Microbial DNA Sequencing in Environmental Studies

THESIS FOR DOCTORAL DEGREE (Ph. D.)

By

Yue Hu

Principal Supervisor:
Docent Anders Andersson
KTH Royal Institute of Technology
School of Biotechnology

Co-supervisors:
Prof. Joakim Lundeberg
KTH Royal Institute of Technology
School of Biotechnology

Dr. Jürg Brendan Logue
Linnaeus University
Department of Biology and Environmental Science

Chair of the session:
Prof. Afshin Ahmadian
School of Biotechnology
KTH Royal Institute of Technology

Opponent:
Prof. Marja Tiirola
University of Jyväskylä, Finland
Department of Biological and Environmental Science

Examination Board:
Docent Silke Langenheder
Department of Ecology and Genetics;
Uppsala University

Docent Annelie Hedström
Luleå university of Technology.
Department of Civil, Environmental and Natural Resources Engineering

Dr. Andreas Sjödin
FOI, Swedish Defence Research Agency

This thesis will be publicly defended on April 21st 2017, 10:00
Location: Air-and-Fire lecture hall
Science for Life Laboratory Stockholm

Docent Lukas Käll
KTH Royal Institute of Technology
School of Biotechnology
Yue Hu (2017): **Microbial DNA Sequencing in Environmental Studies**, Division of Gene Technology, School of Biotechnology, KTH Royal Institute of Technology, Science for Life Laboratory, Stockholm, Sweden

**Abstract**

The field of microbial ecology has just entered a new era of rapid technological development and generation of big data. The high-throughput sequencing techniques presently available provide an opportunity to extensively inventorize the blueprints of life. Now, millions of microbes of natural microbial communities can be studied simultaneously without prior cultivation. New species and new functions (genes) can be discovered just by mining sequencing data. However, there is still a tremendous number of microorganisms not yet examined, nor are the ecosystem functions these carry out. The modern genomic technologies can contribute to solve environmental problems and help us understand ecosystems, but to most efficiently do so, methods need to be continuously optimised.

During my Ph. D. studies, I developed a method to survey eukaryotic microbial diversity with a higher accuracy, and applied various sequencing-based approaches in an attempt to answer questions of importance in environmental research and ecology. In **PAPER-I**, we developed a set of 18S rRNA gene PCR primers with high taxonomic coverage, meeting the requirements of currently popular sequencing technologies and matching the richness of 18S rRNA reference sequences accumulated so far. In **PAPER-II**, we conducted the first sequencing-based spatial survey on the combined eukaryotic and bacterial planktonic community in the Baltic Sea to uncover the relationship of microbial diversity and environmental conditions. Here, the 18S primers designed in **PAPER-I** and a pair of broad-coverage 16S primers were employed to target the rRNA genes of protists and bacterioplankton for amplicon sequencing. In **PAPER-III**, we integrated metagenomic, metabarcoding, and metatranscriptomic data in an effort to scrutinise the protein synthesis potential (i.e., activity) of microbes in the sediment at a depth of 460 m in the Baltic Sea and, thus, disclosing microbial diversity and their possible ecological functions within such an extreme environment. Lastly, in **PAPER-IV**, we compared the performance of *E. coli* culturing, high-
throughput sequencing, and portable real-time sequencing in tracking wastewater contamination in an urban stormwater system. From the aspects of cost, mobility and accuracy, we evaluated the usage of sequencing-based approaches in civil engineering, and for the first time, validated the real-time sequencing device in use within water quality monitoring.

In summary, these studies demonstrate how DNA sequencing of microbial communities can be applied in environmental monitoring and ecological research.

**Keywords:** DNA sequencing; Metabarcoding; Microbial ecology; Baltic Sea; Microbial community; Illumina; Oxford Nanopore; Source tracking; Stormwater
Sammanfattning


Sammanfattningsvis visar dessa studier hur DNA-sekvensering av mikrobiella samhällen kan tillämpas för miljöövervakning och ekologisk forskning.
Nyckelord: DNA sekvensering; Metabarkodning; Mikrobiell ekologi; Östersjön; Mikrobiellt samhälle; Illumina; Oxford Nanopore; Spårning; Dagvatten
Publications included in this thesis:


PAPER-IV: Yue O. O. Hu, Nelson Ndegwa, Johannes Alneberg, Sebastian Johansson, Jürg B. Logue, Mikael Huss, Max Källe, Joakim Lundberg, Jens Fagerberg, Anders F. Andersson. “Stationary and portable sequencing-based approaches for tracing wastewater contamination in urban stormwater systems” (manuscript)

Note: “Yue O. O. Hu” is the name used for Yue Hu’s publications.
Other publications during Ph.D. studies:


II. Sepideh Lamei, Yue O. O. Hu, Tobias C. Olofsson, Anders F. Andersson, Eva Forsgren, Alejandra Vasquez. "Optimised growth conditions and improved molecular identification methods for honeybee specific Lactic Acid Bacteria" PLoS One, 2017; (Accepted for publication)


Note: “Yue O. O. Hu” is the name used for Yue Hu's publications.
Table of Contents

Preface ................................................................. 11

1. Introduction ...................................................... 13
   1.1 Microbes and their roles in the environment ................. 13
   1.2 Use of microbes as indicators ............................... 14

2. Microbial Community Profiling Methods ....................... 16
   2.1 Phenotypic methods .......................................... 16
   2.2 Molecular-based approaches ................................ 17
      2.2.1 DNA fingerprinting-based approaches ................. 18
      2.2.2 DNA hybridization-based approaches .................. 19
      2.2.3 DNA sequencing-based approaches ..................... 19

3. Sequencing Technologies Suitable for Metagenomics .......... 22

4. Selecting and designing primers ................................ 25

5. Sample preparation .............................................. 27
   5.1 DNA extraction .............................................. 27
   5.2 Preparation of sequencing libraries ........................ 28

6. Processing of sequencing data .................................. 33
   6.1 Quality control and reads pre-processing .................. 33
   6.2 Clustering and denoising .................................... 34
   6.3 Annotation ................................................... 36

7. Results and Discussion ........................................... 38
   7.1 Systematic design of 18S rRNA gene primers (PAPER-I) .... 38
   7.2 Microbial plankton biogeography in the Baltic Sea (PAPER-II) 39
   7.3 Inspecting protein synthesis potential of the microbial community at Landsort Deep, the Baltic Sea (PAPER-III) 40
   7.4 Sequencing-based approaches to track faecal contamination in stormwater systems (PAPER-IV) 41

8. Concluding Remarks and Future Outlook ....................... 43

Acknowledgements ................................................ 45

References ......................................................... 51
Preface

In this thesis, I review the brief history of how microbial communities have been studied and show how sequencing technologies have revolutionized the field of microbiology. I also discuss the tools available and procedures followed in sequencing-based microbial studies, from selection of sequencing platform, primer design, sequencing library preparation, to data processing. The four studies presented here provide an idea of how to plan and apply sequencing techniques in environmental and applied microbial studies.
1. Introduction

1.1 Microbes and their roles in the environment

Microbes, the tiny organisms invisible to the naked eye, are ubiquitous, extremely abundant, and hugely diverse. These often neglected organisms carry out essential functions with regard to ecology, ecosystems, climate, as well as the health of other organisms including us human beings.

The sheer number of microbial organisms and species surpasses everything known in biology and even astronomy. For instance, the number of bacteria in the oceans ($13 \times 10^{28}$) is a 100 million times greater than the number of stars in the universe; strung together, all $10^{31}$ viruses existing on Earth would stretch for a 100 million light years [1]; one gram of soil contains more than $10^{10}$ of bacteria [2] and one milliliter of water from the open ocean harbours $5 \times 10^5$ microbes [3,4]. These numbers illustrate the extreme abundance of microbes on Earth but behind these lies so much more. Sixty years ago, Albert Jan Kluyver and Cornelius Bernardus van Niel had estimated that microbes contributed to half of the ‘living protoplasm’ on Earth [5]. In 1998, William B. Whitman estimated that the total amount of carbon (C) from all living microbes amounts to more than 60% of C found in plants or that the respective total amounts of nitrogen (N) and phosphorus (P) present in microbes are 10 times greater than in plants, respectively. Thus, microbes account for the largest pool of nutrients in all living organisms [6].

Fast reproduction rates and high dispersal capabilities along with evolutionary forces (e.g., genetic mutation, drift, and natural selection) have allowed microbes to become hugely diverse over the billions of years of their existence. Their collectively vast gene pool underpins the immense number of functions they carry out and essentially their impact on Earth’s ecosystems [6]. Unicellular microbes were the first life forms to emerge on Earth [7,8]. With time, microbes started using solar energy to fix carbon dioxide and water to synthesise sugars and in doing so producing enough oxygen to change Earth’s atmosphere and allowing oxygen-breathing life forms to come into existence. Nowadays, microbial phytoplankton are still carrying out half of all the photosynthetic processes occurring on Earth, thereby serving as a gigantic C pump for
the planet and foundation for all aquatic food webs [9]. However, microbial plankton not only circulate C but also other elements (e.g., N and P) essential to cellular organisms on a global scale [10,11]. In terrestrial ecosystems, on the other hand, soil microbes are crucial for soil formation and important regulators of plant productivity, abundance, as well as diversity [12]. From grasslands to forests, nitrogen-fixing microbes provide plants with N and P for up to 80% of their annual demand. The interaction of microbes and plants fundamentally affects the dynamics of terrestrial plants, which consequently results in visible changes in our environment [13,14]. Also in the atmosphere do airborne microbes play important roles, such as participating in the process of precipitation [15].

Apart from their role in nature, microbes also assume central roles in a variety of applied and engineering fields. Ever since the Neolithic period, humans have used microbes for the preservation of food (e.g., for fermenting milk or fruit), albeit without knowing their true nature [16]. A new phase of biotechnology was entered during World War I, when methods were developed to produce acetone and glycerol for ammunition. The discovery of Penicillin in 1929 has benefited the health of both animals and human beings from then on [17]. Nowadays, microbes are widely used in many fields that are closely related to our daily life, such as pharmaceutical manufacturing and wastewater processing.

### 1.2 Use of microbes as indicators

As the environment has been shown to structure microbial communities in that environmental conditions select for microbes with specific niches [18–20], the presence and abundance of microbes can be used as indicators when assessing an environment and its condition [21]. For instance, eutrophication occurs when a waterbody receives excess nutrients (mainly P and N) from sewage, fertilizers, detergents, etc., which in turn favours microalgae blooms in the surface water. Such blooms result in large quantities of organic matter (i.e., dead blooming microalgae) sinking to the deeper layers of the water body and eventually the sediments. As this organic matter degrades, oxygen is depleted in the waterbody, potentially causing death of fish and a deterioration in water
quality. Because of their rapid response in both density and diversity to a wide range of nutrient conditions in water bodies, microalgae are great indicators of eutrophication [22,23]. Due to their fast proliferation under favoured conditions, microbes and/or microbial communities can also be used as environmental indicators in other cases, such as for assessing soil quality or for detecting various pollutions such as sewage spills, oil spills or even uranium [21,24–26].
2. Microbial Community Profiling Methods

To objectively evaluate the environmental status of a habitat, it is eminent to obtain integral information about the microbial community inhabiting it, including abundance, composition, diversity, and functions. The approaches to obtain that kind of information can broadly be divided into phenotypic- (17th century until present day) and molecular-based (20th century until present day) ones; the most important of which I will briefly describe in the following sections.

2.1 Phenotypic methods

Phenotypic-based approaches dominated the field of microbiology for most of its time, ever since Antonie Van Leeuwenhoek first observed microbes through his handmade microscope in the 17th century. While subsequent efforts from generations of scientists enabled the identification and characterisation of numerous microbes, they also lead to the realisation that the majority of microbes withstood cultivation and, hence, further identification and characterisation (i.e., the great plate count anomaly; [27–32]). This realisation is at the basis of modern microbiology (i.e., molecular microbiology) in terms of stimulating the development of a series of seminal methodological breakthroughs.

In any case, the conventional phenotypic-based approach to study the microbial diversity in an environment relied on culturing microbes by using a variety of culture media to maximize the recovery of microbial organisms and then continued with phenotypic assays for microbial profiling. For example, the more recent community-level physiological profiling (CLPP) technique yields information with respect to mixed microbial community function and adaptations over space and time [33]. Here, different communities can be compared and classified based on C utilisation patterns using BIOLOG microplates. This technique is highly reproducible and needs little hand-on work. However, such phenotypic-based approaches are limited by culturing and only allow for the detection of fast-growing and non-fastidious microbes [34,35].

A number of phenotypic-based approaches do, though, exist that circumvent the problem of culturing [27–31]. Such are direct counting via
the microscope or flow cytometry or biochemical assays (e.g., fatty acid methyl ester analysis; FAME) [26,35–37]. For instance, plankton is usually monitored by taxonomists counting under the microscope. This is, however, time-consuming and heavily biased towards the observer (and, hence, lacks reproducibility). Flow cytometry is rapid and robust in counting cells, yet, it has rather low resolution in taxonomically classifying microbes. All in all, both direct counting and biochemical assays provide information of low resolution when it comes to what organisms make up microbial communities.

2.2 Molecular-based approaches

Among all the cellular molecules tested to date, nucleic acids have been proven to be the most powerful molecular markers with regard to providing taxonomic information about microbes. It all began in 1965, when Emile Zuckerkandl and Linus Pauling realised that macromolecules (i.e., deoxyribonucleic acid; DNA) carrying the information of genes were ideal for the unraveling of evolutionary history [38]. From then on out, DNA was confirmed to carry the hereditary information of organisms, allowing the inspection of the tree of life. A decade later, in 1977, two independent findings contributed significantly to the progress of molecular-based approaches. First, Frederick Sanger developed a technique allowing the base sequence of DNA to be determined (Sanger sequencing), which revolutionized the way to study and classify microbes [39]. Second, Carl Richard Woese and George Edward Fox developed a framework based on the small subunit of ribosomal ribonucleic acid (16S rRNA) gene within which evolutionary relationships between organisms could be deduced (i.e., they discovered that the tree of life consisted of three domains: Archaea, Bacteria, and Eukarya [40–42]). Somewhat later, also the large subunits of rRNA genes (i.e., 23S, 28S) were shown to uncover phylogenetic relationships among organisms of interest congruent with the phylogenies inferred from the small subunits [43,44]. Another meanwhile widely accepted phylogenetic marker for microorganisms is the internal transcribed spacer (ITS), which targets the spacer DNA fragment situated between the small and large ribosomal subunits and evolves faster than the rRNA gene. The ITS marker is applied to some specific taxonomic groups, such as fungi [45–49]. These ribosomal RNA genes are, in particular, ideal molecular markers for
studying the structure of microbial communities because they are 1) present in all organisms, 2) rarely horizontally transferred, and 3) include both highly conserved and hypervariable regions that facilitate the comparison of sequences over a large phylogenetic range. In general, slowly evolving genes (e.g., ribosomal RNA genes) are more suitable for the study of broad taxonomic groups, while fast-evolving genomic regions (e.g. non-coding regions such as ITS) are preferred in studies focusing on specific taxa within a community.

Molecular-based approaches allow for studying the microbial community without the need of cultivation. Molecular-based, genotypic methods developed so far are based on three types of approaches: DNA fingerprinting, DNA hybridization, and DNA sequencing.

2.2.1 DNA fingerprinting-based approaches

Microbial community fingerprinting is the analysis of a microbial community's structure and dynamics via its genetic fingerprint. Thus, the overall structure of a microbial community is inspected by means of genetic variation within a gene (or a particular region of such) rather than identifying individual microbial species within the community. Variation in the genetic profile is reflected by specific banding patterns of products amplified via the polymerase chain reaction (PCR). The targeted genes are usually either phylogenetic markers (e.g., the 16S rRNA gene) or genes of interest that are responsible for specific metabolic processes. Differences in fingerprints between communities are then reflecting the dissimilarity in community composition between the compared communities. This approach is relatively quick and cheap and able to accommodate a large number of samples. Analysis of the results (i.e., banding patterns among samples) can provide detailed information about community structure in terms of richness, evenness, and composition and can be used to compare samples within the same study [50]. However, these community fingerprinting techniques are semi-quantitative at most, providing an overview of only the most dominant members within a community (i.e., low phylogenetic resolution), and PCR biases may influence their results. Moreover, results from different studies are relatively difficult to compare, especially if different techniques have been used. Representative methods of this type of approach are Terminal
Restriction Fragment Length Polymorphism (T-RFLP), Denaturing Gradient Gel Electrophoresis (DGGE) or Temperature Gradient Gel Electrophoresis (TGGE), Ribosomal Intergenic Spacer Analysis (ARISA) and Single-strand Conformation Polymorphism (SSCP) [51–57].

2.2.2 DNA hybridization-based approaches

DNA hybridization-based approaches, such as microarrays, offer a convenient way to study microbial communities from complex samples in a comprehensive manner. DNA microarray scrutinizes the genetic (or metagenomic) content from sequence libraries based on fluorescence locations and intensities on a chip attached with known oligonucleotide probes [58]. This approach is fairly cheap as long as the microarray is designed and printed. It, moreover, allows for comparing results from different studies (if the same type of microarray is used, that is). However, printed microarrays do not allow for the identification of novel taxa [58], and designing customised arrays requires prior knowledge about the genes of interest [58,59]. Another issue is that of non-specific hybridisation. So far, there are some standard microarrays designed, such as PhyloChip and GeoChip (with thousands of 16S rRNA genes and functional genes printed, respectively) [60,61].

2.2.3 DNA sequencing-based approaches

Profiling of microbial communities via sequencing includes two main approaches: targeted sequencing based on the use of specific phylogenetic marker genes (metabarcoding) and shotgun sequencing of the genomic DNA (gDNA) extracted from the whole community (metagenomics).

Pace and his colleagues were the first to inspect microbial communities (here in the tubeworm) adopting a phylogenetic sequencing-based approach that targeted the 5S rRNA subunit [62]. In the following years, the same research group published a series of studies on microbial communities in different natural environments by sequencing rRNA genes [63–67]. This targeted sequencing approach was then revived with the advent of high-throughput sequencing, which allowed studying microbial communities to a greater depth and greater number simultaneously [68–70]. The challenges that came along with the data generated by high-throughput sequencing are a demand of increased
computational resources and computing skills (i.e., bioinformatics and statistics) in order to fully understand the massive data. Consequently, sequence databases have grown immensely, especially rRNA sequencing databases, yet facilitating the redesign and development of primers that can not only target a greater proportion of microbes but also cope with existing sequencing platforms, which is of great importance (see Chapter 4 and PAPER-I for an overview of primer design and selection.)

If we say metabarcoding can tell us ‘who they are’, then metagenomics can tell us ‘what they can do’; shotgun metagenomic sequencing captures the essential information from functional genes (i.e., the functional potential of organisms) [71–73]. Sequencing of RNA or identification of proteins (metatranscriptomics or metaproteomics, respectively) from environmental samples goes beyond the potential in that it provides insight into the actual activity of genes related to, for example, metabolic activities (i.e., functional traits) [74–76]. Conducting sequencing/identification of bulk genomic/proteomic material extracted directly from the environment allows obtaining knowledge about the composition and function of the investigated microbial community. To gain deeper understanding at the strain or species level, however, other sequencing approaches can be conducted, such as single-cell sequencing and whole-genome isolate sequencing [77–79]. Some newly developed techniques integrate the ideas from different approaches and offer an efficient way to study microbes. For instance, epicPCR (Emulsion, Paired Isolation and Concatenation PCR) conducts the phylogenetic and functional analysis simultaneously at the single-cell level by using a modified PCR technique to concatenate two genes (e.g., a phylogenetic marker and a functional gene) of interest from the same cells into the same amplicons [80].

Although many sequencing-based approaches can offer phylogenetic and functional information at the same time, it does not mean metabarcoding has become superfluous. Compared to other sequencing-based approaches, both the cost of library preparation and of sequencing is lower for metabarcoding, since relatively shallow sequencing is sufficient to obtain taxonomic information from rRNA gene amplicons. And for communities for which reference genomes exist for many community members, the microbial community composition even permits predicting functional traits present in the community by using software programmes.
such as PICRUSt and Tax4Fun [81,82]. The convenient operation and rich information make it a popular approach with ever declining sequencing costs. In this thesis, papers I-IV have all adopted the metabarcoding approach. In papers III-IV we also employed metagenomics and in paper III metatranscriptomics.
3. Sequencing Technologies Suitable for Metagenomics

As previously mentioned, the discovery of phylogenetic marker genes and the invention of Sanger sequencing had great impacts on the study of microbial communities, while high-throughput sequencing techniques revolutionized microbial studies. Although Sanger sequencing has dramatically increased its performance in read yield (up to 384 sequences per run with an average sequence length of 700 bp) and boosts with a sequencing accuracy of 99.999% [83], it cannot compete with high-throughput sequencing technologies in terms of throughput. The first high-throughput sequencing technology, 454 pyrosequencing, was released in 2005 and was popular for analysing microbial community composition due to its lower costs and significant data yield compared to Sanger sequencing [69,84]. Still, due to its higher cost per base than the later developed high-throughput sequencing technologies, it has eventually been outcompeted. Table 1 provides an overview of the currently available sequencing platforms that are suitable for microbial community analysis.

Choosing the right sequencing platform to be used in the study of microbial communities depends on the aim and budget of the study, the type and number of communities to be sampled, which in turn define read length, sequencing depth and coverage. For instance, in papers II-IV, primers 341F-805R were applied to target a 464 base-pair (bp) long region of the 16S rRNA gene in bacterial communities. We aimed at ~100 samples and a sequencing depth of 20,000 reads per sample. We opted for Illumina MiSeq (v3, 600-cycle) combined with the dual-barcoding strategy to sequence 16S amplicons. In PAPER-III, metatranscriptomics and metagenomics were employed to gain information from total RNA and total DNA of a sample taken from Baltic Sea sediments. Here, Illumina HiSeq 2000 (the platform giving highest sequencing depth per cost [85]; Table 1) was chosen in order to gain enough sequencing depth for the possibly highly diverse microbial community. A main objective of PAPER-IV was to evaluate an instrument that could be used in the field, which is why the portable Oxford Nanopore Technology (ONT) MinION was chosen to conduct metagenome sequencing on stormwater samples. Finally, another single-molecule sequencing technology, Pacific
Biosciences (PacBio), that can generate long reads and diminish its sequencing errors with the strategy of circular consensus sequencing, has been successfully applied for reconstructing metagenomes from complex microbial communities [86,87] and also shown its advantages in phylogenetic analysis through yielding full-length rRNA gene sequences [88].
Table 1. Comparison of sequencing technologies suitable for microbial community analysis

<table>
<thead>
<tr>
<th>Sequencer Type/Modea (Kit)b</th>
<th>Illumina HiSeq 2500 (HT v4)</th>
<th>Illumina MiSeq (v3, 600-cycle)</th>
<th>Ion Torrent PGM 318 (400 kit)</th>
<th>PacBio Sequel (SMRT)</th>
<th>Oxford Nanopore MinION (R9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum read length (bp)</td>
<td>125 x 2</td>
<td>300 x 2</td>
<td>400</td>
<td>60,000</td>
<td>200,000</td>
</tr>
<tr>
<td>Maximum insert size (bp)</td>
<td>500</td>
<td>500</td>
<td>247-429</td>
<td>~20,000</td>
<td>~200,000</td>
</tr>
<tr>
<td>Maximum Output (Gb)</td>
<td>1,000</td>
<td>15</td>
<td>4</td>
<td>5-10</td>
<td>2</td>
</tr>
<tr>
<td>Cost($)/Gb</td>
<td>29</td>
<td>55</td>
<td>125</td>
<td>70</td>
<td>500</td>
</tr>
<tr>
<td>Cost($)/device</td>
<td>1,000,000</td>
<td>125,000</td>
<td>80,490</td>
<td>350,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Run time</td>
<td>6d</td>
<td>4h-55h</td>
<td>4h-7h</td>
<td>0.5h-10h</td>
<td>0.1h-48h</td>
</tr>
<tr>
<td>Error rate</td>
<td>0.1-10%</td>
<td>0.1-10%</td>
<td>~1%</td>
<td>~14%</td>
<td>~10%</td>
</tr>
<tr>
<td>Error type</td>
<td>Substitution</td>
<td>Substitution</td>
<td>Indel</td>
<td>Indel</td>
<td>Indel</td>
</tr>
<tr>
<td>Scale of samples / run</td>
<td>10^2-10^4</td>
<td>10^2</td>
<td>10^2</td>
<td>10^2</td>
<td>10^2</td>
</tr>
<tr>
<td>Scale of reads / runc</td>
<td>10^4-10^6</td>
<td>10^4-10^5</td>
<td>10^3</td>
<td>10^3</td>
<td>10^3-10^6</td>
</tr>
<tr>
<td>Portability</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Real-time</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Template amplification strategies</td>
<td>Solid-phase bridge amplification</td>
<td>Solid-phase bridge amplification</td>
<td>Emulsion PCR</td>
<td>Single-molecule long-read sequencing</td>
<td>Single-molecule long-read sequencing</td>
</tr>
</tbody>
</table>

a Besides Illumina MiSeq and ONT MinION, the rest sequencing technologies have multiple types or modes.

b All sequencing technologies offer multiple kits. Here the kit from each sequencing platform that offers the longest maximum read length and biggest data yield were selected for comparison (till 2017 - March).

c The scale of reads per run depends on the scale of samples and/or the read lengths.
4. Selecting and designing primers

PCR primers are critical to PCR amplification-based approaches studying microbial communities by targeting only certain genes or genomic regions (e.g., metabarcoding, epicPCR, or PhyloChip). So far, a large number of primer pairs have been published, with some extensively investigated genes (e.g., 16S rRNA gene) having multiple choices. An ideal primer pair covers (i.e., matches to the genes of) a large fraction of the taxa to be studied and amplifies a DNA region with much phylogenetic information. Moreover, the length of the amplified region needs to be compatible with the sequencing platform. In respect to studying host-associated microbial communities, the specificity of the primer binding region compared to that in the host’s genome is an additionally critical issue. When selecting primers, background information on the targeted lineages is of invaluable help. For instance, when investigating the microbial community from an extreme environment (e.g., deep sea sediments as investigated in PAPER-II) in which archaea might be abundant and active microbial members [89,90], the objective is to choose broad enough (i.e., universal) primers that select for both bacterial as well as archaeal lineages [91]. To rapidly assess and scrutinize the phylogenetic coverage and the specificity of the candidate primer pairs, online tools, such as RDP/Probe Match, BLAST/Primer-BLAST and Silva/Test Probe, are commonly used.

During the early stage of applying high-throughput sequencing technologies to study microbial communities, eukaryotic marker genes and their universal primers were not as well described and systematically designed as the 16S rRNA gene of bacterial and archaeal lineages [92–94]. Until the 18S universal primers published in PAPER-I, only a few 18S primers had been released, and their taxonomic coverage had not been systematically assessed [95–97]. In consequence thereof, there has been a great demand for 18S universal primers that enable unbiased and comprehensive studies of eukaryotic communities.

In PAPER-I, we therefore systematically designed universal 18S primers that can account for the richness present in current databases and result in amplicons suitable for being sequenced with the latest high-throughput sequencing technologies. For that matter, $31,862^b$ non-redundant full-length reference sequences of the 18S rRNA gene were extracted from the
SILVA database, aligned, and subsequently used in DegePrime to design degenerate primer candidates [98,99]. 18-base long oligomers with a maximum degeneracy of 12 were sought when running the above-mentioned program for each position along the alignment. After selecting a number of high-coverage oligomers that targeted conservative regions of the 18S rRNA gene (Figure 1 of PAPER-I), the phylogenetic specificity of the simulated sequencing reads were assessed, when simulating (single- or paired-end) reads using read lengths of various sequencing platforms. Based on these assessments, three pairs of primers were selected for experimental validation. The, primer pair 574_F* - 1132_R worked best and was successfully applied in metabarcoding (Illumina MiSeq) analysis of different environmental and host-associated samples, yielding results that could be expected from the in silico evaluations. This validated primer pair was subsequently applied to investigate the diversity among eukaryotic microbes along the salinity gradient of the Baltic Sea (PAPER-II).

\[a\] In 2014, the available high-throughput sequencing technologies suitable for metabarcoding were 454 pyrosequencing, Illumina, and Ion Torrent.

\[b\] In 2014, the size of SILVA v.111 of non-redundant 18S reference sequences was 31,862. The current version of the same category in SILVA v.128 contains 421,053 sequences.
5. Sample preparation

Sample preparation is a fundamental step in sequencing-based microbial environmental studies, which involves extraction of the genetic material and preparation of the sequencing library. How different types of environmental samples are processed, what the common strategies for sequencing library preparation are, and which steps are susceptible to contamination and/or biases, are the subjects of Chapter 5.

5.1 DNA extraction

Most DNA extraction procedures for environmental samples consist of cell harvesting, cell lysis, removal of non-DNA molecules, and finally DNA collection (note that only cell harvesting and cell lysis are described here). The extracted genetic material (microbial DNA in this thesis) should be representative of all microbial cells present in the samples (unbiased) and sufficient in quantity and quality for the subsequent downstream process of library preparation.

To harvest microbial cells from liquid (e.g., seawater, stormwater) or gas (e.g., air) samples, filtration with size selection is the most common strategy [69,100,101]. Centrifuging can also be an alternative or supplement approach to enable size selection or fractionation when processing liquid samples [102,103]. Besides, flow cytometry integrates the functions of cell counting and sorting, which can also be used for enriching the targeted microbes [104–107]. However, the procedure of DNA extraction from solid samples (e.g., soil, sediment, faeces), in particular regarding whether or not to separate cells from the solid matrix prior to cell lysis, is currently debated, since it has been observed that these processing procedures may result in significantly altered community composition [108,109].

The most common method for cell lysis is to incubate samples in detergents/enzymes followed by mechanical disruption by bead beating. But as a relatively sensitive procedure, both mechanisms and intensities of the cell lysis procedure may result in a discrepancy with respect to the microbial community composition [110]. While there are many commercial kits or robust protocols for extracting DNA available to choose among [110,111], no protocol produces completely unbiased
results or performs equally well on all types of samples [112]. Being consistent in the use of parameters during sample processing enables the comparison of results across samples as well as projects [113].

During the extraction procedure, the contamination issue should sound a note of caution, since the whole downstream process is based on the extracted genetic material, and it is usually the only source material kept for the culture-independent environmental studies. DNA contaminants are present in both commercial extraction kits and laboratory-made reagents, and may greatly affect the subsequent results, especially in cases, where only low microbial biomass samples exist or/and a targeted sequencing-based approach is adopted [114]. Introducing negative controls (such as DNA-free water) along the whole experimental procedure (from sample processing to sequencing) can help keep track of the quality of operation and enable identification of sequences derived from contamination [114].

In environmental surveys, absolute counts of microbes are in some cases prefered, for instance when assessing water quality (see PAPER-IV). Unlike culturing or cell counting approaches, sequencing only offers relative (i.e., not absolute) community composition data, but by adding a DNA standard before DNA extraction, absolute quantification can be achieved [115]. Such a standard can also be used to estimate sequencing error rates [115].

### 5.2 Preparation of sequencing libraries

Sequencing library preparation is a procedure during which the extracted genetic material is converted into a form compatible with the respective sequencing platform. The basic techniques employed in DNA library preparation include fragmentation, amplification, ligation, size selection, and purification. In the following text, each technique is briefly described. Moreover, protocols followed in papers I-IV exemplify how those basic techniques were arranged in preparing certain types of libraries.

The size of DNA fragments to be sequenced is a key parameter of all sequencing platforms (see Table 1). Commercial extraction kits (such as MO BIO PowerWater or Qiagen QIAprep) usually allow extraction of
DNA fragments of up to 50 kb (or even more). Therefore, fragmentation and targeted amplification are usually the first steps to control for DNA fragments in the desired size range (Figures 1-3). Fragmentation comprises three types of approaches: physical (e.g., sonication, point-sink shearing; [116,117]), chemical (e.g., divalent metal cation; [118]), or enzymatic (e.g., transposase, endonuclease). The most widely applied methods are sonication and point-sink shearing with Covaris\textsuperscript{®} tubes, aiming at short (100 bp - 5 kb) or long (6 - 20 kb) gDNA fragments (see papers II, IV; Figures 2,3). Fragmentation using the transposase enzyme, which is used by the Illumina Nextera Tagmentation and ONT MinION Rapid Sequencing Kit (Figure 3), allows adding sequencing adapters to the fragments’ ends simultaneously. This method is rapid and permits low quantities of input material but may, however, introduce small biases [119].

Ligation during sequencing library preparation is usually the step in which the adapter sequences are added that facilitate the upcoming PCR enrichment and/or the hybridization to the probes on the sequencing flowcell. Some adapter sequences contain barcoding sequences, which allow multiple samples to be sequenced within the same run without physical separation. As mentioned earlier, PCR amplification can be used for both targeting and enrichment. To achieve enrichment, PCR amplification uses primers that target the ligated sequence region from both ends of the DNA fragments. It is a simple and robust technique, however, due to its tendency to introduce biases of substitutions and chimeric constructs, it is highly recommended to minimize PCR amplification cycles during library preparation. In addition, when sufficient amount of sample DNA is available (usually 1 ~ 2 µg per sample), amplification-free protocols are preferred for shotgun genomics (e.g. Illumina/TruSeq DNA PCR-free library preparation kit).

\textsuperscript{a}Chimeric constructs (also called chimeras) are artificial sequences that can be formed in the PCR by combining two or more biological sequences. This is rather common in metabarcoding sequencing as similar sequences are co-amplified in the PCR.
Size selection and purification can be conducted by using gel electrophoresis, silica binding, or binding of magnetic beads together with a precipitation buffer [120–123]. This procedure enables changing reaction buffers and enriching for DNA fragments of an appropriate range for subsequent procedures. During the procedure of library preparation, purification and size selection are required for most of the steps (Figures 1-3). For instance, DNA extracted from the environment may carry components that can affect enzymatic reactions (e.g., humic acids can inhibit PCR amplification [124]). Besides, small fragments (such as primer residues) may also result in PCR artifacts (e.g., primer dimers), which may severely decrease the success of the sequencing by yielding a large number of short, useless reads.

The general principles of constructing sequencing libraries are to use a minimal amount of amplification, to control quality stringently (contamination and size control), and to estimate adequate sequencing depth when multiplexing samples. Wisely selecting and arranging proper techniques can help generating high-quality sequencing libraries of low cost in terms of time and money.

Figure 1. Metabarcoding protocol for generating dual-indexed Illumina-ready amplicons. Purification to remove short fragments is required after each PCR step. The protocol was employed in papers I-IV.
**Figure 2.** Metagenomic protocol for generating dual-indexed Illumina-ready libraries. Purification to remove short fragments is required after both adapter ligation and PCR amplification steps. The protocol was adopted for PAPER-III.
Figure 3. Metagenomic protocol for generating ONT MinION-ready libraries. Purification to remove short fragments is required after each PCR, end repairing, and adapter ligation steps. Left: PCR barcoding 2D-read protocol (approx. 7-hours; PAPER-IV); Right: rapid barcoding 1D-read protocol (approx. 10-mins; discussed in PAPER-IV).
6. Processing of sequencing data

Each sequencing platform has its intrinsic mechanism for detecting nucleotides; the raw data usually consists of optical or electrical signals. Through a procedure called “base-calling” these signals are converted into nucleobases (i.e., A, T, C, G).

The typical data processing workflows for metabarcoding and metagenomics projects are demonstrated in Figure 4. Since all projects that make up this thesis have adopted the metabarcoding approach, the key steps of metabarcoding data processing are explained thoroughly in this section. The relevant steps in processing metagenomics data are, on the other hand, only described very briefly.

![Figure 4. Flow diagram of data processing key steps of (A) metabarcoding and (B) metagenomics projects.](image)

6.1 Quality control and reads pre-processing

Each sequencing platform produces errors with unique characteristics (Table 1). Controlling for quality (QC) of raw reads is the primary step in data processing. Doing otherwise may have severe effects on the outcome of the analyses.

Numerous programs have been developed for conducting QC. Some of them provide QC metrics (i.e., read counts, lengths, GC content, etc.) of the sequencing reads (see FastQC or MultiQC). Others, such as Sickle,
Cutadapt, Mothur, QIIME, USEARCH, DADA2, and Poretools [125–129] also offer functions or even whole pipelines to modify reads (quality filtering/trimming, adapter removal, etc.)[125–129]. To diminish the effect of sequencing errors and artifacts, low-quality sequence regions, primers, and sequencing adapters are to be trimmed before carrying out downstream procedures.

In metabarcoding projects, if paired-end sequencing was employed and the paired reads contain overlapping regions (as was the case for papers II-IV), merging paired-end reads after quality trimming can further reduce sequencing errors. The resulting, longer sequences also allow for an improved downstream annotation. Alternatively, if paired-end reads show gaps, using a consensus annotation of both reads can enhance classification accuracy (e.g., PAPER-II). In shotgun sequencing, insert sizes of paired-end reads are typically too large to facilitate merging, however, the longer insert-size makes the downstream assembly procedure more efficient.

### 6.2 Clustering and denoising

Environmental samples harbour a diverse range of microbes, and consequently sequencing results in a diverse number of sequences. To handle the enormous number of various sequences yielded from high-throughput sequencing platforms, clustering or denoising are the favorable options. The inferred cluster representatives or denoised sequences together with their quantities can later on be used to estimate microbial community composition, diversity, and richness.

Clustering is a procedure that lumps similar reads together into a cluster. Clustering has the benefit of diminishing amplification and sequencing errors and easing the computational burden for downstream analyses. The yielded clusters are termed operational taxonomic units (OTUs) and can be understood as individual species groups, species, or species strains. Methods developed for clustering sequences can be categorized into either hierarchical or heuristic [130]. Hierarchical clustering methods (employed in, e.g., Mothur [125]) calculate the differences between each pair of sequences prior to the actual clustering. This is computationally demanding, thus rendering hierarchical clustering methods unsuitable when dealing with large-scale sequencing data sets.
The heuristic clustering methods (e.g., UPARSE [131], UCLUST [126]), on the other hand, bypass the calculation of distances between sequences in that they begin with one sequence cluster and successively map each of the remaining sequences to that cluster based on their similarity to the cluster’s centroid sequence. The sequences that fail to be mapped form new clusters that subsequent sequences will also be mapped to. This approach is less computationally demanding but also less accurate compared to hierarchical clustering, since the order of the input sequences significantly affects the clustering result. Among the heuristic clustering methods, UPARSE performs best in avoiding false-positive OTUs, and this algorithm carries out clustering and chimera-removal simultaneously. Usually, a similarity cutoff of 97% is applied, which is comparable to intra-species sequence similarities of prokaryotic species [132]. It has to be noted though that different species may have 99% similar or even identical 16S sequences [133,134], while some strains may have multiple copies of 16S rRNA gene with some regions showing 5% dissimilarity [135,136]. In papers II and III, UPARSE was employed to generate OTUs from the paired-end merged reads, and a similarity cutoff of 99% was chosen to offer a finer resolution of the inspected Baltic Sea planktonic communities.

Yet, clustering sequences based upon an arbitrary threshold may result in considerable information to be missed with regard to real biological variants that share highly similar but not identical sequences. These may occupy different ecological niches [137] such as be derived from pathogenic or nonpathogenic strains [138,139]. To avoid clustering such true sequences, programs, such as Minimum Entropy Decomposition (MED) [140] or Divisive Amplicon Denoising Algorithm 2 (DADA2) [128], exist to reconstruct the real diversity within a community while reducing sequencing errors. Both MED and DADA2 initially divide sequences into partitions and presume some lower abundance variants to be artificial, and then, MED employs single-site minor-allele frequencies as thresholds to characterize the real variants [140], while DADA2 builds a model of sequencing errors to infer the real variants [128]. Comparing the two, MED resulted in a higher number of false sequences and failing to detect more rare true biological variants than DADA2 [128].

A recent study showed that the overall microbial community composition was highly similar when applying different clustering approaches on the
same set of amplicon reads [141], a conclusion also drawn by us when we compared the results yielded from clustering (UPARSE) and denoising (DADA2) on the PAPER-IV data (in PAPER-IV, only the DADA2 results are shown). However, when we used metabarcoding to study 13 closely related lactic acid bacterial strains, only DADA2 could truly detect the closely related biological sequences [142]. Therefore, when closely related microbial organisms are to be studied, only denoising algorithms perform adequately.

Data stemming from metagenomics sequencing, on the other hand, are usually directly assembled (after passing an initial quality-control step), as many assembly programs have intrinsic denoising implemented in the actual assembly procedure.

6.3 Annotation

While OTUs, either derived from clustering or denoising, provide the basis for subsequent microbial community composition and diversity analyses, annotation offers putting these OTUs into an ecological context, ‘who are they’ (taxonomic annotation or taxonomic classification). And annotation on the functional genes from metagenomic data answers ‘what do they do’ (functional annotation). Annotation allows carrying out ecological analyses and relating the processed samples from different projects to each other. The annotation methods for metabarcoding and metagenomics can be categorized into three general types that are based on alignments/mappings, k-mers (short nucleotide sequences with the length of k bases), and phylogenetic trees. In addition, metagenomic sequences can also be annotated through relating them to signature genes [143]. In the following, only the alignment- and k-mers-based annotation methods are described.

The most well-known alignment tool is BLAST [144], which can conduct a search from various available databases and with great sensitivity. Other programs developed later, such as BLAT [145], USEARCH [126] or DIAMOND [146], have been improved greatly with regard to search speed, albeit with slightly lower sensitivity. In PAPER-II, in order to investigate which eukaryotic plankton inhabit the Baltic Sea, BLAST was employed to annotate the 18S and 16S OTUs (18S rRNA and chloroplast DNA), using two recently published protist databases as references (PR2
[147] and PhyloREF [148]). Here, a read-pair consensus taxonomy strategy and taxonomic-level specific similarity cutoffs were applied to balance precision and sensitivity, strengthening the annotation process. In contrast, as the goal for PAPER-IV was to rapidly assess the faecal contamination level of water samples, shotgun reads yielded from a real-time sequencer were directly aligned to a human gut microbial database with a fast aligner (USEARCH).

Since the alignment-based methods are computationally demanding (and, thus, time-consuming) a good alternative is to use approaches based on k-mers. Such approaches conduct similarity searches by checking for the occurrences of exact k-mers. Two popular rRNA databases (RDP [149] and SILVA [150]) have developed classifiers (the RDP naive Bayesian classifier [151] and SINA [152] have adopted this strategy with different post processing methods, respectively): RDP employs a fast testing strategy, bootstrapping, to offer confidence values of taxonomic assignments, while SINA uses a partial order alignment strategy to allow for a high accuracy of annotation [153].

Although precise annotation for as many reads as possible is desired, it is usually hard to achieve, especially for lower taxonomic levels (i.e., genus or species levels). Nevertheless, some research questions can be answered by using annotation at a coarse taxonomic level (e.g., phylum or class). Besides, depending on the research aim and data sources, data processing pipelines may vary, each introducing different biases that should be taken into consideration when results are analysed and interpreted.
7. Results and Discussion

7.1 Systematic design of 18S rRNA gene primers (PAPER-I)

Microorganisms play important roles in most ecosystems, from oceans to forests to the intestines of animals. Sequencing genetic material extracted directly from the environment by means of a metabarcoding approach allows us to effectively inspect microbial community structure and easily conduct large-scale comparative studies. However, the study of eukaryotic microorganisms has received far less attention compared to prokaryotes, which is partly due to technical difficulties (e.g., complex life cycles, many heterotrophs are hard to isolate, larger genomes than prokaryotes, etc.). With the fast development of sequencing techniques and growing number of 18S rRNA gene sequences available in databases, we aimed at designing a set of 18S primers with broad taxonomic coverage and compatible with current sequencing platforms.

In this project, we downloaded all non-redundant full-length 18S rRNA gene sequences from the SILVA database (version 111) [150]. A primer design program DegePrime was then employed to generate candidate primers at each position of the reference sequences multiple-alignment [154]. The requirement for the generated primers was 18 bases long with maximum 12 degrees of degeneracy. Candidate primers located in the conserved regions of the 18S rRNA gene, while being flanked by variable regions, were selected for further in silico and in vivo tests.

To evaluate the candidate primer pairs for their phylogenetic resolution, we simulated sequencing reads based on the primer pairs targeted sequences from the reference sequences, and the sequencing characters of various sequencing platforms were used as parameters in the reads simulation (Table 1). Here, we simulated paired-end reads with the length of 150 bp or 250 bp to represent the reads yielded by the Illumina platform, and single-end reads with the length of 400 bp to represent the reads yielded by 454 or IonTorrent sequencing platforms.

In order to assess which amplicon type (decided by both primer pair and sequencing read lengths) is more informative in recovering 18S rRNA
gene, we calculated the ratio between the number of unique amplicon sequences and the unique full-length 18S sequences containing the same primer pair sequences. This comparison showed that the V4 and V5 regions are the most informative of the 18S rRNA gene, and the choice of sequencing region is more important than the choice of sequencing techniques (i.e., read lengths), since amplicons yielded by the same primer pair had similar performance when using different sequencing techniques.

In order to assess the taxonomic assignment accuracy, we conducted an *in silico* annotation on the simulated reads. Here, the reference sequences were divided into two portions, one portion was used for amplicon simulation and the other served as the annotation database. The specificity of annotation was assessed by comparing the original taxonomies of the simulated amplicons to that of their best hits. To estimate the phylogenetic assignment accuracy, the genetic distances between the full-length 18S sequences of the pairs of query-hit were calculated. The best-performing primer pairs were further validated in the laboratory. Eventually, the primer pair, 574*-1132, that targets the V4-V6 regions of the 18S rRNA gene, was successfully applied on samples from a range of environments. The amplicons were sequenced on the Illumina MiSeq platform. The yielded eukaryotic community structures were meaningful and informative on the tested environmental samples.

In this study, we demonstrated how to conduct broad-taxonomic range primer design, from preprocessing reference sequences to experimental validation. Most importantly, the presented primer pair facilitates enhanced metabarcoding studies of the eukaryal domain of life.

### 7.2 Microbial plankton biogeography in the Baltic Sea (PAPER-II)

The Baltic Sea is one of the world’s largest brackish water bodies. Its horizontal salinity gradient (around 2000 km long) ranges from 2 psu in the northeast to 30 psu in the southwest. This study is the first sequencing-based spatial survey of eukaryotic and bacterial plankton along that salinity gradient. In total, 21 samples from the Gulf of Bothnia, the Baltic Proper, and Kattegat were collected from the surface water
during one week in July, 2013. Both 16S and 18S rRNA genes were amplified from sample DNA and sequenced on the Illumina MiSeq platform to detect patterns in community composition of pro- and eukaryotic plankton along the transect.

We found that 1) both 16S and 18S community composition significantly correlated with salinity, emphasising the effect of salinity on the biodiversity within this ecosystem, 2) the alpha-diversity (species richness) of both assemblages showed no clear trend along the salinity gradient, and 3) the distribution of the dominant plankton species tracked by monitoring programs (e.g., blooming taxa, metazoa) followed the expected pattern in this survey. Additionally, species novel to the Baltic Sea were detected, such as the coccolithophore *Emiliana huxleyi*. Moreover, this is the first time eukaryotic plankton diversity is shown to neither diminish nor increase at intermediate salinity. These results challenge both Remane's Artenminimum concept (a minimum of diversity at intermediate salinity levels) and the proposed protist species richness peaking at the horohalinicum region (salinity 5-8 psu) [155]. Our study significantly expands our understanding of the biodiversity of the Baltic Sea and the results serve as a good reference for other studies focusing on brackish plankton. Finally, it aids in assessing the monitoring strategies of plankton in the Baltic Sea.

### 7.3 Inspecting protein synthesis potential of the microbial community at Landsort Deep, the Baltic Sea (PAPER-III)

Microbial community analysis is normally based on genomic DNA, which does not distinguish between viable and dead community members. This dilemma is of particular concern when conducting microbial community studies from sites with a high degree of deposited matter, such as in soil or marine sediments. In this project, we inspected the microbial community of the sediment from the deepest point of the Baltic Sea (depth: 460m), the Landsort Deep. To elucidate which microbes may contribute to ecosystem function, we analysed data originating from rDNA amplicons, rDNA reads from a shotgun metagenome, and expressed rRNA from a shotgun metatranscriptome. Since RNA is a much
more unstable molecule than DNA and likely degrades rapidly in the environment after cells have died, RNA is a better indicator of the viable and active members within a community. By calculating the ratio of RNA to DNA of the rRNA gene as an index of microbial protein synthesis potential, we were able to estimate which taxa showed cellular activity and metabolic potential in this extreme environment.

For each major taxonomic group, we performed a test to determine if its median ratio was significantly higher or lower than the microbial community median ratio. We found that not only anaerobic bacteria, such as Deltaproteobacteria and Euryarchaeota, demonstrated high protein synthesis potential in this anoxic, dark habitat. Additionally, some common surface water bacteria, such as Cyanobacteria, displayed surprisingly high protein synthesis potential, indicating that they may play active roles in this ecosystem. Alternatively, they may survive in the deep water and possibly re-enter the photic zone the following season. Our study provides new insight into the complexity of the sediment microbial community at Landsort Deep, with potential for understanding other sediment ecosystems.

7.4 Sequencing-based approaches to track faecal contamination in stormwater systems (PAPER-IV)

Separate sanitary sewer systems consist of wastewater and stormwater systems, in which only the wastewater is processed before being discharged to natural water bodies. Occasionally, misconnections or leaks happen, leading to wastewater entering the stormwater system and being discharged without purification. The traditional method of enumerating faecal indicator bacteria by cultivation for assessing wastewater contamination in water has limited specificity to human faecal contamination and is time consuming. This project aimed at developing fast and accurate sequencing-based methods to detect human faecal contamination in stormwater systems. Here, we compared the culture-based gold standard, the IDEXX Colilert-18 test, with Illumina MiSeq amplicon sequencing and shotgun sequencing on the portable ONT MinION device in an attempt to assess contamination levels of 73 stormwater samples collected from Stockholm City. An additional
wastewater sample collected from Bromma wastewater plant was included as a positive control in the amplicon sequencing analysis.

The overall microbial community composition was highly correlated with the *E. coli* culturing data. Compared to the samples showing low *E. coli* densities, samples with high *E. coli* densities demonstrated similar community composition as the typical wastewater sample. Besides, high correlations were also obtained between the *E. coli* culturing counts and the frequencies of human gut microbiome sequences acquired via amplicon sequencing. Both results indicated that *E. coli* is indeed a good indicator of faecal contamination. From a residential region adjacent to a eutrophic lake (Lake Trekanten), two contamination sources connected to the same stormwater drainage system were distinguished by microbial community analysis based on amplicon sequences, which could not be revealed with the microbial culturing-based method. In addition, the subset of samples that were also shotgun sequenced with ONT MinION showed human gut microbiome reads fractions proportional to the result yielded by Illumina MiSeq amplicon sequencing. Based on our results and the latest published reagent kits, we also compared the three approaches in terms of time and cost, and proposed optimal experimental settings for assessing microbial water quality with the portable sequencing device in the field.

This study demonstrates that sequencing-based methods are not only able to assess faecal contamination levels of water samples but also allows to discriminate the contamination sources based on the microbial community structures. The capability of yielding microbial metagenomic data in real-time allows the portable sequencing device to be used for tracking contamination sources based on microbial community fingerprints or assessing the water quality for urgent cases (e.g., risk assessment of pathogens) in the field.
8. Concluding Remarks and Future Outlook

From observing microbes through a microscope to obtaining massive amounts of genomic information via sequencing, novel methods and technologies offer great opportunities, but also challenges, with regard to studies in the field of environmental microbiology.

Microbial DNA sequencing allows us to inspect the taxonomic structure and metabolic potential of microbial communities without time consuming culturing or direct observation. Modern genomic methods may also result in new taxonomic findings and/or provide new insights with respect to functional genes in less time than classical methods. The high efficiency with which taxonomic information is obtained through DNA sequencing represents a great advantage in microbiological surveys. To enhance the study of micro-eukaryotes through metabarcoding, a set of 18S universal primers that are compatible with current sequencing platforms were presented in PAPER-I. By using this set of 18S primers together with a set of 16S universal primers, the planktonic communities along the Baltic Sea salinity gradient were exhaustively described in PAPER-II. The diversity of the detected plankton from one week’s sampling was on par with the recorded species over the last century, and the distribution patterns of certain planktons inferred with DNA sequencing are congruent with monitoring records. The great power of using DNA sequencing in environmental surveys was also demonstrated in PAPER-IV, where the approaches of metabarcoding, metagenomics and the classical E. coli culturing yielded concordant results in assessing the faecal contamination of stormwater samples. Superior to the culture-based method, DNA sequencing has the capability of tracking actual contamination sources, and the latest developed real-time DNA sequencer, ONT MinION, demonstrates its potential of assessing microbial water quality in the field. Due to its stable structure and low cost in experimental procedures, DNA is preferred over RNA in regular environmental surveys. While to inspect the active members of a community, RNA is a better indicator. In PAPER-III, we combined information from DNA and RNA to elucidate the functional status of microbes in the seafloor sediment. Apart from genomic information, proteomics, metabolomics, and even the conventional microbial culturing have their specific virtues and output in a various aspects of microbiome studies. By integrating proper methods and tools, we are able to answer
questions of great ecological meaning, and also tackle environmental problems to a greater extent.

So far, techniques that can conduct real-time genomic and proteomic sequencing on portable devices have been developed (the current representative is Oxford Nanopore Technologies [156]). With decreasing cost and increasing accuracy, combined with enhanced mobility of techniques, real-time sequencing in the field will most likely be the method of choice in future environmental surveys and studies. However, the lack of standard procedures of sequencing-based approaches reduces the power of combining or comparing different environmental studies. Besides, the massive sequencing data generated by current sequencing techniques also challenges us in the aspects of data storage, processing, and analysis. In order to take advantage of applying sequencing-based approaches in environmental studies, robust and streamlined experimental protocols and data analysis pipelines are desired. The work presented in my studies offers a good reference for establishing optimal DNA sequencing pipelines in environmental surveys, and also for the research aiming at understanding microbial ecology.
Acknowledgements

This acknowledgement is for all the ones who ever given me warmness, happiness and knowledge during my PhD life. With you, living and studying in Sweden is such a wonderful experience.

First of all, I would like to express my sincere gratitude to my main supervisor, Anders Andersson, for his great patience and endless help to me in these four years. Thank you, Anders! I feel so lucky of being your student. You are such an inspiring and witty person, to work with you, I feel pleasant, guided and greatly supported. In these four years, you have created many chances for me in both doing research and exploring the Nordic life. I enjoy all of them a lot! Thank you so much, I would say, both of your diligent working attitude and gentle personality will affect me for the rest of my life.

Next, I would like to thank my co-supervisor, Jürg Brendan Logue. As being both my deskmate and co-supervisor, I never feel you are like a teacher but a true friend! I would like to thank you for your great patience in answering my stupid questions through the last two years. I want to especially thank you for your great effort in polishing my StockholmVatten manuscript and PhD thesis by sacrificing your precious time in writing the grant application. I do love your “Swiss-Irish” personality, which is meticulous, critical, and thoughtful, well, that is maybe Jürg’s personality not because of where do you origin.

The other two people that are crucial for bringing me to Sweden are my co-supervisor, Joakim Lundeberg, and my master supervisor, Xiaoguo Chen. Thank you, Joakim, for giving me the wonderful chance to study in KTH BioTech School. You have praised my improvement through the years, which largely encouraged me to keep on carrying on! Thank you, Chen 老师！Without your significant help and great encouragement, I would not be able to get CSC scholarship and make the decision to study abroad that I would never regret about. In my mind, 一日为师终生为父 is the most proper expression to describe your role in my mind.

Here I would also give my sincere gratitude to my great collaborators. Thank Bengt for bringing Baltic Sea samples to me, and bringing me to the scientific expedition cruise, Aranda, which was a so cool experience, and you are a very supportive collaborator! Thank Sara and Daniel for involving me in the Baltic sediment project, which gave me the first glance to the microbes in the extreme environment. Thank Jens for driving me to take stormwater samples around Stockholm Län and
offering me *E. coli* culturing data. Thank **Patrik** and **Per** for bringing me to such a cool project, the intestinal microbiome of Antarctica birds. Thank **Alejandra** and **Sepideh** to bring lactic acid bacterial strains, which allows me to confirm that DADA2 is trustworthy in distinguishing close biological variants. Thanks **Tülay** and **Anna** to involve me in the Periodontitis project. **Anna**, you are a great friend and also a nice collaborator. You tried to save my time when I was preparing graduation. Now I am graduating, and it is my turn to help you back!

The two people like tutors for me are **Daniel Lundin** and **Mikael Huss**. You two are like bioinformatics wizards to solve my bioinfo problems so efficiently. I wish I could have a better foundation in programming so that I could learn more from you two. However, I have already been very grateful to know you, and to know bioinfo could be such a joyful work. Daniel, thank you for bringing me to the “graveyard” project that gives me the first chance to work with metagenomics and metatranscriptomics. Thank you, Mikael, for giving me suggestions on the “faecal tracking” project and bring me to R workshops in Stockholm. Chatting in Mandarin with you in SciLife makes me feel relaxed during work.

Now, it is the turn of our sweet **EnvGen** group! Thank you all, you make me feel a sense of belonging! I enjoy all group gatherings, and I would like to thank you for all your kindly help to me in both researching and living. **Conny**, thank you for giving me so many fist-to-fist encouragements. You are like a big brother to take care of me in doing research, sports and travelling. I would never forget my swollen arms from skiing training and the wonderful Niagara Fall trip! Being nice to the others and keep on fighting are the best qualities you taught me. **Carlo**, thank you for forwarding me so many postdoc positions in these months. I love to have lunch chats with you, which are always relaxing and inspiring. The amazing sports of climbing and slacklining you showed me bring me novel experiences from the sports emphasising focusing and persistence. **Johannes**, thank you for supporting me during these four years. Although you are younger than me, I feel you are the one taking care of me, in either work or life, with your cute fiancéé **Johanna**. You have witnessed many of my awkward moments and maybe also some improvements, from “Do you know ‘cd’?” in my first metagenomics workshop to read this story at here. I hope I could also witness your wonderful moments, such as your wedding and doctoral defence! **Luisa**, thank you for making me become a better me. Your remarkable working attitude and great confidence largely influenced me. Thank you for guiding me in labs and processing sequencing data when I was a freshman in this field. You are also a great travel companion in both Korea and Canada tours. **Hugo**, thank you for always being gentle and
generous whenever I need your help. You are one of the few friends I made in my first year here, and I feel so lucky to know you and being supported whenever you are around. Nelson, thank you for helping me to conduct the MinION runs in 2015, which required intensive pipetting work and we always worked till the midnight to load the yellow essence yielded from the whole day pipetting. Great that we will soonly submit the manuscript. Let’s holding our thumbs still. John, thank you for encouragement always. I admire your gentle manners and diligent working style. Sophie, you are just like your name, wisdom. Thank you for helping me to accomplish the Baltic transects project. I enjoy those nights we worked in Espresso House or swimming at Eriksdalsbadet. You are so active and smart, and I appreciate to have you as my friend. Engy, I love your smile and gentle encouragements. You are such a sweet and nice girl. Thank you for helping me to rehearse my half-time seminar and gave me such a big gift after that. I still pinned the card from flowers on my desk that you sent me. Thank you so much for giving me unconditional trust. Kajsa, you showed me what is girl's force in both aspects of intelligence and physical strength. Your climbing skills totally amazed me. Olov, the forest princess. Everyone loves to hang out with you because you are so fun and interesting. I hope next winter I could go ski with you again, but of course, I could join you more for the Thursday pea soups, albeit I always order other dishes.

My girls! Emanuela, Stefania, Fulya, Luminita, it is your turn! I am so appreciate and happy to meet you, to know you and become a member of this fantastic group. Every two months gatherings, celebrating birthdays and new years, travelling, clubbing or just simply going out for delicious food, girls, you really make my life outside SciLifeLab (though all of you are/were my SciLife colleagues). Some of the time, when looking at our elongated shadows cast by the streetlights, the soundtrack of Sex and the City, Auld Lang Syne, would just ring in my mind. I could write pages about our stories, but I would leave them only for ourselves and for the future since I know this lovely friendship will continue no matter where we are.

For the friends and colleagues in Alpha-3, I feel very happy to work with you. Max and Sebastian, thanks for your last minute great help in conducting the nanopore sequencing by offering the latest released kits! Anders J, thank you very much for helping me through years. Especially in the last semester of my PhD, you kept on tracking my progress and offered me many of your defence materials as templates to save my time. Aleksandra, since you came, I am no more lonely in front of my desk. Morning Hi, leaving Bye and lunch giggles, make me feel happy and relaxed in every working day. You are like my big sister that always takes care of or teases on me. Sailendra and Niyaz, you guys are always in the
office. It is great to have you as companions for working on weekends or at night. **Erik Sjö**, thank you for bringing me to Swedish folk dance clubs and concerts. I enjoy them a lot! Let’s continue when I finish my PhD. **Yilin, Niklas, Martin, Linnea G**, and **Yitian**, I have a natural close feeling to your group. You guys are very nice and bring me a lot of happiness. Thanks for inviting me to hang out with you, which are very joyful. **Annelie, Fredrik, Maja, Mickan, Kim, Kostas, Linda, Jonas** and **Linnea**, the ST gang, thank you for giving me such a lot of encouragements and laugh. Annelie always hugs or comfort me whenever I feel depressed; Fredrik likes to tease on me but always helps me out; Maja always reminds me to go to the gym after work; Mickan is so energetic and inspiring; Kim cares about everyone’s feeling; Kostas likes to share; Linda likes to propose a lot of questions; Jonas loves to share his ideas; Linnea always looks gorgeous at work. **Erik B, David** and **Mahya**, the experimental genomics gang, and **Mengxiao, Simon**, the MEB buddies, thanks for giving me essential help in labs, and bringing me joy outside labs. Here, I would also express my sincere appreciation to the expertises in the platform. **Kicki, Christian N, Badam, Nemo**, with your kindly help in my first year, I could successfully finish my MiSeq sequencing library. I would also thank **Lina, Anna K, Chuan, Magedalena** and **Mattias**, with your diligent work I could have samples been sequenced efficiently. **Francesco, Jun, Senthil, Phil, Richard, Isak**, and **Denis**, thank you for maintaining data delivery goes smoothly. Many of you usually have lunch or hangout with me in or out SciLife, which bring me a lot of fun. I like you smart guys and girls a lot! **Marcel, Estelle, Maxime** and **Jakub**, you are the expertises from WABI while you are super friendly and always give me a hand whenever I need. Thank you, **Pelin**, for always encourages me!

Outside Alpha-3, there are also a few friends that I would like to express my gratitude, **Robin, Guille, Markus Jani, Johannes Asplund, Johannes Küpper, Roman, Robert, Adi, Valérie**, thanks for inviting me for hiking, house-warming, movies and travelling. After knowing you, life in and outside Sweden becomes so colourful.

Thanks for my Chinese gang in Stockholm (basically from Karolinska Institute or SciLifeLab), **Hong J, Jingwen W, Xiao H, Di W, Xiaolu Z, Xiaoyan Z, Wenjing K, Mei W, Jiayao L, Ying S, Shanshan X, Tianwei X, Yingtong D, Liang Z, Wang Z, Mozhu D, Ruijingfang J, Jingli Y, Jingru W, Mengxi P, Jingyu L, Xiaolin Z, Yi W, Yajing S, Tianqi Z** for taking care of me and encouraging me. Because of you, my homesick has diminished largely. Some of you have been graduated and moved to other countries, while some of you are still fighting at here, I wish all of you the best!
I would like to give my special thank to Pangpang (Fan Z). It has been 12 years since I knew you, every time when I feel down or lonely, you always gave me a hand. I am so glad you are coming to my defence from Denmark! I would also thank my old classmate Gang H, with your generous help with documentary things for me in our Master college in China to allow me focusing on the thesis writing. Martin and John, thank you so much for coming to support me from Germany. You are not only support me but also helped Weishi to face to his flight phobia. I am so touched!

And finally, thanks my beloved parents, Shujie Hu and Lamei Cao, for their endless love and unconditional support. I feel so lucky to be your daughter! 谢谢爸爸妈妈给我无尽的爱与无条件的支持，作为你们的女儿我感到非常的幸运！I would like to give my special appreciation to my husband, Weishi Wang, for your eternal support and the ten years’ accompanying. Thanks for your great understanding and persistence, I could carry out my dream so well. Soonly it is your turn for the doctoral defence; I would support you as you supported me, my best partner!
摘要 (Abstract in Chinese)

微生物生态领域已经进入到一个技术快速发展和数据大量产生的时期。高通量测序技术为深入研究生命蓝图提供了有力的支撑。现在，可以在无需培养的情况下同时对自然样品中成千上万种微生物进行研究。并且，通过挖掘测序数据就可以发现新的物种和新的代谢功能（基因）。尽管如此，仍然有海量的微生物以及其生态功能尚未研究。

在我的博士研究中，开发了一种高精度调查真核微生物多样性的方法，并利用多种测序技术对重要的环境问题以及生态学问题进行了解答。在第一篇论文中，我们设计了一套具有高覆盖度的 18S 核糖体基因引物，该引物扩增的核酸片段长度与当下流行测序技术的读长一致，同时，扩增的基因序列也与目前报道的 18S 参考序列的丰度相匹配。在第二篇论文中，我们首次通过测序对“波罗的海”微生物群落的空间分布展开了调查(包括真核藻类与原核藻类), 并揭示了环境因素对微生物多样性的影响。该研究采用第一篇论文设计的 18S 引物以及一对广覆盖度的 16S 引物对真核及原核藻类的 rRNA 基因进行了扩增并测序。在第三篇论文中，我们通过整合宏基因组，宏条形码以及宏转录组数据,详细研究了“波罗的海”海底沉积物(460 米深)中微生物群落的蛋白质合成能力(即活性), 从而揭示了在极端环境下微生物的多样性以及它们可能的生态功能。在第四篇论文中，我们比较了大肠杆菌培养、高通量测序以及便携式实时测序这三种技术在追踪城市雨水管网中生活污水的能力。从成本、便携度以及准确度三个方面评估了测序技术在土木工程项目中的适用性，并首次验证了实时测序仪在水质监测项目中的可行性。

综上所述，本文的研究工作展示了如何使用 DNA 测序检测微生物菌群的手段来开展环境调查以及生态研究。

关键词：DNA 测序；宏条形码技术；微生物生态；波罗的海；微生物群落；Illumina 测序技术；牛津纳米孔技术；污染源追踪；雨水径流
References


16S rRNA gene data. Methods Ecol Evol 4: 1111–1119. This paper demonstrated how careful analysis of exact 16S rRNA sequences that avoids clustering into OTUs can reveal fine-grained information that can be useful for forensic matching Article. 2013;


