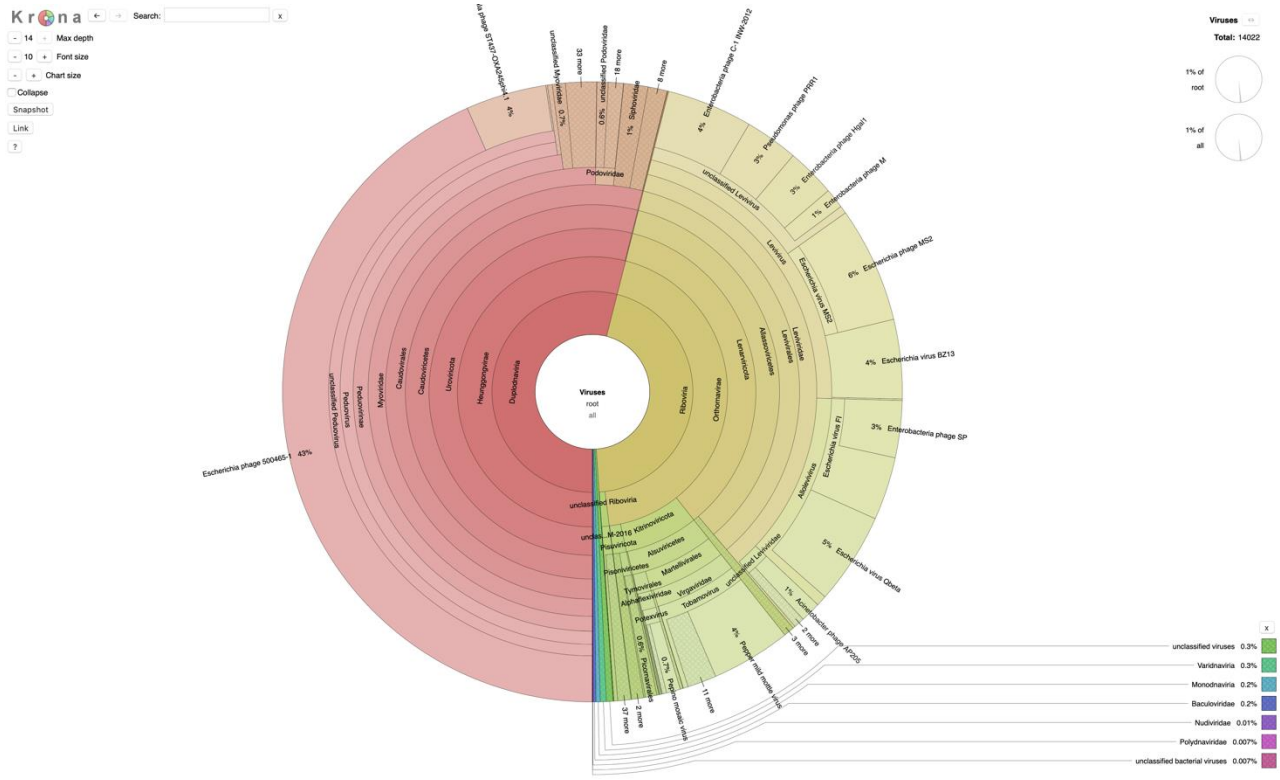
Supplementary: Per Sample Taxonomy Virus DW8_1A



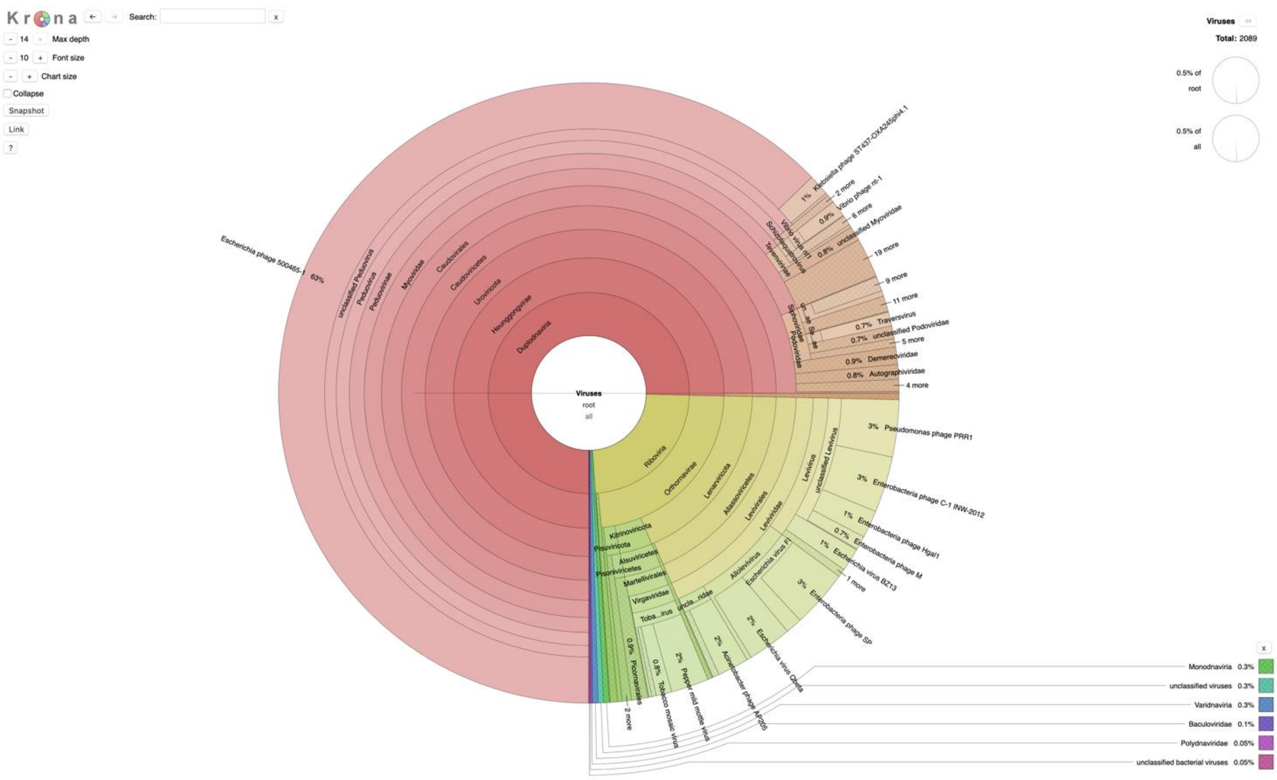
Supplementary: Per Sample Taxonomy Virus DW10_1A



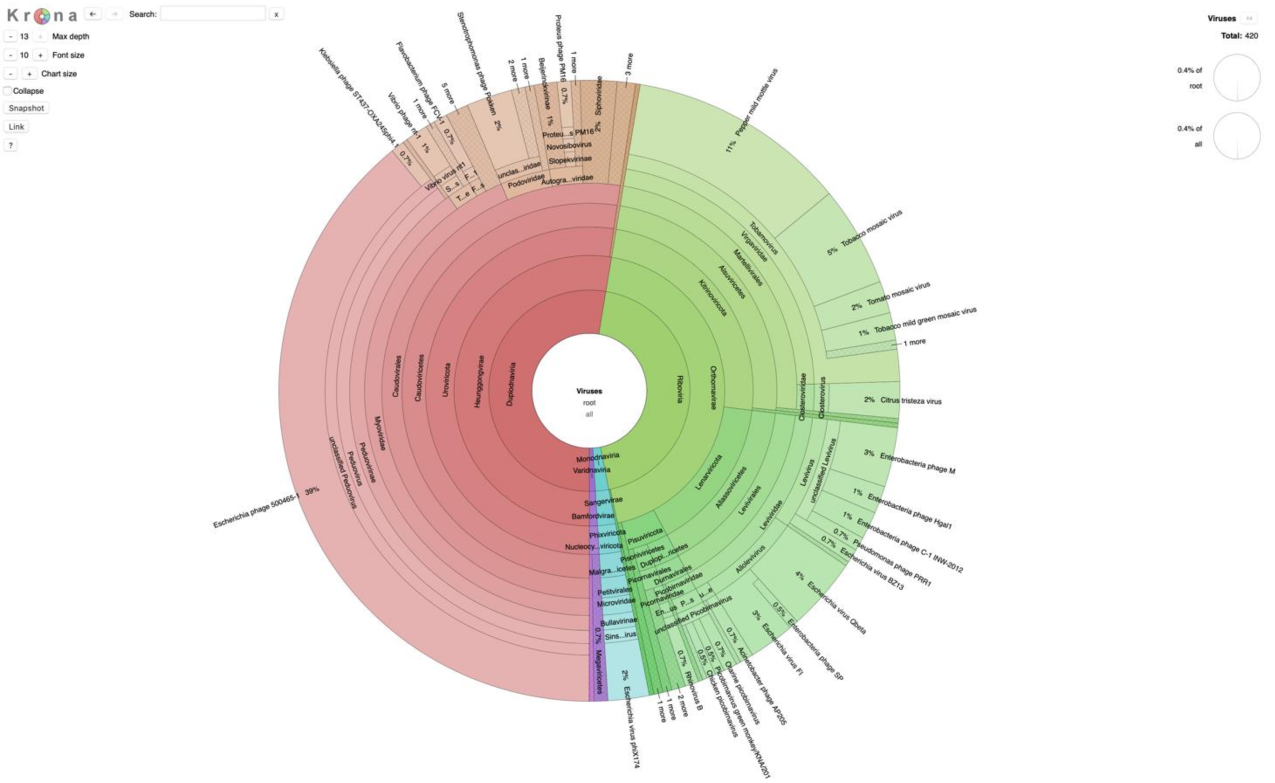
Supplementary: Per Sample Taxonomy Virus DW12_1A



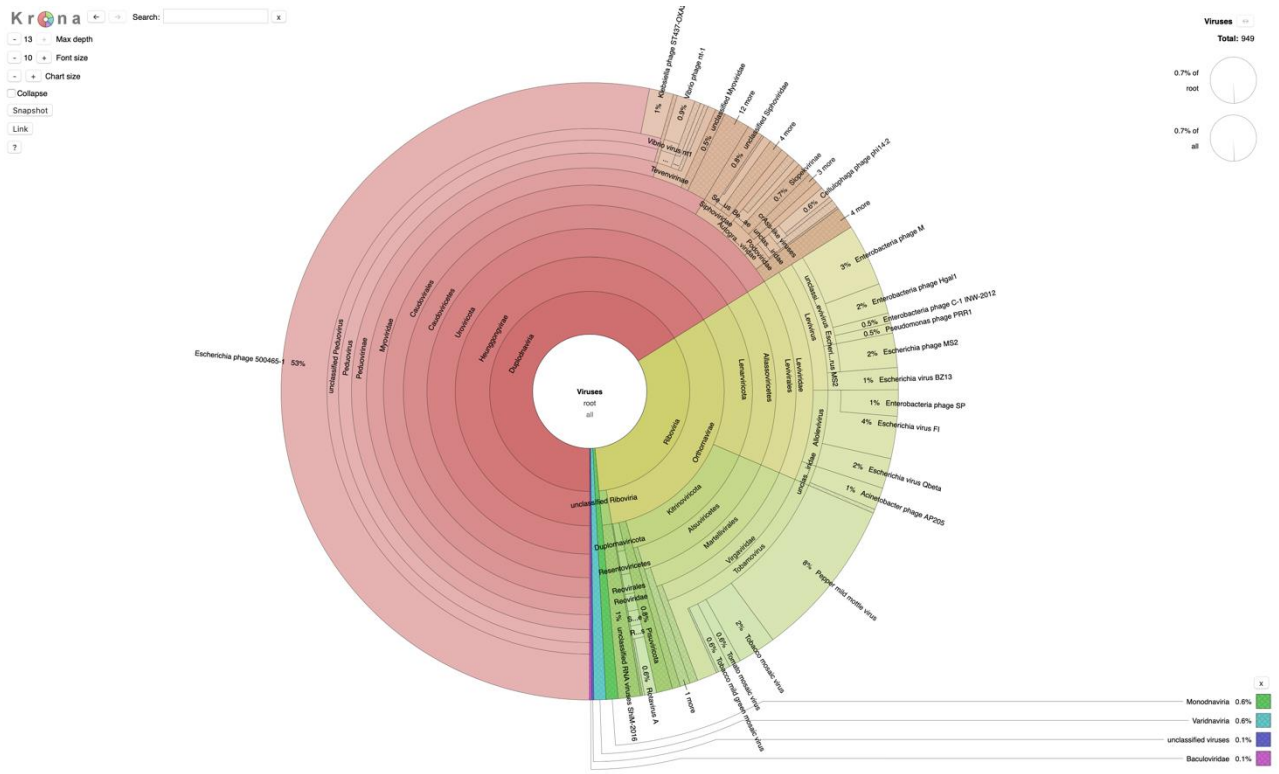
Supplementary: Per Sample Taxonomy Virus DW15_1A



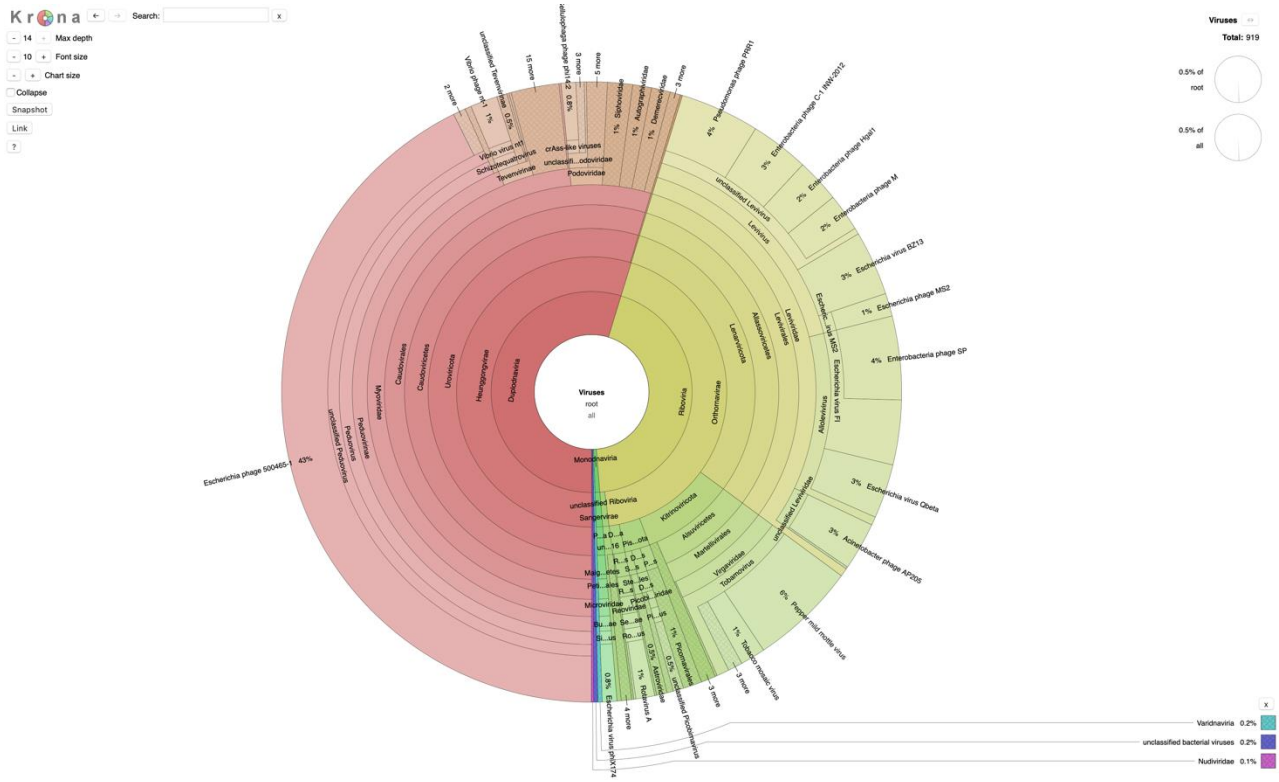
Supplementary: Per Sample Taxonomy Virus RTW1_1A



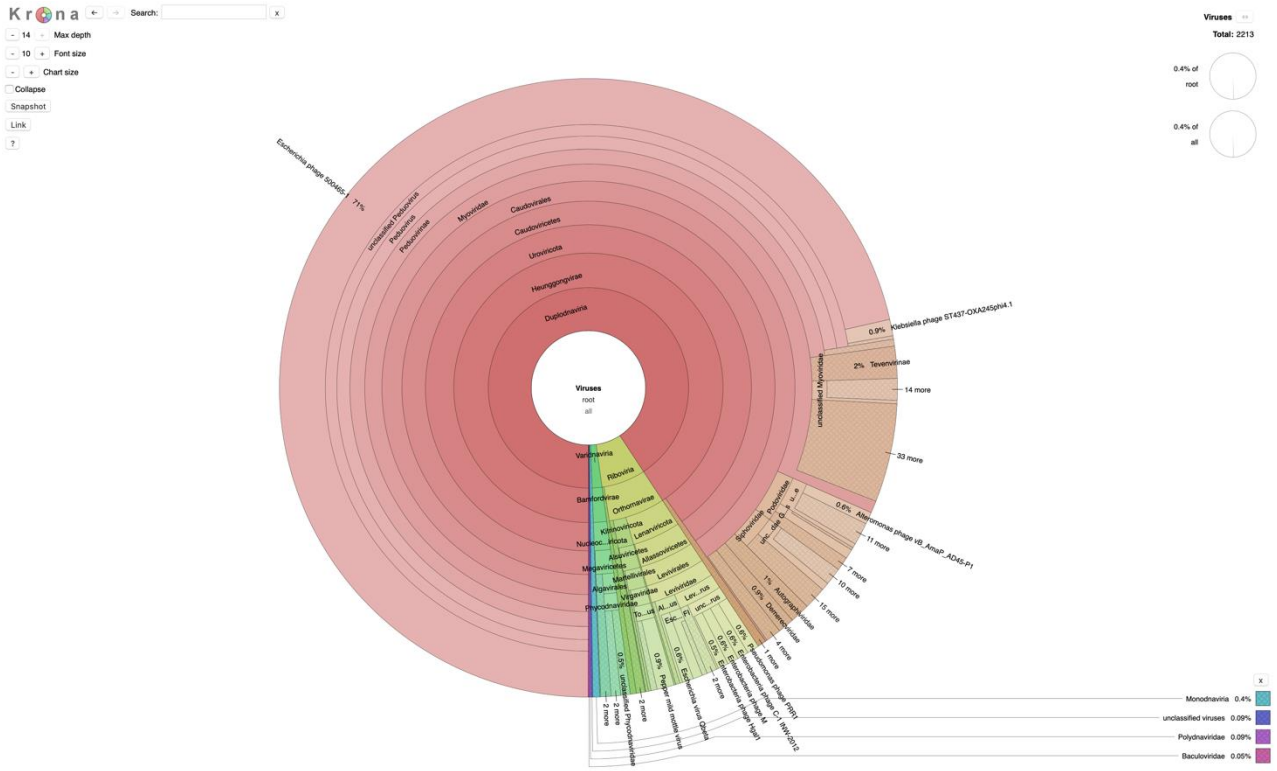
Supplementary: Per Sample Taxonomy Virus RTW2_1A



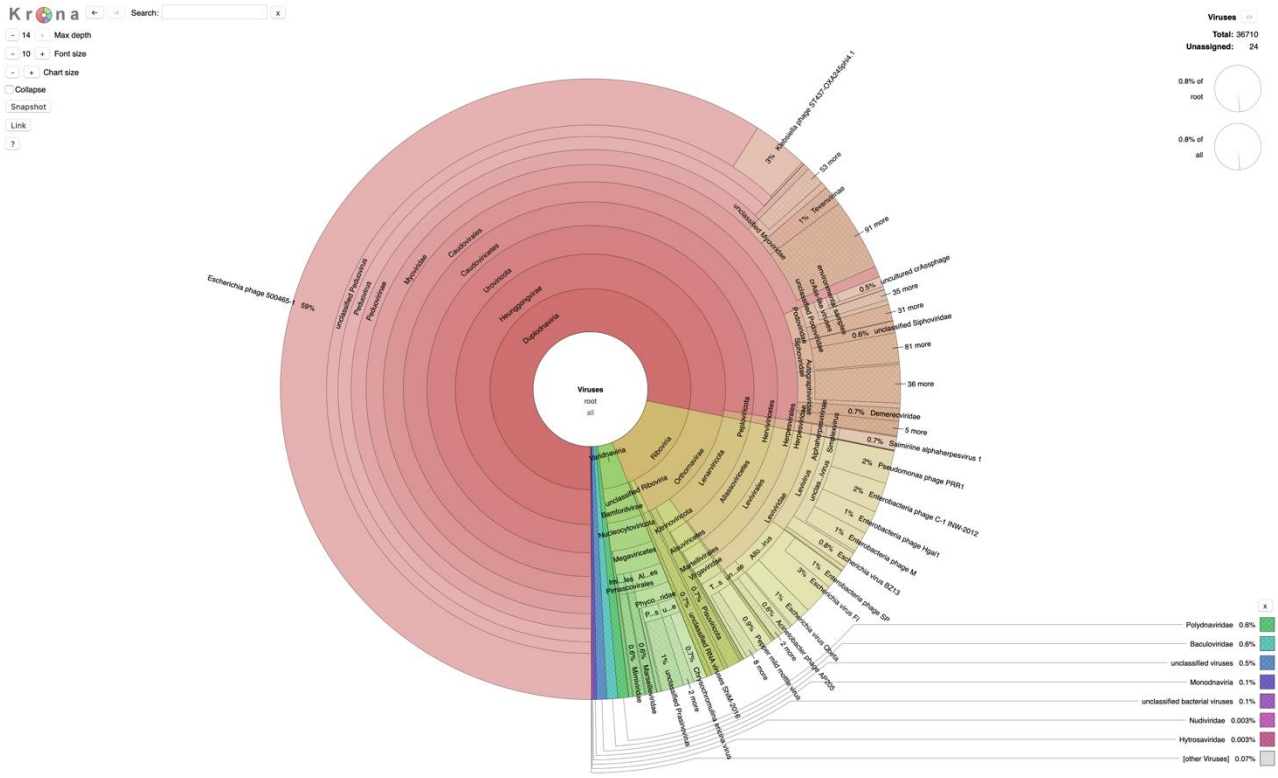
Supplementary: Per Sample Taxonomy Virus RTW6_1A



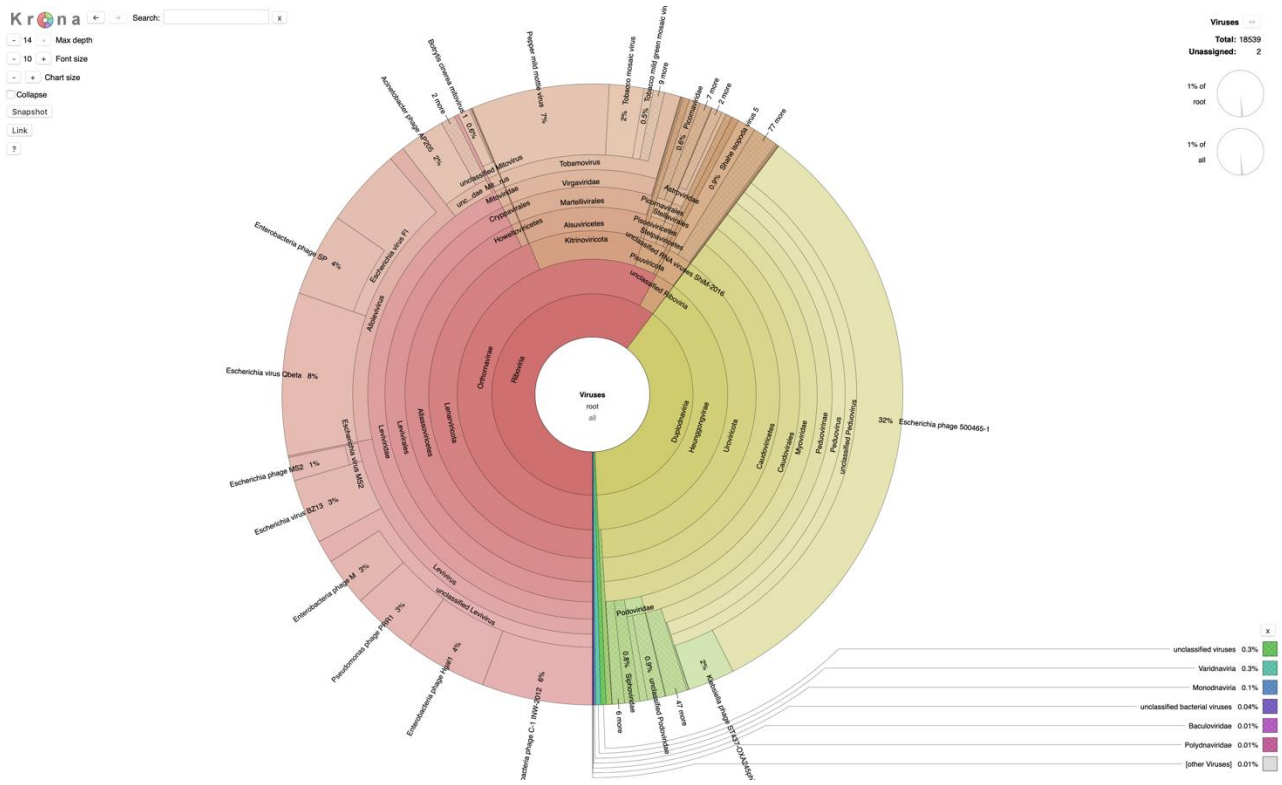
Supplementary: Per Sample Taxonomy Virus RTW7_1A



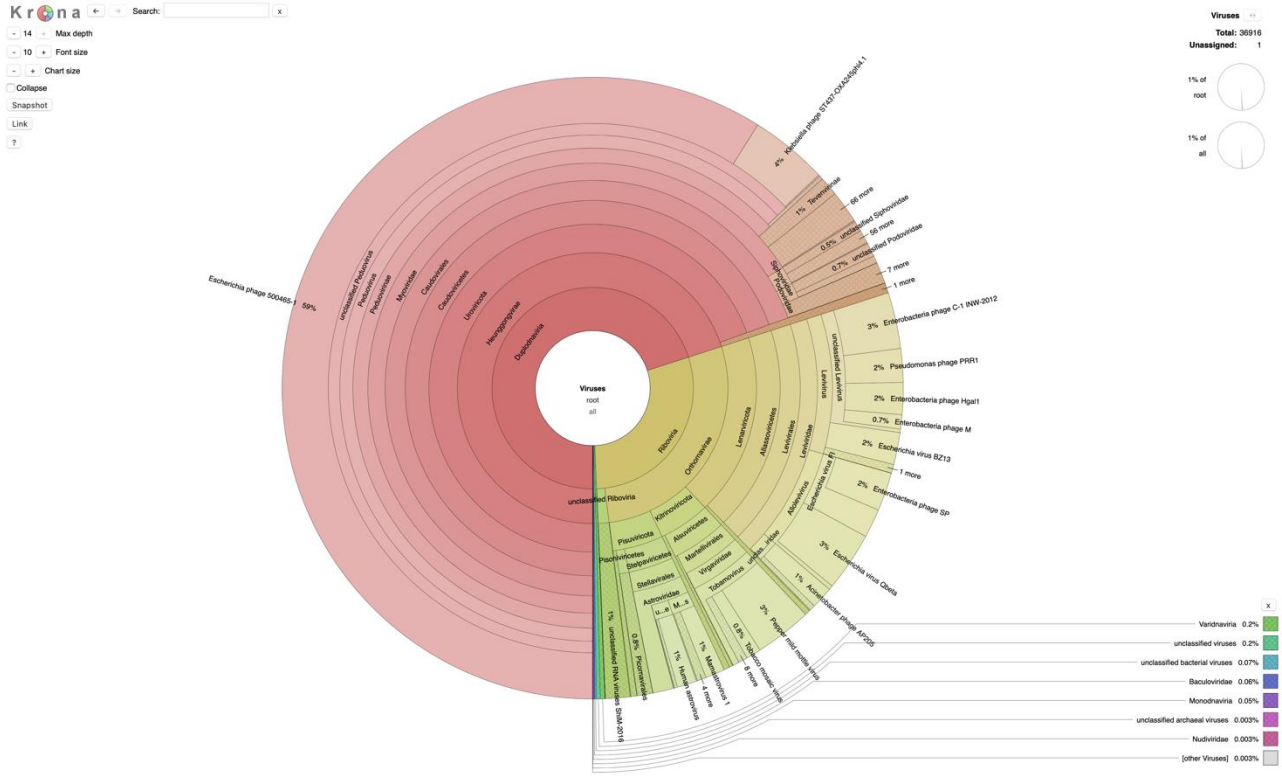
Supplementary: Per Sample Taxonomy Virus RTW8_1A



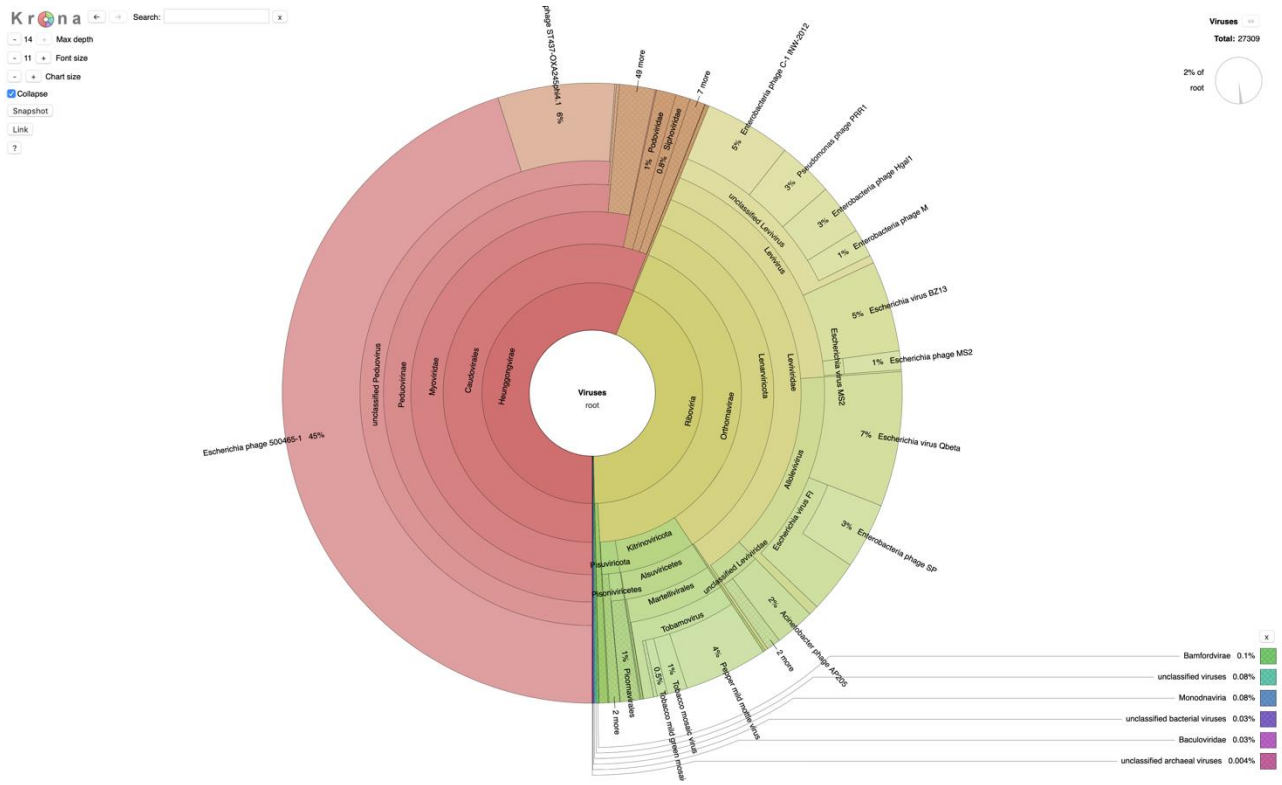
Supplementary: Per Sample Taxonomy Virus RTW10_1A



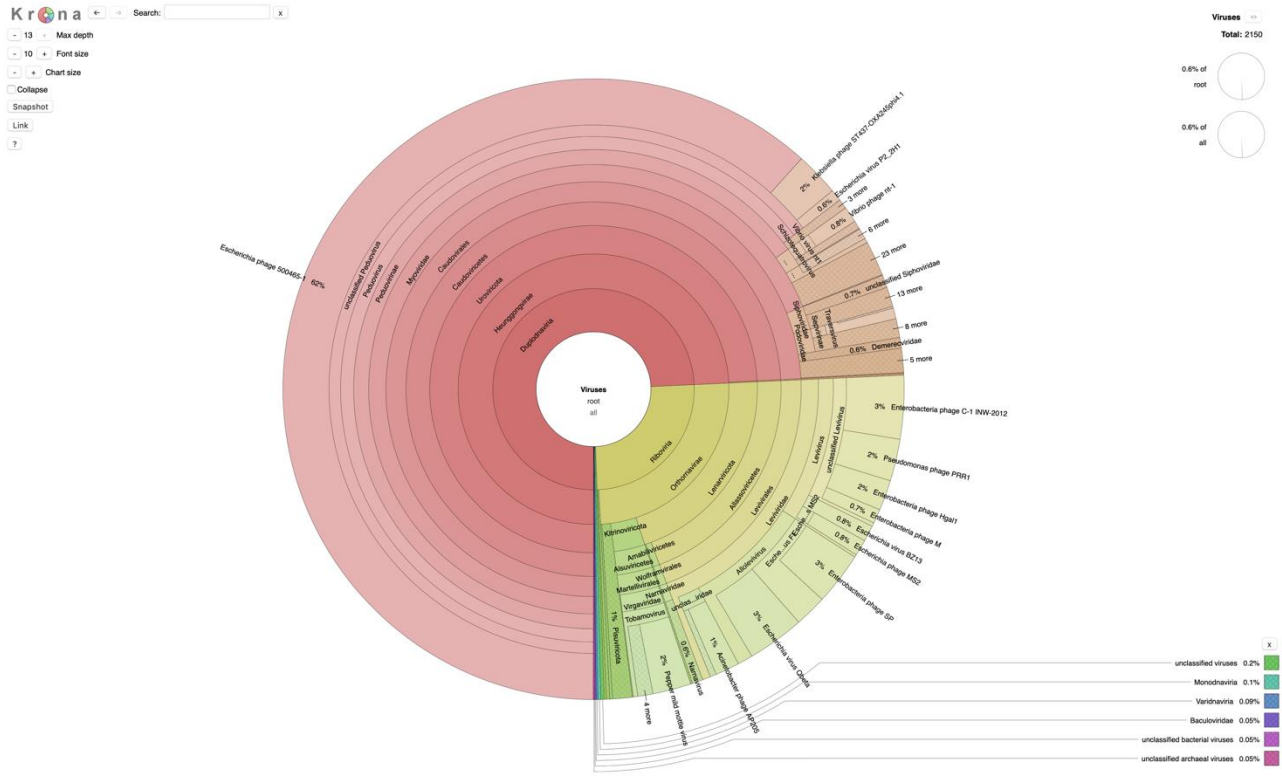
Supplementary: Per Sample Taxonomy Virus RTW11_1A



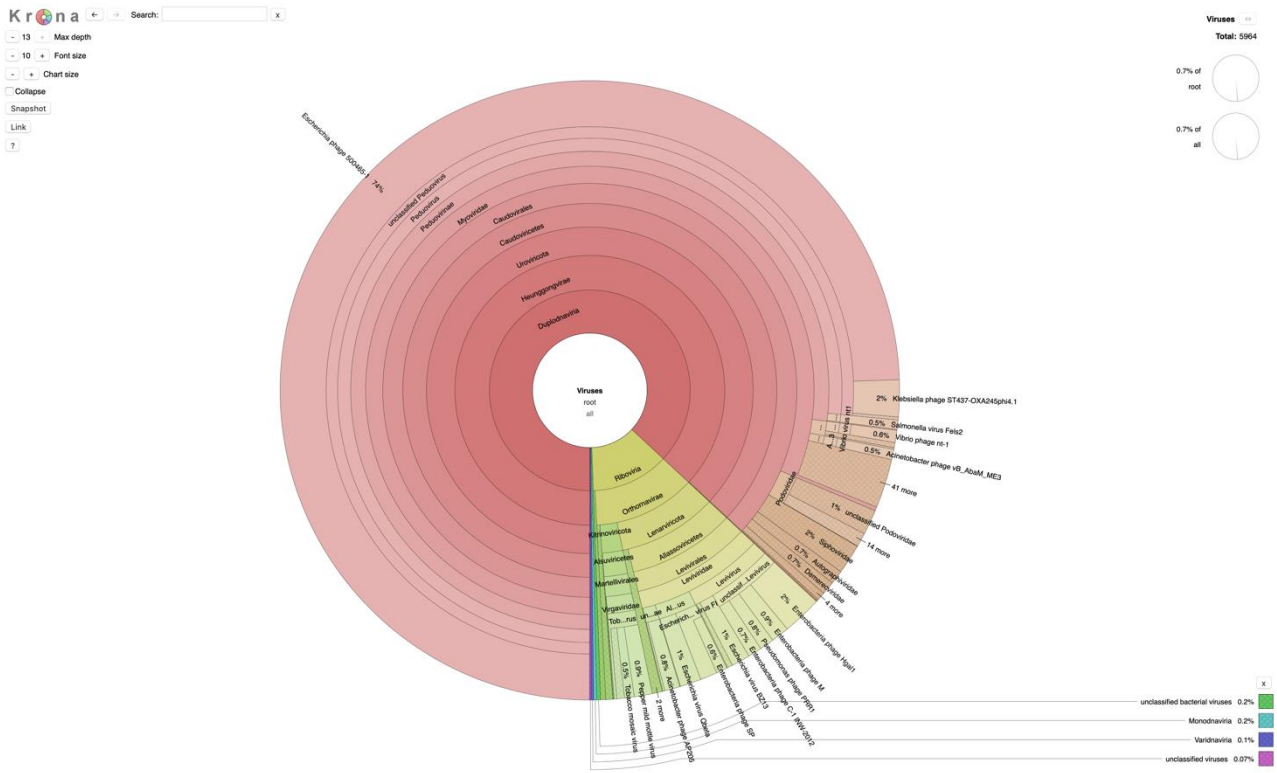
Supplementary: Per Sample Taxonomy Virus RTW12_1A



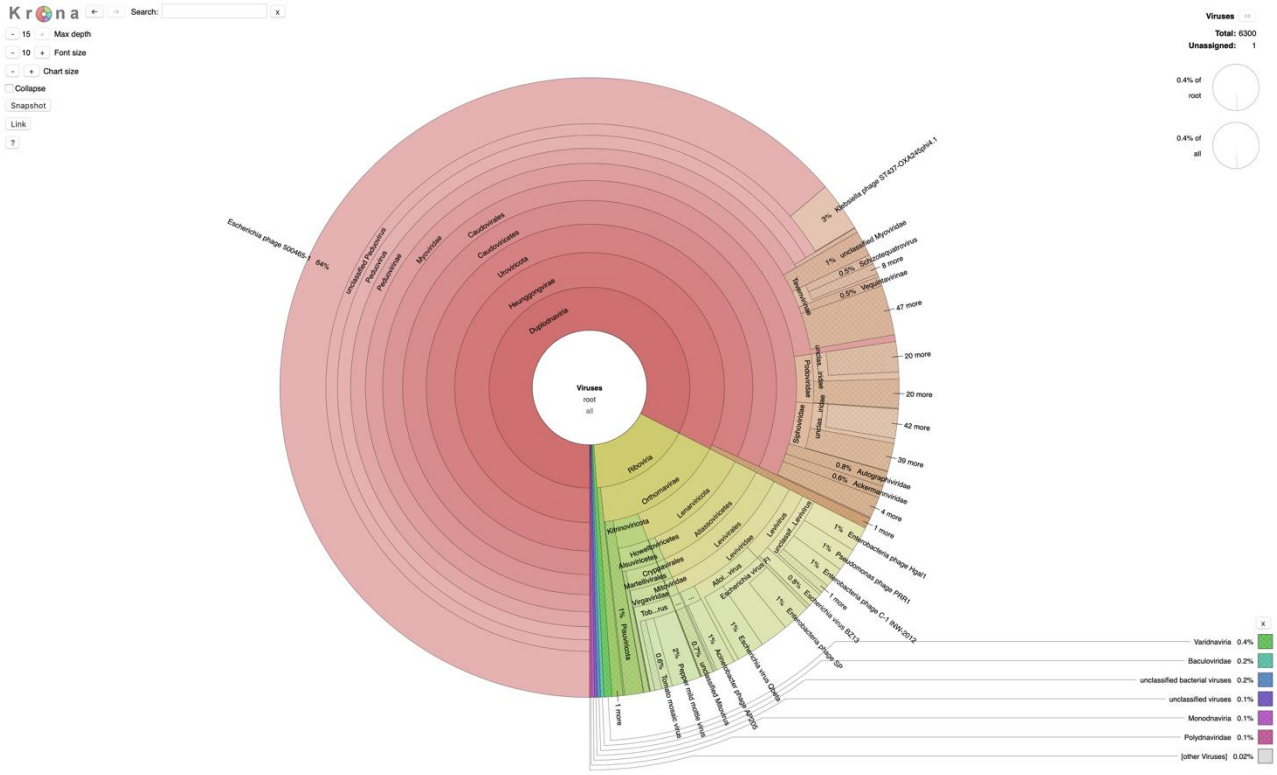
Supplementary: Per Sample Taxonomy Virus RTW13_1A



Supplementary: Per Sample Taxonomy Virus RTW14_1A



Supplementary: Per Sample Taxonomy Virus RTW15_1A



Supplementary: SAJS manuscript

The surveillance of antimicrobial resistance in wastewater from Tshwane district using wastewater-based epidemiology approach

N. G Mbewana-Ntshanka, T.A.M Msagati, A Mutshembele, R.R.E Pierneef, M VanderWalt, T Mooa, P Nkosi, Matodzi, O Lentsoane

College of Science, Engineering and Technology, Institute for Nanotechnology and Water Sustainability, University of South Africa Science Campus Roodepoort, 1709, Johannesburg, South Africa
***Corresponding authors**: msagatam@unisa.ac.za, awelani.mutshembele@mrc.ac.za, PierneefR@arc.agric.za, martiewanderwalt@mrc.ac.za, 67134890@mylife.unisa.ac.za*

Abstract

Background: Antimicrobial resistance (AMR) has become one of the top ten global public health threat. Many countries have recognized the societal and economic burden of AMR. The AMR has reduced the effectiveness of antimicrobial therapies and this results to high mortality, morbidity, and health care expenditure. The propagation of AMR is highly associated with the incorrect antimicrobial regimens as well misuse of antimicrobials. Like all the other developing countries, South Africa falls under the same ambiguous management system of antimicrobials. A lot of research has been focused on the global public health threat "AMR", however, to this day, studies on AMR in wastewater are limited.
Objectives: This paper therefore highlights the imperatives of surveying the AMR pathogens in wastewater (WW) since wastewater (WWTPs) are consecrated as hotspots for the dissemination and propagation of ARB.
Methods: The RNA was extracted from the untreated WW samples that were collected from the Tshwane district in Gauteng province. The metagenomic analysis was proposed for the analysis of the extracted RNA to profile the AMR genes present in the WW.
Results: Based on the filtering criteria, 3 samples (BSW2_1A, RTW1_1A and RTW2_1A) were found to be void of any AMRs. A total of 39 AMR Gene Families and 39 AMR Drug Classes were detected across 17 samples. Certain RTW samples, RTW8_1A and RTW11_1A, dominated in the AMR Gene Family and Drug Class frequencies. Most of the samples showed resistance towards aminoglycoside, carbapenem, cephalosporin, penam, cephamycin, fluoroquinolone, cephalosporin, cephamycin, fluoroquinolone and macrolide antibiotic class. The resistance mechanisms that were mostly detected were antibiotic: efflux, inactivation, target protection, target replacement and reduced permeability to antibiotics.
Conclusion: The metagenomic approach that is discussed in this paper demonstrate the importance of WW surveillance as it can be used as an early detecting system for communicable diseases as well as for monitoring WW from healthcare facilities. By so doing, the new AMR can be identified and monitored at an early stage, then fitting interventions can be employed to mitigate the spread of AMR without using the invasive approaches. Metagenomics of the wastewater pathogens is

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Elucidating the population diversity of microbiota in untreated wastewater in Gauteng

Dineo Raphela

Supervisors: Prof Thulani Makhalanyane & Dr Oliver Bezuidt

Co-supervisor: Dr Rian Pierneef

October 2022

Submitted in partial fulfilment of the requirements for the degree: BScHons(Genetics), Division of Genetics, Department of Biochemistry, Genetics and Microbiology

1

Elucidating the population diversity of microbiota in untreated wastewater in Gauteng

Dineo Raphela, Prof Thulani Makhalanyane, Dr Oliver Bezuidt and Dr Rian Pierneef
Department of Biochemistry, Genetics and Microbiology, Genetic Division, University of Pretoria

Introduction

The analysis of microbial composition in complex environments has become an important part in understanding their functions and interactions in that particular environment. Wastewater environments have been studied for decades but these studies mainly revolved around tracking the circulation of pharmaceuticals, drugs, and pathogens from human sources. The evolution and popularity of metagenomics-based tools have made it possible to study the microbial ecology of wastewater systems without the limitations that come with using culture-dependent methods. Most of the studies done however report on the microbial composition and functions of the environments in mostly westernized regions, these results do not reflect the microbial ecology of wastewater systems in areas of the world where majority of the population is found (Africa, Asia, South America).

To fill this research gap, we proposed a wastewater analysis study with the aim of characterizing the microbial composition of wastewater treatment plants in Gauteng, South Africa.

Results

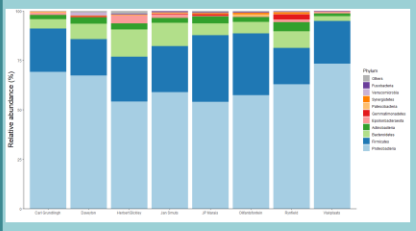
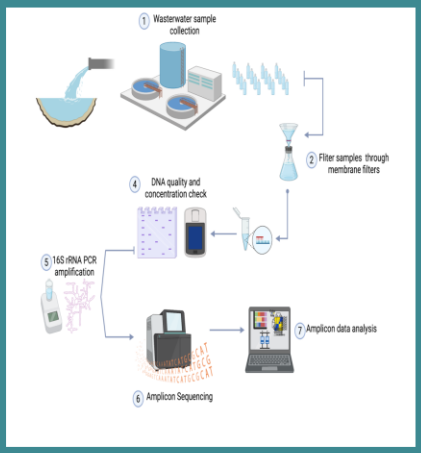


Figure 1: The mean phylum abundance for the top 10 identified phyla in wastewater samples between the different wastewater treatment plants.

- The dominant phylum groups are *Proteobacteria, Firmicutes, Bacteroidetes.* Some groups are more abundant in some WWTPs than others
- These results are similar to those that were reported in other countries.
- The next focus will be alignment mapping using human and environmental databases to identify the sources of these groups.



Conclusion

- The bacterial diversity results are similar to those reported by other studies.
- Future studies need to report on the various variables that affect this diversity.
- The sources of these bacterial groups need to be identified

Acknowledgements

We would like to thank the CSIR for providing the raw sewage samples and 16S rRNA sequence data.

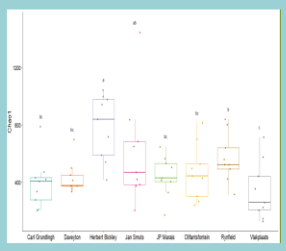


Figure 2: The microbial diversity measure using the Shannon diversity index between the different wastewater treatment plants.

- The overall microbial diversity between the treatments plants is very similar. With the highest diversity observed in Jan-Smuts and Herbert Bickley.
- The high diversity in Jan Smuts could be attributed to the fact that the treatment plant receives influent from the biggest and busiest airport in South Africa (O.R Tambo)