MSc thesis in Civil Engineering

Characterizing Drinking Water Microbiome Using Oxford Nanopore MinION™ Sequencer

Xinyue Xiong

2019

TU Delft

oasen drinkwater

Characterizing Drinking Water Microbiome Using Oxford Nanopore MinION™ Sequencer

By

Xinyue Xiong

in partial fulfilment of the requirements for the degree of

Master of Science
in Civil Engineering

at the Delft University of Technology,
to be defended publicly on Thursday October 10, 2019 at 15:30 PM.

Thesis committee:	Prof. dr. Gertjan Medema,	TU Delft
	Dr. ir. Gang Liu,	TU Delft
	Dr. Thom Bogaard,	TU Delft

TU Delft

oasen drinkwater

Preface

This thesis is the final report of my MSc program in Civil Engineering at Delft University of Technology. This ten-month MSc thesis project is jointly supported by TU Delft and drinking water company Oasen. I would like to express my sincere appreciation to these two institutions for providing me with such a precious opportunity to conduct this scientific research.

Upon the completion of this thesis, I am grateful to everyone who has contributed to this study. Firstly, I would like to express my heartiest thankfulness to my daily supervisor, Lihua Chen, for her patient guidance and assistance in both experiments and academic writing. Moreover, profound gratitude should go to Prof. Gertjan Medema, for his professional instructions and supports in this whole research. I would like to express my gratitude to Dr. Gang Liu, for offering me this opportunity to conduct this project. I would like to express my thankfulness to Dr. Thom Bogaard, for his valuable comments and insightful suggestions on this thesis. Special thanks should also go to Dr. Yujia Zhai and her colleagues in Leiden University, who helped me with part of essential experiments in this research.

Also, I would like to extend my deep gratefulness to my family, especially my parents. They make every effort to support me with not only money but also love. I also owe my heartfelt gratitude to my boyfriend, Qiangqiang Li, and my best friends Shuxin Luo and Yajiao Liu, who encouraged me when I felt depressed.

Most importantly, I would like to thank my great motherland, China. Two years ago, I came to TU Delft for learning the cutting-edge technologies to solve environmental issues in China. She motivated me to carry on every time when I felt like giving up. Ultimately, I managed to finish my thesis by the 70th anniversary of the founding of the People's Republic of China. Happy birthday, my dear China. 此生无悔入华夏，来世愿在种花家。

Finally, thanks to you, who is reading this thesis. If you are reading this line after others, you at least read one page of my thesis. I hope you enjoy your reading.

Xinyue Xiong
Delft, October 1, 2019

Abstract

Biological safety of drinking water is vital for safeguarding public health. Many efforts have been made to explore the microbial universe in drinking water. Nanopore sequencing developed by Oxford Nanopore Technologies is expected to enable PCR-free and rapid identification of species with high accuracy, thus overcoming the impediments of next-generation sequencing. However, the capability of Nanopore sequencing for characterizing the microbiome in drinking water with extremely low biomass content has not been explicitly evaluated. Therefore, this research was carried out to explore the potential of Nanopore sequencing for microbial community characterization and species identification in drinking water. In this study, NanoAmpli-Seq full-length 16S rRNA sequencing and $1D^2$ genomic DNA (gDNA) sequencing were performed on an Oxford Nanopore MinION^{TM} sequencer. DNA samples of artificial microbial communities were sequenced in order to assess the performance of both sequencing strategies. Subsequently, DNA extracted from tap water was subjected to Nanopore sequencing with the two methods. Results showed that NanoAmpli-Seq 16S rRNA sequencing precisely identified abundant species in artificial microbial communities with high level of reproducibility but biased community profiles due to variation in PCR efficiencies of different species, whereas only 10 species were identified in tap water samples. In addition, raw results from $1D^2$ gDNA sequencing provided an unbiased microbial community profile of an artificial community DNA, while polished data improved the species identification accuracy at the expense of the ability to profile the community structure. Furthermore, 45 hours' sequencing generated more reliable results than 5 hours' sequencing with higher profiling accuracy of community structure. Nevertheless, $1D^2$ gDNA sequencing still did not exhibit desirable species identification performance on tap water DNA samples. Notably, despite two enteropathogenic species (*Enterobacter cloacae* and *Laribacter hongkongensis*) were identified, the detection of *Homo sapiens* in the same sample indicated the potential existence of post sample contamination. To conclude, Nanopore sequencing possesses great potential to serve as an efficient tool for study of drinking water microbiology. Specifically, notwithstanding the dissatisfactory performance of NanoAmpli-Seq, its high reproducibility across sequencing runs, adaptability to low DNA quality and quantity, and short turnaround time indicated its potential usefulness to promptly monitor microbial community changes subjected to environmental changes in extremely low-biomass samples (i.e. drinking water). Despite that $1D^2$ gDNA sequencing exhibited superior performance on species identification and microbial community profiling to NanoAmpli-Seq, more endeavors should be made to overcome the hurdles (e.g. demand for high molecular weight gDNA, standard methods for analyzing sequencing data), thereby improving the species identification coverage and microbial community profiling accuracy in drinking water. Understanding the presence and dynamics of the microbial community in DWDS is important for water utilities to gain a better understanding of various microbial processes in drinking water from source to customers' taps, based on which water treatment strategies could be improved and better management of drinking water quality could be performed.

Contents

List of figures

List of tables

List of abbreviations

16S rRNA	16 Subunit ribosomal RNA
1D	One directional
$1D^2$	One directional squared
AMR	Antimicrobial resistance
ARB	Antibiotic resistant bacteria
ARG	Antibiotic resistance gene
AT-rich	Adenine-thymine-rich
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
dsDNA	Double-stranded DNA
DWDS	Drinking water distribution system
Gbp	Giga base pairs
GC content	Guanine-cytosine content
gDNA	Genomic DNA
HMW	High molecular weight
HPC	Heterotrophic plate count
INC-Seq	Intramolecular-ligated Nanopore Consensus Sequencing
kbp	Kilo base pairs
Mbp	Mega base pairs
NDM-1	New Delhi metallo-beta-lactamase 1
NGS	Next-generation sequencing
ONT	Oxford Nanopore Technologies
OPs	Opportunistic pathogens

OPSS	Online Particle Sampling System
OTU	Operational taxonomic unit
PCR	Polymerase Chain Reaction
PES	Polyether sulfone
QC	Quality control
RCA	Rolling circle amplification
SMRT	Single-molecule real-time
SSU rRNA	Small Subunit ribosomal RNA
WIMP	What's In My Pot

1 Introduction

1.1 Drinking water microbiology

Safe and regulation-compliant drinking water can still host an incredible biodiversity of microbes (Bruno et al., 2018). Generally, there would be 10^6-10^9 microbes residing in a liter of drinking water (Hammes et al., 2008; Hull et al., 2019; Lautenschlager et al., 2010; G. Liu et al., 2013b; G. Liu et al., 2018; Prest et al., 2014). Treated water produced by drinking water treatment plant already carries a physical load (particles), a nutrient load (nutrients and biomass), and a microbial load (living microbial cells) before it enters drinking water distribution networks (G. Liu et al., 2013a). As a consequence, a series of physicochemical and biological processes, including growth of planktonic microbes in bulk water, formation and detachment of pipe wall biofilm, and formation and resuspension of loose deposits, may take place throughout drinking water distribution system (DWDS) before drinking water reaches customers' taps (G. Liu et al., 2013c). Therefore, it is not unusual that deterioration in drinking water quality occurs during distribution (G. Liu et al., 2017a; G. Liu et al., 2017b; G. Liu et al., 2018). That is to say, tap water would generally contain more and different microbes than finished water (G. Liu et al., 2014; G. Liu et al., 2013a; Proctor & Hammes, 2015). In the Netherlands, efforts have been made to produce biologically stable treated water for ensuring microbial safety of tap water (Smeets et al., 2009). Nevertheless, there is still potential for bacterial regrowth and microbial contamination due to failure in DWDS and premise plumbing. As tap water comes in direct contact with consumers, maintaining biological safety of not only finished water but also tap water is crucial for ensuring customers' access to biologically safe drinking water via their taps.

Although the majority of microorganisms residing in drinking water are harmless to humans, opportunistic pathogens (OPs) can establish in DWDS, grow as part of drinking water microbiota, and eventually end up in tap water (Hong et al., 2017; Wang et al., 2017). Moreover, stagnation in premise plumbing will further promote microbial growth and thus increase the risk of pathogen proliferation (Bédard et al., 2018; Ling et al., 2018). Being the primary cause of drinking water-related disease outbreaks, the occurrence of OPs in drinking water poses risks to human health and is of growing concern (Craun et al., 2010; Falkinham III et al., 2015; Hong et al., 2017). For instance, it was widely reported that the waterborne pathogenic bacteria *Legionella pneumophila*, *Mycobacterium avium*, and *Pseudomonas aeruginosa* were frequently detected in biofilms in premise plumbing, which pressed public health problems (Falkinham et al., 2015; Hong et al., 2017; Wang et al., 2017). A wide variety of fatal illness could be related to OPs in drinking water, such as Legionnaires' disease caused by *Legionella spp.* and primary amebic meningoencephalitis resulting from *Naegleria fowleri* (Bartrand et al., 2014; Cope et al., 2015; Falkinham, 2015; Wang et al., 2017). Additionally, the emerging antimicrobial resistance (AMR) has aroused great public concern. Moreover, the growing prevalence of antibiotic resistance genes (ARGs) in drinking water microbiome was highlighted by numerous studies (Armstrong et al., 1981; Khan et al., 2016; Schwartz et al.,

2003; Shi et al., 2013; Su et al., 2018; Xi et al., 2009; Xu et al., 2016). As was indicated by Bai et al. (2015), Jia et al. (2015) , Su et al. (2018), and Xi et al. (2009), despite some of the drinking water treatment techniques can effectively eliminate ARGs, certain drinking water treatment processes may still pose selection pressure for ARGs and thus contribute to the enrichment of antibiotic resistant bacteria (ARB) in finished water. Moreover, drinking water distribution systems can act as a reservoir for spread of ARGs and ARB (Zhang et al., 2019). Once the waterborne diseases-related bacteria harbored ARGs, the difficulty in combating those pathogens would dramatically increase. As a consequence, the related potential health risks would also be drastically increased. One of the best-known instance of AMR related water safety issue is the emergence of multidrug resistant superbug NDM-1 in drinking water in New Delhi, India in 2011, which posed grave threat to public health worldwide (A. P. Johnson & Woodford, 2013; Walsh et al., 2011). Therefore, it is of great necessity to assess drinking water biological safety through detection of disease-causing microorganisms.

Besides, the omnipresence of microbes in drinking water and the sensitivity of microorganisms to changes in their habitat enables microbes to serve as indicators for monitoring drinking water quality and DWDS microenvironmental change. Due to the ubiquity of microbes, various biological processes would occur within and across different phases in DWDS, thus potentially contributing to drinking water quality deterioration. From the perspective of drinking water bacteriology, the DWDS microenvironment can be divided into four phases, namely bulk water, suspended solids, pipe wall biofilm, and loose deposits (G. Liu et al., 2013c). Each of the phases carries microorganisms, supports microbial growth and interacts with each other. Microbes in bulk water phase, suspended solids, and loose deposits seed and promote the development of pipe wall biofilm. Whereas biofilm detachment and release of cells will in turn contribute to mobility of biofilm-embedded microbes. Moreover, some pipe materials may enhance biofilm formation and OPs growth (Learbuch et al., 2019), hence would potentially give rise to health risks through release of OPs from biofilm into bulk water or enhance other water quality problems such as red water, corrosion of metal pipes or growth of invertebrates. In addition, free-living amoeba could prey on bacteria from biofilms and thus may carry OPs and protect them from disinfection (Delafont et al., 2013; Wingender & Flemming, 2011). Besides, drinking water microbial community will change in response to changes in environmental conditions. For instance, fluctuations in hydraulic conditions or switching of supply-water quality may affect the growth and release of biofilm-embedded microbes, which will lead to shift in microbial community structures in various phases, potentially associated with health threats (G. Liu et al., 2017b; L. Liu et al., 2016).

Due to lack of desirable approaches to directly detect some waterborne pathogens, determination of potential health risks in drinking water was typically conducted by analysis of specific indicator microorganisms (Saxena et al., 2014). The most commonly used indicator microorganisms for drinking water are thermotolerant or fecal coliforms, and *Escherichia coli* (Bridle et al., 2014; Saxena et al., 2014). Despite the inexpensive and easy detection of those microbial indicators, there is not always good correlation between the detection of indicators and presence of pathogenic microorganisms (Bridle et al., 2014; Hörman et al., 2004; Savichtcheva & Okabe, 2006; Saxena et al., 2014). That is to say, certain pathogens may still

be present in drinking water in the absence of indicator microorganisms (Bridle et al., 2014; Hunter, 1997). Moreover, opportunistic premise plumbing pathogens do not correspond to fecal indicators, thus arise challenges to monitoring OPs through detection of indicator microorganisms (Wang et al., 2017). Besides of selective determination of specific indicator microorganisms, another method that is universally used for indicating general bacteriological quality of drinking water is heterotrophic plate count (HPC) test. HPC test is a culture-based test that is intended for assessing the number of heterotrophs in a given sample (Jamie Bartram et al., 2003; Bridle, 2013). As most of the bacterial pathogens and OPs are heterotrophic bacteria, the HPC result was considered to be an index for the potential risk of such pathogens in drinking water (Allen et al., 2004; J Bartram et al., 2004). HPC has proven a useful tool for determining the variation in bacterial water quality, indirectly indicating fecal contamination, and assessing bacterial regrowth potential in a drinking water sample (Allen et al., 2004; Jamie Bartram et al., 2003). However, there is no direct correlation between HPC bacteria and the presence of OPs , hence there is no direct association between health risks and HPC levels (Allen et al., 2004; Pavlov et al., 2004). Therefore, lack of specificity in identification restricts the application potential for drinking water safety assessment of this method.

Although it has been a consensus that biofilms in DWDS and premise plumbing act as a primary source for microbial contamination in tap water (Falkinham et al., 2015; Williams et al., 2013; Wingender & Flemming, 2011), there is still limited knowledge of the majority of microbial processes occurring during distribution that would give rise to deterioration in drinking water quality. Therefore, it is a requisite to develop trustworthy approaches to investigate the microbiology in drinking water, by which means demonstrate potential health risks associated with microbes, as well as deliver valuable information of drinking water microbiome and knowledge of relevant microbial processes from source to tap to engineers and managers in drinking water field, so as to safeguard biological quality of drinking water (Hull et al., 2019). As was mentioned above, conventional methods for determination of microbial water quality primarily rely on either selective or non-selective culture. However, this strategy is only applicable for nonspecific detection of cultivable microorganisms. With the development of molecular methods, such as PCR and metagenomic sequencing, the gap of detecting uncultivable microorganisms has been filled. With the help of various metagenomics tools, highly specific identification of microbes in a given mixed microbial community can be realized (Simon et al., 2019). Moreover, microbial community composition elucidated using metagenomic sequencing has proven to be a useful tool for assessing drinking water biological stability (Vierheilig et al., 2015). Additionally, the microbiome inventory of drinking water unveiled by molecular biological diagnostic tools can serve as a reference for monitoring perturbations of microbial water quality (Tan et al., 2015), which could also be used as a promising tool for providing early warnings of waterborne disease outbreaks in a given environment. Furthermore, tracking the origin of certain bacteria in drinking water with use of high-throughput sequencing technologies could help water utilities gain a better understanding of various microbial processes in drinking water from source to customers' taps, based on which water treatment strategies could be improved and better management of drinking water quality could be performed (G. Liu et al., 2018)

1.2 High-throughput sequencing technologies

1.2.1 Next-generation sequencing

To date, the most common approaches to characterize microbiomes are molecular methods targeting specific marker genes using high-throughput sequencing technologies. Next generation sequencing **(**NGS) technologies have been extensively used for investigation of environmental microbiomes since their inception from 2005 (Vierheilig et al., 2015). Despite that they can provide high-throughput characterization of microbes, these NGS platforms, such as Illumina, Roche 454, SOLiD and Ion Torrent, have relatively short read length (100-500bp) (Leggett & Clark, 2017; Loit et al., 2019; Shin et al., 2016; Vierheilig et al., 2015) and rely on PCR amplification (Leggett & Clark, 2017; Oikonomopoulos et al., 2016). The taxonomic resolution of NGS at species level is thus limited due to the choice of the primers targeting different SSU rRNA hypervariable regions, sequence assembly accuracy, and PCR amplification biases related to secondary structure or GC content of the resulting amplicons (Cusco et al., 2017; Leggett & Clark, 2017). However, as identification of pathogens always calls for species level resolution, NGS is not competent for pathogen identification. Moreover, NGS methods always require days to weeks for sequencing data acquisition (Loit et al., 2019; Ma et al., 2017; Mitsuhashi et al., 2017), making them unfeasible for rapid microbial community analysis. However, the launching of single-molecule sequencing platforms represented by Pacific Biosciences (PacBio) Single-molecule real-time (SMRT) sequencing and Oxford Nanopore MinION^{TM} overcame the aforementioned pitfalls of most short-read sequencing technologies and offered the possibility of real-time metagenomic analysis. Moreover, owing to the miniaturization of sequencing device and the affordable price, Oxford Nanopore MinION^{TM} sequencer tends to attract more attention than PacBio SMRT in the context of rapid microbial identification and diagnostics (Loit et al., 2019).

1.2.2 Nanopore sequencing

Oxford Nanopore sequencing is a both time- and cost-effective sequencing technology which utilizes protein nanopores or synthetic nanopores with a pore size of several nanometers to read DNA sequences (Deamer et al., 2016; Rhee & Burns, 2006). The schematic workflow of Oxford Nanopore DNA sequencing is depicted in Figure 1.1. During Nanopore sequencing, double-stranded DNA (dsDNA) is unzipped by the motor protein and a single strand is threaded through the nanopore inserted on an insulating membrane across which an electrical potential is applied (de Lannoy et al., 2017). The characterization of nucleobases is based on the ionic current change when different base combinations passing through the nanopore (Leggett & Clark, 2017). Furthermore, the ionic signals are then recorded and translated into sequence of the DNA strand by specific software (de Lannoy et al., 2017).

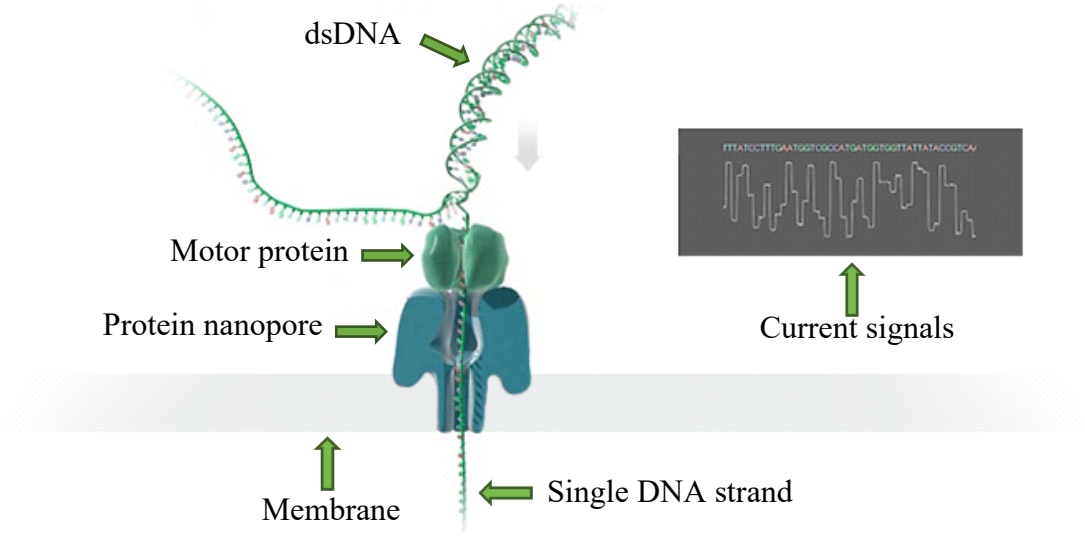

Figure 1.1 Schematic representation of Oxford Nanopore DNA sequencing technology
(Image source: **https://nanoporetech.com/how-it-works**)

Currently, Oxford Nanopore Technologies (ONT) provides two kinds of nanopore sequencing approaches, namely one directional (1D) sequencing and one directional squared ($1D^2$) sequencing Figure 1.2. The two strands of each dsDNA molecule can be considered as a template strand and the corresponding complement strand. During 1D sequencing, each DNA single strand is sequenced separately as individual strands. While in $1D^2$ sequencing library preparation, special adapters are deployed to increase the possibility of the complement strand entering the same channel immediately following the template strand. Moreover, in the succedent basecalling process, the template and complement strands are paired by the basecaller through comparing the time when the strands appear in a certain channel, their sequence lengths, and the complementarity of sequences. As a consequence, higher read accuracy can be achieved through consensus base calling of paired reads in $1D^2$ sequencing.

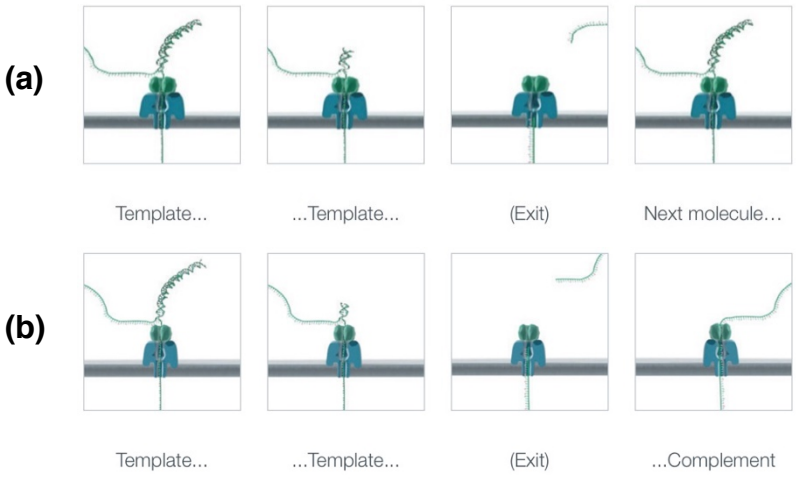

Figure 1.2 Schematic representation of Oxford Nanopore (a) 1D and (b) $1D^2$ sequencing
(Image source: https://nanoporetech.com)

1.3 Oxford Nanopore MinION™ sequencer

MinION™ is the first commercially available Nanopore sequencing device released by ONT, which is a USB-powered sequencer with size comparable to a cellphone (Figure 1.3). A MinION™ sequencing flow cell has a sensor array chip containing 512 channels, with each of which connected to 4 wells (i.e. nanopores). During sequencing, one of the 4 wells is used at a time. Therefore, a maximum number of 512 active pores of a MinION™ flow cell can be used for sequencing simultaneously. MinION™ has numerous advantages over NGS technologies. Firstly, among the most popular high-throughput sequencing platforms, MinION™ is the only portable one. A MinION™ Mk1B sequencing device with a sequencing flow cell inserted in it only weighs 103 g (87 g without flow cell). In addition, it can be connected to a laptop using a USB 3.0 cable (Figure 1.4), which makes it possible to be used in in-field sequencing (Loit et al., 2019). Moreover, MinION™ is capable of achieving sequencing length up to tens of kilobases, enabling deep amplicon sequencing of long marker genes, thus can provide enhanced resolution for bacterial identification at species level (Mitsuhashi et al., 2017), which also contributes to advanced pathogen discrimination. Furthermore, MinION™ is a real-time sequencing platform, which means that the sequencing data can be acquired and processed while sequencing (Benítez-Páez et al., 2016). The real-time data analysis provides information for assessing the quality of the sequencing run as well as helps determining the required length of sequencing experiments during sequencing (Leggett & Clark, 2017; Mitsuhashi et al., 2017). In combination with simple library preparation, the real-time acquisition of sequencing data allows rapid identification of targeted bacteria or genes, which enables prompt response, timely treatment and use of narrow-spectrum antibiotics in disease outbreaks. Finally, being a platform enabling amplification-free sequencing, the PCR-induced biases can be avoided in MinION™ sequencing, which are however inevitable in NGS-based approaches (de Lannoy et al., 2017; Oikonomopoulos et al., 2016).

Figure 1.3 USB-powered MinION™ sequencer
(Image source: **https://nanoporetech.com**)

Figure 1.4 A MinION™ sequencing device plugged in a laptop
(Image source: https://nanoporetech.com/resource-centre/introduction-nanopore-sequencing)

Despite of the aforementioned strengths over other sequencing platforms, the application of MinION™ was limited due to the relatively high per base error rate as compared with <1% error rate of *Illumina* sequencing. The error-prone nature of MinION™ reads made metagenomic sequencing problematic. Nevertheless, refinements in both Nanopore chemistry and computation tools has successfully reduced sequencing error rate by improving basecall accuracy, assembly quality and post-assembly error correction (Jain et al., 2016; Loit et al., 2019). The upgrades of MinION™ 1D chemistry continuously brought about a drop in its raw read error rate from 38% since its release to the currently reported 8% (Hu et al., 2018; Jain et al., 2018; Jain et al., 2016; Ma et al., 2017). In addition, researchers have developed various post-bascalling error correction bioinformatics tools enabling to further restrict the sequencing error rates to 0.5-3% (Calus et al., 2018; Jain et al., 2018; C. Li et al., 2016; Volden et al., 2018).

1.4 MinION™ sequencing applications

Since the advent of Oxford Nanopore sequencing technology, it has shown great potential in a variety of application domains including characterization of microbiomes (Benítez-Páez et al., 2016; Benítez-Páez & Sanz, 2017; Brown et al., 2017; Cuscó et al., 2018; Cusco et al., 2017; Loit et al., 2019; Ma et al., 2017; Mitsuhashi et al., 2017; Shin et al., 2016), detection of antibiotic resistance genes (Ashton et al., 2015; Judge et al., 2015; R. Li et al., 2018; Runtuwene et al., 2018; Tarumoto et al., 2017; van der Helm et al., 2017; Xia et al., 2017), clinical diagnostics (Greninger et al., 2015; Günther et al., 2017; Harstad et al., 2018; Lemon et al., 2017; Quick et al., 2017; Schmidt et al., 2016; Votintseva et al., 2017) and human genome

studies (Bowden et al., 2019; De Coster et al., 2019; Jain et al., 2018; Karamitros et al., 2018; Shafin et al., 2019). Being a portable real-time sequencer, MinION™ has proven utility in challenging field environments, such as in tropical rainforest (Pomerantz et al., 2018), Arctic permafrost (Goordial et al., 2017), Arctic glacier (Edwards et al., 2018), Antarctic dry valleys (S. S. Johnson et al., 2017) and space station (Castro-Wallace et al., 2017). Besides, MinION™ is becoming a ubiquitous tool in studies of microbiomes thanks to its long read length and capability of sequencing without amplification, which enables sequencing of long amplicons and eliminates amplification-related biases, respectively.

1.4.1 Microbiome characterization by MinION™

The most common strategy for characterization of microbes is to sequence specific marker genes, in which hypervariable regions of 16S rRNA gene are the most widely utilized marker genes for bacterial identification (Cuscó et al., 2018; Cusco et al., 2017). Due to the similarity of 16S rRNA amplicon sequences of the closely related microorganisms, the ability of classifying sequencing reads to various taxonomy level is highly dependent on read length (Cusco et al., 2017), while the short-read sequencing technologies can produce only a partial sequence of the 16S rRNA gene, leading to failure in the taxonomy assignment at the species level (Cuscó et al., 2018; Shin et al., 2016). However, it is critical to achieve a higher taxonomic resolution to species level for demands of high accuracy microbial community profiling as well as precise pathogen identification, which is currently possible with Nanopore sequencing. The ultra-long read provided by Nanopore sequencing allow researchers to perform long amplicon sequencing and even whole genome sequencing. Despite the error-prone nature of Nanopore technology, sequencing of long marker genes can compensate for the low per-base accuracy, which makes it a promising tool to conduct high resolution microbial community analysis at the species level (Benítez-Páez & Sanz, 2017; Shin et al., 2016).

Despite the simplification of sequencing library preparation and absence of amplification bias, amplification-free MinION ™ sequencing has not been extensively applied for characterization of microbial communities due to the high demand for input DNA (Tyler et al., 2018). Brown et al. (2017) evaluated the potential of MinION™ whole genome sequencing in microbial community analysis using several mock communities and found that robust taxonomic classification of high-complexity microbial communities was still challenging with PCR-free MinION™ sequencing. Nevertheless, some research has already been performed to assess the performance of long amplicon PCR-based sequencing using MinION™ for microbial community profiling. Shin et al. (2016) evaluated the potential of MinION™ for accurate classification of bacterial community composition in mouse gut by comparison of full-length 16S rRNA amplicon sequencing data from MinION™ and short-read sequencing data from *Illumina*. Despite the relatively high error rate, MinION™ full-length 16S amplicon sequencing was still capable of providing more accurate taxonomy assignment than short amplicon sequencing on *Illumina* platform (Shin et al., 2016). Benítez-Páez and Sanz (2017) successfully reconstructed the structure of two commercially available mock communities using long amplicon sequencing of *rrn* region despite the existence of coverage bias in relative

proportions of some bacterial species. Cuscó et al. (2018) assessed the applicability of long amplicon MinION™ sequencing with full-length 16S rRNA gene and whole *rrn* operon in microbial community analysis. Results demonstrated that taxonomy resolution down to species level was achievable with long amplicon sequencing even in complex microbial communities. Furthermore, the taxonomic resolution was higher when using *rrn* operon while full-length 16S amplicon could provide more reliable abundance profile (Cuscó et al., 2018). In order to further improve the species identification accuracy of full-length 16S amplicon sequencing, researchers developed several workflows aiming at consensus error correction of Nanopore sequencing reads. For instance, the intramolecular-ligated nanopore consensus sequencing (INC-Seq) workflow established by C. Li et al. (2016) was capable of increasing the consensus read accuracy to 97%-98%. Based on the INC-Seq workflow, Calus et al. (2018) developed an improved workflow comprising near full-length 16S amplicon sequencing and *de novo* data processing pipeline called NanoAmpli-Seq, which could achieve a mean sequence accuracy of 99.5 ± 0.08% on artificial microbial communities.

1.4.2 Microbial community profiling with NanoAmpli-Seq

NanoAmpli-Seq is a Nanopore sequencing-based long amplicon sequencing workflow consisting of 16S rRNA gene Nanopore sequencing library preparation (Figure 1.5) and bioinformatics pipeline for taxonomy analysis (Figure 1.6). In this method, near full-length 16S rRNA genes are PCR amplified with use of 5' phosphorylated primer set 8F (5'-AGRGTTTGATCMTGGCTCAG-3') and 1387R (5'-GGGCGGWGTGTACAAG-3'). The resulting 16S amplicons have a length of approximately 1,400 bp. Subsequently, the near full-length 16S amplicons are self-ligated to construct plasmid-like molecules for performing Rolling Circle Amplification (RCA). During RCA, Phi29 isothermal polymerase is used to help generate concatemers consisting of tandem repeats of 16S rRNA genes. The purpose of producing concatemerized amplicon molecules is to enable intra-molecule consensus calling in the subsequent bioinformatics processes to reduce sequence error. Then the hyper-branched RCA products are de-branched and fragmented to be converted back to linear dsDNA molecules for Nanopore sequencing. The desired fragment size range of the final products is from 1,800 bp to 20 kbp, which correspond to 1 to 14 times of the length of full-length 16S rRNA gene. After Nanopore $1D^2$ sequencing, the $1D^2$ basecalled reads are firstly subjected to intra-read iterative consensus calling with use of INC-Seq program developed by C. Li et al. (2016). This first error correction step is capable of increasing the average read accuracy to 97-98%. Moreover, the incorrect amplicon orientation induced by INC-Seq and tandem repeats insertion in stitching sites resulting from re-orientation are solved by chopSeq program. A size filtration step is employed to select for reads of 1,300 to 1,450 bp to get rid of incomplete amplicon sequences. Furthermore, the chopSeq corrected reads are further processed with nanoClust program for OTU analysis. Each of the reads is split into three partitions and the same partitions of different reads are grouped together for OTU clustering. The partition with highest OTU counts is considered to be the optimal partition. Subsequently, the full-length reads are recruited into the optimal partition-based OTUs for consensus alignment to construct consensus sequence of each OTU. With the help of the three bioinformatics error correction

tools, this 16S rRNA gene sequencing pipeline was reported to be able to achieve overall sequence accuracy of ~99.5%. The remaining error in consensus sequences is primarily originated from homopolymer errors during sequencing, which could potentially be resolved by improvement of base calling algorithm.

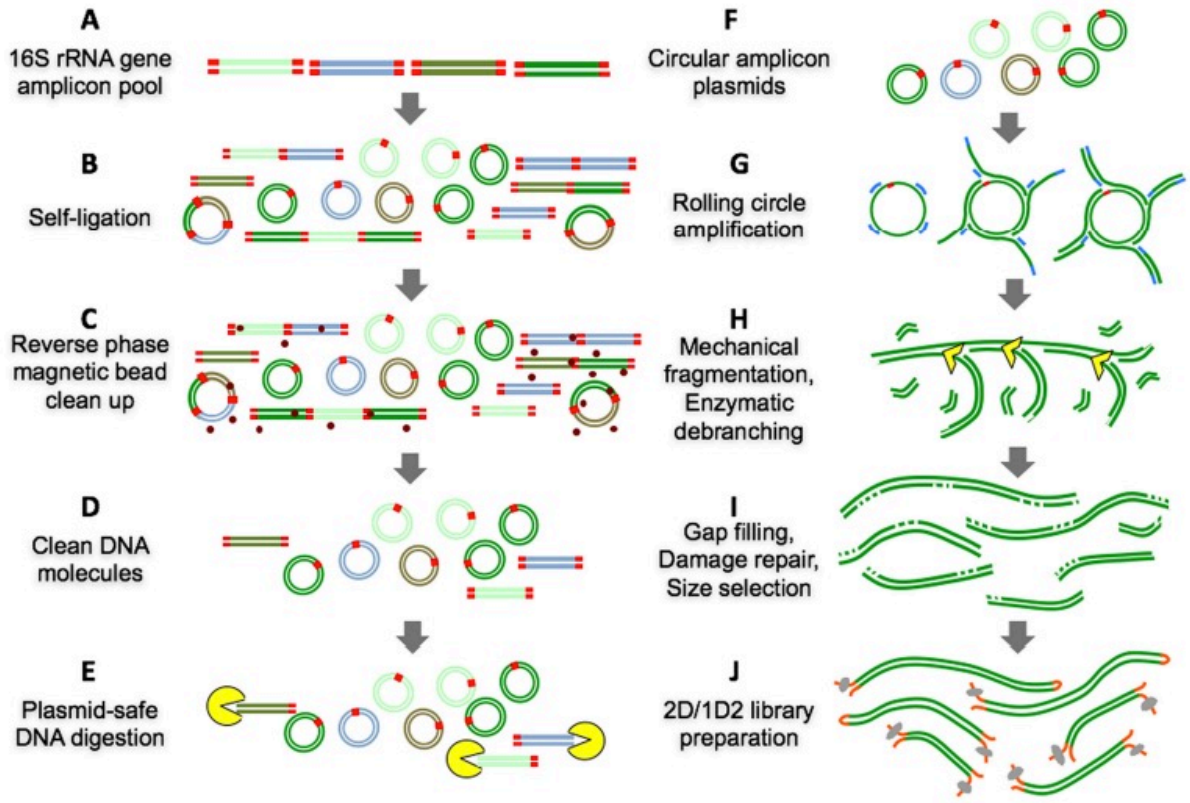

Figure 1.5 Schematic workflow of NanoAmpli-Seq library preparation (Calus et al., 2018)

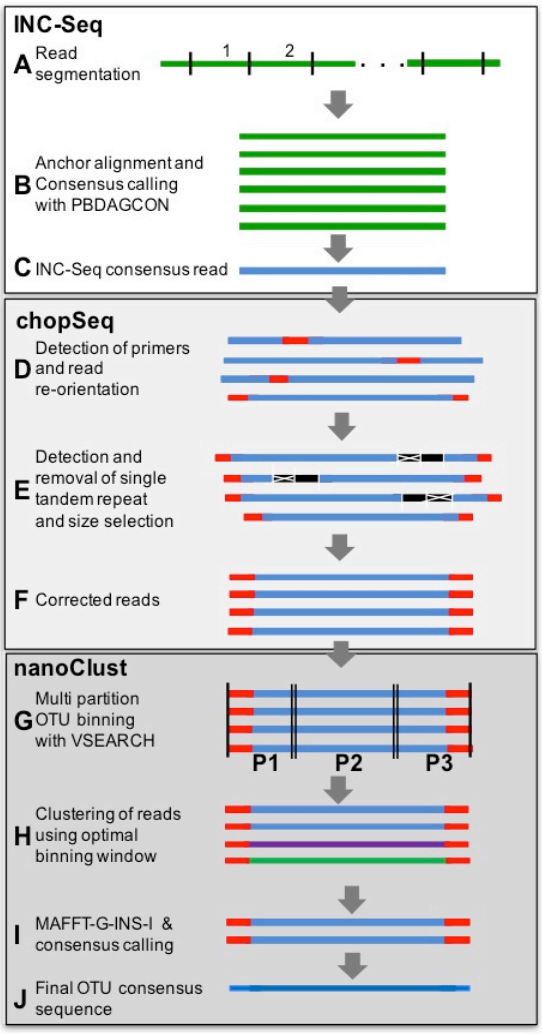

Figure 1.6 Schematic workflow of NanoAmpli-Seq bioinformatics pipeline
(Image source: http://userweb.eng.gla.ac.uk/umer.ijaz/)

1.5 Research objectives and research questions

Since drinking water related studies have imperious demand for rapid analysis and prompt response, the real-time nature of Nanopore sequencing may offer remarkable advantages over other sequencing technologies. Moreover, current short-read sequencing technologies (e.g. Illumina) fail to accurately assign taxonomy of the microbial community at species level, while Nanopore sequencing can generate ultra-long reads whose lengths are long enough to study the entire 16S rRNA gene and even the whole genome, which is expected to overcome the limitation. In addition, the presence of amplification biases for various hypervariable regions of the 16S rRNA gene limited the taxonomic identification accuracy of the PCR-dependent sequencing approaches. Nanopore sequencing can directly sequence genomic DNA without amplification, thus might refrain from the amplification-related biases. However, although Nanopore sequencing has proven to be a promising tool for identification of microbes and pathogens in many application domains (e.g. clinic, human genome, plant, wastewater), there are still some impediments to overcome to apply it in drinking water microbiology investigation. Drinking water has an extremely low microbial biomass content (i.e. 10^3-10^6 cell/mL). However, as Nanopore sequencing has a high demand on both quality and quantity of input DNA, large sample volume is required to collect adequate DNA from drinking water for direct gDNA sequencing. Therefore, Nanopore gDNA sequencing without PCR amplification is not always feasible on drinking water. Nevertheless, efforts have been made to enable highly accurate species identification through Nanopore sequencing of 16S rRNA genes (Calus, 2018; Calus et al., 2018; C. Li et al., 2016). In this sense, NanoAmpli-Seq near full-length 16S amplicon sequencing might serve as a favorable substitute for direct gDNA sequencing in microbial community profiling of low-biomass samples such as drinking water.

Consequently, this research was conducted to explore the potential of full-length 16S amplicon sequencing and direct gDNA sequencing on Oxford Nanopore MinION™ sequencing platform in profiling drinking water microbial community. For this purpose, NanoAmpli-Seq full-length 16S rRNA amplicon sequencing and PCR-free $1D^2$ gDNA sequencing were performed on both artificial microbial community DNA and environmental DNA from tap water with use of MinION™ sequencer. Furthermore, tap water taxonomy classification result obtained from NanoAmpli-Seq full-length 16S rRNA amplicon sequencing was compared with that from Nanopore $1D^2$ gDNA sequencing. The main objective of this research was to develop an approach for rapid and accurate characterization of microbiome in drinking water. Specifically, this study aims to assess the applicability of Nanopore MinION™ sequencing in investigation of drinking water microbial communities, improve the taxonomic resolution with use of full-length 16S rRNA amplicon sequencing, and reduce amplification-related sequencing biases by direct sequencing of gDNA.

Following research questions were answered:

1) What is the potential for species identification and microbial community profiling of an artificial microbial community and drinking water by Nanopore MinION™ full-length 16S rRNA gene sequencing?
2) What is the potential for species identification and microbial community profiling of an artificial microbial community and drinking water by Nanopore MinION™ gDNA sequencing?
3) What is the difference of drinking water microbial community analysis between full-length 16S rRNA amplicon sequencing and direct gDNA sequencing using MinION™ sequencer?

2 Materials and methods

2.1 Experimental design

The schematic workflow of the experimental design is displayed in Figure 2.1. Tap water was chosen as the research object of this research. The biomass in tap water was collected by filtrating a large volume of water through the filter membranes. After DNA extraction, purification and concentrating was performed to obtain concentrated DNA with high purity. In this study, full-length 16S rRNA gene and gDNA were sequenced on MinION^{TM} platform. In terms of gDNA sequencing, sequencing was performed in absence of PCR amplification and the gDNA of the microbial communities was directly sequenced by MinION^{TM} sequencing device using 1D^2 sequencing approach. In terms of full-length 16S rRNA gene sequencing, the experiments were carried out according to the NanoAmpli-Seq protocol developed by Calus et al. (2018). In order to assess the sequencing performance of NanoAmpli-Seq workflow for complex microbial communities, the workflow was applied to three mock bacterial community DNA samples prior to full-length 16S rRNA gene sequencing of tap water DNA samples. As for drinking water microbial community profiling, the microbial community composition derived from full-length 16S rRNA amplicon sequencing was compared with that from gDNA sequencing.

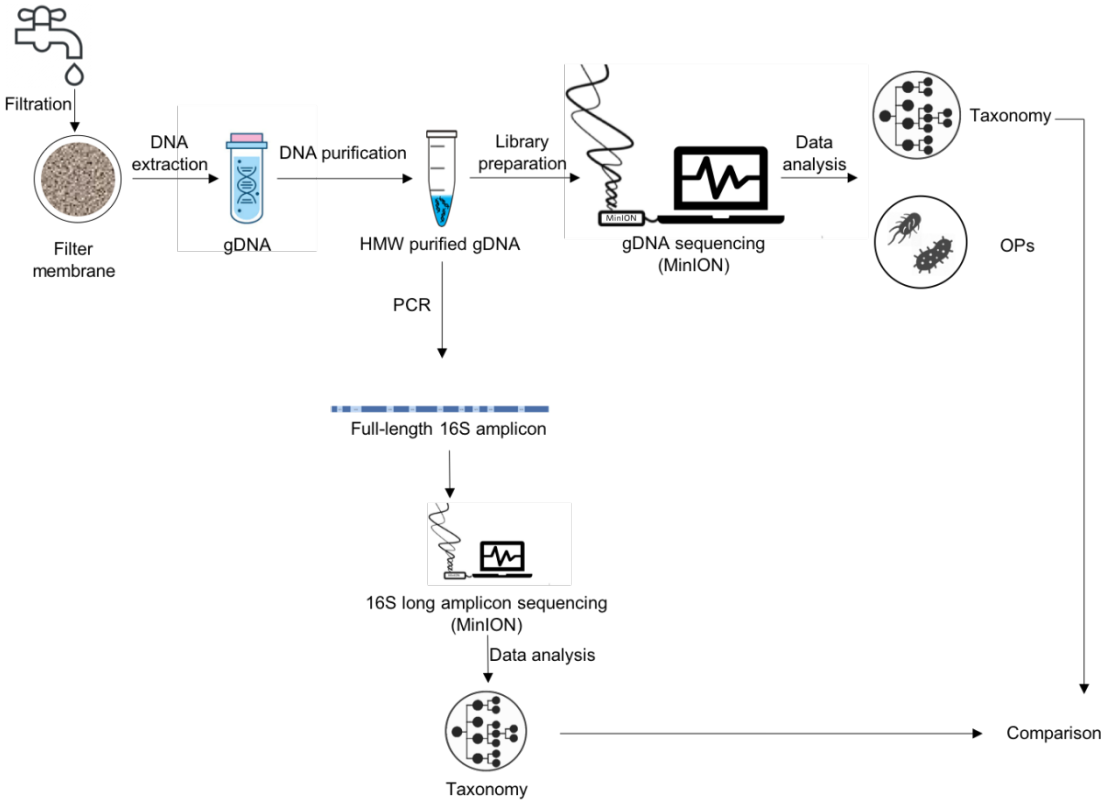

Figure 2.1 Schematic workflow of experimental design

2.2 Mock community construction

Commercial mock microbial community DNA standard (D6305) consisting of genomic DNA of eight bacteria strains (*Bacillus subtilis, Enterococcus faecalis, Escherichia coli, Lactobacillus fermentum, Listeria monocytogenes, Pseudomonas aeruginosa, Salmonella enterica,* and *Staphylococcus aereus*) and two yeast strains (*Cryptococcus neoformans* and *Saccharomyces cerevisiae*) was purchased from Zymo Research. The information of the eight bacteria strains in the mock community was provided in Appendix A (Table A.1). Genomic DNA of *Legionella pneumophila* (DSM-7513) was purchased from DSMZ (Germany). The Genbank accession number of this type strain is AE017354. The commercial mock community DNA and *Legionella pneumophila* DNA were used as templates for PCR amplification. The two DNA standards were amplified separately with use of 5' phosphorylated primer set 8F (5'-AGRGTTTGATCMTGGCTCAG-3') and 1387R (5'-GGGCGGWGTGTACAAG-3') to obtain near full-length 16S rRNA amplicons. Furthermore, the mock bacterial community amplicon pools were constructed by adding 16S amplicons of *Legionella pneumophila* to that of commercial mock community at 0%, 5% and 10% abundance.

High molecular weight mock microbial community DNA standard (D6322) comprised of high molecular weight genomic DNA of seven bacteria strains (*Bacillus subtilis, Enterococcus faecalis, Escherichia coli, Listeria monocytogenes, Pseudomonas aeruginosa, Salmonella enterica,* and *Staphylococcus aereus*) and one yeast strain (*Saccharomyces cerevisiae*) was obtained from Zymo Research. The information of the microorganism species in this high molecular weight DNA standard was shown in Appendix A (Table A.2). The mock community DNA used for assessment of MinION™ genomic DNA sequencing accuracy was generated by mixing the genomic DNA of *Legionella pneumophila* and the high molecular weight mock community DNA.

2.3 Tap water DNA sample collection

Tap water samples were collected directly from a tap at WaterLab in Faculty of Civil Engineering and Geosciences, TU Delft. Before taking samples, tap water was kept running for ~2 minutes to eliminate variation of tap water composition caused by disturbance. Microbes in tap water were concentrated by vacuum filtration of 60 to 100 liters of tap water. Filter membranes used in this study were 0.22 μm pore size Polyether sulfone (PES) filters (Millipore Express® PLUS, USA). The used filter membranes were put in a petri dish and stored at 4 ºC in the refrigerator if not subjected to DNA extraction immediately. DNA in the microbes intercepted by the filter membranes was extracted using the FastDNA® Spin Kit for Soil (MP Biomedicals, USA) with use of MiniBeadBeater-16 (BioSpec, USA) for bead beating. Prior to DNA extraction, filter membranes were cut into small pieces with sterile scissors to ensure sufficient contact with lysing matrix for utmost recovery of DNA from the filters. The DNA extraction protocol provided by the manufacturer was slightly modified in order to increase DNA yield and enhance removal of impurities. Briefly, centrifugation after bead-beating was extended to 15 minutes and the succedent 5 minutes' protein precipitation with PPS solution was performed on ice. Moreover, the SEWS-M solution washing step was repeated for a total of three washes. Furthermore, the air dry of the Spin Filter was enhanced with 60 ºC incubation in a heat block. Ultimately, 60 μL of DES were used for resuspending the Binding Matrix and the incubation before the final elution step was performed at 55 ºC for 5 minutes. The tap water DNA samples were prepared in triplicate and were stored at -20 ºC until use.

2.4 DNA size selection

Two DNA size selection methods using magnetic beads solution were assessed on tap water DNA samples in terms of removal effectiveness of short fragments. One of the methods was 0.4× ratio AMPure XP beads (Backman Coulter, USA) clean-up following the manufacturer's instruction. The other size selection method was performed following the 0.7× ratio modified SPRI beads solution size selection protocol adapted by ONT (Retrieved from https://community.nanoporetech.com/extraction_methods#size_selection&modal=size_selection). The final extracted DNA samples were subjected to 0.4× ratio AMPure XP beads clean-up to enrich for fragments > 1 kbp.

2.5 DNA quantification and qualification

It is recommended that 1.5 µg of high molecular weight DNA (≥ 10 kbp) or 250 fmol of fragmented DNA (< 10 kbp) in 48 µL of buffer solution is used as input DNA for library preparation by the $1D^2$ Ligation Sequencing Kit (SQK-LSK309, Oxford Nanopore Technologies, UK) protocol. Moreover, ONT recommends that the input DNA has an OD 260/280 value of ~1.8 and OD 260/230 value of ~2.0-2.2. The DNA concentration of all DNA samples was quantified using Qubit 3.0 fluorometer (Thermo Fisher Scientific, USA) with dsDNA High Sensitivity Assay kit (Thermo Fisher Scientific, USA) following the manufacturer's instruction. DNA purity was assessed with use of NanoDrop (Thermo Scientific, USA).

2.6 Full-length 16S rRNA gene sequencing

2.6.1 Sequencing library preparation

The DNA samples were sent to BaseClear B.V. for PCR amplification. The 16S rRNA gene of the DNA samples was PCR amplified using 5' phosphorylated primer set 8F (5'-AGRGTTTGATCMTGGCTCAG-3') and 1387R (5'-GGGCGGWGTGTACAAG-3') to generate near full-length 16S amplicons. Sequencing libraries were prepared according to the NanoAmpli-Seq workflow developed by Calus et al. (2018) with some modifications. Briefly, the amplicon pools were diluted to 2-3 ng/µL in nuclease-free water and subjected to self-ligation. 10 µL of Blunt/TA Ligase Master Mix (M0367S, New England Biolabs) were mixed with 90 µL of diluted amplicon pool and incubated at 10 ºC for 15 minutes then at 25 ºC for 10 minutes. 100 µL of concentrated magnetic beads solution was prepared as described by Calus et al. (2018). 0.35× ratio of concentrated beads solution was added to the self-ligation product to remove the multi-molecule hybrids. Briefly, 35 µL of the prepared concentrated beads solution was mixed with 100 µL of the self-ligation product with use of wide-bore pipette tips followed by two minutes' incubation at room temperature. The mixture was then placed on a magnetic rack for one minute to separate beads from the solution. 135 µL of clear supernatant was transferred into a new tube and subjected to a 0.5× ratio AMPure XP beads clean-up following the manufacturer's instruction. The purified amplicons were eluted with 15 µL of warm nuclease-free water as described in Appendix B. The remaining linear amplicons were removed with use of Plasmid-Safe™ ATP-Dependent DNase (E3110K, Epicentre) following the mini-preparation protocol given by the manufacturer except for shortening the

incubation time to 15 minutes. The product was cleaned-up with 0.5× ratio AMPure XP beads according to the manufacturer's instruction and eluted in 10 µL of warm nuclease-free water. The purified self-ligated amplicons were subjected to rolling circle amplification with use of TruePrime® RCA Kit (390100, Expedeon). The amplification reaction was performed according to manufacturer's protocol with exception of adjusting the incubation temperature to 29.5 ºC. Amplification samples were prepared in triplicate. After 150 minutes' incubation, concentrations of triplicate samples were measured with Qubit 3.0 fluorometer. The samples whose concentration were >50 ng/µL were stored on ice for later use while the samples with low concentration were incubated for another 30-45 minutes until the concentration reached ~50 ng/µL. 21 µL of each of triplicate RCA products were combined together and mixed with 2 µL of T7 endonuclease I then incubated for 6 minutes at room temperature for enzymatic debranching. The 65 µL of debranched RCA product was transferred into a g-TUBE (520079, Covaris®) and then centrifugated in Eppendorf Centrifuge 5424R at 1900 rpm for 8 minutes or until the entire mix passed through the orifice. The g-TUBE was reversed and subjected to centrifugation again as previously described. The fragmentation product was cleaned-up with 0.35× ratio concentrated beads solution and eluted in 65 µL of warm nuclease-free water. 63 µL of the purified fragmented RCA product was mixed with 2 µL of T7 endonuclease I and subjected to a secondary enzymatic debranching. After 5 minutes' incubation at 37 ºC, the debranching mix was purified with 0.45× ratio concentrated beads solution and eluted in 55 µL of warm nuclease-free water. Subsequently, DNA end-prep and sequencing adapter ligation were performed according to manufacturer's protocol with doubled incubation time. The detailed description of library preparation protocol was provided in Appendix B.

2.6.2 MinION™ sequencing

Full-length 16S amplicon sequencing library was loaded to a FLO-MIN107 flow cell after performing platform QC analysis on MinKNOW software. The flow cell priming and loading were preformed following the manufacturer's instruction.

2.7 Genomic DNA sequencing

2.7.1 Sequencing library preparation

The input genomic DNA was diluted to 30 ng/µL in nuclease-free water prior to library preparation. The genomic DNA sequencing libraries were prepared with $1D^2$ Ligation Sequencing Kit (SQK-LSK309, Oxford Nanopore Technologies) following the protocol provided by manufacturer with slight modifications to the incubation time. All the incubation time for beads cleaning and adapter ligation were doubled as described in Appendix B.

2.7.2 MinION™ sequencing

Genomic DNA sequencing was conducted on a FLO-MIN107 flow cell after performing platform QC analysis on MinKNOW software. The flow cell priming and loading were preformed following the manufacturer's instruction.

2.8 Data processing and analysis

2.8.1 MinION™ full-length 16S rRNA gene sequencing data analysis

Figure 2.2 shows the overview of full-length 16S rRNA gene sequencing data processing pipeline. The raw output MinION™ $1D^2$ sequencing data obtained from MinKNOW software in fast5 format was firstly 1D basecalled by Guppy 1D basecaller (v3.0.3) with the output flag "--fast5_out" set to generate .fast5 read files for $1D^2$ basecalling. Subsequently, the output 1D basecalling results in fast5 format were $1D^2$ basecalled with use of Guppy $1D^2$ basecaller (v3.0.3) to produce $1D^2$ reads in fastq format and converted to fasta format afterwards using seqtk program. The $1D^2$ reads in the .fasta files were subjected to consensus alignment with INC-Seq program (retrieved from https://github.com/CSB5/INC-Seq) using poa aligner. The minimum number of concatemers were set at 3 and "--iterative" flag was set. Furthermore, the corrected reads were passed to chopSeq program (v0.3, retrieved from https://github.com/umerijaz/nanopore/blob/master/chopSEQ.py) for orientation correction and tandem repeats removal. During chopSeq re-orientation, size filtration flag was set to select for reads from 1250 to 1500 bp. Ultimately, OTU clustering of the chopSeq corrected 16S rRNA reads was achieved with use of the nanoCLUST algorithm (v0.4, retrieved from https://github.com/umerijaz/nanopore/blob/master/nanoCLUST.py). The 16S rRNA reads were split into three partitions (i.e. 1-450, 451-900, 901-1300 bp) before OTU binning with VSEARCH. Post OTU clustering, taxonomy assignment of the OTUs were performed on BLAST or a cloud based data analysis platform provided by ONT named EPI2ME.

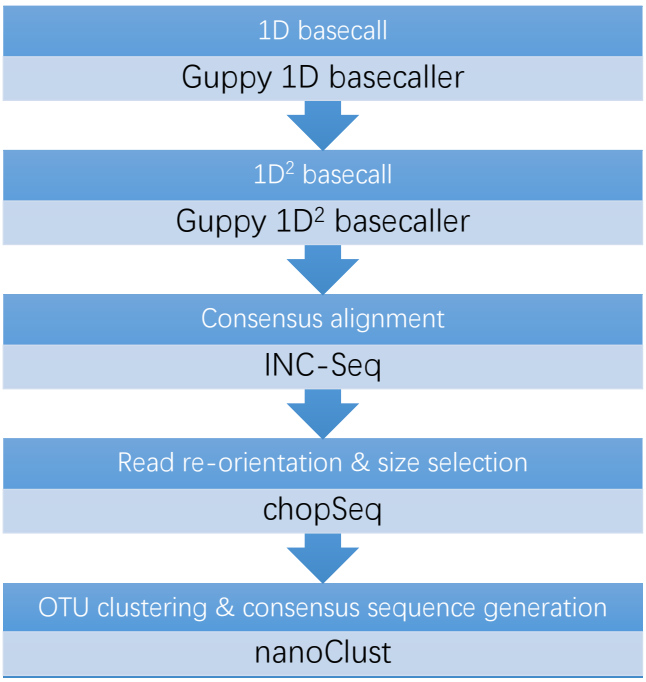

Figure 2.2 Bioinformatics pipeline for processing MinION™ full-length 16S amplicon sequencing data

2.8.2 MinION™ genomic DNA sequencing data analysis

An overview of MinION™ genomic DNA sequencing processing pipeline was shown in Figure 2.3. The raw output MinION™ 1D² sequencing data obtained from MinKNOW software in fast5 format was converted to fastq files using Guppy v3.0.3 basecaller with the output flag "--fast5_out" set to generate .fast5 read files for 1D² basecall. Subsequently, the output 1D reads in fast5 format were 1D² basecalled with use of Guppy 1D² basecaller (v3.0.3) to generate 1D² reads in fastq format. Subsequently, the 1D and 1D² read files were uploaded to EPI2ME platform and the reads were analyzed using What's in my pot (WIMP) workflow for taxonomic classification. Furthermore, the 1D² reads were polished by Racon (v0.5.0) to generate high-quality consensus sequences. Prior to Racon consensus calling, the 1D² reads were *de novo* assembled with use of minimap (v0.2, retrieved from https://github.com/lh3/minimap) and miniasm (v0.3, retrieved from https://github.com/lh3/miniasm). The final assembly in fasta format was converted back to .fastq read files and then subjected to taxonomy classification on EPI2ME using WIMP workflow.

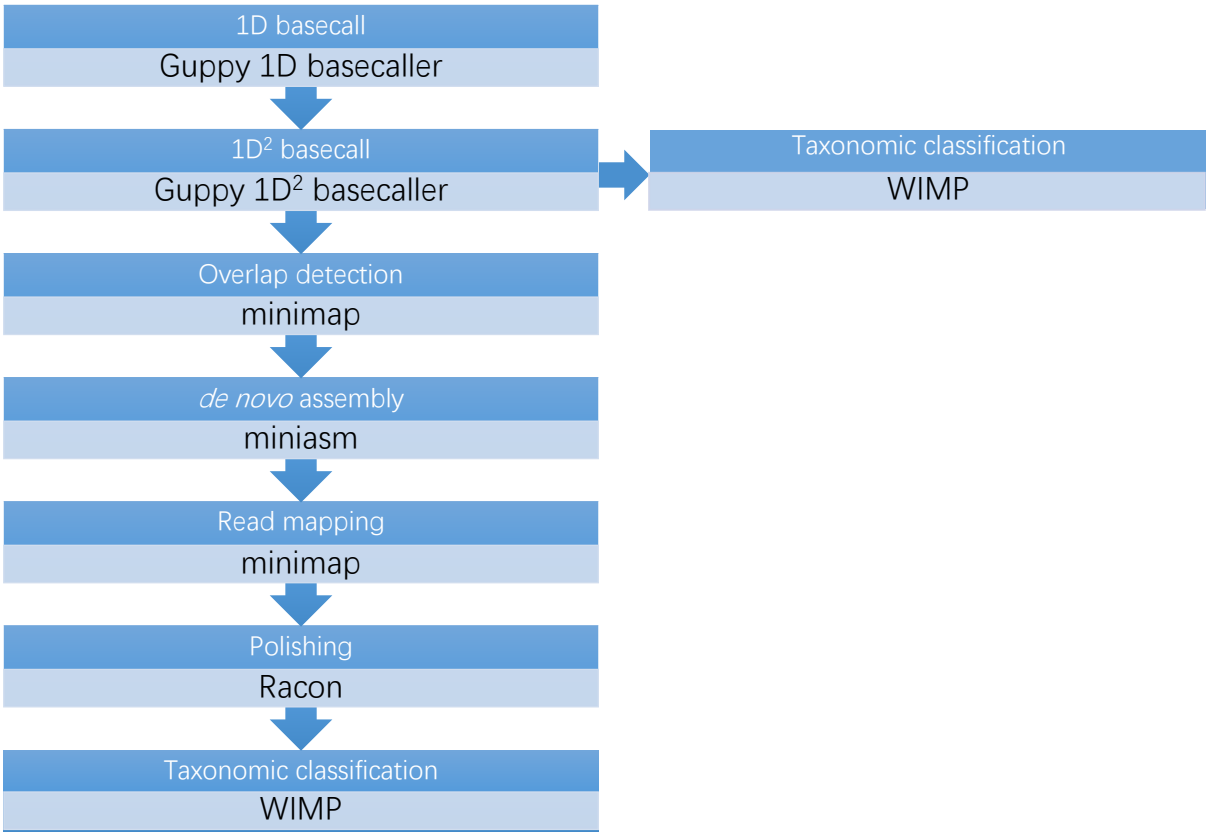

Figure 2.3 Bioinformatics pipeline for processing MinION™ genomic DNA sequencing data

3 Results and discussion

3.1 DNA sample preparation

3.1.1 Tap water DNA sample preparation

Vacuum filtration of 60 L of tap water through a 0.22 μm pore size PES membrane took 7-8 hours. A 0.22 μm pore size PES filter membrane which had filtrated 60 L of tap water was shown in Figure 3.1. After DNA extraction with FastDNA® Spin Kit for Soil, each filter membrane yielded 50-60 ng/μL DNA in approximately 50 μL of elution solution. Three replicate tap water DNA samples were prepared for sequencing in this research. DNA concentrations of the three samples before and after 0.4× ratio AMPure XP beads size selection are shown in Table 3.1.

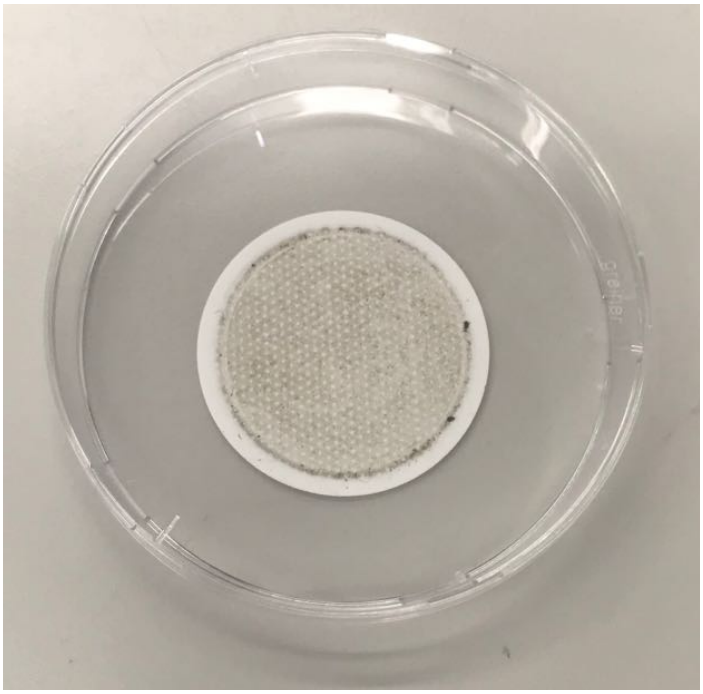

Figure 3.1 Petri dish with a used PES filter membrane

Table 3.1 DNA concentration of tap water DNA samples before and after selection of >1kb fragment size

Sample NO.	1	2	3
Concentration before size selection (ng/μL)	57.0	51.0	48.0
Concentration after size selection (ng/μL)	50.6	47.4	33.4
Recovery rate (%)	88.8	92.9	69.6

3.1.2 DNA size selection

Each of the two DNA size selection methods were conducted on two tap water DNA samples respectively. The short fragments removal effectiveness of the two size selection strategies was confirmed by agarose gel electrophoresis (Figure 3.2). It can be seen that 0.7× ratio modified beads size selection removed most of fragments below 3-4 kbp while 0.4× ratio AMPure XP beads size selection only removed DNA fragments up to 1 kbp. The concentrations of DNA samples before and after size selection were measured to estimate the recovery rate. The recovery rate of the former size selection method was around 30% (Table 3.2) while that of the latter method was over 60% (Table 3.1). In other words, the 0.7× ratio modified beads size selection is capable of obtaining DNA samples with higher average fragment length than 0.4× ratio AMPure XP beads size selection. However, due to the broad size range of DNA extracted by FastDNA® Spin Kit for Soil, removal of fragments up to 3-4 kbp would lead to considerable loss of input DNA, which might affect the microbial community profiling result. For this reason, 0.4× ratio AMPure XP beads size selection was adopted for MinION™ sequencing input DNA preparation in this research.

Table 3.2 DNA concentration of tap water DNA samples before and after 0.7× ratio modified beads selection

Sample NO.	S1	S2
Concentration before size selection (ng/μL)	52.6	49.2
Concentration after size selection (ng/μL)	17.9	14.4
Recovery rate (%)	34.0	29.3

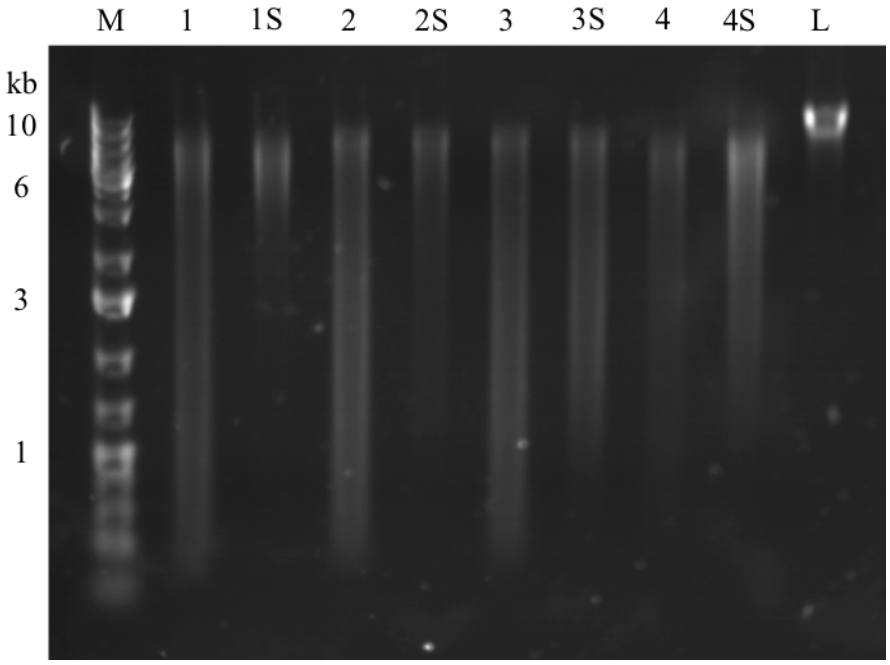

Figure 3.2 Agarose gel electrophoresis of tap water DNA samples. Lane M: GeneRulerTM 1 kbp DNA ladder; lane 1, 2, 3, 4: Original tap water genomic DNA extracted using FastDNA® Spin Kit for Soil; lane 1S, 2S: Tap water genomic DNA post 0.7× ratio modified beads solution size selection; lane 3S, 4S: Tap water genomic DNA post 0.4× ratio AMPure XP beads size selection; lane L: Lambda control DNA provided in Rapid Sequencing kit (SQK-RAD004) with an average fragment size of 48 kbp.

3.2 MinION™ full-length 16S amplicon sequencing

3.2.1 Mock microbial community DNA sequencing

Three MinION™ $1D^2$ sequencing runs were carried out on three mock community full-length 16S amplicon pools containing 0%, 5%, and 10% of *Legionella pneumophila* full-length 16S amplicons, hereinafter referred to as Zymo, Z95L5, and Z90L10, respectively. Number of active pores at the beginning of sequencing experiment of the three sequencing runs were 1275, 878, and 1277.

After 1D and $1D^2$ basecalling by Guppy basecaller, both 1D and $1D^2$ basecalled read files were passed to NanoStat program (Retrieved from https://github.com/wdecoster/nanostat) and EPI2ME platform to generate statistics of the sequencing data as shown in Table 3.3.The three sequencing runs generated 1,008,501, 586,601, and 1,031,626 raw reads with read length ranging from 5 bp to 79 kbp. The number of passed $1D^2$ reads of sample Zymo, Z95L5, and Z90L10 was 101,677, 74,921, and 93,351, which accounted for 10.1%, 12.9%, and 9.0% of the raw 1D reads, respectively. The smaller fraction of $1D^2$ read yield in the sequencing run of sample Z90L10 and Zymo as compared with that of sample Z95L5 was primarily owing to the larger proportion of unwanted short reads in 1D raw reads, as was demonstrated in Figure 3.3. The short reads are prone to ambiguous pairing and are difficult to be successfully basecalled

as paired reads by Guppy $1D^2$ basecalling algorithm, thus the vast majority of short reads whose lengths were only several hundreds of base pairs will not end up in $1D^2$ reads. Nevertheless, the yield of $1D^2$ reads in the Zymo and Z95L5 sequencing run was slightly higher than the 7-9% reported by Calus et al. (2018). This improvement was possibly brought about by the upgrade of Oxford Nanopore $1D^2$ sequencing chemistry. The total number of bases yielded in the three sequencing runs were 5.5 Gbp, 4.6 Gbp, and 6.4 Gbp for 1D data and 543 Mbp, 646 Mbp, and 577 Mbp for $1D^2$ data.

As for the sequencing run of sample Zymo, the mean read length was 5,436 bp for 1D reads and 5,342 bp for $1D^2$ reads. In terms of sequencing run of sample Z95L5, the mean read length was 7,904 bp versus 8,628 bp for 1D and $1D^2$ reads. While the mean 1D and $1D^2$ read length of sequencing run of sample Z90L10 were 6,203 bp versus 6,178 bp. The median read length of $1D^2$ reads varied from 3,388 bp to 6,627 bp, which was slightly lower than those reported by Calus et al. (2018). The length distribution of $1D^2$ reads of the three sequencing runs peaked at around 2,233 bp, 4,653 bp, and 3,621 bp, respectively.

Despite the considerable discrepancy of the read length distribution among the three sequencing runs, the statistics of their quality scores were quite similar (1D mean: 7.6, 7.7, and 7.5; 1D median: 8.1, 8.5, and 8.1; $1D^2$ mean: 8.1, 8.3, and 8.4; $1D^2$ median: 8.9, 9.1, and 9.3). Quality score is a measure of nucleobase identification accuracy of sequencing and is calculated based on the average error probability of the read.

Table 3.3 Statistics of mock community NanoAmpli-Seq full-length 16S amplicon sequencing data

Sample	Dataset	Read count	Number of bases (Mbp)	Read length (bp)			Quality score	
				Mean	Median	Mode	Mean	Median
Zymo	1D	1,008,501	5,482	5,436	3,337	1,683	7.6	8.1
	$1D^2$	101,677 (10.1%)	543	5,342	3,388	2,233	8.1	8.9
Z95L5	1D	582,603	4,605	7,904	5.932	792	7.7	8.5
	$1D^2$	74,921 (12.9%)	646	8,628	6,627	4,653	8.3	9.1
Z90L10	1D	1,031,626	6,399	6,203	4,530	231	7.5	8.1
	$1D^2$	93,351 (9.0%)	577	6,178	4,700	3,621	8.4	9.3

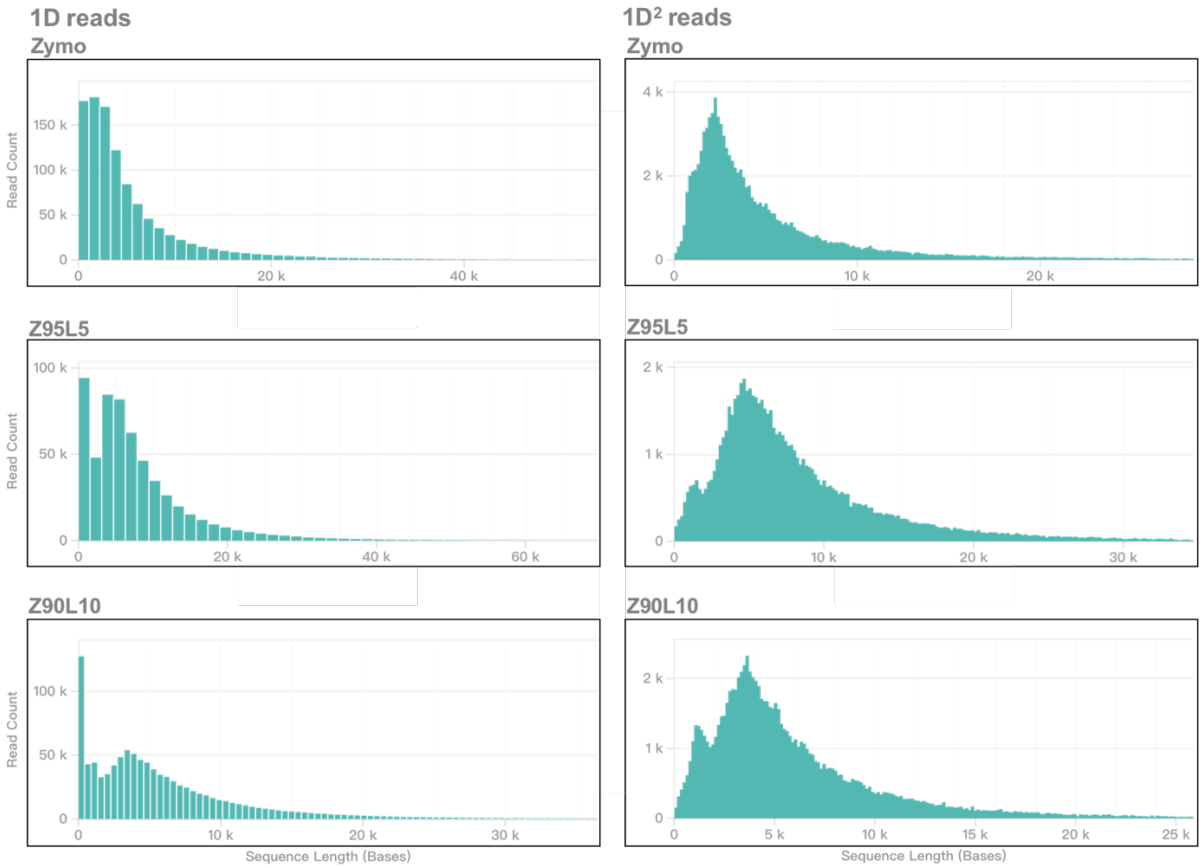

Figure 3.3 Histogram of read length distribution of mock community NanoAmpli-Seq full-length 16S amplicon sequencing 1D and 1D^2 data

The 1D^2 reads were further processed with the NanoAmpli-Seq bioinformatics pipeline as was described in section 2.8.1. The total number of reads after each processing step (i.e. INC-Seq, chopSeq and nanoClust) are displayed in Table 3.4.

The INC-Seq program was set to filter out reads with less than three concatemers. After INC-Seq alignment, the number of remaining reads was 30.2% (30,731 of 101,677), 58.6% (43,893 of 74,921), and 41.6% (38,788 of 93,351) of the 1D^2 reads for sample Zymo, Z95L5 and Z90L10, respectively. This value of sample Zymo was significantly lower than the 36%-75% reported by Calus et al. (2018), which was primarily due to the prevalence of reads shorter than 4500 bp (i.e. length of three concatemers of near full-length 16S rRNA gene) in 1D^2 reads.

As the chopSeq processing involves a size filtration step to select for reads with length ranging from 1,250 to 1,500 bp, the ratio of read counts post chopSeq to that post INC-Seq demonstrated the proportion of near full-length 16S amplicons in INC-Seq aligned reads. The ratio was 64.6% (19,850 of 30,731), 71.8% (31,528 of 43,893), and 69.7% (27,046 of 38,788) for sample Zymo, Z95L5 and Z90L10, respectively.

After nanoClust consensus calling, the total number of reads retained were 12,321, 18,959, and 18,700, which were 62.1%, 60.1% and 69% of chopSeq consensus reads. This ratio was slightly lower than those reported by Calus et al. (2018). Moreover, when comparing with the total number of raw 1D reads, the remaining reads that were ultimately subjected to OTU clustering only accounted for 1.2% (12,321 of 1,008,501), 3.3% (18,959 of 582,603), and 1.8% (18,700 of 1,031,626), which were comparable to those reported by Calus et al. (2018).

Table 3.4 Number of reads after each step of mock community NanoAmpli-Seq full-length 16S amplicon sequencing data processing

Sample	1D	$1D^2$	INC-Seq	chopSeq	nanoClust
Zymo	1,008,501	101,677	30,731	19,850	12,321
Z95L5	582,603	74,921	43,893	31,528	18,959
Z90L10	1,031,626	93,351	38,788	27,046	18,700

The OTU tables generated by nanoClust were provided in Appendix A (Table A.3, Table A.4, Table A.5). The statistics of nanoClust OTU clustering result was shown in Table 3.5. The 12,321 nanoClust consensus called reads of sample Zymo were clustered into 14 OTUs. These 14 OTUs were assigned to 8 species with 6 spurious OTUs and no false negatives. Moreover, 18,959 of Z95L5 nanoClust consensus reads resulted in 17 OTUs, among which 9 were spurious OTUs. The 17 OTUs were classified into 8 species while *Legionella pneumophila* was not detected in this sample. Furthermore, nanoClust OTU clustering of 18,700 nanoClust consensus reads of sample Z90L10 generated 15 OTUs with 7 spurious OTUs and 1 false negative. *Legionella pneumophila* was successfully classified while *Salmonella enterica* was not detected in this sequencing run. Besides, neither of the two yeast strains was detected in the three samples. The reason for this is that 16S rRNA gene only exists in prokaryotic organisms while yeasts are eukaryotic. In other words, PCR amplification using primers targeting the 16S rRNA gene would not function for yeasts. Accordingly, there was no sequences from those two yeast strains in the 16S amplicon pools used for NanoAmpli-Seq sequencing library preparation.

Table 3.5 Number of OTUs of three mock community NanoAmpli-Seq full-length 16S amplicon sequencing runs

Sample	Theoretical	Detected	Spurious	Non-detect
Zymo	8	14	6	0
Z95L5	9	17	9	1
Z90L10	9	15	7	1

Figure 3.4 shows the relative abundance of each bacterial species in the three mock communities. Relative abundance was calculated as the percent composition of a certain strain in the total number of microbes in the community. It can be seen from the plot that the pattern of relative abundance of the eight species from the commercial mock community DNA

standard (i.e. *Bacillus subtilis, Enterococcus faecalis, Escherichia coli, Lactobacillus fermentum, Listeria monocytogenes, Pseudomonas aeruginosa, Salmonella enterica,* and *Staphylococcus aureus*) was consistent among the three sequencing runs. However, the relative abundance of *Legionella pneumophila* in the two samples with 5% and 10% addition of *Legionella pneumophila* largely deviated from the theoretical value. To be more specific, the relative abundance of *Legionella pneumophila* revealed by NanoAmpli-Seq workflow was 0% (0 of 18,959) and 0.24% (44 of 18,700) in 16S amplicon pools with 5% and 10% of *Legionella pneumophila* amplicons, respectively. This pronounced deviation was presumably induced by PCR biases during RCA amplification resulted from differences in GC content (Laursen et al., 2017; Pinto & Raskin, 2012). In addition, the majority of the reads were filtered out during data processing, as was demonstrated by the aforementioned low ratio of nanoClust read count to raw 1D read count. Hence, there is possibility that the bacterial strains at a low relative abundance are not detectable with this method.

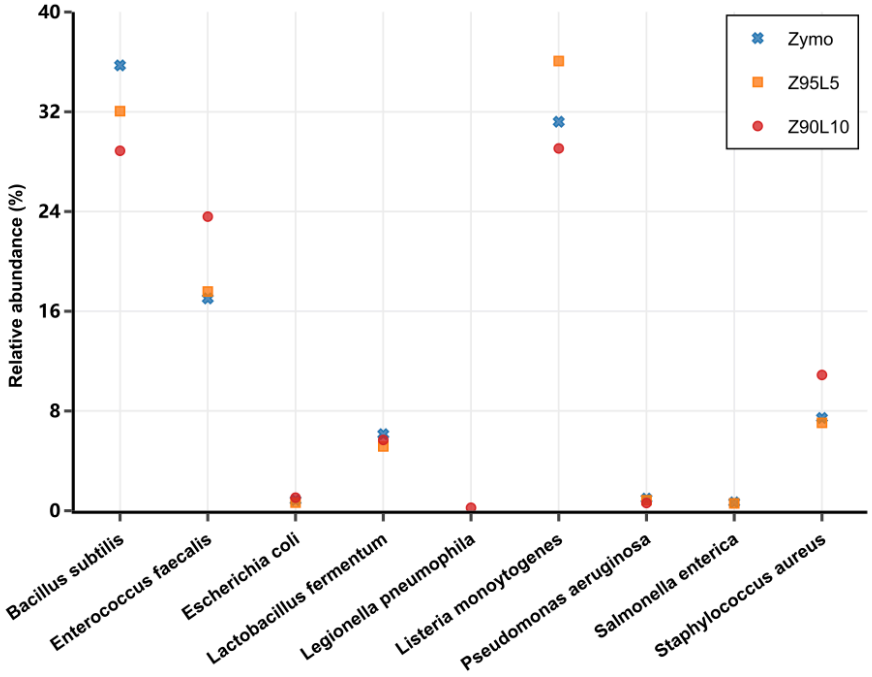

Figure 3.4 Relative abundance of species in the three mock microbial communities revealed by NanoAmpli-Seq full-length 16S rRNA sequencing

3.2.2 Tap water DNA sequencing

Tap water near full-length 16S amplicons were processed according to NanoAmpli-Seq experimental workflow and then subjected to MinION™ $1D^2$ sequencing as described in section 2.6. The raw read files were basecalled by Guppy v3.0.3 1D and $1D^2$ basecaller and then analyzed with NanoStat program for statistics analysis. The statistics of the sequencing data was summarized in Table 3.6. 586,370 of 1D reads with maximum read length of 223 kbp were generated during this sequencing run, while only 7.8% (45,858 of 586,370) of them were $1D^2$ reads. This value was slightly lower than those of mock community 16S amplicon sequencing runs, probably indicating existence of unknown inhibitors in environmental samples. It can be inferred from Figure 3.5 that the large proportion of short reads was also responsible for the low yield of $1D^2$ reads. The total number of bases generated in this sequencing run were 3.7 Gbp and 306 Mbp for 1D and $1D^2$ reads, respectively. The read length

of 1D reads had a mean and median value of 6,317 bp and 4,838 bp while the mean and median read length of $1D^2$ reads were 6,683 bp and 5.246 bp, respectively. After filtering out the short reads in 1D raw records, the majority of the reads had a read length of around 3,900 bp. The mean and median quality scores of 1D reads (7.6 and 8.3) were consistent with those obtained in mock community amplicon sequencing runs (7.5-7.7 and 8.1-8.5). Moreover, the mean and median quality scores of tap water amplicon sequencing and the aforementioned mock community amplicon sequencing runs were also quite similar in terms of $1D^2$ reads (8.4 versus 8.1-8.5, 9.2 versus 8.9-9.3).

Table 3.6 Statistics of tap water NanoAmpli-Seq full-length 16S amplicon sequencing data

Dataset	Read count	Number of bases (Mbp)	Read length (bp)			Quality score	
			Mean	Median	Mode	Mean	Median
1D	586,370	3,704	6,317	4,838	224	7.6	8.3
$1D^2$	45,858 (7.8%)	306	6,683	5,246	3,900	8.4	9.2

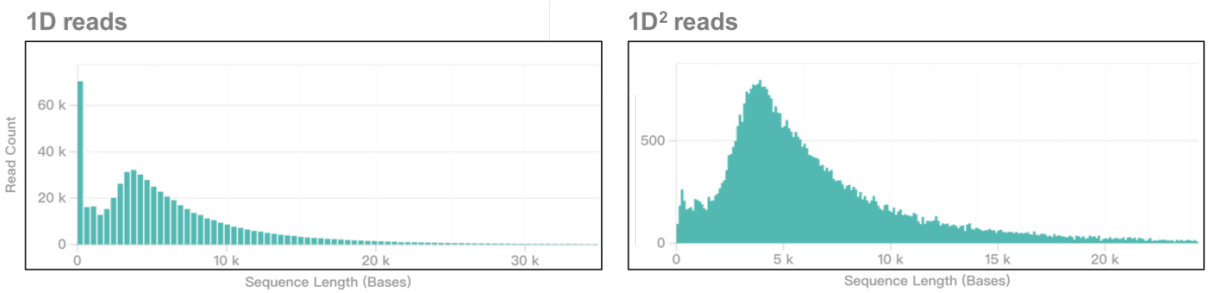

Figure 3.5 Histogram of read length distribution of tap water Nano-Ampli-Seq full-length 16S amplicon sequencing 1D and $1D^2$ data

The $1D^2$ reads were processed by INC-Seq program for consensus alignment and the number of reads passing the three concatemer threshold was 50.4% (23,114 of 45,858) of $1D^2$ reads, which was similar to those reported by Calus et al. (2018). After chopSeq re-orientation and size selection, 53.5% (12,374 of 23,114) of post INC-Seq reads were retained and passed to nanoClust program for OTU analysis. Ultimately, 3,061 reads were generated out of 12,374 chopSeq corrected reads by nanoClust algorithm. The ratio of post-nanoClust reads to post-chopSeq reads were only 24.7%, which was significantly lower than those of mock community 16S amplicon sequencing runs (62.1-69%). The reason for this may lie in the fact that a large proportion of bacterial species in tap water are present at relatively low abundance. Despite that the 16S rRNA reads of those low-abundance species successfully passed through the previous processing steps and ended up in nanoClust, chances are that the reads would finally be discarded during singleton removal before OTU clustering.

Table 3.7 Number of reads after each step of tap water microbial community NanoAmpli-Seq full-length 16S amplicon sequencing data processing

Process name	1D	$1D^2$	INC-Seq	chopSeq	nanoClust
Read count	586,370	45,858	23,114	12,374	3,061
Ratio to previous step (%)	-	7.8	50.4	53.5	24.7

Furthermore, the 3,061 nanoClust consensus reads were clustered into a total of 58 OTUs. Subsequently, EPI2ME 16S classification workflow was used for taxonomy assignment of the consensus reads of each OTU. However, only 12 out of the 58 OTUs were successfully classified at species level while 8 of the OTUs were unclassified (i.e. percent identity < 75%). The 12 species level OTUs were further assigned to 10 species. Moreover, reads assigned to a species level taxonomy only accounted for 23.4% (717 of 3,061) of the total post-nanoClust reads. The OTU table is available in Appendix A (Table A.6). This lower-than-expected taxonomy resolution is most likely to be attributed to the OTU clustering and consensus sequence construction algorithm of nanoClust. As the majority of microorganisms in tap water are always closely related, multi-species binning would probably occur during OTU clustering (Calus, 2018). Consequently, the highly similar 16S rRNA gene sequences from closely related but different species were passed to the next step for consensus calling as multiple reads of one single OTU, thus generating a consensus read for that OTU which has the shared sequences of those different species but with plenty of ambiguous or even biased bases in the unshared regions. Hence, the taxonomic resolution was largely limited in this manner. When looking at phylum level, the majority of the classified reads (2,300 in 3,061 reads) belong to proteobacteria. This result is consistent with other researches regarding drinking water microbiomes (G. Liu et al., 2018).

Table 3.8 Statistics of OTU clustering result of tap water NanoAmpli-Seq full-length 16S amplicon sequencing

Taxonomy labels	Total	Class	Order	Family	Genus	Species	Not classified
OTU count	58	36	28	25	12	12	8
Read count (coverage %)	3,061 (-)	2,069 (67.6%)	1,867 (61.0%)	1,829 (59.8%)	717 (23.4%)	717 (23.4%)	370 (12.1%)

3.3 MinION™ genomic DNA sequencing

3.3.1 Mock microbial community DNA sequencing

Genomic DNA of *Legionella pneumophila* type strain (DSM-7513, DSMZ) and HMW microbial community DNA standard (D6322) were firstly diluted to 30 ng/μL and then mixed at a ratio of 1 to 19 to generate mock community genomic DNA with 5% of *Legionella pneumophila* genomic DNA. Subsequently, the mock community gDNA was fragmented to an average size of 8 kbp with gTUBE and subjected to MinION™ $1D^2$ sequencing as described in Section 2.7.1 and 2.7.2.

The sequencing run started with 1515 active pores. The 45 hours' sequencing generated 2,171,050 1D reads in total (Table 3.9) with read length ranging from 5 bp to 382 kbp. The ratio of $1D^2$ reads in raw 1D reads was 21.2% (460,617 of 2,171,050), which was substantially higher than those of 16S amplicon sequencing runs. This was most likely to be attributed to the higher quality of input DNA. As the libraries for 16S amplicon sequencing runs were subjected to mechanical fragmentation and enzymatic fragmentation, the DNA molecules could be severely damaged. The subsequent end-prep might not be effective enough for repairing all the damaged ends, thus leading to inefficiency in $1D^2$ adapter ligation as well as the following sequencing adapter ligation. Consequently, the $1D^2$ sequencing yield was compromised in NanoAmpli-Seq. The total number of bases yielded by this sequencing run was 14.1 Gbp and 3.5 Gbp for 1D and $1D^2$ reads, respectively. The read length of both 1D and $1D^2$ reads have a wide distribution, as was illustrated in Figure 3.6. The mean and median read length of 1D reads were 6,478 bp and 6,321 bp while those of $1D^2$ reads were 7,599 bp and 7,543 bp. The 1D reads had a highest proportion at around 382 bp while the $1D^2$ reads peaked at 7,725 bp. The mean quality score was 8.2 for 1D reads and 8.5 for $1D^2$ reads. While the median quality score for 1D and $1D^2$ reads were 8.9 and 9.4, respectively. Both the 1D and $1D^2$ read files were uploaded to EPI2ME with quality score threshold set to 7 for comparison of taxonomy classification result. Subsequently, 75.5% (1,639,557 in 2,171,050 reads) of 1D reads and 75.6% (348,244 in 460,617 reads) of $1D^2$ reads were passed to WIMP workflow for taxonomic classification.

Table 3.9 Statistics of mock microbial community genomic DNA sequencing data

Dataset	Read count	Number of passed reads (≥Q7)	Number of bases (Gbp)	Read length (bp)			Quality score	
				Mean	Median	Mode	Mean	Median
1D	2,171,050	1,639,557	14.1	6,478	6,321	382	8.2	8.9
$1D^2$	460,617 (21.2%)	348,244	3.5	7,599	7,543	7,725	8.5	9.4

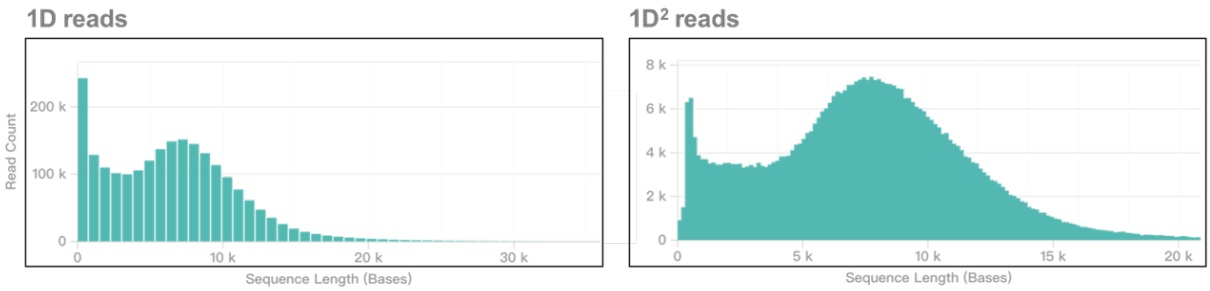

Figure 3.6 Histogram of read length distribution of mock microbial community genomic DNA sequencing 1D and 1D^2 data

3.3.1.1 Effect of sequencing approach on microbial community profiling

All of the nine species were successfully identified both with 1D and 1D^2 reads. Moreover, the observed relative abundance of *Legionella pneumophila* was slightly higher than the theoretical value (8.5% for 1D reads and 6.2% for 1D^2 reads versus 5%). However, there still existed some false positives (3.4% and 2.0% for 1D and 1D^2 reads), which might result from either contamination in the sequencing library or misclassification by the bioinformatics algorithm. The relative abundance of each strain revealed by 1D and 1D^2 data are shown as a bar chart in Figure 3.7. It can be seen from Figure 3.7 that the proportions of false positives and unclassified reads were lower for 1D^2 reads as compared with those of 1D reads, which was probably attributable to the increased read accuracy brought about by 1D^2 basecall.

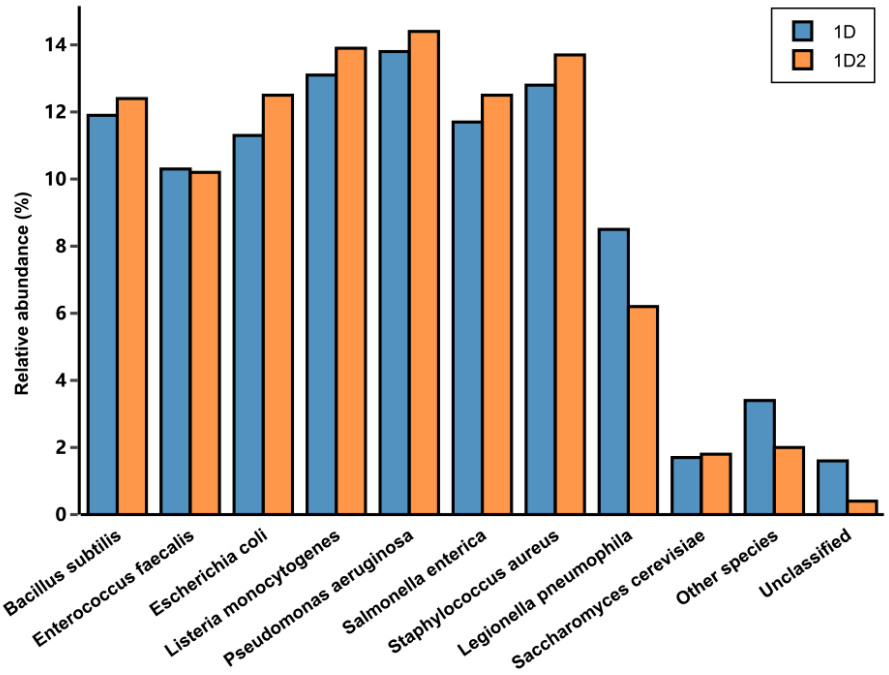

Figure 3.7 Bar chart of relative abundance of species in the mock microbial community revealed by MinION^{TM} genomic DNA sequencing 1D and 1D^2 data

3.3.1.2 Effect of quality threshold on microbial community profiling

In order to assess the impact of quality score threshold on the taxonomy classification result, the $1D^2$ reads were uploaded to EPI2ME to conduct WIMP workflow with quality score threshold set at 7 (default), 8, 9, 10 and 11, respectively. Only reads with quality score satisfying the set quality metric were subjected to taxonomy classification. The observed relative abundance of each species across the five datasets are compared in Figure 3.8. Although the proportion of false positives and unclassified reads decreased with higher quality threshold, there were still approximately 2% of the reads not correctly classified (Figure 3.8 (a)). Additionally, with quality score threshold exceeding 10, the relative abundance of each strain deviated further from the theoretical values (Figure 3.8 (b)). This can be explained by loss of mass data with high quality score threshold (Table 3.10). When the quality score threshold was set at 9, there were still 55.7% (256,726 of 460,617) of $1D^2$ reads retained by the quality score filter and passed to WIMP workflow. However, with the quality score threshold raised to 10, 63.3% (291,688 of 460,617) of the total $1D^2$ reads failed to pass the quality score filter and did not take part in taxonomy assignment. Furthermore, if the quality threshold was set to 11, only 12.2% (56,083 of 460617) of total $1D^2$ reads were passed to WIMP workflow for taxonomy classification. Therefore, this method of improving the quality of the reads fed to WIMP workflow by solely raising the quality score threshold to leave out the low-quality reads is at the expense of losing mass of available data. In addition, AT-rich reads tend to have higher quality scores than GC-rich reads (Krishnakumar et al., 2018). As a consequence, with the increase of quality score threshold, larger fractions of sequences of high GC content microbes did not end up in taxonomy classification as compared with those of microbes with lower GC content, thus resulting in biased microbial community structure profile. Hence, a trade-off between taxonomy assignment accuracy and extent of deviation in relative abundance has to be found to achieve more accurate microbial community structure profiling. According to the results obtained in this sequencing experiment, there is no distinct difference in community structures with quality threshold set at Q7, Q8 and Q9. Thus, setting a quality score threshold at 7 was believed to be capable of providing reliable microbial community structure profile, though further polishing might be needed to improve the species level identification accuracy.

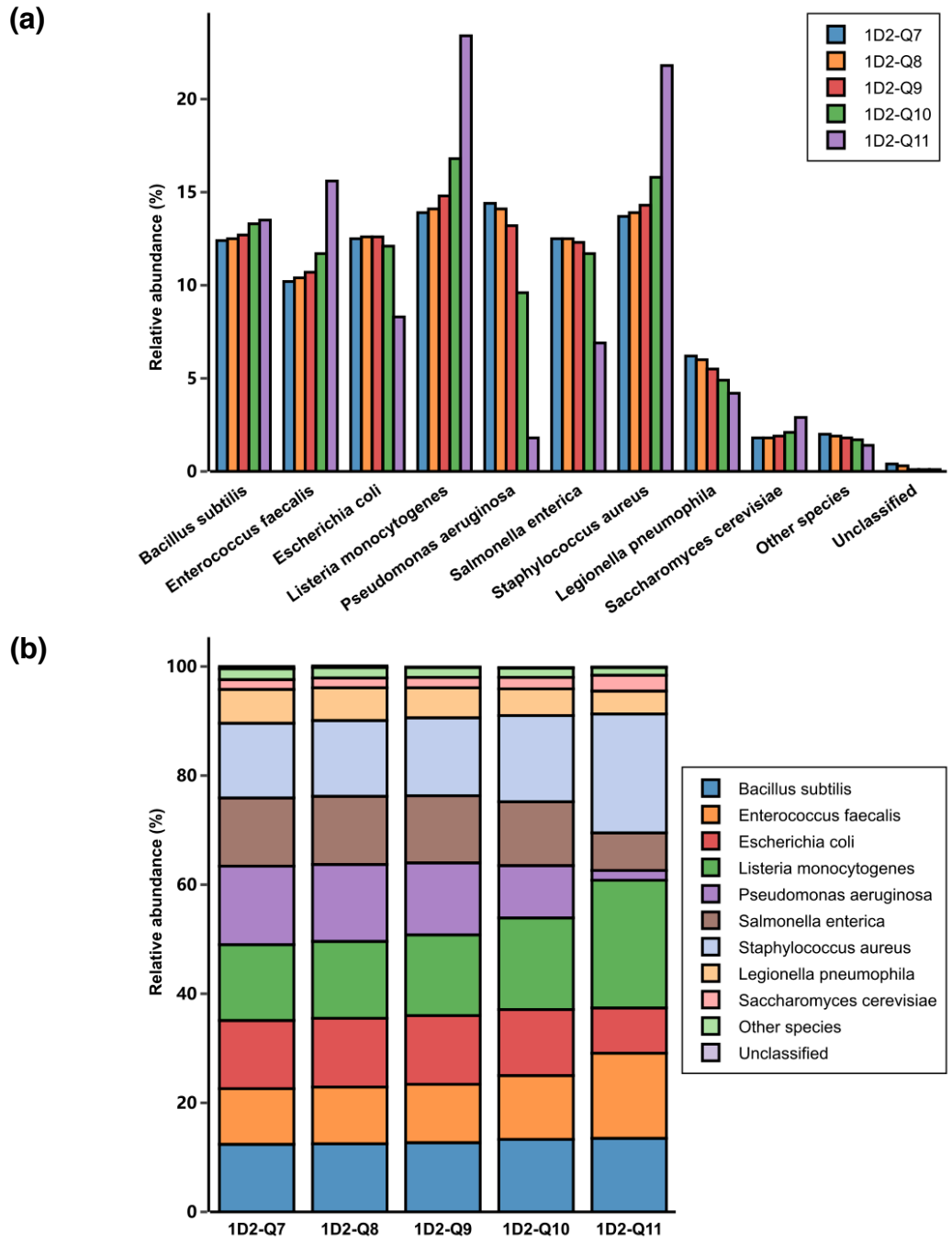

Figure 3.8 Community composition of the mock microbial community revealed by 1D^2 data with quality score threshold set at 7, 8, 9, 10 and 11

Table 3.10 Number of passed 1D^2 reads in MinION^{TM} genomic DNA sequencing with different quality score threshold

Quality threshold	≥Q7	≥Q8	≥Q9	≥Q10	≥Q11
Read count (%)	348,244 (75.6%)	311,903 (67.7%)	256,726 (55.7%)	168,929 (36.7%)	56,083 (12.2%)

3.3.1.3 Effect of data polishing on microbial community profiling

In order to eliminate the presence of false positives and unclassifiable reads in taxonomic classification, the 1D^2 reads were further polished with Racon to generate highly accurate sequences. A total of 379 consensus reads were obtained from the 460,617 1D^2 reads. Those consensus reads were assigned to 9 species by WIMP workflow, as was shown in Table 3.11. Hence, all of the Racon polished reads were correctly classified at species level without any false positives, false negatives or unclassified reads. Nevertheless, the results were no longer capable of quantifying the relative abundance of microbial community members. However, each species is expected to have only one assembly with this bioinformatics pipeline. The generation of multiple assembly reads for a given species is probably attributable to sequence error. It is likely that failure would occur when finding overlaps of the originally overlapping reads due to the high deviation of sequences. Thus, there is possibility that reads from the same species were wrongly differentiated and subsequently archived to different draft assemblies, hence leading to multiple final assemblies for one species. Besides, as singletons were removed during draft assembly construction, low-abundance species would be polished out and thus fail to give an assembly. As a consequence, the detection limit still needs to be determined.

Table 3.11 Taxonomy assignment of 1D^2 genomic DNA sequencing consensus reads polished by Racon

Species	Cumulative reads
Bacillus subtilis	12
Enterococcus faecalis	9
Escherichia coli	84
Listeria monocytogenes	4
Pseudomonas aeruginosa	24
Salmonella enterica	102
Staphylococcus aureus	4
Legionella pneumophila	98
Saccharomyces cerevisiae	42

3.3.1.4 Effect of sequencing duration on sequencing throughput and microbial community profiling

In order to assess the feasibility of shortening the sequencing duration to achieve rapid microbial community profiling, $1D^2$ sequencing data of the first 5 hours was extracted from the whole dataset of the 45 hours' sequencing for analysis. During the first 5 hours, 90,792 of $1D^2$ reads comprising 669 Mbp of nucleotide bases were sequenced, which accounted for 19.7% (90,792 of 460,617) and 19.1% (669 of 3,500) of those of the entire sequencing run (Table 3.12). The mean and median $1D^2$ read length of the first 5 hours were slightly lower than those of 45 hours (mean: 7,364 bp vs 7,599 bp; median: 7,321 bp vs 7,543 bp). Interestingly, both mean and median quality score were higher for the first 5 hours as compared with those of the entire sequencing run (mean: 8.9 vs 8.5; median: 9.9 vs 9.4). Moreover, the $1D^2$ reads of the first 5 hours were subjected to WIMP workflow for taxonomy classification with quality score threshold set at 7. The proportion of reads passing the quality score filter of the first 5 hours' sequencing was 79.0%, which was also higher than that of the entire 45 hours' sequencing run (75.6%).

Table 3.12 Statistics of mock community genomic DNA $1D^2$ sequencing data of 5 hours and 45 hours

Duration	$1D^2$ read count	Number of passed reads (≥Q7)	Percentage of passed reads (%)	Number of bases (Mbp)	Read length (bp)		Quality score	
					Mean	Median	Mean	Median
5 h	90,792	71,702	79.0	669	7,364	7,321	8.9	9.9
45 h	460,617	348,244	75.6	3,500	7,599	7,543	8.5	9.4

The structure of the mock microbial community profiled by the two datasets are compared in Figure 3.9. Histogram showed that there is only negligible differences between the relative abundance of the 9 species revealed by sequencing data of the first 5 hours' and the entire 45 hours. Furthermore, after polishing with Racon, 325 of consensus reads were generated out of the 90,792 $1D^2$ reads, which was further classified into 10 species (Table 3.13). However, *Legionella pneumophila* was not included in the 10 detected species while three of the consensus reads were classified as other species (i.e. *Cronobacter malonaticus* and *Shigella flexneri*), which were not identified by 45 hours' polished data. Considering the sequence similarity of the species, *Cronobacter malonaticus* might be misclassified from the reads of *Enterococcus faecalis*. Similarly, identification of *Shigella flexneri* could arise from misclassification of reads of *Escherichia coli*. As read polishing is realized through consensus calling, the extent of accuracy improvement by polishing is dependent on the sequencing coverage (Logan et al., 2014). The presence of misclassified reads in 5 hours' polished reads indicated that the sequencing coverage of 5 hours' sequencing could not completely

compensate for the sequencing error. Hence, 5 hours' sequencing is not sufficient for accurate species identification of mixed microbial communities.

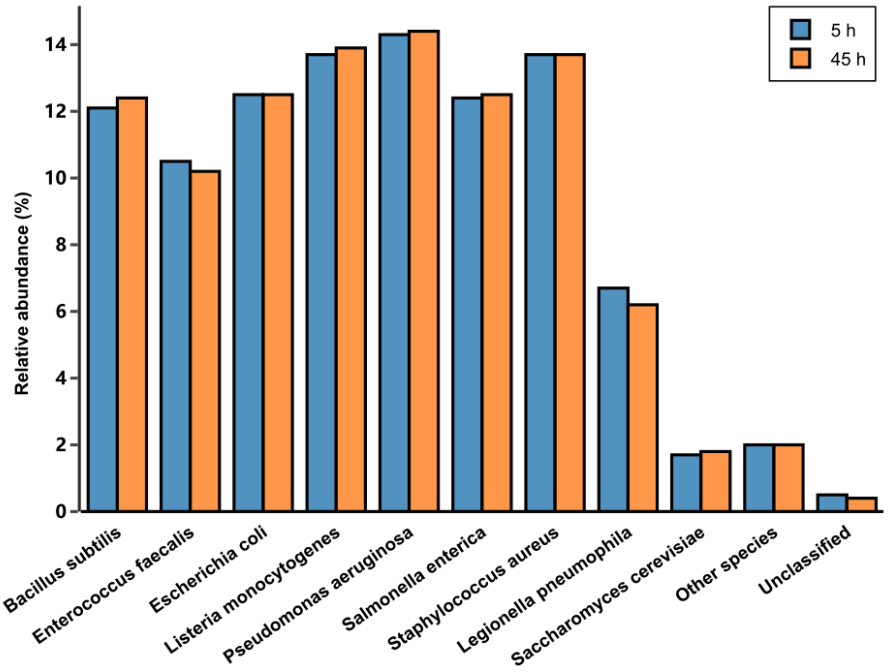

Figure 3.9 Bar chart of observed relative abundance of species in the constructed microbial community with MinION™ genomic DNA 1D² sequencing data of 5 hours and 45 hours

Table 3.13 Taxonomy assignment of consensus reads generated from the first 5 hours' mock community genomic DNA sequencing 1D² reads polished by Racon

Species	Cumulative reads
Bacillus subtilis	41
Enterococcus faecalis	18
Escherichia coli	82
Listeria monocytogenes	22
Pseudomonas aeruginosa	52
Salmonella enterica	85
Staphylococcus aureus	18
Saccharomyces cerevisiae	1
Legionella pneumophila	0
Cronobacter malonaticus	1
Shigella flexneri	2

3.3.2 Tap water DNA sequencing

Triplicate gDNA samples of tap water microbial community were mixed together after 0.4× AMPure XP beads size selection. Subsequently, the mixed gDNA sample was diluted to 30 ng/μL using nuclease-free water and subjected to MinION™ $1D^2$ sequencing on a FLO-MIN107 flow cell after sequencing library preparation as described in Appendix B.4 (gTUBE fragmentation was not performed on tap water gDNA sample). The sequencing run started with 1365 available pores and lasted for 48 hours. A total of 7,410,986 1D reads were generated during this sequencing run, of which 18.6% (1,380,547 reads) were recognized as $1D^2$ reads. The total 1D reads comprises of 11.5 Gbp of bases while the $1D^2$ reads contains 2.4 Gbp of DNA sequence. The maximum sequence length was 564,641 bp, which was substantially higher than those of the other 5 sequencing runs because no intended fragmentation was involved in library preparation. However, both mean and median read length (1,550 bp and 1,820 bp for 1D reads; 1,734 bp and 1,184 bp for $1D^2$ reads) were significantly lower than those of other sequencing runs. Although there exists several ultra-long 1D reads longer than 100 kbp, a major part of the reads were below 4 kbp. The shorter reads in tap water DNA samples might result from occurrence of strong fragmentation during DNA extraction. Besides, most of the ultra-long reads were of low quality ($Q<6$) and did not end up in $1D^2$ reads (Figure 3.10). The mean and median quality score of 1D reads were markedly lower than those of mock community gDNA sequencing run (Mean: 7.8 vs 8.5; Median: 8.2 vs 9.4), which could be attributable to lower quality of input DNA and existence of inhibitors in environmental samples. Besides, prevalence of short reads also contributed to the low quality scores (Krishnakumar et al., 2018). Unexpectedly, both the mean and median quality score saw a pronounced drop after $1D^2$ basecall. It can be seen from Figure 3.11 that the quality score of the $1D^2$ reads of this sequencing run shows a bimodal distribution. Besides the normal peak at 8.8, there is also a distinct peak at around 6, which was not observed in other sequencing runs. The presence of this unwanted peak in quality score distribution is owing to the prevalence of false positive pairings during $1D^2$ basecall. When comparing with mock community gDNA sequencing run, the reads generated in tap water gDNA sequencing run showed a much narrower sequence length distribution which peaked at around 1 kbp (Figure 3.12). Accordingly, there were higher risks that reads of equal length but from different dsDNA molecules were paired for consensus basecalling. As a major part of microbes in drinking water are closely related, their partial gene sequences are also similar to each other's, hence boosting the chances of false positive pairings not differentiated during pre-alignment. As a consequence, those falsely paired but not identical reads that successfully passed the pre-alignment stage would generate $1D^2$ reads with a lower quality score than the original reads, which contributed to the formation of the bimodal distribution.

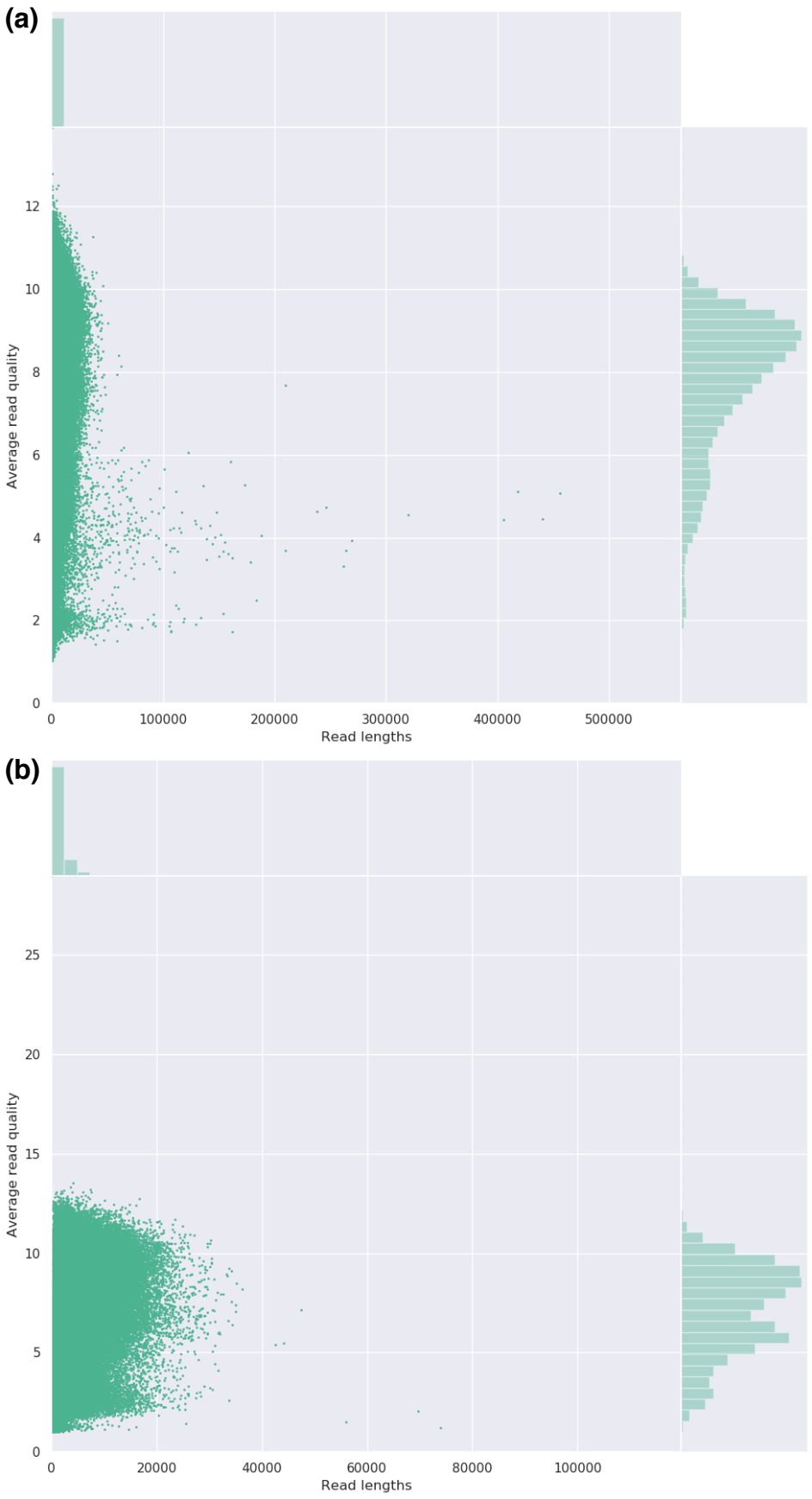

Figure 3.10 Read length versus quality score scatter plots of tap water genomic DNA sequencing (a) 1D and (b) $1D^2$ reads

Table 3.14 Statistics of tap water genomic DNA sequencing data

Dataset	Read count	Number of bases (Gbp)	Read length (bp)			Quality score	
			Mean	Median	Mode	Mean	Median
1D	7,410,986	11.5	1,550	1,820	565	7.8	8.2
$1D^2$	1,380,547	2.4	1,734	1,184	840	7.2	7.5

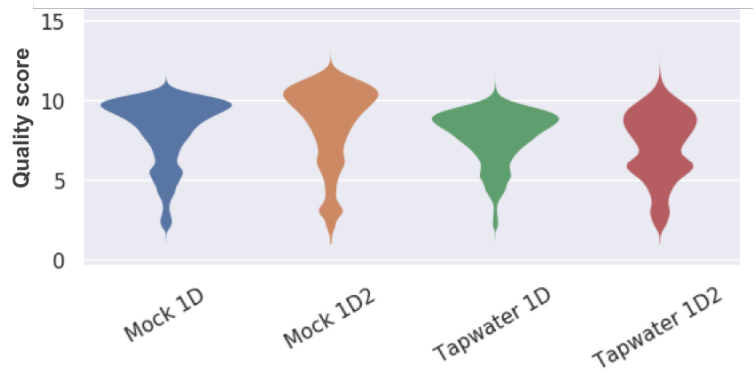

Figure 3.11 Violin plots of quality score distribution of 1D and $1D^2$ reads of two genomic DNA sequencing runs

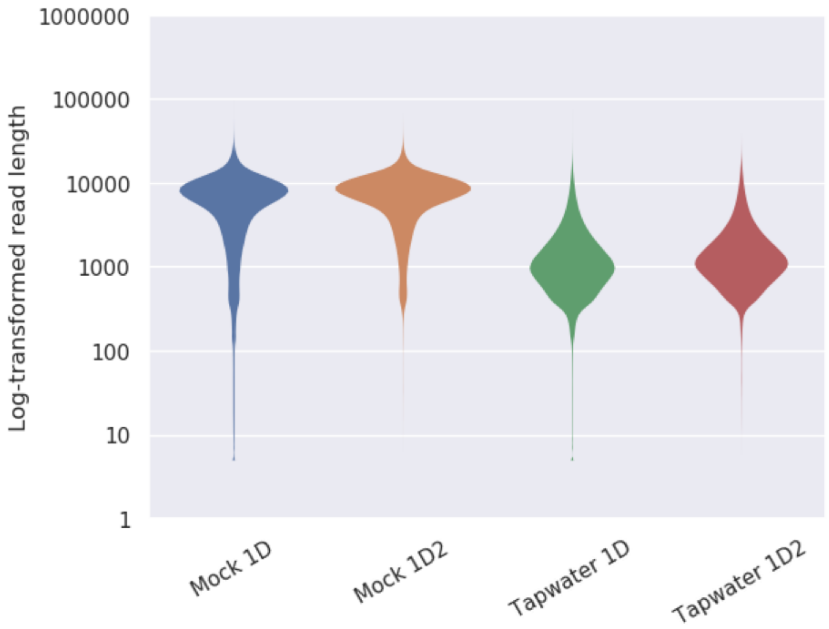

Figure 3.12 Violin plots of log-transformed read length distribution of 1D and $1D^2$ reads of two genomic DNA sequencing runs

The $1D^2$ reads were subjected to WIMP taxonomy classification in EPI2ME with the quality score threshold set at 7 for estimation of microbial community structure. 55.6% (767,624 of 1,380,547) of the $1D^2$ reads have passed the quality score filter and were passed to WIMP. However, only 25.1% (192,914 of 767,624) of the passed reads were assigned to a taxonomy while the rest 74.9% were unclassified. The 192,914 reads were classified into a total of 3303 species. 93.5% (180,454 of 192,914) of the classified reads were recognized as bacteria, 5.0% (9,640 of 192,914) was classified as eukaryota, and the rest were identified as archaea (1,111 reads) and viruses (213 reads). The community composition consisting of the most abundant species (relative abundance ≥0.1%) is shown in Figure 3.13. As a considerable amount of falsely paired reads from highly similar species were left out due to the low quality scores, there was a likelihood of generation of biased community structure profile. Furthermore, given the lowest raw read accuracy of 80% (i.e. Q7), there still remains high probability of occurrence of misclassification across closely related species. As a result, the community structure would be more reliably estimated at higher taxonomic ranks. In addition, it should be noted that *Homo sapiens* was detected at 0.30% relative abundance, which is however not expected to be present in tap water. This unexpected presence of *Homo sapiens* could be explained by contamination from experiment operator during sample preparation (i.e. tap water sample collection, tap water filtration, DNA extraction, and sequencing library preparation). The top 1000 abundant species classified in this experiment are shown in Appendix A (Table A.7).

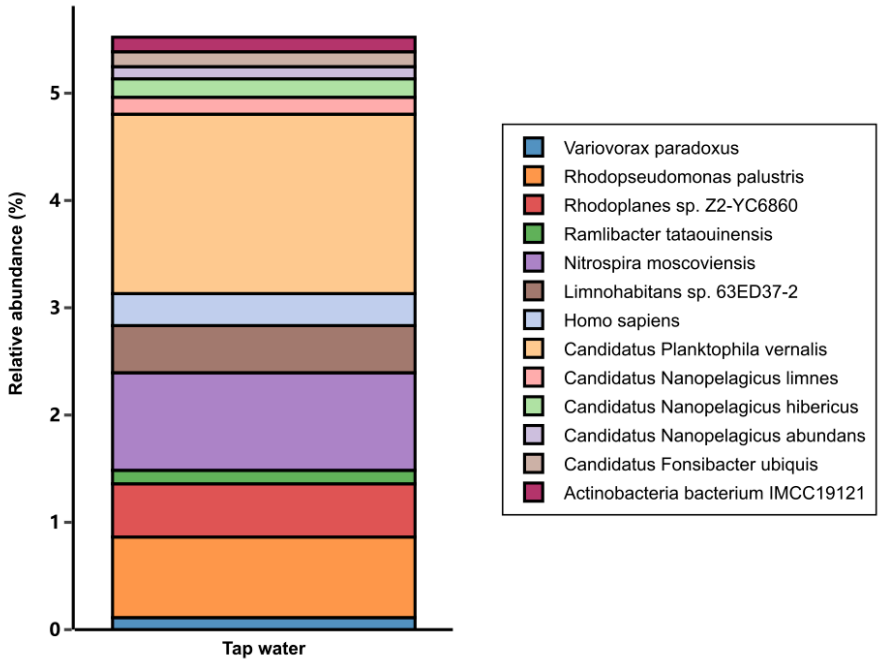

Figure 3.13 Taxonomy assignment of tap water genomic DNA $1D^2$ sequencing reads (only species with relative abundance ≥0.1% were plotted in the chart)

In order to obtain accurate species estimation, the $1D^2$ reads were polished using Racon. A total of 106 read assemblies were generated out of the 1,380,547 $1D^2$ reads and all of them were successfully assigned to a taxonomy. 105 of the consensus reads were classified at species level while the rest 1 read was assigned at phylum level as Proteobacteria. However, it is likely that reads of low-abundance species did not end up in the final assembly due to singleton removal during polishing, thus leading to underestimation of community diversity. The species composition revealed by polished $1D^2$ reads was given in Table 3.15. Almost all of the 20 detected species are harmless bacteria that were commonly found in aquatic environment or

soil, except for *Laribacter hongkongensis*, which is a pathogenic bacterial strain related to gastroenteritis and travellers' diarrhea (Engsbro et al., 2018), and *Enterobacter cloacae*, which was found as a nosocomial pathogen responsible for various infectious diseases (Mezzatesta et al., 2012). Nonetheless, as was discussed in tap water gDNA sequencing raw results, there existed post sample contamination in the sequencing library. Therefore, the identification of those OPs does not conclusively indicate that the original tap water was inhabited by them. Moreover, there was also possibility that the reads were misclassified from their closely related species due to the remaining error that was not resolved by polishing.

Table 3.15 Species composition of tap water microbial community obtained from polished 1D^2 reads

Species	Cumulative reads
Acidovorax avenae	1
Acidovorax sp. T1	1
Candidatus Fonsibacter ubiquis	34
Cupriavidus sp. USMAA2-4	1
Delftia sp. HK171	1
Diaphorobacter polyhydroxybutyrativorans	1
Enterobacter cloacae	1
Hydrogenophaga sp. PBC	1
Hydrogenophage sp. RAC07	2
Laribacter hongkongensis	1
Leptothrix cholodnii	1
Limnohabitans sp. 63ED37-2	48
Nitrospirillum amazonense	1
Polaromonas naphthalenivorans	1
Pseudomonas sp. CCOS 191	1
Ramlibacter tataouinensis	3
Roseateles depolymerans	1
Variovorax boronicumulans	2
Variovorax paradoxus	2
Variovorax sp. PAMC 28711	1

3.4 Comparison of tap water microbial community profiling results obtained by different sequencing strategies on MinION^{TM} platform

Statistics of taxonomy assignment of tap water microbial community by NanoAmpli-Seq 16S rRNA gene sequencing data, 1D^2 gDNA sequencing raw read data and 1D^2 gDNA sequencing polished data were compared in Figure 3.14. NanoAmpli-Seq 16S amplicon sequencing identified 6 phyla, 11 classes, 13 orders, 13 families, 9 genus, and 10 species. Genomic DNA 1D^2 sequencing raw data identified 41 phyla, 93 classes, 209 orders, 446 families, 1199 genus, and 3303 species. While genomic DNA 1D^2 sequencing polished data identified 1 phylum, 3

classes, 6 orders, 6 families, 16 genus, and 20 species. All of the taxa identified by gDNA sequencing polished $1D^2$ data were also observed by raw $1D^2$ reads. Species *Geobacter metallireducens*, *Pirellula staleyi*, and *Rhodoferax ferrireducens* were detected by both 16S amplicon sequencing and gDNA sequencing raw data, but were absent from taxonomy classification result of polished gDNA sequencing reads. Moreover, *Geobacter spp.*, *Helicobacter spp.*, *Methylobacter spp.*, *Nitrospira spp.*, *Pirellula spp.*, and *Rhodoferax spp.* are genera that were observed by both 16S amplicon sequencing and gDNA sequencing raw data. When going to superior taxa, the taxonomic units identified by 16S amplicon sequencing were all included in the result obtained from gDNA sequencing raw data. While comparing at family level, *Burkholderiaceae* and *Comamonadaceae* were recognized by both 16S amplicon sequencing and gDNA sequencing polished reads. Furthermore, the three sets of sequencing data shared 1 order (*Burkholderiales*). All of the 3 classes (*Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria*) and the only phylum (*Proteobacteria*) identified by gDNA polished data were also identified in 16S amplicon sequencing. Hence, this comparative analysis indicated that none of the three sequencing strategies were reliable enough to accurately characterize tap water microbiome at species or genus level. However, the raw $1D^2$ data of gDNA sequencing may offer unbiased microbial community profiles with family level resolution.

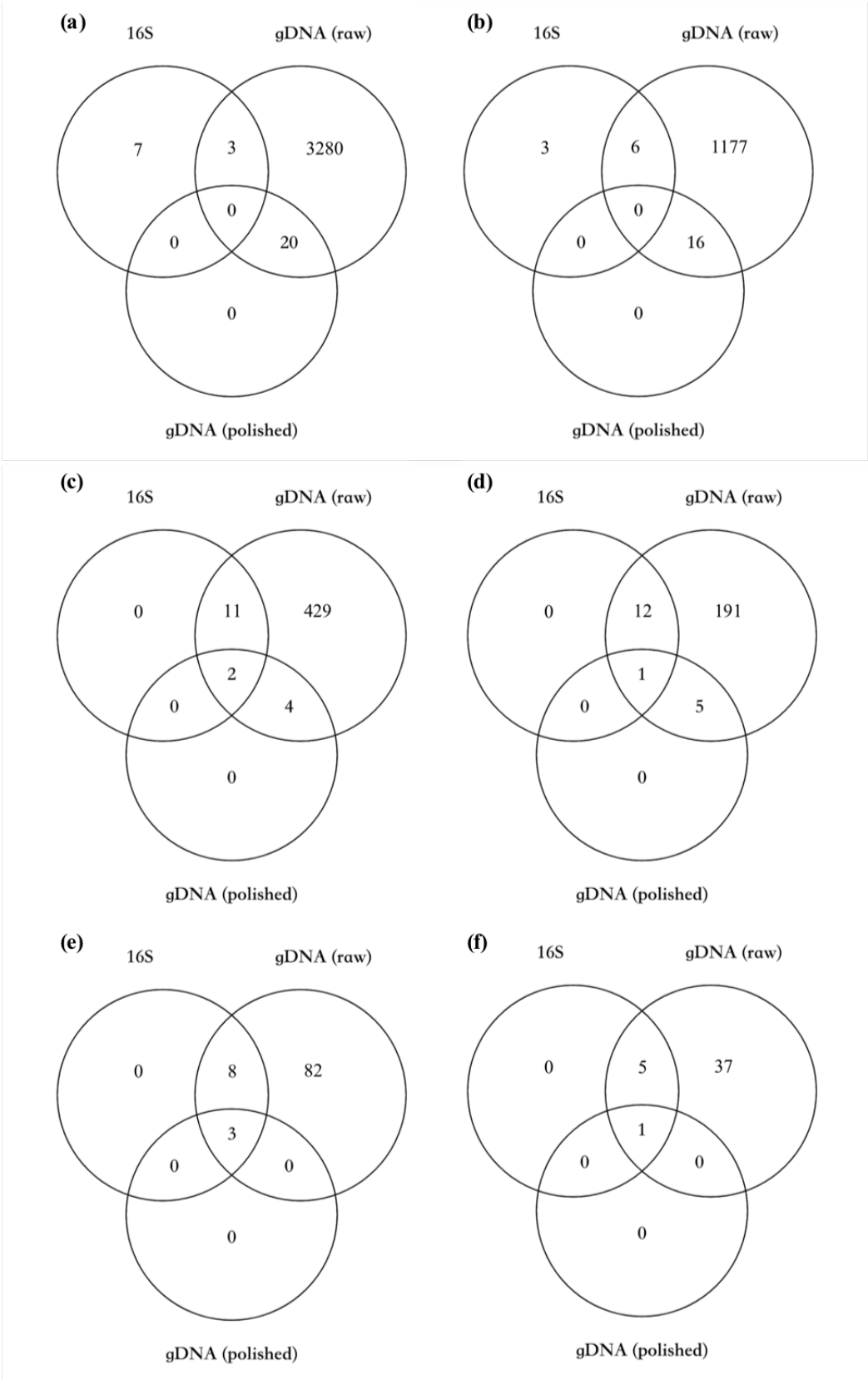

Figure 3.14 Venn diagrams showing the number of shared and specific taxonomic units at (a) species, (b) genus, (c) family, (d) order, (e) class, and (f) phylum level among 16S amplicon sequencing, genomic DNA raw 1D^2 sequencing, and genomic DNA polished 1D^2 sequencing data of tap water

4 Conclusion and recommendation

4.1 Conclusion

This research aimed at proposing a dependable approach for characterizing the microbiome of drinking water samples with extremely low biomass using Oxford Nanopore MinION™ sequencer. In this study, the accuracy of species identification and microbial community profiling for two DNA sequencing strategies, i.e. NanoAmpli-Seq sequencing of full-length 16S amplicons and $1D^2$ sequencing of gDNA, was assessed on MinION™ sequencer using artificial microbial community DNA samples. Moreover, DNA samples from tap water were extracted and sequenced based on the two methods. The results for species identification and microbial community profiling obtained from the two sequencing strategies were compared with each other.

NanoAmpli-Seq full-length 16S amplicon sequencing is capable of precisely identifying bacteria at species level in artificial microbial communities. However, the detection limit still needs to be determined. Additionally, with the presence of PCR amplification in library preparation, it is unable to accurately quantify the relative abundance of each bacteria strain in all samples, which might be contributed by PCR biases related to GC content. Moreover, with two rounds of PCR amplification (i.e. 16S rRNA gene PCR amplification and RCA amplification), the PCR biases tends to multiply and accumulate in the resulting sequencing library, further potentially giving rise to generation of spurious OTUs in the final taxonomy analysis. Notably, though species identification was realized in tap water samples, most of the reads fell into higher-rank taxa with only 23.4% classified to 10 species, which might result from the possible multi-species binning in complicated environmental samples with closely related species. This suggested that NanoAmpli-Seq might not be applicable for classifying bacteria in complex microbial communities at species level. Nevertheless, it should be noted that high level of reproducibility of the relative abundance of bacterial strains from Zymo was observed among the three sequencing runs (Zymo, Z95L5, Z90L10), indicating the possibility for the method acting as a robust approach to rapidly monitor microbial community changes during environmental changes (e.g. source water switching in drinking water systems).

Similarly, direct gDNA sequencing successfully identified bacteria at species level in artificial microbial community. Remarkably, high accuracy of species relative abundance in this mock microbial community was obtained from its raw $1D^2$ sequencing results even with default quality score threshold of 7, suggesting direct gDNA sequencing could be a more reliable approach for profiling microbial communities than NanoAmpli-Seq, notwithstanding 2% of the reads remaining unclassified or misclassified. Interestingly, the unclassified and misclassified reads were eliminated after data polishing, but the relative abundance of species was unable to be depicted. In addition, longer sequencing duration (45 hours) can not only dramatically increase the data throughput, but also significantly improve the species identification accuracy when compared with the results from 5 hours'. However, species identification and microbial community characterization in tap water by gDNA sequencing was still unsatisfactory, despite that more species were identified in tap water by gDNA sequencing than NanoAmpli-Seq, even with two assembly classified as pathogen *Enterobacter cloacae* and *Laribacter hongkongensis*. This could be explained by the presence of predominant short fragment and high similarity of sequences of DNA extracted from tap water. Conceivably, direct gDNA sequencing is a promising method to rapidly and precisely characterize microbial communities in environmental samples. Nevertheless, more efforts

should be made to overcome the challenges for gDNA sequencing (e.g. extracting high amount of HMW gDNA, standard method for accurately analyzing the sequencing data) to make it a powerful tool for species identification and microbial community characterization in complex environmental samples associated with low biomass.

4.2 Recommendations

4.2.1 Recommendation on future work

We would like to address a few difficulties that we came into when characterizing drinking water microbiome using MinION™ and defects of the sequencing approaches that were deployed, as well as make some suggestions on improvements to be made in future work as follows.

First of all, as drinking water has a low biomass content, concentrating microbes by filtrating a large volume of water is required to obtain sufficient quantity of DNA for Nanopore gDNA sequencing. However, if the filtration process is performed manually, it would be both labor-intensive and time-consuming. Moreover, manual filtration in a nonsterile environment would also give rise to risks of contamination in the samples. Therefore, one possibility to circumvent these issue is to conduct drinking water filtration with automated apparatus, for example, online particle sampling system (OPSS), which is originally designed to enable 24-hours continuous sampling and monitoring of drinking water distribution system (X. Li, 2017). Moreover, DNA extraction using most of the commercial DNA isolation kits involves vigorous bead-beating and shaking, thus resulting in highly fragmented DNA, which would undermine the subsequent Nanopore sequencing performance. Hence, more gentle DNA extraction protocols must be employed to avoid unwanted shearing. In addition, size selection based on gel electrophoresis could be adopted to select for fragments larger than 8 kbp. However, this will in turn pose high demands on starting material considering the sharp decline in total DNA amount caused by elimination of short fragments.

Regarding NanoAmpli-Seq 16S rRNA gene sequencing, the main issue lies in OTU clustering of the closely related species. Thus, improvement of OTU clustering and consensus sequence construction algorithm of nanoClust program might be needed in order to fit this sequencing strategy for microbiome characterization of environmental samples. Moreover, optimization in library preparation could also help resolving this issue (Calus, 2018). Furthermore, tap water microbial community profiling through $1D^2$ gDNA sequencing also suffered from high similarity among reads. As the similar reads primarily affect $1D^2$ base calling, this problem has to be solved by improvement of base calling algorithms. Another possibility for addressing this issue is to go for 1D sequencing. ONT has recently released R10 flow cells for 1D sequencing, claiming to be able to achieve 99.999% consensus accuracy, which is comparable to the consensus accuracy of $1D^2$ reads. Thus, the potential of microbial community profiling by gDNA sequencing on R10 flow cells could be explored in future work. In addition, the detection limit of both sequencing strategies has to be verified by additional sequencing experiments to address their applicability to identification of microbes at low abundance.

4.2.2 Recommendation on choice of MinION™ sequencing strategies

Undoubtedly, there is no "one size fits all" approach for microbiological research. By comparing the merits and demerits of both sequencing strategies, we proposed following recommendations on choices of Nanopore sequencing strategies and critical aspects to take into consideration for different application scenarios.

Scenario 1: adequate high quality (i.e. total amount ≥ 1.5 µg and average fragment length ≥ 8 kbp) starting material is available.

Both sequencing strategies exhibit satisfying performance on high quality DNA samples. However, genomic DNA sequencing is favored over amplicon sequencing thanks to its simplicity in both library preparation and subsequent data processing. In addition, absence of PCR amplification also contributed to unbiased microbial community profiles, thus making it superior to the PCR-based NanoAmpli-Seq. When high sequencing throughput in a relatively short period of time is demanded (e.g. initial screening inspection of microbial community changes, OPs or ARGs in drinking water), it is recommended to use 1D sequencing notwithstanding the relatively high raw read error rate. As ONT 1D R9.4.1 sequencing chemistry was reported to achieve consensus accuracy of over 99.9% (Bowden et al., 2019), it is believed to be capable of providing promising microbial community profiling result. Otherwise if the aim is to precisely identify species in a microbial community, $1D^2$ chemistry is preferred due to its greater raw read accuracy as compared to 1D reads (96% versus 90%, claimed by ONT). Since the post-polish accuracy is dependent on the accuracy of pre-polish reads (Wick et al., 2019), $1D^2$ reads is expected to generate higher consensus accuracy than current 1D R9.4.1 chemistry. However, the recently released R10 flow cells could achieve 99.999% consensus accuracy, thus provide an attractive alternative for those who pursue high sequencing accuracy and simplicity in both library preparation and data processing at the same time. In terms of choice of sequencing duration, if one aims to roughly profile microbial communities, $1D^2$ basecalled reads from several hours' sequencing would be sufficient. For more accurate species classification and pathogen identification, 48 hours' or even longer $1D^2$ sequencing run with $1D^2$ basecalled data polishing could be a good option.

Scenario 2: sample amount is limited (i.e. starting DNA significantly less than 1.5 µg) or DNA is strongly fragmented (i.e. average length significantly shorter than 8 kbp).

As short fragment size of input DNA would have an adverse effect on Nanopore sequencing performance, NanoAmpli-Seq 16S amplicon sequencing is recommended in this case. As long as the 16S rRNA genes in the original DNA sample are unsheared, the PCR amplification will function properly regardless of the fragment size distribution. Thus, a Nanopore sequencing run can be carried out successfully as expected. Thanks to its high reproducibility, NanoAmpli-Seq could be used for routine monitoring of community composition and assessing the changes in bacterial communities during treatment or during distribution. It would also be useful for indicating the drinking water bacterial community composition change due to treatment modifications or source water change.

Bibliography

Allen, M. J., Edberg, S. C., & Reasoner, D. J. (2004). Heterotrophic plate count bacteria—what is their significance in drinking water? *International journal of food microbiology, 92*(3), 265-274.

Armstrong, J. L., Shigeno, D. S., Calomiris, J., & Seidler, R. J. (1981). Antibiotic-resistant bacteria in drinking water. *Appl. Environ. Microbiol., 42*(2), 277-283.

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., & O'grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology, 33*(3), 296.

Bai, X., Ma, X., Xu, F., Li, J., Zhang, H., & Xiao, X. (2015). The drinking water treatment process as a potential source of affecting the bacterial antibiotic resistance. *Science of the Total Environment, 533*, 24-31.

Bartram, J., Cotruvo, J., Exner, M., Fricker, C., & Glasmacher, A. (2003). *Heterotrophic plate counts and drinking-water safety*: IWA publishing.

Bartram, J., Cotruvo, J., Exner, M., Fricker, C., & Glasmacher, A. (2004). Heterotrophic plate count measurement in drinking water safety management: report of an Expert Meeting Geneva, 24–25 April 2002. *International journal of food microbiology, 92*(3), 241-247.

Bartrand, T. A., Causey, J. J., & Clancy, J. L. (2014). Naegleria fowleri: An emerging drinking water pathogen. *Journal-American Water Works Association, 106*(10), E418-E432.

Bédard, E., Laferrière, C., Déziel, E., & Prévost, M. (2018). Impact of stagnation and sampling volume on water microbial quality monitoring in large buildings. *PloS one, 13*(6), e0199429.

Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *Gigascience, 5*(1), 4.

Benítez-Páez, A., & Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. *Gigascience, 6*(7), gix043.

Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., de Cesare, M., Oikkonen, L. E., Parkes, D., Freeman, C., Dhalla, F., & Patel, S. Y. (2019). Sequencing of human genomes with nanopore technology. *Nature communications, 10*(1), 1869.

Bridle, H. (2013). *Waterborne pathogens: detection methods and applications*: Newnes.

Bridle, H., Gilbride, K., & Green, J. (2014). Chapter Three - Existing Methods of Waterborne Pathogen Detection. In H. Bridle (Ed.), *Waterborne Pathogens* (pp. 41-64). Amsterdam: Academic Press.

Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C., & Franklin, R. B. (2017). MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience, 6*(3), gix007.

Bruno, A., Sandionigi, A., Bernasconi, M., Panio, A., Labra, M., & Casiraghi, M. (2018). Changes in the drinking water microbiome: effects of water treatments along the flow of two drinking water treatment plants in a urbanized area, Milan (Italy). *Frontiers in microbiology, 9*, 2557.

Calus, S. T. (2018). *Evaluation of nanopore-based sequencing technology for gene marker based analysis of complex microbial communities. Method development for accurate 16S rRNA gene amplicon sequencing.* University of Glasgow,

Calus, S. T., Ijaz, U. Z., & Pinto, A. J. (2018). NanoAmpli-Seq: A workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience, 7*(12), giy140.

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B., Dworkin, J. P., Lupisella, M. L., Smith, D. J., & Botkin, D. J. (2017). Nanopore DNA sequencing and genome assembly on the International Space Station. *Scientific reports, 7*(1), 18022.

Cope, J. R., Ratard, R. C., Hill, V. R., Sokol, T., Causey, J. J., Yoder, J. S., Mirani, G., Mull, B., Mukerjee, K. A., & Narayanan, J. (2015). The first association of a primary amebic meningoencephalitis death with culturable Naegleria fowleri in tap water from a US treated public drinking water system. *Clinical Infectious Diseases, 60*(8), e36-e42.

Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., Calderon, R. L., Roberts, J. M., Beach, M. J., & Roy, S. L. (2010). Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. *Clinical Microbiology Reviews, 23*(3), 507-528.

Cuscó, A., Catozzi, C., Viñes, J., Sanchez, A., & Francino, O. (2018). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. *F1000Research, 7*.

Cusco, A., Vines, J., D'Andreano, S., Riva, F., Casellas, J., Sanchez, A., & Francino, O. (2017). Using MinION to characterize dog skin microbiota through full-length 16S rRNA gene sequencing approach. *bioRxiv*, 167015.

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome research*.

de Lannoy, C., De Ridder, D., & Risse, J. (2017). A sequencer coming of age: de novo genome assembly using MinION reads. *F1000Research, 6*.

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature biotechnology, 34*(5), 518.

Delafont, V., Brouke, A., Bouchon, D., Moulin, L., & Héchard, Y. (2013). Microbiome of free-living amoebae isolated from drinking water. *Water research, 47*(19), 6958-6965. doi:https://doi.org/10.1016/j.watres.2013.07.047

Edwards, A., Debbonaire, A. R., Nicholls, S. M., Rassner, S. M., Sattler, B., Cook, J. M., Davy, T., Mur, L. A., & Hodson, A. J. (2018). In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. *bioRxiv*, 073965.

Engsbro, A. L., Nielsen, K. L., Hornum, M., & Andersen, L. P. (2018). Laribacter hongkongensis: clinical presentation, epidemiology and treatment. A review of the literature and report of the first case in Denmark. *Infectious Diseases, 50*(6), 417-422.

Falkinham III, J. O., Hilborn, E. D., Arduino, M. J., Pruden, A., & Edwards, M. A. (2015). Epidemiology and ecology of opportunistic premise plumbing pathogens: Legionella pneumophila, Mycobacterium avium, and Pseudomonas aeruginosa. *Environmental health perspectives, 123*(8), 749-758.

Falkinham, J. (2015). Common features of opportunistic premise plumbing pathogens. *International journal of environmental research and public health, 12*(5), 4533-4545.

Falkinham, J., Pruden, A., & Edwards, M. (2015). Opportunistic premise plumbing pathogens: increasingly important pathogens in drinking water. *Pathogens, 4*(2), 373-386.

Goordial, J., Altshuler, I., Hindson, K., Chan-Yam, K., Marcolefas, E., & Whyte, L. G. (2017). In situ field sequencing and life detection in remote (79° 26′ N) Canadian high Arctic permafrost ice wedge microbial communities. *Frontiers in microbiology, 8*, 2594.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., & Linnen, J. M. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome medicine, 7*(1), 99.

Günther, T., Haas, L., Alawi, M., Wohlsein, P., Marks, J., Grundhoff, A., Becher, P., & Fischer, N. (2017). Recovery of the first full-length genome sequence of a parapoxvirus directly from a clinical sample. *Scientific reports, 7*(1), 3734.

Hammes, F., Berney, M., Wang, Y., Vital, M., Köster, O., & Egli, T. (2008). Flow-cytometric total bacterial cell counts as a descriptive microbiological parameter for drinking water treatment processes. *Water research, 42*(1-2), 269-277.

Harstad, H., Ahmad, R., & Bredberg, A. (2018). Nanopore-based DNA sequencing in clinical microbiology: preliminary assessment of basic requirements. *bioRxiv*, 382580.

Hong, W., Bedard, E., Prevost, M., Camper, A. K., Hill, V. R., & Pruden, A. (2017). Methodological approaches for monitoring opportunistic pathogens in premise plumbing: A review. *Water research, 117*, 68-86.

Hörman, A., Rimhanen-Finne, R., Maunula, L., von Bonsdorff, C.-H., Torvela, N., Heikinheimo, A., & Hänninen, M.-L. (2004). Campylobacter spp., Giardia spp., Cryptosporidium spp., noroviruses, and indicator organisms in surface water in southwestern Finland, 2000-2001. *Appl. Environ. Microbiol., 70*(1), 87-95.

Hu, Y. O., Ndegwa, N., Alneberg, J., Johansson, S., Logue, J. B., Huss, M., Käller, M., Lundeberg, J., Fagerberg, J., & Andersson, A. F. (2018). Stationary and portable sequencing-based approaches for tracing wastewater contamination in urban stormwater systems. *Scientific reports, 8*(1), 11907.

Hull, N. M., Ling, F., Pinto, A. J., Albertsen, M., Jang, H. G., Hong, P.-Y., Konstantinidis, K. T., LeChevallier, M., Colwell, R. R., & Liu, W.-T. (2019). Drinking Water Microbiome Project: Is it Time? *Trends in Microbiology, 27*(8), 670-677. doi:10.1016/j.tim.2019.03.011

Hunter, P. (1997). *Waterborne disease: epidemiology and ecology*: John Wiley & Sons.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., & Fiddes, I. T. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology, 36*(4), 338.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology, 17*(1), 239.

Jia, S., Shi, P., Hu, Q., Li, B., Zhang, T., & Zhang, X.-X. (2015). Bacterial community shift drives antibiotic resistance promotion during drinking water chlorination. *Environmental science & technology, 49*(20), 12271-12279.

Johnson, A. P., & Woodford, N. (2013). Global spread of antibiotic resistance: the example of New Delhi metallo-β-lactamase (NDM)-mediated carbapenem resistance. *Journal of medical microbiology, 62*(4), 499-513.

Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., & Tighe, S. W. (2017). Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer. *Journal of Biomolecular Techniques: JBT, 28*(1), 2.

Judge, K., Harris, S. R., Reuter, S., Parkhill, J., & Peacock, S. J. (2015). Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy, 70*(10), 2775-2778.

Karamitros, T., van Wilgenburg, B., Wills, M., Klenerman, P., & Magiorkinis, G. (2018). Nanopore sequencing and full genome de novo assembly of human cytomegalovirus TB40/E reveals clonal diversity and structural variations. *BMC genomics, 19*(1), 577.

Khan, S., Knapp, C. W., & Beattie, T. K. (2016). Antibiotic resistant bacteria found in municipal drinking water. *Environmental Processes, 3*(3), 541-552.

Krishnakumar, R., Sinha, A., Bird, S., Jayamohan, H., Edwards, H., Schoeniger, J., Patel, K., Branda, S., & Bartsch, M. (2018). Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Scientific reports, 8*. doi:10.1038/s41598-018-21484-w

Laursen, M. F., Dalgaard, M. D., & Bahl, M. I. (2017). Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. *Frontiers in microbiology, 8*, 1934.

Lautenschlager, K., Boon, N., Wang, Y., Egli, T., & Hammes, F. (2010). Overnight stagnation of drinking water in household taps induces microbial growth and changes in community composition. *Water research, 44*(17), 4868-4877.

Learbuch, K., Lut, M., Liu, G., Smidt, H., & van der Wielen, P. (2019). Legionella growth potential of drinking water produced by a reverse osmosis pilot plant. *Water research, 157*, 55-63.

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany, 68*(20), 5419-5429.

Lemon, J. K., Khil, P. P., Frank, K. M., & Dekker, J. P. (2017). Rapid nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *Journal of clinical microbiology, 55*(12), 3530-3543.

Li, C., Chng, K. R., Hui Boey, E. J., Qi Ng, A. H., Wilm, A., & Nagarajan, N. (2016). INC-Seq: Accurate single molecule reads using nanopore sequencing. *bioRxiv*, 038042.

Li, R., Xie, M., Dong, N., Lin, D., Yang, X., Wong, M. H. Y., Chan, E. W.-C., & Chen, S. (2018). Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data. *Gigascience, 7*(3), gix132.

Li, X. (2017). *Assessing water quality changes and transition effects in distribution systems by online particle sampling system (OPSS).* Delft University of Technology, Retrieved from http://resolver.tudelft.nl/uuid:b44d2955-d8e7-45e8-9c55-384b49260fee

Ling, F., Whitaker, R., LeChevallier, M. W., & Liu, W.-T. (2018). Drinking water microbiome assembly induced by water stagnation. *The ISME journal, 12*(6), 1520.

Liu, G., Bakker, G., Li, S., Vreeburg, J., Verberk, J., Medema, G., Liu, W.-T., & Van Dijk, J. (2014). Pyrosequencing reveals bacterial communities in unchlorinated drinking water distribution system: an integral study of bulk water, suspended solids, loose deposits, and pipe wall biofilm. *Environmental science & technology, 48*(10), 5467-5476.

Liu, G., Lut, M., Verberk, J., & Van Dijk, J. (2013a). A comparison of additional treatment processes to limit particle accumulation and microbial growth during drinking water distribution. *Water research, 47*(8), 2719-2728.

Liu, G., Tao, Y., Zhang, Y., Lut, M., Knibbe, W.-J., van der Wielen, P., Liu, W., Medema, G., & van der Meer, W. (2017a). Hotspots for selected metal elements and microbes accumulation and the corresponding water quality deterioration potential in an unchlorinated drinking water distribution system. *Water research, 124*, 435-445.

Liu, G., Van der Mark, E., Verberk, J., & Van Dijk, J. (2013b). Flow cytometry total cell counts: a field study assessing microbiological water quality and growth in unchlorinated drinking water distribution systems. *BioMed research international, 2013*.

Liu, G., Verberk, J. Q. J. C., & Van Dijk, J. C. (2013c). Bacteriology of drinking water distribution systems: an integral and multidimensional review. *Applied Microbiology and Biotechnology, 97*(21), 9265-9276. doi:10.1007/s00253-013-5217-y

Liu, G., Zhang, Y., Knibbe, W.-J., Feng, C., Liu, W., Medema, G., & van der Meer, W. (2017b). Potential impacts of changing supply-water quality on drinking water distribution: A review. *Water research, 116*, 135-148.

Liu, G., Zhang, Y., van der Mark, E., Magic-Knezev, A., Pinto, A., van den Bogert, B., Liu, W., van der Meer, W., & Medema, G. (2018). Assessing the origin of bacteria in tap water and distribution system in an unchlorinated drinking water system by SourceTracker using microbial community fingerprints. *Water research, 138*, 86-96.

Liu, L., Liu, Y., Lu, Q., Chen, G., & Wang, G. (2016). Assessing comprehensive performance of biofilm formation and water quality in drinking water distribution systems. *Water Science and Technology: Water Supply, 17*(1), 267-278.

Logan, G., Freimanis, G. L., King, D. J., Valdazo-González, B., Bachanek-Bankowska, K., Sanderson, N. D., Knowles, N. J., King, D. P., & Cottam, E. M. (2014). A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. *BMC genomics, 15*(1), 828.

Loit, K., Adamson, K., Bahram, M., Puusepp, R., Anslan, S., Kiiker, R., Drenkhan, R., & Tedersoo, L. (2019). Relative performance of Oxford Nanopore MinION vs. Pacific Biosciences Sequel third-generation sequencing platforms in identification of agricultural and forest pathogens. *bioRxiv*, 592972.

Ma, X., Stachler, E., & Bibby, K. (2017). Evaluation of Oxford Nanopore MinIONTM Sequencing for 16S rRNA Microbiome Characterization. *bioRxiv*, 099960.

Mezzatesta, M. L., Gona, F., & Stefani, S. (2012). Enterobacter cloacae complex: clinical impact and emerging antibiotic resistance. *Future microbiology, 7*(7), 887-902.

Mitsuhashi, S., Kryukov, K., Nakagawa, S., Takeuchi, J. S., Shiraishi, Y., Asano, K., & Imanishi, T. (2017). A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Scientific reports, 7*(1), 5657.

Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., & Ragoussis, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Scientific reports, 6*, 31602.

Pavlov, D., De Wet, C., Grabow, W., & Ehlers, M. (2004). Potentially pathogenic features of heterotrophic plate count bacteria isolated from treated and untreated drinking water. *International journal of food microbiology, 92*(3), 275-287.

Pinto, A. J., & Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PloS one, 7*(8), e43093.

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., Barrio-Amorós, C. L., Salazar-Valenzuela, D., & Prost, S. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience, 7*(4), giy033.

Prest, E., El-Chakhtoura, J., Hammes, F., Saikaly, P., Van Loosdrecht, M., & Vrouwenvelder, J. S. (2014). Combining flow cytometry and 16S rRNA gene pyrosequencing: a promising approach for drinking water monitoring and characterization. *Water research, 63*, 179-189.

Proctor, C. R., & Hammes, F. (2015). Drinking water microbiology—from measurement to management. *Current Opinion in Biotechnology, 33*, 87-94.

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., & Beutler, N. A. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *nature protocols, 12*(6), 1261.

Rhee, M., & Burns, M. A. (2006). Nanopore sequencing technology: research trends and applications. *Trends in biotechnology, 24*(12), 580-586.

Runtuwene, L. R., Tuda, J. S., Mongan, A. E., Makalowski, W., Frith, M. C., Imwong, M., Srisutham, S., Thi, L. A. N., Tuan, N. N., & Eshita, Y. (2018). Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum. *Scientific reports, 8*(1), 8286.

Savichtcheva, O., & Okabe, S. (2006). Alternative indicators of fecal pollution: relations with pathogens and conventional indicators, current methodologies for direct pathogen monitoring and future application perspectives. *Water research, 40*(13), 2463-2476.

Saxena, G., Bharagava, R. N., Kaithwas, G., & Raj, A. (2014). Microbial indicators, pathogens and methods for their monitoring in water environment. *Journal of water and health, 13*(2), 319-339.

Schmidt, K., Mwaigwisya, S., Crossman, L., Doumith, M., Munroe, D., Pires, C., Khan, A., Woodford, N., Saunders, N., & Wain, J. (2016). Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *Journal of Antimicrobial Chemotherapy, 72*(1), 104-114.

Schwartz, T., Kohnen, W., Jansen, B., & Obst, U. (2003). Detection of antibiotic-resistant bacteria and their resistance genes in wastewater, surface water, and drinking water biofilms. *FEMS microbiology ecology, 43*(3), 325-335.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., & Koren, S. (2019). Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit. *bioRxiv*, 715722.

Shi, P., Jia, S., Zhang, X.-X., Zhang, T., Cheng, S., & Li, A. (2013). Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. *Water research, 47*(1), 111-120.

Shin, J., Lee, S., Go, M.-J., Lee, S. Y., Kim, S. C., Lee, C.-H., & Cho, B.-K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific reports, 6*, 29681.

Simon, H. Y., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell, 178*(4), 779-794.

Smeets, P., Medema, G., & Van Dijk, J. (2009). The Dutch secret: how to provide safe drinking water without chlorine in the Netherlands. *Drinking Water Engineering and Science, 2*(1), 1-14.

Su, H.-C., Liu, Y.-S., Pan, C.-G., Chen, J., He, L.-Y., & Ying, G.-G. (2018). Persistence of antibiotic resistance genes and bacterial community changes in drinking water treatment system: From drinking water source to tap water. *Science of the Total Environment, 616*, 453-461.

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

Tarumoto, N., Sakai, J., Sujino, K., Yamaguchi, T., Ohta, M., Yamagishi, J., Runtuwene, L., Murakami, T., Suzuki, Y., & Maeda, T. (2017). Use of the Oxford Nanopore MinION sequencer for MLST genotyping of vancomycin-resistant enterococci. *The Journal of hospital infection, 96*(3), 296.

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., & Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Scientific reports, 8*(1), 10931.

van der Helm, E., Imamovic, L., Hashim Ellabaan, M. M., van Schaik, W., Koza, A., & Sommer, M. O. A. (2017). Rapid resistome mapping using nanopore sequencing. *Nucleic Acids Research, 45*(8), e61-e61.

Vierheilig, J., ., Savio, D., ., Ley, R. E., Mach, R. L., Farnleitner, A. H., & Reischer, G. H. (2015). Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. *Water Science & Technology A Journal of the International Association on Water Pollution Research, In Press, Uncorrected Proof*(11), 1962.

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., & Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences, 115*(39), 9726-9731.

Votintseva, A. A., Bradley, P., Pankhurst, L., del Ojo Elias, C., Loose, M., Nilgiriwala, K., Chatterjee, A., Smith, E. G., Sanderson, N., & Walker, T. M. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *Journal of clinical microbiology, 55*(5), 1285-1298.

Walsh, T. R., Weeks, J., Livermore, D. M., & Toleman, M. A. (2011). Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. *The Lancet infectious diseases, 11*(5), 355-362.

Wang, H., Bedard, E., Prevost, M., Camper, A. K., Hill, V. R., & Pruden, A. (2017). Methodological approaches for monitoring opportunistic pathogens in premise plumbing: a review. *Water research, 117*, 68-86.

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome biology, 20*(1), 129.

Williams, M. M., Armbruster, C. R., & Arduino, M. J. (2013). Plumbing of hospital premises is a reservoir for opportunistically pathogenic microorganisms: a review. *Biofouling, 29*(2), 147-162.

Wingender, J., & Flemming, H.-C. (2011). Biofilms in drinking water and their role as reservoir for pathogens. *International journal of hygiene and environmental health, 214*(6), 417-423. doi:https://doi.org/10.1016/j.ijheh.2011.05.009

Xi, C., Zhang, Y., Marrs, C. F., Ye, W., Simon, C., Foxman, B., & Nriagu, J. (2009). Prevalence of antibiotic resistance in drinking water treatment and distribution systems. *Appl. Environ. Microbiol., 75*(17), 5714-5718.

Xia, Y., Li, A. D., Deng, Y., Jiang, X. T., Li, L. G., & Zhang, T. (2017). MinION Nanopore Sequencing Enables Correlation between Resistome Phenotype and Genotype of Coliform Bacteria in Municipal Sewage. *Frontiers in microbiology, 8*, 2105.

Xu, L., Ouyang, W., Qian, Y., Su, C., Su, J., & Chen, H. (2016). High-throughput profiling of antibiotic resistance genes in drinking water treatment plants and distribution systems. *Environmental pollution, 213*, 119-126.

Zhang, M., Wang, L., Xu, M., Zhou, H., Wang, S., Wang, Y., Bai, M., & Zhang, C. (2019). Selective antibiotic resistance genes in multiphase samples during biofilm growth in a simulated drinking water distribution system: Occurrence, correlation and low-pressure ultraviolet removal. *Science of the Total Environment, 649*, 146-155.

Appendix A

Table A.1 ZymoBIOMICS™ microbial community DNA standard (D6305) information

Species	NRRL accession NO.	GC content (%)	Gram stain
Bacillus subtilis	B-354	43.9	+
Enterococcus faecalis	B-537	37.5	+
Escherichia coli	B-1109	46.7	-
Lactobacillus fermentum	B-1840	52.4	+
Listeria monocytogenes	B-33116	38.0	+
Pseudomonas aeruginosa	B-3509	66.2	-
Salmonella enterica	B-4212	52.2	-
Staphylococcus aureus	B-41012	32.9	+

Table A.2 ZymoBIOMICS™ HMW microbial community DNA standard (D6322) information

Species	NRRL accession NO.	Genomic DNA composition (%)	Genome copy composition (%)	GC content (%)	Gram stain
Bacillus subtilis	B-354	14	13.2	43.9	+
Enterococcus faecalis	B-537	14	18.8	37.5	+
Escherichia coli	B-1109	14	10.9	46.7	-
Listeria monocytogenes	B-33116	14	17.8	38.0	+
Pseudomonas aeruginosa	B-3509	14	7.8	66.2	-
Salmonella enterica	B-4212	14	11.2	52.2	-
Staphylococcus aureus	B-41012	14	19.6	32.9	+
Saccharomyces cerevisiae	Y-567	2	0.63	38.3	Yeast

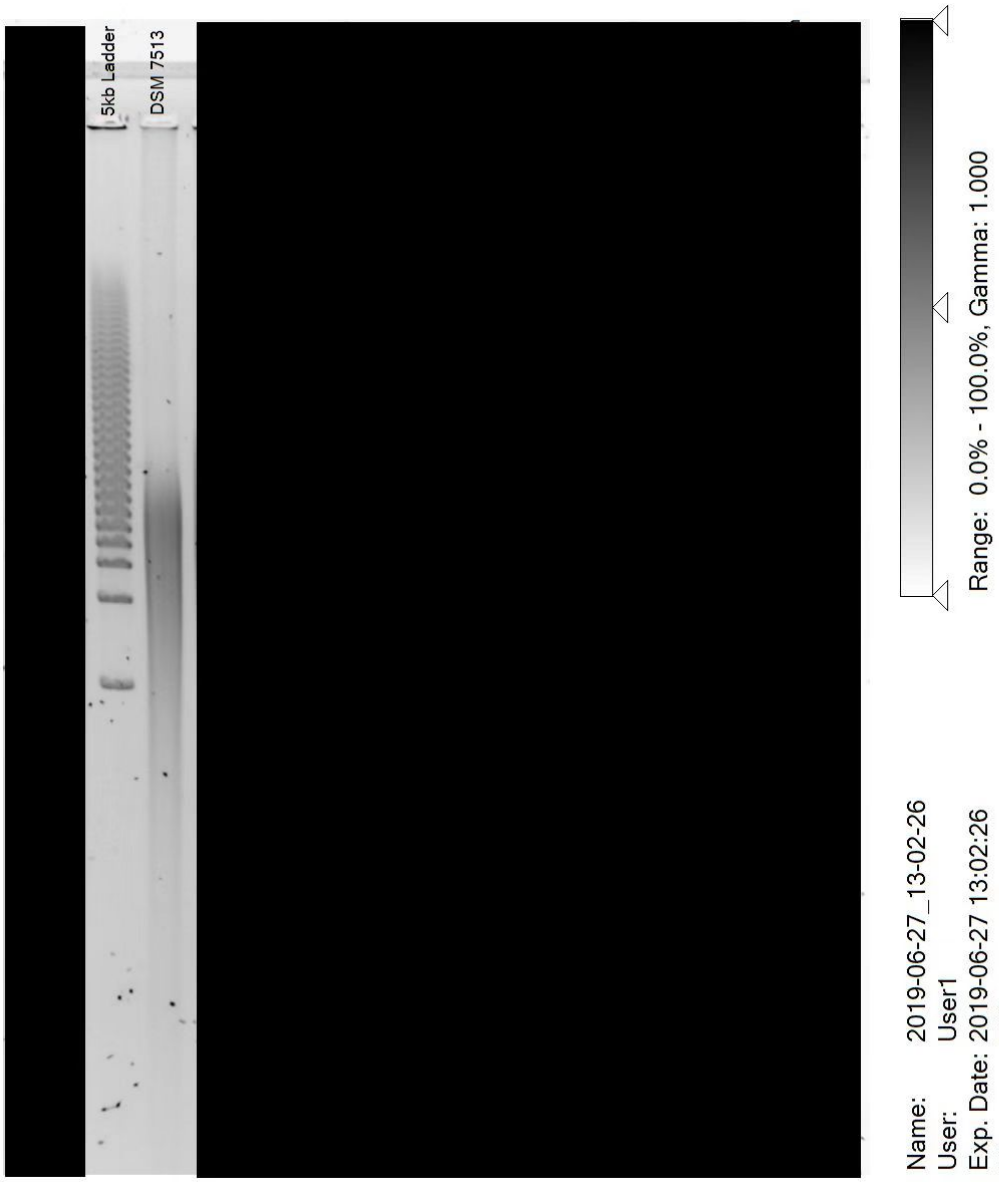

Figure A.1 Agarose gel electrophoresis of *Legionella pneumophila* (DSM-7513) genomic DNA (provided by DSMZ)

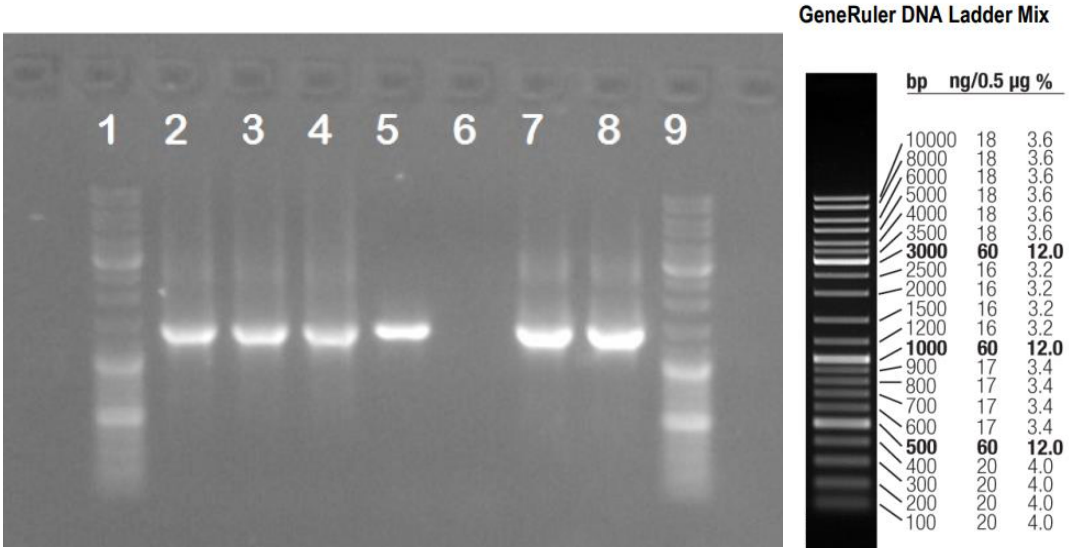

Figure A.2 Agarose gel electrophoresis of near full-length 16S PCR products (Provided by BaseClear B.V.) Lane 1,9: GeneRulerTM 1 kbp DNA ladder mix; lane 2, 3, 4: PCR product of triplicate tap water DNA samples; lane 5: Positive control of E. coli indicating a successful amplification; lane 6: Negative control; lane 7: PCR product of ZymoBIOMICSTM microbial community DNA standard (D6305); lane 8: PCR product of Legionella pneumophila (DSM-7513) genomic DNA.

Table A.3 OTU table of NanoAmpli-Seq full-length 16S amplicon sequencing of ZymoBIOMICS™ microbial community DNA standard (D6305)

Species	OTU ID	Cumulative reads
Bacillus subtilis	OTU_1	4012
	OTU_10	62
	OTU_12	198
	OTU_13	129
Enterococcus faecalis	OTU_2	2098
Escherichia coli	OTU_7	104
Lactobacillus fermentum	OTU_4	751
Listeria monocytogenes	OTU_3	3680
	OTU_14	164
Pseudomonas aeruginosa	OTU_6	119
Salmonella enterica	OTU_9	82
Staphylococcus aureus	OTU_5	867
	OTU_11	50

Table A.4 OTU table of NanoAmpli-Seq full-length 16S amplicon sequencing of mock community amplicon pool consisting of 95% ZymoBIOMICS™ microbial community (D6305) amplicons and 5% *Legionella pneumophila* amplicons

Species	OTU ID	Cumulative reads
Bacillus subtilis	OTU_1	5553
	OTU_10	57
	OTU_14	333
	OTU_16	133
Enterococcus faecalis	OTU_3	3056
	OTU_13	11
	OTU_15	155
	OTU_17	111
Escherichia coli	OTU_6	123
Lactobacillus fermentum	OTU_5	979
Listeria monocytogenes	OTU_2	6428
	OTU_9	128
	OTU_12	283
Pseudomonas aeruginosa	OTU_7	160
Salmonella enterica	OTU_8	112
Staphylococcus aureus	OTU_4	1231
	OTU_11	106

Table A.5 OTU table of NanoAmpli-Seq full-length 16S amplicon sequencing of mock community amplicon pool consisting of 90% ZymoBIOMICS™ microbial community (D6305) amplicons and 10% *Legionella pneumophila* amplicons

Species	OTU ID	Cumulative reads
Bacillus subtilis	OTU_3	5399
Enterococcus faecalis	OTU_1	4254
	OTU_11	71
	OTU_13	78
	OTU_14	8
Escherichia coli	OTU_6	197
Lactobacillus fermentum	OTU_5	1050
	OTU_15	12
Legionella pneumophila	OTU_9	44
Listeria monocytogenes	OTU_2	5293
	OTU_10	111
	OTU_12	30
Pseudomonas aeruginosa	OTU_7	116
Staphylococcus aureus	OTU_4	1947
	OTU_8	90

Table A.6 OTU table of tap water microbial community MinION™ full-length 16S amplicon sequencing

Taxonomy level	Taxonomy	OTU ID	Cumulative reads	Taxonomy	OTU ID	Cumulative reads
Species	*Geobacter metallireducens*	OTU_22	20	*Nitrospira lenta*	OTU_3	92
		OTU_25	29		OTU_4	58
	Helicobacter brantae	OTU_24	39	*Pirellula staleyi*	OTU_34	7
	Limnobacter thiooxidans	OTU_5	223	*Rhodoferax ferrireducens*	OTU_2	186
	Methylobacter psychrophilus	OTU_37	6	*Stenotrophobacter namibiensis*	OTU_48	16
	Pedomicrobium manganicum	OTU_49	24	*Stenotrophobacter roseus*	OTU_40	17
Family	*Chromatiaceae*	OTU_23	28	*Planctomycetaceae*	OTU_30	31
	Comamonadaceae	OTU_1	752		OTU_33	16
		OTU_16	28		OTU_36	13
	Desulfobulbaceae	OTU_7	145	*Sphingobacteriaceae*	OTU_17	7
	Hyphomicrobiaceae	OTU_43	2		OTU_18	20
	Phycisphaeraceae	OTU_55	25		OTU_35	14
					OTU_54	31
Order	*Clostridiales*	OTU_41	26	*Rhizobiales*	OTU_32	7
	Desulfobacterales	OTU_109	5			
Class	*Alphaproteobacteria*	OTU_9	14	*Deltaproteobacteria*	OTU_26	38
		OTU_13	18		OTU_27	28
		OTU_42	56		OTU_50	21
	Gammaproteobacteria	OTU_52	25		OTU_58	2
Phylum	*Planctomycetes*	OTU_47	18	*Proteobacteria*	OTU_6	24

(Table A.6 continued)

Taxonomy level	Taxonomy	OTU ID	Cumulative reads	Taxonomy	OTU ID	Cumulative reads
Superkingdom	Bacteria	OTU_8	95	Bacteria	OTU_39	45
		OTU_10	206		OTU_44	21
		OTU_11	99		OTU_45	38
		OTU_28	8		OTU_46	4
		OTU_31	21		OTU_51	9
		OTU_38	12		OTU_56	22
Not classified	-	OTU_12	39	-	OTU_20	33
		OTU_14	72		OTU_29	25
		OTU_15	147		OTU_53	16
		OTU_19	15		OTU_57	23

Table A.7 Species classification of tap water microbial community MinION™ gDNA sequencing (top 1000 abundant species)

Species	Cumulative reads	Species	Cumulative reads
Limnohabitans sp. 63ED37-2	12,837	*Leptothrix cholodnii*	390
Candidatus Nanopelagicus limnes	6,981	*Polaromonas naphthalenivorans*	389
Candidatus Fonsibacter ubiquis	5,767	*Burkholderia pseudomallei*	387
Candidatus Nanopelagicus abundans	3,804	*Candidatus Planktophila dulcis*	375
Candidatus Planktophila vernalis	3,379	*Candidatus Planktophila limnetica*	363
Homo sapiens	2,291	*Gemmata sp. SH-PL17*	362
Ramlibacter tataouinensis	1,319	*Sandaracinus amylolyticus*	358
Nitrospira moscoviensis	1,214	*Methylibium petroleiphilum*	357
Rhodopseudomonas palustris	1,064	*Comamonas testosteroni*	356
Variovorax paradoxus	1,047	*Acidovorax citrulli*	354
Candidatus Nanopelagicus hibericus	966	*Rhizobium leguminosarum*	353
Rhodoplanes sp. Z2-YC6860	864	*Acidovorax sp. NA3*	350
Actinobacteria bacterium IMCC19121	860	*Cyanobium sp. NIES-981*	288
Limnohabitans sp. 103DPR2	752	*Shewanella baltica*	281
Hydrogenophaga sp. RAC07	697	*Bradyrhizobium japonicum*	272
Hydrogenophaga sp. PBC	670	*Acidovorax avenae*	267
Caulobacteraceae bacterium OTSz_A_272	658	*Betaproteobacteria bacterium GR16-43*	259
Acidovorax sp. RAC01	630	*Lacunisphaera limnophila*	257
Nitrospira defluvii	604	*Acidovorax sp. P3*	255
Rhodoferax sp. DCY110	590	*Pseudomonas fluorescens*	255
Candidatus Nitrospira inopinata	589	*Haliangium ochraceum*	254
Candidatus Planktophila lacus	561	*Aeromonas hydrophila*	247
Candidatus Methylopumilus planktonicus	552	*Rhizobacter gummiphilus*	245
Alicycliphilus denitrificans	498	*Hydrogenophaga crassostreae*	243

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
[Polyangium] brachysporum	242	*Bosea vaviloviae*	203
Paludisphaera borealis	242	*Actinobacteria bacterium IMCC26103*	203
Variovorax sp. PAMC 28711	241	*Mesorhizobium loti*	202
Thauera sp. K11	241	*Herbaspirillum frisingense*	202
Paucibacter sp. KCTC 42545	240	*Comamonas serinivorans*	201
Rhodoferax ferrireducens	239	*Ralstonia pickettii*	199
Blastochloris viridis	238	*Cyanobium gracile*	199
Diaphorobacter polyhydroxybutyrativorans	237	*Tistrella mobilis*	198
Actinobacteria bacterium IMCC25003	237	*Cupriavidus gilardii*	198
Candidatus Nitrosotenuis cloacae	236	*Oligotropha carboxidovorans*	197
Planctomyces sp. SH-PL62	233	*Bradyrhizobium sp. CCGE-LA001*	196
Verminephrobacter eiseniae	232	*Bradyrhizobium sp. ORS 278*	195
Pseudorhodoplanes sinuspersici	228	*Cupriavidus necator*	195
Bradyrhizobium sp. BTAi1	223	*Sinorhizobium meliloti*	193
Delftia sp. Cs1-4	223	*Sulfurifustis variabilis*	193
Bradyrhizobium sp. ORS 285	222	*Methylobacterium sp. 4-46*	192
Opitutus terrae	222	*Singulisphaera acidiphila*	190
Bradyrhizobium oligotrophicum	215	*Hyphomicrobium denitrificans*	189
Bosea sp. RAC05	214	*Methylobacterium extorquens*	189
Roseateles depolymerans	213	*Cupriavidus basilensis*	189
Bdellovibrio bacteriovorus	213	*Rhodoferax antarcticus*	189
Azospirillum thiophilum	211	*Phycomyces blakesleeanus*	187
Rhodospirillum centenum	208	*Methylobacterium aquaticum*	183
Starkeya novella	207	*Thauera sp. MZ1T*	183

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Mitsuaria sp. 7	182	*Comamonadaceae bacterium A1*	157
Thiomonas intermedia	180	*Candidatus Solibacter usitatus*	157
Rhodobacter sphaeroides	178	*Bosea sp. PAMC 26642*	155
Delftia tsuruhatensis	178	*Sphingomonas wittichii*	154
Bordetella hinzii	178	*Pseudomonas mendocina*	154
Polymorphum gilvum	176	*Microvirga ossetica*	152
Bosea sp. AS-1	175	*Sphingopyxis macrogoltabida*	152
Conexibacter woesei	175	*Burkholderia cenocepacia*	152
Xanthobacter autotrophicus	173	*Achromobacter denitrificans*	152
Anaeromyxobacter sp. Fw109-5	173	*Methyloversatilis sp. RAC08*	152
Nitrospirillum amazonense	172	*Phenylobacterium zucineum*	151
Paraburkholderia caribensis	172	*Comamonadaceae bacterium B1*	151
Burkholderia multivorans	171	*Pseudomonas chlororaphis*	151
Sulfuritalea hydrogenivorans	171	*Mesorhizobium sp. B7*	150
Bradyrhizobium sp.	168	*Azoarcus sp. CIB*	150
Burkholderia cepacia	168	*Myxococcus fulvus*	150
Roseomonas sp. FDAARGOS_362	167	*Acidovorax sp. JS42*	148
Thauera chlorobenzoica	166	*Magnetospirillum gryphiswaldense*	147
Methylobacterium sp. PR1016A	165	*Castellaniella defragrans*	147
Opitutaceae bacterium TAV5	165	*Chelatococcus sp. CO-6*	146
Methylobacterium populi	162	*Chondromyces crocatus*	145
Anaeromyxobacter dehalogenans	162	*Neorhizobium galegae*	144
Archangium gephyra	161	*Cupriavidus pinatubonensis*	144
Phycisphaera mikurensis	158	*Thauera humireducens*	144

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Hartmannibacter diazotrophicus	143	*Magnetospirillum sp. XM-1*	131
Variibacter gotjawalensis	142	*Massilia sp. B2*	131
Bordetella genomosp. 13	142	*Lysobacter antibioticus*	131
Candidatus Accumulibacter phosphatis	142	*Caulobacter mirabilis*	130
Pseudoxanthomonas suwonensis	142	*Paraburkholderia xenovorans*	130
Verrucomicrobia bacterium	142	*Brevundimonas subvibrioides*	129
Massilia putida	141	*Pirellula staleyi*	129
Azoarcus sp. KH32C	139	*Mesorhizobium amorphae*	128
Nostoc sp. PCC 7120	139	*Brevundimonas sp. LM2*	128
Bradyrhizobium sp. S23321	138	*Lysobacter capsici*	128
Caulobacter vibrioides	138	*Shinella sp. HZN7*	127
Collimonas fungivorans	138	*Burkholderia ubonensis*	127
Sulfuricaulis limicola	137	*Massilia sp. WG5*	127
Parvibaculum lavamentivorans	136	*Pandoraea pnomenusa*	125
Xanthomonas campestris	136	*Massilia sp. NR 4-1*	125
Methylobacterium nodulans	135	*Pseudogulbenkiania sp. NH8B*	125
Bordetella bronchialis	135	*Azospira oryzae*	125
Delftia sp. HK171	134	*Devosia sp. A16*	124
Dokdonella koreensis	134	*Cupriavidus sp. USMAA2-4*	124
Nitrobacter hamburgensis	131	*Acidovorax sp. NA2*	123
Methylosinus trichosporium	131	*Delftia acidovorans*	122
Aminobacter aminovorans	131	*Cystobacter fuscus*	122
Sinorhizobium fredii	131	*Gemmatimonas aurantiaca*	122
Roseomonas gilardii	131	*Magnetospirillum sp. ME-1*	121

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Acidovorax sp. P4	121	*Synechococcus sp. SynAce01*	110
Burkholderiales bacterium GJ-E10	121	*Aureimonas sp. AU20*	109
Thiobacillus denitrificans	121	*Comamonas kerstersii*	109
Ascoidea rubescens	121	*Bordetella sp. H567*	109
Corallococcus coralloides	120	*Sideroxydans lithotrophicus*	109
Magnetospirillum magneticum	119	*Rhodanobacter denitrificans*	109
Azorhizobium caulinodans	117	*Sphingomonas sp. DC-6*	108
Collimonas arenae	117	*Ottowia sp. oral taxon 894*	108
Brevundimonas naejangsanensis	116	*Janthinobacterium sp. 1_2014MBL_MicDiv*	108
Burkholderia glumae	115	*Klebsiella sp. M5al*	108
Cupriavidus sp. NH9	115	*Pseudomonas pseudoalcaligenes*	107
Bordetella petrii	115	*Hyphomicrobium nitrativorans*	106
Immundisolibacter cernigliae	115	*Sphingomonas sanxanigenens*	106
Bordetella genomosp. 8	114	*Pseudomonas syringae*	106
Aromatoleum aromaticum	114	*Pseudomonas citronellolis*	105
Ralstonia mannitolilytica	113	*Azotobacter chroococcum*	103
Acidovorax ebreus	113	*Rhodothermaceae bacterium RA*	103
Vulgatibacter incomptus	112	*Mesorhizobium opportunistum*	102
Serratia marcescens	111	*Pararhodospirillum photometricum*	102
Ilumatobacter coccineus	111	*Jeongeupia sp. USM3*	102
Eutypa lata	111	*Caulobacter sp. K31*	101
Anthracocystis flocculosa	111	*Burkholderia thailandensis*	101
Vitreoscilla filiformis	110	*Mesorhizobium ciceri*	100
Lysobacter enzymogenes	110	*Bordetella genomosp. 9*	100

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Flavobacterium columnare	100	*Stigmatella aurantiaca*	92
Melittangium boletus	98	*Calothrix sp. PCC 7507*	92
Verrucomicrobia bacterium IMCC26134	98	*Lobosporangium transversale*	92
Bdellovibrio exovorus	97	*Devosia sp. H5989*	91
Methyloceanibacter caenitepidi	96	*Rhizorhabdus dicambivorans*	91
Thiobacimonas profunda	96	*Hyphomicrobium sp. MC1*	90
Marichromatium purpuratum	96	*Gluconacetobacter diazotrophicus*	90
Myxococcus hansupus	96	*Burkholderia vietnamiensis*	90
Mycobacterium rhodesiae	96	*Orrella dioscoreae*	90
Fuerstia marisgermanicae	96	*Sphaerobacter thermophilus*	90
Marssonina brunnea	96	*Pannonibacter phragmitetus*	89
Bordetella flabilis	95	*Achromobacter insolitus*	89
Nonomuraea sp. ATCC 55076	95	*Chromobacterium violaceum*	89
Gemmatimonas phototrophica	95	*Lysobacter gummosus*	89
Methylocystis bryophila	94	*Brevundimonas sp. DS20*	88
Ralstonia insidiosa	94	*Sphingomonas taxi*	88
Herbaspirillum hiltneri	94	*Sphingopyxis granuli*	88
Dyella thiooxydans	94	*Herbaspirillum seropedicae*	88
Methylocystis sp. SC2	93	*Flavobacterium johnsoniae*	88
Mesorhizobium australicum	93	*Isosphaera pallida*	88
Ensifer adhaerens	93	*Sphingomonas sp. MM-1*	87
Paracoccus yeei	93	*Legionella pneumophila*	87
Caulobacter segnis	92	*Xanthomonas translucens*	87
Rhodobacter sp. CZR27	92	*Rubinisphaera brasiliensis*	87

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Thiohalobacter thiocyanaticus	86	*Rhodospirillum rubrum*	79
Paracoccus denitrificans	85	*Aeromonas sp. CU5*	79
Burkholderia ambifaria	85	*Xanthomonas citri*	79
Dechloromonas aromatica	85	*Rhodococcus opacus*	79
Enterobacter cloacae	85	*Burkholderia sp. CCGE1002*	78
Myxococcus stipitatus	85	*Janthinobacterium sp. LM6*	78
Methylobacterium phyllosphaerae	84	*Thioalkalivibrio sulfidiphilus*	78
Sphingomonas panacis	84	*Buchnera aphidicola*	78
Alcaligenes faecalis	84	*Microcystis aeruginosa*	78
Gaeumannomyces tritici	84	*Chelatococcus daeguensis*	77
Micavibrio aeruginosavorus	83	*Rhizobium etli*	77
Clostridium botulinum	83	*Cupriavidus sp. USMAHM13*	77
Frankia inefficax	82	*Neisseria meningitidis*	77
Erythrobacter litoralis	81	*Thiocystis violascens*	77
Pandoraea apista	81	*Mycobacterium avium*	77
Pseudoxanthomonas spadix	81	*Rhodomicrobium vannielii*	76
Desulfovibrio vulgaris	81	*Cupriavidus metallidurans*	76
Streptomyces venezuelae	81	*Pandoraea pulmonicola*	76
Methylocella silvestris	80	*Janthinobacterium agaricidamnosum*	76
Burkholderia oklahomensis	80	*Thioflavicoccus mobilis*	76
Bordetella bronchiseptica	80	*Lodderomyces elongisporus*	76
Candidatus Promineofilum breve	80	*Nitrobacter winogradskyi*	75
Brevundimonas sp. GW460-12-10-14-LB2	79	*Sphingopyxis sp. QXT-31*	75
Pelagibaca abyssi	79	*Salmonella enterica*	75

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Nocardia brasiliensis	75	*Paracoccidioides lutzii*	72
Planctopirus limnophila	75	*Pseudomonas sp. CCOS 191*	71
Rhizobium sp. ACO-34A	74	*Dyella jiangningensis*	71
Celeribacter indicus	74	*Burkholderia sp. OLGA172*	70
Azospirillum humicireducens	74	*Wenzhouxiangella marina*	70
Paraburkholderia sprentiae	74	*Halomonas beimenensis*	70
Bordetella trematum	74	*Dyella japonica*	70
Pantoea ananatis	74	*Pimelobacter simplex*	70
Nocardia farcinica	74	*Caldilinea aerophila*	70
Frankia alni	74	*Martelella endophytica*	69
Blastomyces gilchristii	74	*Chelativorans sp. BNC1*	69
Sinorhizobium sp. RAC02	73	*Defluviimonas alba*	69
Sphingomonas sp. LM7	73	*Sphingomonas sp. KC8*	69
Burkholderia sp. RPE64	73	*Bordetella pseudohinzii*	69
Sulfuriferula sp. AH1	73	*Geobacter sp. M18*	69
Chromatiaceae bacterium 2141T.STBD.0c.01a	73	*Phaeobacter gallaeciensis*	68
Marinovum algicola	72	*Methylococcus capsulatus*	68
Rhodobacter sp. LPB0142	72	*Agrobacterium tumefaciens*	67
Sphingomonas hengshuiensis	72	*Aeromonas salmonicida*	67
Sphingomonas koreensis	72	*Xanthomonas oryzae*	67
Paraburkholderia sp. BN5	72	*Xanthomonas sacchari*	67
Paraburkholderia sp. SOS3	72	*Streptosporangium roseum*	67
Collimonas pratensis	72	*Prochlorococcus marinus*	67
Allochromatium vinosum	72	*Candidatus Koribacter versatilis*	67

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Chryseobacterium indologenes	67	*endosymbiont of unidentified scaly snail isolate Monju*	64
Sinorhizobium americanum	66	*Martelella sp. AD-3*	63
Caulobacter henricii	66	*Pandoraea thiooxydans*	63
Herbaspirillum sp. meg3	66	*Pseudomonas koreensis*	63
Aquaspirillum sp. LM1	66	*Pandoraea faecigallinarum*	62
Pseudomonas protegens	66	*Pandoraea norimbergensis*	62
Thioalkalivibrio nitratireducens	66	*Kitasatospora setae*	62
Klebsiella pneumoniae	66	*Clavibacter michiganensis*	62
Luteimonas sp. 100111	66	*Actinoplanes missouriensis*	62
Stenotrophomonas acidaminiphila	66	*Amycolatopsis mediterranei*	62
Amycolatopsis orientalis	66	*Chloracidobacterium thermophilum*	62
Pseudonocardia dioxanivorans	66	*Gluconobacter oxydans*	61
Rhodothermus marinus	66	*Croceicoccus marinus*	61
Puccinia graminis	66	*Paraburkholderia phymatum*	61
Martelella mediterranea	65	*Nocardia seriolae*	61
Ensifer sojae	65	*Acidobacterium capsulatum*	61
Rhodovulum sulfidophilum	65	*Rhodopirellula baltica*	61
Sphingobium yanoikuyae	65	*alpha proteobacterium HIMB5*	60
Chromobacterium vaccinii	65	*Confluentimicrobium sp. EMB200-NS6*	60
Desulfarculus baarsii	65	*Novosphingobium aromaticivorans*	60
Streptomyces lydicus	65	*Paraburkholderia fungorum*	60
Rhizobium gallicum	64	*Actinoplanes friuliensis*	60
Novosphingobium resinovorum	64	*Nostoc sp. PCC 7524*	60
Thioalkalivibrio versutus	64	*Niastella koreensis*	60

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Blastomonas sp. RAC04	59	*Pseudomonas alcaligenes*	57
Sphingobium sp. SYK-6	59	*Halomonas aestuarii*	57
Sphingopyxis alaskensis	59	*Blastococcus saxobsidens*	57
Burkholderia sp. CCGE1003	59	*Hymenobacter sp. APR13*	57
Herbaspirillum rubrisubalbicans	59	*Rhodotorula graminis*	57
Pseudomonas frederiksbergensis	59	*Brevundimonas vesicularis*	56
Pseudomonas mosselii	59	*Hyphomonas neptunium*	56
Steroidobacter denitrificans	59	*Paracoccus contaminans*	56
Alcanivorax pacificus	59	*Magnetospira sp. QH-2*	56
Desulfococcus oleovorans	59	*Altererythrobacter mangrovi*	56
Streptomyces hygroscopicus	59	*Pseudomonas resinovorans*	56
Amycolatopsis methanolica	59	*Thioalkalivibrio sp. K90mix*	56
Filimonas lacunae	59	*Streptomyces pactum*	56
Labrenzia sp. VG12	58	*Catenulispora acidiphila*	56
Granulibacter bethesdensis	58	*Capronia epimyces*	56
Sulfuricella denitrificans	58	*Setosphaeria turcica*	56
Saccharopolyspora erythraea	58	*Sclerotinia sclerotiorum*	56
Truepera radiovictrix	58	*Colletotrichum graminicola*	56
Diplodia corticola	58	*Thermothelomyces thermophila*	56
Isaria fumosorosea	58	*Yangia sp. CCB-MM3*	55
Dinoroseobacter shibae	57	*Janthinobacterium sp. Marseille*	55
Paraburkholderia phytofirmans	57	*Thioalkalivibrio paradoxus*	55
Bordetella genomosp. 6	57	*Gloeobacter kilaueensis*	55
Janthinobacterium svalbardensis	57	*Thielavia terrestris*	55

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Sphingopyxis terrae	54	*Pseudonocardia sp. AL041005-10*	52
Pseudomonas knackmussii	54	*Gloeobacter violaceus*	52
Pseudomonas sp. CC6-YY-74	54	*Ruegeria pomeroyi*	51
Desulfovibrio desulfuricans	54	*Croceicoccus naphthovorans*	51
Frankia sp. EAN1pec	54	*Gallionella capsiferriformans*	51
Haliscomenobacter hydrossis	54	*Aeromonas media*	51
Pseudocercospora fijiensis	54	*Anaeromyxobacter sp. K*	51
Sordaria macrospora	54	*Leifsonia xyli*	51
Rhodobacteraceae bacterium QY30	53	*Actinoplanes sp. N902-109*	51
Escherichia coli	53	*Oceanithermus profundus*	51
Xanthomonas vesicatoria	53	*Aspergillus aculeatus*	51
Plantactinospora sp. KBS50	53	*Sphaerulina musiva*	51
Thermaerobacter marianensis	53	*Ochrobactrum anthropi*	50
Fluviicola taffensis	53	*Rhodobacter capsulatus*	50
Paracoccidioides brasiliensis	53	*Burkholderia lata*	50
Hoeflea sp. IMCC20628	52	*Alkalilimnicola ehrlichii*	50
Magnetococcus marinus	52	*Myxococcus xanthus*	50
Sulfitobacter sp. AM1-D1	52	*Mycobacterium smegmatis*	50
Sphingomonas sp. JJ-A5	52	*Fimbriimonas ginsengisoli*	50
Pseudomonas brassicacearum	52	*Hymenobacter sp. PAMC 26628*	50
Azotobacter vinelandii	52	*Cutaneotrichosporon oleaginosum*	50
Desulfuromonas soudanensis	52	*Agrobacterium sp. RAC06*	49
Geobacter uraniireducens	52	*Novosphingobium pentaromativorans*	49
Lentzea guizhouensis	52	*Herminiimonas arsenicoxydans*	49

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Yersinia enterocolitica	49	*Colletotrichum higginsianum*	48
Halomonas sp. 1513	49	*Rhizobium phaseoli*	47
Nocardioides sp. JS614	49	*Sinorhizobium sp. CCBAU 05631*	47
Kibdelosporangium phytohabitans	49	*Rhodovulum sp. P5*	47
Kiritimatiella glycovorans	49	*Sphingobium sp. YBL2*	47
Dactylellina haptotyla	49	*Sphingopyxis fribergensis*	47
Leptosphaeria maculans	49	*Burkholderia sp. RPE67*	47
Phyllobacterium sp. Tri-48	48	*Pseudomonas fulva*	47
Asticcacaulis excentricus	48	*Pseudomonas sp. URMO17WK12:I11*	47
Porphyrobacter neustonensis	48	*Zobellella denitrificans*	47
Porphyrobacter sp. LM 6	48	*Acidihalobacter prosperus*	47
Sphingomonas sp. LK11	48	*Halotalea alkalilenta*	47
Burkholderia sp. Bp7605	48	*Geoalkalibacter subterraneus*	47
Pandoraea vervacti	48	*Streptomyces sp. CdTB01*	47
Pseudomonas parafulva	48	*Arsenicicoccus sp. oral taxon 190*	47
Ectothiorhodospira sp. BSL-9	48	*Cellulosimicrobium cellulans*	47
Streptomyces cattleya	48	*Saccharothrix espanaensis*	47
Streptomyces sp. Mg1	48	*Melampsora larici-populina*	47
Mycobacterium sp. JS623	48	*Candidatus Filomicrobium marinum*	46
Stackebrandtia nassauensis	48	*Maricaulis maris*	46
Kribbella flavida	48	*Paracoccus aminophilus*	46
Nocardioides dokdonensis	48	*Roseibacterium elongatum*	46
Botrytis cinerea	48	*Sphingobium sp. TKS*	46
Pseudogymnoascus destructans	48	*Sphingorhabdus flavimaris*	46

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Burkholderia stabilis	46	*Amycolatopsis japonica*	45
Burkholderia sp. HB1	46	*Thermomonospora curvata*	45
Polynucleobacter duraquae	46	*Granulicella mallensis*	45
Raoultella ornithinolytica	46	*Hymenobacter sedentarius*	45
Desulfuromonas sp. DDH964	46	*Thermogutta terrifontis*	45
Streptomyces sp. CNQ-509	46	*Verticillium dahliae*	45
Streptomyces sp. RTd22	46	*Candidatus Nitrosopumilus adriaticus*	45
Kutzneria albida	46	*Sinorhizobium medicae*	44
Symbiobacterium thermophilum	46	*Komagataeibacter xylinus*	44
Flavobacterium commune	46	*Sphingobium chlorophenolicum*	44
Sphingobacteriaceae bacterium GW460-11-11-14-LB5	46	*Sphingomonas sp. NIC1*	44
Colletotrichum orchidophilum	46	*Pseudomonas sp. TCU-HL1*	44
Trichoderma reesei	46	*Haemophilus influenzae*	44
Purpureocillium lilacinum	46	*Vibrio vulnificus*	44
Agrobacterium rhizogenes	45	*Xanthomonas gardneri*	44
Altererythrobacter dongtanensis	45	*Desulfovibrio magneticus*	44
Citromicrobium sp. JL477	45	*Mycobacterium gilvum*	44
Sphingomonas melonis	45	*Alloactinosynnema sp. L-07*	44
Pusillimonas sp. T7-7	45	*Thermobacillus composti*	44
Pseudomonas azotoformans	45	*Flavobacterium psychrophilum*	44
Geobacter pickeringii	45	*Methylobacterium sp. AMS5*	43
Streptomyces albus	45	*Rhizobium tropici*	43
Nocardia soli	45	*Celeribacter manganoxidans*	43
Cellulomonas fimi	45	*Altererythrobacter namhicola*	43

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Sphingobium japonicum	43	*Cellulomonas sp. PSBB021*	42
Sphingobium sp. RAC03	43	*Nakamurella multipartita*	42
Burkholderia sp. YI23	43	*Synechococcus sp. KORDI-49*	42
Bordetella avium	43	*Deinococcus gobiensis*	42
Stenotrophomonas sp. LM091	43	*Alicyclobacillus acidocaldarius*	42
Streptomyces albulus	43	*Niabella ginsenosidivorans*	42
Mycobacterium phlei	43	*Niabella soli*	42
Geodermatophilus obscurus	43	*Fusobacterium nucleatum*	42
Modestobacter marinus	43	*Rhizobium sp. 10195*	41
Actinobacteria bacterium IMCC26077	43	*Sphingobium baderi*	41
Synechococcus sp. RCC307	43	*Sphingobium indicum*	41
Deinococcus radiodurans	43	*beta proteobacterium CB*	41
Grosmannia clavigera	43	*Pseudomonas psychrotolerans*	41
Auricularia subglabra	43	*Pseudomonas oryzihabitans*	41
Acidiphilium multivorum	42	*Pseudomonas balearica*	41
Altererythrobacter marensis	42	*Pelobacter propionicus*	41
Sphingopyxis sp. 113P3	42	*Geobacter bemidjiensis*	41
Pandoraea oxalativorans	42	*Streptomyces sp. 4F*	41
Polynucleobacter asymbioticus	42	*Mycobacterium chubuense*	41
Bordetella holmesii	42	*Mycobacterium goodii*	41
Pseudomonas alkylphenolica	42	*Agromyces aureus*	41
Pseudomonas sp. ATCC 13867	42	*Microbacterium sp. No. 7*	41
Cedecea neteri	42	*Leptolyngbya sp. O-77*	41
Mycobacterium abscessus	42	*Flavisolibacter tropicus*	41

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Hymenobacter sp. PAMC 26554	41	*Mycobacterium sp. djl-10*	39
Chlorobaculum limnaeum	41	*Nocardia terpenica*	39
Neofusicoccum parvum	41	*Frankia casuarinae*	39
Sporothrix schenckii	41	*Intrasporangium calvum*	39
Rhizobium sp. NT-26	40	*Moorea producens*	39
Celeribacter ethanolicus	40	*Chlorobium phaeobacteroides*	39
Rhodovulum sp. MB263	40	*Endocarpon pusillum*	39
Sphingobium cloacae	40	*Penicilliopsis zonata*	39
Burkholderia stagnalis	40	*Tetrapisispora blattae*	39
Nitrosospira briensis	40	*Postia placenta*	39
Streptomyces sp. CLI2509	40	*Ustilago maydis*	39
Beutenbergia cavernae	40	*Candidatus Pelagibacter sp. IMCC9063*	38
Eggerthella lenta	40	*Agrobacterium fabrum*	38
Roseiflexus sp. RS-1	40	*Agrobacterium vitis*	38
Hymenobacter swuensis	40	*Rhizobium sp. N324*	38
Solitalea canadensis	40	*Hyphomonas sp. Mor2*	38
Moesziomyces antarcticus	40	*Loktanella vestfoldensis*	38
Nitratireductor basaltis	39	*Thioclava nitratireducens*	38
Nitrosospira lacus	39	*Acetobacter pasteurianus*	38
Pseudomonas entomophila	39	*Halorhodospira halophila*	38
Serratia plymuthica	39	*Serratia ficaria*	38
Streptomyces bingchenggensis	39	*Methylomonas methanica*	38
Streptomyces scabiei	39	*Desulfococcus multivorans*	38
Streptomyces silaceus	39	*Pseudodesulfovibrio aespoeensis*	38

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Pseudodesulfovibrio indicus	38	*Flavobacterium branchiophilum*	37
Geobacter sulfurreducens	38	*Mucilaginibacter sp. BJC16-A31*	37
Streptomyces glaucescens	38	*Beauveria bassiana*	37
Streptomyces puniciscabiei	38	*Neurospora crassa*	37
Mycobacterium vaccae	38	*Candida dubliniensis*	37
Mycobacterium vanbaalenii	38	*Candidatus Pelagibacter ubique*	36
Clostridium pasteurianum	38	*Pelagibacterium halotolerans*	36
Cordyceps militaris	38	*Roseovarius mucosus*	36
Candidatus Nitrosopumilus sediminis	38	*Acidiphilium cryptum*	36
Beijerinckia indica	37	*Erythrobacter gangjinensis*	36
Rhizobium sp. NXC14	37	*Sphingobium herbicidovorans*	36
Halothiobacillus neapolitanus	37	*Sphingobium hydrophobicum*	36
Desulfomicrobium baculatum	37	*Candidatus Symbiobacter mobilis*	36
Streptomyces formicae	37	*Shewanella putrefaciens*	36
Streptomyces sp. SCSIO 03032	37	*Xanthomonas fragariae*	36
Streptomyces vietnamensis	37	*Streptomyces avermitilis*	36
Thermobispora bispora	37	*Streptomyces gilvosporeus*	36
Frankia symbiont of Datisca glomerata	37	*Streptomyces lincolnensis*	36
Kineococcus radiotolerans	37	*Streptomyces pristinaespiralis*	36
Frondihabitans sp. PAMC 28766	37	*Streptomyces sp. Sge12*	36
Isoptericola variabilis	37	*Bifidobacterium longum*	36
Nocardiopsis dassonvillei	37	*Mycobacterium aurum*	36
Meiothermus ruber	37	*Mycobacterium litorale*	36
Granulicella tundricola	37	*Brevibacterium linens*	36

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Verrucosispora maris	36	*Flavobacterium indicum*	35
Rubrobacter xylanophilus	36	*Pestalotiopsis fici*	35
Chloroflexus aurantiacus	36	*Sulfitobacter pseudonitzschiae*	34
Deinococcus maricopensis	36	*Sphingobium sp. EP60837*	34
Cytophaga hutchinsonii	36	*Pandoraea sputorum*	34
Pedobacter cryoconitis	36	*Paraburkholderia rhizoxinica*	34
Komagataeibacter nataicola	35	*Herminiimonas arsenitoxidans*	34
Neoasaia chiangmaiensis	35	*Nitrosomonas communis*	34
Thalassospira xiamenensis	35	*Nitrosospira multiformis*	34
Altererythrobacter atlanticus	35	*Granulosicoccus antarcticus*	34
Novosphingobium sp. PP1Y	35	*Halothiobacillus sp. LS2*	34
Janthinobacterium sp. B9-8	35	*Lelliottia sp. PFL01*	34
Pseudomonas sp. UW4	35	*Flavobacterium indicum*	35
Stenotrophomonas rhizophila	35	*Pestalotiopsis fici*	35
Geobacter metallireducens	35	*Sulfitobacter pseudonitzschiae*	34
Geobacter sp. M21	35	*Sphingobium sp. EP60837*	34
Streptomyces alboflavus	35	*Pandoraea sputorum*	34
Streptomyces xinghaiensis	35	*Paraburkholderia rhizoxinica*	34
Mycobacterium dioxanotrophicus	35	*Herminiimonas arsenitoxidans*	34
Cellulomonas flavigena	35	*Nitrosomonas communis*	34
Micromonospora aurantiaca	35	*Nitrosospira multiformis*	34
Bacillus thuringiensis	35	*Granulosicoccus antarcticus*	34
Terriglobus roseus	35	*Halothiobacillus sp. LS2*	34
Elizabethkingia anophelis	35	*Lelliottia sp. PFL01*	34

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Pectobacterium carotovorum	34	*Streptomyces rubrolavendulae*	33
Methylomicrobium alcaliphilum	34	*Mycobacterium colombiense*	33
Methylomonas denitrificans	34	*Roseiflexus castenholzii*	33
Cobetia marina	34	*Thermomicrobium roseum*	33
Streptomyces reticuli	34	*Calothrix sp. NIES-4071*	33
Rhodococcus hoagii	34	*Synechococcus sp. WH 7803*	33
Rhodococcus sp. WB1	34	*Paenibacillus mucilaginosus*	33
Serinicoccus sp. JLT9	34	*Terriglobus saanensis*	33
Pseudonocardia sp. HH130629-09	34	*Desulfurispirillum indicum*	33
Ignavibacterium album	34	*Salinibacter ruber*	33
Candidatus Kuenenia stuttgartiensis	34	*Chitinophaga pinensis*	33
Rhizobium sp. CIAT894	33	*Belliella baltica*	33
Erythrobacter flavus	33	*Chryseobacterium sp. StRB126*	33
Burkholderia seminalis	33	*Tuber melanosporum*	33
Methylobacillus flagellatus	33	*Scedosporium apiospermum*	33
Pseudomonas alcaliphila	33	*Burkholderia metallica*	32
Pseudomonas cremoricolorata	33	*Marinobacter salarius*	32
Citrobacter freundii	33	*Woeseia oceani*	32
Coxiella burnetii	33	*Serratia fonticola*	32
Tatlockia micdadei	33	*Methylovulum psychrotolerans*	32
Chromohalobacter salexigens	33	*Campylobacter jejuni*	32
Xylella fastidiosa	33	*Streptomyces albireticuli*	32
Desulfovibrio africanus	33	*Kitasatospora aureofaciens*	32
Geobacter anodireducens	33	*Aeromicrobium erythreum*	32

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Pseudonocardia sp. HH130630-07	32	*Jannaschia sp. CCS1*	30
Rubrobacter radiotolerans	32	*Sphingobium sp. MI1205*	30
Anaerolinea thermophila	32	*Pseudomonas rhizosphaerae*	30
Limnochorda pilosa	32	*Pseudomonas sp. MRSN12121*	30
Dyadobacter fermentans	32	*Pseudomonas sp. StFLB209*	30
Hymenobacter sp. DG25B	32	*Pluralibacter gergoviae*	30
Mucilaginibacter gotjawali	32	*Legionella fallonii*	30
Advenella kashmirensis	31	*Halomonas chromatireducens*	30
Azoarcus olearius	31	*Halobacteriovorax marinus*	30
Pseudomonas fragi	31	*Pelobacter carbinolicus*	30
Marinobacter hydrocarbonoclasticus	31	*Corynebacterium sphenisci*	30
Legionella longbeachae	31	*Gordonia polyisoprenivorans*	30
Streptomyces sampsonii	31	*Mycobacterium sp. YC-RL4*	30
Streptomyces xiamenensis	31	*Brachybacterium sp. VM2412*	30
Sinomonas atrocyanea	31	*Luteipulveratus mongoliensis*	30
Microlunatus phosphovorus	31	*Cnuibacter physcomitrellae*	30
Bacillus coagulans	31	*Cellulosimicrobium sp. TH-20*	30
Listeria monocytogenes	31	*Salinispora tropica*	30
Runella slithyformis	31	*Propionibacterium freudenreichii*	30
Pedobacter heparinus	31	*Leptolyngbya boryana*	30
Arthrobotrys oligospora	31	*Bacillus clausii*	30
Verruconis gallopava	31	*Eubacterium limosum*	30
Trichoderma virens	31	*Deferribacter desulfuricans*	30
Candidatus Nitrosomarinus catalina	31	*Arachidicoccus sp. BS20*	30

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Chlorobaculum parvum	30	*Pontibacter akesuensis*	29
Chlorobaculum tepidum	30	*Akkermansia glycaniphila*	29
Candidatus Protochlamydia naegleriophila	30	*Neurospora tetrasperma*	29
Tateyamaria omphalii	29	*Yarrowia lipolytica*	29
Altererythrobacter epoxidivorans	29	*Vanderwaltozyma polyspora*	29
Sphingorhabdus sp. M41	29	*Leisingera methylohalidivorans*	28
Burkholderia territorii	29	*Methylophilus sp. TWE2*	28
Pseudomonas monteilii	29	*Pseudomonas veronii*	28
Acinetobacter baumannii	29	*Citrobacter amalonaticus*	28
Simiduia agarivorans	29	*Alcanivorax xenomutans*	28
Serratia liquefaciens	29	*Stenotrophomonas sp. WZN-1*	28
Halomonas huangheensis	29	*Geobacter daltonii*	28
Marinobacterium aestuarii	29	*Syntrophobacter fumaroxidans*	28
gamma proteobacterium HdN1	29	*Streptomyces clavuligerus*	28
Myxococcus macrosporus	29	*Streptomyces sp. SAT1*	28
Streptomyces ambofaciens	29	*Streptomyces sp. fd1-xmd*	28
Chloroflexus aggregans	29	*Streptomyces violaceoruber*	28
Synechococcus sp. KORDI-100	29	*Nocardia cyriacigeorgica*	28
Tumebacillus sp. AR23208	29	*Candidatus Rhodoluna planktonica*	28
Paenibacillus sp. 32O-W	29	*Acidipropionibacterium acidipropionici*	28
Clostridium perfringens	29	*Acidimicrobium ferrooxidans*	28
Bernardetia litoralis	29	*Dehalogenimonas formicexedens*	28
Algoriphagus sp. M8-2	29	*Oscillatoria nigro-viridis*	28
Spirosoma montaniterrae	29	*Synechococcus sp. KORDI-52*	28

(Table A.7 continued)

Species	Cumulative reads	Species	Cumulative reads
Deinococcus swuensis	28	*Piscirickettsia salmonis*	27
Bacteroides salanitronis	28	*Xanthomonas albilineans*	27
Flammeovirgaceae bacterium 311	28	*Syntrophus aciditrophicus*	27
Lutibacter profundi	28	*Streptomyces niveus*	27
Seonamhaeicola sp. S2-3	28	*Mycobacterium shigaense*	27
Mucilaginibacter sp. PAMC 26640	28	*Mycobacterium thermoresistibile*	27
Pyrenophora teres	28	*Nocardia nova*	27
Magnaporthe oryzae	28	*Micrococcus luteus*	27
Chaetomium globosum	28	*Xylanimonas cellulosilytica*	27
Heterobasidion irregulare	28	*Sanguibacter keddieii*	27
Antarctobacter heliothermus	27	*Gordonibacter urolithinfaciens*	27
Zymomonas mobilis	27	*Synechococcus sp. CC9605*	27
Burkholderia sp. PAMC 28687	27	*Deinococcus geothermalis*	27
Oceanimonas sp. GK1	27	*Clostridioides difficile*	27
Halioglobus pacificus	27	*Chitinophagaceae bacterium 13*	27
Zhongshania aliphaticivorans	27	*Aequorivita sublithincola*	27
Dickeya solani	27	*Chryseobacterium sp. IHBB 10212*	27
Dickeya zeae	27	*Lacinutrix sp. 5H-3-7-4*	27

Appendix B

B.1 DNA extraction and quantification protocol

Checklist

- ☐ Filter membrane
- ☐ Sterile tweezers and scissors
- ☐ FastDNA™ Spin Kit for Soil
- ☐ 1.5 mL Eppendorf DNA LoBind tubes
- ☐ 2 mL Eppendorf DNA LoBind tubes
- ☐ 15 mL tubes
- ☐ P1000, P100, P10 pipette and tips
- ☐ MiniBeadBeater-16 (Model 607EUR, 3450 RPM)
- ☐ Eppendorf Microcentrifuge 5424R
- ☐ Block heater at 55° C and 60° C
- ☐ Qubit 3.0 fluorometer
- ☐ Qubit dsDNA HS Assay Kit
- ☐ Vortex mixer

1) Cut the filter membrane into small pieces and add to a Lysing Matrix E tube.
2) Add **978 µL Sodium Phosphate Buffer**.
3) Add **122 µL MT Buffer**.
4) Homogenize in the MiniBeadBeater-16 for **40 sec**.
5) Centrifuge at **14000 ×g** for **15 min**. Transfer **supernatant** (~**1000 µL**) to a **2.0 mL microcentrifuge tube**.
6) Add **250 µL PPS** and mix by **inverting** the tube **2 min**. Incubate **on ice** for **5 min**.
7) Centrifuge at **14000 ×g** for **5 min**. Transfer **supernatant** to a **15 mL tube**.
8) Resuspend **Binding Matrix suspension** and add **1.0 mL** to the tube.
9) **Invert** by hand for **5 min**. Place tube in a rack for **5 min** to allow **settling** of silica matrix.
10) Remove and **discard 500 µL** of **supernatant**.
11) **Gently resuspend** Binding Matrix in the remaining supernatant. Transfer **600 µL** of the mixture to a **Spin Filter**. Centrifuge at **14000 ×g** for **1 min**. **Empty** the **catch tube**. Add the **remaining** mixture to the Spin Filter and centrifuge and empty the catch tube as before.
12) Add **500 µL SEWS-M** and **gently resuspend** the pellet. Centrifuge at **14000 ×g** for **1 min**. **Empty** the **catch tube**. **Repeat** this step for a total of **3 times**.

13) Centrifuge at **14000 ×g** for **2 min**. **Discard** the **catch tube** and **replace** with a **new catch tube**.

14) Air dry the Spin Filter for **5 min** with **60º C** incubation.

15) Add **60 µL** of **DES** to resuspend Binding Matrix above the Spin Filter. Incubate for **5 min** at **55º C**.

16) Centrifuge **14000 ×g** for **1 min**. **Discard** the **Spin Filter**.

(DNA quantification)

17) Prepare the working solution by adding **199 µL** of dsDNA HS **Buffer** and **1 µL** of dsDNA HS **Reagent** per DNA sample into a **1.5 mL tube**.

18) (Calibration: Add **190 µL working solution** and **10 µL** of **Qubit standard** into each tube. Mix by **vortexing 2-3 sec**.)

19) Add **198 µL** of **working solution** into each of the Qubit Assay tubes.

20) Add **2 µL DNA** sample and mix by **vortexing 2-3 sec**.

21) Incubate at **room temperature** for **2 min**.

22) On the home screen of Qubit 3.0 fluorometer, press **DNA** and then select **dsDNA High Sensitivity**. Press **Run samples**, select the **sample volume (2 µL)** and **units (ng/µL)**.

23) Insert the sample tube into the sample chamber, close the lid and press **Read tube**. Record the data.

B.2 0.7× ratio modified beads solution size selection protocol

Checklist

- ☐ DNA samples (60 ng/μL, 50μL)
- ☐ Agencourt AMPure XP beads
- ☐ 1 M Tris-HCl
- ☐ 0.5 M EDTA pH 8
- ☐ 5 M NaCl
- ☐ 50% w/v PEG 8000
- ☐ TE buffer pH 8
- ☐ 100% Ethanol
- ☐ Nuclease-free water
- ☐ 1.5 mL Eppendorf DNA LoBind tubes
- ☐ 2 mL Eppendorf DNA LoBind tubes
- ☐ P1000, P100, P10 pipette and tips
- ☐ Magnetic rack
- ☐ Minicentrifuge
- ☐ Block heater at 50 °C

1) Prepare the Custom buffer in a **1.5 mL tube** by mixing:

Reagent	Volume	Volume
1 M Tris-HCl	20 μL	5 μL
0.5 M EDTA pH 8	4 μL	1 μL
5 M NaCl	640 μL	160 μL
50% w/v PEG 8000	440 μL	110 μL
Nuclease-free water	888 μL	222 μL
Total	**1992 μL**	**498 μL**

2) **Shake** the Agencourt AMPure XP bottle to **resuspend** the magnetic beads. Add **500 μL** of Agencourt AMPure XP into a **1.5 mL tube**. Place the tube **on magnetic rack**, wait for the solution to **clear** and **discard** the **supernatant (500 μL)**.

3) **Remove** the tubes from the magnet rack. Resuspend the beads with **500 μL** of **nuclease-free water**.

4) Return the tubes **to the magnet**, allow beads to pellet and **discard** the **supernatant (500 μL)**.

5) **Repeat step 3** and **4** once more (**Total** of **2** washes).

6) **Spin down** and place the tube back **on the magnet**. **Pipette off** any residual water.

7) **Remove** the tubes from the magnet rack. **Resuspend** the bead pellets in the tube with **498 µL** of **Custom buffer**.
8) If not using immediately, **store** the suspension at **4° C**.
9) Bring the custom bead suspension to **room temperature** before use.
10) Prepare 70% ethanol by mixing **700 µL ethanol** and **300 µL nuclease-free water** in a **1.5 mL tube**.
11) Dilute DNA sample to **60 ng/µL** in **50 µL** of **TE buffer** at **pH 8** in a **1.5 mL tube**.
12) Add **0.7x (35 µL) custom bead suspension** to the DNA sample and mix by **flicking** the tube. Incubate at **room temperature** on a **Hula mixer** for **10 min**.
13) Briefly **spin down** and place the tube **on the magnet**. **Discard** the **supernatant (85 µL)**.
14) Keep the tube **on magnet** and add **200 µL** of freshly prepared **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol **(200 µL)**.
15) **Repeat** the wash step **once** more.
16) **Spin down** and place the tube **back on magnet**. **Pipette off** any residual ethanol. **Dry** for **30 sec**.
17) **Remove** the tube from the magnet and resuspend the pellet in **40 µL TE buffer**. Incubate for **1 min** at **50° C** and then **5 min** at **room temperature**.
18) Place the tube **on magnet** and wait for the solution to **clear and colourless**. Pipette off **40 µL** of eluate into a clean **1.5 mL tube**.

B.3 16S amplicon sequencing protocol

Checklist

- ☐ 16S PCR product
- ☐ Agencourt AMPure XP beads
- ☐ 100% Ethanol
- ☐ Nuclease-free water
- ☐ Blunt/TA ligase Master Mix
- ☐ Plasmid-Safe™ ATP-Dependent DNase
- ☐ TruePrime™ RCA Kit
- ☐ 1.5 mL Eppendorf DNA LoBind tubes
- ☐ 0.2 mL PCR tubes
- ☐ P1000, P100, P10 pipette and tips
- ☐ Ice bucket with ice
- ☐ Magnetic rack
- ☐ Minicentrifuge
- ☐ Block heater at 50 ºC
- ☐ Thermal cycler
- ☐ Qubit 3.0 fluorometer
- ☐ Qubit dsDNA HS Assay Kit
- ☐ T7 Endonuclease I
- ☐ NEBNext FFPE DNA Repair Mix
- ☐ NEBNext End repair/dA-tailing Module
- ☐ g-TUBE
- ☐ $1D^2$ Ligation Sequencing Kit (SQK-LSK309)
- ☐ R9.5.1 Flow Cell
- ☐ Eppendorf Microcentrifuge 5424R
- ☐ Vortex mixer

Preparation

1) Prepare **3x1 mL** of **70% ethanol** by mixing **700 µL ethanol** and **300 µL nuclease-free water** in a **1.5 mL tube** and keep it **on ice**.
2) Thaw the reagents and keep all reagents **on ice**.
3) Keep a tube (**1 mL**) of **nuclease-free water** at **50 ºC**.
4) Bring the **Agencourt AMPure XP** to **room temperature**.

Quantification of PCR product

1) Prepare the working solution by adding **199 µL** of dsDNA HS **Buffer** and **1 µL** of dsDNA HS **Reagent** into a **1.5 mL tube**.
2) Add **199 µL** of **working solution** into each of the Qubit Assay tubes.
5) Add **1 µL DNA** sample and mix by **vortexing 2-3 sec**.
6) Incubate at **room temperature** for **2 min**.
7) On the home screen of Qubit 3.0 fluorometer, press **DNA** and then select **dsDNA High Sensitivity**. Press **Run samples**, select the **sample volume (1 µL)** and **units (ng/µL)**.
8) Insert the sample tube into the sample chamber, close the lid and press **Read tube**. Record the data.

Self-ligation

1) Dilute the PCR product to **2 ng/µL** using nuclease-free water.
2) Transfer **90 µL** of **diluted amplicons** into a **0.2 mL PCR tube**.
3) Add **10 µL** of **Blunt/TA Ligase Master Mix**. **Gently mix** flicking the tube followed by **spinning down** for **10 sec**. Incubate the tube for **15 min** at **10 °C** and **gently mix** by flicking the tube. **Spin down** for **10 sec** and incubate for **10 min** at **25 °C**.

(During incubation, perform step 4-6)

Reverse phase clean-up

4) **Vortex AMPure XP beads** and transfer **200 µL** into a new **1.5 mL tube**.
5) Place the tube **on magnet** for **2 min** to allow beads to pellet. Remove **120 µL** of **supernatant** (Be careful: do not disturb the beads) and place it **inside the lid** of the tube. **Discard** remaining **80 µL supernatant** (Be careful: do not disturb the beads). Transfer **100 µL** of clear buffer **from the lid** and discard the remaining buffer.
6) **Remove** the tube from magnet and **gently vortex** to resuspend the beads. Keep at **room temperature** until use.
7) Add **35 µL (0.35x) concentrated beads** from step 5 to the **100 µL self-ligation mix** from step 3. **Gently mix** by pipetting up and down. Incubate for **2 min** at **room temperature**.

(Keep the remaining **65 µL concentrated beads** at **room temperature** for use later)

8) Place the tube **on the magnet** and allow beads to pellet. Transfer **135 µL** of **supernatant** into a new **1.5 mL tube**.

Short amplicon clean-up

9) Add **67.5 µL (0.5x)** of **AMPure XP beads** to the liquid from step 8. **Gently vortex** and incubate for **2 min** at **room temperature**.

10) Place the tube **on the magnet** and allow beads to pellet. **Discard** the **supernatant (202.5 µL)**.

11) Keep the tube **on magnet** and add **200 µL** of freshly prepared ice-cooled **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

12) **Repeat** the wash step **once** more.

13) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Incubate for **10 sec** at **50 °C**. (Be careful: Not to overdry)

14) Resuspend the beads with **15 µL nuclease-free water** kept at 50° C. Incubate for **2 min** at **room temperature**.

15) Return the tube back **to magnet**, allow beads to pellet and transfer **15 µL** amplicons into a clean **1.5 mL tube**.

Removal of linear molecules

16) Combine the following reagents (**Plasmid-Safe™ ATP-Dependent DNase**) in a **0.2 mL PCR tube**. Incubate for **15 min** at **37 °C**.

Reagent	Volume
Purified self-ligated amplicons	15 µL
Nuclease-free water	2 µL
25 mM ATP	2 µL
10x Reaction Buffer	5 µL
Plasmid-Safe DNase (10U)	1 µL
Total	**25 µL**

17) Add **12.5 µL (0.5x)** of **AMPure XP beads** to the liquid from step 16. **Gently vortex** and incubate for **2 min** at **room temperature**.

18) Place the tube **on the magnet** and allow beads to pellet. **Discard** the **supernatant (37.5 µL)**.

19) Keep the tube **on magnet** and add **200 µL** of freshly prepared ice-cooled **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

20) **Repeat** the wash step **once** more.

21) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Incubate for **10 sec** at **50 °C**. (Be careful: Not to overdry)

22) Resuspend the beads with **10 µL nuclease-free water** kept at 50° C. Incubate for **2 min** at **room temperature**.

23) Return the tube back **to magnet**, allow beads to pellet and transfer **2.5 µL** amplicons into each clean **0.2 mL PCR tube**.

24) Quantify **DNA concentration** using **Qubit**. (aim at ~0.5 ng/µL)

Rolling Circle Amplification (RCA)

(Perform RCA in **3 replicates** and include **negative controls** using nuclease-free water.)

25) Prepare the **amplification mix** in a **0.2 mL PCR tube**.

Reagent	Volume/sample	4 samples
Nuclease-free water	9.3 µL	37.2 µL
Reaction buffer	2.5 µL	10 µL
dNTPs	2.5 µL	10 µL
Enzyme 1	2.5 µL	10 µL
Enzyme 2	0.7 µL	2.8 µL
Total	**17.5 µL**	**70 µL**

26) Combine **2.5 µL** of **self-ligated amplicons** from step 23 and **2.5 µL Buffer D** in a **0.2 mL PCR tube**. **Gently mix** by pipetting up and down. Incubate for **no more than 3 min** at **room temperature**.

27) Add **2.5 µL Buffer N** to **5 µL** of **incubated sample** from step 26. **Gently mix** by pipetting up and down.

28) Add **17.5 µL** of **amplification mix** from step 25 to the 7.5 µL DNA mix from step 27. **Gently mix** by pipetting up and down. Incubate for **120-150 min** at **29.5 °C**.

(After **90 min**, test the assay efficiency by measuring the **DNA concentration**. If RCA result in appropriate concentration (**60-80 ng/µL**), then the mixture can be taken to the next step. If the concentration is not sufficient, incubate for another 30-60 min and measure DNA concentration again. Minimum required concentration is **53.3 ng/µL**.)

Enzymatic de-branching

29) Combine three RCA replicates (**3 x 21 µL**) in a **0.2 mL PCR tube**.

30) Add **2 µL T7 endonuclease I** to 63 µL RCA products and **gently mix** with use of wide bore tips. Incubate for **5 min** at **room temperature**.

Mechanical fragmentation & clean-up

31) Transfer **65 µL RCA product** into a **g-TUBE** using wide bore pipette tips. Centrifuge at **1800 rpm** for **6 min** or until the entire mix passes through the hole. **Reverse** the g-TUBE and centrifuge at **1800 rpm** for **6 min** or until entire reaction mix passes through the fragmentation hole.

32) **Gently vortex** the **concentrated beads** and add **22.75 µL (0.35x)** of **concentrated beads** to **65 µL** of **fragmented RCA products**. **Gently mix** by pipetting up and down. Incubate for **2 min** at **room temperature**.

33) Place the tube **on magnet** and allow beads to pellet. **Discard** the **supernatant (87.75 µL)**.

34) Keep the tube **on magnet** and add **200 µL** of freshly prepared ice-cooled **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

35) **Repeat** the wash step **once** more.

36) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Incubate for **10 sec** at **50 °C**. (Be careful: Not to overdry)

37) Resuspend the beads with **65 µL nuclease-free water** kept at 50° C. Incubate for **2 min** at **room temperature**.

38) Return the tube back **to magnet**, allow beads to pellet and transfer **63 µL fragmented RCA product** into a clean **0.2 mL PCR tube**.

39) Quantify **DNA concentration** using Qubit.

Secondary enzymatic de-branching & clean-up

40) Add **2 µL T7 endonuclease I** to 63 µL fragmented RCA product from step 10. Incubate for **5 min** at **37 °C**.

41) Add **29.25 µL (0.45x) concentrated beads** to 65 µL fragmented RCA products from step 11. **Gently mix** by pipetting up and down. Incubate for **2 min** at **room temperature**.

42) Place the tube **on magnet** and allow beads to pellet. **Discard** the **supernatant (94.25 µL)**.

43) Keep the tube **on magnet** and add **200 µL** of freshly prepared ice-cooled **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

44) **Repeat** the wash step **once** more.

45) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Incubate for **10 sec** at **50 °C**. (Be careful: Not to overdry)

46) Resuspend the beads with **55 µL nuclease-free water** kept at 50° C. Incubate for **2 min** at **room temperature**.

47) Return the tube back **to magnet**, allow beads to pellet and transfer **53 µL fragmented RCA product** into a clean **0.2 mL PCR tube**.

48) Quantify **DNA concentration**.

Gap-filling and dA-tailing & clean-up

49) Combine the following reagents in a **0.2 mL PCR tube**. **Gently mix** by pipetting up and down **10 times** using wide bore tips. Incubate for **10 min** at **20 °C** and **10 min** at **65 °C** then **hold** at **4 °C**.

Reagent	Volume
RCA product (26.5 ng/µL)	53 µL
FFPE DNA Repair Buffer	3.5 µL
NEBNext Ultra II End Prep Buffer	3.5 µL
NEBNext FFPE DNA Repair Mix	2 µL
Ultra II End Prep enzyme mix	3 µL
Total	**65 µL**

50) Add **32.5 µL (0.5x) AMPure XP beads** to 65 µL fragmented RCA products from step 20. **Gently vortex**. Incubate for **4 min** at **room temperature**.

51) Place the tube **on magnet** and allow beads to pellet. **Discard** the **supernatant (97.5 µL)**.

52) Keep the tube **on magnet** and add **200 µL** of freshly prepared ice-cooled **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

53) **Repeat** the wash step **once** more.

54) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Incubate for **10 sec** at **50° C**. (Be careful: Not to overdry)

55) Resuspend the beads with **35 µL nuclease-free water** kept at 50° C. Incubate for **4 min** at **room temperature**.

56) Return the tube back **to magnet**, allow beads to pellet and transfer **33 µL fragmented RCA product** into a clean **1.5 mL tube**.

57) Quantify **DNA concentration**.

1D^2 adapter ligation

58) Combine the following reagents in a **1.5 mL tube**.

Reagent	Volume
End-repaired DNA	29 µL
LNB (**mix by flicking after adding LNB**)	12.5 µL
Quick T4 DNA Ligase	5 µL
1D^2 Adapter	3.5 µL
Total	**50 µL**

Gently mix by **flicking** the tube and **spin down**. Incubate for **20 min** at **room temperature**.

59) Resuspend **AMPure XP beads** by **vortexing**. Add **20 µL AMPure XP beads** to the reaction. Gently mix by **flicking** the tube. Incubate on a **rotator mixer** for **20 min** at **room temperature**.

60) **Spin down** and place the tube **on magnet** for **~5 min**. **Discard** the **supernatant (70 µL)**.

61) Keep the tube **on magnet** and add **200 µL** of freshly prepared **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

62) **Repeat** the wash step **once** more.

63) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Allow to dry for **~30 sec**. (Be careful: Not to overdry)

64) **Remove** the tube from the magnet. Resuspend the beads with **63 µL nuclease-free water**. Incubate for **20 min** at **room temperature**.

65) Return the tube back **to magnet**, allow eluate to clear and colourless. Transfer **63 µL** repaired DNA into a clean **1.5 mL tube**.

66) Quantify **DNA concentration** of **1 µL** repaired DNA using Qubit.

Adapter ligation

67) Combine the following reagents in a **1.5 mL tube**. Gently mix by **flicking** the tube and **spin down**. Incubate for **20 min** at **room temperature.**

Reagent	Volume
1D^2 adapted DNA	60 µL
LNB (**mix by flicking after adding LNB**)	25 µL
Quick T4 DNA Ligase	10 µL
Adapter Mix II (AMII)	5 µL
Total	**100 µL**

68) Resuspend **AMPure XP beads** by **vortexing**. Add **40 µL AMPure XP beads** to the reaction. Gently mix by **flicking** the tube. Incubate on a **rotator mixer** for **10 min** at **room temperature**.

69) **Spin down** and place the tube **on magnet** for **~5 min**. **Discard** the **supernatant (140 µL)**.

70) **Remove** the tube from magnet and add **140 µL LFB/SFB**. Resuspend the beads by **flicking** the tube.

71) Return the tube **on magnet** and allow beads to pellet. **Discard** the supernatant (**140 µL**).

72) **Repeat** the wash step **once** more.

73) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual supernatant. Allow to dry for **~30 sec**. (Be careful: Not to overdry)

74) **Remove** the tube from the magnet. Resuspend the beads with **15 µL** Elution Buffer **(EB)**. Incubate for **20 min** at **room temperature**.
(For HMW DNA, incubate at 37 °C can improve the recovery of long fragments.)

75) Return the tube back **to magnet**, allow eluate to clear and colourless and transfer **15 µL ligated DNA** into a clean **1.5 mL tube**.
(Store the library **on ice** until ready to load.)

76) Quantify **DNA concentration** with Qubit. Expected concentration should be **16-20 ng/µL (240-300 ng)**.

Priming and loading the flow cell

77) Mix the Sequencing Buffer (**SQB**) and Flush Buffer (**FB**)tubes by **vortexing**, **spin down** and **return to ice**.

78) **Spin down** the Flush Tether (**FLT**), mix by **pipetting**, and **return to ice**.

79) **Open the lid** of MinION and slide the flow cell' s **priming port cover clockwise** so that priming port is visible.

80) Set a P1000 pipette to 200 µL. **Insert the tip** into the priming port and **turn the wheel** until the dial shows **220-230 µL** or until seeing a small volume of buffer entering the tip. Visually check that there is **continuous buffer** from the priming port across the sensor array.

81) Prepare priming mix: add **30 µL** Flush Tether (**FLT**) directly to the tube of Flush Buffer (**FB**). Mix by pipetting up and down.

82) Load **800 µL** of **priming mix** into the flow cell via the **priming port** (Be careful: do not introduce air bubbles). Wait for **5 min**.

83) Thoroughly mix the Loading Beads (**LB**) by **pipetting** immediately before use.

84) Prepare the library in a **0.2 mL PCR tube**.

Reagent	Volume
Sequencing Buffer (SQB)	37.5 µL
Loading Beads (LB) (**mixed before use**)	25.5 µL
DNA library	12 µL
Total	**75 µL**

85) Gently lift the **SpotON sample port cover** to make the SpotON sample port accessible.

86) Load **200 µL** of **priming mix** into the flow cell via the **priming port** (Be careful: do not introduce air bubbles).

87) **Gently mix** the **prepared library** by pipetting up and down just prior to loading.

88) Add **75 µL** of **library** to the flow cell via the **SpotON sample port** in a **dropwise** fashion. (Ensure each drop flows into the port before adding the next)

89) Gently replace the **SpotON sample port cover** (making sure the bung enters the SpotON port).

90) Close the **priming port cover** and replace the **MinION lid**.

B.4 Genomic DNA 1D² sequencing protocol

Checklist

- ☐ Genomic DNA
- ☐ Agencourt AMPure XP beads
- ☐ 100% Ethanol
- ☐ Nuclease-free water
- ☐ NEBNext FFPE DNA Repair Mix
- ☐ NEBNext End repair/dA-tailing Module
- ☐ Quick T4 Ligase
- ☐ g-TUBE
- ☐ 1D² Ligation Sequencing Kit (SQK-LSK309)
- ☐ R9.5.1 Flow Cell
- ☐ 1.5 mL Eppendorf DNA LoBind tubes
- ☐ 0.2 mL PCR tubes
- ☐ P1000, P100, P10 pipette and tips
- ☐ Ice bucket with ice
- ☐ Magnetic rack
- ☐ Eppendorf Microcentrifuge 5424R
- ☐ Minicentrifuge
- ☐ Thermal cycler
- ☐ Vortex mixer
- ☐ Qubit 3.0 fluorometer
- ☐ Qubit dsDNA HS Assay Kit

Preparation

1) Prepare **1 mL** of **70% ethanol** by mixing **700 µL ethanol** and **300 µL nuclease-free water** in a **1.5 mL tube**.
2) Thaw the reagents and keep all reagents in the **NEBNext FFPE DNA Repair Mix**, **NEBNext End repair/dA-tailing Module**, **Quick T4 Ligase** and **Ligation Sequencing kit (LNB, 1D², LFB/SFB, EB, AMII, SQB, LB, FLT** and **one tube of FB) on ice**.
3) **LNB/FLT**: Spin down and mix by pipetting.

   LFB/SFB, **SQB** and **PFB/FB**: mix by vortexing and spin down.
4) Bring the **Agencourt AMPure XP** to **room temperature**.
5) Check flow cell.

DNA fragmentation

1) Transfer **1.5 µg** genomic DNA in **50 µL (30 ng/µL)** to the **g-TUBE**.
2) Centrifuge at **6000 rpm** for **1 min** at **room temperature**. If DNA remains in the upper chamber, centrifuge again for **1 min**.
3) **Invert** the g-TUBE and centrifuge at **6000 rpm** for **1 min** to collect the fragmented DNA. If DNA remains in the upper chamber, centrifuge again for **1 min**. Remove g-TUBE.
4) Transfer **50 µL fragmented DNA** to a **1.5 mL tube**.

DNA repair and end-prep

5) Combine the following reagents in a **0.2 mL PCR tube**.

Reagent	Volume
1.5 µg (> 10 kbp) or 250 fmol (< 10 kbp) DNA If **8 kbp**, use **27 ng/µL**. If 2 kbp, use 6.8 ng/µL	48 µL
NEBNext FFPE Repair Buffer	3.5 µL
NEBNext FFPE DNA Repair Mix	2 µL
NEBNext Ultra II End Prep Buffer	3.5 µL
Ultra II End Prep enzyme mix	3 µL
Total	**60 µL**

Gently mix by **flicking** the tube and **spin down**. Incubate for **15 min** at **20 ºC** and **5 min** at **65 ºC**. Transfer into a **1.5 mL tube**.

6) Resuspend **AMPure XP beads** by **vortexing**. Add **60 µL AMPure XP beads** to the end-prep reaction. Gently mix by **flicking** the tube. Incubate on a **rotator mixer** for **10 min** at **room temperature**.
7) **Spin down** and place the tube **on magnet** and allow beads to pellet. **Discard** the **supernatant (120 µL)**.
8) Keep the tube **on magnet** and add **200 µL** of freshly prepared **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

9) **Repeat** the wash step **once** more.

10) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Briefly allow to dry. (Be careful: Not to overdry)

11) **Remove** the tube from the magnet. Resuspend the beads with **32 µL nuclease-free water**. Incubate for **4 min** at **room temperature**.

12) Return the tube back **to magnet**, allow eluate to clear and colourless. Transfer **32 µL eluate** into a clean **1.5 mL tube**.

13) Quantify **DNA concentration** of **1 µL** repaired DNA using Qubit. Expected recovery should be **>67%**.

1D² adapter ligation

14) Combine the following reagents in a **1.5 mL tube**.

Reagent	Volume
End-repaired DNA	29 µL
LNB (**mix by flicking after adding LNB**)	12.5 µL
Quick T4 DNA Ligase	5 µL
1D² Adapter	3.5 µL
Total	**50 µL**

Gently mix by **flicking** the tube and **spin down**. Incubate for **20 min** at **room temperature**.

15) Resuspend **AMPure XP beads** by **vortexing**. Add **20 µL AMPure XP beads** to the reaction. Gently mix by **flicking** the tube. Incubate on a **rotator mixer** for **20 min** at **room temperature**.

16) **Spin down** and place the tube **on magnet** for **~5 min**. **Discard** the **supernatant (70 µL)**.

17) Keep the tube **on magnet** and add **200 µL** of freshly prepared **70% ethanol** (Do not disturb the pellet). **Discard** the ethanol (**200 µL**).

18) **Repeat** the wash step **once** more.

19) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual ethanol. Allow to dry for **~30 sec**. (Be careful: Not to overdry)

20) **Remove** the tube from the magnet. Resuspend the beads with **63 µL nuclease-free water**. Incubate for **20 min** at **room temperature**.

21) Return the tube back **to magnet**, allow eluate to clear and colourless. Transfer **63 µL** repaired DNA into a clean **1.5 mL tube**.

22) Quantify **DNA concentration** of **1 µL** repaired DNA using Qubit.

Adapter ligation

23) Combine the following reagents in a **1.5 mL tube**.

Reagent	Volume
1D² adapted DNA	60 µL
LNB (**mix by flicking after adding LNB**)	25 µL
Quick T4 DNA Ligase	10 µL
Adapter Mix II (AMII)	5 µL
Total	**100 µL**

Gently mix by **flicking** the tube and **spin down**. Incubate for **20 min** at **room temperature**.

24) Resuspend **AMPure XP beads** by **vortexing**. Add **40 µL AMPure XP beads** to the reaction. Gently mix by **flicking** the tube. Incubate on a **rotator mixer** for **20 min** at **room temperature**.

25) **Spin down** and place the tube **on magnet** for **~5 min**. **Discard** the **supernatant (140 µL)**.

26) **Remove** the tube from magnet and add **140 µL LFB/SFB**. Resuspend the beads by **flicking** the tube.

27) Return the tube **on magnet** and allow beads to pellet. **Discard** the supernatant (**140 µL**).

28) **Repeat** the wash step **once** more.

29) Briefly **spin down** and put the tube back **on the magnet**. **Pipette off** any residual supernatant. Allow to dry for **~30 sec**. (Be careful: Not to overdry)

30) **Remove** the tube from the magnet. Resuspend the beads with **15 µL** Elution Buffer **(EB)**. Incubate for **20 min** at **room temperature**.
(For HMW DNA, incubate at 37 ºC can improve the recovery of long fragments.)

31) Return the tube back **to magnet**, allow eluate to clear and colourless and transfer **15 µL ligated DNA** into a clean **1.5 mL tube**.
(Store the library **on ice** until ready to load.)

32) Quantify **DNA concentration** of **1 µL** repaired DNA using Qubit.

Priming and loading the flow cell

33) **Open the lid** of MinION and slide the flow cell' s **priming port cover clockwise** so that priming port is visible.

34) Set a P1000 pipette to 200 µL. **Insert the tip** into the priming port and **turn the wheel** until the dial shows **220-230 µL** or until seeing a small volume of buffer entering the tip. Visually check that there is **continuous buffer** from the priming port across the sensor array.

35) Add **30 µL** Flush Tether (**FLT**) directly to the tube of Flush Buffer (**FB**). Mix by **pipetting** up and down.

36) Load **800 µL** of **priming mix** into the flow cell via the **priming port** (Be careful: do not introduce air bubbles). Wait for **5 min**.

37) Thoroughly mix the Loading Beads (**LB**) by **pipetting** immediately before use.

38) Prepare the library in a **0.2 mL PCR tube**.

Reagent	Volume
Sequencing Buffer (SQB)	37.5 µL
Loading Beads (LB) (**mixed before use**)	25.5 µL
DNA library	12 µL
Total	**75 µL**

39) Gently lift the **SpotON sample port cover** to make the SpotON sample port accessible.

40) Load **200 µL** of **priming mix** into the flow cell via the **priming port** (Be careful: do not introduce air bubbles).

41) **Gently mix** the **prepared library** by pipetting up and down just prior to loading.

42) Add **75 µL** of **library** to the flow cell via the **SpotON sample port** in a **dropwise** fashion. (Ensure each drop flows into the port before adding the next)

43) Gently replace the **SpotON sample port cover** (making sure the bung enters the SpotON port).

44) Close the **priming port cover** and replace the **MinION lid**.