Access to clean water, one of the United Nation's Sustainable Development Goals, is challenged by an increase in the presence of emerging microbial and other contaminants due to urbanization, among other factors. Traditionally, the presence of indicator microorganisms is determined using culturing methods. However, these classical methods cannot be used to determine the identities of ‘unknown’ bacteria and is limited to isolating the culturable state of microorganisms. Thus with culturing, the identities of many bacteria, particularly novel or non-culturable, may remain unknown. The use of a DNA-based method, 16S rRNA next-generation sequencing (NGS), can assist with determining the identities of bacterial populations in a water sample. The objective of this 16S rRNA NGS study was to investigate the bacterial community composition and diversity in a range of water sources. Water samples comprising of potable, surface, ground, marine, aquaculture, rain, wetland and swimming bath water matrices were subjected to 16S rRNA NGS using the Illumina 16S rRNA Metagenomics analysis pipeline. Operational taxonomic units were analysed and the identities of bacterial genera determined. In this study, genera of *Acinetobacter*, *Mycobacterium*, *Pseudomonas*, *Legionella*, *Burkholderia*, *Yersinia*, *Staphylococcus* and *Vibrio* were spread across the water matrices. Alpha (within sample) and beta (between samples) diversities for each bacterial community within the tested samples were also determined.

INTRODUCTION

Approximately 70% of the Earth consists of water. Of this, about 68.7% is not available as it is present in glaciers and ice caps, and about 30.1% is in groundwater, while roughly 0.3% is found in surface water, which includes rivers, swamps and lakes (National Geographic, 2022). Water is the basis of all forms of life and is used for various purposes such as domestic (cooking and cleaning), farming, industry and recreation (Evans, 2019). Water is also an important habitat for bacteria and serves as a means for their dissemination in nature (Vaz-Moreira et al., 2014). Some bacteria have been found to evade disinfection in the drinking water treatment process, either through resistance or via biofilm protection (Simoes and Simões, 2013; Li et al., 2017). As a result, access to clean water has been adopted as one of the United Nation's Sustainable Development Goals (United Nations, 2023). In 2017, the World Health Organisation (WHO) published their *Guidelines for Drinking Water Quality* as a tool to ensure that drinking water of good quality is maintained. Currently in South Africa, drinking water is routinely monitored to ensure that it complies with the South African National Standard for drinking water (SANS 241:2015, SABS, 2015). However, exposure to waterborne bacteria is not limited to potable water.

The human gastrointestinal microbiome, shed in faeces, consists of a mixed population of bacteria, some of which are commensals; others exhibit a mutualistic relationship, while some may cause harm in the host (Almeida et al., 2019). Worldwide, inadequate human hygiene practices, open defecation, lack of access to clean water, and inadequately treated water have notably contributed to various opportunistic infections (Garr, 2001; WHO, 2019). Several anthropogenic activities as well as human and animal waste can contaminate surface (rivers and streams), coastal and groundwater sources (Pål et al., 2013; Itchon and Bruni, 2020). In South Africa, wastewater treatment plants treat sewage according to licence agreements instituted by the national Department of Water and Sanitation (DWS) to reduce the pollution risk of effluents to the environment. Growing populations, loadshedding or rolling blackouts, sewer line failures and blockages, sewer pump station overflows and breakages as well as increases in emerging contaminants could have an indirect effect on the efficacy of wastewater treatment (City of Cape Town, 2020).

Traditional culture-based techniques determine the presence and number of microorganisms such as bacteria in a tested sample. However, the method is limited to culturable microorganisms (Verhille, 2013), is laborious and less sensitive than molecular techniques such as the quantitative polymerase chain reaction (qPCR). Despite being a sensitive technique, qPCR can only detect and quantify the target microorganisms such as bacteria that the reaction has been designed for. Thus, unknown/non-suspected bacteria may remain undetected. Urban waters have a diverse array of microorganisms, many of which are still unknown (Jin et al., 2018), thus identifying them using specific approaches such as culturing or qPCR is impractical. This could be overcome with the use of screening molecular techniques such as next-generation sequencing (NGS). 16S rRNA NGS has transformed the field...
of bacterial ecology (Bautista-de los Santos et al., 2016; Brandt et al., 2018; Zhang and Liu, 2019). The composition of a bacterial population can be determined using the unique 16S rRNA-encoding DNA sequences in samples and through comparing it to all known 16S rRNA sequences in a database to determine the presence, abundance and taxonomy level of the bacteria present. Alpha (within sample) and beta (between samples) diversity can be calculated for each bacterial community giving an indication of the diversity and abundance within the bacterial community (Bautista-de los Santos et al., 2016).

This study aimed to use 16S rRNA NGS to investigate the bacterial composition in an array of filtered water samples such as potable, surface, ground, marine, aquaculture, rain, wetland and swimming baths. This study also determined the alpha (within sample) and beta (between samples) diversity for each bacterial community within the tested samples.

**METHODS**

**Sample collection**

Water samples were mainly collected from the Cape Metropole area and consisted of municipal samples as well as citizen science samples. Municipal potable water samples were collected in disposable 250mL PET sterile (gamma radiated) bottles with 3% sodium thiosulphate; while swimming bath sampling was collected in pre-autoclaved bottles, transported on ice and immediately filtered upon arrival at the laboratory. Citizen science samples were collected in a sterile manner at the water source using sterile disposable gloves. The stipulated instructions for water filtration and storage were supplied with the water collection kit. Each water sample (three 60 mL aliquots) was aseptically filtered using a syringe-filtering unit (cellulose acetate membranes with 0.2 µm pores; GVS Filter Technologies, UK) and touching of the syringe and filter attachment points was avoided. All processed filters were placed in sterile plastic zip-lock bags and stored at −20°C until delivered to CPGR. Frozen filters were transported on ice to CPGR and stored at −20°C prior to DNA extraction. Water samples were collected (50 from the municipality and 50 from a variety of water sources supplied by members of the public; see Supplementary Material I) between 2019/10/28 and 2019/11/26.

In tandem, the municipal samples were subjected to standard microbial analyses and assessed according to the required guidelines (WHO, 2006; DoH, 2015; SANS, 2015).

**DNA extraction and QC**

Filters were sterilely removed and DNA extractions carried out using the ZymoBIOMICS 96 MagRead DNA Kit (Zymo Research, USA). As a negative control, DNA was extracted from a sterile filter. DNA quantification was carried out on the GloMax Multi Detection System (Promega, USA) using the fluorescent-based QuantiFlor dsDNA System (Promega, USA).

**Sequencing**

In addition to all extracted DNA samples, the ZymoBIOMICS Microbial Community DNA Standard (Zymo Research, USA) positive control and a no-template control were included in library preparation and sequencing. The 16S V3-V4 rRNA amplicons were produced according to the Illumina 16S Metagenomics Sequencing Library Preparation guide using KAPA HiFi HotStart ReadyMix (Roche, USA) and the Nextera XT Index Kit v2 (Illumina, USA). Libraries were normalised to 5 ng/µL or left at their original concentration if lower than 5 ng/µL. Up to 10 ng of each sequencing library was pooled. The concentration of the final library pool was confirmed by qPCR using the KAPA Illumina Library Quantification Kit (Roche, USA) and diluted to a concentration of 4nM, denatured, diluted to 5 pM and combined with the 10% PhiX control (Illumina, USA). Sequencing was performed on an Illumina MiSeq (Illumina, USA) using an Illumina MiSeq Reagent Kit v2 (500 cycles).

**Data analysis**

The Illumina 16S rRNA Metagenomics analysis pipeline classified sequencing reads using an illumina-curated version of the Greengenes database (DeSantis et al., 2006) and run on all samples using the MiSeq Reporter software version 2.6.2.3 (illumina, USA). The resulting summary text files for each sample contained operational taxonomic unit (OTU) classifications, the related number and percentage of OTU reads. OTUs that were not assigned to a taxonomic classification were labelled ‘Unclassified’, along with the related number and percentage of reads. Programming scripts were used to (i) filter OTUs with less than 1% of the total number of reads; (ii) sum and list OTUs to genus level; while (iii) remaining OTUs with a classification higher than genus were labelled ‘Other, level > genus’; (iv) identify bacterial genera based on WHO, 2006; USEPA, 2012 and WHO, 2017 guidelines; and (v) calculate the Shannon Diversity Index (H) as follows: \[ H = -\sum_{i=1}^{P} P_i \ln P_i \] where \( P_i \) = proportion of the sample represented by Genus \( i \) (Spellerberg and Fedor, 2003). Further data analysis and graphical representation was carried out using Excel, the on-line Venn diagram tool (Ghent University, 2023) or JAMOVI version 1.6.13 (Jamovi, 2023). After testing for normality using the Shapiro Wilk test, Welch’s one-way ANOVA with a Games-Howell post-hoc test was used to determine statistically significant differences between Shannon diversities of the different water matrices.

FASTQ files were uploaded to the CyVerse Data Store (Cyverse, 2023) and analysed using the DNA Subway (Cold Spring Harbor Laboratory, 2023) Metabarcoding pipeline with a simplified version of the QIIME 2 pipeline (Bolyen et al., 2019). The Greengenes database was used for classification. Beta diversity was analysed by principal component analysis (PCA) of the unweighted UniFrac distance which measures the community dissimilarity and incorporates phylogenetic relationships between the features (Lozupone and Knight, 2005).

Sequencing reads were deposited on the NCBI Sequence Read Archive with Project Number PRJNA775828.

**RESULTS AND DISCUSSION**

**Samples, DNA extraction and library preparation**

A mixture of filtered water samples were received from the municipality and from members of the public, which included samples from an abalone farm, an aquaponics system, boreholes, bottled water, household taps, freshwater and marine tanks, harbour and sea samples, rainwater tanks, rivers, springs, an estuary, a vlei and wellpoints. Municipal potable water and swimming bath samples were subjected to standard microbial analyses. All municipal samples tested were compliant (Supplementary Material I), except for two desalination plants which exceeded the SANS241:2015 limits (Supplementary Material I). Both swimming bath samples complied with the national and WHO operational guidelines for microbial testing (Supplementary Material I) (WHO, 2006; DoH, 2015).

DNA concentrations from the extracted water samples ranged from <1 ng/µL to 37 ng/µL (Supplementary Material I). To test the utility of the DNA across this concentration range, 16S rRNA sequencing libraries were produced from all samples and quantified. The concentration of the sterile filter negative control library was 0.317 ng/µL (Supplementary Material I). Thirteen filtered water samples had library DNA concentrations <0.317 ng/µL.
and were therefore excluded from further analysis. Thus only 87 of the 100 samples were subjected for further analyses in the study. These water samples were divided into 8 categories: potable water, surface water (from dams and rivers), groundwater (from springs, wellpoints and boreholes), marine water (from sea, harbour and marine water tanks), aquaculture samples (from abalone and aquaponics systems), rainwater tanks, wetland samples (vlei and estuary) and swimming baths (Fig. 1, Supplementary Material I).

Sequencing and data analysis

The MiSeq sequencing run resulted in 7.72 Gb non-indexed sequence data. The Q-score distribution plot generated a Q30 score of 76.0% for the sequencing run, indicating that the quality of the sequencing data was acceptable.

The sequencing output was run through Illumina’s 16S rRNA metagenomics pipeline. The number of sequencing reads for the detected genera based on WHO (2006), EPA (2012) and WHO (2017) were identified (Supplementary Material I). Supplementary Material II lists the genera identified in each water sample. A very low number of reads was obtained for the no-template and sterile filter negative controls (885 and 7 540 filtered reads, respectively). The percentage of filtered reads classified at the genus level with high confidence for the ZymoBIOMICS positive control are shown in Fig. 2. The ZymoBIOMICS Microbial Community DNA Standard consists of 8 bacterial genera and all 8 genera were detected, ranging from 3.9% of filtered total reads (for Salmonella spp.) to 16.2% of filtered total reads (for Bacillus spp.).

Bacterial diversity of varied water sources

The Shannon Diversity Index was used to calculate an overall score of diversity for each sample at the genus level (Supplementary Material I). To compare diversity ranges within the eight water categories, the mean Shannon Diversity Index was calculated (Fig. 3). The potable water category had the lowest average Shannon diversity. This was significantly lower than rain and surface water, whereas the rain tank category (reflecting samples harvested from rooves and stored prior to purification) had the highest Shannon diversity (Fig. 3).

Beta diversity of the sample types was determined using an unweighted UniFrac PCA. This measured bacterial community differences between the tested samples (Lozupone and Knight, 2005). The majority of the potable samples, together with the swimming bath samples, broadly separated from the other sample types based on the first component contributing to the variation (13.8% of the variation), whereas the rain tank and the majority of the ground and surface water samples grouped further from the other sample types based on the second component (5.338% of the variation) (Fig. 4).
Within the surface water category, 6 river samples were collected within a few days of each other from clearly defined river crossings along the Otter Trail, a hiking trail along the Garden Route coast of South Africa. This dataset offered an opportunity to study the diversity of different genera in nearby rivers over a very short time. Comparison of the genera identified for each sample (Supplementary Material II, Fig. 5) indicated that there was very little overlap between the different samples. This suggests that the composition of the bacterial populations in these river samples may be influenced by distance to the river mouth, localised geology and upstream farming practices (Doherty et al., 2017; Chen et al., 2018).

Identification of bacterial genera

Filtered sequencing data were further analysed to determine the genera based on WHO (2006), USEPA (2012) and WHO (2017) guidelines, in each water category (Table 1). Sequencing reads corresponding to 8 genera (Acinetobacter, Mycobacterium, Pseudomonas, Legionella, Burkholderia, Yersinia, Staphylococcus and Vibrio) were detected in at least one matrix (Table 1; WHO, 2017). In addition, sequencing reads corresponding to the genera Streptococcus, Clostridium, Microcystis, and Arcobacter were detected in more than one water matrix (Table 1; WHO, 2006; USEPA, 2012).

Currently, the Acinetobacter genus consists of 74 different species (LPSN, 2022a; Parte et al., 2020), is considered to be ubiquitous in nature, and its presence in sewage, soil, water, food and clinical environments is noted (Towner, 2006; WHO, 2017). Acinetobacter spp. have been linked to enhanced biological phosphorus removal from wastewater (Wagner et al., 1994; Yang et al., 2015) and have bioremediation potential (Paisio et al., 2016; Al-Hadithi et al., 2017) suggesting a further possibility for their presence in waterbodies. To date, Acinetobacter spp. have been found in potable water (Carvalheira et al., 2021), rivers (Kittinger et al., 2018), wastewater (Yang et al., 2015), groundwater (Pindi et al., 2013) and seawater (Yoon et al., 2007). In this study, Acinetobacter was detected in 4 potable water samples (2 tap and 2 desalination water samples) and 1 surface (river) water sample (Table 1; Supplementary Material I).
The Mycobacterium genus consists of 3 groups: (1) obligate pathogens, (2) potential pathogens and (3) saprophytes (Vaerewijck et al., 2005). Group 1, the *M. tuberculosis* complex (MTC), causes disease in humans and animals and is generally not found in the environment. On the other hand, Group 2 (potential pathogens) consists of the *Mycobacterium avium* complex (MAC) (Hunter, 1997) and is found in terrestrial and aquatic environments. In comparison, Group 3 (saprophytes) are rarely pathogenic (Vaerewijck et al., 2005). Groups 2 and 3 are also known as non-tuberculosis/atypical *Mycobacterium* (NTM), are naturally occurring in a range of water environments, and are able to grow and persist if conditions are suitable (Le Dantec et al., 2002; WHO, 2017). To date, there are 195 validly published species of *Mycobacterium* (LPSN, 2022b). In our study, the genus *Mycobacterium* was identified in 1 sample in each of the following categories: aquaculture, marine, surface, wetland and swimming bath; as well as in 5 potable water samples (Table 1; Supplementary Material I). These findings are in agreement with international studies where *Mycobacterium* was found in aquaculture (Yanong and Pouder, 2010; Gcebe et al., 2018; Mugetti et al., 2020; Mataragka et al., 2022), marine environments (Hashish et al., 2018; Mataragka et al., 2022), surface water (Eun-Sook et al., 2008), wetlands (Kopecky et al., 2011; Tortone et al., 2018), swimming baths (Torvinen et al., 1999; D’Ancona et al., 2014) and extensively found in potable water sources worldwide (Bailey et al., 1970; Von Reyn et al., 1993; Schwartz et al., 1998; Vaerewijck et al., 2005; Tortone et al., 2018; Lin et al., 2020). Furthermore, chlorine disinfection has been found to be ineffective for...
Mycobacterium in potable water samples, with the presence of biofilms adding an extra mechanism for its persistence in water (Simoes and Simões, 2013; Li et al., 2017).

Pseudomonas are free-living micro-organisms and ubiquitous in nature. The genus is found in moist environments, rivers and soils, therapy pools and surrounding areas, and is often associated with biofilm production (Wingender and Flemming, 2011; Koek et al., 2018). To date, there are 310 validly published species of Pseudomonas (LPSN, 2022c). In our study, the genus Pseudomonas was detected in 3 potable water samples, 1 swimming bath and 5 surface water samples, 4 of which were found along the Otter Trail (Table 1; Supplementary Material I).

Of interest, 20 samples (23%) in this study contained Legionella. Twelve of these were potable water samples (2 from water treatment plants and 10 from distribution points or taps) (Table 1; Supplementary Material I). This is possible as Legionella can grow in water, pipe biofilms and soil. High levels of biodegradable organic carbon, warmer temperatures, stagnant water, low residual chlorine concentrations and biofilms may further exacerbate their growth in water (WHO, 2017). Other water categories where Legionella were detected in our study included: 2 samples from an abalone farm and 1 sample each from a dam and river (surface water), spring, wetlands, marine, and rainwater tank (Table 1; Supplementary Material I). Legionella is an environmental microorganism and may proliferate in suitable conditions (WHO, 2017). Currently, there are more than 58 different Legionella spp. with more than 70 serotypes, (Ditommaso et al, 2021; Legionella Control International Ltd., 2021).

Burkholderia was detected in 3 river samples; 2 of which were along the Otter trail (Table 1; Supplementary Material I). Burkholderia was co-detected with Yersinia in 1 river sample. The principle reservoirs of Yersinia spp. are domestic and wild animals but they can also be found in sewage and polluted waters (WHO, 2017). In 2015, a study by Prentice indicated that a species of Yersinia could enter the viable but non-culturable state in water and could also persist in soil, thus emphasizing the importance for molecular monitoring over culturing for this genus. The presence of Burkholderia spp. has been documented in soil and muddy water in tropical regions (WHO, 2017), coastal environments and mussels (Maravic et al., 2012), a water treatment plant (Inglis et al., 2000) and also maintained its culturable state after several years’ incubation in distilled water (Wuthiekanun et al., 1995).

In this study, Staphylococcus was detected in 2 swimming bath samples (Table 1; Supplementary Material I). Staphylococci are ubiquitous in the environment and commonly found in mucous membranes and on human and animal skins (Leroy et al., 2016). As a result of its presence on the skin, it can be transferred to swimming baths, recreational water and spa pools (Koek et al., 2018). Although this microorganism has been detected in drinking water, there is no evidence that it can be transmitted through water consumption (WHO, 2017). Staphylococci are occasionally found in the gastrointestinal tracts of humans; thus its presence in sewage is also possible (WHO, 2017).

There are currently 138 validly published species of Vibrio (LPSN, 2022d). These species are abundant in aquatic environments as well as in the organs and/or tissues of marine algae and animals, including corals, sponges, bivalves, abalone, shrimp, fish, squid, and zooplankton (Nicolas et al., 2002; Cai et al., 2006; Pichon et al., 2013; Economopoulou et al., 2017), which may negatively impact the infected species and may also have a zoonotic potential (Economopoulou et al., 2017). In our study, the Vibrio genus was detected in 2 aquaculture samples where abalone was farmed.

Although Escherichia spp. sequencing reads were not reported in various water samples of this study (Table 1), it was reported in

the ZymoBIOMICS control (Fig. 2). Escherichia spp. and Shigella spp. are closely related and belong to the Enterobacteriaceae family (Devanga Ragupathi et al., 2017). Of the municipal samples collected, one treated desalination plant sample (ASV041) contained 1 CFU/100 mL E. coli (Supplementary Material I). Additionally, poor correlation between 16S microbiome sequencing data and culture-based methods may reflect the limits of detection using 16S sequencing, as low numbers of sequencing reads associated with low concentrations of micro-organisms in communities may be removed during downstream data analysis steps; alternatively deeper sequencing may be needed to detect micro-organisms in very low abundance (Hamady and Knight, 2009; Schang et al. 2016). Intestinal enterococci are also used to monitor water quality (WHO, 2017). They consist of 2 genera, Enterococcus and Streptococcus, and are used as indicators of faecal pollution. Intestinal enterococci are highly prevalent in sewage and sewage spills as well as human and animal waste. They survive for longer periods in water compared to E. coli, and are resistant to chlorination and dry conditions (WHO, 2017). Limayem et. al. (2019) found Streptococcus to be the second-most predominant genus in influent and treated effluent samples collected at reclamation facilities in the USA. Streptococcus were also found in the Nile River (Goja, 2013) as well as in a groundwater sample in South Africa (Ateba and Molaoa, 2013). In our study, Enterococcus was not detected but Streptococcus was identified in 3 river samples and 1 groundwater sample (Supplementary Material I).

Clostridium spp. are ubiquitous in nature and often found in the gastro-intestinal tracts of humans and animals. These species can enter surface water systems via runoff in agricultural settings or through treated wastewater effluents (Fourie, 2017). The genus Clostridium was detected in 4 samples in this study, consisting of 2 surface (river), 1 groundwater and 1 potable water sample (Supplementary Material I, II); these surface and groundwater samples also contained Streptococcus spp. The two surface water samples (ASV005 and ASV006) were submitted with a note of ‘may contain sewage’, suggesting that the presence of both Streptococcus spp. and Clostridium spp. may be linked to faecal contamination, while an explanation for the presence of both genera in a single groundwater sample (ASV016) in our study is unclear.

Cyanobacteria are found in a range of environments, including hot springs, soil, artificially illuminated cave systems, ice layers in polar regions as well as in marine, brackish and freshwater systems (Harding and Paxton, 2001; WHO, 2017; USEPA, 2020). These aquatic organisms produce pigments and generally occur in low numbers in most surface waters; however, eutrophication can result in bloom development. Microcystis is a genus of freshwater cyanobacteria and is capable of causing harmful/toxic as well as non-toxic blooms in ecosystems globally (Park et al., 2018; Pérez-Carrascal et al., 2019). In our study, Microcystis was detected in 1 wetland, 1 potable water (citizen science sample) and 1 surface water (raw sample; water treatment plant) sample (Supplementary Material I, II). All samples were colourless and odourless, suggesting low growth of Microcystis. No sequencing reads corresponding to other cyanotoxin bacteria were detected.

Arcobacter, a genus of emerging concern, has been associated with 3 waterborne outbreaks and its persistence in wastewater is suggestive of an ecological reservoir (Banting and Figueras Salvat, 2017). In our study, Arcobacter was detected in 2 potable, 3 surface, 2 marine, 1 ground and 2 aquaculture water samples, all of which were citizen science samples. This finding is not new as Arcobacter spp. have been identified in several water sources worldwide. These include: tap and river water in Asia (Marita et al., 2004) and Europe (Moreno et al., 2003), seawater in Spain (Collado et al., 2008), groundwater in South Africa and Turkey (Diergaardt et al., 2004; Ertas et al., 2010), and it has also been isolated from abalone in Japan (Mizutani et al., 2019).
Importantly, the viability status of the microorganisms detected in our study was not determined and was not part of the scope of the current study. Rather, it was used to determine the bacteriological composition and diversity within and among the tested water samples.

CONCLUSION

This 16S rRNA amplicon microbiome sequencing study identified 16 genera which included Acinetobacter, Mycobacterium, Pseudomonas, Legionella, Burkholderia, Yersinia, Staphylococcus, Vibrio, Streptococcus, Clostridium, Microcystis and Arcobacter. In addition, this study uncovered a wide range of bacterial diversity at the genus level, both within and between the different water types, with municipal treated potable samples and samples from rain tanks displaying the lowest and highest within-sample type diversity, respectively. Of note, the limit of detection of 16S NGS compared to culture techniques is also an important factor to consider as the indicator micro-organism, *E. coli*, could be present in lower concentrations compared to the more abundant genera. Thus *E. coli* may not be detected with NGS but could be cultured using selective media or detected using a rapid technology like IDEXX Colilert.

ACKNOWLEDGEMENTS

Gratitude is extended to CPGR for funding and to members of the public for providing water samples to incorporate a citizen science component in the study. Additionally, municipal staff are thanked for sampling, sample filtration and microbial analyses of the municipal samples.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES


BABINTZ 02109-9


