Aalborg Universitet

Characterization of Microbial Communities

From Fragments of Genes to Full Genomes

Brandt, Jakob

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Brandt, J. (2019). *Characterization of Microbial Communities: From Fragments of Genes to Full Genomes*. Aalborg Universitetsforlag.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: July 05, 2025

CHARACTERIZATION OF MICROBIAL COMMUNITIES

FROM FRAGMENTS OF GENES TO FULL GENOMES

BY
JAKOB BRANDT

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

CHARACTERIZATION OF MICROBIAL COMMUNITIES

FROM FRAGMENTS OF GENES TO FULL GENOMES

by

Jakob Brandt

AALBORG UNIVERSITY
DENMARK

Dissertation submitted December 2019

.

Dissertation submitted: 31-12-2019

PhD supervisor: Prof. MSO Mads Albertsen,
Aalborg University

Assistant PhD supervisors: Associate Prof. Morten Dueholm,
Aalborg University

Senior Researcher Søren M. Karst,
Aalborg University

PhD committee: Professor Jeppe Lund Nielsen (chairman)
Aalborg University

Tenure Track Scientist Anja Spang
Royal Netherlands Institute of Sea Research

Associate Professor Kasper Urup Nielsen
Aarhus University

PhD Series: Faculty of Engineering and Science, Aalborg University

Department: Department of Chemistry and Bioscience

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-581-9

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Jakob Brandt

Printed in Denmark by Rosendahls, 2020

ENGLISH SUMMARY

Microorganisms inhabit every environment on this globe. They have an enormous impact on biogeochemical cycles, sustain entire ecosystems, keep us healthy and make us sick. Traditionally, studying microorganisms has relied on microscopy and culture-based methods, but the sheer number of microorganisms and their diversity render these approaches impractical for uncovering all the existing microbial diversity. In this PhD project, culture-independent methods primarily relying on DNA-sequencing techniques were used to investigate microbial communities from different environments in unprecedented throughput.

The first part of the project focused on 16S rRNA gene amplicon sequencing of microbial communities in drinking water. For many years, this method has been heavily used for identification of microbes in multiple different environments. Although this method also has been widely used for analyzing the microbiome of low-biomass environments (such as drinking water), no comprehensive attempts had been made to illuminate the inherent method biases relating to drinking water communities. We investigated the impact of DNA extraction and primer choice on the observed microbial community. We found drastic differences relating to both factors. Furthermore, we estimated the detection limit of the 16S rRNA gene amplicon sequencing method relating to drinking water samples. We also demonstrated that contamination for samples with bacterial concentrations in the same range as drinking water was of no concern with our updated workflow.

The second part of the PhD project focused on the relatively newly discovered super-phylum of Asgard archaea, which are the closest known relatives to the eukaryotes in the tree of life. This discovery has brought more attention to the discussions of eukaryogenesis and the topology of the tree of life. Our first aim was to uncover ‘hidden’ diversity from this super-phylum. Applying a recently published high-throughput technique for generating high-quality SSU rRNA sequences using synthetic long-read sequencing by molecular tagging, we were able to retrieve more than 100,000 full-length archaea sequences. From our data-set of 16S rRNA gene sequences, we were able to uncover more than 250 new species-level Asgard archaeal OTUs. Furthermore, a family-level cut-off revealed 33 novel Asgard families. Our next aim was to extract Asgard archaea genomes from complex environments using metagenomics. Using a combination of long and short read technologies with cutting-edge assembly strategies, we successfully retrieved 16 genome bins belonging to Asgard archaea and verified presence of multiple eukaryotic signature proteins common to all.

DANSK RESUME

Mikroorganismer findes i stort set alle miljøer på Jorden. De har en enorm indvirkning på biogeokemiske kredsløb, og de er essentielle for at opretholde økosystemer. Mikroorganismer holder os sunde og gør os syge. Derfor er vigtigheden i at studere Jordens mikroorganismer stor. Traditionelt set har studier af mikroorganismer været afhængige af mikroskopi og kulturbaserede metoder, men det store antal mikroorganismer og deres diversitet har gjort disse metoder upraktiske til at afdække hele den eksisterende mikrobielle diversitet. I dette ph.d.-projekt blev kultur-uafhængige metoder (primært i form af DNA-sekventeringsteknikker) brugt til at undersøge mikrobielle samfund fra forskellige miljøer.

Den første del af projektet vedrørte 16S rRNA amplikon-sekventering af mikrobielle samfund i drikkevand. Denne metode har i mange år været bredt anvendt til identifikation af mikrober i flere forskellige miljøer. På trods af at denne metode også er blevet anvendt i vid udstrækning til analyse af mikrobiomet i drikkevand, var der ikke blevet gjort omfattende forsøg på at belyse de metode-biaser, der vedrører mikrobiomet i drikkevand. Vi undersøgte effekten af DNA-ekstraktion- og primer-valg på det observerede mikrobielle samfund, hvor vi fandt drastiske forskelle relateret til begge faktorer. Desuden estimerede vi detektionsgrænsen for 16S rRNA amplikon-sekventering, der vedrørte drikkevandsprøver, og påviste at kontaminering af prøver med bakteriekoncentrationer i det samme interval som drikkevand ikke påvirkedeanalysen.

Den anden del af ph.d.-projektet fokuserede på den relativt nyopdagede superfylum Asgard archaea, der er de nærmeste kendte slægtninge til eukaryoterne i Livets Træ. Denne opdagelse har fornyet opmærksomheden på diskussionerne vedrørende eukaryogenese og topologien i Livets Træ. Vores første mål var at afdække 'skjult' diversitet fra Asgard archaea. Med anvendelse af en nyligt publiceret metode til sekventering af SSU-rRNA-sekvenser af høj kvalitet, var vi i stand til at producere mere end 100.000 fuld-længde archaea-sekvenser. Ud fra vores datasæt med 16S rRNA-gensekvenser var vi i stand til at finde mere end 250 nye Asgard archaea OTUer på artsniveau, hvorimod et cut-off på familieniveau afslørede 33 nye Asgard-familier. Vores næste mål var at samle Asgard archaea genomer fra komplekse miljøer ved hjælp af metagenomics. Med en kombination af long og short read teknologier samt avancerede assembly-strategier fik vi med succes lokaliseret 16 Asgard genomer hvor tilstedeværelse af adskillige eukaryote signatur proteiner.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude towards my supervisor Mads Albertsen for giving me the opportunity to work on a highly interesting project. Your keen guidance and support have been greatly appreciated throughout the entire period and have shaped me as a person. There has always been plenty of opportunities to participate in top conferences and workshops all around the world, which has made for some very educational and memorable moments.

I would also like to like to send a very big thank you to all my colleagues in the Albertsen Lab as well as the Center for Microbial Communities – you have always made it exciting to go to work every day (of course some days more exciting than others). In particular, I would like to express thanks to my office mates for their great sense of humor, knowledge-sharing, and assistance with R and Unix related problems.

I also want to show gratitude towards the Ettema Lab for providing me with the opportunity to go abroad and be part of their lab. A special thank you should be directed at Will and Felix from the Ettema Lab for taking very good care of me for two months and for teaching me everything I know about handling and performing data analysis in Unix.

Finally, I would like to acknowledge all my friends and family that have supported me throughout the last three years.

TABLE OF CONTENTS

Chapter 1 – Introduction………………………………………….………… 13

1.1 Characterization of Microorganisms……………….……………………… 14

16S ribosomal RNA gene amplicon sequencing…………………….…………. 14

Full-length 16S ribosomal RNA gene sequencing using unique molecular identifiers

………………………………………………………………………………16

Microbial exploration using metagenomics………….………….……….………18

1.2 The Importance of Databases………….……………………………21

Chapter 2 – Aim…………………………………………………………..24

Chapter 3 – Preamble for body of work……………………………………25

3.1 Identification of Microorganisms in Drinking Water……………...…25

3.2 Asgard archaea…………………………………...………………..26

Chapter 4 – Conclusion and Perspective………..…………...…..………… 30

Literature list………..……………………………………...…..………… 32

Chapter 5 – Paper 1…………………………………………...…..………… 40

Chapter 6 – Paper 2……………………………………...…..… ……… 41

Chapter 7 – Paper 3………………………..………………...…..………… 42

Chapter 8 – Paper 4………………………..………………...…..………… 43

TABLE OF FIGURES

Figure 1-1 …………………………………………………………………………..14

Figure 1-2 …………………………………………………………………………..15

Figure 1-3 …………………………………………………………………………..17

Figure 1-4 …………………………………………………………………………..19

Figure 1-5 …………………………………………………………………………..22

Figure 3-1 …………………………………………………………………………..27

CHAPTER 1 INTRODUCTION

We live in a microbial world. Microbes are essential for life as we know it[1] and plays fundamental roles in the major biogeochemical cycles (Gilbert and Neufeld, 2014). Microbes have been estimated to globally constitute $4\text{-}6 \bullet 10^{30}$ cells and contain between $350\text{-}550 \bullet 10^{12}$ kg of carbon (Whitman, Coleman and Wiebe, 1998). The amount of microbial diversity inhabiting the Earth is more debated, with estimates ranging from millions to upward of one trillion species (Locey *et al.*, 2016; Louca *et al.*, 2019). Given the enormous importance of microbes and their omnipresence in our surroundings, it is only natural that humans have tried to discover, understand, grow, characterize and make use of microorganisms ever since the first microorganism was discovered in 1677 (Lane, 2015). Yet, the field of microbiology has changed dramatically from the time of Anton van Leeuwenhoek's initial discovery of microorganisms to today's increasingly molecular-based and data-driven approaches to microbiology. In essence, however, the work presented in this thesis tries to answer some of the same fundamental questions that have intrigued microbiologists from the beginning: Who are there? And what do they do?

The body of work for this PhD-project turned out to be 'thematically' divided. With the first part of the project focusing on the "Who are there?" in relation to drinking water environments. With amplicon sequencing of the 16S ribosomal RNA gene (rRNA) gaining popularity as a potential method for water quality monitoring, a need for methodological standardization had emerged. A comprehensive study was carried out, highlighting the impact different extraction methods and primer sets had on the observed microbial communities. The second part of the project replaced drinking water with the recently discovered super-phylum of Asgard archaea. However, the focus was still very much centered around the question of "who are there?". The amplicon sequencing technique was replaced with a novel method allowing high-throughput sequencing of the full-length 16S rRNA gene. This enabled us to successfully uncover a lot of novel Asgard archaea diversity. As a final study, a metagenomic approach was applied to potentially answer the "what do they do?" question. Based on Nanopore and Illumina data, a hydride genome assembly strategy was used to produce metagenome-assembled genomes (MAGs) where the full-length 16S rRNA gene sequences were leveraged to extract bins from Asgard archaea.

---

[1] As described by Jack A. Gilbert and Josh D. Neufeld in *Life in a World Without Microbes* (Gilbert and Neufeld, 2014).

1.1 CHARACTERIZATION OF MICROORGANISMS

Historically, investigation of microorganisms has relied on microscopes and culturing techniques. But the invention of DNA sequencing revolutionized the way we investigate microbial communities and has complemented the culture-dependent methods with a wide selection of culture-independent methods. This section is not intended to be an exhaustive description of all techniques invented to characterize microorganisms. Rather, this section highlights a selection of techniques or concepts comprising the methodological foundation for this thesis.

16S ribosomal RNA gene amplicon sequencing

In 1977, Carl Woese and George Fox demonstrated how sequencing data from the 16S rRNA gene could be used to reveal the three primary kingdoms of bacteria, archaea and eukarya (Woese and Fox, 1977; Pace, Sapp and Goldenfeld, 2012). Since the method was pioneered in the seventies, sequencing of the 16S rRNA gene has become a standard method in the toolkit of microbiologists. Even today, the method is frequently used as demonstrated in Figure 1-1. In 2019 alone, the number of publications in Web of Science's databases constitute roughly 25% of the total number of publications in the last 20 years. The advent of next-generation sequencing technologies has facilitated large-scale and routine studies of microbial communities using the 16S rRNA gene as a phylogenetic marker gene.

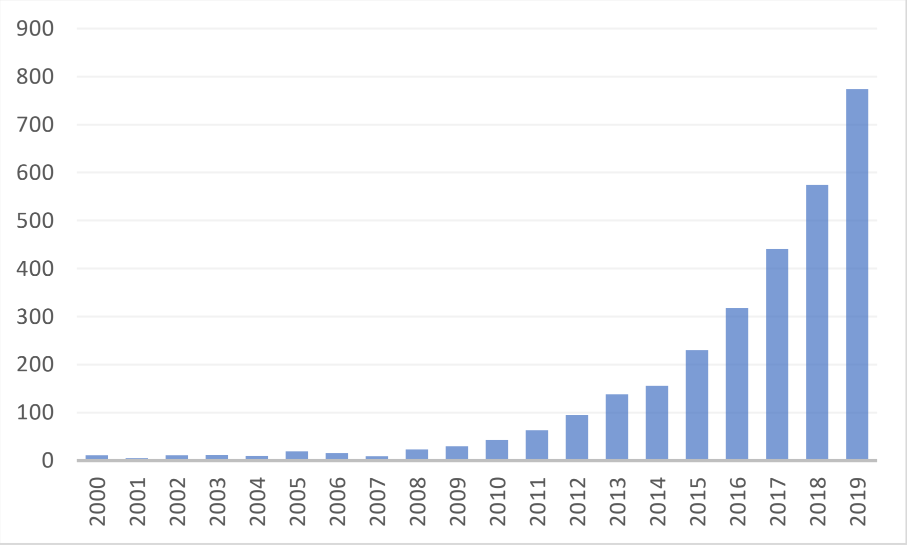


Figure 1-1 Number of publications in Web of Science (all databases) for the last 20 years relating to '16S rRNA amplicon sequencing' (search performed December 2019).

The widespread use of 16S rRNA gene sequencing for identification of bacteria and archaea can be ascribed to a few key features (McDonald *et al.*, 2012; van Loosdrecht *et al.*, 2016). Mainly, the gene encodes a functional RNA-molecule forming a part of the small sub-unit of the prokaryotic ribosome involved in protein synthesis. This is an essential function for all prokaryotic cells, making the 16S rRNA gene an ideal marker gene as the gene is universally conserved across prokaryotes. The gene also spans regions in its nucleotide sequence that are extremely conserved as well as regions that are variable as visualized in Figure 1-2. This facilitates primer-design targeting conserved regions, whereas variable regions are used for discrimination between different organisms.

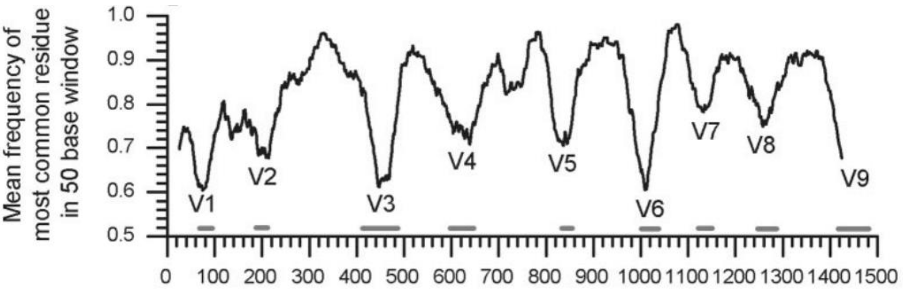


Figure 1-2 Illustration from (Ashelford et al., 2005) showing the average frequency of the most conserved base at a specific position based on all available bacteria sequences of the 16S rRNA gene. Variable regions are denoted from V1 to V9.

Given the widespread use of 16S rRNA gene amplicon sequencing, different variations of the method exist (Rausch *et al.*, 2019). However, all the different methods share the same core principle, which relies on amplifying 16S rRNA gene, or the region(s) of interest with the polymerase chain reaction (PCR) (Mullis *et al.*, 1986). In addition to increasing the number of amplicon copies, the PCR amplification also attaches each target molecule with adaptors. This results in a library with millions of 16S rRNA gene copies attached with adaptors at both the 3’ and 5’ end of the molecules. For Illumina-based sequencing, the adaptors enable the PCR-products to be attached to the flow cell during sequencing and also serve as barcodes allowing multiplexing of numerous samples per sequencing run ((Illumina), 2017). After sequencing, the 16S rRNA amplicon data is required to be computationally processed. A multitude of bioinformatic software are available for processing of amplicon data such as mothur and QIIME, although the overall concept remains the same (Schloss *et al.*, 2009; Caporaso *et al.*, 2010). Reads from the sequencing are quality filtered by removing any low-quality reads. Often, the quality filtered reads are subsequently clustered into operational taxonomic units (OTUs) based on the sequence identity of the reads. For full-length sequences, a sequence identity of 97 % is commonly applied as a proxy for species level OTUs, although different thresholds for OTU-clustering have been discussed before (Yarza *et al.*, 2014; van Loosdrecht *et al.*, 2016). Following clustering, reads belonging to the different OTUs are counted and a single, representative sequence from each OTU is picked out for taxonomic classification. Quite recently, however, there have been calls for using amplicon sequencing variants (ASVs) (Callahan, McMurdie and Holmes, 2017; Dueholm *et al.*, 2019). Using ASVs would allow the region to be resolved down to the level of single-nucleotide differences. This eliminates the need for clustering of OTUs and yields a much better resolution of taxa (Callahan, McMurdie and Holmes, 2017).

Classification consists of matching OTU or ASV sequences against a reference database of existing 16S rRNA gene sequences. The outcome of the bioinformatic steps is an OTU table containing information about the number of times each OTU was observed in each sample.

Substantial advantages are related to the use of 16S rRNA amplicon sequencing over conventional culture-based approaches. In theory, 16S rRNA amplicon sequencing allows for detection of all prokaryotic diversity and not just organisms readily cultured in lab settings. Additionally, the method includes abundance information of each organism. However, analyses with 16S rRNA gene sequences also come with inherent limitations. The predominant limitation is primer bias as no single primer-pair can target all prokaryotes (Albertsen *et al.*, 2015; Tremblay *et al.*, 2015). Furthermore, the abundance information obtained is biased as the number of 16S rRNA gene copies vary from species to species (Větrovský and Baldrian, 2013). In fact, the genomic copy number of the 16S rRNA gene has been demonstrated to vary between 1-15 in bacteria (Kembel *et al.*, 2012). Another bias related to the method – and to DNA-based methods in general – is extraction bias. Studies have highlighted the importance of the DNA extraction step and showed the impact different extraction methods can have on the subsequent analyses (Salter *et al.*, 2014; Albertsen *et al.*, 2015).

Full-length 16S ribosomal RNA gene sequencing using unique molecular identifiers

As mentioned in the paragraph above, amplicon-based identification of microorganisms is complicated by primer-bias, and sequencing data only contain sequence information for a fragment of the 16S rRNA gene. In 2018, a paper by Karst and colleagues described a novel method for full-length sequencing of the 16S rRNA gene without primer bias (Karst *et al.*, 2018). A schematic representation is displayed in Figure 2-1. Primer bias is avoided by using purified RNA as input for an initial gel electrophoresis. For prokaryotes, the cell's total amount of RNA is 82–90% ribosomal RNA (Blazewicz *et al.*, 2013). Hence, by size-selecting the RNA content of a complex sample on a gel, aliquots of rRNA fractions representing the entire community can be obtained. In order to convert the isolated rRNA molecules to double stranded DNA, a polyA-tail is ligated to all rRNA-molecules and functions as a priming site for first-strand cDNA synthesis using reverse transcription. For synthesis of the second-strand cDNA molecule, a second generic priming site is ligated to the opposite end of the polyadenylated cDNA molecule. During both first- and second-strand synthesis, adaptors containing unique tags are incorporated at each end of the target molecules (Burke and Darling, 2016; Salk, Schmitt and Loeb, 2018). The unique tags consist of 15 random nucleotides (this represents more than one billion possible tag combinations, 4^15 = 1,073,741,824), allowing each target molecule to have a unique molecular identifier. The adaptors also include generic priming sites needed for PCR amplification of the tagged target molecules. An initial PCR-amplification is needed to increase the number of uniquely tagged 16S rRNA gene molecules in order to perform another size-selection that removes partial and truncated products. A limited number of the size-selected, uniquely tagged 16S rRNA gene amplicons (100,000 – 1,000,000) are used as input for a PCR amplification, resulting in a clonal amplicon library consisting of thousands of copies of uniquely tagged amplicons.

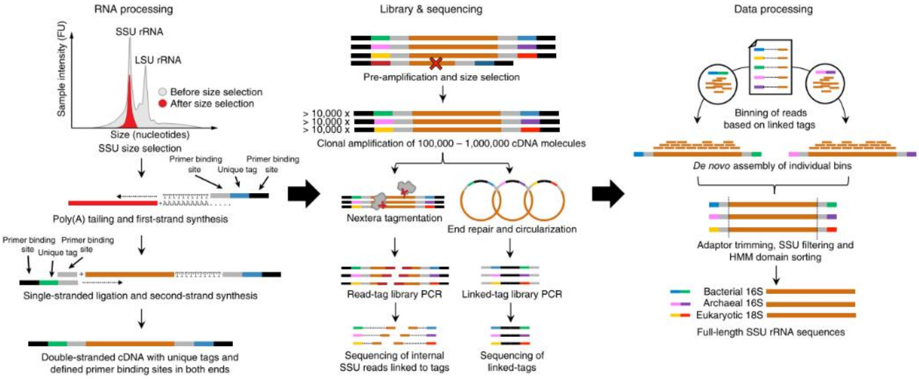

RNA processing
Sample intensity (FU)
SSU rRNA
LSU rRNA
Before size selection
After size selection
Size (nucleotides)
SSU size selection
Primer binding site
Unique tag
Primer binding site
AAAAAAAAAA
TTTTTTTTTT
Poly(A) tailing and first-strand synthesis
Primer binding site
Unique tag
Primer binding site
Single-stranded ligation and second-strand synthesis
Double-stranded cDNA with unique tags and defined primer binding sites in both ends
Library & sequencing
Pre-amplification and size selection
> 10,000 x
> 10,000 x
> 10,000 x
Clonal amplification of 100,000 – 1,000,000 cDNA molecules
Nextera tagmentation
End repair and circularization
Read-tag library PCR
Linked-tag library PCR
Sequencing of internal SSU reads linked to tags
Sequencing of linked-tags
Data processing
Binning of reads based on linked tags
De novo assembly of individual bins
Adaptor trimming, SSU filtering and HMM domain sorting
Bacterial 16S
Archaeal 16S
Eukaryotic 18S
Full-length SSU rRNA sequences

Figure 1-3A grafic representation of the workflow for producoing full-length 16S rRNA gene sequences. Illustration revised from (Karst et al., 2018).

The clonal amplicon library is used as input for two separate library preparations: a read-tag library and a linked-tag library. The read-tag library is produced with an Illumina Nextera DNA library prep kit. The kit applies a transposome that both randomly fragments the full-length 16S rRNA amplicons and tag them with adapter sequences. Thus, paired-end sequencing of the read-tag library generates data where internal amplicon sequences are related to single, unique, tag-reads. In order to connect read-tag sequences belonging to the same parent molecule, a linked-tag library is prepared. The amplicon molecules from the clonal library are circularized with intra-molecular ligation. The ligation is followed by a PCR amplification of the linked-tags, which subsequently can be identified by Illumina sequencing. *De novo* assembly is used to assemble the original 16S rRNA gene sequence. In similar fashion to conventional amplicon data, the full-length 16S rRNA gene sequences are subject to filtration, clustering and classification steps prior to data analyses.

The method is also compatible with the use of primers targeting the full-length of the 16S rRNA gene. Instead of using RNA as starting material for the protocol, purified DNA is used as input for an initial PCR amplification. Custom-made primers are required that (in addition to the primer-region) contain a tail with a generic primer binding site and a unique molecular tag. The initial PCR should be carried out with the lowest possible number of cycles in order to minimize formation of chimeras and number of PCR-introduced errors. The initial PCR step is followed by another PCR amplification that targets all molecules containing the generic primer binding site. The purpose of this PCR step is to obtain a sufficient amount of PCR product for validation, size-selection and quantification. Although susceptible to primer-bias, this approach also results in 16S rRNA gene molecules with unique molecular tags at both ends. Once the target molecules have been tagged and size-selected, the workflow is identical to the primer-free version. Again, a limited number of tagged molecules are used for preparation of a clonal library, which are split in two and used for preparation of read-tag and linked-tag libraries.

The main advantage of the primer-based version is that the input for the method is in the range of a few nanograms of DNA. For the primer-free version, the required input of RNA for the initial size-selection is around 800-1,000 ng. In this way, using primers allows for analysis of even low-biomass samples and environments with restricted sampling options.

A few alternative approaches to high-throughput sequencing of full-length 16S rRNA genes exist. In 2019, two papers demonstrated how protocols using long-read sequencing platforms like PacBio and Oxford Nanopore can enable high-accuracy, full-length sequencing of the 16S rRNA gene (Callahan *et al.*, 2019; Karst *et al.*, 2019). The article by Karst and colleagues - combining long reads and unique molecular identifiers – even permits sequencing of the entire ribosomal operon. Also, a commercial option from Loop Genomics (www.loopgenomics.com) is available.

Microbial exploration using metagenomics

Despite the prevalent use of the 16S rRNA marker-gene for analyses of microbial communities, the information obtained from these experiments primarily answers the question of who are there. In order to investigate what the microorganisms (potentially) can do, we need to sequence not just one gene, but the entire genome. This can be achieved with metagenomic studies. The term 'metagenomics' refers to random sequencing of microbial DNA, without selecting for specific genes (Breitwieser, Lu and Salzberg, 2017). Already back in 1998, the term 'metagenome' was coined to describe the collective genomes of a soil environment (Handelsman *et al.*, 1998). In the last decade, tremendous progress in sequencing technologies and lowered cost of sequencing have facilitated a wealth of studies using a metagenomic approach (Goodwin, McPherson and McCombie, 2016). In addition, relatively new sequencing platforms such as Oxford Nanopore Technologies have had a large impact on the scientific community with long-read sequence becoming more easily accessible and further assisting recovery of genomes from metagenomes (Kono and Arakawa, 2019). The rapid development within the field of sequencing has also led to publications of several tools for processing, visualizing and handling of increasingly large sequencing datasets.

A brief outline of the typical process from sequencing of a complex sample to extraction of a single genome is illustrated in Figure 1-4.

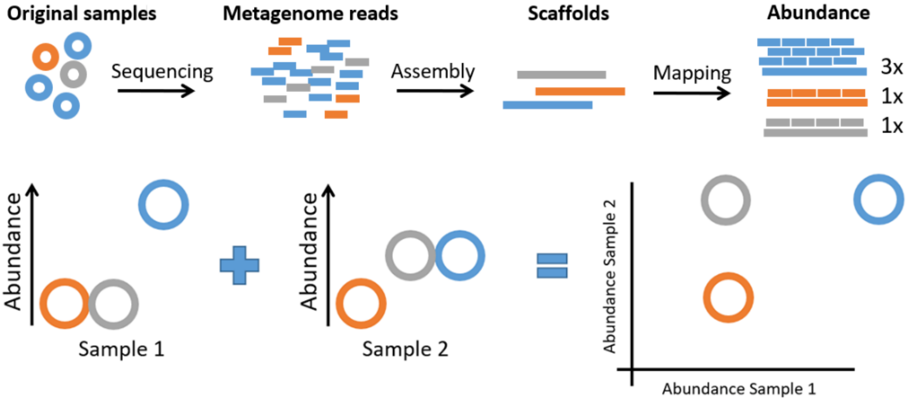


Figure 1-4 A simplified, step-wise overview of the workflow for extracting genome bins from environmental samples. The sample in the illustration comprises bacterial genomes from three different species: blue, grey and orange in varying abundance. Sequencing of the complex sample produces millions of reads that are assembled into large scaffolds based on overlaps between reads. Mapping the reads back to the scaffolds provides information about the abundance of each scaffold. This abundance information can be leveraged in the genome binning process where distinct clusters of scaffolds are separated into genome bins representing different microbial genomes.

As visualized in the figure above, the first part of the workflow is sequencing of samples. Currently, sequencing power and cost no longer constitute a limiting factor for most studies with high-throughput sequencing technologies routinely producing more than 100 gigabases of reads within 24 hours (Lightbody *et al.*, 2018). Prior to assembly, an initial quality assessment and filtering step of the raw metagenome reads may be performed. This is in particular relevant for error-prone platforms such as Oxford Nanopore, which generate reads with an error rate between 5-20% (Kono and Arakawa, 2019). However, this procedure can readily be carried out with software tools like NanoPack (De Coster *et al.*, 2018).

In essence, all assembly methods depend on the basic assumption that highly similar reads originate from the same position within a given genome (Nagarajan and Pop, 2013). Based on similarity between reads, contiguous sequences can be pieced together and in that way recover lost information from the sequencing run (Nagarajan and Pop, 2013). Numerous of different software are available for performing metagenome assemblies like Miniasm (Li, 2016), Unicycler (Wick *et al.*, 2017), CANU (Koren *et al.*, 2017) and Flye (Kolmogorov *et al.*, 2019) – just to name a few. All the programs each apply a unique strategy for the assembly process. However, each strategy can be divided into two paradigms of assembly: overlap-layout-consensus or De Bruijn graph (Nagarajan and Pop, 2013; Vries, Tsang and Grigoriev, 2018). Typically, assemblers using short-reads as input are based on de Bruijn graphs that deconstruct sequencing reads into overlapping k-mers, with each k-mer being a possible sub-string of length ‘k’ for a given read. A critical parameter is choosing an optimal k-mer size, as too long k-mers render it impossible to resolve repeats. Conversely, large k-mers may increase the number of errors in the assembly (Vries, Tsang and Grigoriev, 2018). The overlap–layout–consensus paradigm relies on performing all-vs-all pairwise alignments between the metagenome reads and is commonly used for long-read assemblers (Vries, Tsang and Grigoriev, 2018).

After completion of the assembly process, a genome binning step is required to extract the individual genomes from the assembly and divide them into separate ‘bins’. The main problems in reconstructing genomes from metagenomes relate to the genome binning step (Albertsen *et al.*, 2013; Nelson *et al.*, 2016). Early strategies for separating contigs in the assembly into distinct genome bins involved use of genetic patterns such as tetranucleotide frequencies, GC-content as well as k-mer distribution (Perry and Beiko, 2010; Mande, Mohammed and Ghosh, 2012). However, in recent years the use of differential coverage binning as a sequence-independent binning strategy (see Figure 1-4) has enhanced the recovery of genomes from metagenomic sequencing (Albertsen *et al.*, 2013; Sharon *et al.*, 2013). Nevertheless, genome binning of assemblies originating from complex samples with a vast amount of microbial diversity is still troublesome. Another aspect further complicating genome binning is microdiversity, which is defined as diversity of phylogenetically closely related organisms that display different metabolic activities and thus inhabit distinct niches (Nelson *et al.*, 2016). Discriminating sequencing data from closely related species can be near impossible. Manual selection of genome bins can be performed with tools such as mmgenome2 or Anvi'o (Eren *et al.*, 2015; Karst, Kirkegaard and Albertsen, 2016), but also automated binning pipelines are available (like MetaBat) which use both genomic patterns as well as abundance information (Kang *et al.*, 2015). Typically, the quality of the genome bins is estimated with CheckM (Parks *et al.*, 2015). CheckM estimates both completeness and contamination relying on marker genes that are specific to a genome's inferred lineage within a reference genome tree (Parks et al., 2015).

THE IMPORTANCE OF DATABASES

One of the most fundamental disciplines in microbiology is to identify microbes by matching sequencing data against a reference database and assign a taxonomy. Having a name for a sequenced organism is extremely valuable, as names can be linked with functions. For classifying 16S rRNA gene sequences, numerous reference databases exist. These reference databases range from ecosystem-specific like the MiDAS database for wastewater treatment systems (McIlroy *et al.*, 2017) to global and frequently used reference databases like the SILVA, Greengenes and RDP databases (McDonald *et al.*, 2012; Quast *et al.*, 2013; Cole *et al.*, 2014). Similarly, online resources like the Genomes OnLine Database and the Genome Taxonomy Database are available for genome-based analysis (Chaumeil, Hugenholtz and Parks, 2019; Mukherjee *et al.*, 2019). Whereas the Genomes OnLine Database functions as an extensive catalog of genome and metagenome sequencing projects around the world, the Genome Taxonomy Database specifically aims at establishing a standardized microbial taxonomy based on genome phylogeny. All these databases provide a crucial resource as they constitute a taxonomic framework for interpretation of both marker gene analyses and metagenomic surveys (McDonald *et al.*, 2012). Ideally, an identification of a target organism at the lowest taxonomic rank is desirable. Nevertheless, you can only assign a classification at a low taxonomic rank if the reference database contains a highly similar sequence to your input sequence. The SILVA reference database (version 132) contains around two million 16S rRNA gene sequences. As briefly mentioned in the introduction, this means a potentially large fraction of microbial diversity could still not be represented in the reference databases (Locey *et al.*, 2016; Louca *et al.*, 2019). This claim is supported in Figure 1-5, which is a revised illustration from Schloss and Co. (2016) displaying rarefaction curves for a dataset including 1,411,234 bacterial and 53,546 archaeal 16S rRNA gene sequences.

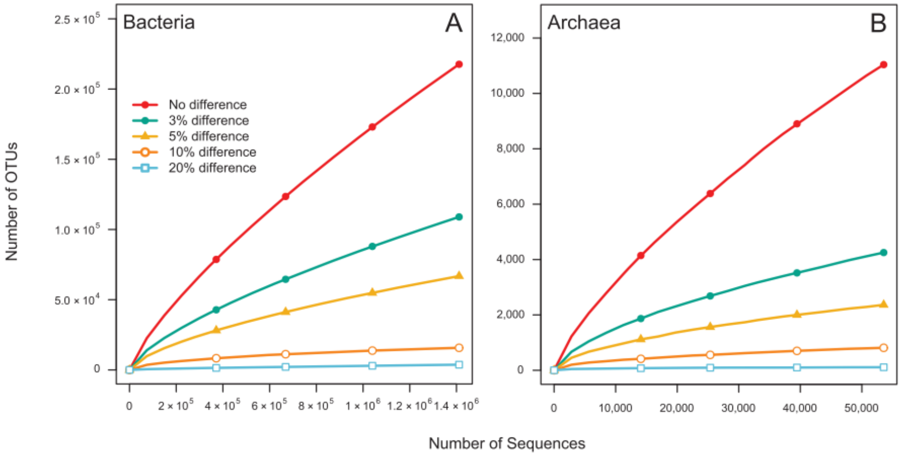


Figure 1-5 Rare faction curves for different OTU cut-offs for bacteria and archaea. The figure is revised from (Schloss et al., 2016).

The reference databases of rRNA gene sequences are also known to be skewed towards well-characterized environments and also susceptible to primer bias and chimerism (Karst *et al.*, 2018). Naturally, the genome reference databases are significantly smaller in magnitude. Currently, the Genome Taxonomy Database (release 04-RS89) includes a total of 143,512 bacterial genomes. These genomes represent roughly 23,458 species and only 2,392 archaeal genomes representing 1,248 species.

Considering that reference databases are/represent an integral resource in microbiology, it is a very important task to populate the databases with high quality sequences from yet undiscovered microbes. This is true for both the marker gene databases as well as the genome databases. As accounted for in section 1.1, the improvements in DNA sequencing coupled with developments of novel techniques will facilitate a large expansion of databases. Examples of this will be evident in **Paper 2** and **Paper 3** where state-of-art techniques have been used to produce large data-sets of full-length 16S rRNA gene sequences and genome bins, respectively.

CHAPTER 2 AIM

The overall aim of this PhD project was to apply high-throughput DNA analyses for identification of novel prokaryotic diversity, and ultimately, characterization of previously unidentified microbes. Specific objectives included the following:

1. Use full-length 16S rRNA gene analysis for selected ecosystems of interest in order to uncover microbial novelty and populate reference databases.

2. Apply state-of-the-art approaches within metagenomics in order to extract high-quality genomes from complex samples of interest.

CHAPTER 3 PREAMBLE FOR BODY OF WORK

The main output of this PhD-project is a collection of papers. In the following sections, a brief context to the different areas that the papers relate to will be provided.

3.1 DNA-BASED IDENTIFICATION OF MICROORGANISMS IN DRINKING WATER

Drinking water (DW) contains microorganisms, a fact often ignored to the public. Even DW entering the distribution system after treatment has bacterial concentrations of 10^3 to 10^5 cells/ml (Pinto, Xi and Raskin, 2012). The presence of microbial life in DW has been connected to harmful effects and contribute to deterioration of the infrastructure in distribution systems (Pinto, Xi and Raskin, 2012). Also, DW can harbor pathogens - estimations suggest 12 to 19 million annual cases of gastrointestinal illnesses are directly related to contaminated DW (Ashbolt, 2015; Shaw *et al.*, 2015). Historically, microbial characterization of DW has been based on culture-dependent methods of only a few indicator organisms. However, in recent years, DNA-based approaches have been more frequently applied for characterization of microbial communities in DW. Nevertheless, the DW environment has recently been described as under-investigated compared to other environments (Bruno *et al.*, 2017; Hull *et al.*, 2019). In 2019, an opinion paper entitled *Drinking Water Microbiome Project: Is it Time?* also argued for the need of more coordinated, large-scale projects of the DW environment (Hull *et al.*, 2019). Sequencing-based approaches to characterizing DW also have significant potential advantages to conventional methods. Mainly, sequencing-based methods allow to potentially detect almost any microorganism in DW. However, low-biomass samples like DW are vulnerable to biases and contamination. This was highlighted in an earlier study from Salter and colleagues (2014) where they demonstrated how contamination from lab reagents and extraction kits can critically impact sequence-based microbiome analyses. In particular, the extraction procedure was reported to contain several contaminating taxa (Salter *et al.*, 2014).

In **Paper 1** (published), we outlined a lack of standardization in the methodology of published literature relating to 16S rRNA gene profiling of DW. In an effort to highlight the need for a more uniform methodology, we examined the performance of two commonly used DNA extraction kits and three popular primer-sets for DW studies. We decided to examine DNA extraction and PCR-amplification. These steps comprise the most crucial parts of the analysis since you only detect what you can amplify, and you only amplify what you can extract (Albertsen *et al.*, 2015). Of the two DNA extraction kits tested, a marked difference was observed concerning yields, reproducibility and number of OTUs identified. A similarly stark contrast was observed for the primer-sets where large differences in OTU abundances were observed. Some of the primer-sets even proved incapable of detecting entire phyla (**Paper 1**).

Another aspect of **Paper 1** was to estimate a lower detection limit of amplicon sequencing - something very few have attempted to. This experiment was a highly relevant contribution to the discussion of the potential applicability of DNA-based monitoring of DW quality. For a sequencing-based method to be a feasible alternative/addition to conventional quality monitoring, ideally it should be able to detect low-abundance pathogens or organisms of ecological importance. The experiment was set up with bacteria-free water samples (1 L) spiked with *Escherichia coli* cells in different concentrations [10^1–10^6 cells/ml] spanning the typical range of DW. Therefore, the experiment could also provide insight into whether microbiome analyses from samples with bacterial concertation within the normal range would be impacted by contamination. The experiment demonstrated that a multitude of contaminating OTUs were discovered for the samples with the lowest concentration of bacteria. However, samples within the typical range of DW was not impacted by contamination (for the top 25 most abundant OTUs).

3.2 ASGARD ARCHAEA

Asgard archaea is a recently discovered super-phylum that has had a large impact on our understanding of evolution of life. The first "Asgard archaea" publication appeared in 2015 and used a metagenomic approach for obtaining genomic information from Lokiarchaeota (Spang *et al.*, 2015). Interestingly, phylogenomic analyses demonstrated that Lokiarchaeota form a monophyletic group with the eukaryotes (see Figure 3-1) as well as encoding a wide range of eukaryotic signature proteins (ESP) in their genomes (Spang *et al.*, 2015). This has changed our understanding of the tree of life and eukaryogenesis (Eme *et al.*, 2017; Spang, Caceres and Ettema, 2017; Castelle and Banfield, 2018). A two-domain-topology of the tree of life – where Eukarya emerge from within the archaea, opposed to the classical three-domain-topology, had been proposed earlier (Guy and Ettema, 2011; Kelly, Wickstead and Gull, 2011). The relatedness of the Asgard archaea to the eukaryotes combined with the many ESPs encoded in Asgard genomes strengthened a 2D-topology for the tree of life. Since 2015, the Asgard super-phylum has been expanded with Thor- Odin- and Heimdallarchaeota (Zaremba-

Niedzwiedzka *et al.*, 2017) and more recently the Helarchaeota (Seitz *et al.*, 2019). It is important to note, however, that the Asgard super-phylum has also caused some debate. More specifically, there have been claims that the affiliation between Eukarya and Lokiarchaeota can be attributed to an artefact from genome reconstruction (Da Cunha *et al.*, 2017). Even harder criticism has been voiced by Garg et al. claiming all Asgard genomes reconstructed from metagenomic data were artefacts and questioning metagenomics as a whole (Garg *et al.*, 2019). Ironically, the article by Garg et al. was published only a few days apart from a study authored by Imachi et al. (2019) that reported a decade-long isolation of a Lokiarchaeota from which a closed genome could be retrieved. Analyses from Imachi et al. indeed showed clear phylogenetic sistering between the isolated Lokiarchaeota archaeon and Eukarya. Their analyses also confirmed the presence of many ESPs also found in other Asgard archaea. Since then, another study has reported near-complete Lokiarchaeota genomes from metagenomic data and demonstrated that eukaryote-like features linked to Lokiarchaeota are not caused by contamination or assembly artefacts (Caceres *et al.*, 2019).

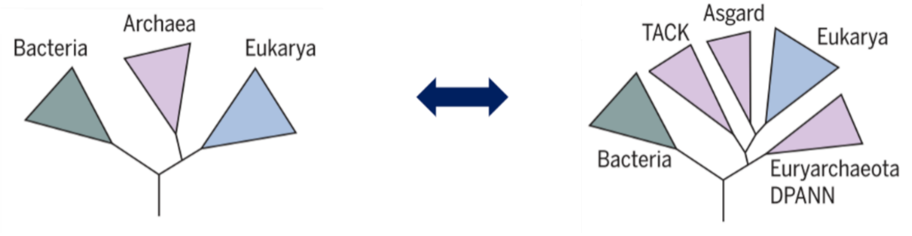


Figure 3-1 Illustration of the topology of the tree of life with a three-domain-topology and two-domain-topology respectively. Phylogenomic analyses have positioned Asgard archaea as a sister-group to eukaryotes favoring the 2D model on the right. The illustration is revised from (Spang, Caceres and Ettema, 2017).

The work presented in this PhD-project on Asgard archaea showcases how the novel techniques described in section 1.1 can be used to do a large-scale screening of potentially novel Asgard archaea (based on the 16S rRNA gene) and identify Asgard archaea bins from large metagenomic datasets.

In **Paper 2** (draft manuscript ready for submission), we aimed at uncovering some of the 'hidden' diversity of Asgard archaea based on 16S rRNA gene sequences. Currently, the amount of publicly available 16S rRNA gene data for Asgard archaea is rather limited. This is clearly reflected in the latest version of the SILVA database (release 138, dated December 18, 2019), which only includes 670 sequences classified as Asgardarchaeota (based on numbers from the non-redundant version where sequences have been clustered at 99% identity). In fact, archaea sequences in general are vastly underrepresented in the database owing to the fact that no good universal archaeal primer-sets exist (Karst *et al.*, 2018). However, by using the full-length 16S rRNA gene sequencing method described in section 1.1, we were able to produce ~32,000 (>1,200 bp) archaea sequences clustered at 99% identity. Comparatively, SILVA138 contains ~20,000 archaea sequences (99% identity). We also made a subset of our dataset consisting exclusively of Asgard archaea sequences for further analyses. A phylogenetic analysis showed that we were able to uncover Asgard sequences broadly across the super-phylum. In addition, the subset was combined with Asgard sequences from the SILVA database and clustered into OTUs with various similarity thresholds to indicate different taxonomic ranks. At a 97% identity threshold, more than 250 OTUs consisted of Asgard sequences exclusively from our study. Correspondingly, for a threshold of 86.5% identity (indicating family rank), 33 OTUs consisted exclusively of our Asgard sequences.

Overall, our study highlighted that plenty of Asgard diversity is still left to be discovered and, at the same time, demonstrated the high throughput of the method.

Paper 3 (in preparation) took basis in a large dataset consisting of both long and short read data from a sediment sample. A metagenomic assembly was constructed from the dataset with the aim of extracting genome bins belonging to Asgard archaea. Extraction of 16S rRNA gene sequences from the assembly was used for mapping against our Asgard archaea 16S rRNA gene sequences from **Paper 2**. In this way, we were able to specifically locate contigs belonging to Asgard bins. From the assembly, we identified 33 contigs encoding Asgard archaea 16S rRNA gene sequences. Of these 33 contigs, 29 of them were assigned to a bin from an automated binning software representing a total of 16 potential Asgard archaea bins.

The Asgard archaea bins were validated (besides documenting the presence of Asgard archaea 16S rRNA genes), by identifying ESPs. We were able to identify a wide range of ESPs for all of our 16 Asgard bins when searched against a list of 347 ESPs published by Han & Collins (2012). Further validation of the Asgard bins were based on phylogeny inferred from large subunit rRNA gene sequences. 23S sequences from our bins were incorporated into a database containing complete and cross-validated rRNA sequences of species from all known phyla (the SEREB database (Bernier *et al.*, 2018)) and clearly illustrated clustering of our bins within the Asgard group. In summary, our analyses – yet again – confirmed Asgard archaea as “true” and not artificial constructs as argued by Garg et al (2019) and Da Cunha et al. (2017). However, quality assessment of the Asgard bins showed a need for additional polishing and curation of the bins. Also, more in-depth analyses of all or a subset of the bins are still to be carried out. Nevertheless, the work presented in **Paper 3** so far clearly demonstrated the potential of a metagenomic approach for retrieving genomes from organisms of interest in complex samples using high-throughput, long and short read sequencing.

A fourth paper was included in this thesis (**Paper 4, co-author**) presenting a minireview on metatranscriptomics. Recently, it has become possible to perform analysis of transcriptomic data obtained from mixed communities. **Paper 4** comments on various computational strategies for handling antisense expression in metatranscriptomic datasets as well as the potential effects it may have on downstream analyses. In the near-future, transcriptomic analysis of communities containing Asgard archaea could provide valuable insight into their ecology.

CHAPTER 4 CONCLUSION

The fields of microbiology and molecular biology are currently advancing at an incredible rate. The rapid progression in these fields has been perfectly exemplified in the three-year run of this PhD-project.

At the start of this PhD-project, the work relied on datasets only containing information about a fragment of a single gene. In **Paper 1**, we used 16S rRNA gene amplicon data from drinking water samples to argue for the necessity of standardization of methodology. We highlighted the large impact different extraction methods and primer-sets may have on downstream analyses. Our message of methodological standardization is highly consistent with the recent call for a larger and more coordinated effort in drinking water microbiome research.

Along the PhD-project, fragments of gene sequences were replaced with full-length (>1200 bp) gene sequences. Using a recently developed method, we generated more than 100,000 16S rRNA gene sequences from the kingdom of archaea (**Paper 2**). With a specific interest in Asgard archaea, we successfully uncovered a plethora of novel Asgard archaea diversity. Classifying the Asgard sequences via phylogenetic analyses, we highlighted that different Asgard archaea were present for basically all the different Asgard phyla.

In the final part of the project, a giant leap was taken technology-wise as sequencing of single genes was substituted with sequencing of entire genomes. Mining a large dataset with long and short-read data from a sediment sample, we were able to identify a total of 16 genome bins containing 16S rRNA gene sequences from Asgard archaea (**Paper 3**). The genome bins were further validated as Asgard archaea by identifying a multitude of eukaryotic signature proteins.

LITERATURE LIST

Bibliography

(Illumina) (2017) 'Illumina sequencing introduction', (October), pp. 1–8. doi: http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

Albertsen, M. *et al.* (2013) 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes', *Nature Biotechnology*, 31(6), pp. 533–538. doi: 10.1038/nbt.2579.

Albertsen, M. *et al.* (2015) 'Back to basics - The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities', *PLoS ONE*. doi: 10.1371/journal.pone.0132783.

Ashbolt, N. J. (2015) 'Microbial Contamination of Drinking Water and Human Health from Community Water Systems', *Current Environmental Health Reports*, 2(1), pp. 95–106. doi: 10.1007/s40572-014-0037-5.

Ashelford, K. E. *et al.* (2005) 'At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies', *Applied and Environmental Microbiology*, 71(12), pp. 7724–7736. doi: 10.1128/AEM.71.12.7724-7736.2005.

Bernier, C. R. *et al.* (2018) 'Translation: The universal structural core of life', *Molecular Biology and Evolution*, 35(8), pp. 2065–2076. doi: 10.1093/molbev/msy101.

Blazewicz, S. J. *et al.* (2013) 'Evaluating rRNA as an indicator of microbial activity in environmental communities: Limitations and uses', *ISME Journal*. Nature Publishing Group, 7(11), pp. 2061–2068. doi: 10.1038/ismej.2013.102.

Breitwieser, F. P., Lu, J. and Salzberg, S. L. (2017) 'A review of methods and databases for metagenomic classification and assembly', *Briefings in Bioinformatics*. doi: 10.1093/bib/bbx120.

Bruno, A. *et al.* (2017) 'Exploring the under-investigated "microbial dark matter" of drinking water treatment plants', *Scientific Reports*, 7, p. 44350. doi: 10.1038/srep44350.

Burke, C. M. and Darling, A. E. (2016) ‘A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq’, *PeerJ*, 4, p. e2492. doi: 10.7717/peerj.2492.

Caceres, E. F. *et al.* (2019) ‘Near-complete Lokiarchaeota genomes from complex environmental samples using long and short read metagenomic analyses’, *bioRxiv*.

Callahan, B. J. *et al.* (2019) ‘High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution’, *Nucleic acids research*, 47(18), p. e103. doi: 10.1093/nar/gkz569.

Callahan, B. J., McMurdie, P. J. and Holmes, S. P. (2017) ‘Exact sequence variants should replace operational taxonomic units in marker-gene data analysis’, *ISME Journal*. Nature Publishing Group, 11(12), pp. 2639–2643. doi: 10.1038/ismej.2017.119.

Caporaso, J. G. *et al.* (2010) ‘QIIME allows analysis of high-throughput community’, *Nat Methods.*, 7(5), pp. 335–336. doi: 10.1038/nmeth.f.303.QIIME.

Castelle, C. J. and Banfield, J. F. (2018) ‘Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life’, *Cell*, pp. 1181–1197. doi: 10.1016/j.cell.2018.02.016.

Chaumeil, P., Hugenholtz, P. and Parks, D. H. (2019) ‘GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.’, *In preparation*, pp. 1–3.

Cole, J. R. *et al.* (2014) ‘Ribosomal Database Project: Data and tools for high throughput rRNA analysis’, *Nucleic Acids Research*, 42, pp. 633–642. doi: 10.1093/nar/gkt1244.

De Coster, W. *et al.* (2018) ‘NanoPack: Visualizing and processing long-read sequencing data’, *Bioinformatics*, 34(15), pp. 2666–2669. doi: 10.1093/bioinformatics/bty149.

Da Cunha, V. *et al.* (2017) ‘Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes’, *PLOS Genetics*. Edited by A. Rokas, 13(6), p. e1006810. doi: 10.1371/journal.pgen.1006810.

Dueholm, M. S. *et al.* (2019) ‘Comprehensive ecosystem-specific 16S rRNA gene databases with automated taxonomy assignment (AutoTax) provide species-level resolution in microbial ecology’, *bioRxiv*. doi: 10.1101/672873.

Eme, L. *et al.* (2017) ‘Archaea and the origin of eukaryotes’, *Nature Reviews Microbiology*, pp. 711–723. doi: 10.1038/nrmicro.2017.133.

Eren, A. M. *et al.* (2015) ‘Anvi ’ o : an advanced analysis and visualization platform for ‘ omics data’, *PeerJ*, pp. 1–29. doi: 10.7717/peerj.1319.

Garg, S. G. *et al.* (2019) ‘Anomalous phylogenetic behavior of ribosomal proteins in metagenome assembled genomes’, *bioRxiv*. doi: : http://dx.doi.org/10.1101/731091.

Gilbert, J. A. and Neufeld, J. D. (2014) ‘Life in a World without Microbes’, *PLoS Biology*, 12(12), pp. 1–3. doi: 10.1371/journal.pbio.1002020.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) ‘Coming of age: Ten years of next-generation sequencing technologies’, *Nature Reviews Genetics*, pp. 333–351. doi: 10.1038/nrg.2016.49.

Guy, L. and Ettema, T. J. G. (2011) ‘The archaeal “TACK” superphylum and the origin of eukaryotes’, *Trends in Microbiology*, pp. 580–587. doi: 10.1016/j.tim.2011.09.002.

Han, J. and J. Collins, L. (2012) ‘Eukaryotic Signature Proteins Of Giardia’, *Journal of Proteomics And Genomics Research*, 1(1), pp. 2–8. doi: 10.14302/issn.2326-0793.jpgr-12-101.

Handelsman, J. *et al.* (1998) ‘Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products’, *Chemistry and Biology*, 5(10). doi: 10.1016/S1074-5521(98)90108-9.

Hull, N. M. *et al.* (2019) ‘Drinking Water Microbiome Project: Is it Time?’, *Trends in Microbiology*. Elsevier Ltd, 27(8), pp. 670–677. doi: 10.1016/j.tim.2019.03.011.

Imachi, H. *et al.* (2019) ‘Isolation of an archaeon at the prokaryote-eukaryote interface’, *bioRxiv*, p. 726976. doi: 10.1101/726976.

Kang, D. D. *et al.* (2015) ‘MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities’, *PeerJ*, 2015(8), pp. 1–15. doi: 10.7717/peerj.1165.

Karst, S. M. *et al.* (2018) ‘Retrieval of a million high-quality , full-length microbial 16S and 18S rRNA gene sequences without primer bias’, *Nature Biotechnology*. doi: 10.1038/nbt.4045.

Karst, S. M. *et al.* (2019) ‘Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing’, *bioRxiv*, p. 645903. doi: 10.1101/645903.

Karst, S. M., Kirkegaard, R. H. and Albertsen, M. (2016) ‘Mmgenome : a Toolbox for Reproducible Genome Extraction From Metagenomes’, *bioRxiv*, Preprint, pp. 2014–2016. doi: 10.1101/059121.

Kelly, S., Wickstead, B. and Gull, K. (2011) ‘Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes’, *Proceedings of the Royal Society B: Biological Sciences*, 278(1708), pp. 1009–1018. doi: 10.1098/rspb.2010.1427.

Kembel, S. W. *et al.* (2012) ‘Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance’, *PLoS Computational Biology*, 8(10), pp. 16–18. doi: 10.1371/journal.pcbi.1002743.

Kolmogorov, M. *et al.* (2019) ‘Assembly of long, error-prone reads using repeat graphs’, *Nature Biotechnology*. Springer US, 37(5), pp. 540–546. doi: 10.1038/s41587-019-0072-8.

Kono, N. and Arakawa, K. (2019) ‘Nanopore sequencing: Review of potential applications in functional genomics’, *Development Growth and Differentiation*, 61(5), pp. 316–326. doi: 10.1111/dgd.12608.

Koren, S. *et al.* (2017) ‘Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.’, *Genome research*, 27(5), pp. 722–736. doi: 10.1101/gr.215087.116.

Lane, N. (2015) ‘The unseen World: Reflections on Leeuwenhoek (1677) “Concerning little animals”’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666). doi: 10.1098/rstb.2014.0344.

Li, H. (2016) ‘Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences’, *Bioinformatics*, 32(14), pp. 2103–2110. doi: 10.1093/bioinformatics/btw152.

Lightbody, G. *et al.* (2018) ‘Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application’, *Briefings in Bioinformatics*, (August), pp. 1–17. doi: 10.1093/bib/bby051.

Locey, K. J. *et al.* (2016) ‘Scaling laws predict global microbial diversity

Scaling laws predict global microbial diversity', *PNAS*, 113(21), pp. 5970–5975. doi: 10.1073/pnas.1521291113.

van Loosdrecht, M. C. M. *et al.* (2016) *Exprimental Methods in Wastewater Treatment*. 1st edn. (C) 2016 IWA Publishing. doi: 10.1017/CBO9781107415324.004.

Louca, S. *et al.* (2019) 'A census-based estimate of Earth's bacterial and archaeal diversity', *PLoS Biology*. doi: 10.1371/journal.pbio.3000106.

Mande, S. S., Mohammed, M. H. and Ghosh, T. S. (2012) 'Classification of metagenomic sequences: Methods and challenges', *Briefings in Bioinformatics*, 13(6), pp. 669–681. doi: 10.1093/bib/bbs054.

McDonald, D. *et al.* (2012) 'An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.', *The ISME journal*. Nature Publishing Group, 6(3), pp. 610–8. doi: 10.1038/ismej.2011.139.

McIlroy, S. J. *et al.* (2017) 'MiDAS 2.0: An ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups', *Database*, 2017(1), pp. 1–9. doi: 10.1093/database/bax016.

Mukherjee, S. *et al.* (2019) 'Genomes OnLine database (GOLD) v.7: Updates and new features', *Nucleic Acids Research*, 47(D1), pp. D649–D659. doi: 10.1093/nar/gky977.

Mullis, K. *et al.* (1986) 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986.', *Cold Spring Harb Symp Quant Biol.*, 51(1), pp. 263–73. doi: 10.1101/sqb.1986.051.01.032.

Nagarajan, N. and Pop, M. (2013) 'Sequence assembly demystified Niranjan', *Nature Reviews Genetics*. Nature Publishing Group, 14(3). doi: 10.1038/nrg3367.

Nelson, W. C. *et al.* (2016) 'Identification and Resolution of Microdiversity through Metagenomic', 82(1), pp. 255–267. doi: 10.1128/AEM.02274-15.Editor.

Pace, N. R., Sapp, J. and Goldenfeld, N. (2012) 'Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life', *Proceedings of the National Academy of Sciences of the United States of America*, 109(4), pp. 1011–1018. doi: 10.1073/pnas.1109716109.

Parks, D. H. *et al.* (2015) ‘CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes’, *Genome Research*, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

Perry, S. C. and Beiko, R. G. (2010) ‘Distinguishing microbial genome fragments based on their composition: Evolutionary and comparative genomic perspectives’, *Genome Biology and Evolution*, 2(1), pp. 117–131. doi: 10.1093/gbe/evq004.

Pinto, A. J., Xi, C. and Raskin, L. (2012) ‘Bacterial community structure in the drinking water microbiome is governed by filtration processes’, *Environmental Science and Technology*, 46(16), pp. 8851–8859. doi: 10.1021/es302042t.

Quast, C. *et al.* (2013) ‘The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools’, *Nucleic Acids Research*, 41(D1). doi: 10.1093/nar/gks1219.

Rausch, P. *et al.* (2019) ‘Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms’, *Microbiome*. Microbiome, 7(1), pp. 1–19. doi: 10.1186/s40168-019-0743-1.

Salk, J. J., Schmitt, M. W. and Loeb, L. A. (2018) ‘Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations’, *Nature Reviews Genetics*. Nature Publishing Group, 19(5), pp. 269–285. doi: 10.1038/nrg.2017.117.

Salter, S. J. *et al.* (2014) ‘Reagent and laboratory contamination can critically impact sequence-based microbiome analyses’, *BMC Biology*, 12(1), p. 87. doi: 10.1186/s12915-014-0087-z.

Schloss, P. D. *et al.* (2009) ‘Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities’, *Applied and Environmental Microbiology*, 75(23), pp. 7537–7541. doi: 10.1128/AEM.01541-09.

Schloss, P. D. *et al.* (2016) ‘Status of the archaeal and bacterial census: An update’, *mBio*, 7(3), pp. 1–10. doi: 10.1128/mBio.00201-16.

Seitz, K. W. *et al.* (2019) ‘Asgard archaea capable of anaerobic hydrocarbon cycling’, *Nature Communications*. Springer US, 10(1), p. 1822. doi: 10.1038/s41467-019-09364-x.

Sharon, I. *et al.* (2013) ‘Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization’, *Genome Research*, 23(1), pp. 111–120. doi: 10.1101/gr.142315.112.

Shaw, J. L. a. *et al.* (2015) ‘Using amplicon sequencing to characterize and monitor bacterial diversity in drinking water distribution systems’, *Applied and Environmental Microbiology*, (July), p. AEM.01297-15. doi: 10.1128/AEM.01297-15.

Spang, A. *et al.* (2015) ‘Complex archaea that bridge the gap between prokaryotes and eukaryotes’, *Nature*, 521(7551), pp. 173–179. doi: 10.1038/nature14447.

Spang, A., Caceres, E. F. and Ettema, T. J. G. (2017) ‘Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life.’, *Science (New York, N.Y.)*, 357(6351). doi: 10.1126/science.aaf3883.

Tremblay, J. *et al.* (2015) ‘Primer and platform effects on 16S rRNA tag sequencing’, *Frontiers in Microbiology*, 6(AUG), pp. 1–15. doi: 10.3389/fmicb.2015.00771.

Větrovský, T. and Baldrian, P. (2013) ‘The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses’, *PLoS ONE*, 8(2). doi: 10.1371/journal.pone.0057923.

Vries, R. P. de, Tsang, A. and Grigoriev, I. V. (2018) *Fungal Genomics methods and protocols*. doi: 10.1007/978-1-61779-040-9.

Whitman, W. B., Coleman, D. C. and Wiebe, W. J. (1998) ‘Prokaryotes: The unseen majority’, *Proc. Natl. Acad. Sci. USA*, 95, pp. 6578–6583.

Wick, R. R. *et al.* (2017) ‘Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads’, *PLoS Computational Biology*, 13(6), pp. 1–22. doi: 10.1371/journal.pcbi.1005595.

Woese, C. R. and Fox, G. E. (1977) ‘Phylogenetic structure of the prokaryotic domain: the primary kingdoms.’, *Proceedings of the National Academy of Sciences of the United States of America*, 74(11), pp. 5088–5090. doi: 10.1073/pnas.74.11.5088.

Yarza, P. *et al.* (2014) ‘Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences’, *Nature Reviews Microbiology*, 12(9), pp. 635–645. doi: 10.1038/nrmicro3330.

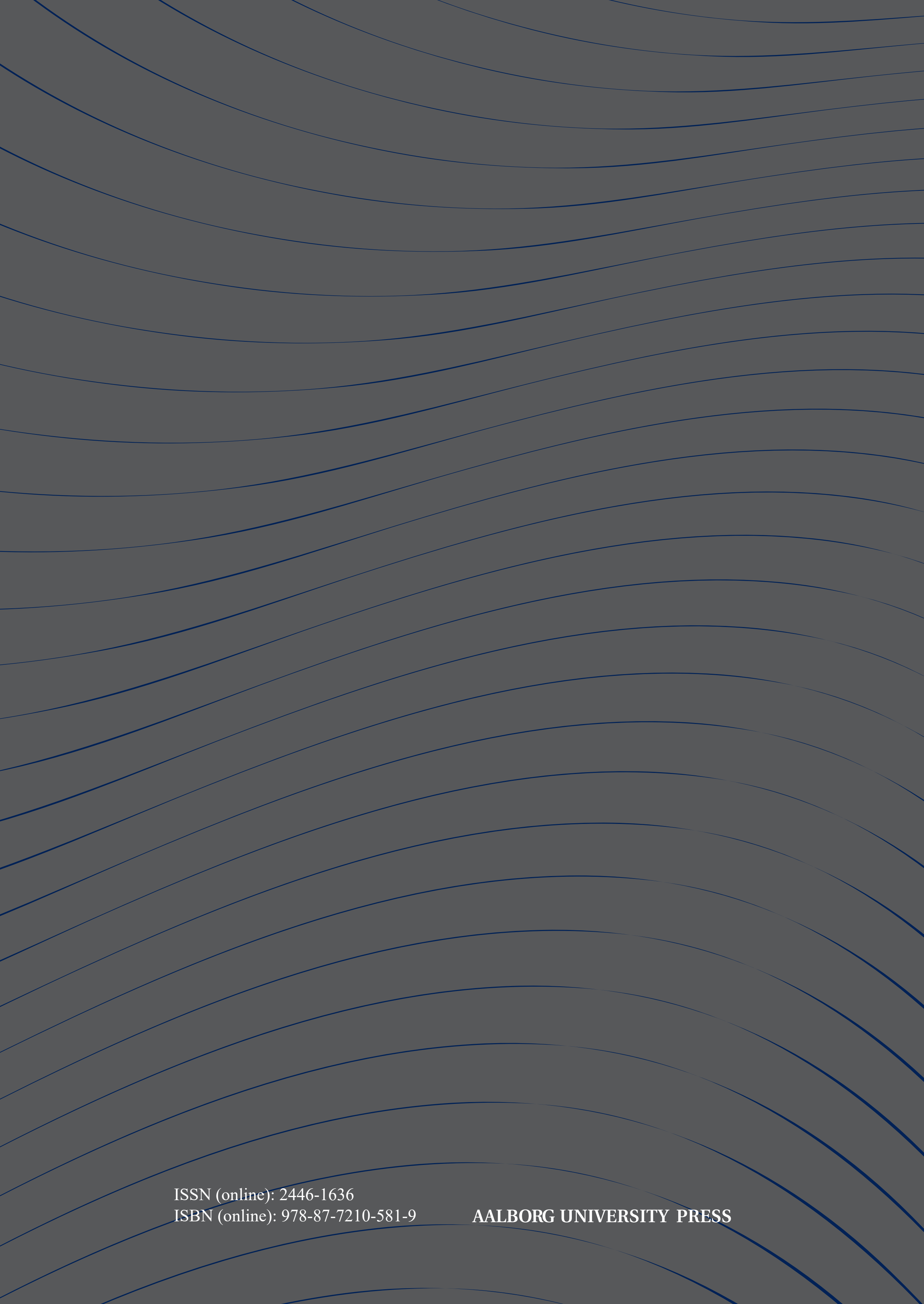
Zaremba-Niedzwiedzka, K. *et al.* (2017) ‘Asgard archaea illuminate the origin of eukaryotic cellular complexity’, *Nature*, 541(7637), pp. 353–358. doi: 10.1038/nature21031.

CHAPTER 5. INVESTIGATION OF DETECTION LIMITS AND THE INFLUENCE OF DNA EXTRACTION AND PRIMER CHOICE ON THE OBSERVED MICROBIAL COMMUNITIES IN DRINKING WATER SAMPLES USING 16S RRNA GENE AMPLICON SEQUENCING

CHAPTER 6. EXPANDING ASGARD ARCHAEA DIVERSITY WITH THOUSANDS OF HIGH-QUALITY 16S RRNA GENE SEQUENCES

CHAPTER 7. GENOMIC EXPLORATION OF ASGARD ARCHAEA FROM COMPLEX SAMPLES

CHAPTER 8. THE SIGNAL AND THE NOISE – CHARACTERISTICS OF ANTISENSE RNA IN COMPLEX MICROBIAL COMMUNITIES



ISSN (online): 2446-1636
ISBN (online): 978-87-7210-581-9

AALBORG UNIVERSITY PRESS