FACULTY OF
BIOSCIENCE ENGINEERING

GENETIC BARCODING OF MARINE ZOOPLANKTON COMMUNITIES IN THE NORTH SEA USING THE MINION SEQUENCER

Word count: 29,515

Stijn Willemse

Student number: 01201148

Promotor: Dr. Eng. Jana Asselman
Copromotor: Dr. Eng. Michiel Vandegehuchte

Master's Dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Bioscience Engineering: Environmental Technology

Academic year: 2017 - 2018

GHENT
UNIVERSITY

PREFACE

This bundle contains my dissertation which is titled “Genetic barcoding of marine zooplankton communities in the North Sea using the MinION sequencer” and was submitted to Ghent University in June 2018 in partial fulfilment of the requirements for the degree of Master of Science in Bioscience Engineering: Environmental. It documents all the laboratory and literature research that I was actively engaged in during my time at GhenToxLab.

When I embarked on this journey, I had little to no knowledge about molecular biology and even less laboratory experience. This educational experience enabled me to uncover and explore my passion for this fascinating field. Even though the learning curve was steep at times, I am delighted by what I have learnt and achieved.

This research was conducted in collaboration with the Flanders Marine Institute (VLIZ) and the framework of the LifeWatch project between August 30 2017 and June 8 2018. They were kind enough to grant me access to the Flemish research vessel Simon Stevin for sampling campaigns and allowed me to make use of their genetics laboratory and the ZooScan. I would like to thank Dr. Eng. Michiel Vandegehuchte for assisting with the coordination of this project as well as Dr. Maarten De Rijcke and Jonas Mortelmans for investing their time in helping and instructing me.

I would like to show my appreciation to Prof. Dr. Colin Janssen and Prof. Dr. Eng. Karel De Schamphelaere for accepting me into their team as thesis student and providing the necessary facilities for me to undertake this project. I also wish to express my sincere gratitude to my supervisor Dr. Eng. Jana Asselman. I could not have completed this endeavour without her excellent guidance and adept advice. I am also grateful for Ilias Semmouri who was always there to lift my spirits whenever experiments failed and my morale was low.

Finally, I want to thank the people who took time from their busy schedules to improve the quality of this dissertation. A big thank you to Tina Mertens from VLIZ for proofreading my literature review and teaching me about policy and to Louise Montgomery for spell checking and proofreading this entire text.

Ghent, June 8, 2018

Stijn Willemse

TABLE OF CONTENT

ABBREVIATIONS

CPR	Continuous plankton recorder
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
eDNA	Environmental DNA
MM	Master mix
MSFD	Marine Strategy Framework Directive
ONT	Oxford Nanopore Technologies
PCR	Polymerase chain reaction
PIC	PCR inhibiting compounds
RV	Research vessel
VLIZ	Flanders Marine Institute

SUMMARY

Biomonitoring is a way to determine the biodiversity and ecological status of an ecosystem. This is important in a world where species are lost at an alarming rate. Animal species can be identified during biomonitoring with various methods. One of these is metabarcoding which uses genetic markers isolated from a DNA extract from the biological community under study. These can then be sequenced with one of the many available platforms. Recently, the British company Oxford Nanopore Technologies has developed a revolutionary technology (MinION) has made sequencing much more accessible. In this research, it was investigated whether it is possible to use this MinION sequencer for metabarcoding of marine zooplankton communities. Each step of the protocol was optimised, after which it was applied to samples taken during a biomonitoring campaign on the North Sea. Different DNA extraction techniques are tested, after which PCR amplification and purification are optimized by selection of an optimal primer set, PCR master mix, inhibition prevention measures and purification protocols. The amplicons were then sequenced and subdivided into species using a reference database. Finally, a comparison was made with a ZooScan analysis in which species are determined based on their morphology. The results largely conform to ZooScan determinations and show no discrepancies from what was found in the literature. Although further research in data processing is required, it can be concluded that metabarcoding with MinION sequencing is a valuable innovation in the field of biomonitoring.

SAMENVATTING

Biomonitoring is een manier om de biodiversiteit en ecologische status van een ecosysteem te bepalen. Dit is belangrijk in een wereld met een alarmerend soortenverlies. Diersoorten kunnen tijdens biomonitoring geïdentificeerd worden met verscheidene methoden. Eén hiervan is metabarcoding die gebruik maakt van genetische merkers geïsoleerd uit een DNA extract van de onderzochte biologische gemeenschap. Deze kunnen dan gesequenced worden met één van de vele beschikbare platformen. Recentelijk heeft het Britse bedrijf Oxford Nanopore Technologies een revolutionaire technologie (MinION) ontwikkeld die het sequencen veel toegankelijker maakt. In dit onderzoek werd uitgezocht of het mogelijk is om deze MinION sequencer te gebruiken voor het metabarcoden van mariene zooplankton gemeenschappen. Hierbij werd elke stap van het protocol geoptimaliseerd waarna het werd toegepast op stalen genomen tijdens een biomonitoring campagne op de Noordzee. Verschillende DNA extractie technieken worden getest, waarna PCR amplificatie en opzuivering werden geoptimaliseerd door selectie van de optimale primer set, PCR master mix, inhibitievoorkomende maatregelen en opzuiveringsprotocols. De amplicons werden daarna gesequenced en in soorten onderverdeeld aan de hand van een referentiedatabase. Tot slot werd tevens een vergelijking gemaakt met een ZooScan analyse die zich baseerr op morfologie voor soortenbepaling. De resultaten komen grotendeels overeen met ZooScan bepalingen en vertonen verder geen afwijkingen ten opzichte van wat in de literatuur werd gevonden. Ondanks dat de methode nog niet volledig op punt staat en verder onderzoek inzake dataverwerking vereist is, kan geconcludeerd worden dat metabarcoding met MinION sequencing een waardevolle innovatie is in het veld van biomonitoring.

1 INTRODUCTION

The oceans are one of the Earth's most valuable natural resources. They provide humanity with many ecosystem services such as recreation, food production, nutrient processing and climate regulation just to name a few. However, human society has profusely disrupted natural ecosystems, including the marine environment. If our species wants to move to a communalistic rather than a parasitic relationship with nature, we must first understand the mechanisms behind biodiversity loss and the role we play in this process.

Bio-monitoring is a prominent method for assessing environmental status and assisting mitigation and restoration efforts. Several strategies for bio-monitoring exist, each with their own strengths and weaknesses. One of them is called metabarcoding. Since the turn of the century, it has emerged as a promising new way for studying species compositions of biological communities. It was introduced to tackle a decline in taxonomic expertise and to provide a solution for some of the shortcomings of traditional morphology-based methods (Hebert, Cywinska, and Ball 2003). Metabarcoding is a genetic approach to species identification. By looking at a gene that is present in all studied organisms, it is possible to distinguish between species based on variable regions in the sequence. These sequences must be registered in a database and linked to their respective species. Unlike morphological identification, this method is not sensitive to morphological ambiguities such as sexual dimorphism or cryptic species and a sample can be processed in a fraction of the time.

The large cost and limited availability of sequencing laboratories have so far impeded a large scale implementation in routine laboratories. But recent developments in sequencing technology could dramatically alter the field bio-monitoring. A UK based company called Oxford Nanopore Technologies (ONT) has introduced a revolutionary new sequencing technology called the MinION sequencer. It is a small pocket-sized device that can be taken to any location and could potentially reduce the cost of sequencing to a fraction of that of conventional methods. At the time of writing this thesis, nanopore sequencing is still in its developmental phase. A lot of research still needs to be done on making the process more streamlined, as well as on how to handle the specific properties of MinION sequencing data.

In this dissertation, a protocol was developed to allow the use of MinION technology for metabarcoding of marine zooplankton communities. Zooplankton are an interesting group for bio-monitoring in the marine environment because they constitute an important link between phytoplankton as primary producers and higher trophic levels. Not only are they an important source of food for organisms living in the upper pelagic zones, they also contribute to the downward flux of biomass to benthic organisms through sedimentation of faecal pellets (Wexels Riser et al. 2002).

Samples are taken on a one day bio-monitoring campaign with the Simon Stevin research vessel (RV) on the Belgian part of the North Sea (BPNS). Then a metabarcoding protocol is developed by assessing different techniques for each step individually, by both literature review and laboratory testing. First, a suitable DNA extraction protocol is described. Then two PCR primer pairs for different barcoding genes and various iterations for PCR reaction conditions are evaluated. After the amplification step is optimised, different PCR product purification protocols are assessed. The pure amplicons are then sequenced by MinION and third-party data processing tools are applied to the output data. The steps in the data processing pipeline for which no reliable or eligible tool was available, custom python algorithms are written. Finally, a list of species names was generated for each sample, which was then compared to past bio-monitoring surveys.

2 LITERATURE REVIEW

2.1 The marine environment

Ever since the industrial revolution humanity has had a deeply transformative and often deleterious effect on the natural environment. From urbanisation to large scale exploitation of natural resources, humans have made global and irrevocable changes to Earth's landscape and ecosystems. There is compelling evidence that human activities are a major driving force behind global disasters such as global warming and have caused the onset of the planet's sixth mass extinction event (Cook et al. 2013; Barnosky et al. 2011). Anthropogenic influences are so profound that geologists have called for the introduction of the Anthropocene to delineate the epoch in which human activity is significantly evident in the geological record (Smith and Zeder 2013).

As a consequence, many ecosystems have suffered immensely during the Anthropocene and the marine environment is certainly no exception. It has endured severe losses in biological diversity as a consequence of many atrocities. Exploitation, dumping of municipal and nuclear waste, oil spills, release of eutrophication facilitating nutrients, introduction of invasive species and overfishing are just a few examples. These effectors can also have an indirect impact as is the case for ocean acidification, increasing water temperatures and changes in salinity (Lotze et al. 2006; Jackson et al. 2001; Brierley and Kingsford 2009).

Yet the oceans are one of the Earth's most valuable natural resources. They were the breeding grounds for the first life on Earth ~3.5 billion years ago. After life's conception, they have provided shelter from the hostile environment at the surface for billions of years as our protective atmosphere developed (Kasting 1993; Schopf and Packer 1987). At present, the marine environment is believed to constitute ~50% of the global oxygen production, cover over 70% of the planet's surface and make up ~99% of the total available living space (A. Longhurst et al. 1995). Also for humanity, they continue to provide many ecosystem services such as recreation, food production, regulation of nutrients and climate regulation just to name a few (de Groot et al. 2012). In the following, two prominent examples will be discussed in more detail and potential threats to ecosystems which provide these services will be assessed. However, this elaboration is not exhaustive.

2.1.1 Climate control: Carbon sink

Water can dissolve CO_2 after which the two react to form carbonic acid (H_2CO_3). The latter dissociates into bicarbonate (HCO_3^-) and then into carbonate ions (CO_3^{2-}):

$$H_2O + CO_2 \rightleftharpoons H_2CO_3 \rightleftharpoons H^- + HCO_3^- \rightleftharpoons 2H^- + CO_3^{2-}$$

The resulting CO_3^{2-} can bind to Ca^{2+} and precipitate from the water column as calcium carbonate ($CaCO_3$). Organisms such as molluscs or certain algae like foraminifera's and coccolithophores, for example, incorporate some mineral form of $CaCO_3$ into their shells. However, before CO_2 reacts with water, it can also be transformed into biomass through photosynthesis and enter the food chain. Though other pathways such as autotrophic carbon fixation (Hügler and Sievert 2010; Subramaniam et al. 2008) have been described, these two are among most the important biological mechanisms for carbon sequestration (Volk and Hoffert 1985).

As a result, the oceans hold 50 times the amount of CO_2 that the atmosphere contains, making it the largest carbon reservoir in the carbon cycle according to the National Oceanic and Atmospheric Administration (NOAA). Primary production from marine phytoplankton has been estimated to be 45-50 Gt C/year or about 50% of the global primary production (A. Longhurst et al. 1995). In most ecosystems production and consumption are in equilibrium and carbon is only sequestered in deep ocean sediments when this balance is shifted during disruptive events (Fasham et al. 2001). However, in some locations, seasonal variations in vertical carbon flux have been observed (Wexels Riser et al. 2002). A well-known example is the North Atlantic where spring blooms create large accumulations of biomass (Antia et al. 2001).

In order for this biological carbon pump to function, some conditions must be met. A peak in primary production alone won't suffice since microscopic algae, as well as their tests (i.e. the Ca- or silica-based skeleton of phytoplankton), don't sink fast enough to avoid dissolution, remineralisation or scavenging. Carbon is most effectively exported in aggregates such as faecal pellets expelled by zooplankton (Laurenceau-Cornec et al. 2015). This reduces the surface area to volume ratio, decreasing buoyancy as well as the surface area exposed to scavengers and corrosive ocean water. Studies have found that for these processes species diversity and food web dynamics are among the determining factors for carbon turnover (Poulton et al. 2007; Legendre and Rassoulzadegan 1996). These findings emphasise the importance of biodiversity and planktonic organisms for the oceans to serve as a carbon sink.

2.1.2 Marine derived goods

The most obvious good the marine environment provides is food. According to a 2016 report by the Food and Agriculture Organization of the United Nations (FAO) fish protein constitutes about 17% of global animal protein intake (FAO 2016). This amounts to over 160 million tonnes of fish produced of which 81.5 million tonnes was captured in marine waters. Finfish makes up the majority of this figure, but it also includes molluscs, crustaceans and other aquatic animals.

However, food is only part of what marine biodiversity has to offer. For example, agar which has been highly desired in microbiology for decades is derived from seaweed (Armisen and Galatas 1987). Also in healthcare, many marine derived goods are used or have shown promise for future application. Peptides extracted from algae, fish, molluscs, crustaceans or marine by-products show several properties beneficial to human health. These include antioxidant activity, anti-hypertensive activity and anti-HIV activity (Rajapakse et al. 2005; Ngo et al. 2012; Lee and Maruyama 1998).

2.1.3 Threats

All of the above-mentioned goods and services depend on marine biodiversity. However, the past few decades saw the emergence of a variety of stressors as a consequence of human progress. A paper published in Science found that diversity is positively correlated with ecosystem services and highlighted their deterioration as a consequence of accelerated erosion of biodiversity (Worm et al. 2006). Overfishing, perturbations in chemical composition and the introduction of invasive species are among the most significant causes they found.

As mentioned above, fish is an important part of the global diet. The United Nations has projected the human population to reach about 9 billion by 2050 (Economic 2006). This means that demand for food and therefore fish will not decrease anytime soon. This raises an important question: How much fish can be captured before ecosystems are irrevocably corrupted and populations collapse? Studies have shown that population collapse often coincides with on the one hand high fishing pressures over several years and on the other a lag in mitigated response (Essington et al. 2015). Also, indirect consequences of overfishing have been identified. Evidence suggests that overfished communities exhibit a lower genetic diversity, eroding their evolutionary potential

(Pinsky and Palumbi 2014). Changes to the local food web can also disrupt and degrade entire ecosystems. Loh et al. saw a threefold increase in overgrowth of Caribbean reef-building corals by palatable sponge species in areas with overfishing (Loh et al. 2015). These effects are not always immediately visible and adversities can lag decades to centuries behind the onset of overfishing (Jackson et al. 2001).

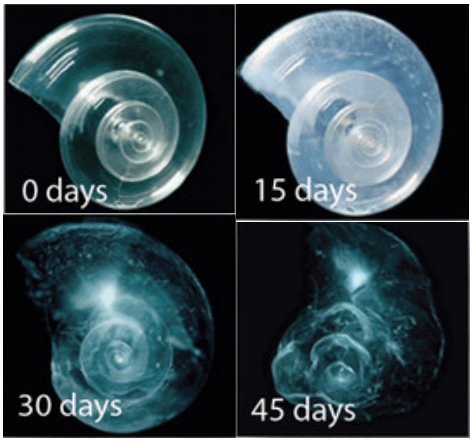

Figure 1: A Pteropod shell placed in water with carbonate levels projected for 2100. Source: NOAA. Photo credit: David Liittschwager/National Geographic Stock

Shifts in marine food webs are not the only concern for ecosystems. It has already been suggested that the oceans contain about 50 times more CO_2 than the atmosphere. This implies that as anthropogenic carbon emissions rise, the oceanic reservoirs become increasingly more saturated and the water's chemical composition will change. A higher concentration of dissolved CO_2 gas will shift the equilibrium towards more carbonic acid, thereby decreasing pH. This phenomenon is known as ocean acidification (Figure 1). Data suggests the ocean pH has already dropped by 0.1 units and that partial pressure of CO_2 has reached levels unprecedented in the last 800,000 years (Solomon et al. 2007). A rise in [H^+] will result in fewer carbonate ions hampering the precipitation of biogenic calcareous structures (Orr et al. 2005). However, many taxa rely on the protection of Ca^+-based skeletons for at least a part of their life cycle. This ultimately alters selective pressures and contributes to significant shifts in biodiversity (Sunday et al. 2017; Fabricius et al. 2014).

The last example of a major threat to the marine environment is invasive species. These are taxa from any branch of the tree of life that naturally don't occur in a certain location. Biological dispersal is a natural phenomenon crucial for species survival. However, the term "invasive species" usually implies much larger, human-induced extensions. One of the most cited human vectors are ships. Some species cause fouling of ship hulls and are thereby transported over vast distances. Similarly, discharging ballast water can introduce taxa or pathogens into foreign habitats, impacting the local environment. For example, evidence has pointed to the invasive comb jelly, *Mnemiopsis leidyi,* as the cause of a sharp drop in biodiversity and the collapse of coastal fisheries in the Black Sea (Shiganova 1998).

These are just a few examples of the many strains marine ecosystems endure. Much more research needs to be done on the precise interplay of different effectors and on subsequent mitigation strategies.

2.2 Legal framework

Evidently, there is an urgent need for effective mitigation and also restoration efforts if we want to protect and preserve marine ecosystems. During the end of the 20th century, the scientific community, as well as the general public, grew increasingly more concerned about surface water quality in Europe. There was a strong demand for a collective approach to prevention and removal of strains on aquatic ecosystems. In 2000 the Water Framework Directive was established which concerns itself with river basins and groundwater quality (*Directive 2000/60/EC* 2000). In June 2008 additional legislation followed when the European Commission proposed the 2008/56/EC Marine Strategy Framework Directive (MSFD) for the protection and reconstruction of marine ecosystems.

The general aim of the MSFD is to achieve good environmental status in EU seas by 2020. Legislators hope to do this by providing methodological standards for quality assessment and setting detailed criteria for quality parameters. Each member state is responsible for developing and implementing a strategy to meet the criteria imposed by the MSFD (*Directive 2008/56/EC* 2008). This strategy should describe how good environmental status will be achieved by using eleven descriptive elements, i.e. biological diversity, invasive species, commercially exploited species (fish, crustaceans and shellfish), the food chain, enrichment by nutrients (fertilisers), the integrity of the seabed, hydrography (currents, salinity, temperature etc., of the seawater), pollution, food safety, marine litter and underwater noise.

Shortly after the MSFD was formulated, the Belgian federal government had transposed the directive into the Royal Decree of June 23, 2010. In 2012 an initial assessment of the environmental status of the Belgian part of the North Sea (BPNS) was conducted and subsequent environmental objectives were established. The findings were published in a 30-page report that is freely available on the Health, Food Chain Safety and Environment public service website of the Belgian federal government. Progress is evaluated every six years and the strategies, methodologies and criteria are adjusted accordingly. The first review report is due later this year (2018).

Evaluation of environmental status is based on both continuous and periodic monitoring programs. They assess a wide variety of parameters which indicate the state of, as well as the pressures on the marine environment. Biomonitoring is an important part of this. It is used to evaluate biodiversity and by extension the ecological status of a certain area. The AZTI Marine Biotic Index (AMBI) is such an ecological indicator and has already been applied for the determination of ecological quality status in the context of the WFD (Muxika, Borja, and Bonne 2005). The core idea is that the presence of sensitive species in a certain location is an indication of a good ecological health. This way a site can be scored based on the absence or presence of key species. A major contributor to biomonitoring efforts in Europe is the European Strategy Forum on Research Infrastructure (ESFRI) project called LifeWatch. They operate in Europe both on land as well as at sea. The Continuous Plankton Recorder survey is another marine biomonitoring program that has made unprecedented contributions to knowledge about plankton biodiversity. Both projects are active in the BPNS and will be discussed in more detail below. First, the study area of this thesis is contextualised in the following paragraph.

2.3 The Belgian part of the North Sea

The North Sea is located on the European continental shelf between the UK, the European mainland and South Norway (see Figure 2). As the North Sea is located on a continental shelf, it is relatively shallow. The water depth rarely exceeds 40 m in the Southern Bight. It increases across the Central North Sea to a maximum of 100 m and reaches up to 200 m in the Northern regions (Paramor et al. 2009). Additionally, the area is surrounded by a wide variety of land masses and water bodies. It connects to the Atlantic through the English Channel beyond the strait of Dover down south and receives fresher water from the Baltic Sea through the Skagerrak strait between Sweden and

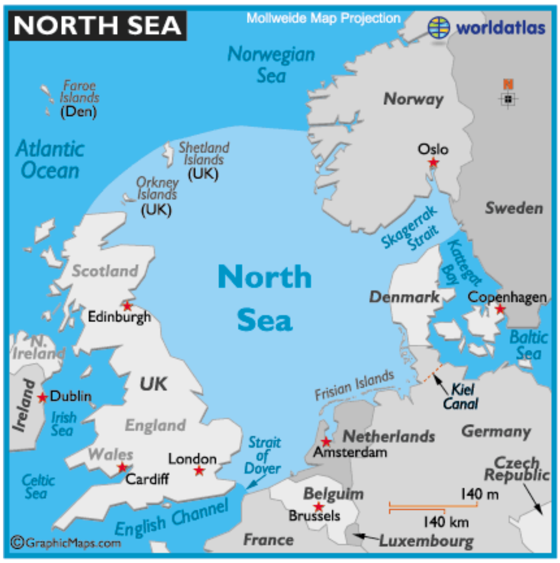

Figure 2: Approximate outline of the North Sea. Source: Worldatlas.

Denmark. In the north, it transitions into the Atlantic Ocean. Finally, the North Sea receives fresh water from multiple big rivers carrying with them material and nutrients from vastly different continental areas (Lacroix et al. 2007; Sanders, Jickells, and Mills 2001). These many interactions have resulted in a relatively heterogeneous distribution of water quality and hydrographic characteristics (Otto et al. 1990). The resulting spatial variability of habitats has created many different ecosystem compositions and dynamics in the rather small area that the North Sea comprises (Richardson et al. 1998; A. R. Longhurst 2010).

The Belgian part of the North Sea (BPNS) is located in the Southern Bight and so this part of the sea is less than 40 m deep. It is located in an interesting region because high salinity Atlantic water extends into the North Sea from the south, and freshwater enters the BPNS in large amounts from the Scheldt estuary up north (Böhnecke 1922; Lacroix et al. 2004). Strong tidal currents and winds in this narrow corner of the North Sea results in an intense mixing of these waters.

2.4 Bio-monitoring

Water quality could easily be monitored with conventional tools that are able to perform online measurement of many parameters such as temperature, turbidity, salinity, etc. Concentrations of organic and inorganic pollutants can easily be determined from routinely taken water samples. However, it is difficult to infer any tangible information about environmental status from these measurements alone. The impact that the prevalence of one chemical has on an ecosystem, can only be assessed after ecotoxicological research. In April 2018 a total of ~18700 substances had been registered under REACH. It is extremely impractical to measure all physical parameters and chemical components and then make a meaningful prediction about ecological status. That's why it's more feasible to look at biodiversity directly through bio-monitoring when investigating ecological status.

2.4.1 Examples of biomonitoring

2.4.1.1 LifeWatch

LifeWatch is a European bio-monitoring program that is part of the *European Strategy Forum on Research Infrastructure* (ESFRI). With this program, Europe aims to strengthen scientific knowledge about biodiversity and to better understand how climate change affects ecosystems globally. LifeWatch has its headquarters in Spain but divisions can be found across all of Europe. Data is gathered through bio-monitoring campaigns and shared with scientists, stakeholders and citizens from an open access datacentre. In Belgium, institutes which are financed by the federal and Flemish government as well as by the Wallonia-Brussels Federation contribute to LifeWatch activity. The Flemish consortium consists of Flanders Marine Institute (VLIZ) and by the Flemish Research Institute for Nature and Forest (INBO). In both cases funds are provided by the Research Foundation - Flanders (FWO).

Plankton is one of the groups of interest for the LifeWatch efforts and monitoring campaign are organised by VLIZ. At the moment species identification is based on morphology. Advantages and objections to this method are discussed below.

2.4.1.2 Continuous Plankton Recorder survey

The continuous plankton recorder (CPR) survey was first initiated in 1931 by marine biologists Sir Alister Hardy and Sir Cyril Lucas. Today it is one of the oldest continuously run bio-monitoring programs in the world. The sampling method has remained largely the same since 1948, enabling representative comparison of analyses made today to those dating back up to 70 years. This gives scientists unique and comprehensive insight into the spatial and temporal dynamics of zooplankton communities.

The CPR was designed to be towed behind merchant ships so that they could be deployed on the existing global fleet. Anyone who wants to contribute to data collection can contact the Sir Alister Hardy Foundation for Ocean Science (SAHFOS) who now manages the program.

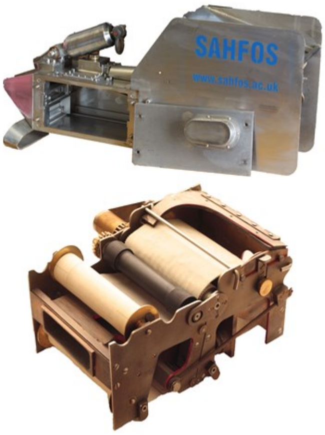

Figure 3: The Continuous plankton recorder (top) and the filter cassette (bottom). Source: SAHFOS

Samples are taken by filtering water on a constantly moving sheet of filter silk. The roll is turned by gears which connect to a small propeller driven by the current. After a trip, the recorder is returned to the lab where the filter roll is divided into sheets representing 10 nautical miles. For each sheet, first the colour is characterised to obtain a semi-quantitative measurement of phytoplankton density. Taxonomists then identify the phytoplankton and zooplankton species under a microscope.

2.5 Plankton

The Britannica Encyclopaedia defines plankton as microscopic organisms that drift or float in the water column and are unable to swim against the current. They are an important source of food for animals in higher trophic levels (Turner 2004). There are many different types of planktonic species. A first division that can be made is phytoplankton versus zooplankton. As the names suggest, phytoplankton are plant-like and zooplankton are animal-like plankton. This distinction does however quickly unravel when looking at single-celled plankton. For example, scientists often use the term protozoa to refer to heterotrophic single-celled eukaryotes. However, the use of this term is discouraged as it wrongly implies that they are animal-like. Regarding single celled organisms as one taxon under this name implies a direct shared lineage with Animalia, which genetic evidence has discarded (Cavalier-Smith 2003). Since the phylogeny of eukaryotic unicellular life does not conform to this arbitrary plant/animal distinction, the term Protista is preferred to describe this group.

2.5.1 Phytoplankton

Phytoplankton are autotrophic eukaryotes that are essential to marine ecosystems as the primary source of food. Furthermore, they are responsible for 50-85% of the photosynthetically produced oxygen on Earth (EarthSky 2015; Ryther 1970). Despite providing crucial ecosystem services, phytoplankton can in some circumstances also be detrimental to the environment. When weather conditions are favourable and nutrients abundant (i.e. eutrophication), algal blooms can occur during which algae rapidly multiply into large populations. These blooms occur naturally each year mainly in spring. During the dark winter months, light is the limiting factor for algal growth, allowing nutrients to accumulate. When light becomes more available in spring, algal growth will explode due to the high abundance of nutrients. This mechanism has been found to be partially influenced by human activity through continental nutrient run-off (Beman, Arrigo, and Matson 2005).

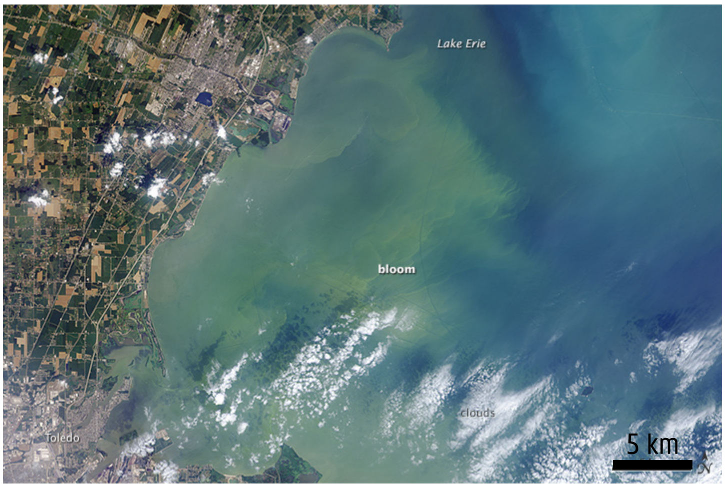

Figure 4: Satellite picture of algal bloom in Lake Erie. Source: NASA Earth Observatory, 2014

Even though the individual organisms are not visible to the naked eye, they occur in such large numbers that blooms can be seen from space (Figure 4). Ocean waters can become brightly coloured during algal blooms because of the pigments algal cells contain. An example of this is the red tide caused by *Korenia brevis* (Flewelling et al. 2005). Though excessively produced during the day, oxygen is only consumed by algae at night when photosynthesis is not possible. Furthermore, when organisms die they introduce a significant amount of organic carbon into the environment, increasing oxygen demand for microbial decomposition. Both mechanisms can cause hypoxia (i.e. a depletion of oxygen) leading to an increased mortality in all oxygen dependent organisms and subsequently disturbing the local food web (Pihl 1994). Some blooms can cause more direct harm as certain algal species produce toxins which can do significant damage to ecosystems and ecosystem services (Jin, Thunberg, and Hoagland 2008). These are called harmful algal blooms (HABs).

In the following, phytoplankton and other Protista will not be discussed in further detail since this is out of the scope of this research topic.

2.5.2 Zooplankton

Zooplankton can be single celled or multicellular organisms. Unicellular zooplankton actually refers to a paraphyletic group which contains heterotrophic Protista and unicellular Animalia (e.g. Myxozoa). This group will not be further discussed as this is out of the scope of this thesis.

Metazoan zooplankton can be further subdivided into two groups, meroplankton and holoplankton. Meroplankton typically spend their larval and juvenile life stages as plankton, after which they grow into adults and revert back to the benthic or nekton environment. Examples of this type of organisms can be found in many different phyla, such as Molluska, Cnidaria, Annelida, Echinodermata, Porifera and Chordata. Holoplankton, on the other hand, spend their entire life cycle in the plankton environment. A prominent example of this are the Copepods, a subclass of Crustaceans.

Copepods are among the most abundant planktonic species in the marine environment. They form a crucial link in the marine food web. Not only are they thought to be the largest group of diatom grazing organisms, young fish rely on them as a main portion of their food source (Rae and Rees 1947). Additionally, copepods are a zooplankton group that exhibit a wide diversity of species. In some locations, a single net sample can contain more than a hundred copepod species (Angel 1997). According to observations made by Rae and Rees 1947, the most dominant species of planktonic copepods are Acartia spp., Pseudocalanus elongates, Microcalanus pusillus, Paracalanus parvus and Temora longicornis. Johns and Reid 2001 ascertained that also Calanus spp. and Oithona spp. were among the top 10 most abundant plankton species.

Other, less abundant zooplankton taxa include echinoderm larvae and Onychopoda (Johns and Reid 2001). Echinoderms are meriplankton, so larvae eventually grow into the various starfish, brittle stars and sea urchins that inhabit the North Sea's benthic environment whereas Onychopoda are another type of holoplanktonic Crustaceans. Prominent species in the BPNS include Evadne spp. and Oithona spp. The most recent comprehensive elaboration on zooplankton in the BPNS (Figure 5), was published by Karl Van Ginderdeuren (Karl Van Ginderdeuren et al. 2012).

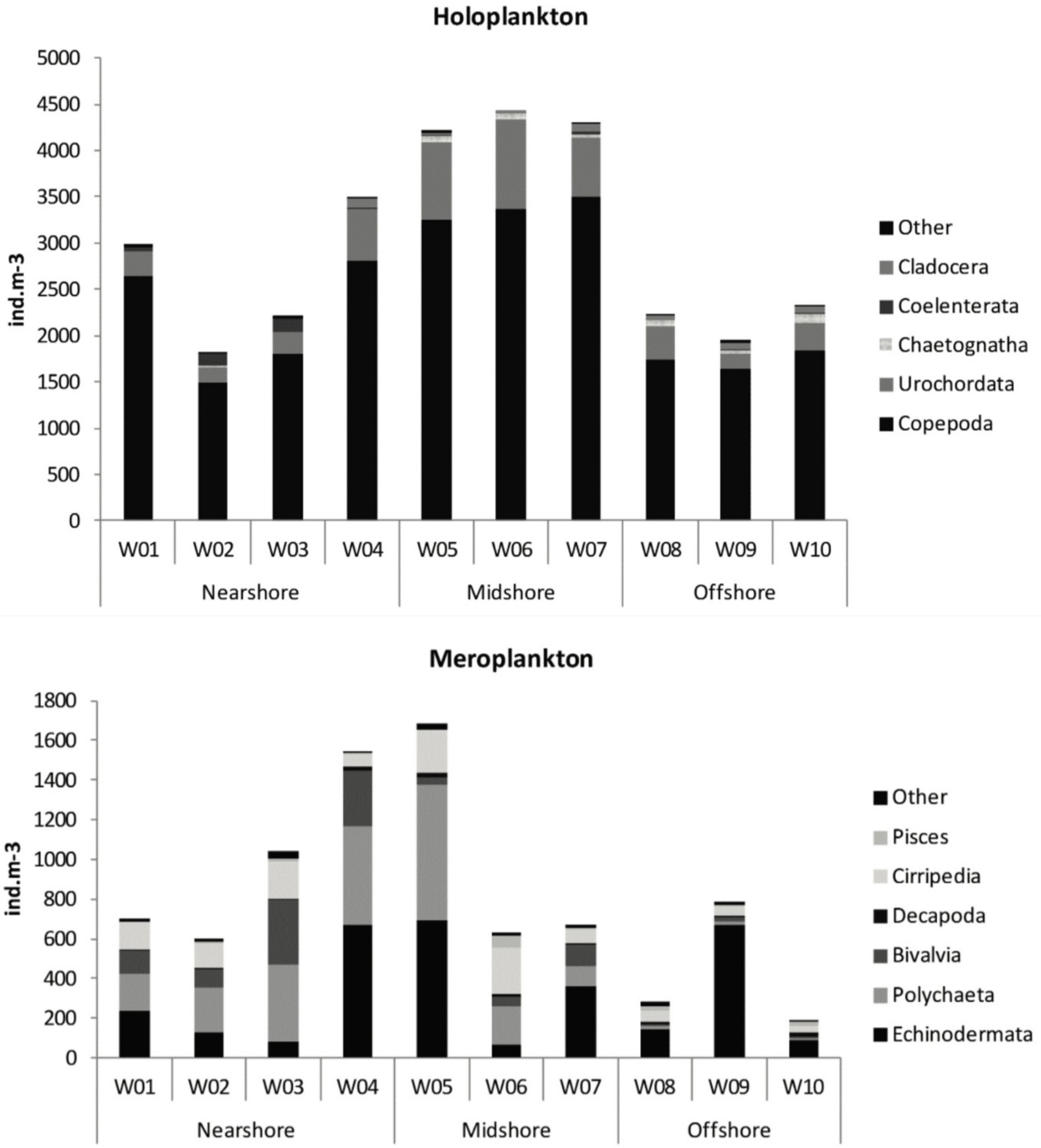

Figure 5: Average absolute abundance of zooplankton (ind./m^3) for each station. (Graph taken from Van Ginderdeuren et al. 2012)

2.6 Methods to study biodiversity

2.6.1 Taxonomy

Taxonomy is the science in which organisms are named and classified into groups (also called taxa), based on similarities in certain characteristics. The goal is to facilitate unambiguous communication about biodiversity by introducing a standardised nomenclature. This avoids the use of multiple names for the same species or superfluous descriptions of a specimen. Taxa can be placed in relation to each other in a hierarchic system. Taxa belonging to the same taxonomic level, share the same rank. The lower the taxonomic level or rank shared by two organisms, the higher the resemblance between them. The conception of these taxa can be done in a number of different ways, such as e.g. appearance, attributes or decent. The first scientist in recorded history to ever classify organisms into groups was Aristotle with his Historia Animalium. He divided organisms based on attributes such as the number of legs, the presence or absence of blood, ability to fly, etc. It was not an evolutionary system, resulting in taxa that contained organisms from widely different evolutionary decent (e.g. butterflies and birds can both fly) (Hull 1965). In modern taxonomy, taxa are conceived and related according to their phylogeny, the science of unravelling the evolutionary history of individuals. This is called Darwinian taxonomy.

2.6.2 Morphological identification

Still, in the modern system of classification, species are traditionally identified through their morphology using determination guides (e.g. Bolton 1994). It is a cheap and straightforward method and many identification guides are available. The downside of this method is that it is labour intensive and time consuming work for which a high level of experience is required. Especially small animals like plankton that require a microscope to be observed are difficult to manipulate and can be fragile. Additionally, human errors can occur, resulting in a faulty identification of certain attributes. This makes the method dependent on the scientist conducting the analysis.

Different techniques have already been devised to automate morphological identification, drastically increasing throughput and objectivity of a sample analysis. VLIZ owns three such devices that are available to Belgian as well as international marine research institutes, namely FlowCAM, Zooscan and Video Plankton Recorder.

2.6.3 FlowCAM

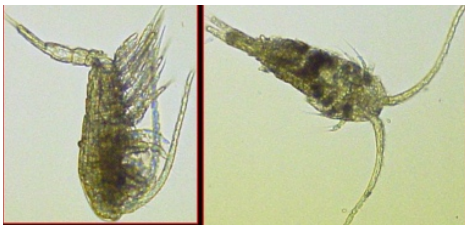

Figure 6: FlowCAM image of zooplankton from the Baltic Sea (4X). Source: Fluid Imaging Technologies, Inc.

The FlowCAM is a device produced by the company Fluid Imaging Technologies Inc. It uses imaging techniques and the VisualSpreadsheet® identification algorithm to semi-automatically detect and classify particles in a liquid sample as it flows through the device (Figure 6). Four magnifications are available allowing measurement in four size ranges between 2 µm and 2 mm. Particles mostly consist of phytoplankton, though zooplankton can also be analysed. Identification is based on morphology or fluorescence in phytoplankton cells. Studies have shown that in practice the taxonomic resolution is quite low, despite having a broad taxonomic range (Álvarez et al. 2014). The FlowCAM does, however, excel in determining cell count (Ide et al. 2008) and has yielded similar results in cell sizing trials as light microscopy, albeit at a fraction of the time (Spaulding et al. 2012) compared to manual measurement. This makes it an excellent addition to traditional identification methods.

2.6.4 ZooScan

Much like a flatbed scanner, the ZooScan produces images (Figure 7) of a thin layer of liquid containing a zooplankton sample. The images have an approximate resolution of 10.6 µm per pixel. This way computer software is able to analyse objects as small as 200 µm, which perfectly accommodates zooplankton community studies. After imaging, the scanning chamber can be drained, recovering the sample without damage, making the analysis non-destructive. Once the image is uploaded to a computer, it can be analysed by a machine learning algorithm that compares each individual to a library database. Such libraries can be compiled for every organism of interest or imported from existing databases. The algorithm is also able to discriminate zooplankton specimens from inorganic detritus like plastic debris and sediment.

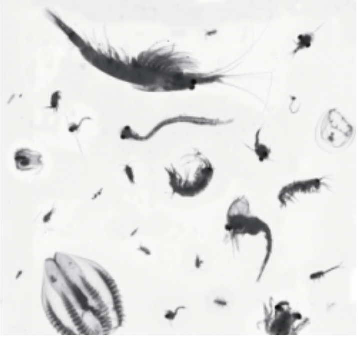

Figure 7: Example of ZooScan image (©VLIZ). Source: LifeWatch Belgium

Several publications have explored the validity of this technique. One study showed that specimen identification and counts were sufficiently accurate for abundant species, but became less reliable for rarer species (Gorsky et al. 2010). Results from a body size estimation survey on the other hand are very promising. This is exciting as many ecological parameters have been shown to be correlated with body size distribution of species (Brown et al. 2004). The ZooScan is a great addition to biomonitoring when high throughput and less accuracy is required.

2.6.5 Video Plankton Recorder

The video plankton recorder was developed to allow in situ imaging and identification of the zooplankton community while being towed by a ship. It contains a camera and stroboscope which acts as a light source. The images are sent to a computer where an algorithm identifies species from 100 µm to a few centimetres, in a similar fashion as the ZooScan does (Davis et al. 2005). The fact that the imaging is performed in situ is a great added value since this data can be coupled with abiotic measurements from the CTD (i.e. instrument which measures conductivity, temperature and pressure). This device has limitations similar to those of the ZooScan in that it generally cannot identify individuals down to a species level.

2.7 Molecular methods

Morphology was long the only thinkable method for species identification but over the past few decades, an alternative approach has emerged. The molecular method looks at the code of life to gain insight into the classification and evolutionary relations between taxa.

2.7.1 Introduction to sequencing

Deoxyribonucleic acid or DNA was first discovered by Friedrich Miescher in his search for the fundamental building blocks of life (Dahm 2005). He obtained his first isolate in 1869 and showed that its properties differed greatly from those of proteins. The significance of this discovery was long underappreciated and it was not until 1944, almost 50 years after his death, that the function of DNA was better understood. Avery, MacLeod, and

McCarty managed for the first time to transform a bacterium *in vitro* from one strain to another by introducing a strand of DNA (Avery, MacLeod, and McCarty 1944). This study showed that even though proteins are the driving force behind functionality in an organism, DNA carries the necessary information.

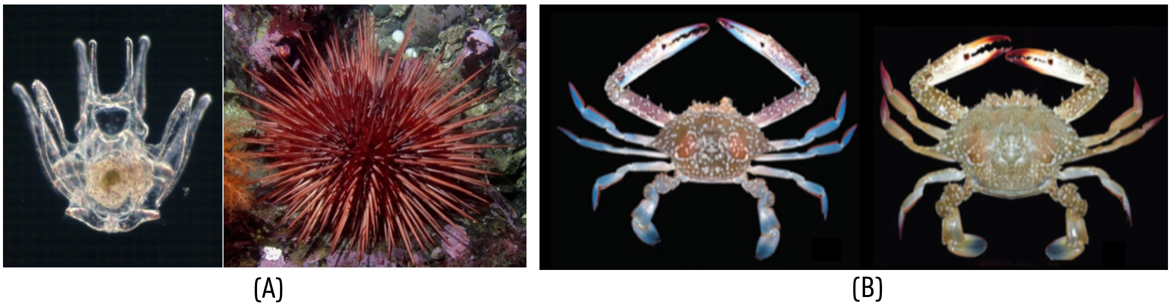

Figure 8: Intraspecies morphological diversity. (A) Strongylocentrotus franciscanus, mid-stage pluteus larva (left), adult (right) Source: University of Saskatchewan. (B) Portunus pelagicus, male (left) and female (right). Source: (Lai, Ng, and Davie 2010)

The current structural model of DNA was first proposed in 1953 by Nobel prize winning molecular biologists James Watson and Francis Crick. They found that DNA is composed of four nucleic acids — adenine (A), guanine (G), thymine (T) and cytosine (C) (J. D. Watson and Crick 1953). At this point, however, it was not yet possible to determine the precise order of these nucleotides in a given strand of DNA. Development in this field slowly progressed during the 1970's. The first RNA genome to be completely sequenced was that of the MS2 bacteriophage and was constructed at Ghent University by Walter Fiers in 1976 (Fiers et al. 1976). One year later Frederick Sanger described the first rapid genome sequencing method (Sanger, Nicklen, and Coulson 1977). Due to first generation sequencing, or Sanger sequencing, it became possible to determine nucleotide sequences — both DNA and RNA — of any organism and link it to its functionality.

The invention of sequencing enabled disruptive research and marked the beginning of numerous novel fields of study such as bioinformatics, synthetic biology and hereditary studies. This technology also enabled the identification of organisms through their genetic code and resulted in the conception of the field of molecular taxonomy.

2.7.2 Molecular taxonomy

Sometimes it is extremely difficult to discern any difference between two species based solely on their morphology. This is the case for cryptic species — organisms that are morphologically indistinguishable, yet genetically differentiated to such extent that they are unable to produce fertile offspring. In such instances, it will however be possible to discriminate between these morphologically identical species using molecular taxonomy. It constitutes the science of classifying organisms and determining their evolutionary distance based on their molecular compositions (RNA, DNA and proteins), using statistics and genetic models. It leans on the assumption that intra-species variations of molecular structure are significantly smaller than those between different species (Hebert, Cywinska, and Ball 2003). Though this is generally the case, this assumption is not always valid. That is why in this context the term operational taxonomic unit (OTU) is used instead of 'species'. It describes a group of organisms that are grouped together based on a resemblance in molecular structure.

The procedure is relatively straightforward. First DNA must be extracted from a tissue sample and separated from the remaining cell content and organic debris. From this genomic DNA (gDNA), a small genetic marker is subsequently isolated and amplified to allow the fragment to be read through sequencing. Identification is digital, making it easier to standardise. Only a tissue sample is needed for determination, so damaged specimens which might have lost characteristic morphological features can still be unequivocally identified. The same can be said for species that show great morphological diversity, such as those that exhibit sexual dimorphism, phenotypic plasticity or have drastically different larval and adult stages (Figure 8). The genetic markers used should be independent of an individual's life stage or gender, e.g. the genetic barcodes discussed below. Hence, this classification method is more accurate and objective. It eliminates the need for specialised taxonomists and is less time consuming, thereby overcoming many of the limitations of traditional taxonomy (Paquin and Hedin 2004).

Promising as this method may sound, it is not fail-proof. During species divergence, it takes some time (usually less than 100 000 years) before a characteristic mutation is cemented into the population gene pool (Tautz et al. 2003). Therefore, in recent species, the morphological differences might be more apparent and unambiguous than their marker genes. Additionally, one of its advantages is also its second pitfall. The genetic marker of an organism won't be influenced by its age, gender or body size. This means that age distributions or gender ratios cannot be inferred from this data, resulting in a loss of information about ecological status.

2.7.3 DNA barcoding

In 2003, Paul Hebert proposed DNA barcoding as a new way of looking at biological identification (Hebert, Cywinska, and Ball 2003). Metabarcoding is an identification method that adheres to the principles of molecular taxonomy. However, the cornerstone of this method is that one or a few genes can be used to identify organisms from all branches of the tree of life. It is estimated that around 8.7 million species exist on Earth (Mora et al. 2011). Since DNA consists of 4 nucleotides (A, T, C or G), this means that a strand of only 12 base pairs (bp) (4^{12} = ~17 million) can yield twice the amount of unique sequences necessary to unequivocally identify all species on Earth. This is, of course, an oversimplification and as will be discussed below, the problem is certainly much more complex. But this illustrates the potential that DNA has as an identification tool.

Paul Hebert suggested that for this concept to work, one or a few standardized barcoding genes should be selected based on scientific literature. From this gene a community based open access database can be constructed, containing a barcode reference library of all available species and relevant metadata (e.g. geographical location, time of sampling,...). This way researchers around the world can submit data to the organisation, contributing to the expansion of the library, which is made publicly accessible. Today, multiple databases are available and used by researchers worldwide. For example, the Barcode of Life library contains COI gene sequences and the SILVA database gathers 18S gene barcodes (Ratnasingham and Hebert 2007; Pruesse et al. 2007; Lai, Ng, and Davie 2010).

Genetic barcoding can be considered as a black box method. DNA is extracted from a bulk environmental sample (i.e. metagenomics) and after an appropriate genetic barcoding protocol, a list of detected OTU's is obtained (Figure 9). This can be achieved without the need for sample matrix clean up, isolation of individual specimens or morphological identification. It is important to keep in mind that this is a method which detects strands of DNA. The presence or absence of a certain OTU can depend on many different factors and is not perfectly correlated with the prevalence of the supposed corresponding taxon.

When a taxon is not detected, this does not necessarily imply that it is absent in the area where the sample was taken. Like with any sampling method, rare species could have avoided being captured or been excluded from subsampling. Yet, even if a specimen is included in the DNA extraction, it does not guaranty detection. Loss of DNA template is inherent to any extraction protocol. Also during subsequent manipulations losses can occur and multiple dilutions and subsamples are made.

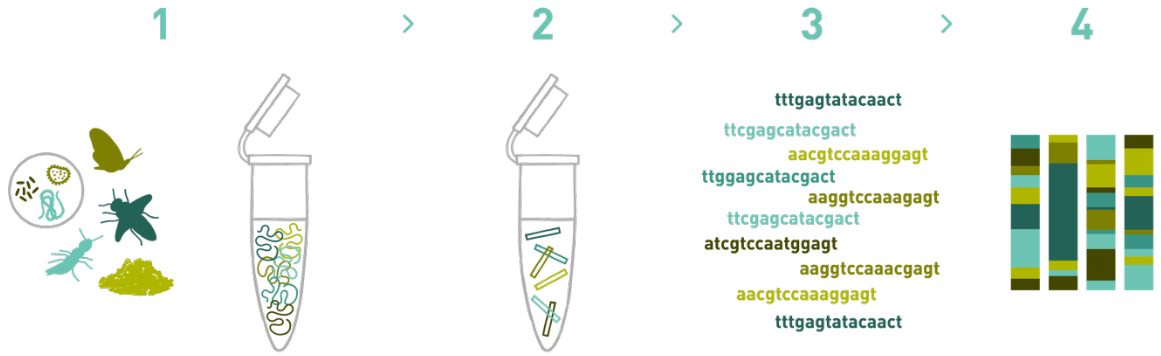

Figure 9: General overview of barcoding workflow. DNA extraction of metazoan community (1), amplification of barcoding (2), sequencing of amplicons (3) and OTU determination and further ecological analysis (4). (Source: AllGenetics)

On the other hand, also false positives can be obtained if the sample was contaminated during handling. DNA is a relatively robust molecule and carry-over between samples should be minimised with the appropriate measures (Kwok 1990). Additionally, during amplification of genetic markers with a polymerase chain reaction (PCR) a range of different biases can be introduced. Firstly, it has been found that primer sets do not anneal to all templates with the same binding energy, thereby shifting amplicon ratios in multi-template reactions such as in metagenomics (Polz and Cavanaugh 1998). Selection of a primer set thereby inherently introduces a taxonomic bias (Clarke et al. 2014). Secondly, Chandler, Fredrickson, and Brockman 1997 have described fluctuations in priming efficiency when template concentrations are low. This introduces a bias for organisms that are less abundant.

As the sample matrix generally is not cleaned up before DNA extraction, some components can be introduced which disrupt the analysis. These are called PCR inhibiting compounds (PIC). As the name suggests, they inhibit the PCR reaction which is a crucial step in a barcoding protocol. These compounds are different for every type of environmental sample, and for every sample matrix, the protocol has to be adjusted accordingly (Rådström et al. 2004).

2.7.4 Barcoding genes

Before a gene region can be considered as a DNA barcode, it has to meet a few conditions. First of all, it must be a gene that is widespread among as many species as possible. Secondly, it must contain conservative regions to facilitate the development of a universal primer set. In between these conservative regions, it must also contain highly variable sequences that allow identification on a high taxonomic level. Lastly, the length of the barcode should be carefully considered. A long sequence can grant more specificity but will increase post-sequencing processing time and complexity. The following sections will elaborate on the two most commonly used barcoding genes.

2.7.4.1 Mitochondrial cytochrome oxidase subunit I gene

The mitochondrial cytochrome oxidase subunit I gene (COI) was the very first marker that Paul Hebert suggested as the core of genetic barcoding when he first proposed the method. In his publication, he demonstrated its viability by creating COI profiles of certain taxa and successfully assigning new samples to their correct taxon (Hebert, Cywinska, and Ball 2003). This was done first on a phylum level, then on an order level and finally on a species level. This tool is so effective that researchers have even managed to identify museum archive specimens up to 7 years old using mini-barcodes on highly degraded DNA (Hajibabaei et al. 2006).

Since COI stems from mitochondrial DNA (mtDNA), its application is limited to Eukaryotes. However, the rate of mutation varies between different kingdoms. In plants, for example, it has been demonstrated that substitution rates in mtDNA are very low (Wolfe, Li, and Sharp 1987). Therefore, this gene is not an eligible candidate for this group. Also, Fungi mtDNA doesn't easily support the application of COI barcoding (Stockinger, Krüger, and Schüßler 2010). Due to this, COI has only been officially recognised as a barcoding gene for Animalia. However, Cnidaria are excluded after evidence suggested that over 94% of Cnidarian species show <2% divergence (Hebert, Ratnasingham, and de Waard 2003).

In 1994 the first primer set for COI (LCO1490/HCO2198) had already been designed and tested as a genetic marker (Folmer et al. 1994). Throughout the past decades, this primer set has repeatedly proven its value for genetic barcoding of a wide variety of animal species and has been called the "Rosetta Stone" for metazoan zooplankton (Bucklin et al. 2010). This has not stopped researchers from designing degenerate forms or even targeting completely different regions in an attempt to improve specificity and ubiquity (Leray et al. 2013).

2.7.4.2 18S nuclear ribosomal RNA gene

The 18S ribosomal RNA (18S rRNA) gene is located on the nuclear genome. All organisms have ribosomes which consist of 2 subunits made up of RNA strands and proteins, the small (SSU) and large (LSU) subunit. The SSU only contains one strand of rRNA, the 18S rRNA strand in Eukaryotes and 16S in Prokaryotes. The gene is present in all living cells so the requirement of ubiquity is certainly met. Also, both conservative and highly variable regions (V1-V9) have been identified and mapped for 18S (De Rijk et al. 1992). Similar to COI, 18S is also a great candidate as a genetic marker for barcoding. The latter, however, has not received as much attention as COI. This might be due to its lower evolutionary rate, which makes it harder to distinguish between closely related taxa (Machida and Tsuda 2010). The flipside of a lower mutation rate is that universal primers are easier to design. Finally, alignment of 18S might be more difficult than with COI because of a higher occurrence of indels (Machida and Knowlton 2012).

2.7.5 Barcoding in the marine environment

A lot of research has been done involving genetic barcoding in the marine environment. As has been mentioned above, Copepods are the most important group of zooplankton species. Hence, most of the following discussion will involve this taxon.

The COI gene was first proposed as a universal genetic marker by Hebert 2003. Since then, many researchers have assessed its usability and found a variety of applications. During a sampling campaign conducted in April 2006 by Bucklin et al. 297 specimens of 175 holoplanktonic species were collected and identified with the help of taxonomists, and barcodes of each group were determined. Based on their results and previous literature they concluded that the COI gene a useful marker for zooplankton communities and even called it the "Rosetta Stone" for marine metazoan (Bucklin et al. 2010). Researchers have exploited its wide marine taxonomic coverage for applications such as metabarcoding of fish gut contents (Harms-Tuohy, Schizas, and Appeldoorn 2016) and identification of marine crustaceans in the North Sea (Raupach et al. 2015).

However, the 18S marker has also proven its worth in marine metagenomics. A study published in 2014 thoroughly investigated the 18S gene in order to design the best universal primer set. It was the first complete 18S gene characterisation and they used all eukaryotic sequences that were available in the SILVA database for their assessment. The primers with the best theoretical matches were then tested on marine sediment samples. Hadziavdic et al. proposes the F-566/R-1200 combination based on their wide taxonomic range (Hadziavdic et al. 2014). Theoretical coverage reached 80% and empirical data showed amplification for 71% of all eukaryotic phyla. The gene has been used in conjunction with COI to shine a light on phylogenetic relations between marine invertebrate species on a higher taxonomic level (Grant and Linse 2009) and to identify prey species in faecal material from European eel larvae (Riemann et al. 2010).

2.7.6 Sequencing platforms

2.7.6.1 First generation: Sanger sequencing

The first generation of rapid polynucleotide sequencing was first developed in 1977 by Frederick Sanger and his team (Sanger, Nicklen, and Coulson 1977). It has since been optimised and the technique that is currently used was first described in 1990 (Cohen, Najarian, and Karger 1990). Dye-terminator sequencing is a PCR based technique in which polymerase randomly incorporates terminal fluorescent dideoxynucleotides (see Figure 10). As they lack an oxygen atom in their ribonucleotide, chain extension is no longer possible and the reaction is terminated. To this chain-terminal nucleotide, one of four colours of fluorescent dye is bound depending on the nucleotide. The result of this PCR reaction is a mix of extension products of different lengths with a fluorescent chain-terminal nucleotide at the 3' end of the strand. The products can now be separated using capillary electrophoresis (CE). In this process, the DNA strands are forced through a gel polymer by an electric field. The speed at which the fragment migrates through the gel depends on its length. Shorter fragments can move through the matrix more freely and reach the detector faster. Here a laser is used to excite the fluorescently labelled terminal nucleotide, causing it to emit photons at a wavelength that is indicative of the nucleotide. Hereby a chromatogram is created that base calling software can translate into a sequence.

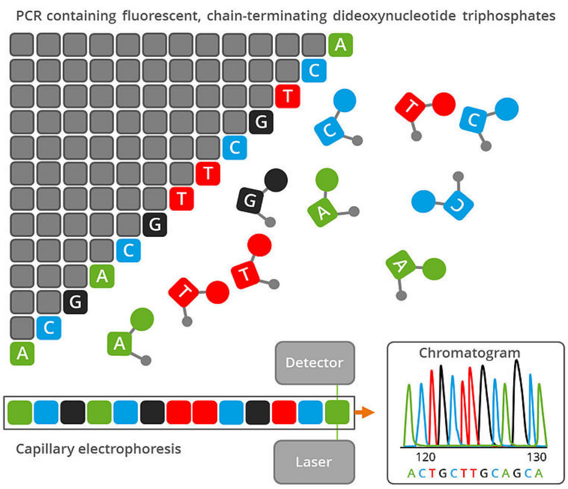
PCR containing fluorescent, chain-terminating dideoxynucleotide triphosphates
A
C
T
G
C
T
T
G
C
A
G
C
A
Detector
Capillary electrophoresis
Laser
Chromatogram
120
130
ACTGCTTGCAGCA

Figure 10: Sanger Sequencing. Source: GATC Biotech

2.7.6.2 Second generation: Sequencing by synthesis

The second generation of rapid DNA sequencing is sequencing by synthesis. It too is a PCR based method in which – unlike Sanger sequencing – nucleotides are identified upon polymerisation. Many different variations of this concept have been proposed. Here only Illumina sequencing is briefly discussed because it is most commonly found in the literature.

Illumina sequencing can only accurately determine a relatively short read length. Therefore, a library preparation step called tagmentation is performed. Transposase enzymes will randomly cut the DNA template into shorter polynucleotides, called tags. These have been supplemented with adapters at each end. By reduced cycle amplification, additional motives are added, such as the sequencing primer binding site, indices and regions complementary to the flow cell oligoes.

An Illumina flow cell consists of a glass slide with two types of oligonucleotides bounds to its surface. During library preparation, DNA fragments have been provided with regions complementary to the 2 types of oligoes at each end of the strand. When single stranded fragments are introduced to the flow cell they will anneal to their respective oligoes (see Figure 11a). At this point, the density of DNA fragments on the flow cell surface is relatively low to facilitate the bridge amplification step.

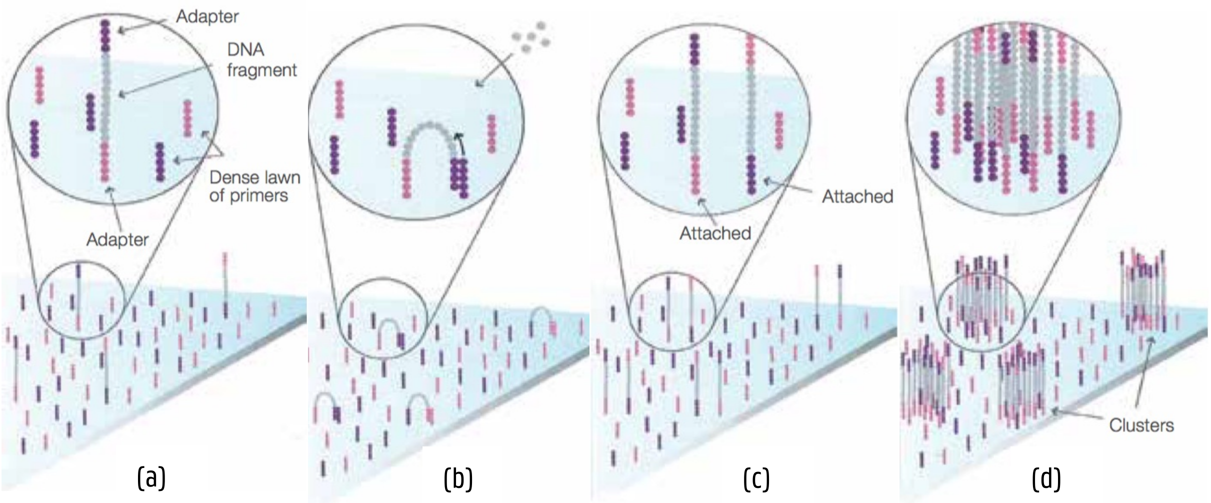

Figure 11: Illumina flow cell during cluster generation. Substrate with oligoes and initially bound DNA fragments (a). Bridge amplification (b). A strand complementary to the original fragment is formed (c). Cluster formation through clonal amplification (d). Source: Technology spotlight Illumina® sequencing.

The region complementary to the second type of oligo at the far end of the DNA fragment will hybridise with its respective oligo, forming a bridge (see Figure 11b). DNA polymerase will copy this strand and after denaturing, a DNA fragment complementary to the original one is created (see Figure 11c). This process is repeated until a cluster is formed (see Figure 11d). Now all reverse strands are stripped from the flow cell leaving behind only the forward strand copies.

Finally, the actual sequencing step can commence. Reads are constructed in cycles that incorporate one terminal fluorescent nucleotide at a time. After each additional nucleotide, a light source will excite fluorescent labels of all the clusters and the emitted wavelengths are registered by a detector. The fluorescent tag will be removed from the nucleotide, allowing further polymerisation in the next cycle. After a certain read length has been reached, reverse strands are again constructed using bridge amplification. The forward strands are washed away and the same procedure is conducted on the reverse strand.

A massive amount of data is created in this process. The index motives that were added to each end of the DNA fragment will now enable computer software to cluster and orient the different reads from the same sample in a pooled sample setup. The short reads are then assembled to rebuild the original full length DNA strands of the sample.

2.7.6.3 Third generation: Nanopore sequencing

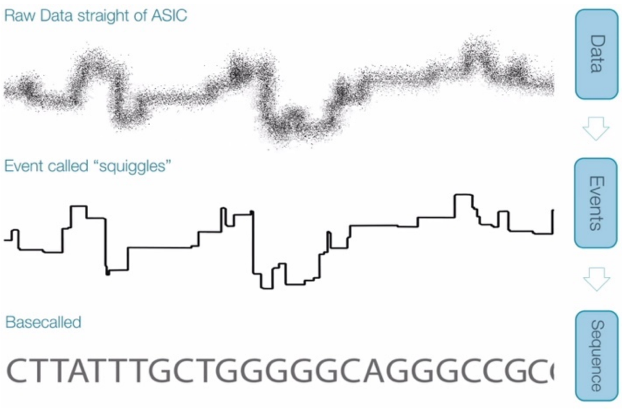

Figure 12: Base calling of ionic current measurement. Source: ONT.

The latest development in sequencing technology is the nanopore approach, by Oxford Nanopore Technologies (ONT). The method is based on characteristic changes in ionic current across a nanoscale aperture when a certain molecule passes through it. The first generation apertures consist of proteins which are embedded in an electrically resistant membrane. An electrical potential across the membrane creates an ionic current that passes through the nanopores. When a molecule is translocated through the hole, the surface area through which ions can migrate decreases, resulting in slight but distinct measurable dips in ionic current. Computer software analyses these dips and translates them into a base call (see Figure 12). The method is scaled and optimised for polynucleotide molecules and can be used for DNA or direct RNA sequencing.

A major advantage of this technology is its scalability. Depending on the application, three formats are available at the moment: MinION™, GridION™ and PromethION™. The sequencer of interest in this study is the MinION™ format. It is a pocket-sized device that weighs less than 100 g and can be connected to any high performance computer with USB 3.0 connectivity. Due to its portability, it is not confined to a laboratory environment and can be taken anywhere. One research team managed to perform a complete barcoding analysis, from DNA extraction to sequencing in a remote rainforest in Tanzania, using a portable sequencing laboratory (Menegon et al. 2017). A similar concept was used for in situ molecular analysis of rare, endangered or undescribed species in the Ecuadorian rainforest (Pomerantz et al. 2017).

Unlike Illumina sequencing, nanopore technology has no predefined read length or experiment duration. Individual strands of DNA are translocated through the nanopore and reads can be analysed in real-time. As a result, the experiment can be run until sufficient data has been generated to answer the research question. The fast acquisition of results has also enabled the real-time monitoring of epidemics and timely implementation of control measures. During the 2015 Ebola outbreak in West Africa, a team transported MinION sequencers and all necessary equipment to the affected areas in regular airline luggage (Quick et al. 2016). They conducted real-time genomic surveillance in remote locations with limited or no access to standard sequencing laboratories. Results could be generated 24h after receiving an Ebola sample, revealing new transmission routes and supporting disease control efforts.

Another advantage is that the read length is only limited by the length of the DNA strand that is presented to the nanopore. Reads of several hundreds of kilobases (kb) in length have been reported (Ip et al. 2015). The development of ultra-long read generating protocols is now paving the way to complete de novo assembly of a human genome, without the need for a reference genome. Assembled genome lengths of 2,867 million bp, covering 85.8% of the human reference genome have been achieved using such protocols (Jain et al. 2018).

Finally, ONT sequencing has the ability to discriminate between 5 different cytosine modifications (Wescoe, Schreiber, and Akeson 2014). This allows direct analysis of different kinds of methylation patterns without the need for bisulphite treatment. Further development of base calling software could potentially revolutionise epigenetic research.

Like with any technology, there are also disadvantages. A high error rate on base calling currently prevents the technology from competing with other existing sequencing platforms in applications that require a high accuracy. Error rates have been found to be as high as 38.2% (Laver et al. 2015). In recent years, improvements in nanopore proteins, library preparation protocols, base calling algorithms and error correction software have significantly lowered this figure. One way to achieve a higher accuracy is to perform a 1D^2 read assay. This particular chemistry allows ONT's new R9.5 nanopore proteins to immediately capture the reverse strand after the forward strand has fully translocated through the nanopore. As every read is sequenced twice, accuracies of over 96% can be achieved.

2.7.6.4 Comparing sequencing platforms

All three sequencing methods are currently being applied. Each has its advantages and pitfalls, which means that the optimal choice is dependent on the research question and setting. For every generation of sequencing methods, different platforms have been developed to cater to the specific needs of many research groups. In the following section, one platform of each method is compared (Table 1). The Applied Biosystems SeqStudio Genetic Analyser (first generation), illumine MiniSeq System (second generation) and the ONT MinION are considered. They are all compact benchtop-sized solutions for laboratories that want to generate and manage their own sequencing data instead of sending samples to an external sequencing lab. Performance parameters were gathered from specification sheets provided by the manufacturers.

Table 1: Performance parameter summary of sequencing platforms

	Read length (bp)	Epi-genetics	Average basecalling accuracy	Real-time data analysis	Direct RNA sequencing	Automated sample loading	Metagenomics	Portable
Seqstudio	Up to 800	No	99.99%	No	Yes	≤ 96 samples	No	No
Miniseq	2 x 150	No	>80% of bp over 99.9%	No	No	≤ 28 for amplicon sequencing	Yes	No
Minion	Limited by strand length	Yes	Up to 96%	Yes	Yes	No	Yes	Yes

2.7.7 Barcoding and environmental studies with nanopore sequencing

Third generation sequencing has only recently been made available for researchers to experiment with. As a consequence, scientific literature on the subject is relatively scarce. Some researchers have investigated the potential nanopore sequencing has for metagenomics and genetic barcoding. As mentioned above, offline *in situ* genetic barcoding has been conducted, which is currently only possible with the ONT MinION. As far as metagenomics is concerned, only studies involving microbial communities have been published (Edwards et al. 2017, 2016). However, at the time of writing this thesis, research on prokaryotic, as well as eukaryotic DNA from scrape samples and environmental DNA (eDNA) taken at the wind mill park in the North Sea is ongoing at Wageningen University by Prof. Dr. Nijland. No papers have been released so far.

The lack of similar research underscores the importance of this thesis. Knowledge gathered from this research could pave the way for a more in-depth development of methodologies and innovation in marine studies.

2.8 Sequence alignment

Sequence alignment is a technology which allows for two or more sequences to be compared and to find matching regions. In the context of metabarcoding, alignment algorithms are used to map query sequences (i.e. sequencing output or reads) to a database containing reference sequences (i.e. annotated DNA sequences of known species). Per query read, several matches can be found, each with its own alignment quality scores. The query read will be registered as originating from the same species as the best matching reference sequence.

A match is scored using a predefined scoring system. For every match, a certain value is added to the score while mismatches, gap openings and gap extensions receive a penalty. These systems can be symmetrical (i.e. the same score or penalty is given to every match or mismatch) or a score matrix can be used which describes a nucleotide dependent score and penalty (Table 6). The algorithm can also take into account sequencing error probabilities reported during base calling (Frith, Wan, and Horton 2010). This is especially important when aligning sequences with a highly variable error rate like nanopore data.

The resulting score on its own doesn't provide much information about the quality of an alignment. A standardised value must be calculated to quantify the significance of an alignment. A commonly used parameter is called the E-value. It represents the probability that a similar or better score can be given to an alignment between two random sequences of the same length as the query and reference sequences. The smaller the value, the more significant a match is. Another important value is the identity or percentage of overlap in an alignment.

3 METHODOLOGY

In this thesis, the possibility of applying third generation sequencing to metagenomics of marine zooplankton communities is investigated. The aim is to develop a protocol that allows for qualitative species level classification without the need for taxonomists or time consuming morphological identification. After a zooplankton community has been sampled, DNA is extracted from it. A barcoding gene is amplified and

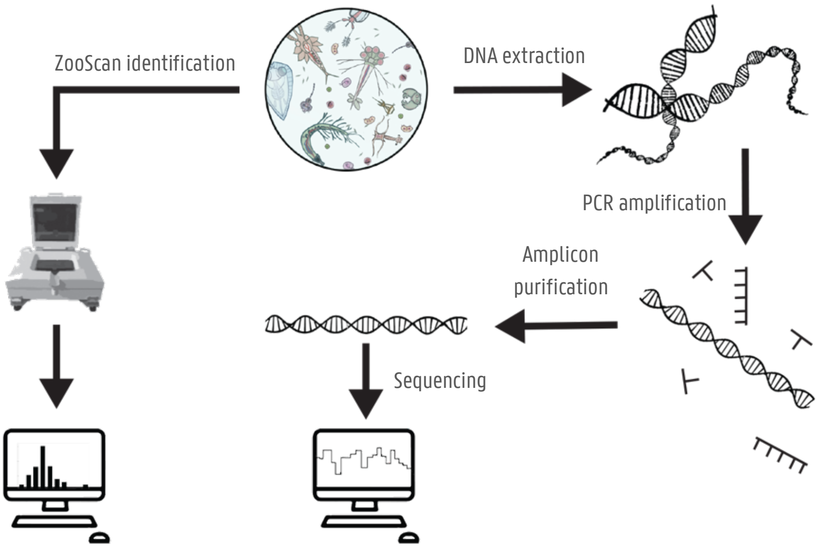

Figure 13: Work flow of DNA extraction and ZooScan.

sequenced with the MinION sequencer. This is followed by data processing which is tailored to the specific characteristics of MinION output data. At the end of the protocol, a list of species that were present in the sample is generated. These results are then compared to those obtained by ZooScan analysis, which is currently being used by LifeWatch. Each stage of the metabarcoding protocol, from DNA extraction to sequencing data processing was first individually evaluated and optimised for marine zooplankton communities in the North Sea. These optimised techniques are then applied to twelve samples taken on a bio-monitoring campaign on the Belgian part of the North Sea (BPNS). The following chapter will elaborate on the methodology used for protocol optimisation.

3.1 Sampling and sample preparation

For the optimisation of DNA extraction and amplification, marine zooplankton tissue samples were used but no metadata was documented. Initially, samples were hand sorted to isolate *Temora sp.* for different research purposes, leaving all other planktonic species for this thesis. Handpicked specimens were collected in an Eppendorf tube. Pipet tips with a small opening (300 µL volume) were used to draw off the residual water, being careful to not entrain any zooplankton specimens. From these individuals, DNA was extracted.

In a second stage, leftover cultures and seawater samples were sieved with a 50 µm mesh. From this filter cake which contained zooplankton, sediment, dead organic matter as well as microplastics, DNA was directly extracted.

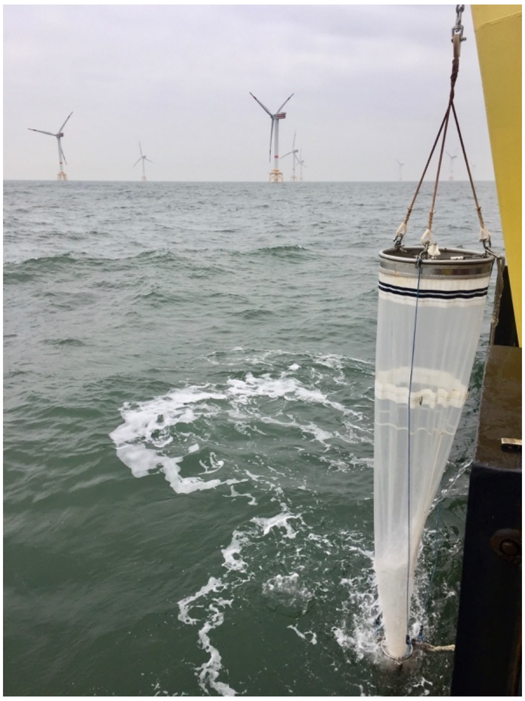

Figure 14: WP2 sampling near C-Power wind mill park on the BPNS.

Finally, after optimisation, samples were taken on the BPNS during a one day cruise (15th of March 2017) on the Flemish research vessel (RV) Simon Stevin. When arriving at the sampling location, a mechanical arm casts the sampling net into the water. Two methods of sampling can be applied. The net can be towed behind the RV for a horizontal sample. For this, CalCoFi nets with a mesh width of 1 mm are used. However, this method has several drawbacks. Horizontal sampling allows only a relatively thin layer of the pelagic zone to be included, which introduces a bias for certain species. Different species live at different depths and many migrate over tens of meters in a diurnal cycle (Roe 1974). Additionally, a mesh width of 1 mm was found to be too large to retain all species of interest. For these reasons vertical sampling, using a WP2 is preferred (see Figure 14). It consists of a ring net, which tapers down into a small collection cylinder at the cod-end where plankton is concentrated. The round opening at the top has a diameter of 57 cm and the mesh width is 200 µm, which is a more desirable size. The sample is taken after the RV has come to a full stop.

At each of the four locations (Figure 15), three samples were taken. From the first cast, 30 mL was put in a Falcon tube and stored for ZooScan analysis. Then, all three samples were sieved with a 150 µm mesh and the filter cake was subsequently divided between two Eppendorf tubes for DNA extraction. The Simon Stevin is equipped with a freezer operating at -16°C that was used to freeze and temporarily store all water and tissue samples for the rest of the campaign. Back on land, the tissue samples were stored at -85°C for extraction the following day. Water samples for the ZooScan were put in a freezer at -20°C.

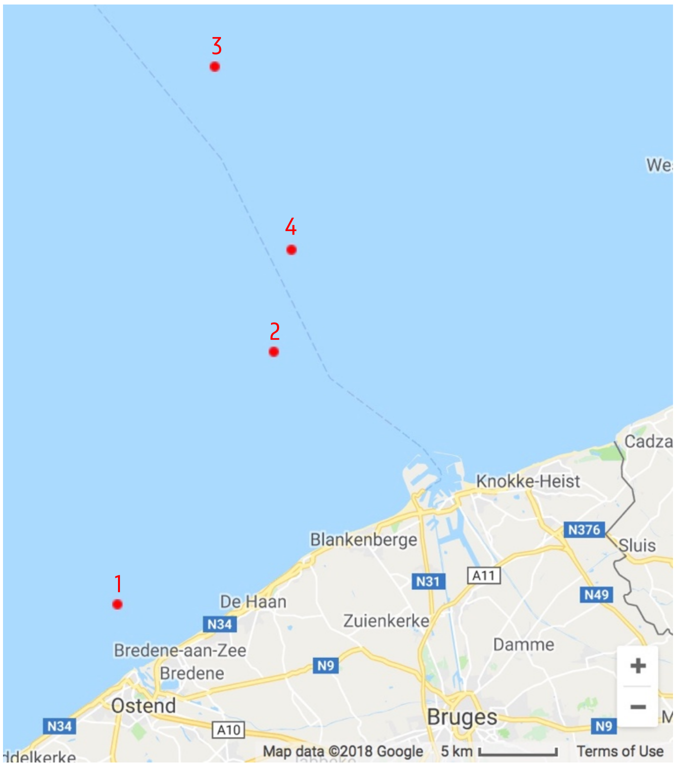

Figure 15: Sampling locations in BPNS

3.2 ZooScan

The ZooScan is a morphology based plankton identifier which scans a thin layer of liquid sample and analyses the image with a machine learning algorithm to name the present taxa.

During the biomonitoring campaign on the BPNS, 30 mL of the zooplankton sample was retained before sieving at each of the four locations. These four samples were later analysed by ZooScan at VLIZ following standard LifeWatch protocol (Gorsky et al. 2010). First, a blank scan is taken, which is to be subtracted from the sample scan. Then a thin layer of the sample is spread out over a flatbed scanner, making sure that plankton specimens do not overlap. Then a high definition image with a resolution of 4800 dpi was uploaded to the computer. A specially designed Java script called Zooprocess (Abramoff, Magalhaes, and Ram 2004) commands Image J to discern the different plankton specimens and store them as individual files in a separate folder (Zooprocess v7.25 at time of analysis). Later a machine learning algorithm called PlanktonIdentifier (v1.3.4) classifies these images into 25 biotic and 6 abiotic (i.e. detritus, fibres, etc.) groups based on a training set it was given. After the algorithm has finished, the data must be validated by manually sieving through the different folders and making sure all specimens are categorised correctly.

3.3 DNA extraction

For DNA extraction, the suitability of two different protocols is evaluated. The first one is the Epicentre® MasterPure™ DNA Purification Kit. The second is a modified CTAB (Cetrimethylammonium bromide) protocol that was optimised by Jana Asselman to be performed in one day. For every DNA extraction procedure, a tissue sample of *Daphnia magna* was included as a positive control.

3.3.1 Epicentre® MasterPure™ DNA Purification Kit

The tissue sample is ground and crushed in 300 µL of Tissue and Cell Lysis Solution containing Proteinase K and incubated at 65°C for 15 min with vortex mixing every 5 min. After cell lysis, the sample is cooled to 37°C and 1 µL of 5 µg/µL RNase A is added. After mixing vigorously and incubating for 30 min at 37°C, the samples are put on ice for 3-5 min before proceeding with DNA precipitation.

To the lysis solution, 175 µL of MPC Protein Precipitation Reagent is added. The mixture is vortex mixed for 10 seconds and subsequently centrifuged at 4°C and 10,000 rcf for 10 min. The supernatant is transferred to a clean autoclaved Eppendorf tube, leaving behind pelleted debris.

Finally, DNA is precipitated by adding 500 µL of isopropanol and centrifuging at 4°C and 10,000 relative centrifugal force (rcf) for 10 min. Without dislodging the DNA pellet, the supernatant is removed and the pellet is rinsed twice with 70% ethanol. If during this process the pellet is dislodged, the sample is briefly centrifuged. After all residual ethanol is removed, the pellet is allowed to air dry until the pellet becomes translucent. Finally, the DNA is resuspended in 35 µL of TE buffer.

3.3.2 Modified CTAB protocol

To a thawed sample, 300 µL of CTAB extraction buffer at 65°C is added. The sample is ground using an autoclaved pestle and incubated for 1 hour at 65°C with vortexing every 20 min. During this step the cells will rupture as CTAB binds to the cell membrane, releasing their content. Through consecutive purification steps, DNA is separated from remaining cell debris, proteins and other contaminants.

In this study, purified DNA is obtained using a phenol-chloroform extraction protocol. First 300 µL of phenol:choroform:isoamyl alcohol (25:24:1) is added, mixed and centrifuged for 20 minutes at 15,000 rcf and 4°C. The organic phase will dissolve proteins and hydrophobic lipids while leaving both RNA and DNA in the aqueous layer (Kirby 1964). Cell and other debris will collect at the interface. The supernatant is recovered in a clean 1.5 mL microcentrifuge tube. To reduce RNA contamination, 2 µL of 5 µg/µL of RNase A is added and the mixture incubated for 20 min at room temperature.

To this solution 500 µL of chloroform:isoamyl alcohol (24:1) is added to separate additional contaminants. The liquids are mixed and centrifuged at 15,000 rcf for 15 min. If the supernatant still contains noticeable colouring, this step is repeated until the supernatant appears colourless. Only then can the protocol be continued.

The aqueous layer is transferred to a clean microcentrifuge tube and 27µL of 3M Na acetate and 500 µL of 2-propanol is added. The solution is gently mixed by inverting and then incubated for 10 min at room temperature. DNA will start to precipitate and will subsequently be pelleted by spinning the tube at 4°C, 10,000 rcf for 15 min. The liquid is removed leaving a purified DNA pellet.

The last step is to wash the pellet with 500 µL of 70% molecular grade ethanol (ethOH), spin for 15 min at 4°C and 10,000 rcf and discard the liquid phase. This procedure is repeated in two consecutive iterations. The pellet is allowed to air dry until it becomes translucent and is subsequently resuspended in 50 µL of TE buffer. The result is a high molecular weight DNA extract.

3.4 PCR amplification

After DNA is extracted from the zooplankton community, a barcoding gene of choice has to be amplified. For this part of the protocol, two sets of primers are evaluated. The first is the mlCOIintF/jgHCO2198 combination after Leray et al. 2013. It amplifies a section of around 300bp from the mitochondrial COI gene. The second pair is the F-566/R-1200 set designed by (Hadziavdic et al. 2014), which amplifies a ~650 bp region of the 18S rRNA gene.

Both sets were tested with ThermoFisher Scientific Maxima Hot Start PCR Master Mix on both handpicked and sieved samples, as well as on *Daphnia magna* tissue as a positive control.

Table 2: PCR primer sets

Genetic marker	Primer set	Sequence (5' - 3')	T_m (°C)	Reference
COI gene	mlCOIintF	GGWACWGGWTGAACWGTWTAYCCYCC	54	(Leray et al. 2013)
	jgHCO2198	TAIACYTCIGGRTGICCRAARAAYCA	51	
18S rRNA gene	F-566	CAG CAG CCG CGG TAA TTC C	50	(Hadziavdic et al. 2014)
	R-1200	CCC GTG TTG AGT CAA ATT AAG C	48	

As PCR is considerably sensitive to inhibition, three master mixes were evaluated to find the most optimal mastermix – primer combination and the best conditions for this specific analysis. Additionally, several methods to mitigate the effect of PIC were investigated.

3.4.1 Maxima™ Hot Start PCR Master Mix

Each PCR reaction tube is prepared with 25µL of ThermoFisher Scientific Maxima Hot Start PCR Master Mix (2X), 5 µL of 5 µM forward and reverse primer stock solution, 2 µL of 1/20 diluted high molecular weight DNA template and 13 µL of DNase free water to a final volume of 50 µL.

For the 18S amplification, the thermal cycler is set to 95°C for 4 minutes to activate DNA polymerase. The actual PCR reaction consists of 35 cycles of melting at 95 °C for 30 s, annealing at 44 °C for 30 s and extension at 72 °C for 1 min, followed by a 10 min final extension at 72 °C. The melting temperature is calculated according to the manufacturer's recommendation as the average of the two melting temperatures minus 5°C.

The same program is used during the first test of the COI amplification, only the annealing temperature was set to 47 °C since the primer set has a different melting temperature. This resulted in only a faint band after gel electrophoresis, which prompted a second attempt with the following protocol taken from the original publication (Leray et al. 2013). Since a Hot Start master mix was used, the initial DNA polymerase activation step of 95 °C for 4 min was kept. The actual PCR reaction consists of 16 cycles of melting at 95 °C for 10 s, annealing at 62 °C for 30 s from - which 1°C is subtracted every cycle - and extension at 72 °C for 1 min. Finally, 25 cycles were performed with the same settings, except for annealing temperature which was fixed at 46 °C. This program resulted in no amplification at all. After this, no further iterations for the COI primer set were attempted.

3.4.2 DreamTaq™ Hot Start PCR Master Mix

The protocol is kept largely the same, conforming to the manufacturer's instructions. Only the cycler program differs in two ways. Firstly, the initial DNA polymerase activation step is set to 95°C for 3 min and secondly, the melting temperature does not need to be adjusted to facilitate annealing. This equates to an annealing temperature of 49°C for the 18S primer set and 52°C for COI.

3.4.3 AmpliTaq Gold® Fast PCR Master Mix

Each PCR reaction tube is prepared with 10µL of ThermoFisher Scientific AmpliTaq Gold® Fast PCR Master Mix UP, 2 µL of 5 µM forward and reverse primer work solution, 1 µL of 1:20 diluted high molecular weight DNA template and 5 µL of DNase free water to a final volume of 50 µL.

The thermal cycler is programmed according to the manufacturer's instructions. The protocol starts with an initial DNA polymerase activation step at 95°C for 10 min, followed by 35 cycles of melting at 96°C for 3 s, annealing at 49°C for 3 s and extension at 68°C for 15 s. A 15 s extension duration is selected as recommended for an amplicon length of 1 kb. The program concludes with a final extension step at 72°C for 10 s.

3.4.4 Mitigating PCR inhibition

One of the easiest ways to avoid inhibition is to dilute the DNA extract until the concentration of PIC is too low to have any effect on the PCR reaction. In this study, a series of 20, 30, 40, 60 and 80 fold dilution was made for samples 5 and 19 of the biomonitoring campaign.

For certain samples (14, 15 and 23) the 1:80 dilution did not suffice. In these cases, an extra purification step was introduced to remove the offensive organic contaminants that may have led to PCR inhibition. From the DNA extract, 5 µL is taken and diluted in 395 µL of water. To this 500 µL of 24:1 chloroform:IAA is added and the mix is vigorously mixed by vortex. The organic and water phase is separated by spinning at 15,000 rcf and 4°C for 15 min. The supernatant is transferred to a clean Eppendorf tube after which 27 µL of 3 M Na Acetate and 500 µL of 2-propanol is added. The samples are inverted to mix and incubated at room temperature for 10 min to precipitate the DNA. The DNA is then pelleted by spinning at 4°C and 10,000 rfc for 15 min. The supernatant is completely removed and 2 consecutive EtOH wash steps are introduced by adding 500 µL of 70% EtOH, spinning down the pellet at 4°C and 10,000 rcf for 15 min and removing the supernatant. After the pellet has been allowed to dry, it is resuspended in 100 µL of nuclease-free water. Note that initially 5 µL of DNA extract was taken so this protocol results in a 1:20 dilution. From this purified dilution, 5 µL is dissolved in 15 µL of nuclease-free water to come to a final dilution of 1:80. This is only an approximation since some DNA is lost during the purification step.

According to a paper which characterised and tested the 18S F-566/R-1200 primer set on seafloor samples from the North Sea, BSA the addition of 1 µg/µL of BSA reduced the effects of PIC originating from the sediment (Hadziavdic et al. 2014). Since the bio-monitoring samples contain a large amount of sediment, this approach is also tested. The three samples were amplified in duplicate, once with only treatment described above and once with the addition of 5 µL of 10 µg/µL BSA stock (50 µL total PCR volume).

3.5 Gel electrophoresis

Gel electrophoresis is performed in order to see if the PCR reaction was successful and produced the desired length of amplicons. Before starting with gel preparation, make sure the 60 mL tray is securely clamped into its holder, with the comb inserted. The gel is prepared by adding 0.9 g agarose to 60 mL TBE buffer and heating the mix in a microwave until the solution becomes homogeneous and transparent. When the gel has cooled to about 60°C, 6 µL of GelRed is added and after gently swirling the mixture, the gel was poured into the tray. The result is a 1.5% gel with 15 wells.

Amplicon length is measured by injecting 6 µL of ThermoFisher Scientific GeneRuler 50 bp DNA Ladder into the first well. After separation, this produces a DNA ladder of 50 - 1000 bp. This is a suitable range since the COI and 18S amplicons are ~300 and ~650 bp long respectively.

The samples are prepared by mixing 1 µL of ThermoFisher Scientific DNA gel loading dye (6X), 2 µL of PCR product and 3 µL of nuclease-free ultrapure water. Mixing is done by pipetting up and down.

The gel is put in a TBE buffer bath and an electric potential of 50V is applied for at least 1 hour to separate the different lengths of DNA strands. Figure 16 shows some of the gel's attributes.

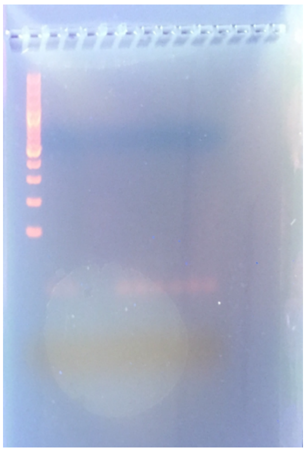

Figure 16: Example of a gel with 15 wells at the top, upper (blue) and lower (yellow) bands of loading dye and DNA ladder fluorescing under UV light (left).

3.6 PCR product purification

Before a PCR product can be applied in downstream amplifications and sequencing, interfering components such as leftover primers and deoxyribose nucleoside triphosphates (dNTP) must first be removed. For PCR purification four protocols are tested. Two involve the use of GE Healthcare illustra GFX PCR DNA and Gel Band Purification Kit, one consists of an enzymatic reaction with the ThermoFisher Scientific CleanSweep™ PCR Purification reagent and the last one makes use of CleanPCR magnetic silica beads from CleanNA.

3.6.1 Gel band extraction with illustra GFX kit

First PCR product is separated with gel electrophoresis. Using a scalpel, the amplicon bands are then cut out of the gel under a UV light. DNA is subsequently extracted from these gel cut-outs using the GE Healthcare illustra GFX PCR DNA and Gel Band Purification Kit.

Firstly, all samples are weighed in Eppendorf tubes. For each 10 mg of gel, 10 µL capture buffer type 3 is added. However, if the gel weighs less than 300 mg, still 300 µL is used. The samples are allowed to incubate at 60°C for 15-30 min with mixing by inversion every 3 min until the gel has completely dissolved into the liquid. Capture buffer type 3 contains a pH indicator that will turn dark pink or red when the pH is not optimal. If it does, a small volume (10 µL) of 3 M sodium acetate pH 5 is added to the mix until the colour turns back to yellow or pale orange. Once the pH is adjusted to the preferred range, the mix is transferred into an assembled GFX MicroSpin column and collection tube. After one minute of incubation at room temperature, the tubes are spun at 16000 rcf for 30 s. The flow-through is discarded and the MicroSpin column placed back into the collection tube. Then 500 µL of buffer type 1 is added to the column, which is subsequently spun at 16,000 rcf for 30 s. After washing, the column is transferred into a clean autoclaved Eppendorf tube. During the elution step, elusion buffer type 6 is preferred because it is better suited for downstream applications like genome sequencing. For optimal recovery, a volume of 50 µL is added to the tube. After a 1 min incubation period at room temperature, the columns are centrifuged at 16,000 rcf for 1 min.

3.6.2 PCR DNA purification with illustra GFX kit

The GE Healthcare illustra GFX PCR DNA and Gel Band Purification Kit can also be directly applied to PCR product. The main protocol is largely the same as with gel extraction. Only during the first step, a fixed volume (500 µL) of capture buffer type 3 is added to 10 µL of PCR product. However, in order to find the optimal protocol for the experimental setup of this study, a few different iterations are tested (see Table 3).

Table 3: Iterations of PCR product purification protocol

Sample label	S0	S15	S30	D0	D15	D30	B1	B2
Number of wash steps	1	1	1	2	2	2	2	2
Air dry after wash step (min)	0	15	30	0	15	30	0	0
Elution buffer type	6	6	6	6	6	6	6	4

The parameters that are tested are number of washing steps (1 or 2), minutes of air drying after washing to allow evaporation of residual ethanol and the elution buffer used. Two elution buffers are provided with the kit. The first consists of 10 mM Tris-HCl, pH 8.0 (elution buffer type 4) and is recommended for multiple downstream applications and long term storage. The second is sterile nuclease-free water (type 6), which is recommended only for samples to be sequenced.

3.6.3 Enzymatic PCR DNA Purification

The third protocol that is tested is the ThermoFisher Scientific CleanSweep™ PCR Purification reagent. It consists of a mixture of enzymes which hydrolyse single stranded DNA primers and dephosphorylate excess dNTP. This ensures that they will not interfere with downstream sequencing. This method is tested on amplicons from two handpicked copepods samples and two sieved samples.

To 30 µL of PCR product, 12 µL of CleanSweep™ reagent is added and mixed by pipetting up and down. Using a thermal cycler, the mixture is incubated at 37°C for 15 min to facilitate the enzymatic reaction, followed by a 15 min incubation period at 80°C to inactivate the reagent. The sample should now be ready for sequencing.

3.6.4 CleanPCR magnetic bead separation

The last protocol for PCR clean-up made use of CleanPCR silica coated magnetic beads from CleanNA. These are suspended in a buffer which causes double stranded DNA (dsDNA) to bind to the silica. To one volume of PCR product, 1.8 volumes of suspended beads are added. After an incubation period of 5 min at room temperature, the beads are pelleted on a magnetic rack until the solution is clear. By aspiring off the supernatant, amplicons which are bound to the pelleted beads are separated from leftover primers and dNTPs. The pellet is then washed twice by adding 200 µL of 70% ethanol, incubating for one minute and discarding the liquid. The pellet is then allowed to air dry for about 10-15 minutes, being careful not to let it dry to the point where it starts to crack. This can impede elution in the following step. When the pellet is dry, it is resuspended in 40 µL of nuclease free water and incubated for 3 minutes at room temperature to eluate dsDNA from the beads. Finally, the beads are pelleted again on a magnetic rack and the amplicon containing eluate is recovered.

3.7 Qubit™ assay

For dsDNA and RNA quantification, the Qubit 2.0 fluorometer is used in combination with broad range (BR) solutions. First, a working solution is prepared. For every sample and calibration solution, 1 µL of Qubit™ dsDNA or RNA BR reagent and 199 µL of Qubit™ dsDNA or RNA BR buffer are added. To calibrate the fluorometer, two standard solutions are prepared by adding 10 µL of standard to 190 µL of working solution. Next 1-20 µL of sample is diluted with working solution to a final volume of 200 µL. To perform a dsDNA assay, 5 µL of PCR product or 1 µL of DNA extract are taken. For RNA quantification, 10 µL of PCR product or 1 µL of DNA extract were required. Then all samples and standard solutions are mixed by vortex for 2-3 seconds and subsequently incubated for 2 min at room temperature. To avoid skewing results due to nucleotide stain degradation, the samples are protected from light during incubation. After sample preparation, the tubes are analysed with the Qubit 2.0 fluorometer.

3.8 *In silico* PCR

The 18S primer set is mapped to the SILVA database, which contains rRNA sequences from all branches of the tree of life. This server has several tools to explore and collect data. One of these tools is called TestPrime 1.0. It has been used by Hadziavdic et al. 2014 to develop and test universal 18S barcoding primer sets for eukaryotes.

Both forward and reverse primers are submitted to be mapped to the SSU-132 database since 18S rRNA is part of the small subunit (SSU) of a eukaryotic ribosome. Then the refNR collection is selected. This dataset is derived from the ref collection by clustering different accession numbers of the same taxon. Therefore, the phylogenetic tree is more uniformly represented. For alignment, a single nucleotide mismatch was allowed for a more realistic simulation of PCR behaviour (Chandler, Fredrickson, and Brockman 1997).

3.9 DNA sequencing

Before starting with the ONT library preparation protocol, DNA must be quantified and the quality recorded. The aim is to recover 0.2 pmoles at the end of the library preparation. As loss of DNA is inherent to this type of protocol, ONT recommends to start with a mass of about 1-1.5 µg of input DNA with a mean length of 3kb. However, since the length of amplicons here is around 650 bp, the input mass can be slightly lower. A recovery of 0.2 pmoles corresponds to ~260 ng of prepped DNA. As far as quality is concerned, NanoDrop values for input DNA should be equal to an OD 260/280 of 1.8 and OD 260/230 of 2.0-2.2.

For these experiments, the SQK-LSK108 1D sample library preparation chemistry is used in combination with a FLO-MIN 106 R9.4 MinION flow cell.

3.9.1 Quality control flow cell

The flow cell platform quality control (QC) should be performed before starting the library prep to avoid losing the sample if flow cell performance is unsatisfactory. Oxford Nanopore Technologies guarantees 800 functioning nanopores upon arrival, but this warranty is voided if QC was not done within 10 days of receiving the flow cell.

To perform a QC, simply mount the flow cell on the MinION device and connect it to a computer. If the flow cell was just removed from cold storage, then wait about 10 min to allow it to reach room temperature. Open the MinKNOW GUI and follow the instructions on the screen. Select Platform QC under 'Choose Operation' and press execute. When the QC is finished, the flow cell can be put back in the fridge for storage.

3.9.2 1D amplicon by ligation SQK-LSK108

After the flow cell QC has been performed and the sample properly processed, everything is ready to start the library preparation protocol. Oxford Nanopore Technologies has produced several library prep kits for various different applications. For this study, the ONT 1D amplicon by ligation SQK-LSK108 kit was selected. Figure 17 shows an overview of the workflow for high molecular weight genomic DNA (gDNA). Although in this case amplicons are used, the principle remains the same.

First, the input DNA are prepared for adapter ligation by first blunting the ends and attaching a dA-tail to the 3'-ends. This will prevent concatenation of amplicons during ligation while at the same time assisting adapter complex attachment by hybridisation with the T-overhang of the adapter sequence. The dA-tails thereby assure that adapters are correctly oriented at both ends of the DNA fragment. The adapter complex contains a motor protein which will dock onto a nanopore protein and facilitate the translocation of a DNA strand through the pore.

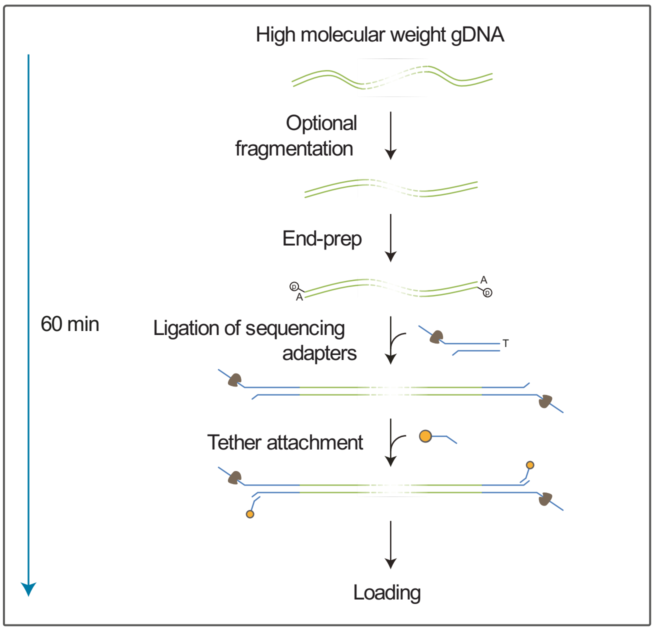

Figure 17: Workflow of ONT 1D amplicon by ligation SQK-LSK108 kit. Source: ONT

When these adapters are successfully ligated to the sample DNA, it is purified by magnetic bead separation (see below) and eluted with the ONT Elution Buffer. This solution will keep the adapter proteins stable and also contain a tether molecule which hybridises to the adapter's bottom strand. This tether will attach to the nanopore membrane, hereby confining the molecule's movement and increasing DNA capture by approximately a thousand fold.

In the ONT protocol it is recommended to use AMPure XP beads for sample purification. However, in this research, the significantly cheaper CleanNGS magnetic beads from CleanNA were preferred. Multiple researchers have reported good results with this alteration in the protocol.

3.9.2.1 End-prep

During the first step, the amplicon ends are prepared for adapter ligation. To 45 µL of input DNA (~1 µg), 7 µL of NEBNext Ultra II End-prep reaction buffer, 3 µL NEBNext Ultra II End-prep enzyme mix and 5 µL of nuclease-free water are added. The reagents are gently mixed by flicking the tube and spinning down the liquid. The mix is transferred to a 0.2 mL PCR tube and incubated in a thermal cycler for 5 min at 20°C, followed by 5 min at 65°C. If condensation on the tube wall is observed after incubation, the liquid is spun down in a microfuge.

The DNA is separated from the liquid using CleanNGS beads from CleanNA. The sample is now transferred into a clean 1.5 ml LoBind tube. Then 60 µL of resuspended CleanNGS beads are added and the mix homogenised by pipetting. The DNA is allowed to bind to the beads for 5 min on a rotating Hula mixer to prevent beads from settling. After incubation, the sample is placed on a magnetic rack to pellet the beads. Only when the solution has completely cleared and all of the beads have collected at the side of the tube, the supernatant is aspired off and discarded. Without disturbing the pellet, 200 µL of freshly made 70% ethanol solution is added to the tube to wash the pellet. After a few moments, the wash solution is removed and this wash step is repeated. The tube is spun briefly to collect all leftover ethanol at the bottom, after which it is removed while keeping the tube on a magnetic rack. Residual ethanol will disrupt the following enzymatic reactions, so the pellet should be allowed to dry sufficiently (~30 s). However, it is important that the pellet does not start cracking since this will hamper the resuspension and elution of the beads, resulting in a loss of sample. The pellet is resuspended in 31 µL of nuclease-free water and incubated at room temperature for 2 min. The DNA mass is quantified by Qubit using 1 µL of solution. Approximately 0.33 pmoles (~430 ng for 650 bp) should remain in the end-prepped sample.

3.9.2.2 Adapter ligation

The second step is adapter ligation. To the 30 µL of the end-prepped sample from the previous step, first 20 µL of ONT adapter mix 1D and then 50 µL of Blunt/TA ligation master mix are added with mixing by flicking in between every consecutive step. The total volume of 100 µL mix is collected at the bottom by spinning the tube. Then the sample is incubated for 10 min at room temperature.

3.9.2.3 Beads purification

During beads purification, all excess proteins, nucleotides and salts need to be removed from the library. This is done by another purification step using magnetic beads.

To the sample, 40 µL of resuspended beads are added and the mix is homogenised by pipetting up and down. The sample is incubated on a rotating Hula mixer for 5 min to allow the DNA to bind to the beads. After incubation, the beads are pelleted on a magnetic rack and the supernatant aspired off. The pellet is washed by adding 140 µL of ABB buffer, resuspending the beads by flicking the tube, to then magnetically pellet them again and removing the supernatant. This wash step is performed twice before the DNA is eluted by incubating the beads in 15 µL of ELB elution buffer for 10 min. Finally, the beads are magnetically pelleted and the eluate transferred to a clean LoBind tube. The used beads can be discarded.

As mentioned before, the aim is to recover 0.2 pmoles of DNA which in this case corresponds to ~260 ng as measured by Qubit.

3.9.2.4 Priming and loading the SpotON flow cell

During priming, the storage buffer in the flow cell is flushed with priming buffer. This contains the chemistry that will allow sequencing. First, the priming port cover is slid clock-wise to reveal the priming port. A P1000 pipette is set to 200 µL and inserted into the priming port. By turning the wheel until the dial shows 220-230 µL a small volume of buffer is drawn into the pipette as well as any air bubbles that are present near the priming port.

The priming buffer is prepared by mixing 576 µL of Running Buffer with Fuel mix (RBF) with 624 µL of nuclease-free water. Through the priming port, 800 µL of priming buffer is loaded into the flow cell. By incubating for 5 minutes any remaining storage buffer is allowed to diffuse out of the nanopore array and into the priming buffer. In the meantime, loading beads are added to the DNA library by mixing 35 µL of RBF, 25.5 µL of Library Loading Beads (LLB), 2.5 µL of nuclease-free water and 12 µL of DNA library. Flow cell priming is completed by first gently lifting the SpotOn port cover and then loading 200 µL of priming buffer into the priming port (not the SpotOn port). Then the 75 µL sample can be loaded onto the SpotOn port dropwise, making sure that each drop has completely entered the flow cell before adding another. Finally, after both the priming port and the SpotOn cover have been put back in place, the sequencing run can be initiated with the MinKNOW GUI.

Throughout the entire process of priming and loading the flow cell, the introduction of air bubbles should be avoided as this will damage the nanopore array.

3.9.2.5 Washing the flow cell

The flow cell can be washed and used again for a subsequent sequencing run. This is done by opening the priming port cover and adding 150 µL of Solution A from the ONT wash kit and waiting for 10 minutes. For immediate reuse, 150 µL of Solution B can be added, but this does not allow for a new platform QC. Only the storage solution S contains the short validation sequences that MinKNOW uses to assess pore quality. Therefore, 150 µL of Solution S is not only added before storage or return to ONT but also in between consecutive runs.

3.10 Data processing

During a sequencing run, raw data is recorded in the form of conductivity measurements. Every passing of a DNA strand through a pore is stored in a separated fast5 file. This file contains raw data and metadata such as the read ID, the sequencing run ID, start time of record, channel number, etc.

The first step in data processing is to assign a nucleotide sequence to these conductivity measurements, i.e. base calling. This can be done either live in MinKNOW during sequencing or afterwards with the ONT Albacore base calling tool. In this study, the last option was chosen. When a fast5 file is base called, a Phred quality score for each nucleotide is calculated as

Qscore = -10 x log(Pe)

with Pe the estimated probability of an erroneous base call. Read sequences and their per base Phred score are stored in a copy of the original fast5 file. One can choose to only store reads and some essential metadata in a separate fastq file or opt for both fast5 and fastq output formats.

Once all data has been base called, the quality of the dataset can be assessed. Several third party tools are available and have been tested here.

If data quality is sufficient for downstream processing, adapter sequences are trimmed from the reads. This leaves a fastq or fasta file containing only the input amplicon sequences. These are now ready for mapping and OTU determination.

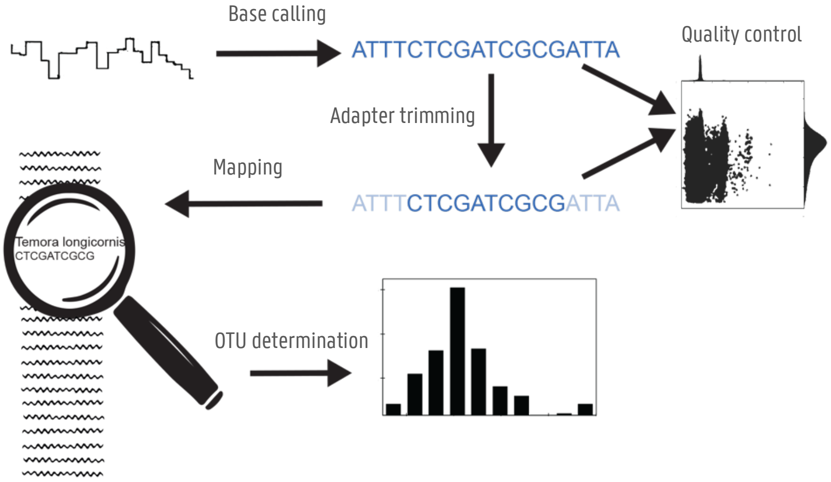

Figure 18: Work flow of data processing from base calling to OTU determination.

3.10.1 Albacore base calling

MinION reads are base called with the command line based Albacore algorithm from ONT. At the time of analysis, the most recent version available was ONT Albacore Sequencing Pipeline Software (version 2.2.7). Base called reads are stored in copies of the original fast5 file as well as in one single fastq file. The computer on which Albacore is run can handle up to 8 worker threads. However, seven were chosen here to leave one thread for reading and writing data, making the workflow more efficient. These parameter preferences result in the following command line:

```
read_fast5_basecaller.py    --input    {path\to\fast5}    --recursive    --save_path
{path\to\Albacore\workspace} --output_format fast5,fastq --reads_per_fastq_batch 0 -
-flowcell FLO-MIN106 --kit SQK-LSK108 --worker_threads 7
```

3.10.2 Read quality control

Before continuing with downstream data processing, the quality of a sequencing run must first be assessed. A variety of third party tools have been developed for this. Four of these tools have been evaluated for their functionality, user-friendliness, efficiency and customisability of resulting graphs. All tools and supporting documentation can be downloaded from Github and self-written scripts can be found in the annex.

3.10.2.1 poRe

PoRe is an R based package which features a graphic user interface (GUI) that enables the extraction of metadata from a set of fast5 files (M. Watson et al. 2015). This algorithm is run after base calling with the poRe parallel GUI (version 0.21). After metadata has been extracted, graphs can be plotted in the GUI or in R by either writing custom code or by making use of functions included in the poRe R package.

3.10.2.2 Poretools

Poretools (currently most recent version 0.6.0) is a command line based python tool that produces graphs based on the sequencing summary file generated during Albacore base calling (Loman and Quinlan 2014). This means that no additional metadata extraction needs to be performed. In the package, functions are included for plotting a cumulative yield, read length histogram, channel usage, mean Phred quality score, etc. It also allows to plot quality score distribution over position in reads and to perform squiggle analyses. This last data, however, is not stored in the sequencing summary and must be extracted from a set of fast5 files, which is computationally intensive.

3.10.2.3 NanoPlot

Similar to Poretools, NanoPlot (currently most recent version 1.13.0) is a command line based tool that will plot from the sequencing summary. It consists of just one command that will generate a read length histogram, bivariate plot of length over mean quality, reads per channel, cumulative yield and violin plots of read length and quality.

3.10.2.4 PycoQC

PycoQC (currently most recent version 1.1a1) is a python 3 module that also takes metadata from the sequencing summary. It can be imported into an integrated development environment (IDE) such as Eclipse (here Eclipse Oxygen.3a). The advantage of this is that this tool can be used in conjunction with other python modules such as os, seaborn or matlibplot for directory organisation or plot customisation, respectively. This way a batch of datasets can be processed in single python script run.

3.10.2.5 Phred quality over base position in read

None of the above-mentioned tools are able to plot the per base Phred score as a function of position in read, directly from a FASTQ file. Instead, they extract the necessary data from a set of FAST5 files. This means that this type of plot cannot be made after adapter trimming for comparison with the initial data. To resolve this, a python script has been developed to do just that (qualpos.py in annex). From every read in the FASTQ dataset, the Phred score of each base is taken and an average is plotted as a function of position in read (blue line in Figure 29). To get an idea of the variance on these values, a boxplot is constructed every 50th position.

3.10.3 Adapter trimming

The last step before species identification is adapter trimming. During library preparation, essential motor protein holding adapters are ligated to the ends of every template. These adapter sequences will interfere with downstream data processing and must therefore be removed first. For this, the third party tool PoreChop (here version 0.2.3) was developed. It aligns a subset of 10,000 reads with a database of known adapter sequences. Once a specific set of adapter sequences has been identified, they are removed from the ends of the read. By default it will also search for adapter sequences in the middle of reads to split chimeric sequences that could have been created during ligation. If the split reads are smaller than 1000 bp, they will by default be omitted from the output file. Since the DNA fragments here are around 650 bp long, this value was set to 550. The following command line was used to start adapter trimming:

```
porechop -i {path/to/fastq} -o {MyTrimmed.fastq} --min_split_read_size 550
```

3.10.4 Mapping and OTU determination

3.10.4.1 Building a reference database

Before mapping can be performed, a database containing all relevant sequences must first be constructed. For this study, a database of 18S sequences from marine zooplankton species was assembled using the National Center for Biotechnology Information (NCBI) GenBank, World Register of Marine Species (WoRMS) and LifeWatch Taxonomic Backbone databases. First a list of relevant taxa was compiled by consulting the WoRMS database and papers mentioned in section 2.5.2. Then the taxon ID of each of these groups was searched on NCBI. Next the NCBI nucleotide database was searched for records containing both the 18S gene and one of these taxon ID's (e.g. syntax "18S rRNA AND txid6657[organism:exp]" will search for 18S of Crustaceans). To avoid lengthy full chromosome sequence records an additional filter was applied restricting hits to a length of 2500 bp. This value was chosen because it corresponds to the upper limit of 18S sequence lengths (KWON, Ogino, and Ishikawa 1991). As the Nematode file contains almost 20 000 records, of which many terrestrial or freshwater species, a subset is selected. Species hits from the NCBI nucleotide database search, are uploaded to the LifeWatch Taxonomic Backbone website and only those that matched with the WoRMS marine database are included in the OTU mapping reference file. When this reference is complete, an index file must be made which links record accession numbers to their respective species names (make_index.py in annex). Finally, reference sequence lengths are plotted in a density distribution graph with python (analyse_DB.py in annex).

3.10.4.2 LAST alignment

Many different alignment algorithms have been developed, each with their own unique characteristics. In the past, LAST has often been the aligner of choice for MinION data, because of its ability to cope with long reads containing many errors (Laver et al. 2015; Quick, Quinlan, and Loman 2014). Independent developers and research scientists implement these conventional alignment algorithms in their own code to help answer their research questions. NanoOK and Streamformatics are two examples of such third party tools that have been made available on GitHub by independent developers. However, these scripts are often unstable and difficult to use on data or for research purposes it was not specifically designed for. Mapping was first attempted with both NanoOK and Streamformatics but a week of debugging did not lead to a successful installation or implementation of the software. Instead, custom code (last_select.py and last_analyse.py in annex) was written to process and analyse LAST output.

For LAST to work, it must first create its own indexed database from a given FASTA file. In the package a tool called lastdb is included which facilitates indexation of the reference database discussed in the previous paragraph. In the command line, Q0 indicates that the input database has a FASTA format and w3 forces initial matches to start at every third position of the sequence. This was done to increase speed and lower memory usage. Due to the datasets' containing long query reads, this does not significantly compromise alignment accuracy. This corresponds to the following command line:

```
lastdb -Q 0 -w3 {MyDB} {MyDB.fasta}
```

One last thing that needs to be considered before lastal can be applied is the scoring system. During mapping, the algorithm will score each alignment based on these settings. Points are added to the score when two nucleotides match and a penalty or cost is applied for mismatches or when a gap has to be created in the sequence. Errors can occur during PCR amplification or during sequencing so each experiment is unique in the errors it may introduce to the sequence. Therefore, the scoring system must be adjusted accordingly. Nanopore data for example is known to be more prone to indels than other sequencing methods (Mikheyev and Tin 2014). To maximise alignment accuracy, the gap creation penalty should be lower than the default value for Illumina data. For this reason, LAST has developed a tool called LAST-train which infers a scoring system from sample read data by iteratively fine tuning substitution, insertion and deletion parameters (Hamada et al. 2017) until a sufficient align quality has been reached. In the paper of Hamada et al. 2017, a training set of 1-10 million bases was sufficient to find a high fidelity scoring matrix. Here a subset of reads with a total of about 10 million bases was created in a FASTA file and processed using the following command:

```
last-train {MyDB} {train_set.fasta}
```

Now LAST alignment can be performed with the lastal tool. The alignment was performed first with the scoring matrix calculated by last-train (-p score.txt) and then repeated with default scoring settings to assess the difference. The command line for default alignment is:

```
lastal -T1 -Q1 -P5 -f BlastTab {path/to/barcodeDB} {path/to/reads} | python
last_select.py > {MyData.txt}
```

In both cases, four other arguments were given. Firstly -T1 to try overlap alignment instead of default local alignment. Secondly, -Q1 to indicate that the query reads are in Sanger-FASTQ format, which contains error probability values per base. LAST will take these values into account to improve alignment. The computer on which the algorithm is run has 6 CPU cores. Entering P5 as the third argument, will designate 5 CPU cores to this algorithm, leaving one to read and write data. Finally, '-f BlastTab' tells lastal to write output data in BLAST tab format. This output data stream is then piped into a custom python script (last_select.py in annex) which will select the best match for every query sequence, based on E-values. Then the OTU is added by searching the subject sequence's accession number in the index file discussed in the previous paragraph (make_index.py). Also, the number of hits the query sequence got are reported. This output is then written to a file in a given directory. The resulting command line will be discussed in paragraph 4.6.4.

3.10.4.3 Data analysis

Data analysis is performed on these output files using a custom python script (last_analyse.py in annex). For every unique species name, the number of corresponding query sequences is extracted and exported to a separate file. Then a bivariate plot of E-value as a function of alignment length is made. A density function along the parameters' respective axes shows the distribution of alignment lengths and E-values in the matched reads dataset. These plots are used to compare the default alignments to alignments made with a score matrix tailored to nanopore data. An additional function is built into the script which allows for optional pruning of the dataset, based on a set of cut-off values. This way, the results of an analysis with and without pruning can be compared.

3.10.5 Species diversity visualisation

First, the species list that was generated during alignment analysis is opened in Excel. The data was cleaned by removing any words that do not belong to the species name such as 'sp.', 'cf.', 'environmental' and 'PREDICTED:'. This list was then uploaded to NCBI's Taxonomy Browser to categorise the species in a phylogenetic tree. The entire tree is then downloaded in PHYLIP format and opened in FigTree v1.4.3 for visualisation and annotation. A full phylogenetic tree is generated and included in the annex (S5_full.pdf and S7_full.pdf). A second circular tree with rudimentary annotation is also generated to be reported in this thesis. Lastly, a bar plot is drawn showing the 10 most detected species in each sample.

4 RESULTS

Th aim of this study is to describe a protocol that allows for the identification of marine zooplankton species through metabarcoding. First, each step in the protocol was evaluated. Once the technique was optimised, it was applied to twelve samples taken on a bio-monitoring campaign. In the following chapter, the results of each test are reported, followed by observations made while processing the twelve bio-monitoring samples.

4.1 DNA extraction

Both the Epicentre® MasterPure™ DNA Purification protocol and a modified CTAB protocol were evaluated with five *Daphnia Magna* tissue samples each. DNA purity was subsequently assessed with a Thermofischer Scientific NanoDrop 2000 (nucleic acid, factor 50). Quality values and nucleic acid quantities of different samples are reported in the annex. There are a few noticeable differences between the two protocols. Firstly, the nucleic acid yield after a CTAB purification (823,68 ± 423.64 ng/µL) is an order of magnitude larger than after the MasterPure™ DNA Purification protocol (79.84 ± 35.83 ng/µL). Secondly, the latter exhibits a 260/230 ratio (0.82 ± 0.15) which is consistently lower than required for downstream applications (Desjardins and Conklin 2010). The purity of CTAB, however, does seem to be sufficient for further sample preparation (2.02 ± 0.11). In both cases, 260/280 values were satisfactory (2.02 ± 0.07 for CTAB and 2.00 ± 0.04 for MasterPure™).

Based on these results, the CTAB protocol was selected to extract DNA from the bio-monitoring samples. The quality values by NanoDrop, as well as the yield as measured by Qubit, are reported in the annex. As these samples were processed, a striking observation was made. After phenol:chloroform:IAA extraction, a wide range of colour intensity was observed between the different samples (Figure 19). More specifically between the four locations where twelve zooplankton community samples were collected. The sample colour ranged from almost translucent to a deep red, with each sampling location having its own distinct hue. The colour intensity decreased markedly after chloroform:IAA purification, but not enough to be considered pure. Therefore, a second chloroform:IAA treatment was applied.

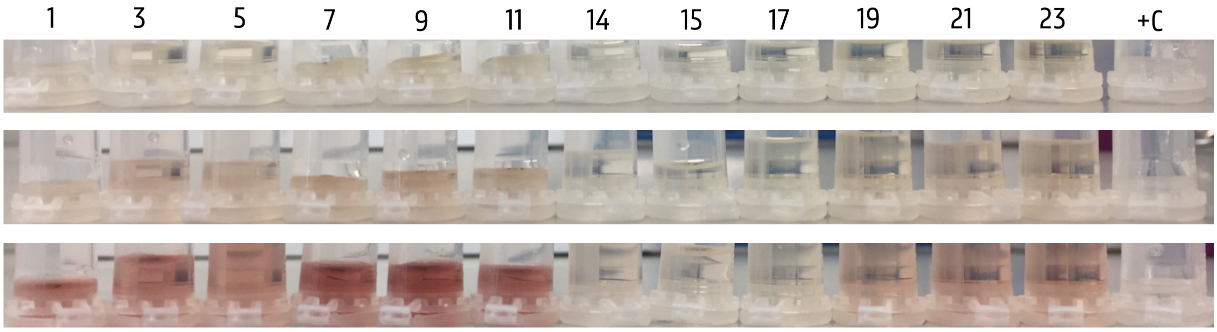

Figure 19: Colour of contaminants present in the sample matrix after phenol extraction (bottom), the first chloroform:isoamyl alcohol (24:1) purification (middle) and after a second chloroform:IAA treatment (top).

NanoDrop and Qubit assays conducted after DNA extraction and purification showed that the quantity and quality of the recovered DNA were satisfactory. Some RNA contamination was observed but considered to be sufficiently low for downstream analysis (see annex).

4.2 PCR amplification

The second step in the protocol is PCR amplification. Two primer sets and three master mixes were evaluated consecutively. During these tests, DNA extract from *Daphnia magna* was used as a positive control. As a negative control, two reactions were performed in the absence of primers (no primer control) and another two without a DNA template (no template control). Marine detritus proved to be a tough matrix for molecular research and some contaminants seemed to interfere with enzymatic reactions. Therefore, two PCR enhancing techniques were investigated to prevent inhibition and increase amplicon yield.

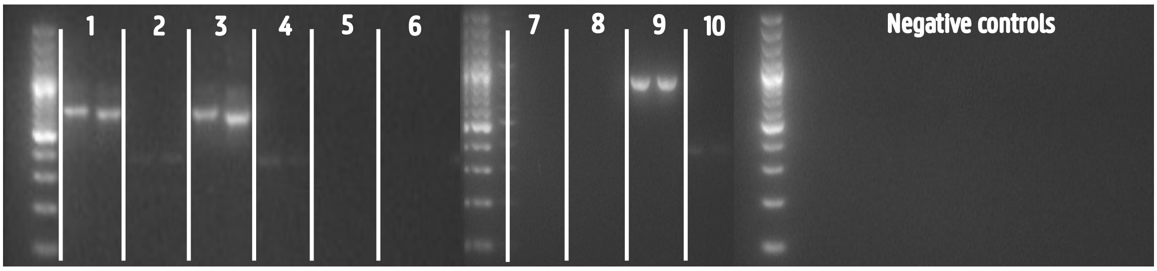

Figure 20: Primer tests gel electrophoresis. (1, 3) 18S handpicked, (2, 4) COI handpicked, (5, 7) 18S Sieved, (6,8) COI Sieved, (9) 18S Daphnia and (10) COI Daphnia. Ladder: 100-3000bp.

The two primer sets mlCOIintF/jgHCO219 and F-566/R-1200 (for COI and 18S respectively) were tested on DNA extracts obtained from three types of samples. Firstly, *Daphnia magna* as a positive control, secondly handpicked copepods, and lastly sieved samples which included marine detritus. Results from gel electrophoresis are shown in Figure 20. The COI primer set amplifies faintly around 300-400 bp, while 18S amplicons are much more prominent and have a length of around 600-700 bp. All but the sieved samples show successful amplification.

Secondly, three types of master mix (MM) were assessed for their resilience against PIC. In each test the gel lanes (Figure 21) were ordered as follows: two no-template controls, two no-primer controls, two positive controls (*Daphnia magna* DNA) and finally two replicates of two sieved marine zooplankton extracts (four in total) to which a 1:20 dilution was applied. For the Maxima™ MM (Figure 21a), all environmental samples, as well as the positive control, were amplified. In the second iteration, the Dreamtaq™ MM (Figure 21b) only amplified the positive control and finally, Amplitaq gold® MM (Figure 21c) wasn't able to amplify any of the DNA templates.

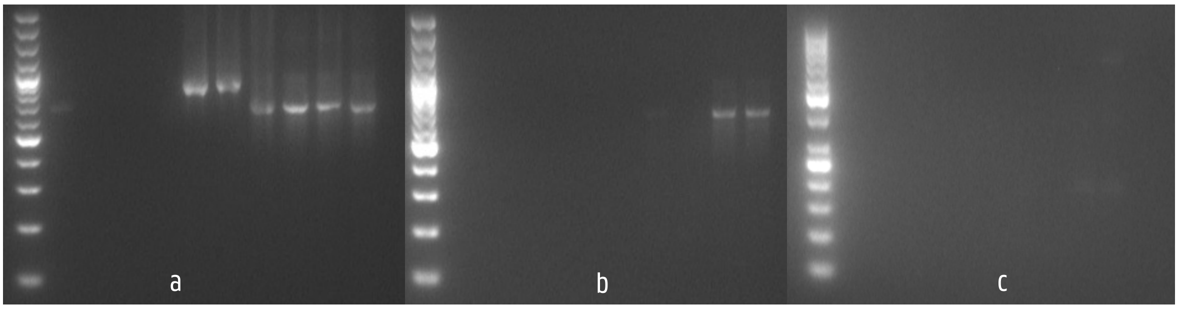

Figure 21: Gel electrophoresis on: (a) Maxima™ hot start, (b) Dreamtaq™, and (c) Amplitaq gold®. DNA ladder: 100-3000 bp.

Lastly, a dilution factor of 1:20 was not enough to facilitate amplification in marine samples taken on the bio-monitoring campaign. Countermeasures for PCR inhibition were investigated. A series of dilution factors were tested on samples from two different locations, with the Maxima™ MM. Results are shown in Figure 23. The minimal dilution factor necessary for amplification was different for both samples. Some amplification can be seen as a factor of 1:30 for sample 19, whereas it takes a 1:80 dilution for sample 5. It was concluded that a 1:80 dilution should be used before adding the template to the PCR reaction mix. This technique was applied to all bio-monitoring samples and was largely successful (see annex).

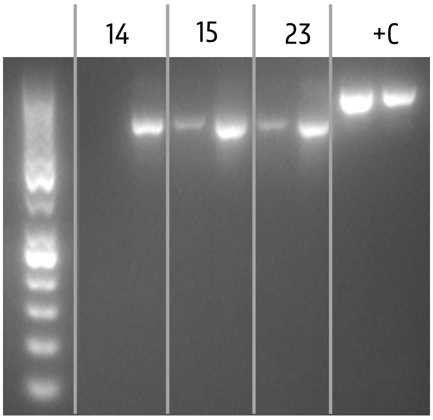

Figure 22: Gel electrophoresis of samples 14, 15 and 23 with (right) and without BSA (left) addition. Ladder: 50-1000 bp

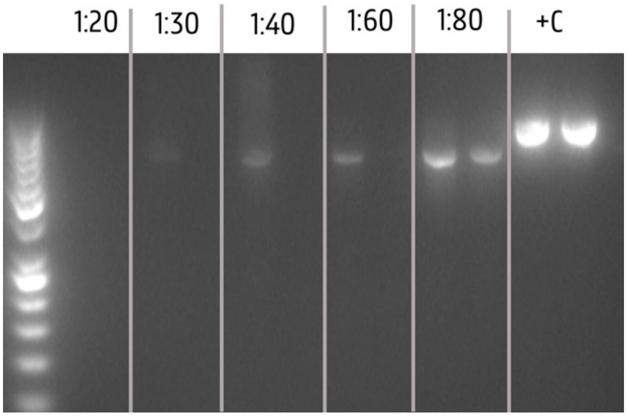

Figure 23: Gel electrophoresis of PCR with several dilution factors. In each frame, two bands represent two environmental samples (19 left and 5 right) taken from different locations. Ladder: 50-1000 bp

For samples 14, 15 and 23 this 1:80 dilution factor still was not enough to produce amplicons. These samples were first treated with an additional chloroform:IAA extraction (the third in total). Then they were amplified, once without any additional treatment and once with the addition of 1µg/µL BSA. In all three instances, the addition of BSA resulted in a noticeably brighter band on the gel (Figure 22). Additionally, the yield of these BSA spiked reactions was significantly higher, as illustrated by a Qubit assay (Table 4). However, NanoDrop values show that the quality of the solution - especially 260/230 - is too low to be applied in downstream sequencing, so additional purification steps must be implemented before these amplicons can be used.

Table 4: Concentration and quality values of PCR products with and without BSA.

Sample	*Treatment*	*Qubit DNA concentration (ng/µl)*	*NanoDrop 260/280*	*NanoDrop 260/230*
14	1:80 dilution	Still no amplification		
15		4,4	1,86	2,06
23		7,04	1,86	2,07
14	1:80 dilution + BSA	22,6	1,71	0,95
15		24,2	1,71	0,91
23		27,88	1,70	0,86

It is interesting to note that the samples that were the most difficult to amplify were also those that showed the least amount of colouring during DNA extraction (Figure 19). One could easily presume that those samples which visually appear to be the most contaminated will actually be the most problematic. However, these observations show that you should never judge a book by its colour.

4.3 PCR product purification

Four PCR product purification methods were performed and evaluated with Qubit and/or Nanodrop assays. The first two constituted a filter tube method in which amplicons were captured from either GE bands or directly from PCR product. Then an enzymatic approach was assessed with the ThermoFisher Scientific CleanSweep™ PCR Purification reagent. The last test involved purification with magnetic beads.

Nearly all trials involving the GE Healthcare illustra GFX PCR DNA and Gel Band Purification Kit failed. Only from two PCR products, DNA of sufficient quality and quantity could be recovered. Both cases constituted direct PCR product purification of *Daphnia magna* 18S amplicons. The washing step was however different between the two samples (S15 vs. D0 in Table 3). Measurements from all samples are listed in the annex.

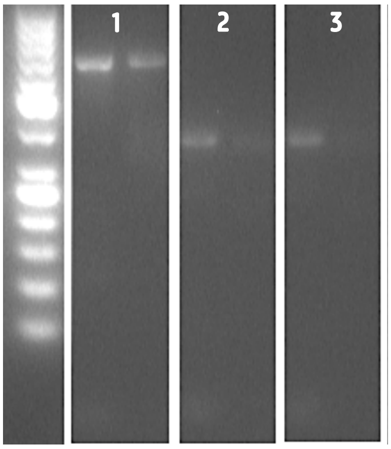

Figure 24: CleanSweep™ assay. Three samples, before (left) and after (right) treatment. DNA ladder: 50-1000 bp

The ThermoFisher Scientific CleanSweep™ PCR Purification reagent, on the other hand, proved to be successful (Figure 24). Amplicon bands are always slightly less bright after CleanSweep™ treatment. This is because adding the enzyme reagent to a PCR product dilutes the amplicons 5 to 7. At the bottom of each lane, a primer band can be seen before but not after CleanSweep™ purification, suggesting that the reaction was successful. NanoDrop values before and after treatment, however, indicate that this method does compromise on quality, especially the 260/230 values (annex). They dropped from 2.13 ± 0.01 before purification to 1.86 ± 0.10 after. After these tests, CleanSweep™ was used to purify PCR product from bio-monitoring amplicons. The results are largely the same (see annex).

Table 5: Quantity and quality of amplicons recovered from magnetic bead purification.

	Before purification			After purification			
Sample	*Mass in 45 µL (ng)*	*260/280*	*260/230*	*Mass in 45 µL (ng)*	*260/280*	*260/230*	*Recovery*
D1	2817	1,86	2,13	1305	1,92	2,42	46,33%
D2	1908	1,71	0,93	936	1,49	0,44	49,06%
D3	2799	1,86	2,13	1575	1,87	2,23	56,27%
D4	2394	1,81	2,12	1197	1,86	2,24	50,00%

Lastly, purification with magnetic beads was assessed on amplicons from *Daphnia magna* samples, one of which was amplified with BSA in the PCR reaction mix (D2). The quality of the amplicons that were recovered is excellent for downstream sequencing (Table 5) but at the expense of DNA recovery which is only about 50%. A second observation is that the quality of BSA amplified samples does not seem to improve, even with magnetic bead purification. The substance also adversely influences the purification protocol. Samples which contain BSA do not form compact pellets on the magnetic rack as non-BSA samples do. Instead, the pellets are smeared out and adhere to the tube wall (Figure 25), as well as to pipet tips if they come in contact with them.

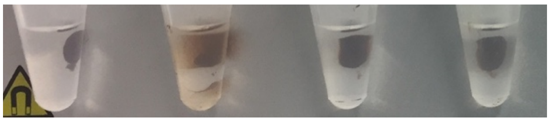

Figure 25: Pelleting on magnetic rack during purification of PCR product. The second tube contains BSA.

As PCR had such a low yield with most bio-monitoring samples, the concentration of amplicons was often not high enough to obtain 1 µg in 46 µL as required for the ONT sequencing by ligation kit. Therefore, multiple PCR reaction tubes of the same sample were pooled after which the CleanPCR kit was used to increase the concentration.

4.4 *In silico* PCR

Now that a primer set is found, *in silico* PCR is performed to confirm its ubiquity and selectiveness for the organisms of interest. A search in SILVA showed that the 18S primer set is indeed very selective for eukaryotes (SILVA_output.xlsx in annex). Of all other domains, only 12 bacterial records returned a match, 10 cyanobacterial and two proteobacterial records.

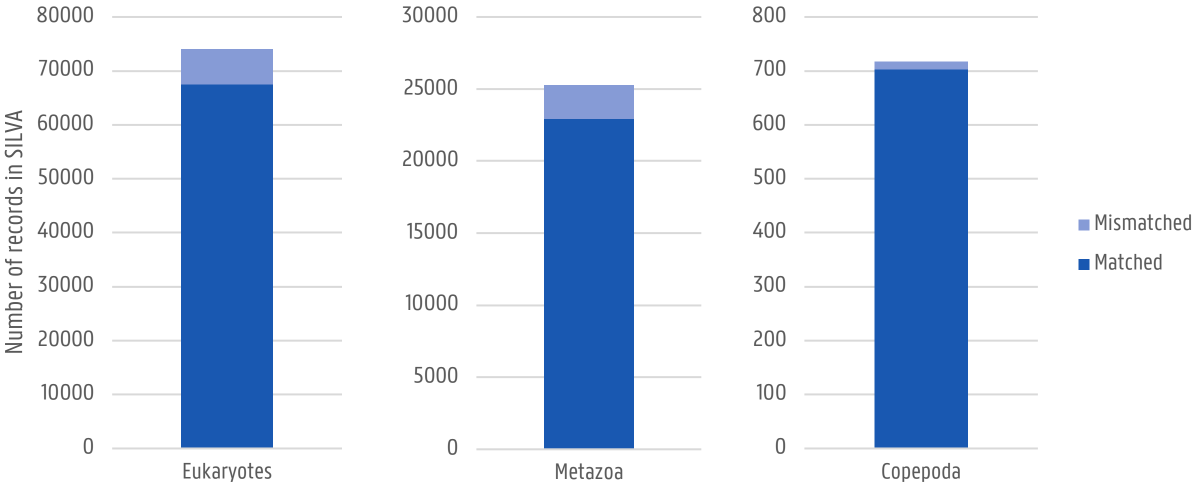

Figure 26: TestPrime search on SILVA database. Results for Eukaryotes, Metazoans and Copepods.

On the other hand, 92.6% of all eukaryotic records matched with the F-566/R-1200 primer set (Figure 26). Looking at the animal kingdom, 90.7% is represented. As discussed before, Copepods are the most abundant taxon to be found in zooplankton communities. The SILVA refNR collection contains 717 copepod records, out of which 703 (98%) could theoretically be detected with these primers.

These results suggest that F-566/R-1200 is sufficiently universal among the taxa that are relevant for this study, while at the same time excluding those which could introduce a significant amount of noise in the data (bacteria and archaea).

4.5 ZooScan

Four samples, each from a different location, were analysed with the ZooScan. Eleven biotic and three abiotic categories were found in these samples. The analysis returns a number of specimens observed for every category. To find the species density at the sample location, a small calculation must be made. At every location, 30 mL of sample was retained from the 500 mL collected with the WP2. These 500 mL is a concentrate of the entire water column that was sieved through this net. This volume can be calculated by multiplying the sampling depth with the surface area of the WP2 mouth, which is 1 m^2. This results in the following formula:

$$Density = \frac{Number\ of\ specimens * 500\ mL}{30\ mL * 1m^2 * sampling\ depth}$$

One of the four samples (Station 1, see Figure 15) could not be analysed with the ZooScan due to a recurring Java script error. Density values of the three remaining samples are shown in Figure 27.

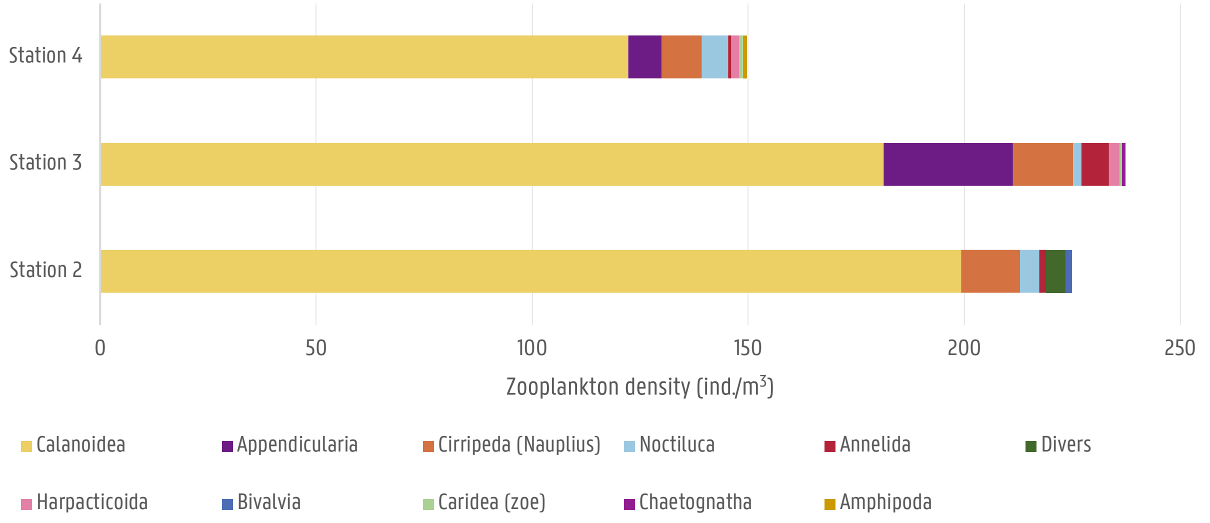

Figure 27: Zooplankton density from three locations as determined by ZooScan.

4.6 DNA sequencing

A total of 12 samples were taken from four locations in the BPNS. The samples were subsequently processed according to the consensus protocols previously discussed. Qubit and Nanodrop measurements from each stage of the protocol can be found in the annex. Due to an unexpected low yield, just four samples produced enough amplicons to perform sequencing, of which two were successful (sample 5 and 17). Data processing results including quality control (QC) and OTU identification are reported below.

4.6.1 Quality control tool selection

Several third party tools for performing data quality control have been assessed. The preferred option here was PycoQC. It is a python module that can be imported in a self-written python script in conjunction with other modules like ‘thread’, ‘os’ or ‘matplotlib’. This gives the researcher optimal flexibility and customisability. A collection of datasets can be imported and directories organised using ‘os’, data can subsequently be processed in parallel with ‘thread’ and plotted with ‘matplotlib’ or ‘seaborn’ for full control over data visualisation. A Phred score per base plot is not supported by this tool, so for this, a custom python script was written (qualpos.py in annex).

4.6.2 Sequencing quality check

Figure 28a and c show bivariate plots of read length versus mean read quality of sample 5 and 17 respectively. Above and next to this plot, the read length and mean quality score density distributions are included along their respective axes. Some interesting statistics can be derived from this data. Firstly the percentage of reads with a

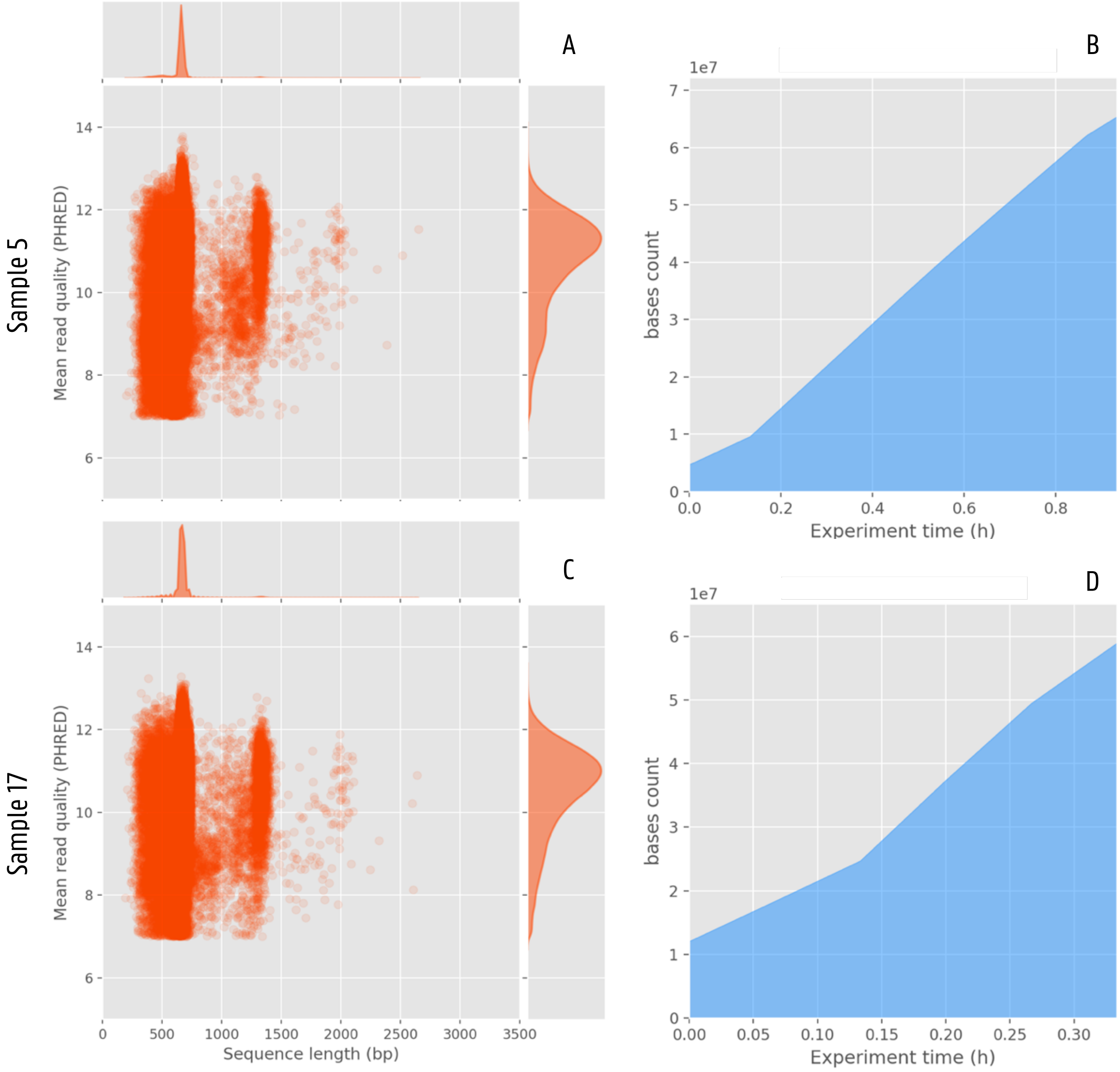

Figure 28: QC plots for samples 5 (top) and 17 (bottom). Bivariate plots of read length versus mean Phred quality score including density functions along their respective axes on the left. Cumulative yield of bases sequenced on the right. The sample 5 dataset counts 330,402 reads and sample 17 contains 261,928.

mean Phred quality score equal to or higher than 10 (i.e. the Q10 value), is 72.9% for sample 5 and 77.8% for sample 17. Secondly, the read length distributions in both datasets show a sharp peak around the predicted amplicon length of 650 bp. In the sample 5 dataset, 90.8% of reads have a length between 550 and 750. For the sample 17 dataset this is 86.7%. Two patterns can be observed in the bivariate plot. The first is that three additional, though much smaller peaks reoccur every 650 bp. The second pattern is that the variance of Phred quality scores seems to linearly decrease for each consecutive peak.

Figure 28b and d show the cumulative yield of bases in sample 5 and 17 respectively. During mux scanning, the graph appears linear. After a few minutes, the scan is complete and sequencing continues at a slightly higher rate. As the sequencing experiment for sample 17 was stopped earlier than that of sample 5, less bases were sequenced.

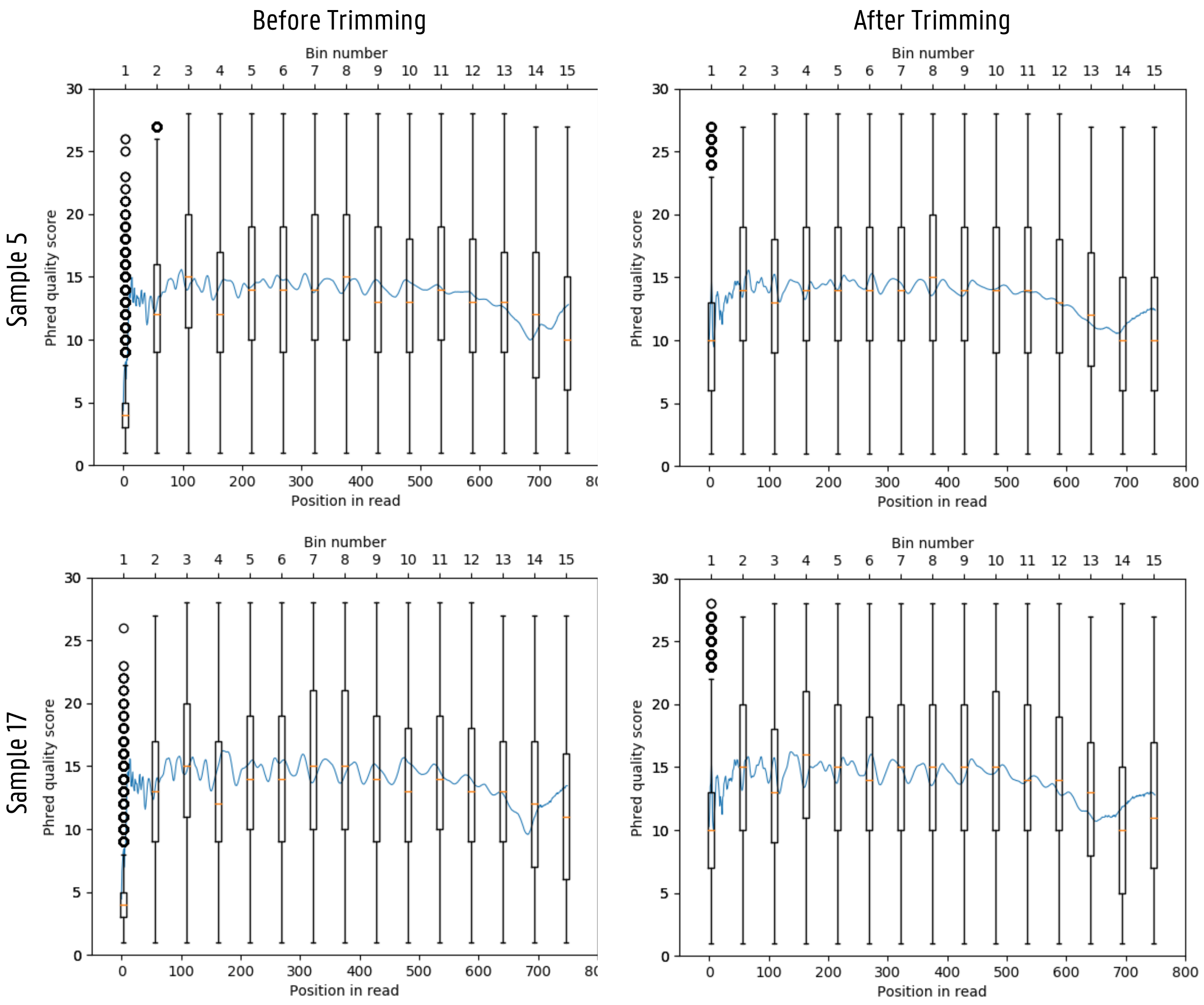

Figure 29: Quality score per position in read of sample 5 (top) and sample 17 (bottom) before (left) and after (right) trimming. A plot of average values for every base (blue line) and boxplots for 50 bp bins are shown.

Looking at the mean Phred score per base of the two datasets, it is quite evident that there exists a large variance on the confidence of base calls (Figure 28). Error probabilities range from 0.2% (Phred = 27) up to 79% (Phred = 1). This underlines the importance of using an alignment algorithm that takes these values into account when searching for a match. This variance, as well as the mean per base Phred score, do however stay largely constant across the length of the read suggesting that the experiment was consistent. Only near the edges, especially at the start of a read, quality does seem to slope down significantly.

4.6.3 Adapter trimming

PoreChop was used to trim adapter sequences from the ends of the reads to reduce noise during alignment. After aligning 10,000 reads to a database of known adapter sequences, the following set was found in both samples:

- SQK-NSK007_Y_Top: AATGTACTTCGTTCAGTTACGTATTGCT
- SQK-NSK007_Y_Bottom: GCAATACGTAACTGAACGAAGT

The algorithm proceeded by mapping these sequences to every read in both datasets. Not only the ends of reads were trimmed but also when an adapter sequence was found in the middle of a read, it was treated as a chimeric sequence and split into two separate records. In sample SW5, 277 098 out of 330,318 reads had adapters trimmed from their start, 81,823 from their end and 3,459 were split based on middle adapters (10,538,2377 bp removed in total). For sample SW17 which contained 261,861 reads, this was 228,815 from the start, 52,587 from the end and 2,949 split with a total of 8,395,665 bp removed.

After adapter trimming (Figure 28 right) the per base quality of the dataset has significantly increased near the edges. The mean Phred score now remains above 10 across the entire read length.

4.6.4 Mapping

By following the procedure described in paragraph 3.10.4.1, a reference database was constructed which contains 52,337 records of 18S rRNA sequences from 18,001 different marine species. No additional redundancy elimination was performed since some intraspecies variation of the 18S gene does exist. This way, the highest possible degree of specificity is maintained. For each of the 18,001 species, the longest reference sequence in the database was taken and plotted in a density distribution (Figure 30). The main peak can be seen around ~1,800 bp, which is the expected length of a full 18S rRNA sequence (KWON, Ogino, and Ishikawa 1991). This suggests that the vast majority of references contain the full F-566/R-1200 barcoding region. A reference sequence is increasingly less likely to contain this region as it becomes shorter.

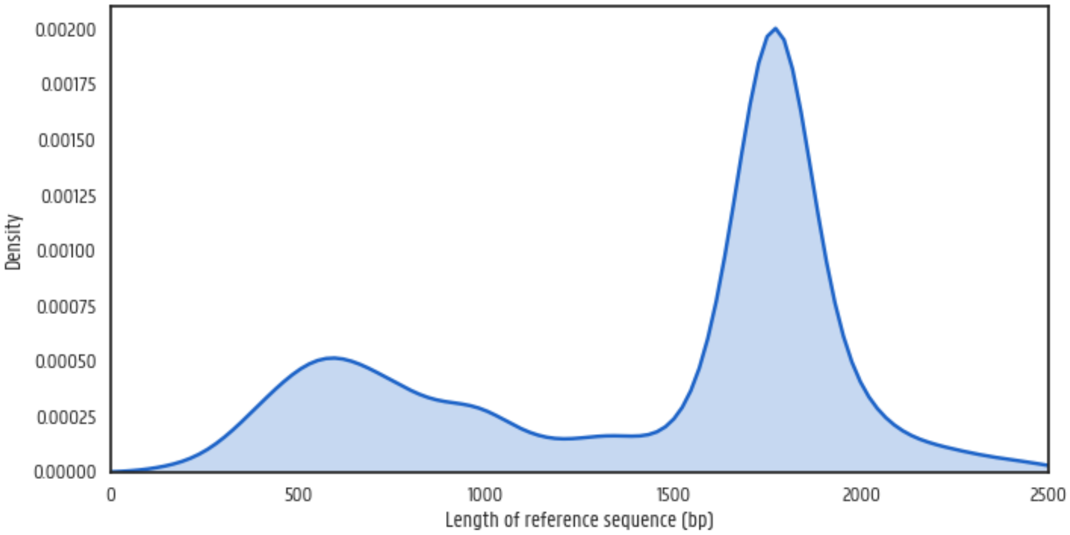

Figure 30: Density distribution of reference sequence lengths form barcoding database.

Based on LAST-train, the optimal parameters for alignment were a gap existence cost of 14, a gap extension cost of 4, an insertion existence cost of 14 and extension cost of 5. The base by base score matrix is shown in Table 6. It is fairly symmetrical indicating that there is no bias towards any of the nucleotides during base calling.

Table 6: Score matrix as determined by LAST-train.

		Reference sequence			
		A	C	G	T
Query sequence	A	6	-18	-7	-18
	C	-18	6	-23	-14
	G	-7	-25	6	-21
	T	-22	-14	-19	6

The match quality was assessed by plotting the E-value of each read that returned a match as a function of its alignment length. Additional density functions of these parameters are drawn along their respective axes (Figure 31). It is quite evident that far fewer reads reside in the top left corner (high E-value, short alignment length) after training LAST to nanopore data. At the same time, training also resulted in more reads getting matched to the reference database (Table 7). This indicates that significantly more reads returned a full barcode-length alignment which in turn leads to a more significant match (i.e. smaller E-value). After analysing the results of the alignment, a list is generated which contains all the species that were detected in the dataset, together with the amount of corresponding reads. When a species is detected by only one or a few reads, it inspires little confidence in these results. This taken into consideration, it is interesting to see that even though generally speaking the reads seem to return more significant matches after data training (Figure 31), the fraction of species that was detected by only a small number of reads does not go down (Table 7).

Finally, the scored dataset was pruned based on the results presented in this chapter. The first cut-off parameter is alignment length. As has been discussed before, the 18S barcode is inherently less variable than other barcoding genes, making it a less optimal choice for species level categorisation. One way to circumvent this is by increasing the query sequence length. The F-566/R-1200 primer set yields only a ~650 bp region of the gene. Therefore, only alignments which span the full length of this barcode are retained and matches with a shorter alignment length are omitted. The second threshold is linked to the quality of the alignment, in particular, the identity between query and reference sequences. From the sequencing summary file, it can be calculated that 96.5% of the reads in dataset 5 and 97.7% in dataset 17 have a mean Phred score of 8 or higher. This value corresponds to a 15.8% error-probability. Based on these observations, matches were only kept if their identity was 84% or higher.

Table 7: Results of alignment analysis

		READS IN DATASET	MATCHED READS	MATCHED SPECIES	SPECIES WITH 1 MATCH	SPECIES WITH ≤10 MATCHES
SAMPLE 5	Default	330,318	314,402	1,653	33.1%	70.4%
	Scored		328,482	2,330	34.3%	74.9%
	pruned		191,271	438	29.5%	63.2%
SAMPLE 17	Default	261,861	255,906	1,137	33.5%	72.6%
	Scored		260,625	1,321	39.1%	74.8%
	pruned		201,104	369	30.4%	72.4%

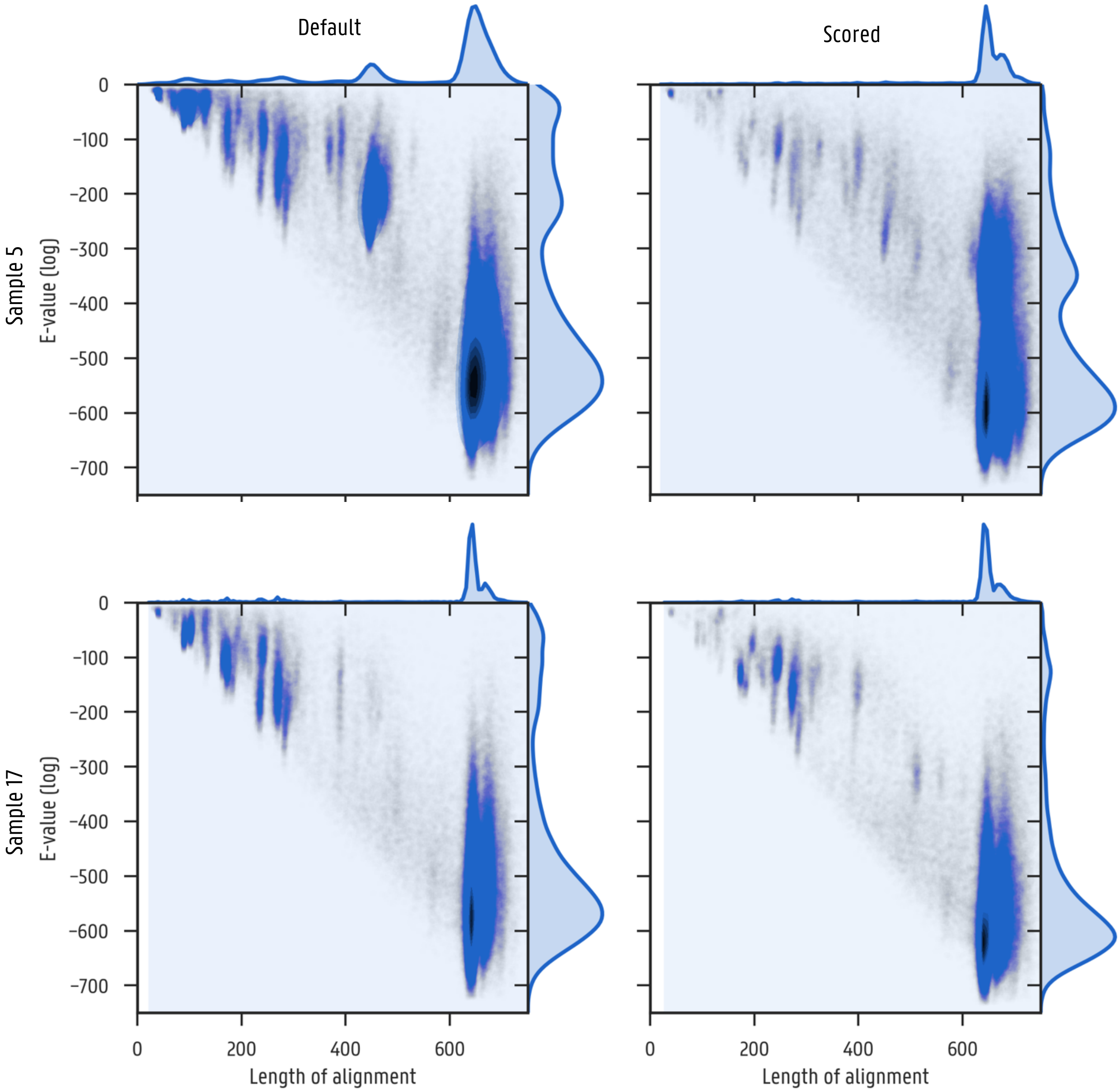

Figure 31: Bivariate plot of log E-value over alignment length. A comparison of LAST alignment with default settings (left) and a score matrix trained for nanopore data (right). Each for sample 5 (top) and sample 17 (bottom).

Pruning has drastically reduced the amount of reads that return a match, especially in sample 5. The effect is even more pronounced when looking at the amount of species that were detected (Table 7). After pruning the fraction of species that was matched by only one and ≤10 has decreased slightly but not dramatically. The match quality plots of both datasets (Figure 32 left) show that all LAST matches now reside in the bottom right corner of the graph, which represents alignments with high confidence. Finally, a bivariate plot was made showing the alignment's position in both the query and the reference sequence (Figure 32 right). Virtually every match originates from a full barcode alignment at the 18S region the F-566/R-1200 primer pair was designed to amplify.

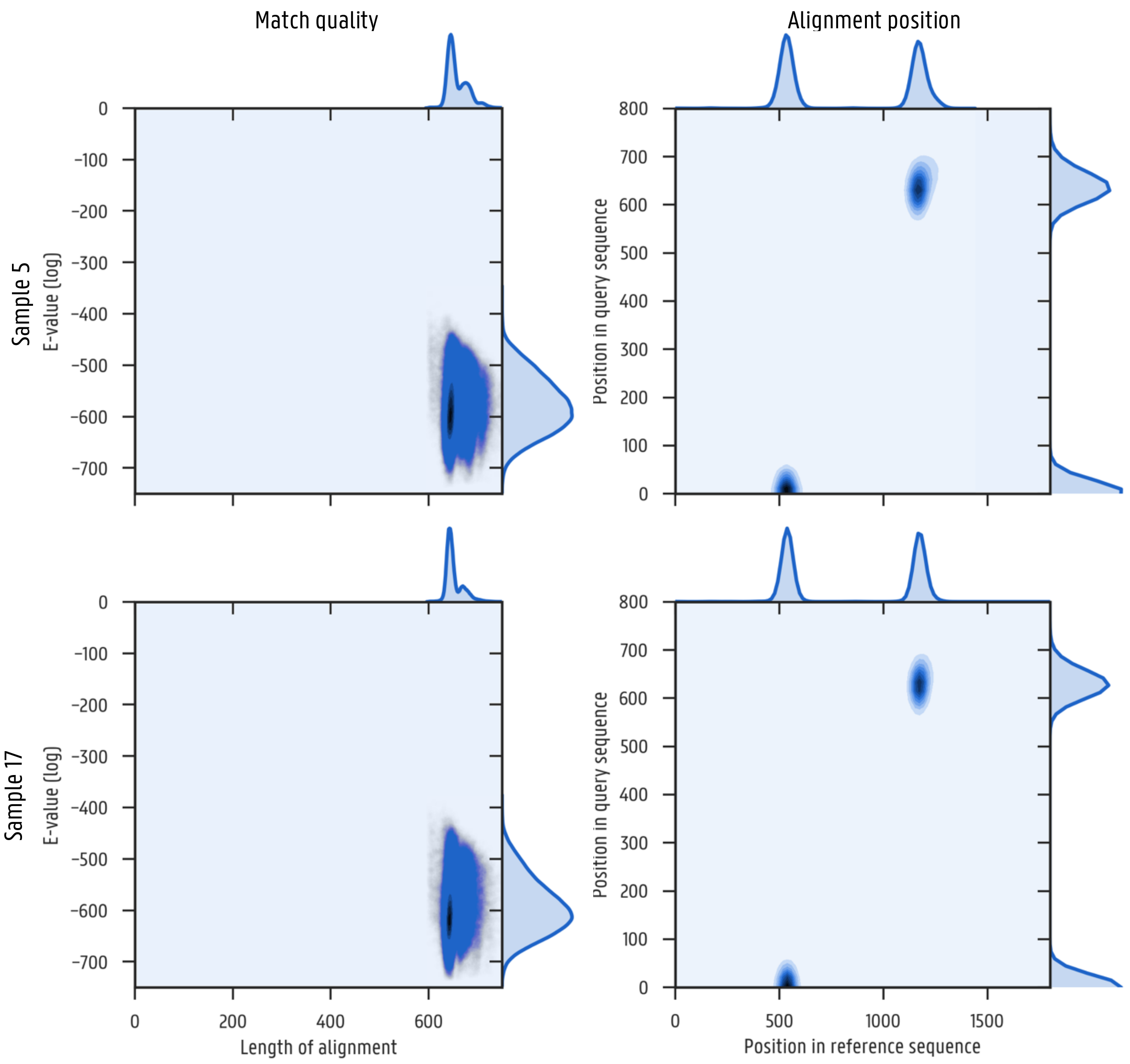

Figure 32: Results after database pruning. A match quality plot is given on the left and a plot representing the position of the alignment in both the read and query sequence is shown on the right. Both sample 5 (top) and sample 17 (bottom) were analysed.

4.6.5 Detected species

As has been discussed before, several species were detected by just one or several reads in the dataset. In order to have a relatively high degree of certainty about species that are reported as present in the sample and for a lack of better solutions, a conservative threshold of 10 matches per species was arbitrarily chosen. This resulted in a list of 168 and 110 species for sample 5 and 17 respectively.

In both cases the copepods are well represented, especially Calanoidea (Figure 33) with *Acartia clausii*, *Calanus helgolandicus*, *Centropages abdominalis*, *Temora turbinate* and *Temora longicornis* being the most detected species (Figure 34). Between the two sample locations, there is a noticeable difference in species diversity. Sample 5 contains a much wider variety of taxa, mainly due to a larger number of crustacean species. Particularly the order of acorn barnacles (Sessilia) seems to be much more diverse close to the coastline (station 1 in Figure 15). Polychaeta, on the other hand, comprises the second largest class. Looking more seaward (sample 17 taken at station 3 in Figure 15), a significant rise in Actinopteri fish species is observed. Also, a new phylum has surfaced, namely the Chaetognatha which were also detected by ZooScan analysis of this location. Nematoda species,

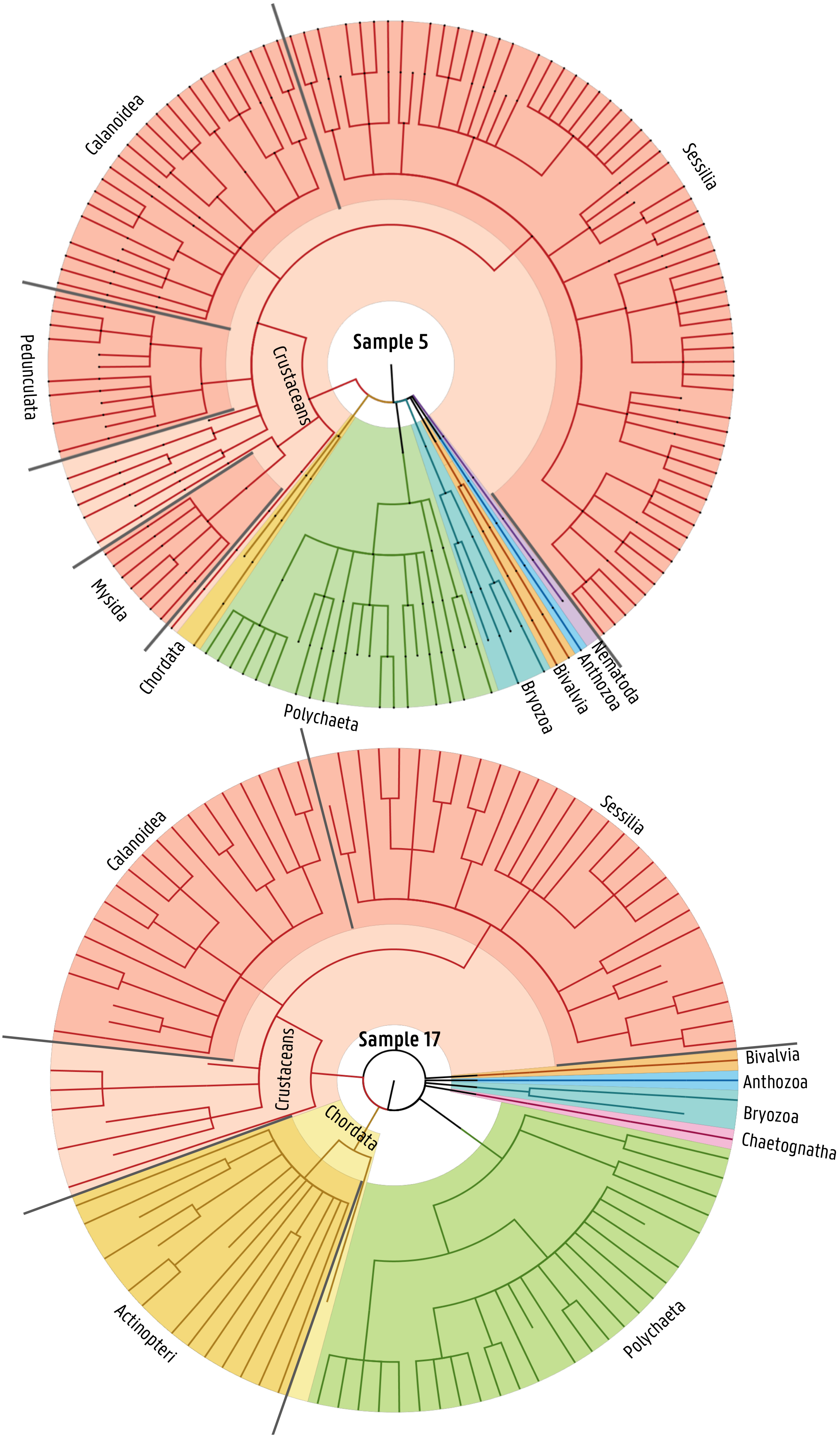

Figure 33: Species diversity as determined by MinION sequencing in sample 5, taken at station 1 and sample 17, taken at station 3

however, are no longer present. Finally, it was surprising to find that no Echinoderm species were detected in either of the datasets, though their absence was affirmed by ZooScan.

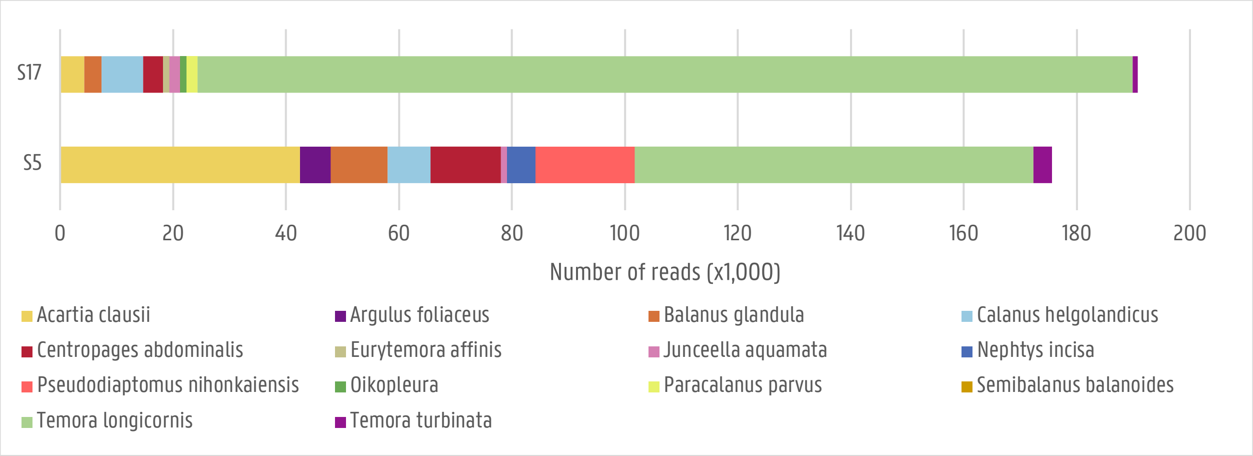

Figure 34: Ten most detected species in sample 5 (station 1) and 17 (station 2).

5 DISCUSSION

5.1 DNA extraction

Two DNA extraction protocol were evaluated on tissue samples from *Daphnia magna*. The most suitable option was then applied to twelve samples taken on the BPNS during a bio-monitoring campaign.

Firstly, the Epicentre® MasterPure™ DNA Purification protocol yielded consistently poor results in these trials. Spectrophotometric quantification of nucleic acid suggests that the amount of DNA recovered is substantially lower than after the CTAB protocol. The difference is even more striking if one considers that in this protocol, the DNA pellet was resuspended in a lower volume of TE buffer (35µL instead of 50 µL for CTAB). The troubleshooting chapter which is included in the extraction kit attributes a low DNA yield to several potential issues, three of which apply to this specific situation. The problem could lay with cell lysis through inhibition from contaminants or the use of Proteinase K with decreased efficacy. This could be solved by increasing Proteinase K, incubation time, vortex mixing or sample tissue homogenisation during cell lysis. The two other common explanations are (1) loss of the DNA pellet after precipitation and (2) contamination with exogenous or endogenous nucleases. However, these last two also apply to CTAB extraction and no problems were observed there.

Secondly, the 260/230 absorbance ratio of these samples was insufficient for downstream applications. According to Thermo Scientific NanoDrop documentation, low 260/230 values indicate contamination with a 230 nm wavelength absorbing contaminants such as EDTA, phenol, TRIzol, Guanidine HCl and glycogen. In the CTAB protocol, phenols and similar organic compounds are thoroughly cleansed from the sample during phenol extraction. The MasterPure™ protocol is potentially not as effective at removing these contaminants. Proteins can be another important source of contamination and are more likely to be the perpetrator here. Peptide bonds in the mix have been found to increase absorption around 230 nm (Kalb and Bernlohr 1977). This is a second indication of some sort of problem with Proteinase K during cell lysis. These interpretations can be taken into consideration for future research but in this thesis, no further assessments were made. The CTAB protocol was chosen to extract DNA from the twelve samples taken on the BPNS.

5.2 PCR amplification

In order to obtain optimal PCR amplification for this particular sample matrix, two primer sets, three master mixes and two PIC mitigation strategies were assessed. The best combination of techniques was subsequently used to amplify one barcoding gene in all bio-monitoring samples.

5.2.1 Primer selection

Two barcoding genes were evaluated each with their own primer set. The COI primer set (mlCOIintF/jgHCO219) did not seem to perform very well under the experimental conditions of these tests. PCR is very sensitive and the reaction conditions must be just right for it to be successful (Dieffenbach and Dveksler 2003). There can be many reasons why a primer set does not perform as well as reported in its initial publication. For example, the optimal DNA polymerase for a primer set can be unstable in the presence of sample specific inhibitors, forcing the researcher to use a less compatible enzyme (Hoorfar et al. 2004). However, the fact that the positive control (freshwater organism) performed equally as bad suggests that the contaminant is not of marine origin. After all, this primer set was tested on the gut content of marine fish. Therefore, the problem is more likely to be inherently linked to the experimental setup, e.g. a difference in $MgCl_2$ or nucleotide concentrations between MMs. Since the 18S primer set (F-566/R-1200) did produce the desired results, no further testing was done on the COI primers.

In addition to a higher PCR product yield, F-566/R-1200 produces much larger amplicons (650 bp instead of around 300). This is especially advantageous in the context of this research because nanopore sequencing is known to have high error rates, especially when the ONT 1D sequencing kit is used. A longer read means that if a match is found, it has more significance, which in turn increases specificity.

As was discussed in the literature review, the 18S barcoding gene has one great disadvantage over COI. It has a lower mutation rate and thereby a lower specificity, making it less useful for species level identification. This is especially true when downstream sequencing is error-prone, as is the case with MinION.

The biggest asset of third generation sequencing is that the read length is only limited by the DNA strand introduced to it. A longer barcode length means a higher specificity and therefore more confidence in OTU determination. This could be used to counteract a loss of accuracy due to high error rates in nanopore sequencing data. In addition to a higher specificity, using longer MinION reads also eliminates cumbersome post-sequencing read assembly of Illumina sequencing (Bálint Miklós et al. 2014). It would be interesting to look into COI primer sets that produce longer barcodes, such as LCO1490/HCO2198 (about 700 bp). Some researchers have even successfully amplified the entire mitochondrial genome for species identification from environmental DNA (eDNA) (Deiner Kristy et al. 2017). Another option would be to sequence entire 18S or even 23S ribosomal subunit genes. The greatest impediment of this technique would be the availability of comprehensive databases with named OTU's.

5.2.2 PCR reaction

From the three master mixes that were assessed, only the Maxima™ MM was able to amplify DNA from environmental samples. The Amplitaq gold® MM was designed and optimised to have highly increased reaction kinetics as opposed to more common MM (~40 min instead of ~2 hours for 35 cycles). It is assumed that there is a trade-off between reaction speed and reaction stability, making the reaction much more sensitive to inhibition. This could explain why the reaction failed. The Dreamtaq™ MM showed some amplification for metazoan tissue samples, albeit much fainter. Presumably, this reagent could still be useful after optimisation, but no further testing was done on this subject. Going forward, the Maxima™ MM was selected for PCR amplification.

Now that a viable combination of primers and MM has been found, it is time to take a look at PICs. All tests on zooplankton tissue samples were successful, but the reaction fails once marine detritus is introduced to the sample. Many sources of inhibiting compounds have been reported, including organic matter and sediment, which partly constitute marine detritus (Wilson 1997). Also, a variety of dissolved salts are known to have adverse effects on DNA extraction or amplification. However, their composition or inhibition mechanisms are largely unknown. Avoiding the effects of PIC can, therefore, be seen as a black box issue.

Dilution allowed for some amplification to occur, but often still resulted in low yields. Ideally, this strategy is supplemented with other techniques, such as the addition of BSA. This resulted in a stark increase in amplicon yield but rendered the product unusable for sequencing. None of the PCR purification strategies were able to recover a pure amplicon solution.

For this study, only dilution was applied to bio-monitoring samples, but BSA might still be an eligible strategy. It is recommended for future research to assess yield and PCR product purity as a function of different BSA concentrations in the reaction mix. Possibly a lower concentration could achieve the same pronounced effect on yield while maintaining a usable product quality.

5.3 PCR product purification

Four techniques for PCR purification were assessed. The first two involved separation on filter tubes, first from GE band cut-outs and then directly from PCR product. One possible explanation for why all trials with this kit failed is the fact that it has recently passed its expiration date. However, the kit was still sealed at the start of these tests. Secondly, this can only have slightly reduced its potency, since some of the *Daphnia magna* samples were still able to be recovered. This means another factor must be at play such as an unknown contaminant interfering with DNA binding to the silica filter tube.

Enzymatic digestion (CleanSweep) and separation on magnetic silica beads (CleanPCR, CleanNA) constituted the other two strategies. Both were successful, each with its own advantages and pitfalls. CleanPCR results in very pure product but recovery is rather low. CleanSweep is quick and does not react with double stranded amplicons, but it diluted the product and decreases the quality by introducing proteins and CleanSweep buffer into the mix. The proteins are denatured after the reaction is complete. Hence, these should normally not interfere with sequencing applications, but this has not been verified. The PCR yield was in many cases too low for direct sequencing which resulted in the need for amplicon pooling and concentration after CleanSweep purification. CleanPCR silica beads were used for this, so it is impossible to assess whether the presence of remaining reagent buffer and denatured enzymes would interfere with the ONT sequencing protocol. It can, however, be concluded that magnetic silica beads from CleanNA yield the best purity of all techniques tested, although a significant loss of DNA must be taken into account.

5.4 DNA sequencing

The MinION sequencing results are all perfectly in line with what can be expected of this sequencing chemistry. First of all, the vast majority of reads had a length similar to what was reported in the F-566/R-1200 design paper. Secondly, the mean Phred quality score is in the same range as data reported in similar MinION research (Benítez-Páez and Sanz 2017; Leija-Salazar et al. 2018).

Two striking observations were made from the bivariate plots. The first one is that there is not just one peak in the read length distribution but that the peak reoccurs every 650 bp. It is possible that a non-optimal end-prep and dA-tailing before adapter ligation has resulted in the concatenation of amplicons. This would explain why the peaks reoccur periodically. Also, this mechanism has been found to result in sometimes adapter sequences in the middle of chimeric reads, which is why PoreChop has developed an algorithm to remove and split them into separate reads. For the purposes of this study, hybrid reads do not pose a problem as they can easily be discriminated from the data of interest. However, they do interfere with alignment algorithms when the data of interest has a wider length distribution that overlaps with these chimeras (Li et al. 2016).

A second observation was that the variance of the mean read Phred quality scores distribution decreases when reads get longer. The mean Phred score of a read is calculated by averaging the Phred score of all bases in that read. The population from which this mean is calculated increases as the read gets longer, thereby resulting in a lower variance on this estimate.

The mean per base quality in Figure 29 decreased significantly near the edges of the read. In the Nanopore Community this is commonly reported and generally assumed to be an inherent property of the adapter sequences which held the motor protein. Adapter trimming with PoreChop seemed to mostly resolve the issue, thereby affirming this hypothesis. Second generation sequencing data is also known to have this problem and similar adapter trimming software has been developed for this (Kong 2011).

5.5 Mapping

Alignment was performed with the LAST alignment algorithm, once with default settings and once after training a score matrix to nanopore data. From the results, it is evident that adjusting the alignment parameters drastically influences the outcome of the alignment. The implementation of a score matrix has markedly improved the alignment since more reads could be mapped, while reducing the fraction of low confidence matches. However, the algorithm still returns short matches with low identity. Additionally, it seems unlikely that 1,300-2,300 OTUs are present in a single sample and implementing the training algorithm only exacerbates this phenomenon.

When taking a closer look at the data, it appears that a large number of OTUs is detected by only one or a few reads. Two possible explanations for a high number of these ‘rare OTUs’ come to mind. Either a significant amount of viable environmental DNA (eDNA) is captured with the zooplankton community or ambiguous matches are returned. A false match can occur when the alignment was of low significance or if this method is inherently incapable of discriminating between close relatives, i.e. several OTUs actually represent one species of a genetically conservative taxon. This bias has already been documented in zooplankton communities, even with vastly more accurate second generation sequencing (Brown Emily A. et al. 2015).

By implementing a cut-off rate for certain alignment quality parameters, matches of low significance can be excluded. If these ‘rare species’ originate from a false alignment with a low quality read, pruning the dataset should resolve the issue. If on the other hand eDNA contamination or ambiguous taxa are the cause, their share should not drop disproportionately when omitting matches of low significance. After pruning, a slightly lower occurrence of ‘rare species’ was observed but they still constitute a major portion of the end result.

Since low quality alignments can mostly be ruled out, these results suggest that another source for single read matches must be at play. This could be eDNA, ambiguous taxa or another unknown mechanism. No further pruning protocols were explored since this is out of the scope of this thesis. Future research can look into statistical analyses which take into account all the quality parameters of every match a read returns. For example, a read might match several times with one species, but only once with the best scoring alignment. Another possibility would be to compare the per base matches and mismatches with the per base Phred score estimate of the query sequence. When a statistically sound OTU determination algorithm is developed, it should be compared to proven technology (e.g. Illumina metabarcoding or morphological identification) for validation.

On the other hand, some valid OTUs were potentially excluded here due to a stringent alignment length cut-off. Omitting matches that are shorter than the full barcode length (in this case 650 bp) leans on the assumption that all OTUs return the same barcode length. This is often not the case, as can be seen in Figure 22 and 23 where *Daphnia magna* (positive control) shows a longer barcode region than marine zooplankton. Interspecies variability of amplicon length has been taken into account by arbitrarily setting the cut-off value at 600 instead of 650. However, a more accurate OTU determination could be achieved by building a reference database which holds just the barcode region of every OTU and performing a global alignment. This forces the alignment to span the entire barcode length of each individual OTU without the need for a fixed cut-off value.

5.6 ZooScan

The ZooScan was used to determine the abundance of 25 taxa at each of the sampling locations. One of the samples (station 1) could not be analysed for reasons still unknown. The program was difficult to use without any prior experience. At the end of the analysis, the data must be validated by manually skimming through the categories and checking whether every specimen has been correctly classified, which requires some degree of taxonomic skills. Even with human verification, accuracies of similar learning sets have been found to be only around 85% (Grosjean et al. 2004). All of these minor drawbacks make the technology not very accessible at the moment.

There are a few additional considerations to be made about the nature of ZooScan data. The biotic categories in which sample specimens are binned are still of a high taxonomic level. In the context of bio-monitoring, little information can be inferred about the zooplankton community under study from this data. For instance, it is important to be able to discern invasive from native species within a taxon, which is not possible with a ZooScan analysis. For research purposes, the training set is quite rudimentary. Similar broad range image recognition software can these days be found in almost every middle to high end smartphone on the market. Training sets with higher taxonomic detail for Copepod taxa have been successfully applied (Gorsky et al. 2010), but they are at the expense of a broad taxonomic coverage. The reason why it is difficult to advance this technology for qualitative analyses is that it requires a large amount of time and money to create better, more specific training sets.

However, the quantitative aspect of ZooScan data is exceptionally useful. In just under an hour it is possible to obtain population densities from 25 taxa. Metabarcoding data can provide a number of reads but it is quite difficult to infer absolute abundances without any knowledge about body size. Secondly, for some taxa the ZooIdentifier is able to discriminate between adult and larval stages, providing insight into fecundity. Lastly, even though some surmountable biases have been described in the literature (Vandromme et al. 2012), ZooScan has been found to be a meritable tool for body size estimation, which is another important ecosystem parameter (Woodward et al. 2005). All of this information is impossible to obtain from genetic barcoding. Therefore, it is argued that when used in conjunction with metabarcoding, ZooScan analysis can provide information about crucial ecosystem function.

5.7 Detected species

Ideally, the results of this barcoding analysis should be validated by proven technology such as morphological identification or metabarcoding with Illumina sequencing. Even though this was not possible within the timeframe of this thesis, a broader sense of validity can be inferred by comparing the results with relevant scientific literature and ZooScan analysis.

Sample 17 of the biomonitoring campaign was sampled at station 3 of which also a ZooScan analysis was successfully performed. Almost all metazoan groups that were observed by ZooScan (Calanoidea, Appendicularia, Cirripedia, Annelida, Harpacticoida, Bivalvia, Chaetognatha) are included in the final pruned OTU list of sample 17 (see annex). Only two discrepancies were discovered. Firstly, no Decapoda OTUs were found in S17, even though a specimen belonging to the infraorder Caridea was present in the ZooScan sample. Secondly, diatom OTUs were completely absent from S17, while several Noctiluca specimens appeared on the ZooScan. The absence of confirmed species in the metabarcoding analysis can have various causes. In the case of Noctiluca, it turns out that NCBI does not contain any records when searching for 18S genes in the Noctiluca genus (txid = 2965). Consequently, this taxon was not included in the reference database and cannot be detected. This underscores the need for a more comprehensive database and adequate validation of the results. On the other hand, some species were only registered by metabarcoding. The most prominent example is the Cnidarian, *Junceella aquamata*, which is the fifth most detected species in S17.

In the literature, some studies have been published about the same area. The most relevant studies were those of Karel Van Ginderdeuren (K Van Ginderdeuren et al. 2013). He has studied the BPNS by means of CPR record analysis, ZooScan and other morphological identification techniques. A detailed comparison of those observations and the results reported here is out of the scope of this thesis. However, some broad scale similarities were found and are discussed below.

Two major differences in species diversity were observed between sample 5 (closest to shore) and sample 17 (further from the coastline). Sessilia were found to be very abundant near the coast. This is not surprising considering that they are benthic species and thereby thrive at locations where more substrate is available to them. Similar observations have been reported in the past (Figure 5). Another taxon that is generally more associated with a nearshore habitat are Mysida ((K Van Ginderdeuren et al. 2013)), or opossum shrimp, which also consisted of a more diverse community in sample 5 (Figure 33). The total number of detected copepod reads was noticeably larger in sample 17, which conform to observations by Van Ginderdeuren that midshore samples contain more copepod individuals than those from onshore locations like sample 5. Looking more seaward, Actinopteri seemed to be much more prevalent as compared to sample 5. Fish OTUs can be interpreted as either meroplanktonic larvae, tychoplanktonic eggs or eDNA contamination in the sample. Results from (K Van Ginderdeuren et al. 2013) show that some Pisces species were detected mid-shore and offshore, but very few to none were reported in nearshore plankton communities. Polychaeta on the other have been found to be ubiquitously represented, which is in line with the results shown in Figure 33.

Finally, it is important to note that abundance in numbers is not an optimal measure for species diversity. However, it does provide some insight into the locations where the respective taxa are known to flourish. It can be concluded that in broad terms the observations made in this research conform to previous studies at the same locations. This provides at least some validation for the results.

6 CONCLUSIONS

The aim of this thesis was to assess the viability of using MinION sequencing for metabarcoding of marine zooplankton communities. First, several techniques were evaluated for every step of the metabarcoding protocol and the optimal strategy was applied to samples taken on a one day bio-monitoring campaign with the Simon Stevin research vessel (RV) on the Belgian part of the North Sea (BPNS). After sequencing, a number of third party software, as well as self-written processing tools, were applied to the output data.

A protocol was successfully described and applied to the bio-monitoring samples. First plankton is collected with a WP2 net for vertical sampling. The sample is sieved and a CTAB extraction protocol was chosen to obtain purified DNA from the filter cake. The F-566/R-1200 primer pair successfully amplified a barcode region from the 18S rRNA gene when used with the ThermoFisher Scientific Maxima Hot Start PCR Master Mix. To mitigate the effects of PCR inhibiting compounds, the DNA extract had to be diluted 1:80. This enabled amplification, but yields remained rather low. After PCR product purification with ThermoFisher Scientific CleanSweep™ PCR Purification reagent and concentration with CleanPCR magnetic silica beads (CleanNA) only four samples had enough amplicons remaining for MinION sequencing. Of these, two were successfully sequenced and processed.

The read quality was in line with what can be found in literature and determined OTUs have mostly been verified by ZooScan analysis and literature review. Data does point to a few flaws in the current methodology. Even after data pruning based on alignment length and identity, a large amount of OTUs were still detected by just one or a few reads. Several explanations can be found at this point. It can be an indication of eDNA contamination, an inherent incapacity of the method to distinguish between genetic close relatives or bad alignments. The latter could be resolved in future research by selecting OTUs based on statistical models which take into account quality parameters of every database match a query sequence returns, instead of simply keeping the highest quality alignment. Due to the high variability (Phred score 1-27) on per base quality during base calling, it would be interesting to look into algorithms which compare these values with matches and mismatches in the alignment. Also, barcoding regions that are longer or have a higher mutation rate should be assessed to counteract the relatively low accuracy associated with nanopore sequencing data, thereby providing more confidence in OUT determination. Finally, further research is needed on reference database assembly. Every reference sequence should ideally consist only of the barcoding region, thereby allowing global alignment.

Overall, this study shows that applying MinION sequencing to metabarcoding of marine zooplankton communities has merit. It allows researchers of small laboratories to gain access to sequencing technology and perform metabarcoding analyses at any location and in a relatively short amount of time. It is believed that as the price of this technology goes down and when used in conjunction with valuable quantitative data from ZooScan analyses, this method can disrupt the field of bio-monitoring and allow large scale routine monitoring programs that could drastically increase our understanding of biodiversity dynamics.

7 ACKNOWLEDGEMENTS

The author thanks the LifeWatch observatory as part of the Flemish contribution to the LifeWatch ESFRI by Flanders Marine Institute for the use of the ZooScanner.

8 REFERENCES

Abramoff, MD, PJ Magalhaes, and SJ Ram. 2004. 'Image Processing with ImageJ. Biophotonics Int. 11: 36–42'. *Google Scholar*.

Álvarez, Eva, Marta Moyano, Ángel López-Urrutia, Enrique Nogueira, and Renate Scharek. 2014. 'Routine Determination of Plankton Community Composition and Size Structure: A Comparison between FlowCAM and Light Microscopy'. *Journal of Plankton Research* 36 (1): 170–84. https://doi.org/10.1093/plankt/fbt069.

Angel, MV. 1997. 'Pelagic Biodiversity'. *Marine Biodiversity: Patterns and Processes. Cambridge University Press, Cambridge, UK*.

Antia, Avan N, Wolfgang Koeve, Gerhard Fischer, Thomas Blanz, Detlef Schulz-Bull, Jan Schölten, Susanne Neuer, Klaus Kremling, Joachim Kuss, and Rolf Peinert. 2001. 'Basin-wide Particulate Carbon Flux in the Atlantic Ocean: Regional Export Patterns and Potential for Atmospheric CO2 Sequestration'. *Global Biogeochemical Cycles* 15 (4): 845–62.

Armisen, Rafael, and Fernando Galatas. 1987. 'Production, Properties and Uses of Agar'. *Production and Utilization of Products from Commercial Seaweeds. FAO Fish. Tech. Pap* 288: 1–57.

Avery, Oswald T, Colin M MacLeod, and Maclyn McCarty. 1944. 'Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III'. *Journal of Experimental Medicine* 79 (2): 137–58.

Bálint Miklós, Schmidt Philipp-André, Sharma Rahul, Thines Marco, and Schmitt Imke. 2014. 'An Illumina Metabarcoding Pipeline for Fungi'. *Ecology and Evolution* 4 (13): 2642–53. https://doi.org/10.1002/ece3.1107.

Barnosky, Anthony D., Nicholas Matzke, Susumu Tomiya, Guinevere O. U. Wogan, Brian Swartz, Tiago B. Quental, Charles Marshall, et al. 2011. 'Has the Earth's Sixth Mass Extinction Already Arrived?' *Nature* 471: 51. http://dx.doi.org/10.1038/nature09678.

Beman, J. Michael, Kevin R. Arrigo, and Pamela A. Matson. 2005. 'Agricultural Runoff Fuels Large Phytoplankton Blooms in Vulnerable Areas of the Ocean'. *Nature* 434 (7030): 211.

Benítez-Páez, Alfonso, and Yolanda Sanz. 2017. 'Multi-Locus and Long Amplicon Sequencing Approach to Study Microbial Diversity at Species Level Using the MinION^{TM} Portable Nanopore Sequencer'. *GigaScience* 6 (7): 1–12.

Böhnecke, Günther. 1922. *Salzgehalt Und Strömungen Der Nordsee, von Dr. Günther Böhnecke...* ES Mittler und Sohn.

Bolton, Barry. 1994. *Identification Guide to the Ant Genera of the World.* Harvard University Press.

Brierley, Andrew S, and Michael J Kingsford. 2009. ‘Impacts of Climate Change on Marine Organisms and Ecosystems’. *Current Biology* 19 (14): R602–14.

Brown Emily A., Chain Frédéric J. J., Crease Teresa J., MacIsaac Hugh J., and Cristescu Melania E. 2015. ‘Divergence Thresholds and Divergent Biodiversity Estimates: Can Metabarcoding Reliably Describe Zooplankton Communities?’ *Ecology and Evolution* 5 (11): 2234–51. https://doi.org/10.1002/ece3.1485.

Brown, James H, James F Gillooly, Andrew P Allen, Van M Savage, and Geoffrey B West. 2004. ‘Toward a Metabolic Theory of Ecology’. *Ecology* 85 (7): 1771–89.

Bucklin, Ann, Brian D. Ortman, Robert M. Jennings, Lisa M. Nigro, Christopher J. Sweetman, Nancy J. Copley, Tracey Sutton, and Peter H. Wiebe. 2010. ‘A “Rosetta Stone” for Metazoan Zooplankton: DNA Barcode Analysis of Species Diversity of the Sargasso Sea (Northwest Atlantic Ocean)’. *Deep Sea Research Part II: Topical Studies in Oceanography* 57 (24–26): 2234–47.

Cavalier-Smith, Thomas. 2003. ‘Protist Phylogeny and the High-Level Classification of Protozoa’. *European Journal of Protistology* 39 (4): 338–48.

Chandler, DP, JK Fredrickson, and FJ Brockman. 1997. ‘Effect of PCR Template Concentration on the Composition and Distribution of Total Community 16S RDNA Clone Libraries’. *Molecular Ecology* 6 (5): 475–82.

Clarke, Laurence J., Julien Soubrier, Laura S. Weyrich, and Alan Cooper. 2014. ‘Environmental Metabarcodes for Insects: In Silico PCR Reveals Potential for Taxonomic Bias’. *Molecular Ecology Resources* 14 (6): 1160–70.

Cohen, A.S., D.R. Najarian, and B.L. Karger. 1990. ‘Separation and Analysis of DNA Sequence Reaction Products by Capillary Gel Electrophoresis’. *SECOND INTERNATIONAL SYMPOSIUM ON HIGH PERFORMANCE CAPILLARY* 516 (1): 49–60. https://doi.org/10.1016/S0021-9673(01)90203-1.

Cook, John, Dana Nuccitelli, Sarah A Green, Mark Richardson, Bärbel Winkler, Rob Painting, Robert Way, Peter Jacobs, and Andrew Skuce. 2013. ‘Quantifying the Consensus on Anthropogenic Global Warming in the Scientific Literature’. *Environmental Research Letters* 8 (2): 024024.

Dahm, Ralf. 2005. ‘Friedrich Miescher and the Discovery of DNA’. *Developmental Biology* 278 (2): 274–88.

Davis, Cabell S, Fredrik T Thwaites, Scott M Gallager, and Qiao Hu. 2005. ‘A Three-axis Fast-tow Digital Video Plankton Recorder for Rapid Surveys of Plankton Taxa and Hydrography’. *Limnology and Oceanography: Methods* 3 (2): 59–74.

De Rijk, Peter, Jean-Marc Neefs, Yves Van de Peer, and Rupert De Wachter. 1992. ‘Compilation of Small Ribosomal Subunit RNA Sequences’. *Nucleic Acids Research* 20 (Suppl): 2075.

Deiner Kristy, Renshaw Mark A., Li Yiyuan, Olds Brett P., Lodge David M., and Pfrender Michael E. 2017. ‘Long-range PCR Allows Sequencing of Mitochondrial Genomes from Environmental DNA’. *Methods in Ecology and Evolution* 8 (12): 1888–98. https://doi.org/10.1111/2041-210X.12836.

Desjardins, Philippe, and Deborah Conklin. 2010. ‘NanoDrop Microvolume Quantitation of Nucleic Acids’. *Journal of Visualized Experiments: JoVE*, no. 45.

Dieffenbach, C. W., and G. S. Dveksler. 2003. *PCR Primer: A Laboratory Manual.* Cold Spring Harbor: Cold Spring

Harbor Laboratory Press.

Directive 2000/60/EC. 2000. *Official Journal of European Union*. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000L0060.

Directive 2008/56/EC. 2008. *Official Journal of the European Union*.

EarthSky. 2015. 'How Much Do Oceans Add to World's Oxygen?' *Earth* (blog). 7 August 2015. http://earthsky.org/earth/how-much-do-oceans-add-to-worlds-oxygen.

Economic, United Nations Department of. 2006. *World Population Prospects: The 2004 Revision. Sex and Age Distribution of the World Population*. Vol. 2. United Nations Publications.

Edwards, Arwyn, Aliyah R Debbonaire, Birgit Sattler, Luis AJ Mur, and Andrew J Hodson. 2016. 'Extreme Metagenomics Using Nanopore DNA Sequencing: A Field Report from Svalbard, 78 N'. *BioRxiv*, January. https://doi.org/10.1101/073965.

Edwards, Arwyn, Andre Soares, Sara Rassner, Paul Green, Joao Felix, and Andrew Mitchell. 2017. 'Deep Sequencing: Intra-Terrestrial Metagenomics Illustrates The Potential Of Off-Grid Nanopore DNA Sequencing'. *BioRxiv*, January. https://doi.org/10.1101/133413.

Essington, Timothy E., Pamela E. Moriarty, Halley E. Froehlich, Emma E. Hodgson, Laura E. Koehn, Kiva L. Oken, Margaret C. Siple, and Christine C. Stawitz. 2015. 'Fishing Amplifies Forage Fish Population Collapses'. *Proceedings of the National Academy of Sciences* 112 (21): 6648. https://doi.org/10.1073/pnas.1422020112.

Fabricius, K. E., G. De'ath, S. Noonan, and S. Uthicke. 2014. 'Ecological Effects of Ocean Acidification and Habitat Complexity on Reef-Associated Macroinvertebrate Communities'. In *Proc. R. Soc. B*, 281:20132479. The Royal Society.

FAO. 2016. 'The State of World Fisheries and Aquaculture 2016. Contributing to Food Security and Nutrition for All.' Rome. http://www.fao.org/3/a-i5555e.pdf.

Fasham, M. J. R., B. M. Balino, M. C. Bowles, R. Anderson, D. Archer, U. Bathmann, P. Boyd, et al. 2001. 'A New Vision of Ocean Biogeochemistry after a Decade of the Joint Global Ocean Flux Study (JGOFS)'. *AMBIO: A Journal of the Human Environment* 2001 (Sp. No. 10): 4–31. http://eprints.uni-kiel.de/38192/.

Fiers, Walter, Roland Contreras, Fred Duerinck, Guy Haegeman, Dirk Iserentant, Jozef Merregaert, W Min Jou, Francis Molemans, Alex Raeymaekers, and A Van den Berghe. 1976. 'Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene'. *Nature* 260 (5551): 500–507.

Flewelling, Leanne J, Jerome P Naar, Jay P Abbott, Daniel G Baden, Nélio B Barros, Gregory D Bossart, Marie-Yasmine D Bottein, et al. 2005. 'Red Tides and Marine Mammal Mortalities: Unexpected Brevetoxin Vectors May Account for Deaths Long after or Remote from an Algal Bloom'. *Nature* 435 (7043): 755–56. https://doi.org/10.1038/nature435755a.

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. 'DNA Primers for Amplification of Mitochondrial Cytochrome c Oxidase Subunit I from Diverse Metazoan Invertebrates.' *Molecular Marine Biology and Biotechnology* 3 (5): 294–99.

Frith, Martin C., Raymond Wan, and Paul Horton. 2010. 'Incorporating Sequence Quality Data into Alignment Improves DNA Read Mapping'. *Nucleic Acids Research* 38 (7): e100–e100. https://doi.org/10.1093/nar/gkq010.

Gorsky, Gaby, Mark D Ohman, Marc Picheral, Stéphane Gasparini, Lars Stemmann, Jean-Baptiste Romagnan, Alison Cawood, Stéphane Pesant, Carmen García-Comas, and Franck Prejger. 2010. 'Digital Zooplankton Image Analysis Using the ZooScan Integrated System'. *Journal of Plankton Research* 32 (3): 285–303.

Grant, Rachel Anne, and Katrin Linse. 2009. 'Barcoding Antarctic Biodiversity: Current Status and the CAML Initiative, a Case Study of Marine Invertebrates'. *Polar Biology* 32 (11): 1629.

Groot, Rudolf de, Luke Brander, Sander van der Ploeg, Robert Costanza, Florence Bernard, Leon Braat, Mike Christie, et al. 2012. 'Global Estimates of the Value of Ecosystems and Their Services in Monetary Units'. *Ecosystem Services* 1 (1): 50–61. https://doi.org/10.1016/j.ecoser.2012.07.005.

Grosjean, Philippe, Marc Picheral, Caroline Warembourg, and Gabriel Gorsky. 2004. 'Enumeration, Measurement, and Identification of Net Zooplankton Samples Using the ZOOSCAN Digital Imaging System'. *ICES Journal of Marine Science* 61 (4): 518–25. https://doi.org/10.1016/j.icesjms.2004.03.012.

Hadziavdic, Kenan, Katrine Lekang, Anders Lanzen, Inge Jonassen, Eric M. Thompson, and Christofer Troedsson. 2014. 'Characterization of the 18S RRNA Gene for Designing Universal Eukaryote Specific Primers'. *PloS One* 9 (2): e87624.

Hajibabaei, M., M.A. Smith, D.H. Janzen, J.J. Rodriguez, J.B. Whitfield, and P.D.N. Hebert. 2006. 'A Minimalist Barcode Can Identify a Specimen Whose DNA Is Degraded'. *Molecular Ecology Notes* 6 (4): 959–64. https://doi.org/10.1111/j.1471-8286.2006.01470.x.

Hamada, Michiaki, Yukiteru Ono, Kiyoshi Asai, and Martin C Frith. 2017. 'Training Alignment Parameters for Arbitrary Sequencers with LAST-TRAIN'. *Bioinformatics* 33 (6): 926–28. https://doi.org/10.1093/bioinformatics/btw742.

Harms-Tuohy, Chelsea A, Nikolaos V Schizas, and Richard S Appeldoorn. 2016. 'Use of DNA Metabarcoding for Stomach Content Analysis in the Invasive Lionfish Pterois Volitans in Puerto Rico'. *Marine Ecology Progress Series* 558: 181–91.

Hebert, Paul DN, Alina Cywinska, and Shelley L Ball. 2003. 'Biological Identifications through DNA Barcodes'. *Proceedings of the Royal Society of London B: Biological Sciences* 270 (1512): 313–21.

Hebert, Paul DN, Sujeevan Ratnasingham, and Jeremy R. de Waard. 2003. 'Barcoding Animal Life: Cytochrome c Oxidase Subunit 1 Divergences among Closely Related Species'. *Proceedings of the Royal Society of London B: Biological Sciences* 270 (Suppl 1): S96–99.

Hoorfar, Jeffrey, B Malorny, A Abdulmawjood, N Cook, M Wagner, and P Fach. 2004. 'Practical Considerations in Design of Internal Amplification Controls for Diagnostic PCR Assays'. *Journal of Clinical Microbiology* 42 (5): 1863–68.

Hügler, Michael, and Stefan M. Sievert. 2010. 'Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean'. *Annual Review of Marine Science* 3 (1): 261–89. https://doi.org/10.1146/annurev-marine-120709-142712.

Hull, David L. 1965. 'The Effect of Essentialism on Taxonomy--Two Thousand Years of Stasis (I)'. *The British Journal for the Philosophy of Science* 15 (60): 314–26. http://www.jstor.org/stable/686538.

Ide, Keiichiro, Kazutaka Takahashi, Akira Kuwata, Miwa Nakamachi, and Hiroaki Saito. 2008. 'A Rapid Analysis of Copepod Feeding Using FlowCAM'. *Journal of Plankton Research* 30 (3): 275–81.

Ip, Camilla LC, Matthew Loose, John R. Tyson, Mariateresa de Cesare, Bonnie L. Brown, Miten Jain, Richard M. Leggett, David A. Eccles, Vadim Zalunin, and John M. Urban. 2015. 'MinION Analysis and Reference Consortium: Phase 1 Data Release and Analysis'. *F1000Research* 4.

Jackson, Jeremy BC, Michael X Kirby, Wolfgang H Berger, Karen A Bjorndal, Louis W Botsford, Bruce J Bourque, Roger H Bradbury, Richard Cooke, Jon Erlandson, and James A Estes. 2001. 'Historical Overfishing and the Recent Collapse of Coastal Ecosystems'. *Science* 293 (5530): 629–37.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, Andrew D. Beggs, Alexander T. Dilthey, and Ian T. Fiddes. 2018. 'Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads'. *Nature Biotechnology*.

Jin, Di, Eric Thunberg, and Porter Hoagland. 2008. 'Economic Impact of the 2005 Red Tide Event on Commercial Shellfish Fisheries in New England'. *Ocean & Coastal Management* 51 (5): 420–29.

Johns, D. G., and P. C. Reid. 2001. 'An Overview of Plankton Ecology in the North Sea'.

Kalb, Vernon F., and Robert W. Bernlohr. 1977. 'A New Spectrophotometric Assay for Protein in Cell Extracts'. *Analytical Biochemistry* 82 (2): 362–71. https://doi.org/10.1016/0003-2697(77)90173-7.

Kasting, J.F. 1993. 'Earth's Early Atmosphere'. *Science* 259 (5097): 920–26. https://doi.org/10.1126/science.11536547.

Kirby, K. S. 1964. 'Isolation and Fractionation of Nucleic Acids'. In *Progress in Nucleic Acid Research and Molecular Biology*, 3:1–31. Elsevier.

Kong, Yong. 2011. 'Btrim: A Fast, Lightweight Adapter and Quality Trimming Program for next-Generation Sequencing Technologies'. *Genomics* 98 (2): 152–53.

Kwok, Shirley. 1990. 'Procedures to Minimize PCR-Product Carry-Over'. *PCR Protocols: A Guide to Methods and Applications*, 142–45.

KWON, O-Yu, Kimihiro Ogino, and Hajime Ishikawa. 1991. 'The Longest 18S Ribosomal RNA Ever Known'. *The FEBS Journal* 202 (3): 827–33.

Lacroix, Geneviève, Kevin Ruddick, Nathalie Gypens, and Christiane Lancelot. 2007. 'Modelling the Relative Impact of Rivers (Scheldt/Rhine/Seine) and Western Channel Waters on the Nutrient and Diatoms/Phaeocystis Distributions in Belgian Waters (Southern North Sea)'. *Continental Shelf Research* 27 (10): 1422–46. https://doi.org/10.1016/j.csr.2007.01.013.

Lacroix, Geneviève, Kevin Ruddick, José Ozer, and Christiane Lancelot. 2004. 'Modelling the Impact of the Scheldt and Rhine/Meuse Plumes on the Salinity Distribution in Belgian Waters (Southern North Sea)'. *Journal of Sea Research* 52 (3): 149–63. https://doi.org/10.1016/j.seares.2004.01.003.

Lai, Joelle CY, Peter KL Ng, and Peter JF Davie. 2010. 'A Revision of the Portunus Pelagicus (Linnaeus, 1758) Species Complex (Crustacea: Brachyura: Portunidae), with the Recognition of Four Species.' *Raffles Bulletin of Zoology* 58 (2).

Laurenceau-Cornec, Emmanuel C, TW Trull, Diana M Davies, Stephen G Bray, Jacqui Doran, Frederic Planchon, Francois Carlotti, MP Jouander, A-J Cavagna, and Anya Waite. 2015. 'The Relative Importance of Phytoplankton Aggregates and Zooplankton Fecal Pellets to Carbon Export: Insights from Free-Drifting Sediment Trap

Deployments in Naturally Iron-Fertilised Waters near the Kerguelen Plateau'. *Biogeosciences* 12: 1007–27.

Laver, Thomas, J. Harrison, P. A. O'neill, Karen Moore, Audrey Farbos, Konrad Paszkiewicz, and David J. Studholme. 2015. 'Assessing the Performance of the Oxford Nanopore Technologies Minion'. *Biomolecular Detection and Quantification* 3: 1–8.

Lee, Tae-Gee, and Susumu Maruyama. 1998. 'Isolation of HIV-1 Protease-Inhibiting Peptides from Thermolysin Hydrolysate of Oyster Proteins'. *Biochemical and Biophysical Research Communications* 253 (3): 604–8.

Legendre, Louis, and Fereidoun Rassoulzadegan. 1996. 'Food-Web Mediated Export of Biogenic Carbon in Oceans: Hydrodynamic Control'. *Marine Ecology Progress Series* 145 (1/3): 179–93. http://www.jstor.org/stable/24857325.

Leija-Salazar, Melissa, Fritz J Sedlazeck, Katya Mokretar, Stephen Mullin, Marco Toffoli, Maria Athanasopoulou, Aimee Donald, et al. 2018. 'Detection of GBA Missense Mutations and Other Variants Using the Oxford Nanopore MinION'. *BioRxiv*, January. https://doi.org/10.1101/288068.

Leray, Matthieu, Joy Y. Yang, Christopher P. Meyer, Suzanne C. Mills, Natalia Agudelo, Vincent Ranwez, Joel T. Boehm, and Ryuji J. Machida. 2013. 'A New Versatile Primer Set Targeting a Short Fragment of the Mitochondrial COI Region for Metabarcoding Metazoan Diversity: Application for Characterizing Coral Reef Fish Gut Contents'. *Frontiers in Zoology* 10 (1): 34.

Li, Chenhao, Kern Rei Chng, Esther Jia Hui Boey, Amanda Hui Qi Ng, Andreas Wilm, and Niranjan Nagarajan. 2016. 'INC-Seq: Accurate Single Molecule Reads Using Nanopore Sequencing'. *GigaScience* 5 (1): 34.

Loh, Tse-Lynn, Steven E. McMurray, Timothy P. Henkel, Jan Vicente, and Joseph R. Pawlik. 2015. 'Indirect Effects of Overfishing on Caribbean Reefs: Sponges Overgrow Reef-Building Corals'. Edited by Patricia Gandini. *PeerJ* 3 (April): e901. https://doi.org/10.7717/peerj.901.

Loman, Nicholas J, and Aaron R Quinlan. 2014. 'Poretools: A Toolkit for Analyzing Nanopore Sequence Data'. *Bioinformatics* 30 (23): 3399–3401.

Longhurst, Alan R. 2010. *Ecological Geography of the Sea*. Elsevier.

Longhurst, Alan, Shubha Sathyendranath, Trevor Platt, and Carla Caverhill. 1995. 'An Estimate of Global Primary Production in the Ocean from Satellite Radiometer Data'. *Journal of Plankton Research* 17 (6): 1245–71. https://doi.org/10.1093/plankt/17.6.1245.

Lotze, Heike K, Hunter S Lenihan, Bruce J Bourque, Roger H Bradbury, Richard G Cooke, Matthew C Kay, Susan M Kidwell, Michael X Kirby, Charles H Peterson, and Jeremy BC Jackson. 2006. 'Depletion, Degradation, and Recovery Potential of Estuaries and Coastal Seas'. *Science* 312 (5781): 1806–9.

Machida, Ryuji J., and Nancy Knowlton. 2012. 'PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences'. *PLoS One* 7 (9): e46180.

Machida, Ryuji J., and Atsushi Tsuda. 2010. 'Dissimilarity of Species and Forms of Planktonic Neocalanus Copepods Using Mitochondrial COI, 12S, Nuclear ITS, and 28S Gene Sequences'. *PLoS One* 5 (4): e10278.

Menegon, Michele, Chiara Cantaloni, Ana Rodriguez-Prieto, Cesare Centomo, Ahmed Abdelfattah, Marzia Rossato, Massimo Bernardi, Luciano Xumerle, Simon Loader, and Massimo Delledonne. 2017. 'On Site DNA Barcoding by Nanopore Sequencing'. *PloS One* 12 (10): e0184741.

Mikheyev, Alexander S, and Mandy MY Tin. 2014. ‘A First Look at the Oxford Nanopore MinION Sequencer’. *Molecular Ecology Resources* 14 (6): 1097–1102.

Mora, Camilo, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. 2011. ‘How Many Species Are There on Earth and in the Ocean?’ *PLOS Biology* 9 (8): e1001127. https://doi.org/10.1371/journal.pbio.1001127.

Muxika, I., A. Borja, and W. Bonne. 2005. ‘The Suitability of the Marine Biotic Index (AMBI) to New Impact Sources along European Coasts’. *Ecological Indicators* 5 (1): 19–31.

Ngo, Dai-Hung, Thanh-Sang Vo, Dai-Nghiep Ngo, Isuru Wijesekara, and Se-Kwon Kim. 2012. ‘Biological Activities and Potential Health Benefits of Bioactive Peptides Derived from Marine Organisms’. *International Journal of Biological Macromolecules* 51 (4): 378–83. https://doi.org/10.1016/j.ijbiomac.2012.06.001.

Orr, James C., Victoria J. Fabry, Olivier Aumont, Laurent Bopp, Scott C. Doney, Richard A. Feely, Anand Gnanadesikan, et al. 2005. ‘Anthropogenic Ocean Acidification over the Twenty-First Century and Its Impact on Calcifying Organisms’. *Nature* 437 (September): 681. http://dx.doi.org/10.1038/nature04095.

Otto, L., J.T.F. Zimmerman, G.K. Furnes, M. Mork, R. Saetre, and G. Becker. 1990. ‘Review of the Physical Oceanography of the North Sea’. *Netherlands Journal of Sea Research* 26 (2): 161–238. https://doi.org/10.1016/0077-7579(90)90091-T.

Paquin, P, and M Hedin. 2004. ‘The Power and Perils of “Molecular Taxonomy”: A Case Study of Eyeless and Endangered Cicurina (Araneae: Dictynidae) from Texas Caves’. *Molecular Ecology* 13 (10): 3239–55.

Paramor, Odette, K.A. Allen, Margrethe Aanesen, C Armstrong, Troels Hegland, W Le Quesne, Gerjan Piet, et al. 2009. *North Sea Atlas*.

Pihl, Leif. 1994. ‘Changes in the Diet of Demersal Fish Due to Eutrophication-Induced Hypoxia in the Kattegat, Sweden’. *Canadian Journal of Fisheries and Aquatic Sciences* 51 (2): 321–36.

Pinsky, Malin L., and Stephen R. Palumbi. 2014. ‘Meta-analysis Reveals Lower Genetic Diversity in Overfished Populations’. *Molecular Ecology* 23 (1): 29–39.

Polz, Martin F, and Colleen M Cavanaugh. 1998. ‘Bias in Template-to-Product Ratios in Multitemplate PCR’. *Applied and Environmental Microbiology* 64 (10): 3724–30.

Pomerantz, Aaron, Nicolas Penafiel, Alejandro Arteaga, Lucas Bustamante, Frank Pichardo, Luis A Coloma, Cesar L Barrio-Amoros, David Salazar-Valenzuela, and Stefan Prost. 2017. ‘Real-Time DNA Barcoding in a Remote Rainforest Using Nanopore Sequencing’. *BioRxiv*, January. https://doi.org/10.1101/189159.

Poulton, Alex J., Tim R. Adey, William M. Balch, and Patrick M. Holligan. 2007. ‘Relating Coccolithophore Calcification Rates to Phytoplankton Community Dynamics: Regional Differences and Implications for Carbon Export’. *The Role of Marine Organic Carbon and Calcite Fluxes in Driving Global Climate Change, Past and Future* 54 (5): 538–57. https://doi.org/10.1016/j.dsr2.2006.12.003.

Pruesse, Elmar, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. 2007. ‘SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB’. *Nucleic Acids Research* 35 (21): 7188–96.

Quick, Joshua, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, and Amy Mikhail. 2016. ‘Real-Time, Portable Genome Sequencing for

Ebola Surveillance'. *Nature* 530 (7589): 228.

Quick, Joshua, Aaron R Quinlan, and Nicholas J Loman. 2014. 'A Reference Bacterial Genome Dataset Generated on the MinION™ Portable Single-Molecule Nanopore Sequencer'. *Gigascience* 3 (1): 22.

Rådström, Peter, Rickard Knutsson, Petra Wolffs, Maria Lövenklev, and Charlotta Löfström. 2004. 'Pre-PCR Processing'. *Molecular Biotechnology* 26 (2): 133–46.

Rae, K.M., and C.B. Rees. 1947. 'Continuous Plankton Records: The Copepoda in the North Sea, 1938-1939'. *Hull Bulletins of Marine Ecology* 2 (11): 95–132.

Rajapakse, Niranjan, Eresha Mendis, Hee-Guk Byun, and Se-Kwon Kim. 2005. 'Purification and in Vitro Antioxidative Effects of Giant Squid Muscle Peptides on Free Radical-Mediated Oxidative Systems'. *The Journal of Nutritional Biochemistry* 16 (9): 562–69.

Ratnasingham, Sujeevan, and Paul DN Hebert. 2007. 'BOLD: The Barcode of Life Data System (Http://Www. Barcodinglife. Org)'. *Molecular Ecology Resources* 7 (3): 355–64.

Raupach, Michael J, Andrea Barco, Dirk Steinke, Jan Beermann, Silke Laakmann, Inga Mohrbeck, Hermann Neumann, Terue C Kihara, Karin Pointner, and Adriana Radulovici. 2015. 'The Application of DNA Barcodes for the Identification of Marine Crustaceans from the North Sea and Adjacent Regions'. *PloS One* 10 (9): e0139421.

Richardson, Katherine, Torkel Gissel Nielsen, Flemming Bo Pedersen, Jens Peter Heilmann, Bo Løkkegaard, and Hanne Kaas. 1998. 'Spatial Heterogeneity in the Structure of the Planktonic Food Web in the North Sea'. *Marine Ecology Progress Series*, 197–211.

Riemann, Lasse, Hanna Alfredsson, Michael M Hansen, Thomas D Als, Torkel G Nielsen, Peter Munk, Kim Aarestrup, Gregory E Maes, Henrik Sparholt, and Michael I Petersen. 2010. 'Qualitative Assessment of the Diet of European Eel Larvae in the Sargasso Sea Resolved by DNA Barcoding'. *Biology Letters* 6 (6): 819–22.

Roe, H. S. J. 1974. 'Observations on the Diurnal Vertical Migrations of an Oceanic Animal Community'. *Marine Biology* 28 (2): 99–113.

Ryther, John H. 1970. 'Biological Sciences: Is the World's Oxygen Supply Threatened?' *Nature* 227 (5256): 374–75.

Sanders, R, T Jickells, and D Mills. 2001. 'Nutrients and Chlorophyll at Two Sites in the Thames Plume and Southern North Sea'. *Journal of Sea Research* 46 (1): 13–28. https://doi.org/10.1016/S1385-1101(01)00068-5.

Sanger, Frederick, Steven Nicklen, and Alan R Coulson. 1977. 'DNA Sequencing with Chain-Terminating Inhibitors'. *Proceedings of the National Academy of Sciences* 74 (12): 5463–67.

Schopf, J.W., and B.M. Packer. 1987. 'Early Archean (3.3-Billion to 3.5-Billion-Year-Old) Microfossils from Warrawoona Group, Australia'. *Science* 237 (4810): 70–73. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0023163693&partnerID=40&md5=c30071b5da6247303df77206dfa24d4f.

Shiganova, T. A. 1998. 'Invasion of the Black Sea by the Ctenophore Mnemiopsis Leidyi and Recent Changes in Pelagic Community Structure'. *Fisheries Oceanography* 7 (3-4): 305–10.

Smith, Bruce D, and Melinda A Zeder. 2013. 'The Onset of the Anthropocene'. *Anthropocene* 4: 8–13.

Solomon, Susan, Dahe Qin, Martin Manning, Zhenlin Chen, Merlinda Marquis, Kristen B. Averyt, M. Tignor, and Henry L. Miller. 2007. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel* *on Climate Change, 2007*. Cambridge University Press, Cambridge.

Spaulding, Sarah A, David H Jewson, Rebecca J Bixby, Harry Nelson, and Diane M McKnight. 2012. 'Automated Measurement of Diatom Size'. *Limnology and Oceanography: Methods* 10 (11): 882–90.

Stockinger, Herbert, Manuela Krüger, and Arthur Schüßler. 2010. 'DNA Barcoding of Arbuscular Mycorrhizal Fungi'. *New Phytologist* 187 (2): 461–74.

Subramaniam, A., P. L. Yager, E. J. Carpenter, C. Mahaffey, K. Björkman, S. Cooley, A. B. Kustka, et al. 2008. 'Amazon River Enhances Diazotrophy and Carbon Sequestration in the Tropical North Atlantic Ocean'. *Proceedings of the National Academy of Sciences* 105 (30): 10460. https://doi.org/10.1073/pnas.0710279105.

Sunday, Jennifer M., Katharina E. Fabricius, Kristy J. Kroeker, Kathryn M. Anderson, Norah E. Brown, James P. Barry, Sean D. Connell, Sam Dupont, Brian Gaylord, and Jason M. Hall-Spencer. 2017. 'Ocean Acidification Can Mediate Biodiversity Shifts by Changing Biogenic Habitat'. *Nature Climate Change* 7 (1): 81.

Tautz, Diethard, Peter Arctander, Alessandro Minelli, Richard H Thomas, and Alfried P Vogler. 2003. 'A Plea for DNA Taxonomy'. *Trends in Ecology & Evolution* 18 (2): 70–74.

Turner, Jefferson T. 2004. 'The Importance of Small Planktonic Copepods and Their Roles in Pelagic Marine Food Webs'. *Zoological Studies* 43 (2): 255–66.

Van Ginderdeuren, K, K Hostens, G Van Hoey, and M Vincx. 2013. 'The Mesozooplankton Species Association in the Belgian Part of the North Sea'.

Van Ginderdeuren, Karl, Frank Fiers, Annelies De Backer, Magda Vincx, and Kris Hostens. 2012. 'Updating the Zooplankton Species List for the Belgian Part of the North Sea.' *Belgian Journal of Zoology* 142 (1).

Vandromme, Pieter, Lars Stemmann, Carmen Garcìa-Comas, Léo Berline, Xiaoxia Sun, and Gaby Gorsky. 2012. 'Assessing Biases in Computing Size Spectra of Automatically Classified Zooplankton from Imaging Systems: A Case Study with the ZooScan Integrated System'. *Methods in Oceanography* 1–2 (April): 3–21. https://doi.org/10.1016/j.mio.2012.06.001.

Volk, Tyler, and Martin I Hoffert. 1985. 'Ocean Carbon Pumps: Analysis of Relative Strengths and Efficiencies in Ocean-driven Atmospheric CO2 Changes'. *The Carbon Cycle and Atmospheric CO: Natural Variations Archean to Present*, 99–110.

Watson, James D, and Francis HC Crick. 1953. 'The Structure of DNA'. In , 18:123–31. Cold Spring Harbor Laboratory Press.

Watson, Mick, Marian Thomson, Judith Risse, Richard Talbot, Javier Santoyo-Lopez, Karim Gharbi, and Mark Blaxter. 2015. 'PoRe: An R Package for the Visualization and Analysis of Nanopore Sequencing Data'. *Bioinformatics* 31 (1): 114–15. https://doi.org/10.1093/bioinformatics/btu590.

Wescoe, Zachary L, Jacob Schreiber, and Mark Akeson. 2014. 'Nanopores Discriminate among Five C5-Cytosine Variants in DNA'. *Journal of the American Chemical Society* 136 (47): 16582–87.

Wexels Riser, Christian, Paul Wassmann, Kalle Olli, Anna Pasternak, and Elena Arashkevich. 2002. 'Seasonal Variation in Production, Retention and Export of Zooplankton Faecal Pellets in the Marginal Ice Zone and Central Barents Sea'. *Seasonal C-Cycling Variability in the Open and Ice-Covered Waters of the Barents Sea* 38 (1): 175–88. https://doi.org/10.1016/S0924-7963(02)00176-8.

Wilson, Ian G. 1997. 'Inhibition and Facilitation of Nucleic Acid Amplification.' *Applied and Environmental Microbiology* 63 (10): 3741.

Wolfe, K H, W H Li, and P M Sharp. 1987. 'Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs'. *Proceedings of the National Academy of Sciences* 84 (24): 9054. http://www.pnas.org/content/84/24/9054.abstract.

Woodward, Guy, Bo Ebenman, Mark Emmerson, Jose M Montoya, Jens M Olesen, Alfredo Valido, and Philip H Warren. 2005. 'Body Size in Ecological Networks'. *Trends in Ecology & Evolution* 20 (7): 402–9.

Worm, Boris, Edward B Barbier, Nicola Beaumont, J Emmett Duffy, Carl Folke, Benjamin S Halpern, Jeremy BC Jackson, Heike K Lotze, Fiorenza Micheli, and Stephen R Palumbi. 2006. 'Impacts of Biodiversity Loss on Ocean Ecosystem Services'. *Science* 314 (5800): 787–90.