Nanopore sequencing of ocean microbiomes

A thesis submitted to the School of Environmental Sciences at the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Emma Langan

December 2022

Abstract

Ocean microbiomes are responsible for the majority of global primary production, and are crucial for global biogeochemical cycles. Eukaryotic phytoplankton contribute to the carbon cycle which is important for the climate, with polar communities making a disproportionate contribution. These communities are threatened by climate change, and it is important to understand their distribution and interactions so that these impacts can be modelled and monitored.

Ocean microbiomes have not been well characterised, but nanopore sequencing could be used to study them with long-reads and in situ sequencing. This project piloted the use of nanopore sequencing for studying ocean microbes to improve our understanding of their genomes, communities, and interactions.

A genome assembly was produced for a haploid *Emiliania huxleyi* strain, to complement the single publicly available diploid genome assembly. The new assembly uses a hybrid approach and represents a significant improvement on the diploid assembly with contiguity and completeness comparable to other recent haptophyte genome assemblies.

In situ nanopore sequencing and real-time taxonomic classification onboard the RRS Discovery in the Southern Ocean established the utility of nanopore sequencing on polar ocean research cruises, and provided insights into polar ocean microbiomes. Additional samples were collected for land-based sequencing which identified key communities such as diatoms, and produced assembly-free functional annotations.

An improved protocol was developed to reduce sampling requirements, and reliance on toxic reagents, for potential use by citizen scientists. The improved protocol was successfully implemented in an in situ sequencing experiment with real-time analysis on the Norfolk coast, and a time-course analysis to investigate population flux.

These experiments showed the benefits of nanopore sequencing for researchers studying ocean microbiomes, and provided insights into their genomes, and communities. With constant improvements to the technology, nanopore sequencing will only become more useful for the study of ocean microbial communities.

Contents

List of Figures

List of Tables

Acknowledgements

Thank you first to my supervisory team, Thomas Mock, Richard Leggett, Clara Manno, and Vincent Moulton - I really appreciate all the guidance and support you've given me over the last 5 and a bit years, I've learned so much from all of you. And to Darren Heavens at EI, thank you for helping me with everything from fieldwork to thesis brainstorming sessions, I'm so glad you joined in with the madness.

I was incredibly lucky to be part of the DY098 Antarctic research cruise which was a wonderful experience, thanks to everyone involved for making it such a happy trip - I'll treasure the memories forever. Huge thanks as well to Rob Utting, Andy MacDonald and the rest of the technicians team at UEA, without your help my MinION would never have made it off dry land. I'm also grateful to Alan, Igor, and the rest of the team at JGI who welcomed me and made my time in California so enjoyable, I'm glad I got to come and meet you all before the world shut down.

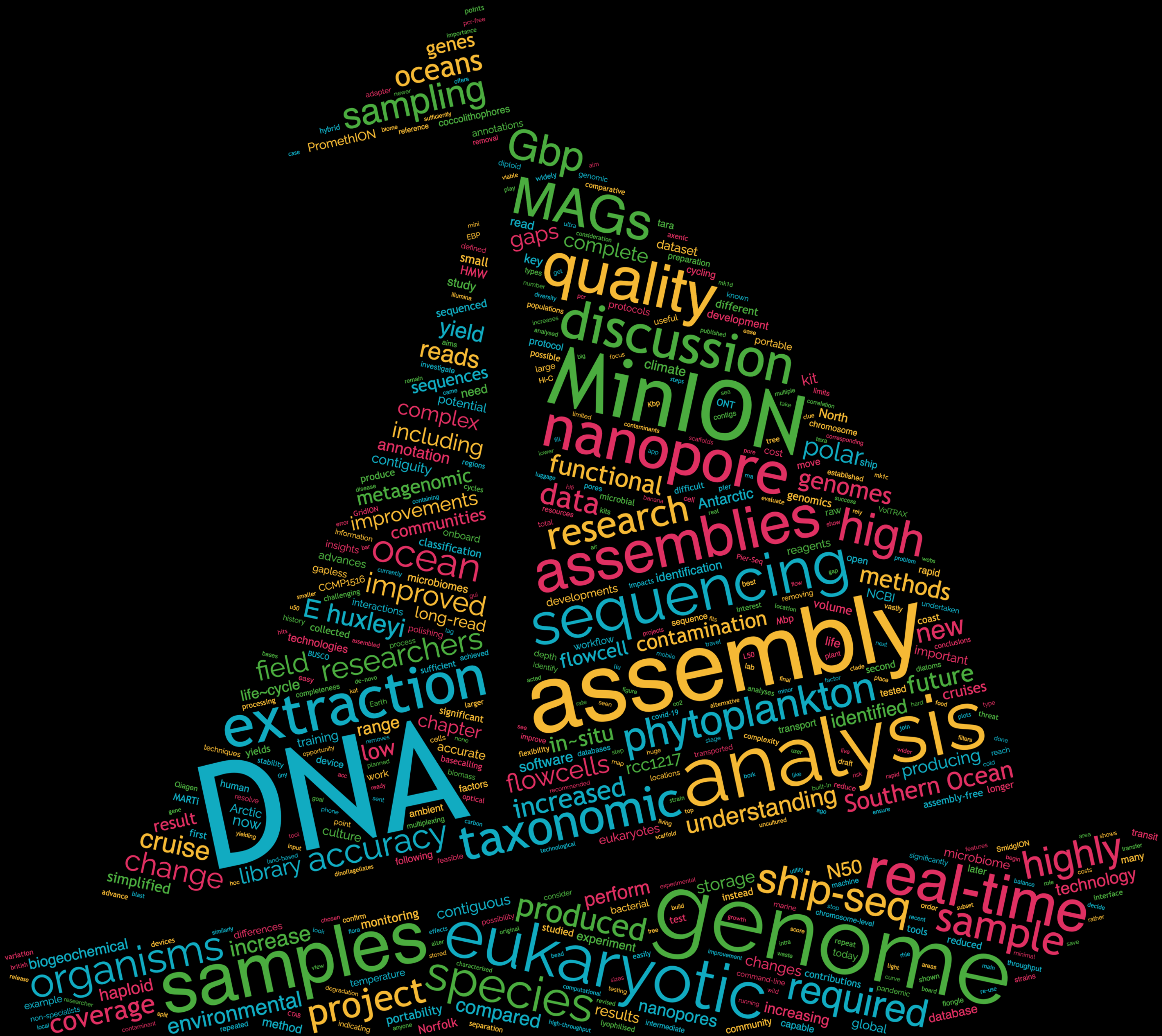
Thank you to all of my office and lab mates at UEA and EI: Jade, Richmal, Amanda, Reuben, Krisztina, Kat, Nicola, Miles, Celia, Ned (especially for your speedy MARTi updates which saved the day), Sam, Mark, Yuxuan, Emily and Roanne for all the tea, cake, chats, and general PhD first-aid.

To my lockdown housemates Bridie, Callum, Claudia, Greg, Sam (and Tom), I wouldn't have wanted to be stuck in the house 23 hours a day for months on end with anyone else, thank you for the film nights, cocktails, and generally helping me stay sane.

Finally, I wouldn't be here without the unfailing support of my parents, grandmother, brother, cat, and extended family and friends - thank you for everything.

This PhD was funded by the Natural Environment Research Council (NERC).

For George Bowes, who always helped me to find out why



DNA
assembly
analysis
genome
MinION
samples
eukaryotic
sequencing
assemblies
nanopore
extraction
quality
discussion
MAGs
taxonomic
phytoplankton
real-time
species
organisms
ship-seq
sample
highly
high
ocean
research
accuracy
improved
researchers
produced
project
oceans
sampling
Gbp
functional
genomes
polar
data
flowcells
increased
Southern Ocean
understanding
required
coverage
environmental
contamination
methods
future
new
in-situ
identified
complete
gaps
complex
including
metagenomic
improvements
communities
E huxleyi
field
change
cruise
library
eukaryotes

1

Introduction

1.1 Ocean microbiomes

1.1.1 Microbes

Microbes are ubiquitous, colonising soil, air, oceans, deserts, frozen tundra, volcanic magma, as well as plants and animals including humans. They include archaea, bacteria, unicellular eukaryotes, and viruses, filling a wide range of niches and roles. It is easy to imagine that such populations have little effect on our day to day lives but in fact microbes are the basis for life on Earth, as they are responsible for the cycling of essential nutrients including Oxygen, Carbon, Nitrogen, and Sulphur, without which plants and animals could not survive. A microbiome is a community of microbes living together in one habitat, characterised by complex interactions between diverse microbe populations. Microbiomes are found in soil, oceans, human guts, and have a huge impact on the habitat in which they live, and the wider world.

Ocean microbiomes are crucial in supporting life not just in the oceans where they perform primary production converting sunlight into energy, but also life on land as they are responsible for approximately half of global O_2 production (Field et al. 1998). Archaea and viruses are involved in nutrient cycling in the oceans, with viruses also responsible for population change through infection of other microbes, and gene mixing within populations (Danovaro et al. 2017). Just as on land, some microbes cause diseases in humans, either through direct infection through exposure to seawater or from consumption of contaminated seafood such as shellfish (Bresnan et al. 2020). Bacterial and eukaryotic ocean microbes are responsible for primary production as well as nutrient cycling, microbes which perform primary production in the oceans are known as phytoplankton (Falkowski 1998).

1.1.2 Phytoplankton

Phytoplankton are found in virtually all aquatic environments on earth, and are responsible for approximately half of global primary production. In ocean environments, they provide the overwhelming majority of O_2 and energy from primary production (Field et al. 1998; Falkowski 1998). Cyanobacteria are the most abundant prokaryotic phytoplankton (Falkowski et al. 2004), and are responsible for around 25% of global primary production. Found in every known aquatic habitat, including oligotrophic and psychrotrophic environments, cyanobacteria are highly diverse with wide ranging physiological adaptations (Bullerjahn and Post 2014) and play a key role in ocean nutrient and energy supply (Mutalipassi et al. 2021).

Eukaryotic phytoplankton are responsible for a significant amount of primary production and are also involved in biogeochemical cycles and form the basis of the food-web as they are preyed on by zooplankton which are in turn preyed on by fish and marine mammals (Smetacek and Nicol 2005). Eukaryotic phytoplankton are also of critical importance in the biogeochemical cycling of key nutrients such as nitrogen, phosphorous, iron and silica (Katz et al. 2004). Through endosymbiotic relationships with nitrogen-fixing bacteria, they provide nitrogen in forms usable by other organisms by conversion to nitrate and nitrite (Foster et al. 2011).

Phytoplankton are also important in the biological carbon pump, see figure 1.1.1, whereby carbon is exported to the sea-bed and thus removed from circulation. This export to the sea-bed is mediated by the sinking of phytoplankton after death and, as such, species with shells which increase sinking such as diatoms and coccolithophores are particularly important (LeMoigne et al. 2015). Eukaryotic phytoplankton mediated CO_2 export is a key determinant of CO_2 levels in the atmosphere and oceans (Katz et al. 2005), with two of the key eukaryotic phytoplankton involved in this process are diatoms and haptophytes.

1.1.3 Phytoplankton evolution

The evolution of eukaryotic phytoplankton involved endosymbiosis events resulting in the incorporation of photosynthetic organelles, see figure 1.1.2. Of the approximately 30,000 species of phytoplankton which have been described, around 90% are eukaryotic. Of these, diatoms account for more than half of the species (Hopes and Mock 2015), although the estimated species numbers are believed to be an underestimate (Guiry 2012). Phytoplankton have been highly

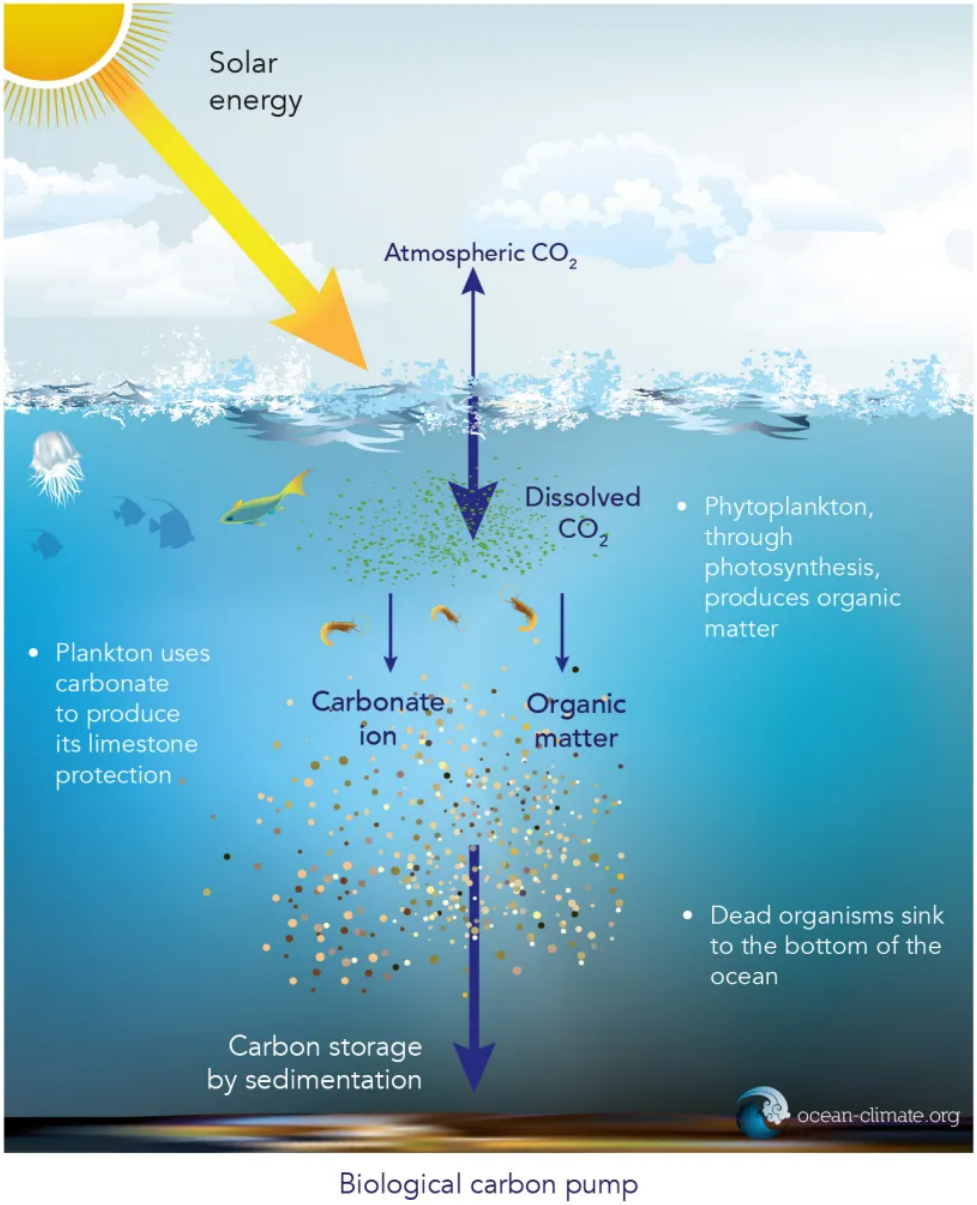


Figure 1.1.1: A schematic showing the biological carbon pump, the processes for uptake and storage of atmospheric CO_2 in the oceans.

successful colonisers of a wide variety of niches, due in large part to their adaptability which is a result of their unusual evolutionary history. There are four major forces of evolution seen in all organisms: mutation, selection, genetic drift and gene flow. In the case of phytoplankton these are present alongside endosymbiosis and vertical and horizontal gene transfer (HGT). Through these processes, phytoplankton have developed with mosaic genomes containing a combination of genes from different organisms (Armbrust 2009).

Three phytoplankton phyla - chlorophyta, glaucophyta, and rhodophyta - resulted from endosymbiosis events in which a cyanobacterium was engulfed by a single-

celled heterotrophic eukaryote, leading to the development of membrane-bound organelles known as plastids (Falkowski et al. 2004). In these new heterotrophic eukaryotes, secondary endosymbiosis occurred, leading to two separate lineages with new plastids formed from green or red algae (Falkowski et al. 2004; Katz et al. 2004; Rynearson and Palenik 2011). From this developed two groups with green plastids, euglenozoa and chlorarachiniophycae, and four red plastid groups, chryptophyta , dinoflagellata, haptophyta (which includes coocolithophores), and heterokontophyta (which inlcudes diatoms).

Tertiary symbiosis has been found in some heterotrophic dinoflagellates, as a result of the engulfing of haptophytes or diatoms by a dinoflagellate (Falkowski et al. 2004). There is evidence that in diatoms, as well as some other red lineages, there was originally a green plastid which was superseded by a red plastid (Frommolt et al. 2008; Moustafa et al. 2009). The reason for this is thought to be a change in ocean Fe levels following the volcanic eruption which triggered the Permian-Triassic mass extinction event. This mass extinction left vacant many ecological niches which the red algal lineages were able to take advantage of, explaining their current dominance (Erwin 1990; Falkowski et al. 2004).

Following endosymbiosis, there is significant gene loss from the engulfed cell, as well as transfer of genes from the engulfed cell to the host cell, with complex interactions between the two (Keeling 2010). Algal genomes are made up of host, plastid and bacterial genes likely acquired by HGT (Bowler et al., 2008; Raymond and Kim, 2012). The complex evolutionary history of phytoplankton has likely equipped them with increased adaptability, allowing them to thrive in a wide range of ecological niches, including extreme environments.

1.1.4 Diatoms

Some of the most dominant phytoplankton groups in modern oceans are diatoms, see figure 1.1.3 and haptophytes. Diatoms are critical components of food-webs and biogeochemical cycles, and account for around 40% of described marine phytoplankton species (Falkowski et al. 2004). They are generally autotrophic and can live in a vast range of ecological niches, from lakes and oceans to the air, and with widely varying lifecycles.

There are planktonic and benthic diatoms which live in fresh and marine water systems; endophytic and endozoic diatoms which live within plants and animals; and those which are epiphytic or epizoic, living on the exterior of plants and animals (Hasle et al. 1996). Diatoms have special cell walls, called frustules, which are made from silica, and are traditionally divided into two groups: centric

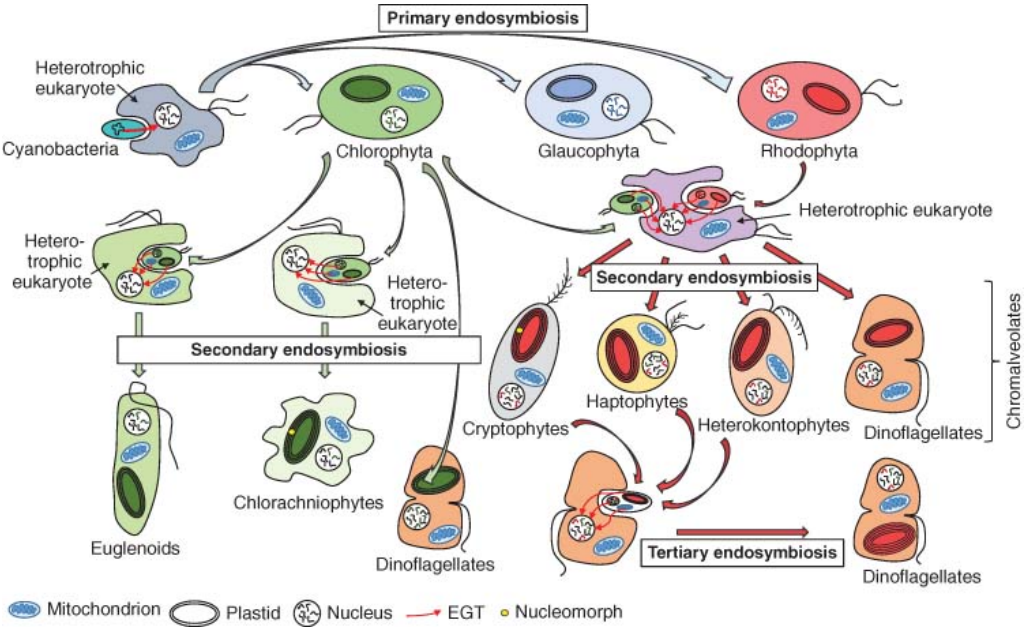


Figure 1.1.2: Evolution of algae according to primary, secondary and tertiary endosymbiotic events. EGT, endosymbiotic gene transfer. Figure copied from Hopes and Mock 2015

and pennate, based on their axis of symmetry. Pennate diatoms are in turn divided into raphid and araphid, depending on whether they have a raphe, or slit through the frustule, which is useful for motility (Kooistra et al. 2007). They are heavily involved in the silica cycle due to their creation of the frustules from dissolved silicic acid in the ocean (Armbrust 2009).

Diatoms appear to have an evolutionary history involving multiple symbiosis events (Moustafa et al. 2009). This has resulted in the retention of bacterial-derived genes which give them the ability to survive in a wide range of environments, including extreme conditions such as polar oceans where they are particularly dominant (Hopes and Mock 2014). Genomic analysis of polar diatom *Fragilariopsis cylindrus* found adaptations to cope with low light, high salinity, low nutrient levels, and low temperatures and the associated reduced enzyme kinetics. *F. cylindrus* was found to have the ability to switch from photosynthesis to cellular respiration during the dark polar winter, and back again when light levels increase in the summer. Other adaptations included high levels of ice-binding proteins in low temperature, high salinity conditions, alongside increased expression of antioxidant proteins and proteins associated with nutrient uptake (Mock et al. 2017).

Recent research has found that diatoms are more abundant, and more important, in oligotrophic waters than had previously been thought, as diatom species use molecular nutrient fixing mechanisms in oligotrophic waters to survive and provide

nutrient cycling. One example is the formation of complexes with symbiotic nitrogen-fixing bacteria, creating conditions which support life for other organisms (Tréguer et al. 2018).

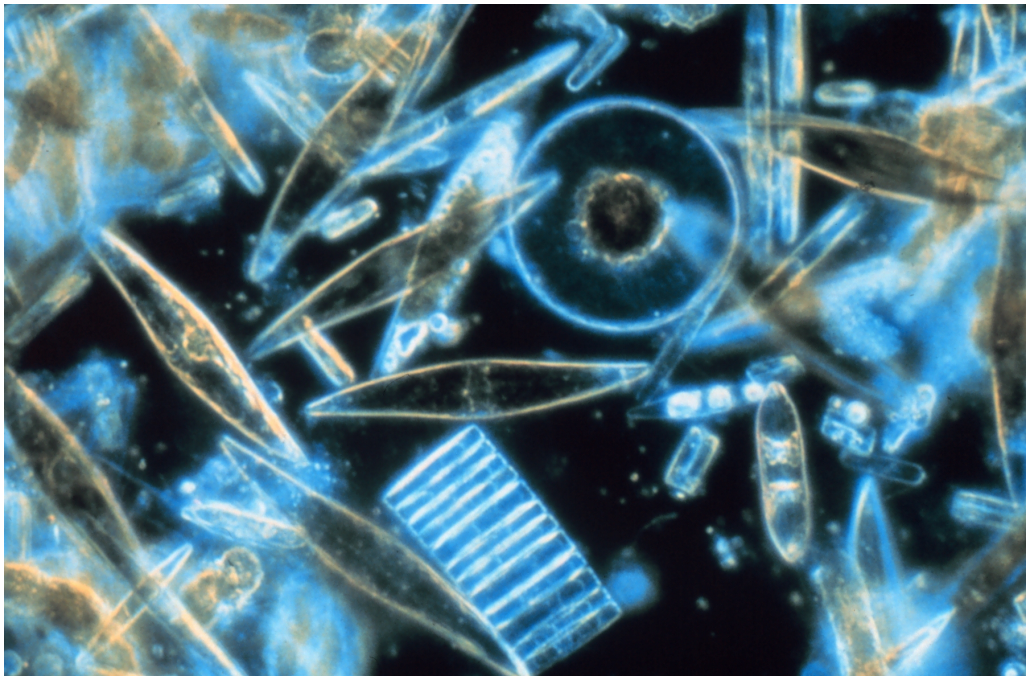


Figure 1.1.3: Diatoms under microscopy. By Prof. Gordon T. Taylor, Stony Brook University - corp2365, NOAA Corps Collection, Public Domain, `https://commons.wikimedia.org/w/index.php?curid=246319`.

1.1.5 Haptophytes

Haptophytes are responsible for 30-50% of total chlorophyll α biomass in modern oceans, some species of which form large blooms which can be observed from space (Liu et al. 2009) - see figure 1.1.4. They generally have 2 flagella and are divided into 2 classes, Pavlovophyceae, and Prymnesiophyceae - based on whether the flagella are unequal, or equal in length (Vargas et al. 2007).

Prymnesiophyceae include coccolithophores which, at certain points in their life cycle, have calcium carbonate coccoliths, or armoured plates, covering their cell membrane. The production of coccoliths is responsible for approximately 50% of oceanic $CaCO_3$ precipitation, while the weight of the coccoliths also results in coccolithophores sinking on death, sequestering carbon on the seabed. This combination of carbon release and sequestration makes coccolithophores important contributors to biogeochemical cycles, and changes to their distribution and life cycle could have wide ranging impacts (Milliman 1993).

Haptophytes have a distinct evolutionary history, representing a branch of the eukaryotic phylogenetic tree. This history is complex, with previous research revealing a mosaic genome (Cuvelier et al. 2010). Genomic analysis of *Emiliania huxleyi*, a temperate coccolithophore, found evidence for the existence of a pan genome, consisting of core genes present in all variants, and genes that were present only in some variants. This genetic differentiation may be responsible for the dominance and adaptability of *E. huxleyi* (Read et al. 2013).

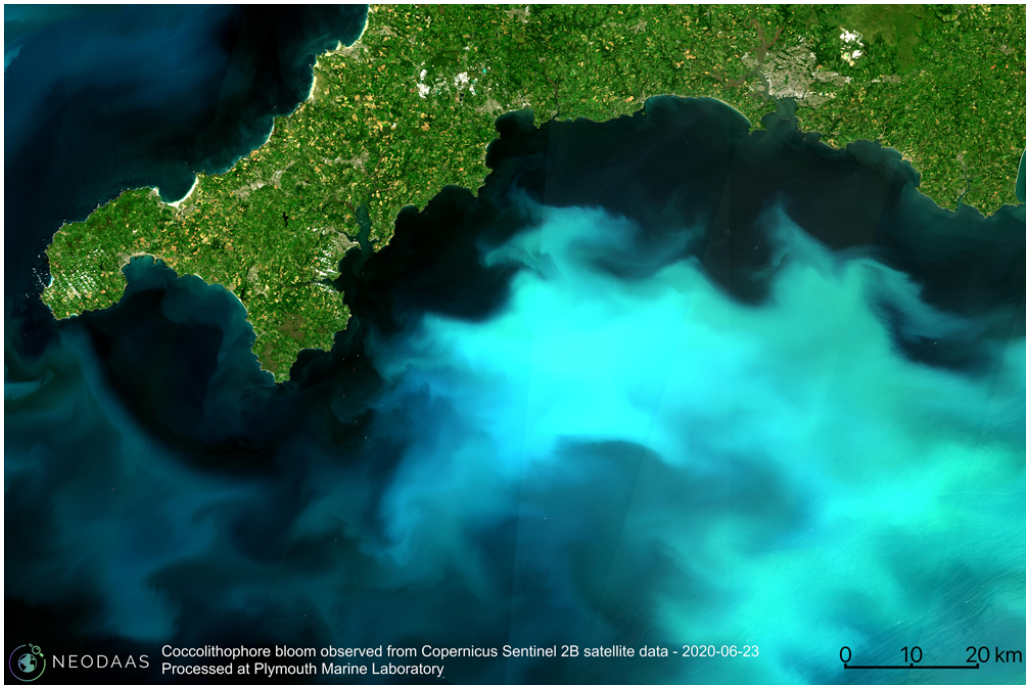


Figure 1.1.4: Coccolithophore bloom observed off the South-West coast of England from Copernicus Sentinel 2B satellite data - 2020-06-23. Processed and published by Plymouth Marine Laboratory

1.1.6 Phytoplankton and climate change

The impacts of climate change, including increasing temperatures, water freshening, acidification, and deoxygenation, will alter the composition and distribution of phytoplankton communities, resulting in shifts in biogeochemical cycles and food-webs (Hays, Richardson, and Robinson 2005). Different conditions found across oceans result in a variety of ecosystems, each with its own unique composition and characteristics.

Warmer water is more stratified than colder water, meaning that tropical oceans have little mixing between the different strata. This results in reduced nutrient flux from one layer to another and can result in reduced nutrient concentration at the surface (Fernández-González et al. 2022). Phytoplankton generally congregate

near the surface as they rely on sunlight for photosynthesis, so reduced nutrient content here limits growth. Changes to currents also affect nutrient availability through upwellings, where deep, nutrient rich water is brought to the surface (Hoegh-Guldberg and Bruno 2010). These features result in areas of high and low productivity, with climate change causing expansion in areas of low productivity (oligotrophic zones) (Polovina, Howell, and Abecassis 2008). Figure 1.1.5 shows stratification compared to temperature in the global oceans.

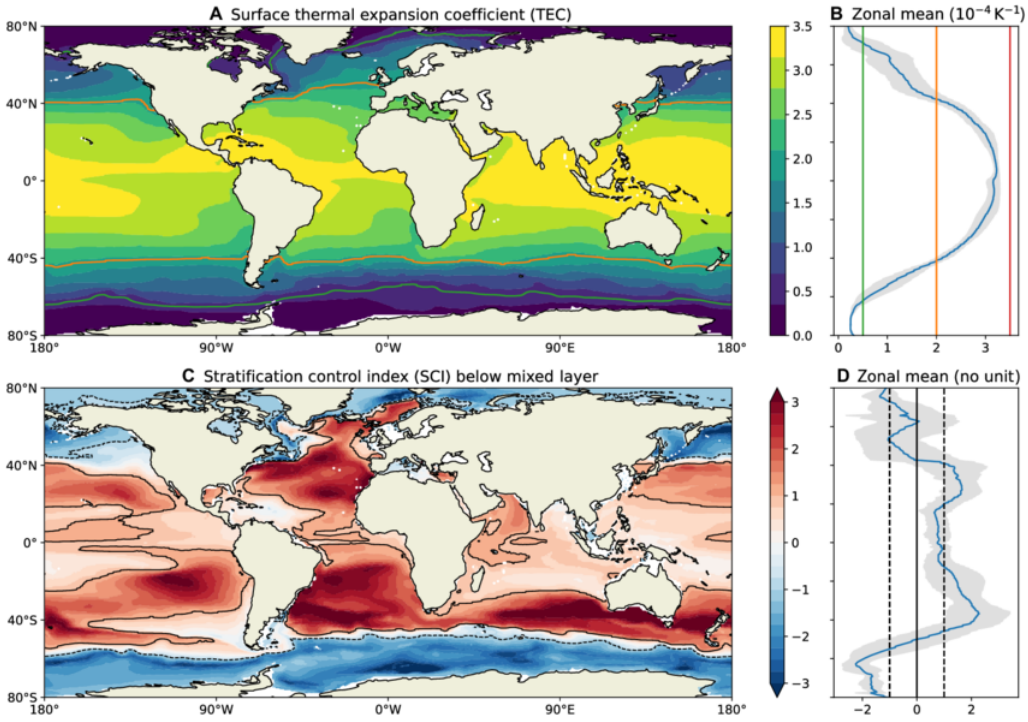


Figure 1.1.5: Stratification control and surface TEC (Thermal Expansion Coefficient) in the ocean. (A) Surface distribution of the TEC, showing a notable correlation with sea surface temperature. (B) Zonal-mean TEC showing an order of magnitude of var. (C) SCI (Stratification Control Index). Blue, stratification dominated by salinity (beta regions); red, dominated by temperature (alpha regions). (D) Zonal-mean SCI. All figures are based on the Estimating the Circulation and Climate of the Ocean (ECCO) state estimate, version 4, release 4 (66). For each year, the SCI was computed on the layer found between 10 and 30 m below the mixed layer for the month of deepest mixed layer. The SCI distribution is obtained by averaging over the 21 years available in ECCO. Figure copied from Roquet et al. 2022

Recent modelling has found that, in response to changing conditions, phytoplankton communities are likely to become increasingly unstable due to altered diversity. This was established through a model analysing the interactions of 35 different phytoplankton, as opposed to models using only 2 or 3 types which give a very limited picture of expected structural change within phytoplankton communities (Henson et al. 2021). Such research shows that we still have relatively little understanding of the complex interactions which underpin ocean microbiomes. This makes it difficult to predict what the impacts

of climate change will be on communities of phytoplankton, and what effect changing phytoplankton populations will produce. To produce improved predictions of the impacts and knock on effects of climate change, therefore, an improved understanding of ocean microbiomes, especially phytoplankton communities, is urgently required.

1.1.7 Metagenomics

Genomic analysis has provided important insights into the evolution, adaptability, and life cycle of critical phytoplankton such as diatoms and haptophytes, but there are many more questions still to answer.

One limiting factor in phytoplankton research so far has been that most analysis has been based on the study of model organisms, such as *Phaeodactylum tricornutum*, *F. cylindrus*, and *E. huxleyi*. This can be extremely useful but has limitations: they have often been grown in the laboratory for many years, undergoing untold genetic changes; they can only provide a small snapshot of phytoplankton evolution and genetics - there are tens of thousands of species of phytoplankton of which only a tiny fraction have been successfully grown in culture and sequenced (Obiol et al. 2020); and the role of interactions between different species, or other microbes cannot be studied through a monoculture. A recent study showed that responses of phytoplankton to changing conditions change depending on whether they are in monoculture or mixed communities (Wolf et al. 2019). Metagenomics is one way to avoid some of these pitfalls and improve our understanding of the intricate networks which shape phytoplankton communities.

1.2 Genome assemblies

1.2.1 The need for phytoplankton genome assemblies

There are very few genome assemblies available for phytoplankton. This limits the research which can be done to understand their evolution, variability, and adaptation to changing conditions, which is of particular relevance given their importance in global biogeochemical cycles and the increasing impacts of climate change. In order to better understand phytoplankton, more publicly available genome assemblies of phytoplankton species are needed. *De novo* genome assembly is the generation of a genome from DNA sequences without an already

assembled reference to compare against. *De novo* genome assemblies are not straight forward to produce, particularly for eukaryotes which tend to have complex genomes.

Only a tiny proportion of the world's organisms have been sequenced, while an even smaller number have a fully assembled genome published. To address this, the Earth BioGenomes Project (EBP) was established to sequence the genomes of all eukaryotic life on earth. This is an extremely ambitious goal, with considerable barriers from sample identification and collection, to production of high quality assemblies. The aim is to produce a complete genome assembly to the highest standard possible for a single member of each of the >9000 eukaryote families, followed by less high quality assemblies for a single species for each genera, and still less high quality assemblies for each of the remaining species (Lewin et al. 2018).

As part of the EBP, the Darwin Tree of Life Project (DTL) undertook to sequence genomes of all 70,000 eukaryotic organisms in Britain and Ireland, starting with 2000 species, which will provide reference genomes for around one third of eukaryotic families and act as a proof of concept for the wider EBP (Life Project Consortium et al. 2022). If completed, the EBP will eventually provide genome assemblies for all eukaryotic phytoplankton species, including diatoms and hatophytes, but the scale of the project means that even if it is successful it will take many years. In the meantime, the work being done by the DTL is advancing the science of DNA extraction, sequencing, and assembly of non-model organisms with physiological and genomic complexities, and could provide a useful basis for the continued development of phytoplankton genomics.

1.2.2 Sequencing history

DNA was first isolated by Friedrich Miescher in 1869 (Miescher-Rüsch 1871) and its structure was famously discovered in 1953 (Watson and Crick 1953), but it would be nearly 20 years before the nucleotides making up the DNA strands could be 'read', through what is now called sequencing. RNA sequencing came first, with the first tRNA sequence, from *Saccharomyces cerevisiae* was produced in 1965 (Holley et al. 1965), followed by the first DNA sequence in 1972 (Jou et al. 1972). The development of Sanger sequencing in 1977, based on the incorporation of chain-terminating dideoxynucleotriphosphates (ddNTPs), (Sanger, Nicklen, and Coulson 1977), revolutionised the field, and it became the most widely used method for the next 30 years. Sanger sequencing could sequence reads of 500-600 bp, with high levels of accuracy at 99.99%. It was

used to produce the first completely sequenced genome, of bacteriophage PhiX in 1978 (Sanger et al. 1978), alongside a range of other genomics projects, including the first fully sequenced prokaryote genome, *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995), and the first full eukaryotic genome, the nematode *Caenorhabditis elegans* in 1998 (Sequencing Consortium* 1998). Higher level organisms followed, with the first plant genome *Arabidopsis thaliania* in 2000 (Initiative 2000), and the first drafts of the human genome published in 2001 (Consortium et al. 2001), before the genome assembly was declared complete in 2003 (Waterston, Lander, and Sulston 2003).

For large, complex genomes, such plants and humans, Sanger sequencing is slow and costly, with the Human Genome Project taking 13 years and costing over £10 million. This limited the use of DNA sequencing to model organisms and large, well-funded projects, with little prospect of sequencing expanding into general use. *De novo* genome assembly of Sanger sequences is largely achieved using overlap-layout-consensus (OLC) algorithms, as used by the Celera assembler. These use all-against-all alignments to identify overlapping reads and building a graph of nodes representing reads and edges linking nodes where the reads overlap over a sequence longer than a minimum cut-off. The layout step involves finding the Hamiltonian path, where each node is visited once, followed by finding the consensus based on the multiple sequence alignments. This is a computationally intensive and time consuming method which works best on long-read sequences.

1.2.3 Next-Generation Sequencing: Advances and challenges

Next-generation sequencing (NGS) increased sequencing yields and reduced costs, allowing high throughput sequencing for the first time. In a similar principle to Sanger sequencing, NGS relies on the use of DNA polymerase to incorporate labelled dNTPs into a strand of DNA (Behjati and Tarpey 2013). The first NGS technology became commercially available in 2005, with the 454 Pyrosequencer used to sequence part of the Neanderthal genome in 2006 (Noonan et al. 2006). This was followed by the Solexa Genome Analyzer (now Illumina) in 2006 which offered researchers the ability to sequence up to 1 Gbp per sequencing run. Illumina sequencing works through the use of random fragmentation followed by 3‘ and 5‘ adapter ligation, before fragments are amplified using PCR and purified. This is followed by a process of cluster generation, where fragments bind complementary oligonucleotides and undergo bridge amplification prior to sequencing. Sequencing is performed through the detection of reversible terminator-bound dNTPs as they are incorporated into template DNA strands,

figure 1.2.1 shows an overview of the Illumina sequencing process. NGS can now sequence reads of 150-300bp, with terabases of data produced per run (Bronner et al. 2013).

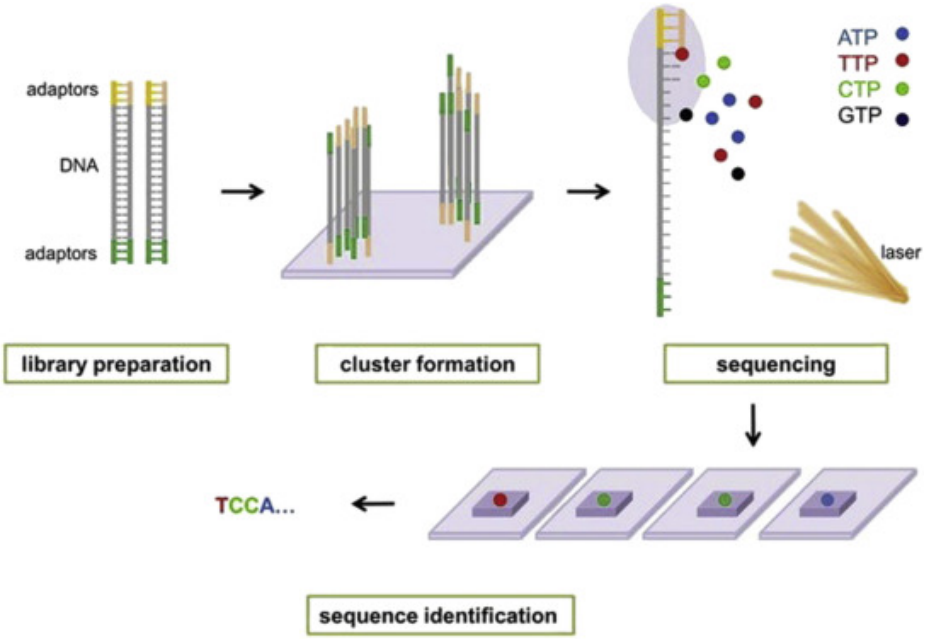


Figure 1.2.1: An overview of Illumina sequencing, showing library preparation, cluster generation, sequencing, and detection. From (Zhou and Li 2015)

Reducing costs combined with increasing throughput and read length transformed genomics, leading to a rapid increase in the number of genomes published, as well as allowing for other purposes, such as whole-genome metagenomics and amplicon-based identification of microbiomes through 16S rRNA, 18S rRNA, and ITS rRNA sequencing. This resulted in a large increase in the genomic data available on phytoplankton, including the publication of coccolithophore and diatom genome assemblies such as the *E. huxleyi* (Read et al. 2013) and *F. cylindrus* genomes (Mock et al. 2017), and a range of metagenomic research including the Tara Oceans project, which used Illumina sequencing to study plankton at a large number of study sites. This constitutes the most wide-ranging and in-depth study of ocean microbes to date (Bork et al. 2015).

Tara Oceans sampled at 210 sites, down to 2000m in depth over four years from 2009 to 2013, covering all of the major oceanic zones. Ribosomal RNA (rRNA) was used for identification of species in the photic zone while metagenomic analyses allowed for profiling of viruses and prokaryotes which resulted in key insights into the determinants of community composition. From this, it was found that temperature is the main driver of community composition, further underlining the importance of including microbes in climate change models (Lima-Mendez et al. 2015). An investigation of the Agulhas rings, where there

is complex nitrogen-cycling due to strong vertical mixing, found that there is selective pressure for nitrogen-fixing which leads to a population shift. The Tara Oceans project demonstrated the complex interactions between population diversity and physical and chemical oceanographic structures. It also highlighted the need for high-resolution taxonomic studies - previous fossil studies in the area had failed to find the choke-point of diversity, because the acquisition of a nitrogen-fixing gene would not make one strain morphologically distinguishable from another (Villar et al. 2015).

For *de novo* assembly of NGS reads, a widely used assembly method is based on de Bruijn graphs, seen in the assembler Velvet (Zerbino and Birney 2008). This breaks the sequencing reads down into k-mers of a fixed length, and creates a graph where each node represents a k-mer, and edges represent an overlap of the k-mer length minus one. A eulerian path can then be found which travels over each edge exactly once, which is less computationally expensive than performing the all-vs all alignments necessary to find the OLC Hamiltonian path. This type of assembly is highly effective for short, simple genomes, but breaking down the sequences into k-mers results in a loss of positional information and it can be very hard to resolve repeat sequences. One study in 2011 found that NGS *de novo* genome assemblies were generally shorter than the reference genome by around 16%, missing many repeat sequences, and 99% of validated duplicated sequences (Alkan, Sajjadian, and Eichler 2010). While read lengths have increased, NGS genome assemblies still have limited success for complex genome assembly (Wang et al. 2021).

There are techniques available which can enhance the performance of NGS in these cases, for example chromatin conformation capture techniques inclduing Hi-C (Belton et al. 2012). Hi-C provides information about the 3D conformation of chromosome and genome structures inside the nucleus, providing information on chromosomal arrangements, as well as an overall picture of the genome structure. Hi-C works through covalently linking spatially adjacent chromatin (the DNA, RNA, and proteins which make up chromosomes in eukaryotes) by cross-linking with formaldehyde, before digestion with a restriction enzyme and filling in the resulting ends with nucleotides, including a biotinylated residue. These fragments undergo ligation under high dilution, creating conditions which favour chimeric ligation events, joining fragments that are cross-linked. This produces a sample of ligation products which were originally physically close together in the nucelus, with the junction marked with biotin. The probability that 2 chromatin fragments were spatially adjacent in the nucelous is proportional to the abundance of the ligation products. This information can be used for the investigation of chromatin itself and improving understanding of interaction between genes and regulatory elements,

as well as giving an overview of the overall genomic structure of eukaryotic cells. Hi-C data can be used to improve genome assembly, by giving information about chromosomal arrangements, which has improved genome assemblies of a variety of organisms, including human, mouse, and *Drosophila* (Burton et al. 2013). The data has also been used in research into chromosomal structure, evolution, and disease, including detection of chromosomal rearrangements in tumours (Harewood et al. 2017).

A key aspect of standard NGS protocols is the use of polymerase chain reaction (PCR) (Mullis and Faloona 1987), which is a method for producing millions of copies of a specific region of DNA from a single strand through the use of DNA polymerase. This is widely used in NGS sequencing, although PCR-free protocols such as Illumina Tru-Seq can be used instead. PCR is extremely useful in diagnostic and forensic applications, which aim to detect the presence of a known sequence (Zhu et al. 2020). It is widely used in metagenomic analysis where conserved gene sequences are amplified and sequenced, before being matched against databases for identification of bacteria (16S rRNA), eukaryotes (18S rRNA), and bacteria, fungi and archaea (ITS rRNA). This is a quick, fairly cheap, and highly accurate, method for determining the composition of a metagenomic sample (Scholz, Lo, and Chain 2012). Downsides of PCR include GC bias, polar phytoplankton genomes, for example, often have a high AT content, resulting in biases when PCR is used. Similarly, *E. huxleyi* has an unusually high GC content in its genome, which also runs the risk of introducing bias if PCR is used (Chen et al. 2013). This limits the utility of NGS sequencing for many phytoplankton; protocols such as Illumina Tru-Seq offer PCR-free library preparation but sequencing still requires bridge amplification, so are not truly PCR-free. Additionally, PCR can only amplify known sequences, which can bias metagenomic studies towards sequences which have already been identified, and it is extremely sensitive to contamination.

1.2.4 Long-read sequencing

An alternative to short-read NGS technologies is long-read sequencing. Long-read sequencing technology offers the ability to sequence strands of DNA which are tens or hundreds of thousands of basepairs long as a single molecule, which is extremely useful for genome assembly. Long-read sequencing does not rely on PCR and can also be used to detect epigenetic modifications, such as methylation, which can give researchers information about changes to genomes in response to changing conditions (Dijk et al. 2018). Methylation is a chemical modification to a DNA base which prevents gene expression, either by preventing

the binding of transcription factors, or through recruitment of proteins which actively repress genes. Methylation detection can be used to identify changes within or between populations which can indicate which genes are involved in certain niche-specific adaptations, and show how populations adapt to different environmental conditions. One recent study which analysed the methylation of metagenome-derived genomes identified a methylation motif in the *Pelagibacter* genome which is related to the control of the cell cycle and in host-pathogen interactions with pelagiphage viruses, indicating that epigenetics could be a useful technique for investigating marine microbiome communty interactions and growth (Seong et al. 2022).

One long-read sequencing technology is PacBio single molecule real-time (SMRT) sequencing (Eid et al. 2009). SMRT sequencing works by using DNA polymerase to incorporate fluorescently labelled dNTPs into a DNA strand, each dNTP emits a different light frequency which allows for real time detection of nucleotides based on wavelength. HiFi sequencing (Wenger et al. 2019) is a further development of SMRT, using circular consensus sequencing and offers sequencing of reads on average 10-25 kbp long, with high accuracy at above 99%, albeit slightly lower than is achieved with NGS or Sanger sequencing. HiFi sequencing has been used to provide high coverage long-read sequencing data to improve genome assemblies for model organisms mouse and corn, alongside two complex genomes, an octoploid strawberry and diploid frog (Hon et al. 2020). PacBio offers high throughput long-read sequencing, with yields of up to 250 Gbp of sequencing data in a 30 hour run.

Another long-read sequencing option is nanopore sequencing, offered by Oxford Nanopore Technologies (ONT). As shown in figure 1.2.2, nanopore sequencing works by feeding a DNA strand through a biological pore one base at a time, identifying bases by their unique ionic current (Deamer, Akeson, and Branton 2016), which means that in theory an entire DNA strand could be sequenced, with accuracy rates, although lower than NGS or SMRT HiFi, now over 99% (Cuber et al. 2022). The use of ionic current detection also allows for the identification of bases which have been modified, for example via methylation, allowing for epigentic studies. An extension of this is the potential for the identification of non-canonical bases (Liu et al. 2021) which is becoming increasingly effective, due to improvements in basecalling accuracy and increased data available for training (White and Hesselberth 2022). An adapted basecalling model trained to recognise the potentially small differences between altered bases is required, but this could lead to improved genome assemblies for organisms such as dinoflagellates which present with non-canonical bases, and for which it is hard to produce accurately basecalled sequencing data (Mendez et al. 2015).

A particular draw of nanopore sequencing is the revolutionary portability and low cost of ONT sequencing instruments. ONT produce the MinION which is a small (10 x 3 x 2 cm, 90 g). portable DNA sequencing machine, which is powered by a USB connection and can be run on a laptop computer. The MinION Mk1c is a slightly larger (14 x 11 x 3 cm) all in one device with integrated basecalling and analysis computing capabilities. Each uses flow cells which can produce up to 50 Gbp. For non-portable high throughput sequencing, the GridION which allows 5 MinION flowcells to be run at once, while the PromethION offers yields of up to 14 terabytes of data per run on 48 high throughput flowcells. The combination of low cost long-read sequencing and portable platforms with real-time analysis makes nanopore sequencing a particularly promising option for sequencing phytoplankton communities.

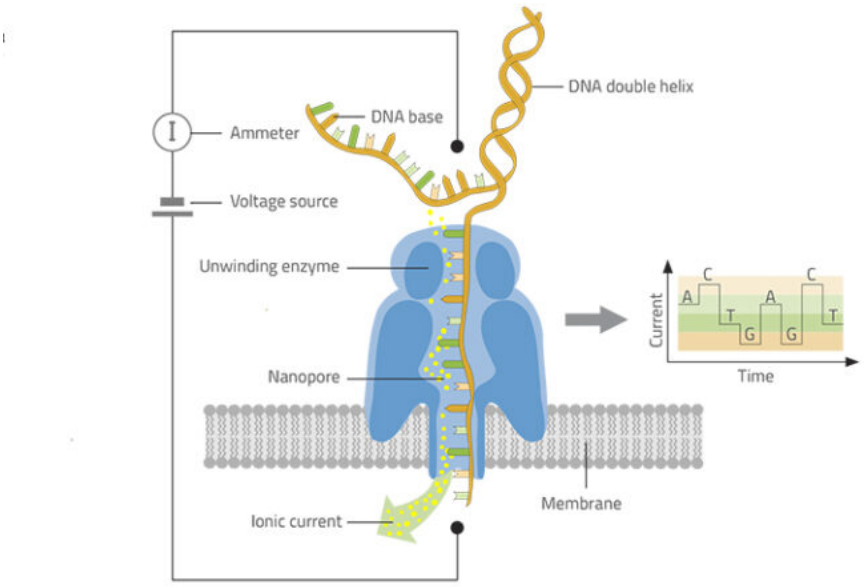


Figure 1.2.2: Illustration of how a nanopore DNA sequencer works. A DNA double strand is separated into single strands by a DNA helicase enzyme and current is applied to the membrane to pass the strand through the nanopore and the changes to the ionic current as each base passes through is measured and translated into the DNA sequence. From Kerstin Göpfrich from Science in School `https://www.scienceinschool.org/article/2018/decoding-dna-pocket-sized-sequencer`

Genome assembly methods for long-read genomes are often developed from OLC assembly algorithms used for Sanger sequencing, since these were developed for long-read sequences. Canu (Koren et al. 2017) is a particularly effective long-read assembler which reduces the impacts of the increased error rate in long-read sequencing by first correcting reads before trimming them to remove adapters and other potentially low-quality sequences. The corrected, trimmed reads are then assembled into contigs based on a modified OLC strategy, with consensus sequences and alternate paths generated. This produces high quality assemblies of complex genomes from long-read sequences, particularly for high

coverage which allows for greater use of long reads. The recommended minimum coverage is 30-60x. Nanopore sequencing has been used to produce high quality genomes with high levels of contiguity for a wide range of species, including a recent telomere-to-telomere highly complex banana genome assembly (Belser et al. 2021).

1.3 Nanopore sequencing for phytoplankton

The portability offered by the ONT MinION opens a wide range of opportunities for field-based DNA sequencing, and has so far been demonstrated on the International Space Station (Castro-Wallace et al. 2016), on glaciers in the Arctic and Antarctic (Edwards et al. 2016; Johnson et al. 2017), and in disease outbreaks during the 2015 ebola virus outbreak (Quick et al. 2016), and in Brazil during the zika outbreak (Faria et al. 2016). The low cost, high portability model has democratised DNA sequencing - individual laboratories can purchase a MinION kit and produce full genome assemblies for $1000 or less, and perform in-house sequencing, rather than relying on sequencing centres. This has particular benefits for research into non-model organisms where there are generally fewer researchers and funding, limiting access to expensive high-throughput sequencing.

1.3.1 Nanopore sequencing for *de novo* genome assembly

Benefits of long-read sequencing include that it becomes easier to assemble genomes with longer reads. Long-reads can often span repeat sequences and transposable elements, allowing these to be resolved, and the organisation of reads into chromosomes to create a genome assembly is simplified allowing for the production of higher quality genome assemblies. In order to harness long-read sequencing data to produce genome assemblies, assembly algorithms had to be developed and adapted to work with relatively error-prone long-reads. Once a genome assembly has been produced, the challenge facing researchers is how to determine whether a genome assembly is of good quality, whether it is an accurate representation of the genome of the organism which was sequenced, and whether it is complete.

The quality and contiguity of a genome assembly can be assessed through a range of analyses including length metrics such as N50, and Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Manni et al. 2021). The N50 of an assembly is the size of contig or scaffold above which 50% of the genome

is assembled. Alongside the number of sequences taken to reach the N50, this can be a useful indication of how much of the genome has been assembled, and how long the resulting contigs are. BUSCO analysis is a non-annotation based measure of completeness of the genome, using single-copy orthologs, which are highly conserved genes which are present in at least 90% of species within a chosen category, such as eukaryotes. These genes are matched against an assembly, without requiring gene annotation, and the presence, absence, fragmentation, or duplication of each is reported. Good assemblies are generally defined as those where less than 1% of the genes are fragmented, and at least 90% are complete. Duplication can indicate polyploidy or contamination. Further methods for assessing genome assembly quality include the k-mer analysis toolkit (KAT) (Mapleson et al. 2016) which breaks down the raw reads into k-mers and searches for them in the assembly, producing a graph which shows the number of kmers which are absent, present, or present multiple times, plotted as k-mer multiplicity against the number of distinct k-mers. This can help to identify where elements of the sequence have been duplicated, missed or contain contamination.

The relatively high error rate of nanopore sequencing means that hybrid assemblies are often more effective than genome assemblies based on nanopore reads alone. Hybrid assemblies may include polishing steps using NGS reads to correct errors in the nanopore reads without sacrificing the benefits of long-read sequqencing, or for more complex genomes, nanopore sequencing data can also be supplemented with long-range data, such as Hi-C or optical mapping (Michael and VanBuren 2020). These can result in the production of highly contiguous, highly accurate genome assemblies. A recent telomere-to-telomere assembly of the human X chromosome was produced through a combination of nanopore sequencing, SMRT sequencing, linked-read sequencing, and optical mapping. Through manual curation, and significant financial and time investment, this allowed researchers to produce a highly contiguous genome (Miga et al. 2020).

Long-read sequencing offers the prospect of producing improved metagenomic assembled genomes (MAGs), which are full genome assemblies produced from metagenomic data. This could rapidly increase the number of genome assemblies available for under-represented groups such as phytoplankton, allowing for comparative analyses between strains and species, and improving our understanding of their evolution and life-cycles. It would also increase the depth of genomics databases, helping to improve alignment-based identification methods. Long reads could be especially useful for producing MAGs, because they are easier to assemble than short-reads, particularly with metagenomic

samples where there is little indication as to which reads go together to create each assembly (Moss, Maghini, and Bhatt 2020).

1.3.2 *In situ* sequencing and analysis

Using the MinION *in situ* for DNA sequencing of metagenomic marine samples to investigate polar phytoplankton could improve our understanding of their communities, and complex interactions, and provide sequencing data for genome assemblies and functional analysis. The portability and limited equipment required for sample preparation would allow immediate sequencing of samples, thus preventing degradation and community alteration (Fonseca et al. 2016). The rapidity would allow for the use of DNA sequence data to be used in determining the progress of a research cruise, based on the identification of species of interest. It would also allow for resources to be saved as researchers would know whether sampling had been effective. Long-read nanopore sequencing of metagenomic samples has been recently used to investigate marine viral communities. A long-read, low-input, metagenomic sequencing approach for observation of natural marine viral communities was developed, called VirION. It was found that abundance estimation from long-read sequencing was comparable to that of short-read sequencing, and that long-read sequencing also captured more diversity and resulted in increased contig lengths, making genome assembly easier. Long-read sequencing also captured entire genomic islands, and bridged regions which are difficult to assemble from short-read sequencing (Warwick-Dugdale et al. 2018). This indicates that nanopore sequencing of marine phytoplankton could be effective, although there are aspects which could be more challenging for phytoplankton than for viruses.

Long-read sequencing requires high molecular weight (HMW) DNA, which is DNA that is in long fragments. Extraction of HMW DNA is more complicated for phytoplankton than for samples such as those used in the VirION project, because viruses have no cell walls or frustules and so gentle detergents can be used for cell lysis which do not damage DNA. In many diatom species, the DNA has a high AT content which, as AT base pairs are attached with a double bond compared to the triple bond seen in GC base pairs, increases the DNA fragility (Rynearson and Palenik 2011). This means that it is often difficult to achieve HMW DNA extraction from polar diatoms, especially as the frustule must also be broken before DNA can be extracted. Other challenges include acquiring sufficient quantities of DNA from dilute ocean samples. Tens of litres of seawater might be required to yield 400 ng of DNA, the minimum for ONT MinION sequencing, which is time-consuming and resource-intensive (Warwick-Dugdale

et al. 2018). VirION relies on amplification of DNA samples using PCR to achieve a sufficient DNA yield, but for phytoplankton this may not be suitable due to GC bias. For nanopore sequencing of important phytoplankton communities, therefore, it is important to establish effective HMW DNA extraction techniques and to determine whether sufficient DNA can be acquired given the particularly low phytoplankton density in polar oceans.

Successful phytoplankton DNA extraction techniques tend to be based on methods developed for the extraction of DNA from plants which use chemicals to breakdown cell wall polysaccharides and remove them from the final DNA sample (Rogers and Bendich 1994). These generally work very well for phytoplankton which have no hard outer layer, but can be ineffective for those which do, including diatoms, coccolithophores, and dinoflagellates. As these groups contain some of the most important phytoplankton, it is important to ensure that DNA extraction protocols are effective for them. Methods for lysis of these groups include freeze-thaw, sonication, grinding in liquid N_2, harsh chemicals, and bead-beating. All of these methods can result in DNA degradation, which can be particularly problematic given the fragility of DNA in some diatoms. Freeze-thaw and grinding in liquid N_2 are often ineffective at disrupting cells, while sonication and harsh chemicals are particularly likely to result in significant DNA degradation. Bead beating has been widely used in phytoplankton DNA extraction in recent years, particularly in 16S rRNA, 18S rRNA, and ITS rRNA sequencing which does not rely on HMW DNA (Yuan, Li, and Lin 2015). For long-read sequencing requiring HMW DNA, extraction protocols such as CTAB remain the most widely used, although there is some evidence that bead beating does not have much effect on molecular weight, which could allow for its use in phytoplankton sequencing experiments. This could have the effect of simplifying DNA extraction for phytoplankton samples, particularly in the field, by allowing for the use of fewer toxic reagents and more rapid, streamlined protocols requiring less equipment.

Many metagenomic phytoplankton studies have relied on 16S rRNA, 18S rRNA, and ITS rRNA sequencing as described previously. This is an effective method for determining composition of a sample but it allows only for identification, without genome assembly, comparative analysis, or phylogenetic study. Further limitations include high similarity between 16S or 18S genes in some species (Scholz, Lo, and Chain 2012). Whole genome sequencing (WGS) offers an alternative which can be used for more in depth analyses as well as species identification. There are challenges associated with metagenomic WGS, however, including taxonomic identification. Any amount of DNA sequencing data is of little use if the organisms sampled cannot be identified, and in metagenomic samples

it can be difficult to determine which species each read belongs to (Nielsen et al. 2014). Many approaches for identification of species from sequencing data are based on alignment of reads to a reference genome. This is effective where there is a good quality genome assembly for the organism in question, but is less useful where there are limited reference genomes available, such as with phytoplankton where only a fraction of the species present in a sample are likely to be represented in a genomic database. In the case of a species such as *E. huxleyi* which has high levels of variance between strains, there is only one publicly available reference genome which is likely to read to strains being misclassified and potential insights into their distributions could be missed.

Metagenomic identification of eukaryotes is further complicated by the dominance of bacterial sequences in genomic databases, while the complex evolutionary history of diatoms and haptophytes can lead to erroneous identification of bacteria from their genomes. An alternative to reference genome alignment-based classification is metagenomic binning, where sequences are grouped by their ancestry, which allows known and unknown genomes to be separated and analysed further. This is generally achieved through assumptions of coabundance and genome similarities of closely related organisms. Metagenomic analysis of eukaryotes is more complex than for prokaryotes, due to a number of factors including larger genome size with increased complexity, alongside a lack of analysis tools geared towards eukaryotes.

Real-time sequencing analysis of nanopore sequencing data has many benefits for researchers, including quality assessment to avoid wasting resources, the ability to end a sequencing run once a chosen parameter has been met, and for rapid diagnostics. For research into phytoplankton and ocean microbiomes, particularly polar oceans, real-time sequencing would allow for results to be produced within hours of sampling in the field. This would give researchers the opportunity to assess sampling sufficiency, determine whether further sampling in a given area would be useful, and allow evidence-based decision making on future sampling locations. DNA sequencing samples collected from polar oceans during research cruises has been difficult because it can take months for the samples to get back to the lab, with potential DNA degradation occurring during storage, and the potential for lab analysis to find that the samples taken are insufficient or otherwise unsuitable. Real-time sequencing analysis can be carried out using the MinION Mk1C with built-in basecalling software, and a laptop computer running an analysis tool such as MARTi `https://marti.cyverseuk.org` (Leggett et al. 2018) which uses alignment of the basecalled reads against the BLAST-nt database to provide real-time taxonomic identification.

1.4 Outlook and statement of aims

Nanopore sequencing offers significant advantages for research into ocean microbiomes and phytoplankton communities. The long reads can help to produce more, higher quality genome assemblies, while the portable, real-time sequencing offered by the ONT MinION will allow researchers to answer questions such as which species are present in locations of interest, how they are responding to climate change, how continued climate change is likely to affect them, and how changes to their populations are likely to affect the wider ecology. Another exciting prospect is the potential for nanopore sequencing as a tool for citizen science and public engagement, as it is low cost and simple to use and could allow groups to produce data with a high spatial or temporal resolution. There are challenges associated with this, however, including the difficulty of extracting HMW DNA from phytoplankton, and metagenomic samples, and the production of high quality eukaryotic phytoplankton assemblies due to their highly complex genomes. Other challenges include the development of a workflow for in field sampling, DNA extraction, sequencing, and analysis which can be easily adapted to a range of conditions and used by non-specialists.

To address these questions, the aim of this project was to establish the utility of nanopore sequencing for the study of ocean microbiomes, particularly eukaryotic phytoplankton. This was split into the following goals:

1. Genome assembly
    (a) The production of a high quality *E. huxleyi* genome assembly from nanopore sequencing data
2. *In situ* sequencing of polar microbes
    (a) *In situ* metagenomic nanopore sequencing and real-time taxonomic classification of polar ocean samples onboard a research cruise in the Southern Ocean
    (b) Land-based nanopore sequencing of polar ocean samples for benchmarking, assembly, and functional annotation.
3. Investigation of the potential of nanopore sequencing as a tool for citizen science and public engagement in ocean health
    (a) Development and field-based testing of an improved workflow for *in situ* nanopore sequencing and real-time taxonomic classification of ocean microbes by citizen scientists
    (b) Land-based nanopore sequencing of additional samples collected across a time course to assess changes in populations over time.

References

Alkan, C., S. Sajjadian, and E. E. Eichler (2010). “Limitations of Next-Generation Genome Sequence Assembly”. In: *Nature Methods* 8.1, pp. 61–65. DOI: 10.1038/nmeth.1527.

Armbrust, E. V. (May 2009). “The life of diatoms in the world’s oceans”. In: *Nature* 459.7244, pp. 185–192. ISSN: 0028-0836. DOI: 10.1038/nature08057.

Behjati, S. and P. S. Tarpey (2013). “What Is Next Generation Sequencing?” In: *Archives of disease in childhood - Education &; practice edition* 98.6, pp. 236–238. DOI: 10.1136/archdischild-2013-304340.

Belser, C., F.-C. Baurens, B. Noel, G. Martin, C. Cruaud, B. Istace, N. Yahiaoui, K. Labadie, E. Hřibová, J. Doležel, A. Lemainque, P. Wincker, A. D’Hont, and J.-M. Aury (2021). “Telomere-To-Telomere Gapless Chromosomes of Banana Using Nanopore Sequencing”. In: *Communications Biology* 4.1, p. 1047. DOI: 10.1038/s42003-021-02559-3.

Belton, J.-M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker (2012). “Hi-C: a Comprehensive Technique To Capture the Conformation of Genomes”. In: *Methods* 58.3, pp. 268–276. DOI: 10.1016/j.ymeth.2012.05.001.

Bork, P., C. Bowler, C. De Vargas, G. Gorsky, E. Karsenti, and P. Wincker (2015). “Tara Oceans studies plankton at Planetary scale”. In: *Science* 348.6237, p. 873. ISSN: 10959203. DOI: 10.1126/science.aac5605.

Bresnan, E., C. Baker-Austin, C. Campos, K. Davidson, M. Edwards, A. Hall, A. McKinney, and A. Turner (2020). “Impacts of climate change on human health, HABs and bathing waters, relevant to the coastal and marine environment around the UK”. In: *MCCIP Science Review 2020*, pp. 521–545. DOI: 10.14465/2020.arc22.hhe.

Bronner, I. F., M. A. Quail, D. J. Turner, and H. Swerdlow (2013). “Improved Protocols for Illumina Sequencing”. In: *Current Protocols in Human Genetics* 79.1, nil. DOI: 10.1002/0471142905.hg1802s79.

Bullerjahn, G. S. and A. F. Post (2014). *Physiology and molecular biology of aquatic cyanobacteria*.

Burton, J. N., A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure (2013). “Chromosome-Scale Scaffolding of De Novo Genome Assemblies Based on Chromatin Interactions”. In: *Nature Biotechnology* 31.12, pp. 1119–1125. DOI: 10.1038/nbt.2727.

Castro-Wallace, S. L., C. Y. Chiu, K. K. John, S. E. Stahl, K. H. Rubins, A. B. R. McIntyre, J. P. Dworkin, M. L. Lupisella, D. J. Smith, D. J. Botkin, T. A. Stephenson, S. Juul, D. J. Turner, F. Izquierdo, S. Federman, D. Stryke, S. Somasekar, N. Alexander, G. Yu, C. Mason, and A. S. Burton (Sept. 2016).

"Nanopore DNA Sequencing and Genome Assembly on the International Space Station". In: *bioRxiv*, p. 077651. DOI: 10.1101/077651.

Chen, Y.-C., T. Liu, C.-H. Yu, T.-Y. Chiang, and C.-C. Hwang (Apr. 2013). "Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly". In: *PLoS ONE* 8.4. Ed. by Y. Xu, e62856. DOI: 10.1371/journal.pone.0062856.

Consortium, I. H. G. S. et al. (2001). "Initial Sequencing and Analysis of the Human Genome". In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Cuber, P., D. Chooneea, C. Geeves, S. Salatino, T. J. Creedy, C. Griffin, L. Sivess, I. Barnes, B. Price, and R. Misra (2022). *Comparing the accuracy and efficiency of third generation DNA barcode sequencing: Oxford Nanopore Technologies versus Pacific Biosciences*. DOI: 10.1101/2022.07.13.499863.

Cuvelier, M. L., A. E. Allen, A. Monier, J. P. McCrow, M. Messié, S. G. Tringe, T. Woyke, R. M. Welsh, T. Ishoey, J.-H. Lee, B. J. Binder, C. L. DuPont, M. Latasa, C. Guigand, K. R. Buck, J. Hilton, M. Thiagarajan, E. Caler, B. Read, R. S. Lasken, F. P. Chavez, and A. Z. Worden (2010). "Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton". In: *Proceedings of the National Academy of Sciences* 107.33, pp. 14679–14684. ISSN: 0027-8424. DOI: 10.1073/pnas.1001665107. eprint: https://www.pnas.org/content/107/33/14679.full.pdf.

Danovaro, R., E. Rastelli, C. Corinaldesi, M. Tangherlini, and A. Dell'Anno (2017). "Marine archaea and archaeal viruses under global change". In: *F1000Research* 6.

Deamer, D., M. Akeson, and D. Branton (2016). "Three Decades of Nanopore Sequencing". In: *Nature Biotechnology* 34.5, pp. 518–524. DOI: 10.1038/nbt.3423.

Dijk, E. L. van, Y. Jaszczyszyn, D. Naquin, and C. Thermes (2018). "The Third Revolution in Sequencing Technology". In: *Trends in Genetics* 34.9, pp. 666–681. DOI: 10.1016/j.tig.2018.05.008.

Edwards, A., A. R. Debbonaire, S. M. Nicholls, S. M. Rassner, B. Sattler, J. M. Cook, T. Davy, L. A. Mur, and A. J. Hodson (Sept. 2016). "In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota". In: DOI: 10.1101/073965.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner (2009). "Real-Time Dna Sequencing From

Single Polymerase Molecules". In: *Science* 323.5910, pp. 133–138. DOI: 10.1126/science.1162986.

Erwin, D. (1990). "The end-Permian mass extinction". In: *Annual Review of Ecology and Systematics* 21.1, pp. 69–91.

Falkowski, P. G. (1998). "Biogeochemical Controls and Feedbacks on Ocean Primary Production". In: *Science* 281.5374, pp. 200–206. DOI: 10.1126/science.281.5374.200.

Falkowski, P. G., M. E. Katz, A. H. Knoll, A. Quigg, J. a. Raven, O. Schofield, and F. J. R. Taylor (2004). "The Evolution of Modern Eukaryotic Phytoplankton". In: *Science* 305.July, pp. 354–360. ISSN: 1095-9203. DOI: 10.1126/science.1095964.

Faria, N. R., R. d. S. d. S. Azevedo, M. U. G. Kraemer, R. Souza, M. S. Cunha, S. C. Hill, J. Theze, M. B. Bonsall, T. A. Bowden, I. Rissanen, I. M. Rocco, J. S. Nogueira, A. Y. Maeda, F. G. d. S. Vasami, F. L. d. L. Macedo, A. Suzuki, S. G. Rodrigues, A. C. R. Cruz, B. T. Nunes, D. B. d. A. Medeiros, D. S. G. Rodrigues, A. L. N. Queiroz, E. V. P. d. Silva, D. F. Henriques, E. S. T. da Rosa, C. S. de Oliveira, L. C. Martins, H. B. Vasconcelos, L. M. N. Casseb, D. d. B. Simith, J. P. Messina, L. Abade, J. Lourenco, L. C. J. Alcantara, M. M. d. Lima, M. Giovanetti, S. I. Hay, R. S. de Oliveira, P. d. S. Lemos, L. F. d. Oliveira, C. P. S. de Lima, S. P. da Silva, J. M. d. Vasconcelos, L. Franco, J. F. Cardoso, J. L. d. S. G. Vianez-Junior, D. Mir, G. Bello, E. Delatorre, K. Khan, M. Creatore, G. E. Coelho, W. K. de Oliveira, R. Tesh, O. G. Pybus, M. R. T. Nunes, and P. F. C. Vasconcelos (Mar. 2016). "Zika virus in the Americas: Early epidemiological and genetic findings". In: *Science* 352.6283, pp. 345–349. DOI: 10.1126/science.aaf5036.

Fernández-González, C., G. A. Tarran, N. Schuback, E. M. S. Woodward, J. Arístegui, and E. Marañón (2022). "Phytoplankton responses to changing temperature and nutrient availability are consistent across the tropical and subtropical Atlantic". In: *Communications Biology* 5.1, p. 1035.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski (1998). "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components". In: *Science* 281.5374, pp. 237–240. ISSN: 0036-8075. DOI: 10.1126/science.281.5374.237. eprint: http://science.sciencemag.org/content/281/5374/237.full.pdf.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and

J. C. Venter (1995). “Whole-Genome Random Sequencing and Assembly of haemophilus Influenzae Rd”. In: *Science* 269.5223, pp. 496–512. DOI: 10.1126/science.7542800.

Fonseca, R. R. da, A. Albrechtsen, G. E. Themudo, J. Ramos-Madrigal, J. A. Sibbesen, L. Maretty, M. L. Zepeda-Mendoza, P. F. Campos, R. Heller, and R. J. Pereira (Dec. 2016). “Next-generation biology: Sequencing and data analysis approaches for non-model organisms”. In: *Marine Genomics* 30, pp. 3–13. DOI: 10.1016/j.margen.2016.04.012.

Foster, R. A., M. M. M. Kuypers, T. Vagner, R. W. Paerl, N. Musat, and J. P. Zehr (Mar. 2011). “Nitrogen fixation and transfer in open ocean diatom–cyanobacterial symbioses”. In: *The ISME Journal* 5.9, pp. 1484–1493. DOI: 10.1038/ismej.2011.26.

Frommolt, R., S. Werner, H. Paulsen, R. Goss, C. Wilhelm, S. Zauner, U. G. Maier, A. R. Grossman, D. Bhattacharya, and M. Lohr (2008). “Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis”. In: *Molecular Biology and Evolution* 25.12, pp. 2653–2667.

Guiry, M. D. (Oct. 2012). “HOW MANY SPECIES OF ALGAE ARE THERE?” In: *Journal of Phycology* 48.5, pp. 1057–1063. ISSN: 00223646. DOI: 10.1111/j.1529-8817.2012.01222.x.

Harewood, L., K. Kishore, M. D. Eldridge, S. Wingett, D. Pearson, S. Schoenfelder, V. P. Collins, and P. Fraser (2017). “Hi-C As a Tool for Precise Detection and Characterisation of Chromosomal Rearrangements and Copy Number Variation in Human Tumours”. In: *Genome Biology* 18.1, p. 125. DOI: 10.1186/s13059-017-1253-8.

Hasle, G. R., E. E. Syvertsen, K. A. Steidinger, K. Tangen, and C. R. Tomas (1996). *Identifying marine diatoms and dinoflagellates*. Elsevier.

Hays, G. C., A. J. Richardson, and C. Robinson (2005). “Climate change and marine plankton”. In: *Trends in ecology & evolution* 20.6, pp. 337–344.

Henson, S. A., B. Cael, S. R. Allen, and S. Dutkiewicz (2021). “Future phytoplankton diversity in a changing climate”. In: *Nature communications* 12.1, p. 5372.

Hoegh-Guldberg, O. and J. F. Bruno (June 2010). “The Impact of Climate Change on the World’s Marine Ecosystems”. In: *Science* 328.5985, pp. 1523–1528. DOI: 10.1126/science.1189930.

Holley, R. W., J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir (1965). “Structure of a Ribonucleic Acid”. In: *Science* 147.3664, pp. 1462–1465. DOI: 10.1126/science.147.3664.1462.

Hon, T., K. Mars, G. Young, Y.-C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, D. Kudrna, M. A. Hardigan, C. C. Steiner, S. J. Knapp, D. Ware, B. Shapiro, P. Peluso, and D. R. Rank (2020). “Highly Accurate Long-Read Hifi Sequencing

Data for Five Complex Genomes". In: *Scientific Data* 7.1, p. 399. DOI: 10.1038/s41597-020-00743-4.

Hopes, A. and T. Mock (2014). "Diatoms: glass-dwelling dynamos". In: *Microbiology Today* 41.February, pp. 20–23.

– (2015). "Evolution of microalgae and their adaptations in different marine ecosystems". In: *eLS*, pp. 1–9.

Initiative, T. A. G. (2000). "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana". In: *Nature* 408.6814, pp. 796–815. DOI: 10.1038/35048692.

Johnson, S. S., E. Zaikova, D. S. Goerlitz, Y. Bai, and S. W. Tighe (Apr. 2017). "Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer". In: *Journal of Biomolecular Techniques : JBT*, jbt.17–2801–009. DOI: 10.7171/jbt.17-2801-009.

Jou, W. M., G. Haegeman, M. Ysebaert, and W. Fiers (1972). "Nucleotide Sequence of the Gene Coding for the Bacteriophage Ms2 Coat Protein". In: *Nature* 237.5350, pp. 82–88. DOI: 10.1038/237082a0.

Katz, M. E., Z. V. Finkel, D. Grzebyk, A. H. Knoll, and P. G. Falkowski (2004). "Evolutionary trajectories and biogeochemical impacts of marine eukaryotic phytoplankton". In: *Annu. Rev. Ecol. Evol. Syst.* 35, pp. 523–556.

Katz, M. E., J. D. Wright, K. G. Miller, B. S. Cramer, K. Fennel, and P. G. Falkowski (June 2005). "Biological overprint of the geological carbon cycle". In: *Marine Geology* 217.3-4, pp. 323–338. DOI: 10.1016/j.margeo.2004.08.005.

Keeling, P. J. (2010). "The endosymbiotic origin, diversification and fate of plastids". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365.1541, pp. 729–748.

Kooistra, W. H., R. Gersonde, L. K. Medlin, and D. G. Mann (2007). "The origin and evolution of the diatoms: their adaptation to a planktonic existence". In: *Evolution of primary producers in the sea*, pp. 207–249.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". In: *Genome research* 27.5, pp. 722–736. DOI: 10.1101/gr.215087.116.

Leggett, R. M., C. Alcon-Giner, D. Heavens, S. Caim, T. C. Brook, M. Kujawska, S. Martin, L. Hoyles, P. Clarke, L. J. Hall, and M. D. Clark (2018). "Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics". In: *bioRxiv*. DOI: 10.1101/180406. eprint: https://www.biorxiv.org/content/early/2018/10/12/180406.full.pdf.

LeMoigne, F. A. C., A. J. Poulton, S. A. Henson, C. J. Daniels, G. M. Fragoso, E. Mitchell, S. Richier, B. C. Russell, H. E. K. Smith, G. A. Tarling, J. R. Young, and M. Zubkov (June 2015). "Carbon export efficiency and phytoplankton community composition in the Atlantic sector of the Arctic Ocean". In: *Journal*

of Geophysical Research: Oceans 120.6, pp. 3896–3912. DOI: 10.1002/2015jc010700.

Lewin, H. A., G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, and G. Zhang (2018). "Earth Biogenome Project: Sequencing Life for the Future of Life". In: *Proceedings of the National Academy of Sciences* 115.17, pp. 4325–4333. DOI: 10.1073/pnas.1720115115.

Life Project Consortium, T. D. T. of, M. Blaxter, N. Mieszkowska, F. D. Palma, P. Holland, R. Durbin, T. Richards, M. Berriman, P. Kersey, P. Hollingsworth, W. Wilson, A. Twyford, E. Gaya, M. Lawniczak, O. Lewis, G. Broad, K. Howe, M. Hart, P. Flicek, and I. Barnes (2022). "Sequence Locally, Think Globally: the Darwin Tree of Life Project". In: *Proceedings of the National Academy of Sciences* 119.4, nil. DOI: 10.1073/pnas.2115642118.

Lima-Mendez, G., K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-espinosa, S. Roux, F. Vincent, and L. Bittner (2015). "Determinants of community structure in the grobal plankton interactome". In: *Science* 348.6237, p. 1262073. ISSN: 0036-8075. DOI: 10.1126/science.1262073. arXiv: arXiv:1011.1669v3.

Liu, H., I. Probert, J. Uitz, H. Claustre, S. Aris-Brosou, M. Frada, F. Not, and C. de Vargas (2009). "Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans". In: *Proceedings of the national academy of sciences* 106.31, pp. 12803–12808.

Liu, Y., W. Rosikiewicz, Z. Pan, N. Jillette, P. Wang, A. Taghbalout, J. Foox, C. Mason, M. Carroll, A. Cheng, et al. (2021). "DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation". In: *Genome biology* 22.1, pp. 1–33.

Manni, M., M. R. Berkeley, M. Seppey, and E. M. Zdobnov (2021). "Busco: Assessing Genomic Data Quality and Beyond". In: *Current Protocols* 1.12, nil. DOI: 10.1002/cpz1.323.

Mapleson, D., G. G. Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo (2016). "Kat: a K-Mer Analysis Toolkit To Quality Control Ngs Datasets and Genome Assemblies". In: *Bioinformatics* nil.nil, btw663. DOI: 10.1093/bioinformatics/btw663.

Mendez, G. S., C. F. Delwiche, K. E. Apt, and J. C. Lippmeier (2015). "Dinoflagellate gene structure and intron splice sites in a genomic tandem array". In: *Journal of Eukaryotic Microbiology* 62.5, pp. 679–687.

Michael, T. P. and R. VanBuren (2020). "Building Near-Complete Plant Genomes". In: *Current Opinion in Plant Biology* 54.nil, pp. 26–33. DOI: 10.1016/j.pbi.2019.12.009.

Miescher-Rüsch, F. (1871). *Ueber die chemische Zusammensetzung der Eiterzellen*.

Miga, K. H., S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T. A. G. Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy (2020). "Telomere-To-Telomere Assembly of a Complete Human X Chromosome". In: *Nature* 585.7823, pp. 79–84. DOI: 10.1038/s41586-020-2547-7.

Milliman, J. D. (1993). "Production and accumulation of calcium carbonate in the ocean: Budget of a nonsteady state". In: *Global Biogeochemical Cycles* 7.4, pp. 927–957.

Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland, B. J. Ward, A. E. Allen, C. L. Dupont, S. Frickenhaus, F. Maumus, A. Veluchamy, T. Wu, K. W. Barry, A. Falciatore, M. I. Ferrante, A. E. Fortunato, G. Glöckner, A. Gruber, R. Hipkin, M. G. Janech, P. G. Kroth, F. Leese, E. A. Lindquist, B. R. Lyon, J. Martin, C. Mayer, M. Parker, H. Quesneville, J. A. Raymond, C. Uhlig, R. E. Valas, K. U. Valentin, A. Z. Worden, E. V. Armbrust, M. D. Clark, C. Bowler, B. R. Green, V. Moulton, C. van Oosterhout, and I. V. Grigoriev (Jan. 2017). "Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus". In: *Nature* 541.7638, pp. 536–540. ISSN: 0028-0836. DOI: 10.1038/nature20803.

Moss, E. L., D. G. Maghini, and A. S. Bhatt (2020). "Complete, Closed Bacterial Genomes From Microbiomes Using Nanopore Sequencing". In: *Nature Biotechnology* 38.6, pp. 701–707. DOI: 10.1038/s41587-020-0422-6.

Moustafa, A., B. Beszteri, U. G. Maier, C. Bowler, K. Valentin, and D. Bhattacharya (June 2009). "Genomic footprints of a cryptic plastid endosymbiosis in diatoms". In: *Science* 324.5935, pp. 1724–1726. ISSN: 00368075. DOI: 10.1126/science.1172983.

Mullis, K. B. and F. A. Faloona (1987). "[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction". In: *Methods in Enzymology*. Methods in Enzymology. Elsevier, pp. 335–350. DOI: 10.1016/0076-6879(87)55023-6.

Mutalipassi, M., G. Riccio, V. Mazzella, C. Galasso, E. Somma, A. Chiarore, D. de Pascale, and V. Zupo (2021). "Symbioses of cyanobacteria in marine environments: Ecological insights and biotechnological perspectives". In: *Marine Drugs* 19.4, p. 227.

Nielsen, H. B., M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. L. Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Q. dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. L. Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, and S. D. Ehrlich (July 2014). "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes". In: *Nature Biotechnology* 32.8, pp. 822–828. DOI: 10.1038/nbt.2939.

Noonan, J. P., G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J. K. Pritchard, and E. M. Rubin (2006). "Sequencing and Analysis of Neanderthal Genomic Dna". In: *Science* 314.5802, pp. 1113–1118. DOI: 10.1126/science.1131412.

Obiol, A., C. R. Giner, P. Sánchez, C. M. Duarte, S. G. Acinas, and R. Massana (2020). "A metagenomic assessment of microbial eukaryotic diversity in the global ocean". In: *Molecular Ecology Resources* n/a, pp. 1–14. DOI: 10.1111/1755-0998.13147. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13147.

Polovina, J. J., E. A. Howell, and M. Abecassis (Feb. 2008). "Ocean's least productive waters are expanding". In: *Geophysical Research Letters* 35.3. DOI: 10.1029/2007gl031745.

Quick, J., N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, et al. (2016). "Real-time, portable genome sequencing for Ebola surveillance". In: *Nature* 530.7589, p. 228. DOI: 10.1038/nature16996.

Read, B. A., E. huxleyi Annotation Consortium, J. Kegel, M. J. Klute, A. Kuo, S. C. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J.-M. Claverie, S. Frickenhaus, K. Gonzalez, E. K. Herman, Y.-C. Lin, J. Napier, H. Ogata, A. F. Sarno, J. Shmutz, D. Schroeder, C. de Vargas, F. Verret, P. von Dassow, K. Valentin, Y. V. de Peer, G. Wheeler, J. B. Dacks, C. F. Delwiche, S. T. Dyhrman, G. Glöckner, U. John, T. Richards, A. Z. Worden, X. Zhang, and I. V. Grigoriev (2013). "Pan Genome of the Phytoplankton Emiliania Underpins Its Global Distribution". In: *Nature* 499.7457, pp. 209–213. DOI: 10.1038/nature12221.

Rogers, S. O. and A. J. Bendich (1994). "Extraction of total cellular DNA from plants, algae and fungi". In: *Plant Molecular Biology Manual*. Ed. by S. B. Gelvin and R. A. Schilperoort. Plant Molecular Biology Manual. Dordrecht: Springer

Netherlands, pp. 183–190. ISBN: 978-94-011-0511-8. DOI: 10.1007/978-94-011-0511-8_12.

Roquet, F., D. Ferreira, R. Caneill, D. Schlesinger, and G. Madec (2022). "Unique thermal expansion properties of water key to the formation of sea ice on Earth". In: *Science Advances* 8.46, eabq0793.

Rynearson, T. A. and B. Palenik (2011). "Learning to read the oceans: genomics of marine phytoplankton". In: *Advances in marine biology*. Vol. 60. Elsevier, pp. 1–39. DOI: 10.1016/B978-0-12-385529-9.00001-9.

Sanger, F., A. Coulson, T. Friedmann, G. Air, B. Barrell, N. Brown, J. Fiddes, C. Hutchison, P. Slocombe, and M. Smith (1978). "The Nucleotide Sequence of Bacteriophage Φx174". In: *Journal of Molecular Biology* 125.2, pp. 225–246. DOI: 10.1016/0022-2836(78)90346-7.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). "Dna Sequencing With Chain-Terminating Inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

Scholz, M. B., C.-C. Lo, and P. S. Chain (2012). "Next Generation Sequencing and Bioinformatic Bottlenecks: the Current State of Metagenomic Data Analysis". In: *Current Opinion in Biotechnology* 23.1, pp. 9–15. DOI: 10.1016/j.copbio.2011.11.013.

Seong, H. J., S. Roux, C. Y. Hwang, and W. J. Sul (2022). "Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics". In: *Microbiome* 10.1, p. 157.

Sequencing Consortium*, T. C. elegans (1998). "Genome Sequence of the Nematode c. Elegans : a Platform for Investigating Biology". In: *Science* 282.5396, pp. 2012–2018. DOI: 10.1126/science.282.5396.2012.

Smetacek, V. and S. Nicol (Sept. 2005). "Polar ocean ecosystems in a changing world". In: *Nature* 437.7057, pp. 362–368. ISSN: 1476-4687. DOI: 10.1038/nature04161.

Tréguer, P., C. Bowler, B. Moriceau, S. Dutkiewicz, M. Gehlen, O. Aumont, L. Bittner, R. Dugdale, Z. Finkel, D. Iudicone, O. Jahn, L. Guidi, M. Lasbleiz, K. Leblanc, M. Levy, and P. Pondaven (Jan. 2018). "Influence of diatom diversity on the ocean biological carbon pump". In: *Nature Geoscience* 11.1, pp. 27–37. ISSN: 1752-0894. DOI: 10.1038/s41561-017-0028-x.

Vargas, C. de, M.-P. Aubry, I. Probert, and J. Young (2007). "Origin and evolution of coccolithophores: from coastal hunters to oceanic farmers". In: *Evolution of primary producers in the sea*. Elsevier, pp. 251–285.

Villar, E., G. K. Farrant, M. Follows, L. Garczarek, S. Speich, S. Audic, L. Bittner, B. Blanke, J. R. Brum, C. Brunet, R. Casotti, A. Chase, J. R. Dolan, O. Jahn, J.-L. Jamet, H. Le Goff, C. Lepoivre, S. Malviya, E. Pelletier, S. Roux, S. Santini, E. Scalco, S. M. Schwenck, A. Tanaka, P. Testor, A. Zingone, C. Dimier, M. Picheral, E. Boss, C. De Vargas, G. Gorsky, H. Ogata, M. B. Sullivan, S.

Sunagawa, P. Wincker, E. Karsenti, C. Bowler, F. Not, P. Hingamp, D. Iudicone, and J.-b. Romagnan (2015). "Environmental characteristics of Agulhas rings affect interocean plankton transport". In: *Science* 348.6237, pp. 1–12.

Wang, P., F. Meng, B. M. Moore, and S.-H. Shiu (2021). "Impact of Short-Read Sequencing on the Misassembly of a Plant Genome". In: *BMC Genomics* 22.1, p. 99. DOI: 10.1186/s12864-021-07397-5.

Warwick-Dugdale, J., N. Solonenko, K. Moore, L. Chittick, A. C. Gregory, M. J. Allen, M. B. Sullivan, and B. Temperton (June 2018). "Long-read metagenomics reveals cryptic and abundant marine viruses". In: *bioRxiv*. DOI: 10.1101/345041.

Waterston, R. H., E. S. Lander, and J. E. Sulston (2003). "More on the Sequencing of the Human Genome". In: *Proceedings of the National Academy of Sciences* 100.6, pp. 3022–3024. DOI: 10.1073/pnas.0634129100.

Watson, J. D. and F. H. C. Crick (1953). "Molecular Structure of Nucleic Acids: a Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356, pp. 737–738. DOI: 10.1038/171737a0.

Wenger, A. M., P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, and M. W. Hunkapiller (2019). "Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome". In: *Nature Biotechnology* 37.10, pp. 1155–1162. DOI: 10.1038/s41587-019-0217-9.

White, L. K. and J. R. Hesselberth (2022). "Modification mapping by nanopore sequencing". In: *Frontiers in Genetics* 13, p. 1037134.

Wolf, K. K. E., E. Romanelli, B. Rost, U. John, S. Collins, H. Weigand, and C. J. M. Hoppe (2019). "Company Matters: the Presence of Other Genotypes Alters Traits and Intraspecific Selection in an Arctic Diatom Under Climate Change". In: *Global Change Biology* 25.9, pp. 2869–2884. DOI: 10.1111/gcb.14675.

Yuan, J., M. Li, and S. Lin (2015). "An Improved Dna Extraction Method for Efficient and Quantitative Recovery of Phytoplankton Diversity in Natural Assemblages". In: *PLOS ONE* 10.7, e0133060. DOI: 10.1371/journal.pone.0133060.

Zerbino, D. R. and E. Birney (2008). "Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs". In: *Genome Research* 18.5, pp. 821–829. DOI: 10.1101/gr.074492.107.

Zhou, X. and Y. Li (2015). "Techniques for oral microbiology". In: *Atlas of Oral Microbiology. Elsevier Inc*, pp. 15–40.

Zhu, H., H. Zhang, Y. Xu, S. Laššáková, M. Korabečná, and P. Neužil (2020). "Pcr Past, Present and Future". In: *BioTechniques* 69.4, pp. 317–325. DOI: 10.2144/btn-2020-0057.

2

Producing an *Emiliania huxleyi* genome assembly

2.1 Introduction

This chapter presents work on the production of a genome assembly for the haploid *Emiliania huxleyi* strain RCC1217. The genome assembly is a hybrid assembly incorporating nanopore, Hi-C data, and Illumina data. I performed the cell culturing, DNA extraction, nanopore sequencing and initial assembly from RCC1217 cultures grown in the Mock Lab at the University of East Anglia. The Hi-C scaffolding was carried out by Dovetail Genomics in 2019 using this nanopore assembly with cultures provided by Dmitry Filatov at the University of Oxford. I then used Illumina sequencing data produced by the Genomic Pipelines group at the Earlham Institute in 2015. I then performed quality assessment, and contamination removal to produce the final assembly.

2.1.1 Importance of *Emiliania huxleyi* and the need for genome assemblies

Coccolithophores are important calcifying phytoplankton species, which have made well documented impacts on the global climate, spanning over 200 million years and linking atmospheric, geospheric, and hydrospheric conditions (Paasche 2001). The relationship between coccolithophores and the carbon cycle is complicated as they are responsible for both CO_2 production and sequestration, through calcification, and their sinking to the sea-bed on death respectively (Read et al. 2013); in some conditions, they can be responsible for up to 20% of total fixation of atmospheric carbon (Poulton et al. 2007). The most studied coccolithophore is *Emiliania huxleyi*, see figure 2.1.1, which has recently been reclassified as *Gephyrocapsa huxleyi* (Bendif et al. 2019). Globally distributed with frequent blooms in UK waters, it plays a critical role in global primary production, biogeochemical cycles, DMS production, and as the base of

the marine food-web (LeMoigne et al. 2015). *E. huxleyi* is also of increasing interest as an indicator of climate change, due to recent poleward expansion meaning it can now be found in the Southern Ocean (Winter et al. 2014). Along with other coccolithophores it is at risk from climate change, which is resulting in increased ocean acidification and the resulting fall in the $CaCO_3$ saturation rate (Zhang and Cao 2016).

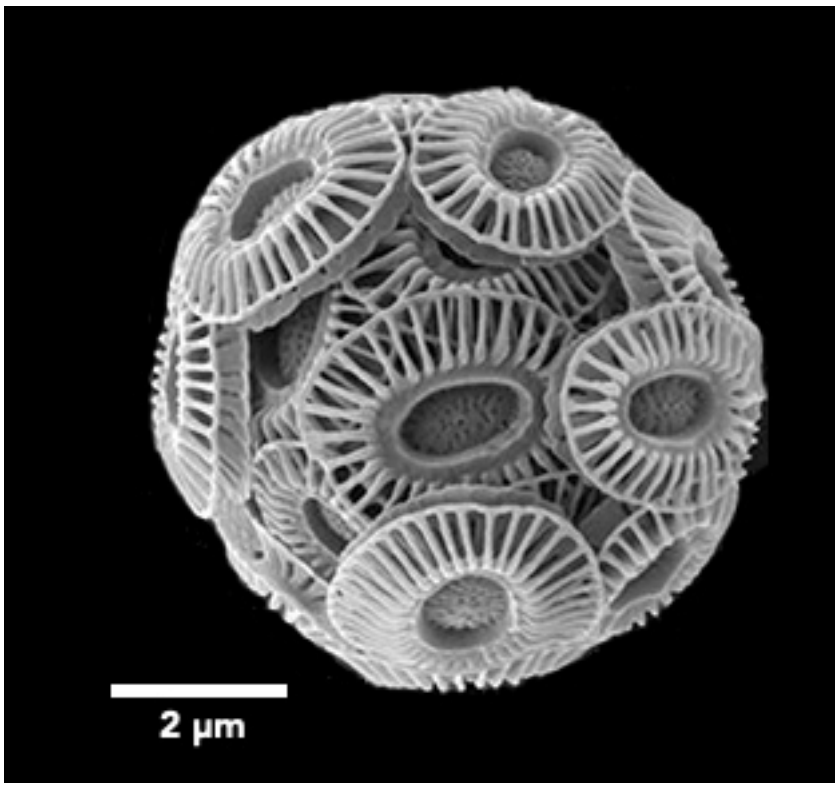


Figure 2.1.1: *A scanning electron micrograph of a single E. huxleyi cell. By Dr. Jeremy Young, University College London - Extracted from this Commons file, CC BY-SA 4.0,* https://commons.wikimedia.org/w/index.php?curid=10975103

Coccolithophores are part of the haptophyte clade, they have a complex evolutionary history involving symbiosis events and possibly horizontal gene transfer, resulting in a mosaic genome (Cuvelier et al. 2010). Analysis of the only publicly available coccolithophore genome assembly, the diploid *E. huxleyi* CCMP1516, alongside sequencing data from additional isolates, showed that *E. huxleyi* has a core set of genes which are present in all strains, alongside a set of genes which are found only in some strains. This is supported by previous studies which found wide phenotypic variation between *E. huxleyi* strains (Reid et al. 2011), and it is this variability which is believed to underpin the adaptability and persistent dominance of *E. huxleyi*.

Advances in genome sequencing are increasingly allowing whole-genome analysis of previously unstudied species, leading to new developments in

phytoplankton research. Recent analysis based on whole-genome sequencing of *Gephyrocapsa* species, including *E. huxleyi*, has provided insights into *Gephyrocapsa* evolution and speciation within the genus. Sequencing data was used to produce full phylogenies for 5 *Gephyrocapsa* species, which formed a consensus phylogeny for the genus and confirmed existing morphological taxonomies. This research also demonstrated that *E. huxleyi* differentiation from *Gephyrocapsa* species occurred within the *Gephyrocapsa* genus lineage, indicating that it should, in fact, be named *Gephyrocapsa huxleyi* (Bendif et al. 2019). Further investigation of 43 *Gephyrocopsa* genomes sequencing data identified potential drivers for evolution of marine phytoplankton. Speciation events were found to coincide with periods of glaciation which, which suggests that isolation between populations and increased differences between ecological niches help to drive speciation (Filatov et al. 2021). These new findings demonstrate that whole-genome sequencing is a hugely powerful tool for increasing our understanding of speciation and evolutionary history, and that the relative lack of sequence data for eukaryotic phytoplankton is a limiting factor in advancing research. Advances in long-read sequencing technology such as nanopore mean that it is increasingly feasible to produce high quality genome assemblies to complement whole-genome sequencing data, which could be functionally annotated and used to provide increased insights into the molecular basis for speciation.

Coccolithophores such as *E. huxleyi* have a haplo-diplontic life cycle, featuring heavily calcified diploid and lightly or non-calcified haploid stages, with reproduction in both haploid and diploid stages (Cros et al. 2000). The differences in calcification mean that diploid and haploid phases can make contrasting contributions to the carbon cycle, since calcification results in the production of CO_2. Previous research has found that differences in calcification, cell size, and morphology, which are especially marked between haploid and diploid life stages, affect the distribution of coccolithophores, perhaps contributing to the distribution of species such as *E. huxleyi* across a wide range of ecological niches (Young et al. 2003). Several studies confirm that coccolithophores occupy different niches at different points in the life cycle, and previous genetic models support the idea that the haplo-diplontic life cycle provides an evolutionary advantage by allowing coccolithophores to colonise a wider range of ecological niches (Rescan, Lenormand, and Roze 2016).

Haploid coccolithophores have increased protein turnover, primary metabolism, and motility compared to the diploid phase, which allows them to surive under more stressful conditions. For example, it was found that in *E. huxleyi*, blooms are often terminated in the wild by fatal viral infections, which haploid populations are

not affected by, which has led to the development of haploid communities under viral pressure, known as the Cheshire Cat escape (Frada et al. 2008). There is also evidence that the haploid phase exhibit different responses to UV, and have different temperature, salinity, and nutrient tolerances which extends their range compared to diploid only populations (Ruan et al. 2023; Rokitta et al. 2014).

Coccolithophore research to date has focussed mainly on the diploid phase, however, so we do not currently have a full picture of the drivers of coccolithophore distribution and niche expansion, and the contributions of haploid populations (Taylor, Brownlee, and Wheeler 2017). One reason for the lack of research into the haploid phase is that the majority of coccolithophore research is based on *E. huxleyi* which is ubiquitous and a key species for climate studies, but is difficult to study in the haploid form because the unmineralised morphology is not easy to identify using microscopy (Frada et al. 2008). As an alternative to microscopic analysis, genomics could help to improve our understanding of the haploid stage in *E. huxleyi*, through the production of a genome assembly of a haploid strain.

Recent research has established that there are marked differences in the nutrient assimilation, gene expression, and photosynthesis in haploid *E. huxleyi* cells. It has been suggested that the process of calcification is energy-intensive, allowing the haploid cells to expend more energy on other cellular processes such as motility and metabolism. The vertical distribution of haploid *E. huxleyi* cells is not well characterised, but new studies allow us to infer their distribution based on differences in UV tolerance. Haploid cells appear to be more sensitive to high UV levels, but also exhibits a rapid recovery from UV radiation induced inhibition of photosynthesis compared to diploid cells. As diploid cells tend to be evenly distributed across surface waters with high UV levels, this evidence indicates that haploid cells tend to live lower in the water column based on their lower acute UV radiation tolerance, with rapid recovery allowing them to withstand vertical mixing in the water column. (Ruan et al. 2023)

A haploid *E. huxleyi* assembly would allow investigations into the genetic basis for the haplo-diplontic life cycle, the process of calcification, and niche preferences; provide a useful resource for the study of speciation within coccolithophores; and expand the potential for metagenomic identification of coccolithophores through increased sequences in genomics databases. This has not been carried out to date, however, due to difficulties in extracting high quality DNA, and assembling the genome which is extremely complex, with a high density of unclassified repeats, tandem repeats, and low-complexity regions which are almost impossible to resolve using short-read sequencing. The CCMP1516 genome assembly is incomplete with a high number of contigs due to these challenging features, which means it cannot give a full picture of the genes and genome structure

(Read et al. 2013). Long-read sequencing allows improved resolution of complex genomes, with promising developments seen recently in the human genome among others (Miga et al. 2020). Given the research need for high quality eukaryotic phytoplankton genomes, and a haploid *E. huxleyi* genome assembly in particular, one of the aims of this project was to produce a high quality genome assembly for the haploid strain RCC1217 using nanopore sequencing, as covered in Chapter 2.

2.1.2 Barriers to producing a genome assembly

The extraction of high quality DNA from phytoplankton is not straight forward, due in part to their physiology. Coccolithophores have calcium carbonate coccoliths, or armoured plates, over their cell membrane which, as with the silica frustules of diatoms and the thecal plates of dinoflagellates, can interfere with the cell lysis required for DNA extraction. Polysaccharides contained in phytoplankton cells also cause problems, widely seen in plant and fungal DNA extraction, as they are not removed during the extraction process and remain in the end product, reducing the purity of the DNA sample. The RCC1217 *E. huxleyi* strain is haploid, and therefore non-calcified, but it has non-mineralising organic scales covering the polysaccharide-containing cell membrane (Paasche 2001), which may contribute to difficulty in minimal damage cell lysis and in the extraction of high quality DNA.

DNA extraction methods which can be effective for removing polysaccharides include cetyl trimethylammonium bromide (CTAB) with phenol and chloroform. The CTAB extraction protocol has been used for the extraction of high purity, relatively HMW DNA from phytoplankton, plants, and fungi for many years. It has been found to result in reduced enzyme inhibition compared to other methods, allows for the simple removal of many polysaccharides, and the addition of phenol and chloroform denatures proteins and emulsifies them to allow their removal from the extracted DNA (Rogers and Bendich 1994). Other extraction methods which may be effective include those which use anion-exchange resin to bind the DNA while impurities are removed such as Qiagen Genomic-Tips, or which employ a resin-based phase separation to enhance DNA extraction with chloroform such as Cytiva's illustra Nucleon Phytopure.

Even with high quality DNA extraction, production of a high quality *de novo* genome assembly is far from simple, especially for complex eukaryotic genomes. Previous studies have found that the *E. huxleyi* pan genome has a high GC content, and is dominated by repetitive elements (Read et al. 2013) which would

complicate genome assembly. Plant genomes are notoriously difficult to assemble because of heterozygosity, unpredictable ploidy, frequent large complex repeats, and transposable elements which present a range of challenges for genome assembly (Claros et al. 2012), and many of these features are also seen in phytoplankton (Falkowski et al. 2004).

Prior to the widespread adoption of NGS methods, sequencing was largely reserved for model organisms, which are generally haploid - such as bacteria and yeasts - or inbred eukaryotic lines with low-variation which resulted in a lack of development for complex assembly tools until recently (Tigano, Sackton, and Friesen 2017). As a result of this, many *de novo* sequencing projects have relied on a combination of both long and short-read sequencing methods, particularly for complex eukaryotic genomes. Short Illumina reads are used as the basis for the assembly, due to low error rates, with long reads used for scaffolding and to bridge repeats and other genomic features which confound assembly from short-reads (Miller et al. 2017). Recent advances in long-read sequencing accuracy, and in base calling and assembly algorithms, however, mean that *de novo* genome assembly is now possible using only long reads. A nematode genome with complex DNA repeats which had not previously been resolved, was recently assembled using MinION reads alone, with 40x coverage (Eccles et al. 2018). This was achieved using Canu, a long-read assembler optimised for noisy, single-molecule sequencing data (Koren et al. 2017). Contamination is a significant problem in *de novo* genome assembly, particularly where there are few genome assemblies available for related organisms, and where it is challenging to produce large quantities of high quality DNA to give high coverage and to produce a high quality genome assembly (Cornet and Baurain 2022).

A crucial limiting factor in research into *E. huxleyi*, coccolithophores, and more widely, other phytoplankton such as diatoms, has been a lack of effective methods for the extraction of HMW gDNA, and appropriate sequencing technology to resolve the complexity of its genome. Adaptation of HMW DNA extraction methods for use with phytoplankton, coupled with long-read nanopore sequencing and assembly using long-read assemblers, could allow for the production of high quality genome assemblies. As such, this chapter covers the development of a DNA extraction, sequencing, assembly, and quality assessment protocol for the production of a high quality *E. huxleyi* RCC1217 genome assembly for annotation and analysis.

2.2 Methods

The finalised protocol is summarised in figure 2.2.1.

2.2.1 Culturing

E. huxleyi was cultured from 50 mL of an existing RCC1217 culture, grown in 500 mL culture flasks containing 250 mL f/2 medium as recommended by the Provasoli-Guillard National Center of Marine Phytoplankton (NCMA) (`https://ncma.bigelow.org/algae-media-recipes`), from Guillard and Ryther 1962. The cultures used to test DNA extraction protocols were incubated at 20 ℃ under 24 hour light conditions and grown to the exponential phase before 50 mL was used for new cultures as above, and the remaining 250 mL was used for DNA extractions.

2.2.2 Establishing a DNA extraction method

CTAB

For CTAB extraction, 250 mL *E. huxleyi* culture was vacuum filtered using 33 mm diameter, 0.45 µm pore filters and filters stored at -80 ℃ until DNA extraction.

The CTAB extraction protocol was adapted from (Phillips, Smith, and Morden 2001). 3 mL per filter of 3% CTAB was added to a falcon tube along with filter(s) containing *E. huxleyi* biomass and incubated at 65 ℃ for at least 1 hour. 1 volume of 24:1:1 Phenol:chloro:isoamyl alcohol was added and mixed gently by inverting the tube before being centrifuged for 30 minutes at 10000 rcf at room temperature. This resulted in 2 distinct phases, and the aqueous phase was carefully removed to a new tube before the rest was discarded. 2/3 volume of -20 ℃ isopropanol was added and incubated on ice for 15 minutes with occasional gentle inversion before centrifuging at 10000 rcf for 30 minutes at 4 ℃ which produced a pellet. Over ice, the supernatant was discarded and the pellet washed twice with ice-cold EtOH. The pellet was left to dry until transparent (no longer than 30 minutes) and low TE buffer added until concentration was around 500 ng/µL based on NanoDrop measurements. DNA quality was assessed using NanoDrop and gel electrophoresis. NanoDrop measures DNA purity through the ration of absorbance at 260/280 nm and 260/230 nm, with a 260/280 measurement of around 1.8 and a 260/230 measurement of around 2.0 indicating pure DNA. Gel electrophoresis was performed using 0.4% agarose gel in a ThermoScientific

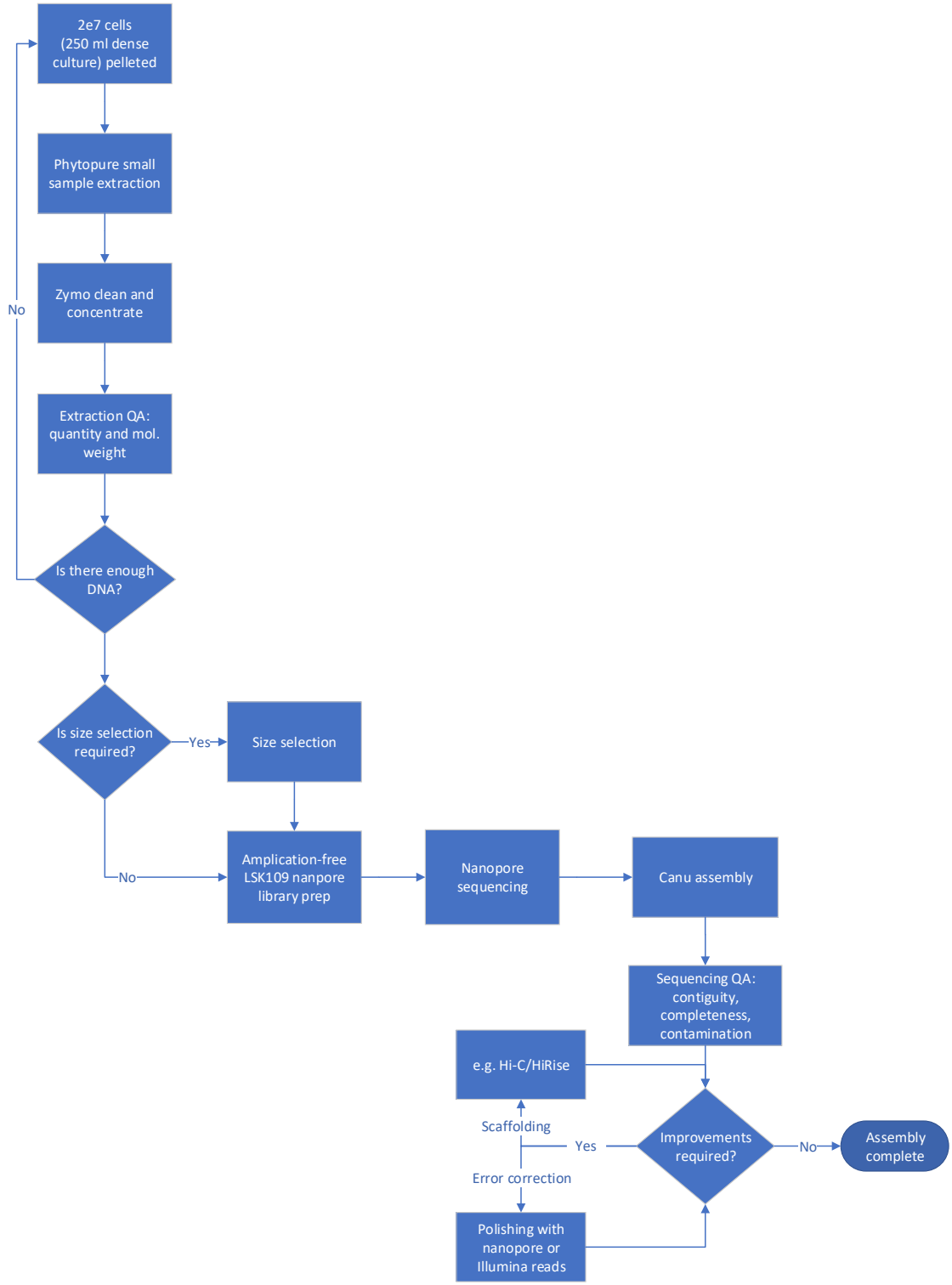


Figure 2.2.1: Summary of the protocol for genome assembly for *E. huxleyi*. Rectangles represent processes, diamonds represent decision points

horizontal midi gel electrophoresis system. The sample was run with a lambda DNA/HindIII Marker2 ladder at 40V for 2 hours and 30 minutes, and the resulting bands visualised under UV light.

Adaptations to the CTAB protocol were trialled:

1: As above with the addition of 20 µL proteinase k and 2 µL of RNase A to the filters before incubation overnight at 60 ℃.

2: As above with the addition of 20 µL proteinase k to the filters before incubation overnight at 60 ℃

Qiagen Genomic-tips

Adapted from the yeast and bacterial protocols recommended by Qiagen in the Genomic-tips manual `https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/qiagen-genomic-tips`. Two protocols were trialled, with and without cell lysis using a French Press pressure cell. A 20/g Midi Genomic Tips kit was used alongside Zymo clean and concentrate columns.

A pellet of around $2x10^7$ *E. hux* cells was produced from 250 mL of culture by centrifuge at 1300 rcf for 15 minutes at 4 ℃ and resuspended in 3.5 mL TE buffer. Where used, cell lysis was performed using a French Press at 3200 psi. 80 µL lysozyme, 45 µL proteinase k, and 250 µL zymolyase were added to the suspension and incubated at 37 ℃ for 30 minutes. 1.2 mL of the bacterial lysis buffer (3 M guanidine HCl; 20% Tween-20) was added and incubated at 50 ℃ for 30 minutes. The Genomic-tips were equilibrated with 4 mL equilibration buffer (750 mM NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol, 0.15% Triton X-100). The sample was vortexed to resuspend any precipitate and applied to the Genomic-tip to pass through the column, assisted using an adapted plunger at a rate of no more than 15 drops per minute. After the sample passed through, the Genomic-tip was twice washed with 7.5 mL of wash buffer (1.0 M NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol) before elution with 5 mL of elution buffer (1.25 M NaCl; 50 mM Tris·Cl, pH 8.5; 15% isopropanol). The elutant was precipitated with 3.5 mL isopropanol and the DNA quantity checked using the NanoDrop. Where the DNA concentration of the elutant was greater than 10 µg/mL alcohol precipitation was performed by centrifuging at 4000 rcf for 15 minutes at 4 ℃, removing the supernatant and washing with 2 mL 70% ice-cold EtOH, and drying the pellet before resuspending in TE buffer. Where the DNA concentration of the elutant was less than 10 µg/mL, Zymo clean and concentrate columns were used

and the DNA eluted into DNase-free water. DNA quality was assessed using NanoDrop and by Earlham Institute using capillary electrophoresis with a Femto Pulse.

Phytopure

Adapted from the Cytiva illustra Nucelon Phytopure genomic DNA extraction kit manual `https://www.cytivalifesciences.com/en/us/shop/molecular-biology/extraction/genomic-dna/illustra-nucleon-phytopure-p-05551#related-documents`.

A pellet of around 2x10^7 *E. hux* cells was produced from 250 mL of culture by centrifuge at 1300G for 15 minutes at 4 ℃ and DNA was extracted as described in the manual for small samples (0.1g), excluding the tissue grinding step. The final DNA pellet was suspended in Tris to a concentration of around 10 µg/mL. Zymo clean and concentrate columns were used to improve the DNA quality for sample 1. The quality of the extracted DNA was assessed by fluometric quantification using the Qubit Fluorometer, and by Earlham Institute using pulsed-field capillary electrophoresis (PFCE) with an Agilent Femto Pulse. Pulsed field electrophoresis allows for the separation of fragments above $\tilde{2}$0 kbp, which is not possible with traditional gel electrophoresis, by alternating the direction of the electric field. PFCE carries out the electrophoresis in capillaries rather than agarose gel which gives results in much shorter time frames. Size selection of DNA fragments greater than 40 kbp was performed using the Blue Pippin size selection system for sample 2 to remove short reads which are preferentially run during nanopore sequencing and can result in long reads not being sequenced before pores degrade.

2.2.3 Nanopore Sequencing

Nanopore sequencing was carried out using the Phytopure extraction DNA. Library preparation was carried out according to the Oxford Nanopore technologies (ONT) SQK-LSK109 protocol for amplification-free nanopore DNA sequencing. Two sequencing runs were carried out on 2018 ONT MinION flowcells and run for around 48 hours, until all of the available pores were depleted.

2.2.4 Assembly

Nanopore only assembly

Reads passing the default quality filter from the two sequencing samples were combined for assembly. Porechop (v0.2.1, `https://github.com/rrwick/Porechop`) was used to remove adapters from the nanopore sequencing reads before assembly.

De novo assembly was performed using each of three assemblers to to determine the best assembly method. These were Miniasm (Li 2016), which uses a simple OLC assembly strategy; Canu (Koren et al. 2017), which corrects and trims reads prior to a modified OLC strategy; and Flye which assembles reads first into error-prone disjointigs before producing a repeat graph which is resolved to produce the final contigs (Kolmogorov et al. 2019), see figure 2.2.2. Miniasm was used in *de novo* assembly mode with the nanopore sequencing setting. Canu was used with the nanopore raw reads setting. Flye was run with the nanopore raw reads setting.

The assemblies were quality checked by investigating contiguity, read lengths, and N50, alongside the QUAST genome quality assessment tool (Gurevich et al. 2013). BUSCO (Manni et al. 2021) was used to determine the completeness of the genome using the eukaryotic database (BUSCO v3.0, odb9). Comparison against an already published *E. huxleyi* genome, CCMP1516 (Read et al. 2013) was carried out using the dnadiff tool from the MUMmer (v3.2.3) suite, which aligns two sequences to identify points of difference and outputs alignment statistics. (Kurtz et al. 2004). BUSCO was also used to determine the completeness of the CCMP1516 genome, as the reported quality checks had been carried out using a now obsolete pipeline, CEGMA (Parra, Bradnam, and Korf 2007), which could not be replicated.

Assembly improvement

Dovetail Genomics carried out Hi-C sequencing and HiRise scaffolding using the canu assembly produced above as a basis to produce an improved assembly. Illumina sequencing data produced as part of a previous project was used to polish the assembly with Pilon (1.2.3) (Walker et al. 2014). The resulting assembly was quality checked as above.

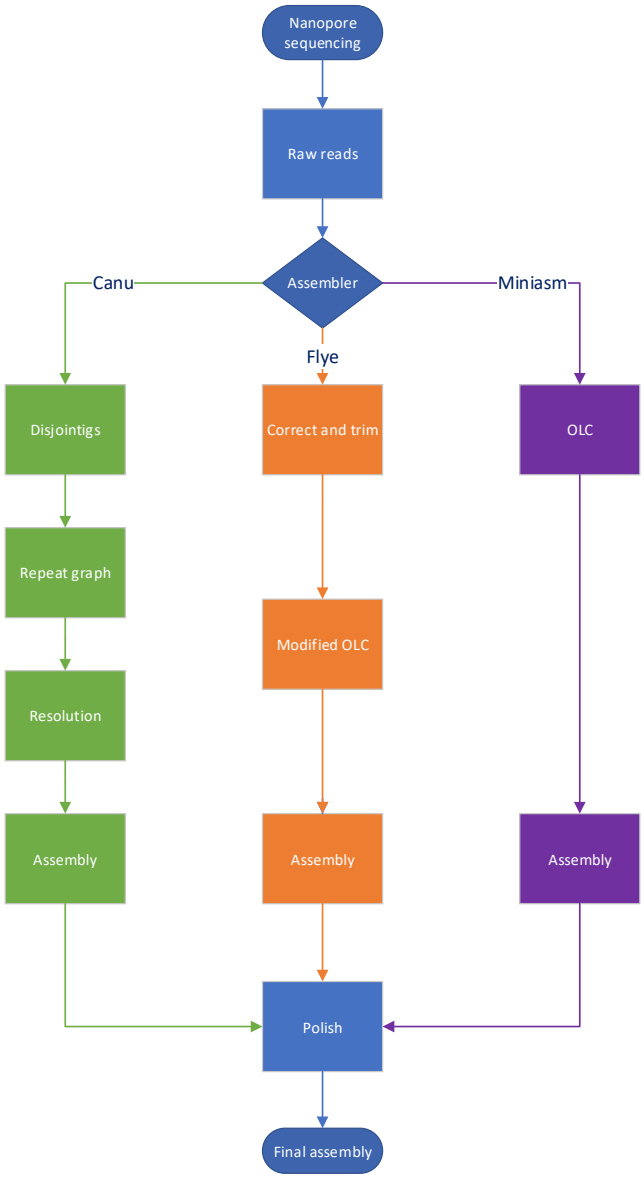


Figure 2.2.2: A diagram showing the process used by Canu, Flye, and Miniasm assemblers to assemble nanopore reads.

2.2.5 Contamination identification and removal

Contamination was examined by BLASTing reads against the NCBI nt database and potential contaminants were screened to identify true contaminants, as opposed to unidentified *E. huxleyi* sequences, or products of horizontal gene transfer (HGT). Filtering of contaminants was carried out by removing sequences with a minimun length of 500 bp which had a 95% or above identity to something

that was not *E. huxleyi*. This resulted in the complete removal of 3 contigs, and the removal of parts of 21 other contigs. The quality of the resulting assembly was assessed as described above.

2.3 Results

2.3.1 Establishing a DNA extraction method

The purity and molecular weight of DNA produced from each extraction method is summarised in table 2.1. Phytopure DNA extraction produced the highest quality DNA and was selected for nanopore sequencing.

Table 2.1: 260/280 and 260/230 ratios of the CTAB, Genomic-tips, and Phytopure DNA extractions as measured by the NanoDrop to assess DNA purity, the highest measured molecular weight of the extracted DNA as measured by gel or capillary electrophoresis, and whether there was clear degradation in the sample.

Protocol	260/280	260/280	Mol Weight	Degradation?
CTAB	1.97	1.47	ca. 23 kbp	Yes
Genomic-tips	1.82	1.91	20-30 kbp	Some
Phytopure	1.82	2.01	>100 kbp	No

CTAB

CTAB DNA extraction produced relatively poor quality DNA, with high 260/280 ratios of around 2 and low 260/230 ratios, around 1.5 as measured by the NanoDrop. Adapted protocols including RNase A and proteinase k incubation did not result in improvements. This indicates a problem with the DNA extraction process, for example the presence of carbohydrates remaining in the sample after extraction, or incomplete removal of phenol. There was a mixture of relatively high molecular weight (HMW) DNA (>20 kbp) and extremely degraded DNA (<100 kbp) when examined using gel electrophoresis.

Genomic-tips

The Genomic-tips DNA extraction produced higher quality DNA, with 260/280 ratios of around 1.8 and 260/230 ratios of around 1.9 indicating an improved purity of DNA. The molecular weight was relatively high, with a broad peak around

20-30 kbp when analysed using capillary electrophoresis - see figure 2.3.1. There was no difference in DNA quality with and without the French Press cell lysis.

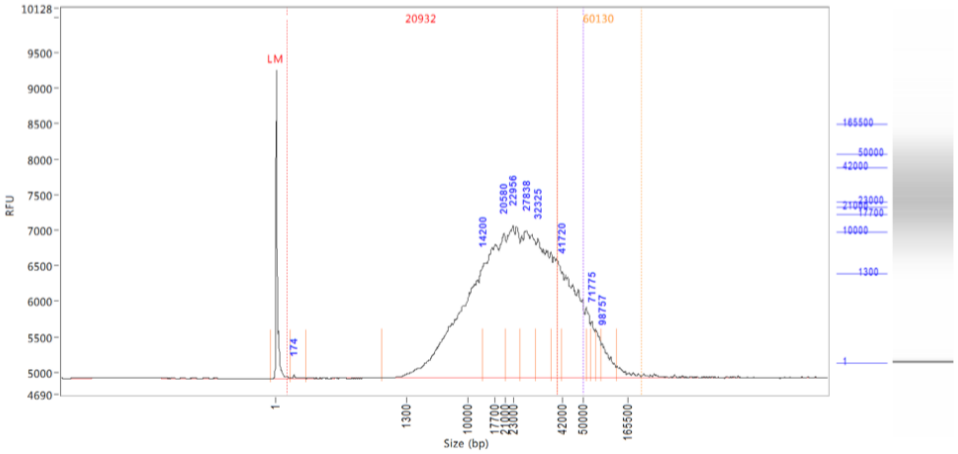


Figure 2.3.1: Capillary electrophoresis output from Femto Pulse showing the molecular weight of DNA extracted using the Genomic-tips protocol against the relative fluorescent units (RFU) measuring DNA quantity. There is a peak from around 20-30 kbp.

Phytopure with clean and concentrate

Phytopure DNA extraction produced very high quality DNA. The 260/280 ratio was 1.82 and the 260/230 ratio was 2.1.

Sample 1 - Before using the clean and concentrate columns, capillary electrophoresis showed a sharp peak of around 133 kbp, with a long tail below 50 kbp. After the clean and concentrate columns had been used, there was a sharp peak at around 150 kbp, with a larger tail from 50 kbp down to around 10 kbp - see figure 2.3.2. The DNA extraction produced around 19 µg DNA, as measured by the Qubit Fluorometer, which was reduced to around 5 µg after clean up and library preparation with 1.6 µg loaded for sequencing.

Sample 2 - Before size selection, there was a sharp peak at around 136 kbp and a small second peak at around 2-5 kbp. Blue Pippin size selection at 40 kbp reduced small DNA fragments and resulted in a strong peak at around 120 kbp, with a longer tail from around 50 kbp down to 20 kbp, see figure 2.3.2. Around 3 µg of DNA was present after size selection and library preparation from around 17 µg extracted as detected by the Qubit Fluorometer.

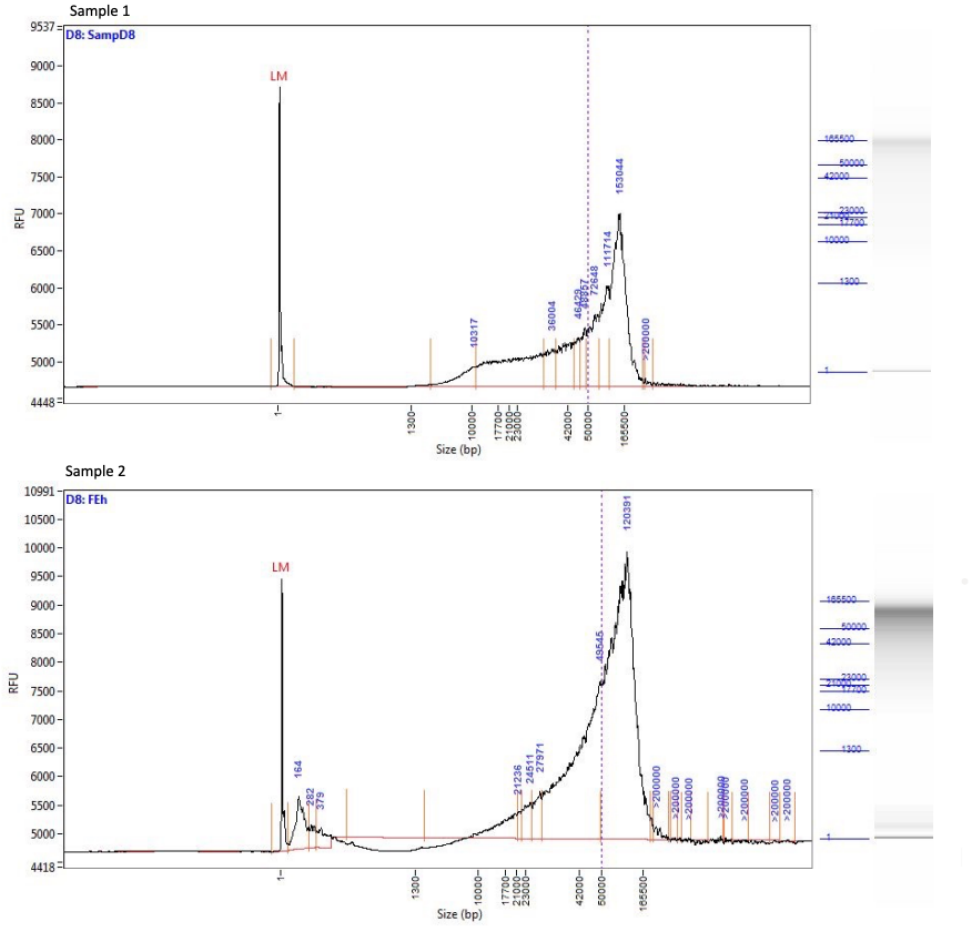


Figure 2.3.2: Capillary electrophoresis output from Femto Pulse (Agilent) showing for sample 1 and sample 2 the molecular weight of DNA extracted using the Phytopure protocol against the relative fluorescent units (RFU) measuring DNA quantity. Samples were assessed after size selection and clean up. The graphs show sharp peaks at around 150 and 135 kbp respectively.

2.3.2 Nanopore Sequencing

Nanopore sequencing of the two Phytopure samples produced around 2.42 Gbp of sequencing data after 48 hours. As summarised in table 2.2, nanopore sequencing produced relatively long reads with a mean read length of around 10 kbp for sample 1 and 13 kbp for sample 2, with an N50 of 26 kbp, and 25.5 kbp respectively. The longest read was 148 kbp in sample 1 and 235 kbp in sample 2. This compared favourably with nanopore sequencing results in Chapters 3 and 4 which used metagenomic samples, not optimised for HMW DNA, indicating that sequencing outputs are improved with high-purity, HMW inputs.

Table 2.2: Run metrics, showing the yield, total number of reads, longest read, N50, N90, and number of reads longer than 50 kbp for each nanopore sequencing sample.

	Sample 1	Sample 2
Total yield (Gbp)	1.25	1.17
Reads	123068	86822
Mean length	10130	13519
Longest read	148262	235175
N50 length	26107	25535
Number of reads >= N50 length	14613	18254
N90 length	6641	11687
Number of reads >= N90 length	50993	41421
Number of reads > 50 kbp	3593	723

2.3.3 Assembly

Nanopore-only assembly

Assemblies of the nanopore sequence data were generated using the tools Miniasm (Li 2016), Canu (Koren et al. 2017) and Flye (Kolmogorov et al. 2019). The genome assemblies produced from each assembler were analysed for contiguity and other descriptive statistics as summarised in table 2.3. Miniasm and Canu produced assemblies of a similar length at around 138 Mbp, while Flye produced a longer assembly of 174 Mbp. Previous studies have found haploid *E. huxleyi* genome sizes to range up to 133 Mbp (Read et al. 2013). The Flye assembly had greatest contiguity with 354 contigs, compared to 504 and 1204 for Miniasm and Canu respectively. Flye also had the largest N50, N90, and mean contig length, followed by Miniasm and then Canu.

The assemblies were compared to the already published *E. huxleyi* CCMP1516 assembly using dnadiff, with the results summarised in table 2.4. It can be seen that the Canu assembly had the greatest alignment to the CCMP1516 genome, with the greatest proportion of aligned contigs and bases. The CCMP1516 assembly had 7795 scaffolds, 16921 contigs, scaffold N50 of 404.8 Kb, and contig N50 of 29.7 Kb with a total size of 167 Mb. The Canu assembly had the highest alignment to the CCMP1516 assembly, with a 92.76% percentage identity and 74.65% of bases aligned. Flye and Miniasm had lower alignments, both around 85% percentage identity and 64% of bases aligned. BUSCO completeness scores for the Miniasm and Flye assemblies were zero, while the Canu assembly

Table 2.3: Assembly size, total number of reads, longest read, N50, N90, and number of reads longer than 50 kbp for each assembly.

	Miniasm	Flye	Canu
Assembly size (Mbp)	137	174	139
Contigs	504	354	1214
Mean contig length	271878	492597	114836
Longest contig	2861626	4169934	3822639
N50 length	404907	1254394	201592
Number of contigs >= N50 length	107	43	187
N90 length	131791	220559	55300
Number of contigs >= N90 length	339	164	671
Number of contigs > 50 kbp	481	306	708

had a completeness of 13%, with 40 complete BUSCOs. The increased alignment to the CCMP1516 genome and 40 complete BUSCOs compared to zero for the other assembies indicated that the Canu assembler had produced an assembly which was more likely to be a true representation of the genome, while the increased contiguity shows that more of the reads had been assembled together. Accordingly, the Canu assembly was selected for improvement and sent to Dovetail Genomics for scaffolding using Hi-C sequencing technology and the HiRise assembly software.

Table 2.4: Percentage of bases aligned to the CCMP1516 genome assembly, and the average 1:1 alignment percentage identity for each RCC1217 assembly.

Assembler	Bases aligned (%)	Average alignment (%)
Miniasm	64.50	84.45
Flye	64.32	85.5
Canu	74.65	92.76

Hybrid assembly

The assembly produced by Dovetail, which used the Hi-C alignment and HiRise assembly, was smaller than the Canu assembly, indicating that there had been a loss of information. The HiC-Canu assembly was polished using pilon to incorporate Illumina sequencing data for error correction which resulted in a 142 Mbp assembly which showed improvements in contiguity, N50 and all other

metrics compared to the Canu assembly. The Canu, HiC-Canu, and polished HiC-Canu assemblies are compared in table 2.5.

Table 2.5: Assembly metrics showing the assembly size, total number of reads, longest read, N50, N90, and number of reads longer than 50 kbp for the Canu assembly, the HiC-Canu assembly, and the polished HiC-Canu which was improved with Illumina sequencing data.

	Canu	HiC-Canu	Polished HiC-Canu
Assembly size (Mbp)	139	40	142
Contigs	1214	593	241
Mean contig length	114836	68205	589933
Longest contig	3822639	457573	10650876
N50 length	201592	108588	5228308
Number of contigs >= N50 length	187	113	10
N90 length	55300	37664	2668444
Number of contigs >= N90 length	671	345	24
Number of contigs > 50 kbp	708	298	47

BUSCO was used to assess the completeness of the polished HiC-Canu genome assembly, and compared to the results to the Canu, HiC-Canu, and CCMP1516 assemblies, along with N50, contiguity, and contig N50 as can be seen in table 2.6. The BUSCO completeness for the polished HiC-Canu assembly was 58.7% which is far higher than either of the previous Canu assemblies, and also higher than the published CCMP1516 assembly. The contiguity was improved compared to the other assemblies, and the N50 was vastly higher after polishing.

Table 2.6: BUSCO completeness percentage, N50, number of contigs, and number of contigs with a length greater than the N50 for each of the Canu based assemblies, and the published CCMP1516 assembly.

Assembly	BUSCO (% complete)	N50 (Mbp)	Contigs	Contigs >= N50
Polished HiC-Canu	58.7	5.20	241	10
HiC-Canu	13.5	1.08	593	113
Canu	13.2	2.01	1214	187
CCMP1516	41.6	4.04	7795	109

2.3.4 Contamination identification and removal

Contamination was identified based on BLAST-nt analysis which identified 3 contigs which were entirely aligned to alphaproteobacteria genomes, and also showed that 156 contigs contained partial matches to organisms other than *E. huxleyi*. The 3 entirely non-*E. huxleyi* contigs were removed, and the remaining alignments were screened to identify potentially missed *E. huxleyi*, chloroplast, and mitochondrial alignments. Contaminants were identified and filtered out of the assembly where there existed a BLAST hit to something other than *E. huxleyi* which was at least 500 bp long and had at least 95% identity. This increased the proportion of blast alignments matching to *E. huxleyi* from 53.8% before filtering to 81.6% afterwards. The pre- and post-filtration results are summarised in tables 2.7 and 2.8. It can be seen that filtration produced an assembly of 130 Mbp, 262 contigs, and an N50 of 4.9 Mbp which is a slight reduction in contiguity, size, and N50 compared to the unfiltered assembly. There was a slight decrease in genome completeness as measured by BUSCO, possibly due to decreased contiguity and genome size, with 2 extra missing BUSCOs, 2 extra fragmented BUSCOs, and four fewer duplicated BUSCOs in the filtered assembly.

Table 2.7: Assembly size, total number of reads, longest read, N50, N90, and number of reads longer than 50 kbp for the polished HiC-Canu assembly before and after filtration of contaminants.

	Pre-filtration	Post-filtration
Assembly size (Mbp)	142	130
Contigs	241	262
Mean contig length	589933	497558
Longest contig	10650876	10650876
N50 length	5228308	4909455
Number of contigs >= N50 length	10	9
N90 length	2668444	936142
Number of contigs >= N90 length	24	30
Number of contigs > 50 kbp	47	64

This left some contigs with partial alignments to organisms other than *E. huxleyi* but these are generally short and have a low percentage identity. Figure 2.3.3 shows the blast alignments in the assemblies before and after contamination removal plotted by alignment length against percentage identity. From this it can be seen that there are far fewer alignments which are not to *E. huxleyi* in the

Table 2.8: Number of complete BUSCOs, duplicated and single, fragmented and missing BUSCOs and the total number of BUSCOs searched for the polished HiC-Canu assembly before and after filtration of contaminants.

	Pre-filtration	Post-filtration
Complete BUSCOs	178	174
Complete single	170	170
Complete duplicated	8	4
Fragmented BUSCOs	47	49
Missing BUSCOs	78	80
Total BUSCOs searched	303	303

post-removal assembly, and the majority of those have low percentage identity and length.

2.4 Discussion

2.4.1 Establishing a DNA extraction method

CTAB extraction produced DNA with a molecular weight peaking around 23 kbp, with low purity, and significant degradation. While CTAB is widely used for phytoplankton DNA extraction, it is often in conjunction with PCR, and generally with DNA sequencing technologies such as Sanger or Illumina (Rogers and Bendich 1994), as opposed to nanopore sequencing as in this project. DNA purity is particularly important in nanopore sequencing, as it relies on DNA strands passing through pores which may be clogged or rendered inactive by contaminants. The aim was to produce a high quality genome assembly and, given that one key advantage of nanopore sequencing is the ability to sequence long strands of DNA, it was decided that higher molecular weight DNA was required. Achieving a high yield was particularly important for *E. huxleyi* sequencing because of the high GC content, which prevented the use of PCR to increase the DNA available for sequencing due to biased amplification(Chen et al. 2013). There is some evidence that CTAB solution can cause DNA degradation (Doktorovova et al. 2013), as can long incubation at temperatures above 50℃. As such, alternative extraction protocols were investigated.

The Qiagen Genomic-tips extraction method was chosen because it was specifically developed to reduce shearing and optimise molecular weight. It was trialled in conjunction with cell lysis by the French Press, as it was unclear

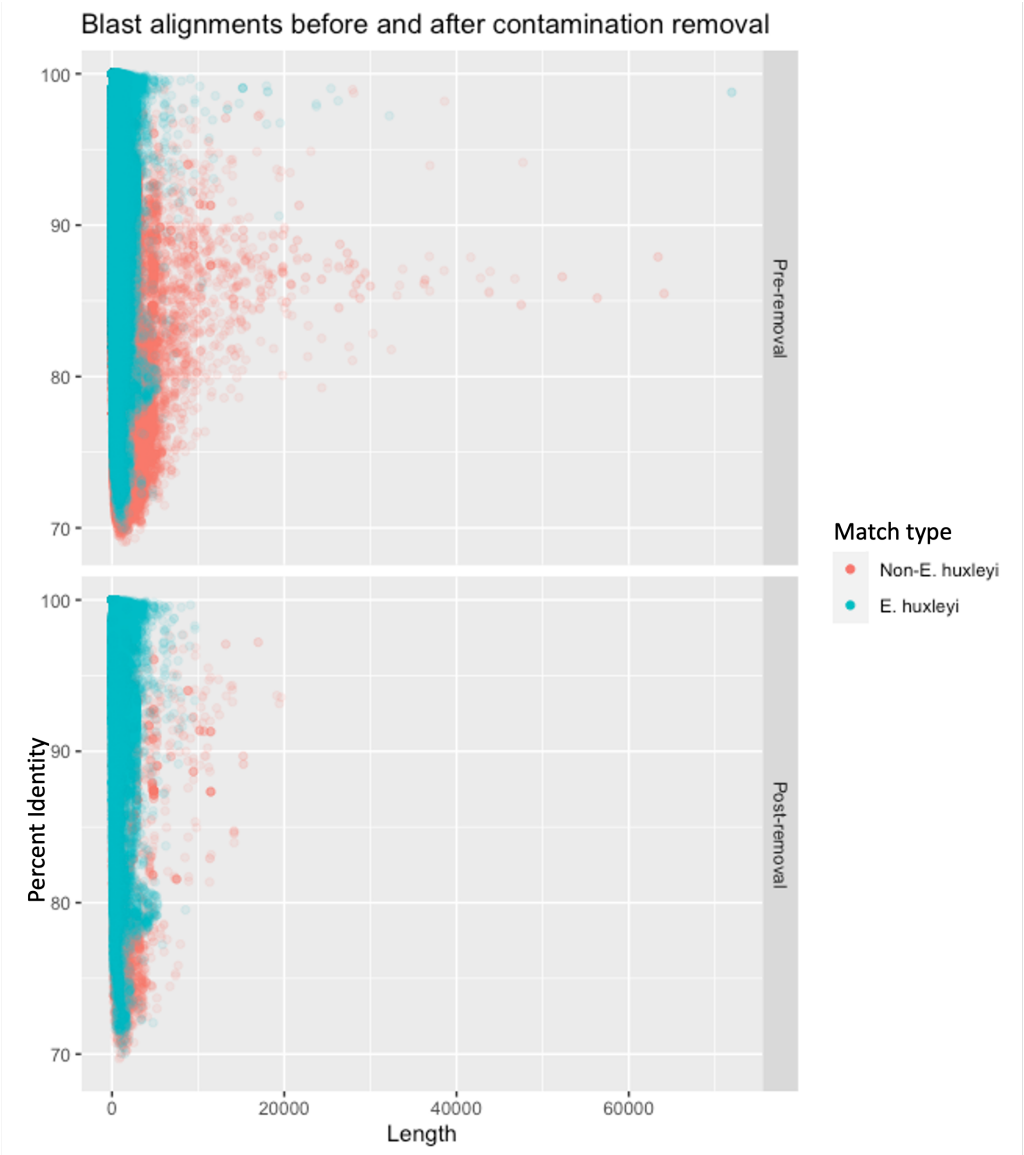


Figure 2.3.3: A scatter plot showing length of alignment against the percentage identity of the alignment in the RCC1217 assembly before and after removal of sequences which were above the contamination threshold.

whether the lysis buffer would be sufficient to lyse the cells alone. Pressure can be an effective method for cell lysis where DNA integrity is important, as it is not reliant on agitation which can result in shearing, as seen in freeze-thaw, grinding, and bead-beating lysis methods. In this instance, there was no appreciable difference between DNA yield and quality between samples treated with pressure lysis, indicating that the Genomic-tips lysis buffer is sufficient. For the diploid strain which has coccoliths, or for diatoms and dinoflagellates, there may be a benefit to employing pressure lysis with a French Press prior to beginning the DNA extraction protocol.

The Genomic-tips protocol uses a buffer containing guanidine HCl and Tween-20 to release DNA and remove proteins without degrading the DNA, then passes the sample through a column containing Anion Exchange Resin which, in low salt and low pH conditions, binds the DNA and allows the removal of RNA, proteins, and other contaminants before the DNA is eluted in a high salt buffer. This protocol produced DNA of improved purity and slightly increased molecular weight compared to the CTAB extraction protocol. The CTAB protocol relies on phenol and chloroform to remove impurities, which can themselves contribute to DNA impurity if they are not fully washed out, while the genomic tips protocol binds the DNA to a resin while impurities are washed away which appears to be more effective for *E. huxleyi* DNA extraction. This may be due to the complex cell wall and scales, meaning there are more impurities to be removed than are seen in, for example, bacterial cells. The molecular weight of the DNA was also slightly increased compared to the CTAB extraction, perhaps because of reduced incubation times at high temperatures, or the absence of potentially genotoxic reagents. The DNA produced from the Genomic-tips extraction method was superior to that achieved with CTAB extraction, but the molecular weight and purity were not optimal for genome assembly, so further alternatives were investigated.

The illustra Nucleon Phytopure DNA extraction protocol was developed for the rapid extraction of high quality, HMW DNA from plant and fungal cells. Following lysis with potassium SDS, DNA extraction is performed with ice-cold chloroform, and a resin gives clear phase separation for easy removal of the DNA-containing layer, which maximises the volume of sample which can be removed without sacrificing purity. This protocol produced highly pure, extremely HMW DNA. There is no phenol in the protocol, which may explain improved purity compared to the CTAB extractions which likely had phenol contaminaiton. Ice-cold chloroform, along with a resin phase separation, appears to be highly effective at the removal of proteins and polysaccharides, resulting in high purity DNA. The Phytopure extraction protocol produced the highest molecular weight DNA of any extraction protocols tested, with a peak of over 100 kbp compared to 20-30 kbp for other methods. This may be due to a range of factors, including the rapidity of the extraction process, reduced incubation time at temperatures above 50 ℃ compared to CTAB, lack of physical binding and release compared to the Genomic-tips extraction, or the use of ice-cold reagents such as chloroform and isopropanol, among others.

Further work on DNA extraction would include testing these protocols on other phytoplankton, including diatoms, dinoflagellates and diploid *E. huxleyi* to establish which is the best all-round protocol, and whether a different approach is

necessary for phytoplankton with hard shells. It would be interesting to investigate whether lysis using the French Press allowed for the use of DNA extraction protocols such as Genomic-tips or Phytopure which use gentler processes and reagents to optimise HMW DNA. Future work would also involve establishing the best DNA extraction method for metagenomic samples, where the aim is less to optimise extremely HMW DNA extraction, and more to extract high quality, high yield DNA at representative levels for different organisms. It is possible that, for example, the gentler more HMW extraction protocols would result in samples with higher proportions of bacterial DNA, as it is generally easier to lyse. Ideally this work would produce a reproducible, expandable decision process for selecting the best extraction protocol for a phytoplankton sequencing experiment based on whether it is single species or metagenomic, the biology of the organism(s) in question, and the intended use of the sequencing data.

2.4.2 Nanopore sequencing for genome assemblies

Nanopore sequencing of the DNA extracted using the Phytopure protocol produced around 2.4 Gbp of sequencing data. Based on a haploid genome size of around 130 Mbp for *E. huxleyi*, this gives a coverage of around 18x, which is relatively low for the purposes of producing a high quality genome assembly (Nagarajan and Pop 2013). The yield of just over 1 Gbp per flowcell was in line with real-world expectations of performance at the time in 2018. There have since been significant improvements in ONT flowcells, and a realistic expected yield from this library would now be around 10 Gbp per flowcell. This increase in data would allow for the production of an initial assembly that would likely be of far higher quality than the initial assembly produced in this experiment. This is because increased coverage makes it simpler to identify and remove errors as they stand out, and there is less ambiguity in the way the reads fit together to create an assembly (Smits 2019). This results in increased contiguity, and fewer errors, and can also help to identify contamination and misassembly.

2.4.3 Quality assessment and evaluation of genome assemblies

It is difficult to assess the quality of *de novo* genome assemblies, particularly where coverage is low, as statistics such as contiguity and N50 can seem impressive even where there are serious errors (Smits 2019). As such, other factors were considered when deciding on which assembly to use to produce a hybrid genome assembly, incorporating Hi-C and Illumina sequencing data. The

initial Canu assembly produced from the nanopore sequencing data alone had relatively low contiguity at 1214 contigs, compared to the Flye and Miniasm assemblies produced from the same data, but it had a higher alignment rate against an already published *E. huxelyi* genome assembly, strain CCMP1516. Given that the contiguity and other statistical factors can be misleading in assessing whether a *de novo* assembly is a good representation of the genome, other indicators were taken into consideration. Alignment to a previously published *E. huxleyi*, strain CCMP1516 genome, and BUSCO completeness scores were considered, and based on these the Canu assembly was chosen for assembly improvement. The Canu assembler is the most established of the three tools tested, and is particularly optimised for complex, noisy sequences (Koren et al. 2017), while minimap requires a second tool to finish the assembly fully and Flye was very new at the time of this experiment, and was not optimised for complex noisy sequences. The quality of the assembly based on contiguity is not particularly high, but strong alignment to the CCMP1516 genome assembly does indicate that the Canu assembly is more likely to be representative of the *E. huxleyi* genome. BUSCO is a tool which assesses the completeness of a genome assembly, based on the presence of genes which are near universal in each kingdom (Manni et al. 2021). The BUSCO completeness score for the Canu assembly was very low at 13%, but compared to the Miniasm and Flye assemblies which both had completeness scores of zero, this was taken as an indication of improved assembly quality.

The difference in BUSCO score between the assemblers, with 0% for Miniasm and Flye, as oppposed to 13% for Canu shows that Canu is better at resolving complex sequences from long, error-prone reads, likely due to the read-correction and trimming stage which is not employed by either of the other assemblers. The BUSCO score of the Canu assembly improved vastly after polishing with highly accurate nanopore reads, from 13% to 58.7%, which indicates that the low BUSCO completeness scores are likely to be partly due to the high error-rate seen in nanopore sequencing at the time, which could result in BUSCOs being identified as missing or fragmented. The accuracy of nanopore sequencing has significantly improved since 2018, from 95% to 99.8% with new flow cells and super accurate basecalling software (Stefan et al. 2022).

Investigations in the Leggett Group at Earlham Institute have compared *Arabadopsis* genome assembly results from the LSK-109 kit and basecaller used in the *E. huxleyi* assembly, the same kit with ONT's newer super-accurate basecaller, and using a new flowcell alongside the recently released Q20+ kit which is designed to provide >99% raw read accuracy and new flowcell. The results, showed a 3.7 fold increase in yield from the new flowcell, and an

approximately 60% decrease in unaligned bases to the reference assembly with the Q20 kit, and 66.7% decrease in unaligned bases with the super-accurate basecaller (S. Martin and R. Leggett, personal communication, 08/05/2023).

As such, if this sequencing experiment were repeated now, the increased yield and accuracy would likely result in a vastly improved initial assembly. A 3.7 fold yield increase applied to this experiment would result in an expected yield of around 9 Gbp giving a coverage of over 70x for *E. huxleyi* genome, and which is in line with 10 Gbp per flowcell real world expectations. It is more complex to calculate the impact of the improvement in accuracy, due to the lack of a high quality reference genome against which to compare. Assuming that the that 74.65% of bases in the Canu assembly aligned to CCMP1516 were not erroneous, then 83.5% of basescould be expected to align instead. This indicates that it would be possible produce a high quality reference alignment if the experiment were repeated today.

The initial Canu assembly was used by Dovetail Genomics to produce a scaffolded hybrid Canu-HiC assembly, using chromatin conformation capture which gives information about the proximity of sequences from the spatial organisation of chromatin in the nucleus. This assembly showed a significant improvement in contiguity, N50, and other descriptive statistics for the assembly but it gave only a very slight increase in BUSCO completeness. This is not surprising, as errors from nanopore sequencing would not be corrected during this process, which is mainly based around understanding the spatial arrangement of chromatin. To correct these errors, Illumina sequencing data, made up of highly accurate but very short (150 bp) reads, was used to polish the assembly using pilon, which corrects incorrect bases, and also identifies misassemblies, and fills in gaps to produce an improved assembly (Walker et al. 2014). This process produced an assembly which had improved contiguity compared to the Canu-HiC assembly, but also had hugely improved completeness at 58.7%, indicating that errors in assembly and basecalling had negatively impacted the assembly quality. The completeness score of 58.7% is relatively low compared to most high quality published genomes, but it was higher than the CCMP1516 genome assembly which was around 40% complete. There are very few published phytoplankton genomes, and no other coccolithophore genomes published, so it is difficult to say whether intrinsic factors make the production of a higher quality genome assembly more challenging for *E. huxleyi*, such as high GC content. The GC content of the assembly was high, at around 65%, which is very similar to the reported GC content of the CCMP1516 genome assembly at 68%.

2.4.4 Contamination identification and removal

It is challenging to identify contamination in *de novo* phytoplankton genome assemblies. There are few published genome assemblies to compare to, and the complex evolutionary history including HGT and symbiosis mean that potential contaminants may in fact be genuine components of the genome. For this reason it was important to approach identification and removal of potential contaminants cautiously. In this case, BLAST was used to establish alignments of each contig to species other than *E. huxleyi*. Three contigs which aligned almost entirely to single alphaproteobacterial genomes, constituting almost entire genome assemblies, were easily removed as they were clearly not part of the *E. huxleyi* genome. There remained many contigs which had partial non-*E. huxleyi* alignments, although the overwhelming majority of these were short low identity hits which raised the question of whether they were truly contaminants. In many cases these could be simply random alignments of short DNA fragments, and particularly where the percentage identity was low, they may simply highlight the lack of full contiguous coccolithophore genome assemblies against which the assembly can align. As such, the potential contaminants were screened to remove candidates that were likely to be a genuine component of the genome, and a conservative threshold was set to remove contig fragments with a percentage identity above 95% and a length of greater than 500 bp.

The resulting assembly had slightly reduced contiguity compared to the unfiltered assembly, alongside a small reduction in N50. There was a reduced number of non-*E. huxleyi* alignments found in the BLAST-nt database, and increased the proportion of *E. huxleyi* alignments in the assembly to over 80%, compared to around 50% in the unfiltered assembly. The remaining non-*E. huxleyi* alignments are very short with low percentage identity, which means that they may be genuinely part of the *E. huxleyi* genome, or were identified as potentially resulting from horizontal gene transfer or symbiosis. This means that there may still be contamination present, but without more information, such as more sequencing data from an axenic culture, it is not possible to say for certain.

Future work on this assembly could include improvements from a larger HMW nanopore sequencing dataset, possibly from an axenic RCC1217 culture which would allow for contamination to be addressed and ruled out. The important next step is annotation, which could be done through the Joint Genome Institute using their fungal annotation pipeline (Min, Grigoriev, and Choi 2017) which uses a combination of gene prediction approaches to overcome the challenges posed by the complex structure of eukaryotic genes. Annotation would allow for phylogenetic analyses to improve understanding of the complex evolutionary

history of haptophytes, and provide insights into the genetic mutations which allow it to adapt to a wide range of conditions. The CCMP1516 genome and annotation (Read et al. 2013) provided valuable insights into the adaptability and evolutionary history of *E. huxleyi* and this assembly could allow for similarly ground-breaking insights, alongside comparative investigations between the two assemblies to answer questions around ploidy variation.

2.4.5 Potential research impacts of the RCC1217 assembly

This project resulted in the production of a haploid *E. huxleyi* RCC1217 genome assembly, which could have significant impacts for the field of phytoplankton research. The assembly is the second *E. huxleyi* assembly, with the first, the diploid CCMP1516 currently the only publicly available *Gephyrocapsa* genome assembly. The RCC1217 assembly is an improvement on the CCMP1516 assembly, with a far greater contiguity and BUSCO completeness. Other publicly available haptophyte genome assemblies were recently published, comprising one chromosome-level *Isochrysis galbana* (Riccio et al. 2022) assembly, with a genome size of 94 Mbp, and a scaffold level *Diacronema lutheri* (Hulatt, Wijffels, and Posewitz 2021) assembly with 1904 scaffolds, genome size 18 Mbp. There are also three contig-level assemblies. Therefore the RCC1217 is superior to the majority of publicly available haptophyte genome assemblies, although it is inferior to the *I. galbana* assembly.

The colonisation of a wide range of niches is key to *E. huxleyi*'s success; it is ubiquitous and dominant in many ecosystems, with one contributor to niche expansion believed to be the haplo-diplontic life cycle and resulting range of morphologies and calcification (Young et al. 2003). The RCC1217 *E. huxleyi* genome assembly will allow researchers to improve our understanding of the habitat requirements and preferences of haploid coccolithophores, and the complex carbon cycle contributions *E. huxleyi*, which alter depending on life cycle phase. Further work on the assembly would include annotation to which would also allow research into the genetic drivers for morphological and calcification differences between haploid and diploid coccolithophores. Functional annotation of the assembly would also allow for research into the genetic basis for phytoplankton speciation.

A further impact of this research is the demonstration of the utility for nanopore sequencing in complex genome assemblies. The initial nanopore assembly was around 1200 contigs prior to scaffolding and polishing, which is in itself an improvement on the CCMP1516 and *D. lutheri* assemblies, at a cost of around

£1200 for two flowcells. The sequencing yield was relatively low and at the time nanopore sequencing accuracy was relatively low, necessitating the production of a hybrid assembly. Advances in nanopore sequencing flowcells and software mean that yield and accuracy are both vastly improved, easily producing 10-20 Gbp per flowcell, and short-read polishing is becoming increasingly redundant. The use of scaffolding from Hi-C long range data is particularly for complex genomes, but with sufficiently long reads to span repeat regions, and areas of low complexity this too can increasingly be dispensed with. A key obstacle to this will be the development of reliable protocols for the extraction of extremely HMW DNA, which is especially challenging for phytoplankton, although promising developments include . Assuming successful HMW DNA extraction, it is likely that in the near future, nanopore sequencing on one or two flowcells coupled with a straightforward assembly process could produce a high quality, even chromosome-level assembly for complex eukaryotic phytoplankton, for a cost of around £600-120. This would lower the barrier for widescale production of high quality genome assemblies, with the potential to produce a vast resource for the study of phytoplankton to improve our understanding of their evolution, life cycle, and the complex interactions between them and their environment.

2.4.6 Conclusion

The polished assembly is of good quality compared to the CCMP1516 assembly in terms of contiguity, and completeness, and upon publication would be only the second coccolithophore genome to be available for research. The two assemblies are highly similar, and it is unlikely that identical contamination would be present in both, although neither culture was axenic before sequencing, which indicates that the contamination levels are likely to be low. This assembly has relatively low contiguity and completeness compared to many published genome assemblies for other species, especially those with less complex genomes. It compares favourably, however, to recently published genomes for other haptophytes which are also at contig level, with only one chromosome-level haptophyte assembly publicly available currently. The assembly has utility for structural analysis and taxonomic classification, and represents a baseline for future improvements. Future work to annotate the assembly would allow for investigation of *E. huxleyi* genes, allowing for insights into its evolution, adaptability, survival mechanisms, and life cycle. Analysis of this genome assembly could improve our understanding of haploid coccolithophores, and through comparison to CCMP1516, provide insights into the genetic basis for the haplo-diplontic life-cycle and its effects on coccolithophore ecology, This offers a significant step forward for our understanding of *E. huxleyi* and coccolithophores.

References

Bendif, E. M., B. Nevado, E. L. Wong, K. Hagino, I. Probert, J. R. Young, R. E. Rickaby, and D. A. Filatov (2019). "Repeated species radiations in the recent evolution of the key marine phytoplankton lineage Gephyrocapsa". In: *Nature communications* 10.1, pp. 1–9.

Chen, Y.-C., T. Liu, C.-H. Yu, T.-Y. Chiang, and C.-C. Hwang (Apr. 2013). "Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly". In: *PLoS ONE* 8.4. Ed. by Y. Xu, e62856. DOI: 10.1371/journal.pone.0062856.

Claros, M. G., R. Bautista, D. Guerrero-Fernández, H. Benzerki, P. Seoane, and N. Fernández-Pozo (Sept. 2012). "Why Assembling Plant Genome Sequences Is So Challenging". In: *Biology* 1.2, pp. 439–459. DOI: 10.3390/biology1020439.

Cornet, L. and D. Baurain (2022). "Contamination Detection in Genomic Data: More Is Not Enough". In: *Genome Biology* 23.1, p. 60. DOI: 10.1186/s13059-022-02619-9.

Cros, L., A. Kleijne, A. Zeltner, C. Billard, and J. Young (2000). "New examples of holococcolith–heterococcolith combination coccospheres and their implications for coccolithophorid biology". In: *Marine Micropaleontology* 39.1-4, pp. 1–34.

Cuvelier, M. L., A. E. Allen, A. Monier, J. P. McCrow, M. Messié, S. G. Tringe, T. Woyke, R. M. Welsh, T. Ishoey, J.-H. Lee, B. J. Binder, C. L. DuPont, M. Latasa, C. Guigand, K. R. Buck, J. Hilton, M. Thiagarajan, E. Caler, B. Read, R. S. Lasken, F. P. Chavez, and A. Z. Worden (2010). "Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton". In: *Proceedings of the National Academy of Sciences* 107.33, pp. 14679–14684. ISSN: 0027-8424. DOI: 10.1073/pnas.1001665107. eprint: https://www.pnas.org/content/107/33/14679.full.pdf.

Doktorovova, S., A. M. Silva, I. Gaivão, E. B. Souto, J. P. Teixeira, and P. Martins-Lopes (2013). "Comet Assay Reveals No Genotoxicity Risk of Cationic Solid Lipid Nanoparticles". In: *Journal of Applied Toxicology* 34.4, pp. 395–403. DOI: 10.1002/jat.2961.

Eccles, D., J. Chandler, M. Camberis, B. Henrissat, S. Koren, G. Le Gros, and J. J. Ewbank (2018). "De novo assembly of the complex genome of Nippostrongylus brasiliensis using MinION long reads". In: *BMC biology* 16.1, p. 6.

Falkowski, P. G., M. E. Katz, A. H. Knoll, A. Quigg, J. a. Raven, O. Schofield, and F. J. R. Taylor (2004). "The Evolution of Modern Eukaryotic Phytoplankton". In: *Science* 305.July, pp. 354–360. ISSN: 1095-9203. DOI: 10.1126/science.1095964.

Filatov, D. A., E. M. Bendif, O. A. Archontikis, K. Hagino, and R. E. Rickaby (2021). "The mode of speciation during a recent radiation in open-ocean phytoplankton". In: *Current Biology* 31.24, pp. 5439–5449.

Frada, M., I. Probert, M. J. Allen, W. H. Wilson, and C. de Vargas (2008). "The "Cheshire Cat" escape strategy of the coccolithophore Emiliania huxleyi in response to viral infection". In: *Proceedings of the National Academy of Sciences* 105.41, pp. 15944–15949.

Guillard, R. R. L. and J. H. Ryther (1962). "Studies of Marine Planktonic Diatoms: I. Cyclotella Nana Hustedt, and Detonula Confervacea (CLEVE) Gran." In: *Canadian Journal of Microbiology* 8.2, pp. 229–239. DOI: 10.1139/m62-029.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler (2013). "Quast: Quality Assessment Tool for Genome Assemblies". In: *Bioinformatics* 29.8, pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086.

Hulatt, C. J., R. H. Wijffels, and M. C. Posewitz (2021). "The genome of the haptophyte Diacronema lutheri (Pavlova lutheri, Pavlovales): A model for lipid biosynthesis in eukaryotic algae". In: *Genome Biology and Evolution* 13.8, evab178.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner (2019). "Assembly of Long, Error-Prone Reads Using Repeat Graphs". In: *Nature Biotechnology* 37.5, pp. 540–546. DOI: 10.1038/s41587-019-0072-8.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". In: *Genome research* 27.5, pp. 722–736. DOI: 10.1101/gr.215087.116.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg (2004). "Versatile and open software for comparing large genomes". In: *Genome Biology* 5.2, R12. DOI: 10.1186/gb-2004-5-2-r12.

LeMoigne, F. A. C., A. J. Poulton, S. A. Henson, C. J. Daniels, G. M. Fragoso, E. Mitchell, S. Richier, B. C. Russell, H. E. K. Smith, G. A. Tarling, J. R. Young, and M. Zubkov (June 2015). "Carbon export efficiency and phytoplankton community composition in the Atlantic sector of the Arctic Ocean". In: *Journal of Geophysical Research: Oceans* 120.6, pp. 3896–3912. DOI: 10.1002/2015jc010700.

Li, H. (2016). "Minimap and Miniasm: Fast Mapping and De Novo Assembly for Noisy Long Sequences". In: *Bioinformatics* 32.14, pp. 2103–2110. DOI: 10.1093/bioinformatics/btw152.

Manni, M., M. R. Berkeley, M. Seppey, and E. M. Zdobnov (2021). "Busco: Assessing Genomic Data Quality and Beyond". In: *Current Protocols* 1.12, nil. DOI: 10.1002/cpz1.323.

Miga, K. H., S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W.

Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T. A. G. Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy (2020). "Telomere-To-Telomere Assembly of a Complete Human X Chromosome". In: *Nature* 585.7823, pp. 79–84. DOI: 10.1038/s41586-020-2547-7.

Miller, J. R., P. Zhou, J. Mudge, J. Gurtowski, H. Lee, T. Ramaraj, B. P. Walenz, J. Liu, R. M. Stupar, R. Denny, et al. (2017). "Hybrid assembly with long and short reads improves discovery of gene family expansions". In: *BMC genomics* 18.1, p. 541.

Min, B., I. V. Grigoriev, and I.-G. Choi (2017). "Fungap: Fungal Genome Annotation Pipeline Using Evidence-Based Gene Model Evaluation". In: *Bioinformatics* 33.18, pp. 2936–2937. DOI: 10.1093/bioinformatics/btx353.

Nagarajan, N. and M. Pop (2013). "Sequence Assembly Demystified". In: *Nature Reviews Genetics* 14.3, pp. 157–167. DOI: 10.1038/nrg3367.

Paasche, E. (2001). "A Review of the Coccolithophorid Emiliania Huxleyi (Prymnesiophyceae), With Particular Reference To Growth, Coccolith Formation, and Calcification-Photosynthesis Interactions". In: *Phycologia* 40.6, pp. 503–529. DOI: 10.2216/i0031-8884-40-6-503.1.

Parra, G., K. Bradnam, and I. Korf (2007). "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes". In: *Bioinformatics* 23.9, pp. 1061–1067.

Phillips, N., C. M. Smith, and C. W. Morden (2001). "An Effective Dna Extraction Protocol for Brown Algae". In: *Phycological Research* 49.2, pp. 97–102. DOI: 10.1111/j.1440-1835.2001.tb00239.x.

Poulton, A. J., T. R. Adey, W. M. Balch, and P. M. Holligan (2007). "Relating Coccolithophore Calcification Rates To Phytoplankton Community Dynamics: Regional Differences and Implications for Carbon Export". In: *Deep Sea Research Part II: Topical Studies in Oceanography* 54.5-7, pp. 538–557. DOI: 10.1016/j.dsr2.2006.12.003.

Read, B. A., E. huxleyi Annotation Consortium, J. Kegel, M. J. Klute, A. Kuo, S. C. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J.-M. Claverie, S. Frickenhaus, K. Gonzalez, E. K. Herman, Y.-C. Lin, J. Napier, H. Ogata, A. F. Sarno, J. Shmutz, D. Schroeder, C. de Vargas, F. Verret, P. von Dassow, K. Valentin, Y. V. de Peer, G. Wheeler, J. B. Dacks, C. F. Delwiche, S. T. Dyhrman, G. Glöckner, U. John, T. Richards, A. Z. Worden, X. Zhang, and I. V. Grigoriev (2013). "Pan Genome of the Phytoplankton

Emiliania Underpins Its Global Distribution". In: *Nature* 499.7457, pp. 209–213. DOI: 10.1038/nature12221.

Reid, E. L., C. A. Worthy, I. Probert, S. T. Ali, J. Love, J. Napier, J. A. Littlechild, P. J. Somerfield, and M. J. Allen (2011). "Coccolithophores: Functional Biodiversity, Enzymes and Bioprospecting". In: *Marine Drugs* 9.4, pp. 586–602. DOI: 10.3390/md9040586.

Rescan, M., T. Lenormand, and D. Roze (2016). "Interactions between genetic and ecological effects on the evolution of life cycles". In: *The American Naturalist* 187.1, pp. 19–34.

Riccio, G., K. A. Martinez, A. Ianora, and C. Lauritano (2022). "De Novo Transcriptome of the Flagellate Isochrysis galbana Identifies Genes Involved in the Metabolism of Antiproliferative Metabolites". In: *Biology* 11.5, p. 771.

Rogers, S. O. and A. J. Bendich (1994). "Extraction of total cellular DNA from plants, algae and fungi". In: *Plant Molecular Biology Manual*. Ed. by S. B. Gelvin and R. A. Schilperoort. Plant Molecular Biology Manual. Dordrecht: Springer Netherlands, pp. 183–190. ISBN: 978-94-011-0511-8. DOI: 10.1007/978-94-011-0511-8_12.

Rokitta, S. D., P. Von Dassow, B. Rost, and U. John (2014). "Emiliania huxleyi endures N-limitation with an efficient metabolic budgeting and effective ATP synthesis". In: *BMC genomics* 15.1, pp. 1–14.

Ruan, Z., M. Lu, H. Lin, S. Chen, P. Li, W. Chen, H. Xu, and D. Qiu (2023). "Different photosynthetic responses of haploid and diploid Emiliania huxleyi (Prymnesiophyceae) to high light and ultraviolet radiation". In: *Bioresources and Bioprocessing* 10.1, p. 40.

Smits, T. H. M. (2019). "The Importance of Genome Sequence Quality To Microbial Comparative Genomics". In: *BMC Genomics* 20.1, p. 662. DOI: 10.1186/s12864-019-6014-5.

Stefan, C. P., A. T. Hall, A. S. Graham, and T. D. Minogue (2022). "Comparison of Illumina and Oxford Nanopore Sequencing Technologies for Pathogen Detection From Clinical Matrices Using Molecular Inversion Probes". In: *The Journal of Molecular Diagnostics* 24.4, pp. 395–405. DOI: 10.1016/j.jmoldx.2021.12.005.

Taylor, A. R., C. Brownlee, and G. Wheeler (2017). "Coccolithophore cell biology: chalking up progress". In: *Annual review of marine science* 9, pp. 283–310.

Tigano, A., T. B. Sackton, and V. L. Friesen (2017). "Assembly and RNA-free annotation of highly heterozygous genomes: The case of the thick-billed murre (Uria lomvia)". In: *Molecular Ecology Resources* 18.1, pp. 79–90. DOI: 10.1111/1755-0998.12712. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12712.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl (2014). "Pilon: an

Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". In: *PLoS ONE* 9.11, e112963. DOI: 10.1371/journal.pone.0112963.

Winter, A., J. Henderiks, L. Beaufort, R. E. M. Rickaby, and C. W. Brown (Mar. 2014). "Poleward expansion of the coccolithophore Emiliania huxleyi". In: *Journal of Plankton Research* 36.2, pp. 316–325. ISSN: 14643774. DOI: 10.1093/plankt/fbt110.

Young, J., M. Geisen, L. Cros, A. Kleijne, C. Sprengel, I. Probert, and J. Østergaard (2003). "A guide to extant coccolithophore taxonomy". In: *Journal of Nannoplankton Research, Special Issue* 1, pp. 1–132.

Zhang, H. and L. Cao (2016). "Simulated effect of calcification feedback on atmospheric CO2 and ocean acidification". In: *Scientific reports* 6.1, pp. 1–10.

3

Ship-Seq: The ups and downs of nanopore sequencing onboard a research ship

3.1 Introduction

This chapter presents work done on nanopore sequencing of polar ocean microbes during and following a research cruise. The fieldwork section covering sample collection, filtration, *in situ* sequencing and analysis and was carried out entirely by me, onboard the RRS Discovery in January and February 2019. The land-based DNA extractions were carried out by me. The multiplex nanopore sequencing of samples from the 12 stations was carried out by Darren Heavens at Earlham Institute in May 2020 due to covid-19 restrictions on lab occupancy. The Illumina sequencing was performed by Genomics Pipelines at Earlham Institute. I performed the single sample nanopore sequencing of samples from stations 5 and 11 and carried out the analysis on all of the sequencing data.

3.1.1 Research Cruise DY098: The Scotia Arc

The Ship-Seq project was conceived to test and evaluate the utility of *in situ* nanopore sequencing and analysis of polar ocean samples. This was a proof-of-concept experiment in using portable nanopore sequencing to monitor polar ocean microbial communities, particularly eukaryotic phytoplankton, which had not previously been attempted. Samples were collected on the British Antarctic Survey (BAS) research cruise DY098 on the RRS Discovery, see figure 3.1.1, which took place in the austral summer over January and February 2019 in the Southern Ocean and South Atlantic, including South Georgia and the South Sandwich Islands. Twelve stations were sampled, with three sequenced onboard and the rest retained for laboratory-based sequencing to provide deeper insights into the sampled microbiomes, and to help with evaluation of *in situ* sequencing

Figure 3.1.1: RRS Discovery at Port Stanley prior to departure. Photograph by Phil Keating

results. The area sampled is highlighted in figure 3.1.2, from which it can be seen that the samples were taken across the southern boundary of the SBACC between South Georgia and the South Sandwich Islands, covering a range of ecological niches.

The Southern Ocean, usually defined as south of 35 °S and encircling the continent of Antarctica, is characterised by a series of fronts, or areas of strong flow which act as boundaries of different masses of water with their own distinct temperature, salinity, and nutrient levels (Mills 2005). Fronts in the Southern Ocean often coexist with "jets" which are strong currents that contribute to the Antarctic Circumpolar Current (ACC) and serve to further divide the watermasses between fronts. This means that within the Southern Ocean there are a series of clearly delineated niches which are able to support a range of different ecosystems (Grant et al. 2006). Figure 3.1.2 shows the ACC and its direction of travel around the Southern Ocean, with the Sub-Antarctic Front (SAF) and the southern boundary of the ACC (SBACC) which are biologically and biogeochemically important. Climate change is already contributing to changes to Southern Ocean fronts and currents, with fronts moving towards the south pole, which could have devastating impacts on the entire ecosystem, from phytoplankton to whales and albatrosses (Constable et al. 2014).

The Scotia Sea is an important area of the Southern Ocean, it contains a number of delineated fronts and is characterised by a strong flow brought about by the physical geography of the Drake Passage and the Scotia Ridge which combine to redirect the ACC. Coupled with the high winds common to the Southern Ocean, these conditions result in a series of fronts delineated by areas of relative calm

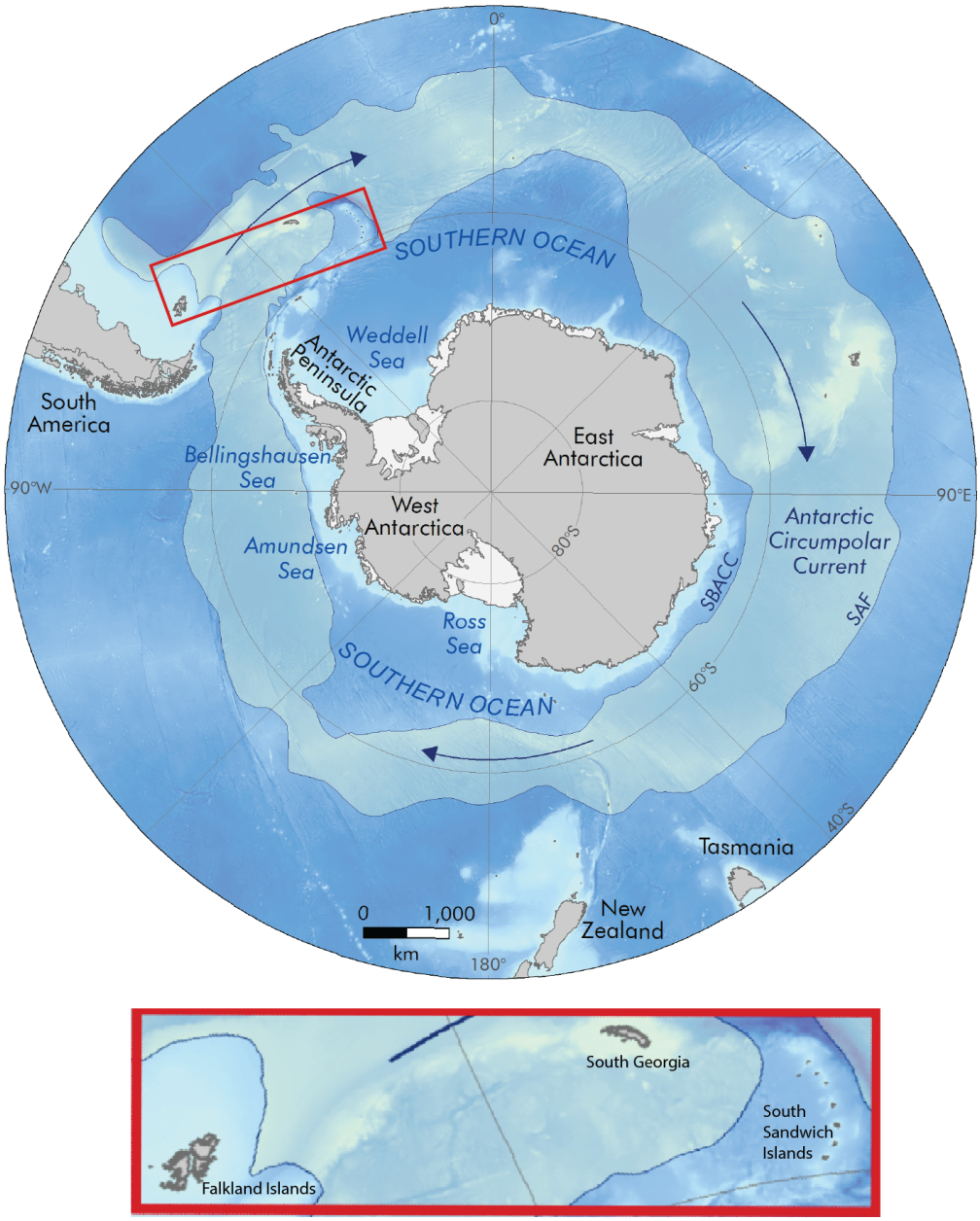


Figure 3.1.2: Map showing Antarctica, the ACC, the SAF and SBACC, and the direction of current. The sampling area for the DY098 cruise is highlighted in red and shown expanded in the inset map beneath. The DY098 cruise travelled from the Falkland Islands to South Georgia, and the South Sandwich Islands, before returning to the Falkland Islands. It can be seen that the sampling area crosses the SBACC. Adapted from map produced by the Mapping and Geographic Information Centre, British Antarctic Survey, 2021. Bathymetry data from the GEBCO Compilation Group (2021) GEBCO 2021 Grid (doi:10.5285/c6612cbe-50b3-0cff-e053-6c86abc09f8f). Coastline data from the SCAR Antarctic Digital Database, accessed 2021.

(Sokolov and Rintoul 2009). These fronts result in distinct habitats with individual biomes (Hunt and Hosie 2005).

Generally the Southern Ocean displays high nutrient, low chlorophyll conditions, thought to be due to iron-limitation preventing the accumulation of phytoplankton blooms (Banse 1996). The Scotia Sea is an exception to this, with higher levels of primary productivity (Whitehouse et al. 2012). There are regular large blooms observed around South Georgia, however, likely due to the presence of iron from the shelf, and diatoms are particularly dominant (Schlosser et al. 2018). Regular large blooms mean that this area is an important contributor to CO_2 sequestration and nutrient cycling, as nutrients are released by the phytoplankton and dying phytoplankton sink to the ocean floor. The blooms generally peak in December, and sometimes again between January and April, which is hypothesised to correspond to the diatom growth depleting silica availability in the first peak and rebounding later when silica availability increases again (Borrione and Schlitzer 2013).

Previous analysis in the Scotia Arc had identified *Emiliania huxleyi* in the north of the region and *Fragilariopsis* species towards the south, correlated to sea surface temperature and silicate concentration, with indications that they contributed significantly to the phytoplankton population (Hinz et al. 2012). As they are an important contributor to the carbon cycle, and at particular risk from climate change, these became a particular focus of this study.

Samples were taken at the locations shown in figure 3.1.3, covering both open ocean and shoreline sites, and encompassing a range of nutrient and temperature conditions.

3.1.2 DNA sequencing for monitoring diatoms and polar phytoplankton

We know surprisingly little about phytoplankton diversity and abundance. Most studies have focussed on laboratory cultures, introducing bias toward culturable species into diversity estimates, which is problematic since the vast majority of phytoplankton have not been cultured and are thought to be unculturable. The Tara Oceans project showed that some phytoplankton previously assumed to be minor contributors to their ecosystems were in fact extremely important, highly abundant species with high diversity (Bork et al. 2015). Polar oceans are even less well characterised than most, and polar phytoplankton, including diatoms, can be extremely difficult to grow in culture. It is clear, therefore, that culture-based methods for studying phytoplankton are insufficient to provide

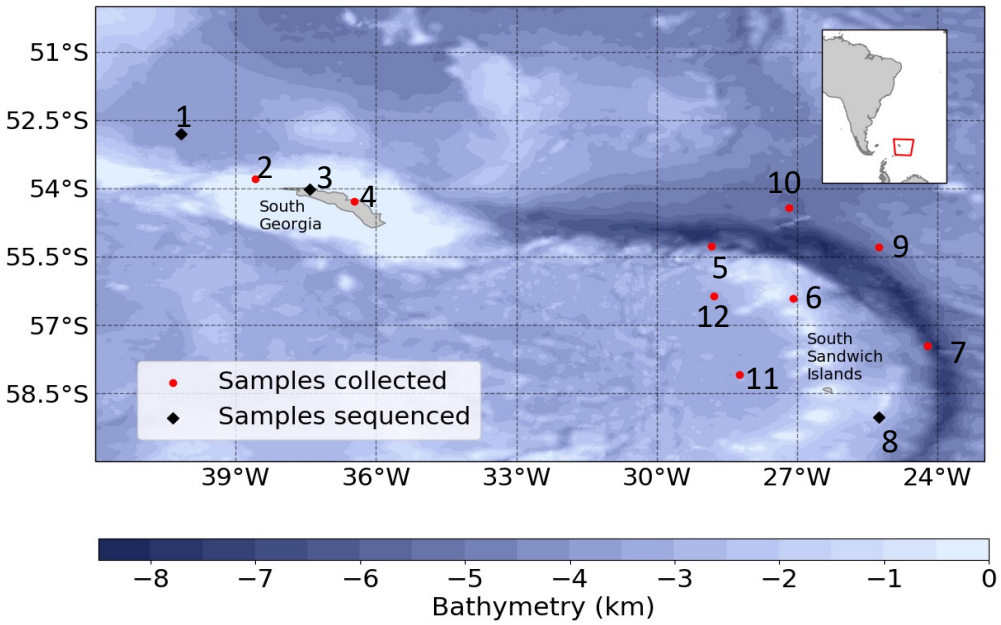


Figure 3.1.3: Map showing sample site locations. Black points indicate sampling stations where samples were used for *in situ* sequencing. Red points indicate sampling stations that were not used for *in situ* sequencing but were stored for later analysis.

a complete understanding of the diversity and community structures of polar ocean microbiomes. DNA sequencing studies remove the reliance on culturing to investigate phytoplankton communities, providing a more accurate picture of populations and diversity by identifying species present based on their genomes.

Standard metagenome studies have involved sampling in the field, storing samples and transporting them back to base, and then sending samples to a dedicated laboratory for metabarcode sequencing on an NGS platform such as Illumina to identify species. Challenges associated with using this method to study polar phytoplankton communities include the high cost, the potentially long storage and transit times if samples are collected in the Southern Ocean, and the inability to tell whether sampling has been successful during the sampling period. Metabarcoding is an effective method for species identification but does not allow for more in depth analyses, and short-read sequences are not ideally suited to *de novo* assembly and analysis of complex genomes. The portability of the ONT MinION allows researchers to perform sequencing immediately after sampling, preventing sample deterioration. Sequencing could be used to direct sampling efforts based on species identified, or to determine whether sampling in a given location has been sufficient. As such, nanopore sequencing allows researchers of polar phytoplankton to better understand which species are present, how they are responding to climate change, how continued climate change is likely to

affect them, and how changes to their populations are likely to affect the wider ecology. There are challenges associated with this, however, including the extraction of sufficient high molecular weight (HMW) DNA from polar phytoplankton. A further challenge is the analysis of the data produced through DNA sequencing, to determine which species are present and to gain a better understanding of the evolutionary history and phylogenetics of polar phytoplankton both in the field and in standard laboratories. Genome assemblies are especially useful for identifying species in metagenomic data but currently there are only around 10 fully assembled diatom genomes publicly available (Tirichine, Rastogi, and Bowler 2017). More genome assemblies are under production, and those numbers will increase as costs decrease and techniques improve but this leaves a huge number of species of phytoplankton as yet unobserved and unidentified, particularly in polar oceans where there has been relatively less intensive research. It is clear that only a tiny fraction of the phytoplankton diversity has been sequenced and there is a huge gap in our knowledge and understanding of phytoplankton, and polar phytoplankton especially.

3.1.3 Polar oceans

The polar ocean microbiome is hugely diverse and contributes significantly to critical ecological processes such as global biogeochemical cycles and primary production. Polar oceans are a unique habitat with low temperatures, extreme seasonal variation in irradiance, and seasonal sea-ice expansion and contraction, and the phytoplankton which live there are specifically adapted to these conditions (Cota 1985; Fiala and Oriol 1990; Mock and Valentin 2004; Ryan, Ralph, and McMinn 2004; Boyd 2002).

3.1.4 Psychrophiles

Low temperatures are a significant barrier to life, they reduce physical and chemical efficiency of cellular processes and survival under these conditions requires a range of specific evolutionary adaptations. Metabolism, diffusion rates, membrane permeability, and cellular structural integrity are all negatively affected by low temperatures. Low temperatures also result in increased viscosity of water which, combined with decreased diffusion rates and enzyme kinetics can impair uptake of essential substrates (Rodrigues and Tiedje 2008). Cold environments often present stressors other than temperature, for example in polar oceans there are extremes of irradiance, high salinity, excessive UV radiation and relatively

low levels of important nutrients such as iron (Cota 1985; Fiala and Oriol 1990; Mock and Valentin 2004; Ryan, Ralph, and McMinn 2004; Boyd 2002) as well as freezing temperatures. In order to survive in these conditions, polar phytoplankton species have developed a variety of adaptive traits including: cryoprotectants such as antifreeze and cold shock proteins; membrane modifications; cold-adapted enzymes; and sophisticated modulations of photosynthetic apparatus (Mock and Kroon 2002; Tehei and Zaccai 2005; Morgan-Kiss et al. 2005), such species are known as psychrophiles.

Psychrophilic diatoms are particularly dominant in polar oceans, due both to the high silica concentration and their adaptability which has allowed them to successfully and rapidly develop strategies for survival in extreme conditions (Armbrust 2009). Relatively little is known about polar diatoms, with most research limited to a few species. Research on *F. cylindrus* has demonstrated the mechanisms by which diatoms have adapted to dominate the polar oceans and given insights into the genetic basis for diatom adaptability. The *F. cylindrus* genome has genes coding for a range of proteins not found in temperate diatom species which have been linked to functions including ice-binding, anti-freeze, and Fe-fixing. These are likely to be critical for psychrophilic survival, forming the basis of life in polar oceans. F. cylindrus was also found to display differential expression of certain alleles under stress conditions. This could increase the adaptability of diatoms by allowing genes to be expressed or not expressed only in the presence of given environmental stressors (Mock et al. 2017).

3.1.5 Climate change and polar phytoplankton

Ocean temperature affects a range of factors which determine the species which can inhabit an environment, including stratification, vertical mixing, and wind and ocean currents. Stratification of the ocean layers is particularly pronounced in warm waters compared to cold, and there is reduced mixing between layers in warm water. This results in reduced nutrient flux from one layer to another and can result in reduced nutrient concentration at the surface. Phytoplankton generally congregate in the pycnocline, near the surface, as they rely on sunlight for photosynthesis, so reduced nutrient content here limits growth. Changes to currents also affect nutrient availability through upwellings, where deep, nutrient rich water is brought to the surface (Hoegh-Guldberg and Bruno 2010). These features result in areas of high and low productivity, and climate change is causing expansion in areas of low productivity (oligotrophic zones) (Fernández-González et al. 2022).

Polar oceans and psychrophilic phytoplankton are at particular risk from climate change, as polar oceans are disproportionately warming (Bindoff et al. 2007). As temperatures rise, the effects on psychrophilic diatoms and other phytoplankton are likely to be negative. It is unlikely that psychrophilic sea-ice species will be able to adapt, as populations with cold-adaptations have less genetic redundancy and so may not be able to adapt to changing temperatures (Mock and Kirkham 2012). Studies have shown that tropical diatoms have been able to adapt to increased ocean temperatures, albeit with trade-offs in ability to withstand high irradiance, and reductions in photosynthetic efficiency and growth rate. This would result in reduced competetive fitness and, consequently, researchers predict a steep decline in phytoplankton diversity as a result of climate change (Jin and Agustí 2018). It is expected that global phytoplankton primary production will decrease as climate change intensifies, although most models predict that primary production will increase in the Southern Ocean (Laufkötter et al. 2015). In order to determine the likely effects of climate change on phytoplankton populations, it could be useful to compare polar and temperate or tropical diatoms (Mock and Kirkham 2012).

It is also important to consider the potential effects of reduced phytoplankton populations on climate change, through the biological carbon pump. Phytoplankton convert CO_2 to organic carbon, much of which is cycled through the ecosystem and re-released but when phytoplankton die and sink, rather than being consumed by zooplankton, the carbon sinks with them (Katz et al. 2005). This means that carbon is stored, effectively forever, in the sea-bed as opposed to in the atmosphere, preventing it contributing to the greenhouse effect. In this way, phytoplankton currently act as a mitigating factor against climate change, especially in polar oceans which are disproportionate contributors. There is wide variation in carbon export contribution between diatom species and this is not yet well understood or described. As a result, it is difficult to quantify the contribution of diatoms to carbon export and therefore it is unclear what effect reduced diatom populations would have on climate change (Tréguer et al. 2018). It is clear from this that good models of how climate change will affect phytoplankton populations and the likely ramifications of this on climate change, are essential to determining the likely impact of climate change on our oceans and the Earth as a whole.

In order to fully understand the contributions and impacts of polar phytoplankton, DNA sequencing studies are essential. Long-read metagenomic sequencing studies offer the ability to capture the complex interactions and community composition, while portable sequencing removes the need for storage, transport, and long wait times before results. The aim of Ship-Seq, therefore, was to

establish the utility of *in situ* nanopore sequencing and real-time analysis of polar ocean communities onboard research ships, and provide insights into polar microbe communities through *in situ* and land-based sequencing and analysis.

3.2 Methods

3.2.1 Protocol

The protocol devised for this set of experiments is shown in figure 3.2.1. This was originally intended as a pilot study with improvements to be made based on testing in the field, prior to a second round of fieldwork. Unfortunately, the second field trip was cancelled due to restrictions related to the Covid-19 pandemic.

3.2.2 Nanopore sequencing on the RRS Discovery

Sampling

12 seawater samples were collected from the chlorophyll maxima using a Seabird SBE19plus current temperature depth profiler (CTD) equipped with the following sensors: Temperature, Conductivity, Digiquartz Pressure, Dissolved O2 , Fluorimeter, Altimeter, UWIRR PAR, DWIRR PAR, Backscatter, Transmissometer, 20L water samplers, LADCP, see figure 3.2.2. The 20 l NISKIN sampling bottles were used to collect seawater at the chlorophyll max depth, and temperature, depth, and conductivity measurements were recorded. Sample collections were carried out at stations detailed in table 3.1 and shown in figure 3.1.3. 100 litres of seawater were collected from the CTD in 10 or 20 litre carboys, rinsed with seawater from the Niskin bottle being used for collection. These carboys were placed in the 2 ℃ cold room until filtration could take place (night collections were filtered the following day as the process took approximately 10 hours). Filtration was carried out in the 2 ℃ cold room using 142 mm cellulose acetate filters with 0.45 µm pore size were used in a Sartorius pressure filter holder (filtration stand). A peristaltic pump was connected to the filtration stand to pump the water from the carboys through the filter until the filter had clogged (tube connectors would no longer stay attached, or the pump could not pump water through). The filtration equipment is shown in figure 3.2.3. The filter was then removed from the stand and cut into 8 pieces and either frozen at -80 ℃ or used immediately for DNA extraction.

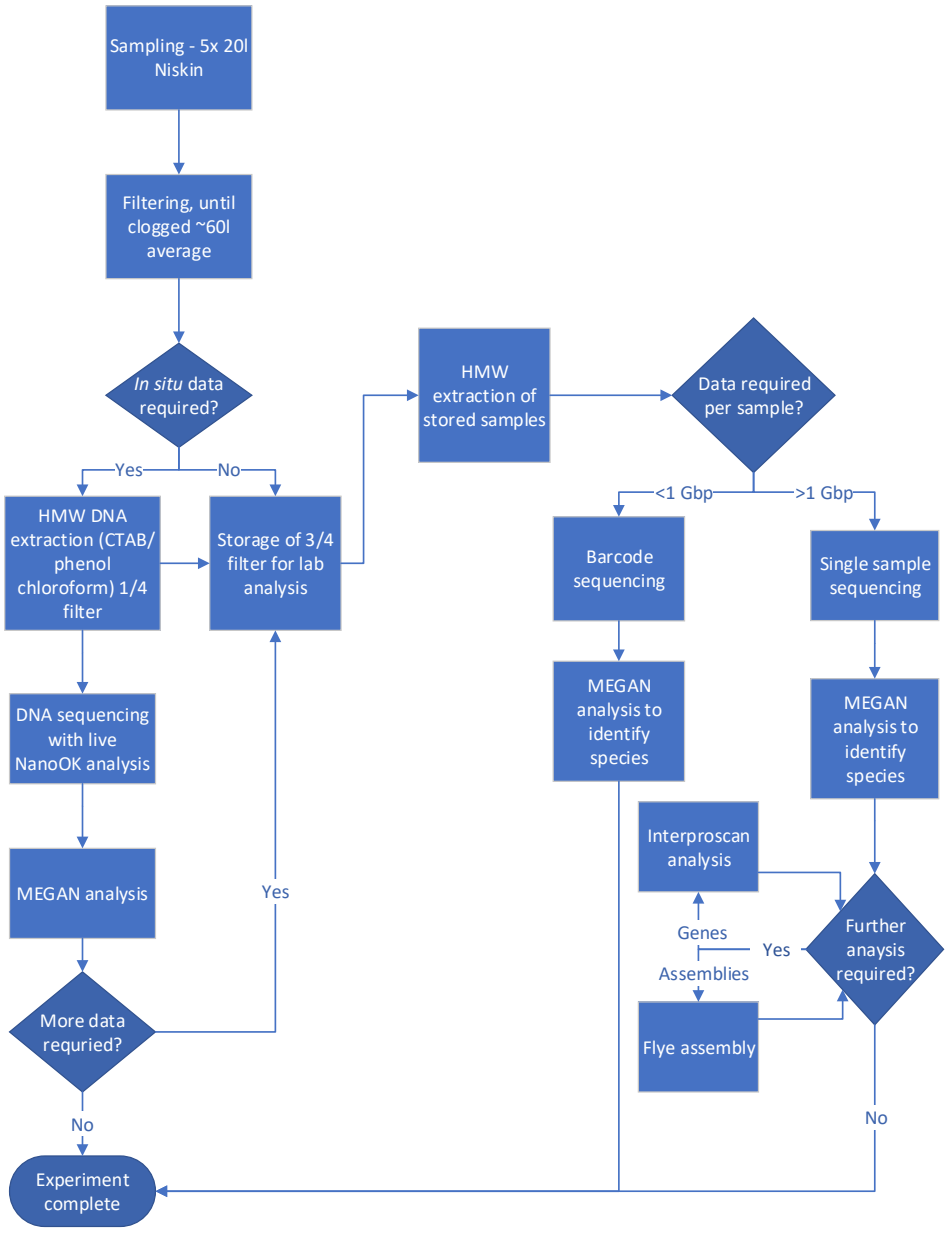


Figure 3.2.1: Workflow diagram showing the protocol for sampling, DNA extraction, sequencing, and analysis onboard a research ship. Squares represent processes, diamonds represent decision points.

Figure 3.2.2: The CTD being brought in by technicians on the RRS Discovery after sampling

Table 3.1: Summary of sample collections showing station number, latitude and longitude of station, depth at which the NISKIN bottles were fired, the date sampling took place, the volume filtered and whether a sample from the station was sequenced onboard.

Station	Lat	Long	Depth (m)	Date	Volume (l)	*In situ* sequencing?
1	-52.8	-40.2	30	05/1/19	80	Yes
2	-53.8	-38.6	20	06/1/19	75	No
3	-54.0	-37.4	50	11/1/19	80	Yes
4	-54.3	-36.4	8	15/1/19	60	No
5	-55.3	-28.8	80	26/1/19	90	No
6	-56.4	-27.1	65	27/1/19	65	No
7	-57.4	-24.2	42	30/1/19	90	No
8	-59.0	-25.3	22	31/1/19	15	Yes
9	-55.2	-25.2	28	02/2/19	65	No
10	-54.3	-27.1	33	03/2/19	100	No
11	-58.1	-28.2	63	06/2/19	90	No
12	-56.2	-28.4	59	07/2/19	100	No

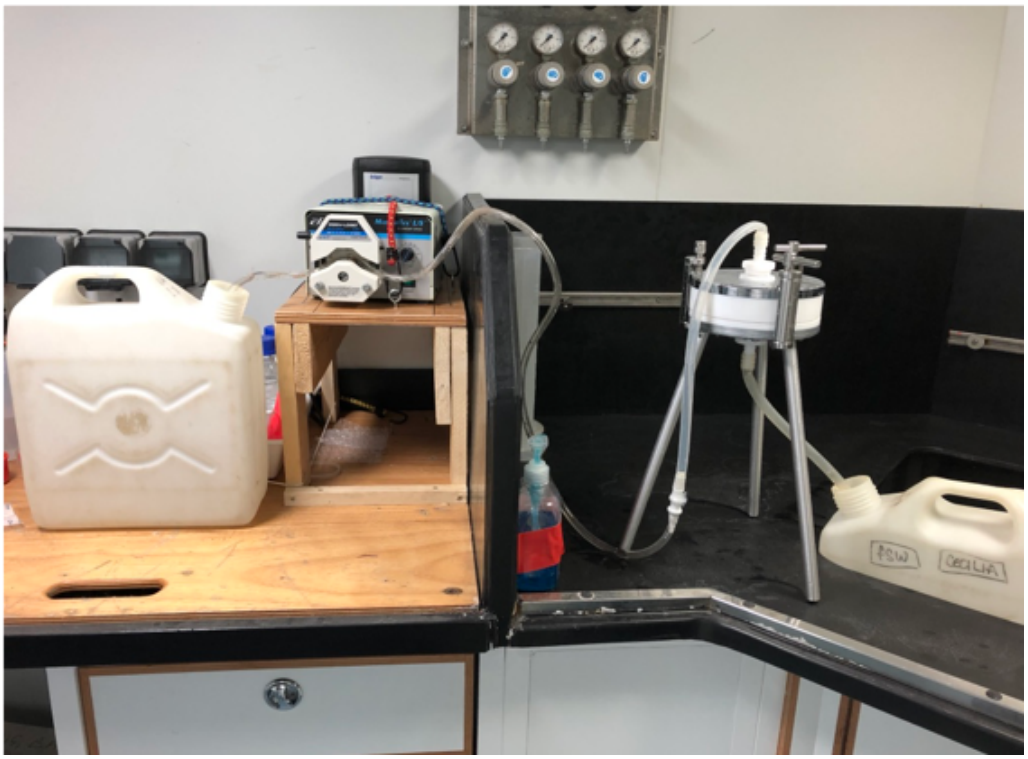


Figure 3.2.3: Filtration stand and peristaltic pump arrangement set up in the 2 ℃ cold room. 10 litre carboy containing seawater from CTD with flexible tube carrying water via a peristaltic pump to a filtration stand containing a 142 mm diameter, 0.45 µm filter. The water is pumped through the filter and drains into the sink, once the filter is clogged or all of the water collected has been filtered, the filter is removed and cut into 8 pieces which are either stored at -80 ℃ or immediately processed for DNA sequencing.

DNA extraction and sequencing

CTAB DNA extractions were carried out using filters from Stations 1, 3, and 8 as follows. In a fume hood, 4 x 1/8 of a filter from each sample were placed in 4 x 2 mL Eppendorfs containing 1.5 mL CTAB, 150 µl 2-mercaptoethanol and 15 µl 10% SDS. These were placed in a thermomixer with 2mL tube attachment at 65 ℃ for 4 hours. In a fume hood, 1 mL of the incubated CTAB mixture was added to 6 clean 2 mL Eppendorf tubes, the used tubes containing the filter were discarded, and 1 mL of phenol:chloroform:isoamyl alcohol (25:24:1) was added to each tube. This mixture was centrifuged in a microfuge until clear phase separation was present (around 30 minutes). In a fume hood, the upper phase was removed to 4 clean 2 mL Eppendorf tubes and the lower phase discarded, and 750 µl (2/3 vol) -20 ℃ isopropanol was added to each tube and left to sit for 15 minutes. This mixture was then centrifuged until a pellet was formed (>1hr). In a fume hood, the supernatant was discarded and the pellet was washed with 70% EtOH. The pellet was then allowed to airdry before resuspension in 50 µl low TE buffer. The sample was either immediately used for DNA library preparation

for MinION sequencing or frozen at -80 ℃. The DNA samples are summarised in table 3.2 below:

Table 3.2: Samples sequenced onboard the ship, showing the Station sampled, the mean DNA extracted per sample across the 4 replicates, and the number of active pores reported for the flowcell. The flowcell pore count at the start of sequencing influences the potential yield of the sequencing run.

Station	Mean DNA (µg)	Pores
1	8.5	1010
3	1.82	646
9	2.19	649

DNA library preparation and MinION sequencing was carried out according to Oxford Nanopore Technologies protocol 1D Genomic DNA by Ligation (SQK-LSK109) – Version GDE_9063_v109_revB_23May2018. The optional DNA shearing step (Covaris g-tube) was omitted. An offline version of MinKNOW 1.15.6 (Bream version 1.15.10.20 and GUI version 2.2.16), was provided upon request by Oxford Nanopore Technologies to allow sequencing to commence when there was no internet connection. Figure 3.2.4 shows a sequencing experiment underway, with the equipment for DNA extraction, library prep, DNA sequencing, and real-time analysis

Real-time analysis of the Nanopore MinION data was performed using NanoOK RT version 1.27 (Leggett et al. 2015). A network cable was used to connect 2 laptops and a shared folder was created which both could read/write to. Laptop 1 ran MinKNOW and was connected to the MinION, and laptop 2 ran NanoOK RT. A script was used to synchronise the sequencing data from the MinKNOW working folder on laptop 1 to the shared folder where NanoOK RT would detect the arrival of new reads. NanoOK RT performed BLAST-based analysis using a pre-downloaded local copy of the NCBI nt database. The NanoOKReporter GUI was used for real-time reporting of species abundance while sequencing proceeded. Data was stored on 2 external 1TB hard-drives, with one copy from the shared folder and one from the MinKNOW working folder of Laptop 1.

On-board the ship there was limited internet access and no facility for high performance computing, so analysis was limited to NanoOK RT identification of what was in the samples.

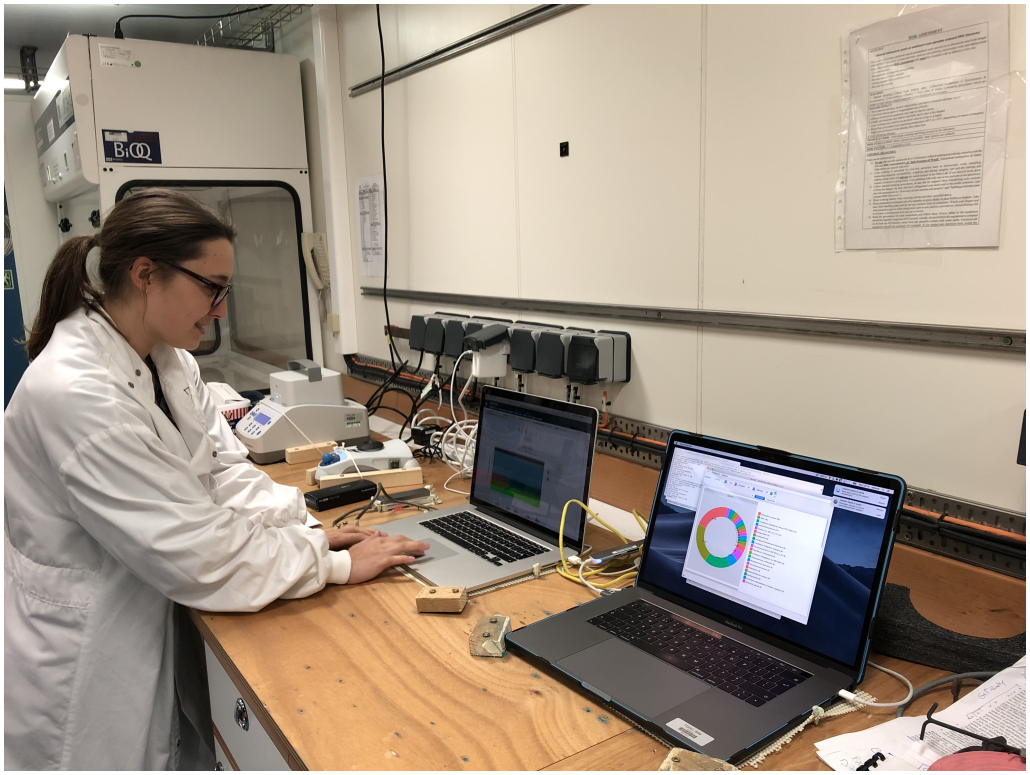


Figure 3.2.4: *In situ* nanopore sequencing experiment with real-time analysis onboard the RRS Discovery. Visible equipment from left to right: fume hood, heat block, vortexer, ONT MinION, sequencing laptop, analysis laptop.

Further analysis

Further analysis of the *in situ* sequencing data was performed upon return to Norwich. Porechop (v0.2.1), `https://github.com/rrwick/Porechop` was used to remove adapters which could have resulted in erroneous BLAST-nt hits, and BLAST re-run using a newer version of the nt database (downloaded on 21/02/2019). The resulting output was analysed using NanoOK RT as above. MEGAN (Huson et al. 2007) was used, to investigate and compare distributions between samples. A minimum support of 0.1% was used for the lowest common ancestor assignment to reduce unsupported BLAST-nt hits and increase confidence in taxa assognments.

3.2.3 Nanopore sequencing of 12 Southern Ocean Samples in the Laboratory

Nanopore sequencing of samples from all 12 stations

As discussed in Chapter 4, a range of improved DNA extraction techniques were tested to optimise the DNA extraction and sequencing workflow. A modified CTAB DNA extraction protocol was selected for the land-based DNA sequencing of samples collected on the DY098 cruise.

In a fume hood, 4 x 1/8 of a 142mm filter containing phytoplankton from each of 12 stations sampled in the DY098 cruise were placed in 2 mL Eppendorf tubes containing 1.5 mL CTAB, and 15 µl 10% SDS. 125 µL DNeasy beads was added and the tube placed in the PowerLyzer and homogenised for 5 min at 2000 RPM, before incubation in a thermomixer with 2mL tube attachment at 65 ℃ for 4 hours. In a fume hood, 1 mL of the incubated CTAB mixture was added to a clean 2 mL Eppendorf tube, the used tubes containing the filter were discarded, and 1 mL of chloroform kept at -18 ℃ was added to each tube. This mixture was centrifuged in a microfuge until clear phase separation was present (around 30 minutes). In a fume hood, the upper phase was removed to 4 clean 2 mL Eppendorf tubes and the lower phase discarded, and 750 µl (2/3 vol) -20 ℃ isopropanol was added to each tube and left to sit for 15 minutes. This mixture was then centrifuged until a pellet was formed (>1hr). In a fume hood, the supernatant was discarded and the pellet was washed with 70% EtOH. The pellet was then allowed to airdry before resuspension in 50 µl low TE buffer. The sample was then immediately used for DNA library preparation for Nanopore MinION sequencing.

DNA library preparation and Nanopore MinION sequencing was carried out according to Oxford Nanopore Technologies protocol 1D Genomic DNA by Ligation (SQK-LSK109) with barcoding expansion kits EXP-NBD104 and EXP-NBD114. The optional DNA shearing step (Covaris g-tube) was omitted. The 12 samples were barcoded and sequenced on a single flowcell on a GridION, with basecalling performed using MinKNOW and demultiplexing by Guppy (v3.2.8).

In depth sequencing was performed for Station 5, using the DNA extraction protocol described above, with single sample library preparation using the same Ligation kit without barcoding. Each sample was sequenced on a separate flowcell on a GridION, with basecalling performed using MinKNOW and demultiplexing by Guppy (v3.2.8).

The resulting data was matched against the NCBI nt database using BLAST and used for taxonomic analysis with MEGAN, as outlined for the ship-based sequencing. The reads were also used for producing metagenomic assemblies.

The vegan package in R was used to determine the statistical significance of metadata variables using non-metric multidimensional scaling. Linear models were produced in R to analyse the correlation between metadata variables and the abundance of diatoms. MEGAN was used to produce rarefaction curves investigate and visualise correlation of species occurrence to metadata variables.

Illumina sequencing

DNA extractions were carried out as for nanopore sequencing and the resulting DNA used for overlapping 2x250 paired-end Illumina sequencing at the Earlham Institute. The Earlham Institute Genomics Pipelines team performed quality checking and clean-up before library preparation, pooling and sequencing the samples on a NovaSeq.

The resulting sequence data was analysed using FastQC (Andrews 2010) and adapter sequences were trimmed using Trimmomatic (Bolger, Lohse, and Usadel 2014). Overlapping reads were merged using FLASH (Magoc and Salzberg 2011) to give a range of read lengths including a large number over 400 bp. Of these merged reads, 200,000 over 300 bp in length were randomly subsampled from each sample to use for classifcation. These were BLASTed against the NCBI nt database (download on 21/05/2021) and used for taxonomic analysis with MEGAN, as outlined above.

Assemblies

Metagenomic assemblies were produced from the land-based Nanopore sequencing data using metaFlye (Kolmogorov et al. 2020) and the resulting assemblies BLASTed against the nt database to identify the resulting assemblies and assess quality. Matches were filtered using cut-offs of 90% identity and length of greater than 1000 bp, and the best hit for each read was used for further investigation.

De novo specific organism assemblies were produced from the land-based Nanopore sequencing data using Flye (Kolmogorov et al. 2019) and the resulting assemblies were aligned against reference genomes using MiniMap2 (Li 2018) to

assess assembly quality. Alvis (Martin and Leggett 2021) was used to visualise alignment.

Metagenomic assemblies were produced from the Illumina sequencing data using metaSPAdes (Nurk et al. 2017) and the resulting assemblies classified using the BLAST nt database to identify the resulting assemblies and assess quality.

Assembly-free functional annotation

Interproscan (Jones et al. 2014) was used to search the Pfam database using the trimmed reads from the land Nanopore sequencing data collected onboard DY098. This produced a list of identifiers for each read with potential genes based on the nucleotide sequence. Identifiers included Pfam IDs which were extracted from this list used to perform correlation analyses and identify whether there was any statistically significant variation in Pfam IDs across the different sample sites. Pfam IDs were also correlated against metadata collected during the DY098 cruise.

3.3 Results

3.3.1 Nanopore sequencing on the RRS Discovery

DNA extraction and nanopore sequencing

Three samples were sequenced successfully onboard the DY098 cruise, summarised in table 3.3. Station 1 produced around 300 Mbp of sequenced DNA over 0.5 million reads. Station 3 (Rosita) produced approximately 1.5 Gbp of sequenced DNA over 1 million reads. Station 9 produced around 2 Gbp sequenced DNA in 1.8 million reads, of which over 1.2 million were longer than 20 kbp. The N50 for the samples ranged from 690 bp to 2000 bp.

Real-time analysis of sequencing data provided species level identification in NanoOK RT for the three samples. After around 2 hours of sequencing the overall species distributions remained fixed with only small changes to the proportions of each species. Across the three samples a similar proportion of reads were unknown, with approximately 85% of reads assigned a taxonomic match at species level. Figure 3.3.1, below shows the NanoOK RT output while it was running during the sequencing of sample 1.

Table 3.3: Run metrics for the ship-based samples showing the yield in Gbp, mean length, longest read length, and N50 in bp, number of reads, and number of reads longer than 20kbp for each sample

Station	Yield (Gbp)	Mean	Longest	N50	Reads	Reads >20 kbp
1	0.28	567	51142	691	45094	373772
3	1.51	1460	58852	2019	1035055	185256
8	2.05	1136	71806	1556	1805691	1280761

Taxonomic analysis

In each sample the largest hit (>30%) was, somewhat surprisingly, *Melanaphis sacchari*, a sugarcane aphid, as can be seen in figure 3.3.1. During the ship-based sequencing experiments it was assumed that these were likely due to adapter contamination of the *M. sacchari* genome assembly in the NCBI nt database and they were ignored. Investigation after the cruise confirmed that these matches were due to 30-32 bases at the beginning of the associated reads which matched an old nanopore library adapter, and porechop was run to remove the *Melanaphis sacchari* matches. Increasing the minimum match length in NanoOK RT would also have removed these classifications.

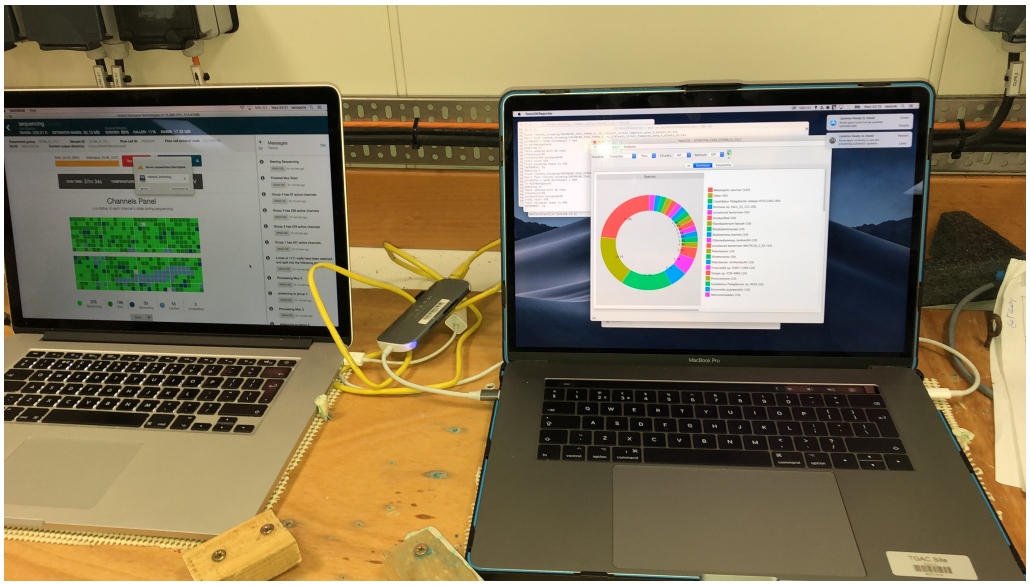


Figure 3.3.1: MinKnow and NanoOK RT running during sequencing of Station 1

The results of the sequencing data after the removal of erroneous *M. sacchari* hits are shown in figure 3.3.2 below. Station 1 in the open ocean and collected at the northernmost station on the cruise was over 50% '*Candidatus* Pelagibacter', part of a clade of highly abundant bacteria found worldwide in freshwater and salt water. *Emiliania huxleyi* was detected at this station. Station 3 in the harbour

at South Georgia, was well over 60% C. pelagibacter, and was dominated by prokaryotes and archaea, in contrast to the open ocean samples. Station 8 in open ocean at the southernmost sampling station, was made up of just under 50% C. pelagibacter and had the greatest overall population diversity, greatest abundance of eukaryotic species and the greatest abundance and diversity of diatom species.

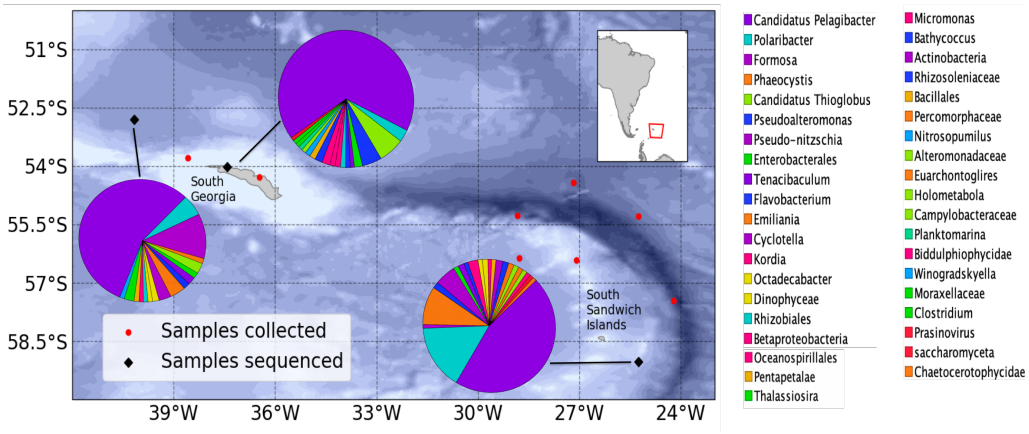


Figure 3.3.2: Map and pie charts showing species distribution at sample stations 1, 3, and 8. Pie charts produced using MEGAN, map produced using Cartopy.

3.3.2 Nanopore sequencing of 12 Southern Ocean samples in the laboratory

DNA extraction and nanopore sequencing of samples from all 12 stations

The extraction yields, mean read length, longest read length, N50, total number of reads, and reads longer than 20 Kbp are shown in table 3.4 below. Figure 3.3.3 shows the TapeStation output plot for station 5, which describes the molecular weight distribution of the extracted DNA plotted as molecular weight against sample intensity (FU). TapeStation outputs for all samples can be seen in Appendix A. Station 5 has a sharp first peak followed by a second peak of nearly 700 FU at around 27.5 kbp. Other stations all had a first peak of very short DNA at around 100 bp with an intensity of approximately 500 FU, which then fell with a smaller second peak or plateau between 10 and 25 kbp. Station 5 had the highest molecular weight DNA, with the least noise at lower weights, which is why it was selected for in depth sequencing.

Samples 11 and 12 yielded the most sequence data with 1.3 and 1.5 Gbp respectively; all other samples yielded in the range 0.3-0.7 Gbp. 8 out of 12 samples had a read N50 of between 2 and 5 Kbp, with 4 between 5 and 8kbp, and one at over 10 kpb. The yield per sample was equivalent to the yield achieved

Table 3.4: Run metrics for the laboratory-sequenced samples showing the yield in Gbp, mean length, longest read length, and N50 in bp, total number of reads, and number of reads longer than 20kbp for each sample

Station	Yield (Gbp)	Mean	Longest	N50	Reads	Reads >20 kbp
1	0.578	2302	87375	6025	243589	2260
2	0.265	1002	111209	2143	244666	659
3	0.662	4554	93769	8274	142932	2339
4	0.712	4532	147354	8290	154425	3240
5	0.616	8046	103594	11272	75698	4611
6	0.627	3400	75525	4764	180088	783
7	0.334	2613	80645	3662	123856	367
8	0.294	1508	65288	2370	188200	209
9	0.387	2638	78422	3765	142849	616
10	0.527	2572	90116	3845	199175	686
11	1.305	3221	90953	4248	395096	2000
12	1.513	3642	100317	4760	406403	2263

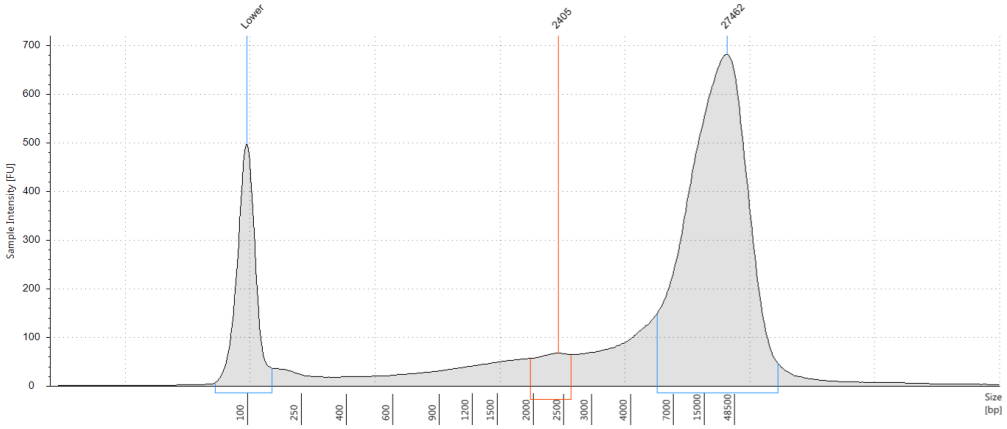


Figure 3.3.3: Agilent TapeStation output for station 5. Shows the molecular weight of the DNA against the sample intensity (FU), giving a visualisation of the molecular weight distribution of the DNA fragments in the sample. TapeStation outputs for all samples can be seen in Appendix A.

onboard the RRS Discovery, despite the use of a single flow cell for all samples. The N50 was increased compared to the ship-based sequencing, by a minimum of 2000 bp, and a maximum of 9000 bp.

The Station 5 sample was sequenced on its own using a whole flowcell in addition to sequencing as part of a multiplexed run. A comparison of sequencing metrics for this station can be found in table 3.5. The yield was 12.7 Gbp, compared to a

yield of 0.6 Gbp for the multiplexed sample. The N50 was lower in the in-depth sequencing experiment, at 7.6 kbp compared to 11.2 kbp for the multiplexed run; as such while the number of reads in total increased by 42-fold, the number of reads longer than 20 kbp only increased by around 12-fold.

Table 3.5: Yield, N50, total number of reads, and number of reads longer than 20kbp for in depth sequencing of Stations 5 in depth sequencing against the mulitplexed sequencing data for Station 5.

Sequencing	Yield (Gbp)	N50 (bp)	Reads	Reads > 20 kbp
In depth	12.7	7632	3216763	58206
Multiplexed	0.616	11272	75698	4611

Taxonomic classification of nanopore sequencing data from all 12 stations

Rarefaction curves were produced for all of the 12 stations as can be seen in figure 3.3.4. None of the curves are plateauing, which means that the full species diversity has not been captured by the available sequencing data. Stations 11 and 12, which had the highest yield, also have the highest identified number of species.

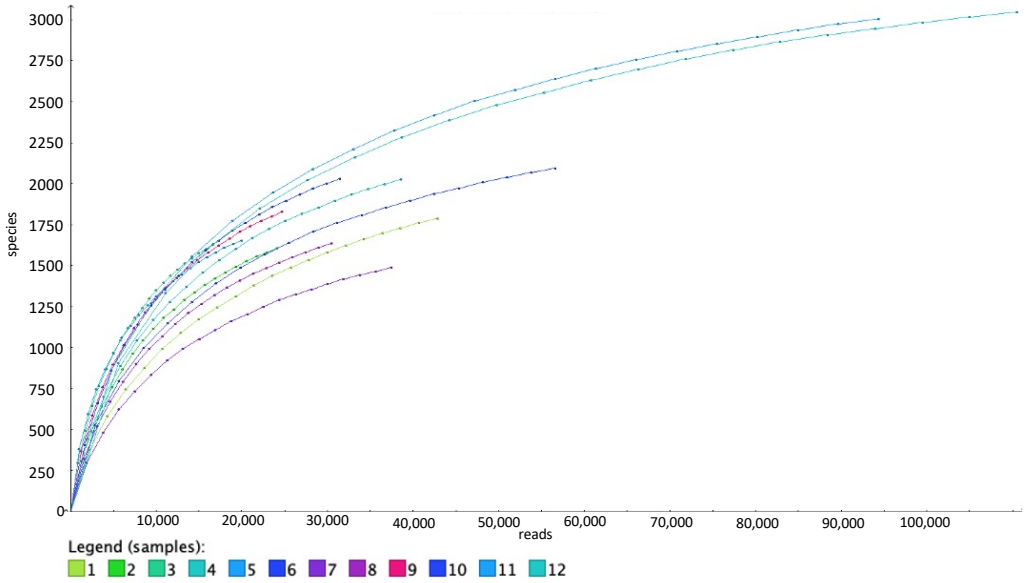


Figure 3.3.4: Rarefaction curves for each station sampled. Plots represent the number of species identified against the number of reads analysed for each station. Produced using MEGAN.

Taxonomic identification at superkingdom level was produced for all 12 stations, with the number of classified reads for each superkingdom shown in figure 3.3.5. From this it can be seen that the majority of reads for each sample are unclassified,

with for example around 150,000 classified reads from station 12 which had 400,000 sequenced reads. The overwhelming majority of the classifications in all samples is for bacteria, with smaller number, in the hundreds to low thousands, of eukaryotic classifications for each sample and a small number of archaea in 3 samples.

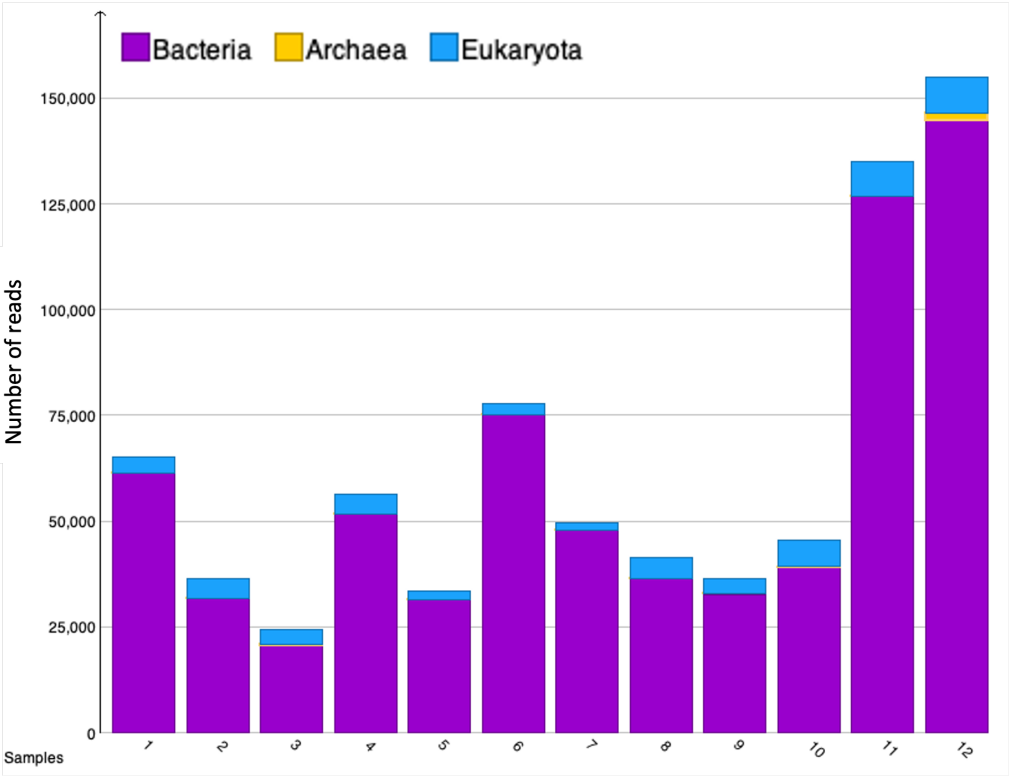


Figure 3.3.5: Number of classified reads for each sample across the bacteria, eukaryote, and archaea superkingdoms. Produced using MEGAN.

Taxonomic identification at genus level for all of the stations is shown in figure 3.3.6. As with ship-based sequencing, the most commonly identified at genus level was '*Candidatus* Pelagibacter'. Most samples were dominated by '*Candidatus* Pelagibacter', *Polaribacter*, and other polar marine bacteria. Figure 3.1.3 shows the location of each sample station. Stations 1 and 5 showed a lower proportion of '*Candidatus* Pelagibacter'. Formosa, a marine bacteria, was only found in significant numbers at station 1. *C. Pseudothioglobus* was most prevalent at stations 4 and 6 which are both close to shore. Eukaryotic phytoplankton including haptophytes and diatoms were found at all of the stations, these were investigated further, as covered below.

In depth sequencing of station 5 was used for taxonomic analysis to examine whether the taxonomic identification was limited by the amount of DNA sequenced. Figure 3.3.7 shows that there is little difference in taxonomic identification at genus

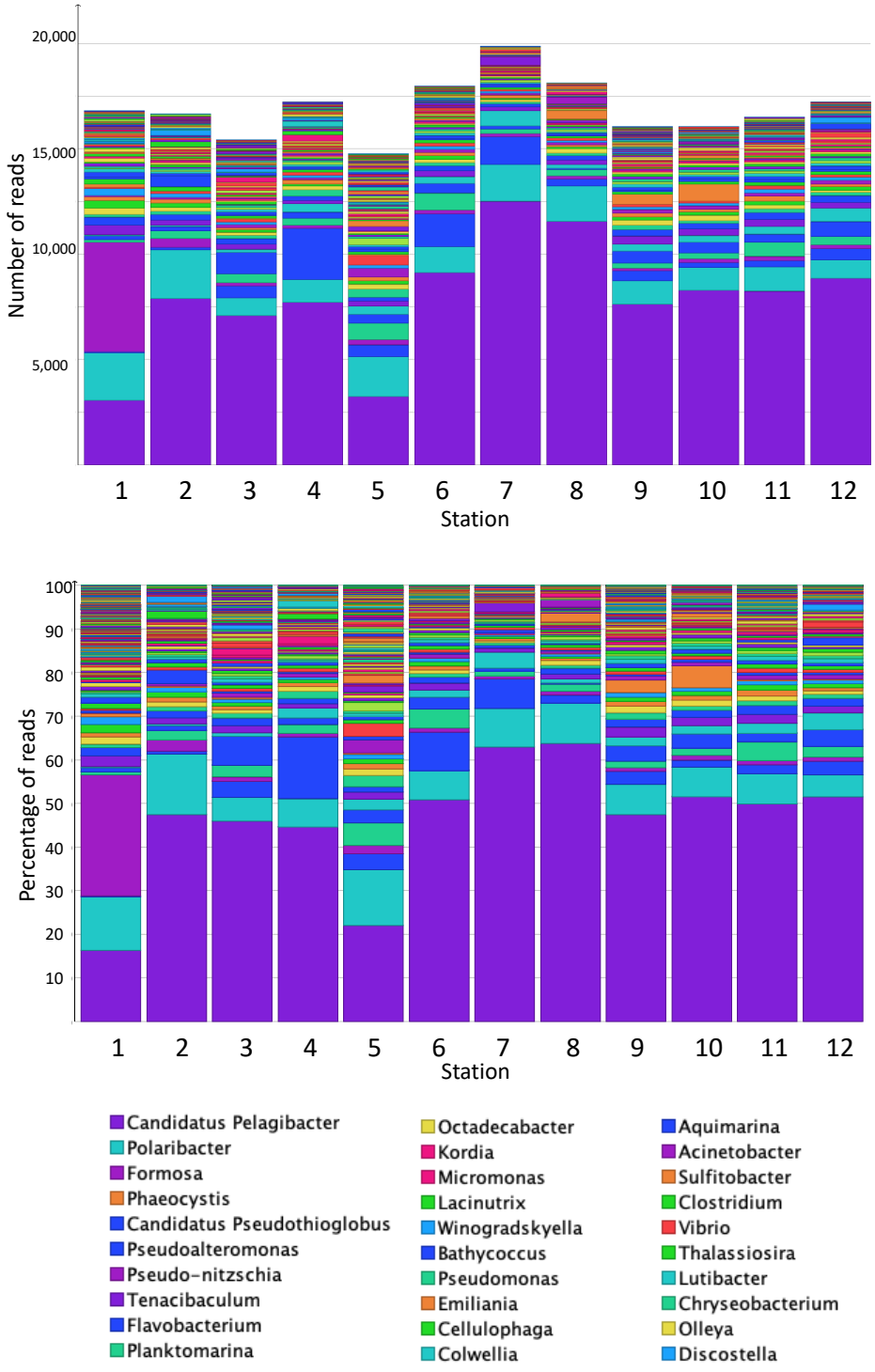


Figure 3.3.6: Stacked barcharts showing the genus level taxonomic identification for Stations 1-12 as absolute numbers of matches per sample and as percentage of the total reads. Legend identifies the top 30 genera by colour. Produced using MEGAN.

level between the in depth sequencing data and the multiplexed sequencing data for the most abundant genera, with the two samples diverging at lower levels of abundance. The order of genera is identical between the two samples for the top 70-80% of matches. A rarefaction curve of the two samples, see figure 3.3.8, shows that there is a far higher number of species identified in the in depth sequencing. A similar outcome was found when a genus-level rarefaction curve was produced, with more than 2500 genera identified in the in depth sample, compared to just over 500 in the multiplexed sample, although the abundance of the genera above 500 was very low, which would give low confidence in results based on taxonomic identification.

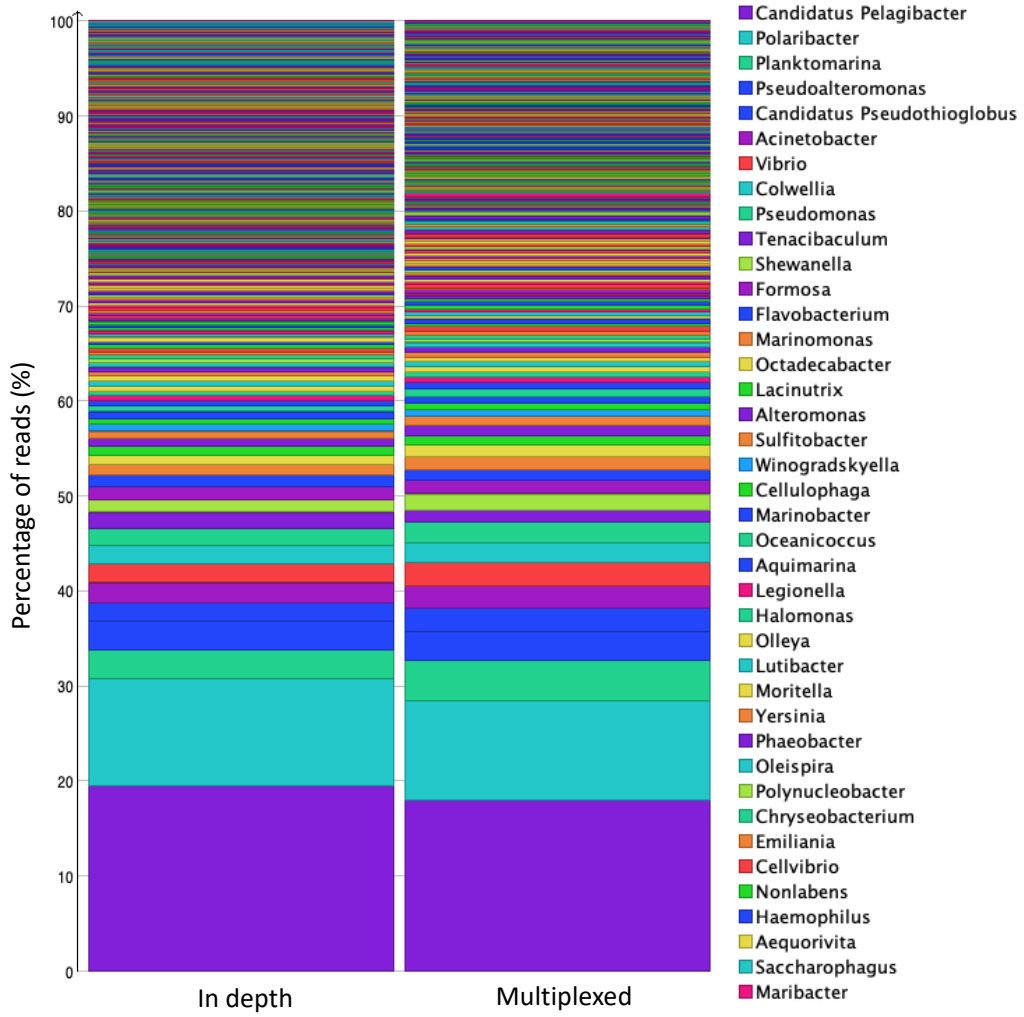


Figure 3.3.7: Stacked bar chart showing the percentage of matches at genus level for the in depth and multiplexed sequencing data for station 5. The legend shows the top 40 genera matches by colour. Produced using MEGAN.

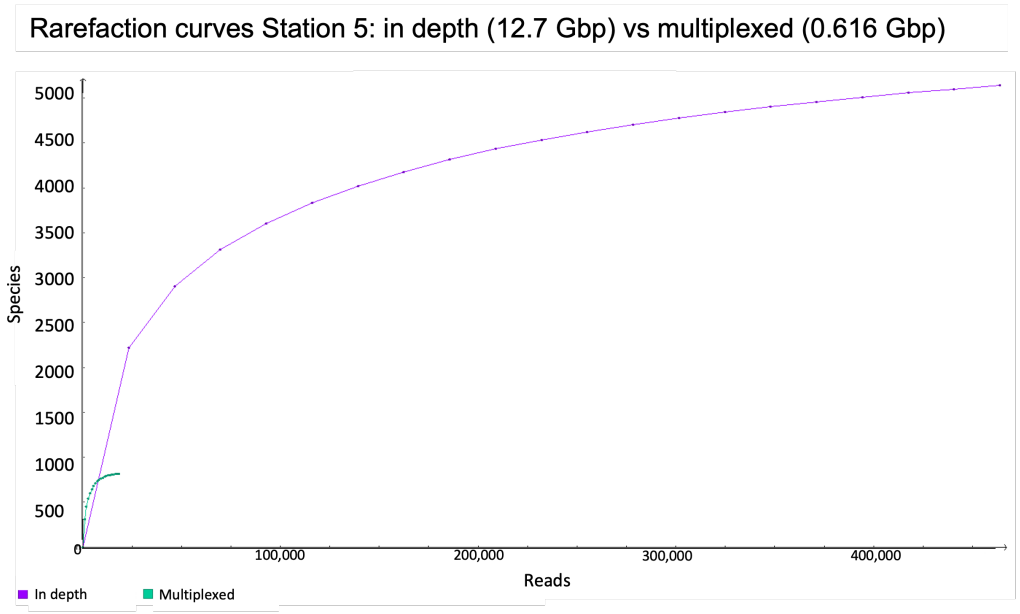


Figure 3.3.8: Rarefaction curves showing the number of species identified against the number of matches for the in depth and multiplexed sequencing data from station 5. Produced using MEGAN.

Targeted analysis of diatom and coccolithophore presence

The prevalence of diatoms and coccolithophores was examined more closely, as they were of particular interest in this project. Figure 3.3.9 shows the number of matches per sample to diatoms and coccolithophores which were above the 0.1% threshold. The overall number of reads matching to each diatom is no more than 500 reads per sample, with most diatoms in the low hundreds or tens of reads in each sample. This is largely true of *Emiliania* as well, with the exceptions of stations 9 and 10. The most prevalent of this subgroup is *Emiliania*, which is a temperate coccolithophore. *Emiliania* was found at every station, occurring most abundantly at stations further away from shore, with the highest abundances seen at stations 9 and 10 which are in the open ocean, above 55.5 °N, with a depth of around 5m. *Thalassiosira*, *Discostella*, *Cyclotella*, and *Skeletonema* were found mainly at stations 1-4 which are all in the North Eastern part of the sample area near South Georgia, with low numbers found at the other stations. *Pseudo-nitzschia* and *Kordia* were found at all stations, with the highest abundance at station 8, the southernmost sampling point. Fragilariopsis was mainly identified at stations 6-12, in the western half of the sample area near the South Sandwich Islands, with the lowest abundance at stations 9 and 10 which have large numbers of *Emiliania* matches. *Phaeodactylum* and *Biddulphophycidae* were found in low numbers at all stations.

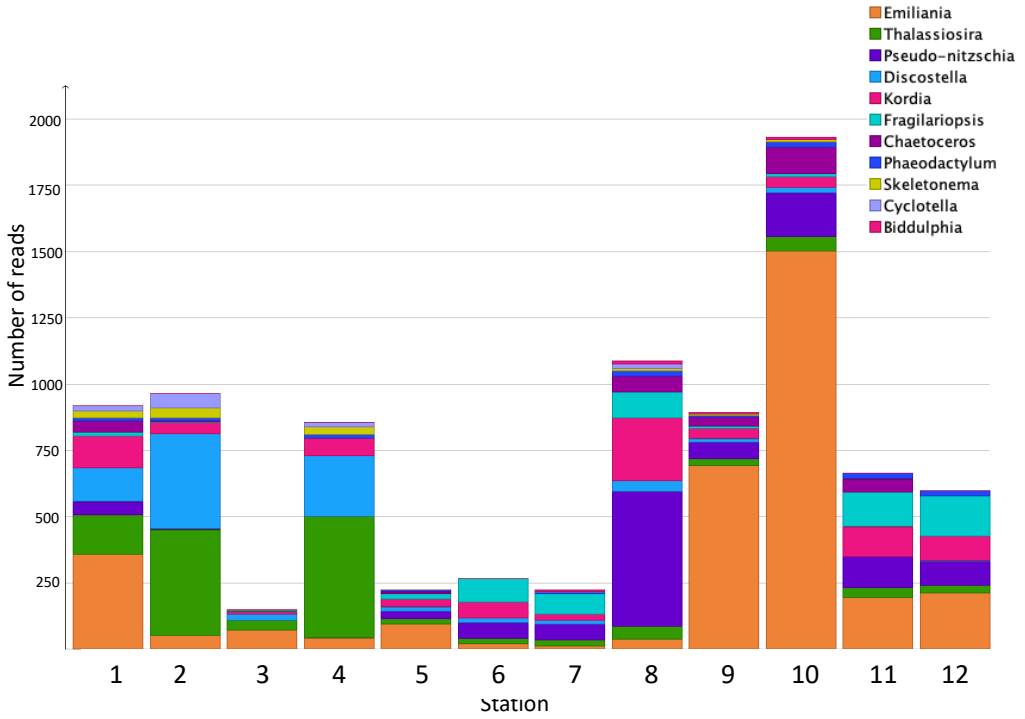


Figure 3.3.9: Stacked bar chart showing number of matches to diatom and coccolithophore genera per sample, legend identifies genus by colour. Produced using MEGAN.

Taxonomic analysis of Illumina sequencing data

MEGAN was used to compare taxonomic classifications between the Illumina sequencing dataset and the nanopore sequencing dataset across a similar number of reads. As can be seen in figure 3.3.10, analysis of data from the two platforms results in very similar taxonomic identification at genus level. There are some differences, however, for example the Illumina data shows a higher proportion of *Emiliania* and *Micromonas* matches. The order of matches is identical between the two up to around 90% of the reads, where divergence due to small numbers of hits in each can be seen. Rarefaction curves, see figure 3.3.11, showed that in each sample 10-100x more genera were identified in the nanopore samples than the Illumina, indicating that low saturation has little effect on the taxonomic identification of the most abundant genera. As they represent very small numbers of hits divergences in genera with lower abundancies may be due to genuine differences between the partial samples, as opposed to differences between the two sequencing platforms.

A comparison of ship-based nanopore sequencing data, laboratory-based nanopore sequencing, and Illumina sequencing data of samples from the three stations which were sequenced onboard the ship was produced. As can be seen

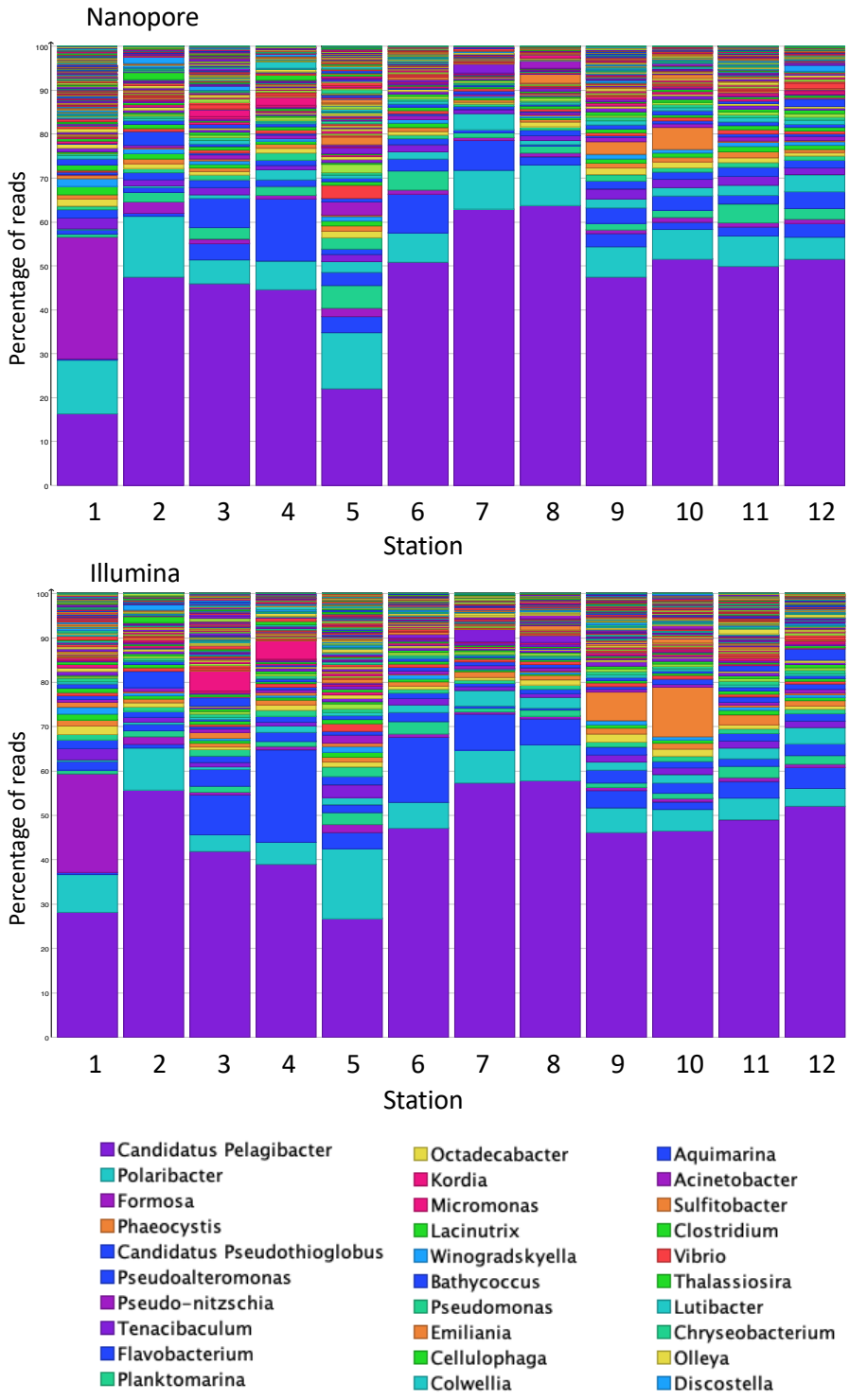


Figure 3.3.10: Stacked bar charts showing the percentage of each sample matching to each genus for Nanopore and Illumina sequencing data. The legend shows the top 30 genera matched.Produced using MEGAN.

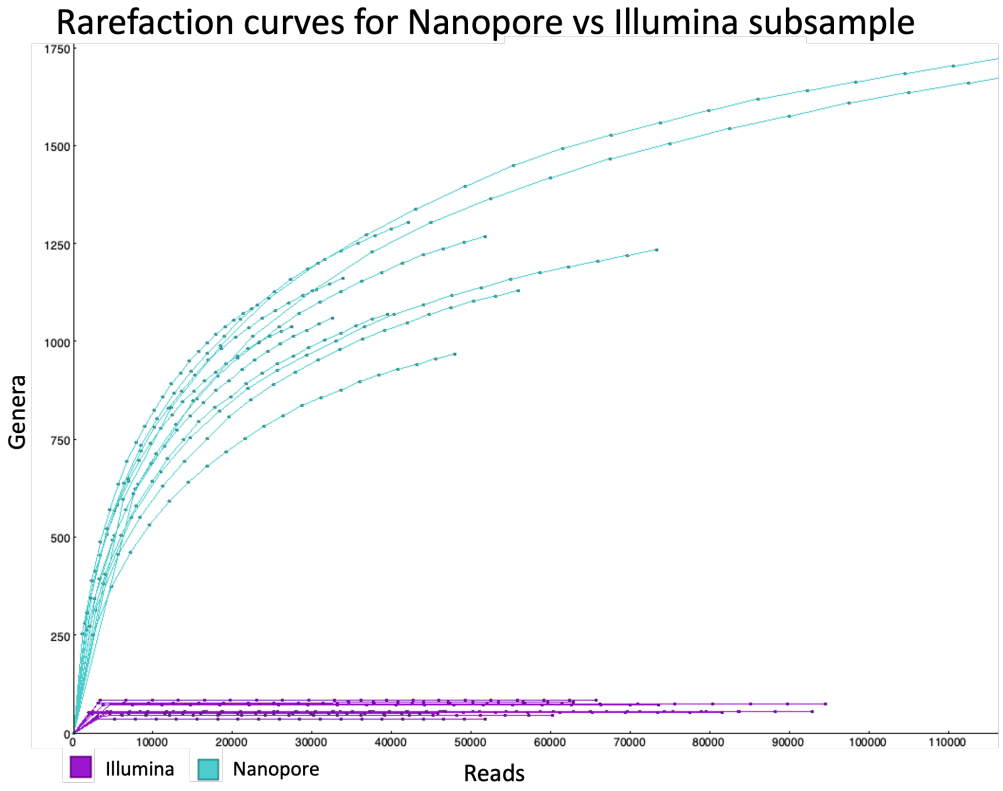


Figure 3.3.11: Rarefaction curves showing the number of genera identified against the number of matches for the 12 samples sequenced using nanopore and Illumina. Nanopore sequencing samples shown in blue, and Illumina samples shown in purple.

in figure 3.3.12, the different sequencing methods result in broadly similar results at each station with both nanopore experiments showing very similar results, in the same order with small variations in proportion, particularly in genera with low abundance. As seen in figure 3.3.10, the Illumina results are broadly similar to the nanopore sequencing data at high abundances with increasing divergence at low abundance. There is greater variation between Nanopore and Illumina than between Nanopore samples of different DNA quantities.

Assemblies

De novo metagenomic assemblies were produced from the in depth nanopore sequencing data for Station 5. Table 3.6 compares the processed reads to the assembled contigs. The mean read length and N50 increased by around ten fold in the assembled contigs compared to the processed reads, and the number of sequences reduced from over 3 million to just under 8000. The longest contig is over 3 million basepairs long after assembly, compared to a longest read size of

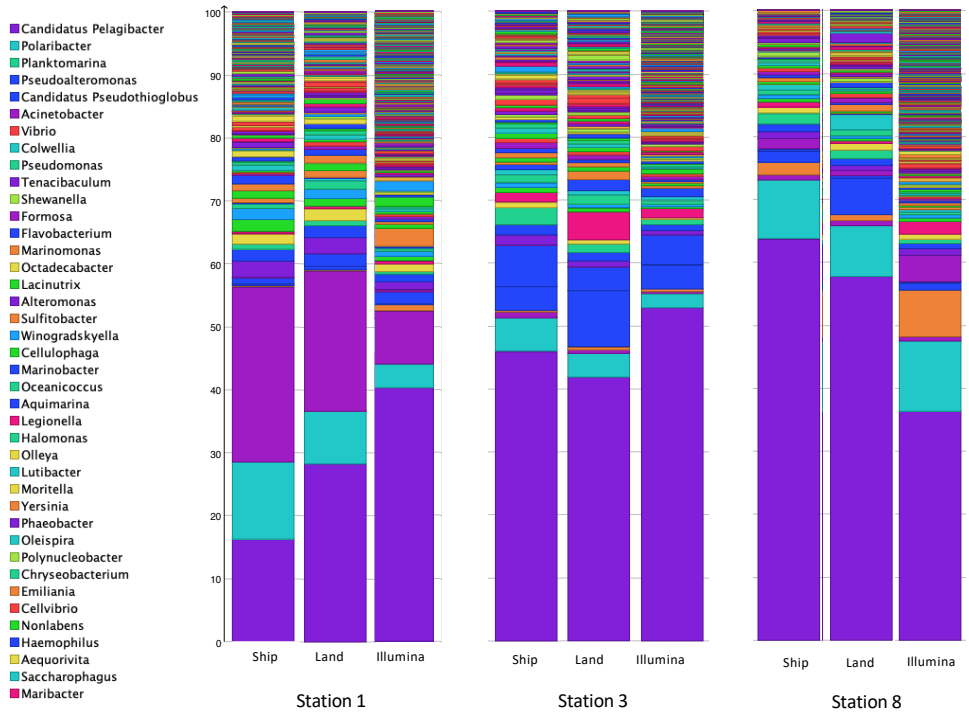


Figure 3.3.12: Stacked bar charts comparing the percentage of reads in each sample matching to each genus for the ship-based and land-based nanopore data and Illumina sequencing data produced from Stations 1, 3, and 8. Produced using MEGAN.

120,000 before assembly and just under half of the assembled contigs are longer than 20,000 bp.

BLASTing contigs against NCBI nt resulted in matches with a total length of 3,129,369 bp, with a mean of 7346 bp, to 100 different organisms. 39 of these were uncultured bacteria, viruses, eukaryotes, or diatoms. Excluding unknowns, the organisms with the longest matches and the highest total number of bases matched were *Psuedalteromonas arctica*, and '*Candidatus* Pelagibacter ubique'. No complete metagenomic assembled genomes (MAGs) were produced. The longest assembled reads did not necessarily yield good matches in the BLAST-nt database - for example the best match for the longest read, at over 3 million bp, resulted in a 2000 bp match to *Thalassiosira minima* for which there is no full genome assembly available.

Reads aligning to *P. arctica* and '*Candidatus* Pelagibacter ubique' were used for single species assembly. These were assessed using the length of the assembly compared to the reference, and the number and length of contigs. The alignment of the assembly against the published reference genome was visualised using

Table 3.6: Comparison of processed reads to the metagenomic assembly produced from in depth Nanopore sequencing of the Station 5 sample.

	Processed reads	Assembled reads
Number of sequences (reads/contigs)	3216763	7818
Mean length	3784.62	34757.89
Shortest	1	46
Longest	121004	3118335
N50 length	7632	78456
Number of reads >= N50 length	445146	694
N90 length	1741	18082
Number of reads >= N90 length	1742060	3526
Number of reads > 20000 bp	58206	3312

Alvis. Tables 3.7 and 3.8 show the comparison between the published reference genome and the genome produced from in-depth sequencing data, and figure 3.3.13 shows a visualisation of the alignment of the '*Candidatus* Pelagibacter' MAG against the reference genome. The *P. arctica* MAG was too incomplete for visualisation to be worthwhile. From the Tables, it can be seen that neither MAG is complete. The '*Candidatus* Pelagibacter' MAG is 78% of the size of the reference genome from a total unassembled read length of 405x genome size, and consists of 19 contigs compared to 1 single complete contig in the reference genome. This indicates that the genome is both incomplete and less well assembled than the reference genome. The *P. arctica* MAG is 10% of the size of the reference genome from a total unassembled read length of 4.5x total genome size and is too incomplete for the contig numbers to be of use in assessing assembly quality. Figure 3.3.13 shows that the MAG is incomplete, with gaps reasonably evenly distributed over the reference assembly without large gaps. This indicates that the reads have been correctly classified and with more data it may be possible to produce a complete MAG from the sample.

Assembly-free functional annotation

Assembly-free functional annotation was used to identify genes present in each sample. Pfam IDs were used to identify genes and their function. There were 9655 unique genes found from 69,286 in total, of which 9953 were of unkown function. The number of genes identified per sample varied from 4300 to 6600, with no clear correlation with the number of nanopore sequencing reads or

Table 3.7: Comparison of the published reference genome assembly for '*Candidatus* Pelagibacter' against the unassembled reads and the MAG produced from in depth sequencing of the Station 5 sample

'*Candidatus* Pelagibacter'	Published	Unassembled reads	Nanopore MAG
Number of contigs	1	168301	19
Length of assembly	1308759	531145966	1027872
Mean contig length	1308759	3156	54099
Shortest contig	1308759	30	15464
Longest contig	1308759	40740	174625
Length of contig at N50	1308759	5356	65285
Number of contigs at N50	1	29726	5

Table 3.8: Comparison of the published reference genome assembly for *P. arctica* against the unassembled reads and the MAG produced from in depth sequencing of the Station 5 sample.

P. arctica	Published	Unassembled reads	Nanopore MAG
Number of contigs	68	17960	18
Genome size	4628018	20646604	470513
Mean contig length	68059.09	1150	26139.61
Shortest contig	508	30	2492
Longest contig	508328	29472	60075
Length of contig at N50	116979	3183	31637
Number of contigs at N50	12	1792	6

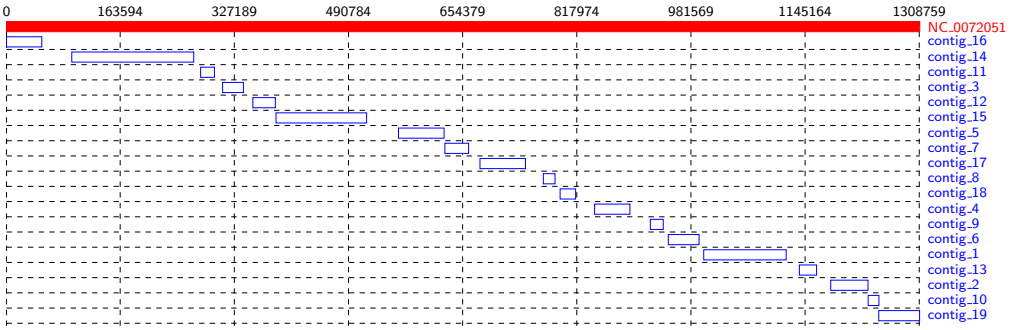


Figure 3.3.13: Visualisation of the alignment of the '*Candidatus* Pelagibacter ubique' MAG against the published reference genome assembly. The reference assembly is shown in red, and the MAG alignmnet shown in blue, with each contig shown on a separate line. This allows us to see how much of the genome is covered and how much is missing from the MAG. Produced using Alvis.

species number identified, but a clear correlation to total nanopore sequencing yield, see figure 3.3.14. There was no significant correlation found between the genes and any species or genus. The most abundant genes correspond to the following genes: ABC transporter, Elongation factor Tu GTP binding domain, HSP70 protein, 4Fe-4S binding domain, and Response regulator receiver domain. There was no clear difference in genes between sample stations. Genes related to cold-shock, iron-binding, silicon transport, and zinc-binding were identified but there was no statistically significant correlation between their abundance and any taxonomic classification or sampling station.

Identified genes were compared between reads which were identified as being bacterial and those which were identified as eukaryotic. As was shown in figure 3.3.5, of the approximately 2.5 million sequenced reads, 12865 were identified as bacterial and 3545 were identified as eukaryotic. From these, 4287 unique genes were found, of which 2071 were present in both bacterial and eukaryotic reads, 1850 were unique to bacterial reads, and 366 were unique to eukaryotic reads. The functions of the 10 most abundant genes in bacterial reads, and the top 10 bacterial only genes are shown in table 3.9, and the 10 most abundant Pfam IDs in eukaryotic reads, and the top 10 eukaryotic only Pfam IDs are shown in table 3.10. From these, it can be seen that the most abundant gene functions differ between bacterial and eukaryotic reads, and that these are different to the most abundant genes found only in bacteria or eukaryotes. Bacterial-only genes include bacterial polymerases and helicases, and polysaccharide biosynthesis, indicating that the genes correspond to the taxonomic identificaiton. Eukaryotic-only genes include those associated with photosystems, which indicates the presence of phytoplankton. Ubiquitin, actin, myosin, and histone genes were also identified; these are associated with eukaryotes which agrees again with the taxonomic identification.

Table 3.9: Count and description for the 10 most abundant genes identified in bacterial reads, and the top 10 bacterial only genes.

Top 10 of all bacterial genes		Top 10 bacterial-only Pfam IDs	
PfamID	Description	PfamID	Description
PF00005	ABC transporter	PF01255	Ptve. undecaprenyl diphosphate synthase
PF00133	tRNA synthetases class I	PF01293	Phosphoenolpyruvate carboxykinase
PF00113	Enolase; C-terminal TIM barrel domain	PF02719	Polysaccharide biosynthesis protein
PF00464	Serine hydroxymethyltransferase	PF02773	S-adenosylmethionine synthetase
PF00171	Aldehyde dehydrogenase family	PF01960	ArgJ family
PF06418	CTP synthase N-terminus	PF07733	Bacterial DNA polymerase III
PF02786	Carbamoyl-phosphate synthase L chain	PF08245	Mur ligase middle domain
PF00009	Elongation factor Tu GTP binding domain	PF05496	Holliday junction DNA helicase ruvB
PF00012	Hsp70 protein	PF06415	BPG-independent PGAM N-terminus
PF00709	Adenylosuccinate synthetase	PF00004	ATPase family; cellular acitivities

Table 3.10: Count and description for the 10 most abundant genes identified in eukaryotic reads, and the top 10 eukaryotic only genes.

Top 10 of all eukaryotic Pfam IDs		Top 10 eukaryotic-only Pfam IDs	
PfamID	Description	PfamID	Description
PF00115	Cytochrome C and Quinol oxidase	PF00240	Ubiquitin family
PF00012	Hsp70 protein	PF00223	Photosystem I psaA/psaB protein
PF00510	Cytochrome c oxidase subunit III	PF00125	Core histone H2A/H2B/H3/H4
PF00361	Proton-conducting membrane transporter	PF00022	Actin
PF00146	NADH dehydrogenase	PF00063	Myosin head (motor domain)
PF00240	Ubiquitin family	PF16211	C-terminus of histone H2A
PF00346	Respiratory-chain NADH dehydrogenase	PF12774	Hydrolytic ATP binding site of dynein motor
PF00223	Photosystem I psaA/psaB protein	PF12780	P-loop containing dynein motor region D4
PF00006	ATP synthase alpha/beta family	PF00493	MCM2/3/5 family
PF00116	Cytochrome C oxidase subunit II	PF12781	ATP-binding dynein motor region D5

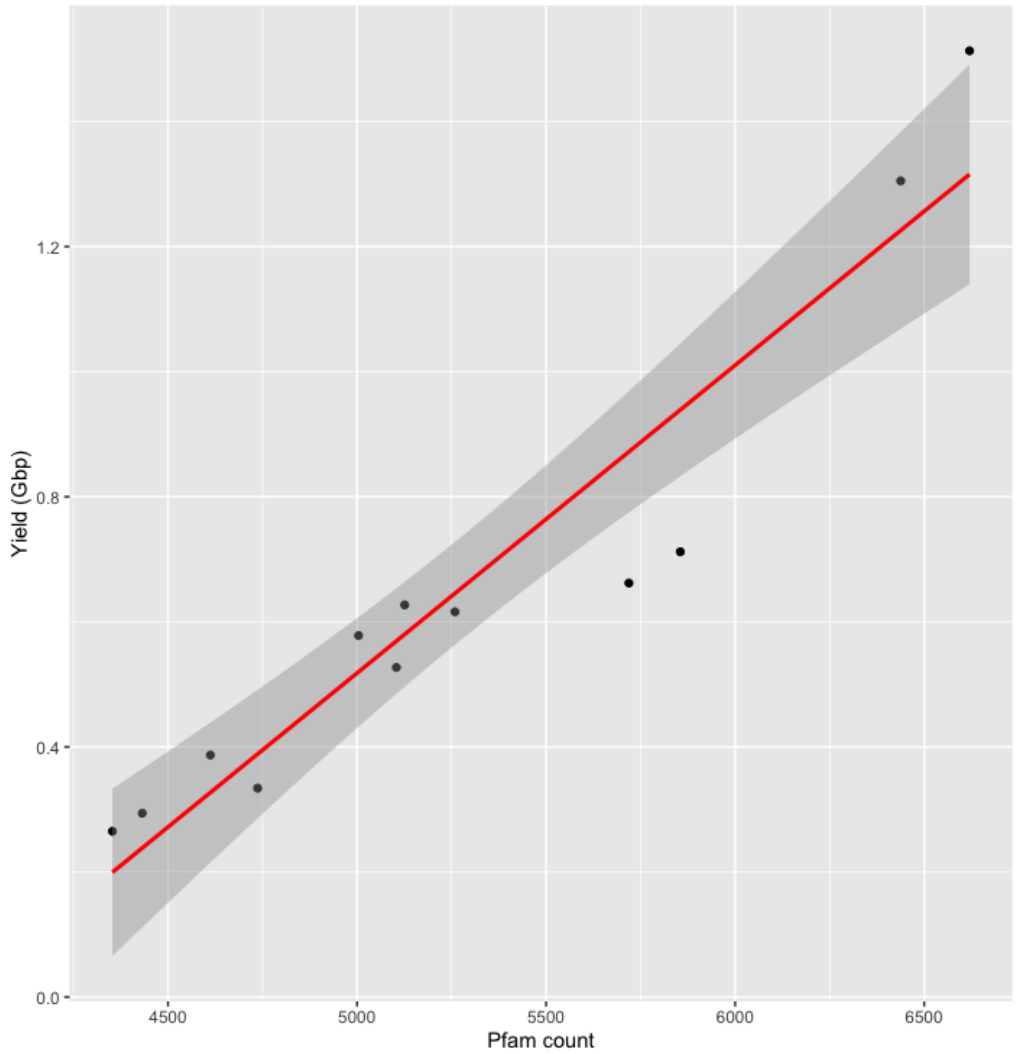


Figure 3.3.14: Scatter plot showing number of genes found per sample, against the total yield of the nanopore sequencing sample, with a linear model in red with 95% confidence intervals in grey showing a strongly positive relationship between number of genes and sample yield.

3.3.3 Case Study: Using environmental data to give context to metagenomic analyses of phytoplankton communities

Environmental and nutrient data from the DY098 cruise was used to investigate potential correlations between nanopore sequencing analysis such as taxonomic identification and assembly free functional annotation, and environmental variables. This was carried out as a proof-of-concept case study to establish analysis protocols prior to a planned second research cruise which was cancelled due to covid-19.

Metadata used: dissolved organic carbon - DOC (mg/l); depth of NISKIN bottle when fired - Depth (m); latitude and longitude of sample station - lat, long (decimal); ammonia - NH4, nitrate - NO2, nitrite - NO3, PO4, μM/L-1; silicic acid - Si (μM); Salinity (g/L); and Temperature (℃). Salinity, temperature, and depth were collected by CTD sensors, data provided by BAS Polar data Center, while particulate and dissolved organic matter and nutrient were collected and analysed by Flavia Saccomandi and Cecilia Silvestri of ISPRA (instituto Italiano per Le Risorse ambientali). See Appendix B for details on metadata collection, analysis and processed results.

A correlation plot was produced using MEGAN to assess the impact of each metadata variable on the presence of diatom and coccolithophore genera (figure 3.3.15). It can be seen from this that there is no clear distinction between clustered genera. Broadly, *Fragilariopsis, Emiliania, Chaetoceros, Biddulphia*, and *Pseudo-nitzschia* matches have a negative to neutral correlation with latitude, indicating decreasing abundance closer to the pole, while *Cyclotella, Kordia, Thalassiosira, Discostella, Phaeodactylum*, and *Skeletonema* have a positive to neutral correlation with increasing latitude. Fragilariopsis appears to be strongly positively correlated with increasing depth, and lower levels of NH4. Other strong correlations include a negative correlation between *Biddulphia* abundance and PO4 levels, and *Cyclotella* abundance and NO3 measurements. It should be noted, however, that the overall abundance of all diatoms but especially *Cyclotella* and *Biddulphia* are very low and as a result the significance of such associations may be overestimated.

The count of reads with a taxonomic match to diatom at each station was plotted against silica levels to assess the impact of silica on diatom abundance, and a linear model applied. As can be seen in figure 3.3.16, there is a weakly positive relationship between silica levels and number of diatom matches per sample, although this may be skewed by two outliers. Similarly, the relationship between diatom matches and temperature was modelled as can be seen in figure 3.3.17.

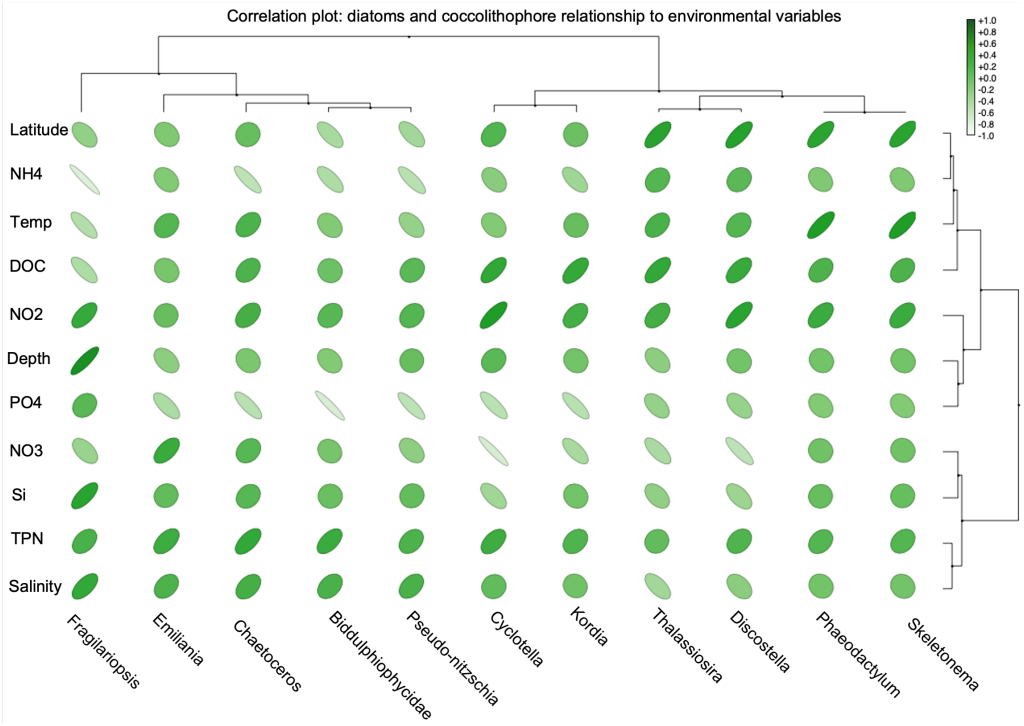


Figure 3.3.15: Correlation plot showing the relationship of each diatom and coccolithophore genera against each metadata variable. Circular points represent zero correlation, right-leaning, less circular shapes indicate positive correlation, while left-leaning, less circular shapes indicate negative correlation. As shown in the legend, correlation is also demonstrated by colour, with pale shades indicating negative correlation and darker shades indicating positive correlation. Clustering has been used to help to identify relationships between different variables.

This showed a weak negative correlation between number of diatoms matched and temperature of the sampling point.

The statistical significance of each metadata parameter was assessed using the vegan package in R, through nonmetric multidimensional scaling. Silica was the only statistically significant variable, with no significant effect found from latitude, longitude, depth, salinity, PO4, NO3, NO2, NH4, particulate organic carbon, total particulate nitrogen, dissolved organic carbon, and photosynthetic active radiation levels.

Correlation of genera present to metadata was largely unsuccessful. This is likely to be due to a small sample size, and potentially by a relative lack of variance in the conditions of the sampling locations. There were weak correlations between the overall number of reads matching diatoms and increasing silica and decreasing temperature levels. This is in line with diatoms reliance on silica for their frustules and their ability to thrive at low temperatures, and with previous findings in the Scotia Sea (Hinz et al. 2012) but they were not statistically significant correlations. Increased sample size and a wider sampling range may allow for statistical

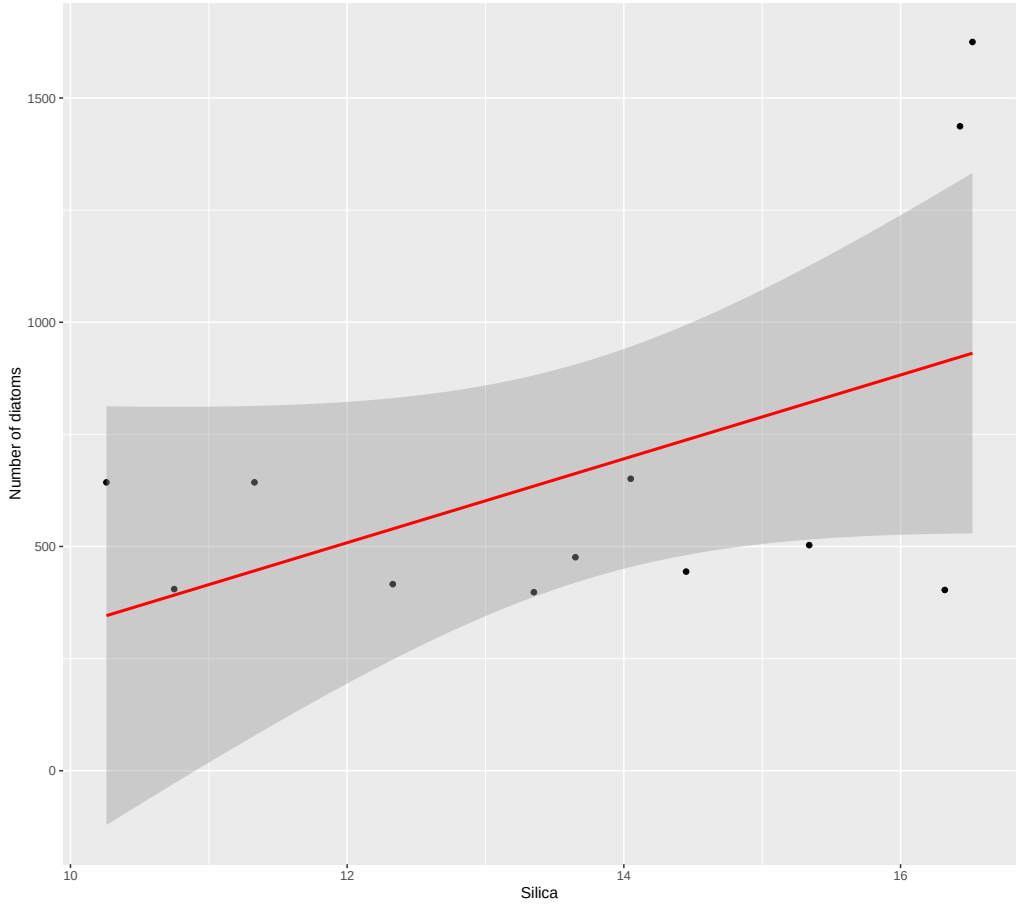


Figure 3.3.16: Scatter plot showing diatom number against silica levels, with a linear model in red with 95% confidence intervals in grey showing a weakly positive relationship between silica levels and number of diatom matches per sample.

analysis of environmental factors contributing to population distribution, perhaps contributing to a predictive process as described above to target organisms of interest such as diatoms.

3.4 Discussion

3.4.1 Nanopore sequencing on the RRS Discovery

Sampling

The 12 stations sampled were distributed along the DY098 research cruise transect. The route covered areas of open ocean and the shoreline of South Georgia and the South Sandwich Islands, and a range of depths. This was intended as a proof-of-concept study to be used to test the feasibility of nanopore

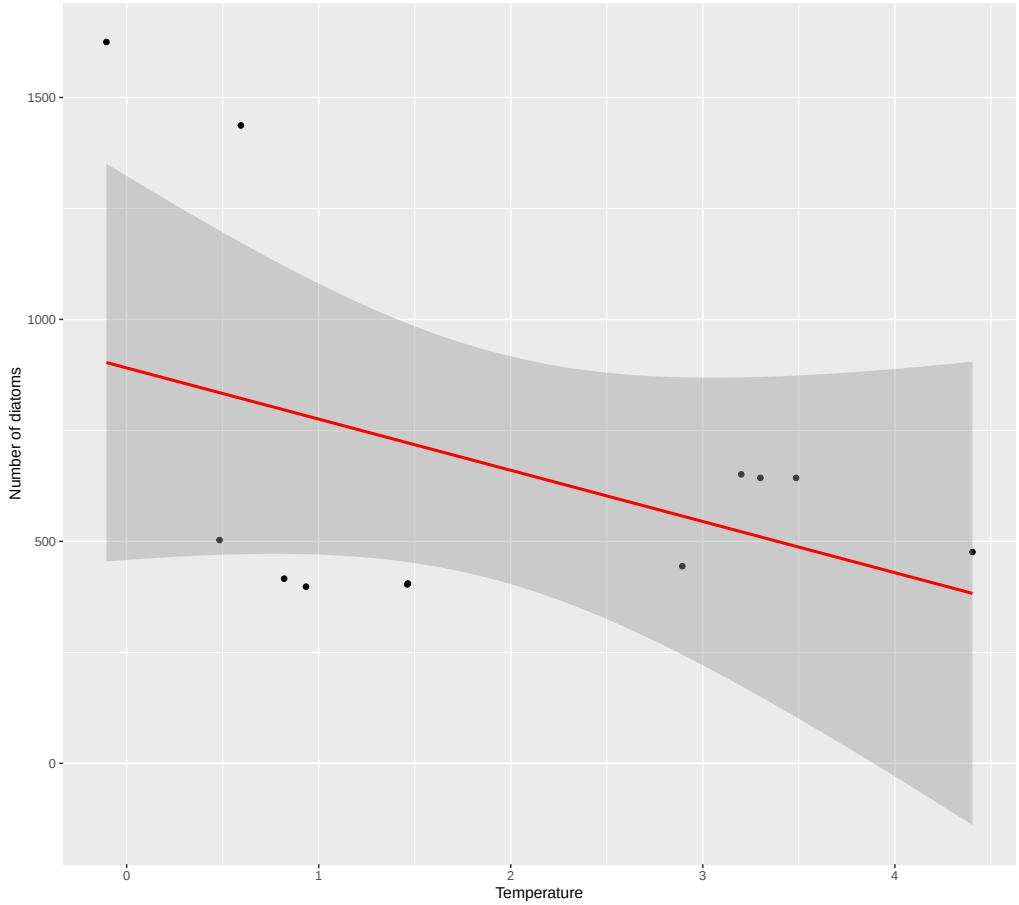


Figure 3.3.17: Scatter plot showing diatom number against temperature, with a linear model in red with 95% confidence intervals in grey showing a weakly negative relationship between temperature and number of diatom matches per sample.

MinION sequencing onboard a research vessel, and to understand the challenges and limitations before a second research cruise. As such the data analysed in this chapter was intended to be a test dataset for establishing analysis methods for a larger dataset from a more carefully selected range of sampling sites.

Due to restrictions imposed in response to the covid-19 pandemic, it was not possible to carry out this second cruise and so this proof-of-concept study was used for analysis. As such, the analysis has been carried out on a sub-optimal number of data points, and the distribution of sampling points is not necessarily ideal. Metadata was collected and used to analyse trends but there was insufficient data to draw any statistically significant conclusions. Similarly, assembly-free functional annotation was carried out as part of an attempt to investigate trends in certain genes present at different environmental conditions, or where specific organisms were present but this did not yield any significant results.

Future work would include a further research cruise, performing more onboard sequencing and collecting a larger number of samples for sequencing on land, ideally across a wider range of latitudes. More onboard sequencing would provide better comparisons between sampling stations and could allow for targeted sampling based on previous results, and a larger number of samples covering a wider latitudinal range would provide a rich dataset for in depth analysis.

Depending on the purpose of the research cruise, the sequencing data could be used to direct the cruise route. It is possible that by correlating microbial populations to zooplankton such as krill, that DNA sequencing onboard a research vessel could help to target sample sites for larger organisms. In depth sequencing after return to the lab would allow for metagenomic assemblies to be produced and genetic analyses to be carried out. This could give information on differences within species in different locations and provide insights into adaptations.

The Southern Ocean is a particularly important habitat for key zooplankton species such as krill and salps. As part of a second research cruise it would be interesting to investigate the potential to use phytoplankton presence as a proxy for zooplankton presence, for example by correlating taxonomic identification against krill catches, and also to investigate the microbiomes of the zooplankton themselves. In this instance it may be beneficial to quantify abundance based on sequencing data, perhaps by using mock cultures with known quantities as a quality control. Additionally, the microbiome of krill could be investigated, for example through sequencing krill or salps from samples which are regularly collected on research cruises in the Southern Ocean for population and diversity analyses and also through the sequencing of krill faecal pellets, which can be collected using sediment traps (Pauli et al. 2021).

DNA extraction and nanopore sequencing

In situ DNA extraction was successful, with relatively high yields of DNA, well above the 400 ng recommended minimum for the MinION flow cell. The DNA *in situ* sequencing using the MinION was also successful, with comparable taxonomic identification results to those achieved on land, and from Illumina sequencing of the same samples. This indicates that nanopore sequencing is an effective method for rapid profiling of marine microbe communities using relatively small quantities of sea water (< 10 litres per µg of DNA extracted on average). In future, this protocol could help researchers to make evidence based sampling decisions while in the field, and provide useful insights into population differences between sample sites. Based on this experiment, improvements to the protocol

will be made to streamline and simplify it so that it can be used more easily by researchers with less training, opening *in situ* sequencing up for wider use. These refinements form the basis for Chapter 4, which involved field testing an improved protocol.

The relatively low sequencing yield, considering that nanopore flow cells can now deliver over 10 Gbp of sequencing data, and short N50, combined with the relatively high error-rate of nanopore sequencing means that the data is not well suited for use in metagenome assemblies, evolutionary anlaysis and other in depth analyses. In order to make nanopore sequencing data more broadly useful, DNA length and sequencing yield would need to be improved. DNA length could potentially be improved through different DNA extraction processes. Sequencing yield improvements could come from improved DNA purity, better flowcell preservation under transport, increasing the number of flowcells used, or optimising loading the flowcells.

CTAB DNA extraction as used onboard the RRS Discovery is widely used for phytoplankton DNA extraction, due to its ability to recover high yields of DNA and remove polysaccharides (Rogers and Bendich 1994) but, as discussed in Chapter 2, CTAB extractions did not have a high purity for *E. huxleyi* extractions. The illustra Nucelon Phytopure kit has been successfully used for HMW DNA extraction from single species phytoplankton samples, see Chapter 2, and this is being tested as a potential improvement on the CTAB extraction for mixed samples. CTAB DNA extraction also uses extremely hazardous chemicals, including phenol and β-mercaptoethanol and is time and energy intensive, requiring a long incubation at 65 ℃. Hazardous chemicals present an obvious obstacle to use in the field, where they have to be transported, stored, used safely in non-standard conditions, and returned for disposal without release into the environment. The time and energy requirements could pose problems to researchers working in deep field conditions. Phytopure DNA extraction requires only 10 minutes incubation at high temperature, can be completed in around an hour, and requires fewer hazardous chemicals. If it is effective at HMW DNA extraction from marine metagenomic samples it would open the door to a wide range of field-based experiments and allow for more in depth research into phytoplankton genomics.

In situ metagenomic sequencing was successfully performed three times during the DY098 cruise. Despite relatively poor flowcell performance, possibly due to difficulties maintaining a cold chain during transport to the ship, the results were still sufficient for taxonomic identification. The real-time taxonomic analysis of *in situ* sequencing data was largely successful. After the removal of *M. sacchari*, the taxonomic identification at species and genera level were very similar to

taxonomic identification of Nanopore sequencing data produced on land, including the in depth sequencing data.

The nanopore flow cells and reagents used for sequencing during the DY098 cruise had to be kept refrigerated and frozen respectively. This posed problems due to the long travel time to the Falklands where DY098 started. Although the flowcells and reagents were reasonably effective after transport, the sub-optimal transport of these consumables likely resulted in reduced DNA sequencing yield, largely through deterioration of the biological pores in the flow cells. Three of the six flow cells had lost too many pores to be usable, while three were somewhat depleted but still usable. Since the DY098 cruise, Oxford Nanopore Technologies have released flow cells and reagents which can be stored at temperatures up to 30 ℃ for up to a month. These changes would make transport far easier, as they could potentially be transported in hand luggage and their use would likely increase the sequencing yield significantly compared to that achieved on the DY098 experiments. The newer MinION Mk1C includes a touch screen computer which can run the sequencing and basecalling internally, removing the requirement to carry multiple computers for sequencing and analysis. This will be of particular benefit for remote fieldwork.

Taxonomic analysis

Results were available within hours of sampling, giving valuable feedback on the success of sampling and DNA extraction. There were clear observable differences in population between each sequenced sample point, with increased numbers of polar-associated organisms, diatoms, and eukaryotic phytoplankton identified at station 8 which was the furthest South, and a more heavily bacterial make-up when sampling in silty shallow water at station 3, compared to the open ocean at station 1. *E. huxleyi* was identified at all stations, but at a much higher abundance in the northernmost station, as expected given that it is temperate. This is in line with previous findings in the Scotia Sea which found *E. huxleyi* towards the north of the area (Hinz et al. 2012). The high prevalence of prokaryotic sequences indicates that the use of a larger filter pore size may be beneficial for future experiments where eukaryotic species are of interest, as this would allow more of the smaller prokaryotic cells to be filtered out, increasing the proportion of larger eukaryotic cells. The smallest diatom cells are approximately 2 μm, so pore sizes between 0.45 and 2 μm could be trialled for future work. The addition of an automated adapter trimming step may be of benefit in future protocols, to prevent erroneous taxonomic identification. Additionally, use of a longer minimum match length would reduce the likelihood of incorrect assignments. This is particularly

important in metagenomic analysis as it is possible that an incorrect taxonomic match could seem plausible and not be recognised as an error.

This experiment was the first of its kind and clearly shows the potential of *in situ* sequencing with real-time analysis for use on research vessels. The protocol has been improved since 2019, it is now simplified, takes less time and requires less equipment and fewer reagents - see Chapter 4. Further work would involve a second cruise to test the improved protocol, collect a larger dataset, and potentially re-sample some of the same locations to give information on changing populations. There are many research cruises in polar oceans each year - if widely used, portable DNA sequencing with real-time analysis could provide an invaluable source of information on the polar ocean microbiome.

3.4.2 Nanopore sequencing of 12 Southern Ocean samples in the laboratory

DNA extraction and nanopore sequencing

The adapted CTAB DNA extraction method used to extract DNA from the 12 frozen samples produced DNA with a higher molecular weight than was achieved onboard the RRS Discovery, which resulted in a higher N50 and mean length compared to the ship-based samples. This indicates that the added bead beating stage did not degrade the DNA, and it is possible that the reduced incubation time was beneficial for DNA quality. The taxonomic identifications resulting from the two nanopore datasets were very similar, which may indicate that N50 is not the most important factor in taxonomic identification. Further work would include investigation of alternative DNA extraction methods, and an *in situ* experiment with more samples for comparison.

The use of multiplex nanopore sequencing of samples from all of the 12 stations allowed for cost-effective analysis of the samples collected. The flowcells used for laboratory-based nanopore sequencing were newer versions than those used onboard the RRS Discovery, with a higher yield, improved pores, and improved stability. They also had not been transported for fieldwork, under challenging conditions. The yield was around 8 Gbp of sequenced DNA, which is 4-5 times the yield achieved for Stations 3 and 8, and >10 times the yield achieved for Station 1, during ship-based sequencing.

Taxonomic analysis

The proportion of reads classified at superkingdom level was low, possibly due to sampling organisms which are not represented in the BLAST-nt database. The overwhelming majority of classifications for all samples were bacterial, which is unsurprising considering that, although eukaryotic phytoplankton make up the majority of the ocean microbiome biomass (Bar-On and Milo 2019), there is evidence that bacterial species richness is greater than that of eukaryotes in aquatic environments (Wan et al. 2023), and there is increased sequencing data available for bacterial species compared to eukaryotic species. Additionally, bacterial DNA is more easily extracted from samples compared to eukaryotic DNA. The majority of eukaryotic phytoplankton have not been sequenced, and so may be difficult to identify in sequencing studies. Increased genomic data for eukaryotic phytoplankton from projects such as Tara Oceans (De Vargas et al. 2015), and the production of MAGs could increase the proportion of reads which are classified.

The rarefaction curves produced for the 12 station samples show that none of the samples are sufficient to reflect species richness of the sample. The rarefaction curve for the in depth sequencing data from station 5 shows some flattening, indicating that it is reasonably reflective of species richness. It was established from comparing taxonomic identifications that approximately the top 80% of genus level matches were in identical order in the multiplexed sample compared with the in depth sample, with only a small variation in proportion as the abundances get lower. This indicates that for straight forward taxonomic identification of metagenomic nanopore sequencing data, multiplex sequencing to produce around 0.5-1.5 Gbp per sample is sufficient to capture an overview.

Station 1 was the most northerly sampling point and may have a higher proportion of non-polar adapted organisms present, while station 5 had relatively few shorter reads compared to other samples which could affect the proportion of bacteria to eukaryotes detected. This is because short reads are more likely to have random alignments to a sequence since they are only 100-300 bp long, such hits are often bacterial due to the dominance of bacterial sequences in sequencing databases. Short stretches of eukaryotic phytoplankton DNA might also be identified as bacterial based on their retention of bacterial DNA. Such erroneous alignments are far less likely with longer reads, as the probability of a random alignment decreases. Formosa was only found at Station 1, since it is a temperate bacteria, this may again be due to that sampling point being the most northerly

Targeted analysis of diatoms and coccolithophores was, on the whole, less successful. Low read numbers coupled with a small sample size means that

it is difficult to draw any firm conclusions regarding distribution of diatoms and coccolithophores. *Emiliania* was found at all stations, including the furthest south, while *Fragilariopsis* was found at the most northerly station which may indicate a wider distribution of both than was previously identified (Hinz et al. 2012). There were some potentially interesting insights, for example the correlation between groups of diatoms or between a diatom and another organism. It is possible that with more data, it would be possible to predict the presence of lower abundance organisms of interest based on the distribution of other more abundant organisms. This could potentially be used to target high yield sequencing efforts on the appropriate samples by using multiplex or flongle sequencing to give an overview of the species present and select samples for further investigation.

Illumina sequencing and taxonomic analysis

Illumina sequencing was used to benchmark nanopore sequencing outputs, as it is a more established technology with high accuracy. There were differences between taxonomic identifications for Illumina and nanpore sequencing data, but generally they were identical in order of genera with some variation in proportion up to around 80% of the sample. In the remaining 20%, where the number of reads matching each genus is very low, there was more variation both in genera found, and their proportions. Based on the vast difference in number of genera identified in the Illumina data compared to nanopore data, and the small variation between the taxonomic identification between the two, it appears that low saturation has little effect on taxonomic identification. At very low read numbers it is possible that the differences are genuinely based on presence in one sequencing sample but not another from the same source but given the very close agreement between the in depth sequencing data and the multiplexed sequencing, it is more likely to be that the use of different sequencing platforms has an effect on taxonomic identification. This may be due to increased accuracy in Illumina sequencing or the increased read length in nanopore sequencing, with the low number of genera identified in the Illumina data compared to nanopore indicating that nanopore sequencing is likely to be more representative at similar read numbers.

The question of whether Illumina or nanopore sequencing is likely to be the most accurate for taxonomic identification of metagenomes appears to depend on the taxa being investigated. One recent study which found that Illumina sequencing identification of a known bacteria was 96.7, compared to 90.3% for nanopore (Stefan et al. 2022), but a study invesitgating the identification of eukaryotes from metagenomic data found that long-read sequencing data, such as that

produced by nanopore, is more accurate (Pearman, Freed, and Silander 2020). The key interests of this project are eukaryotic phytoplankton, which indicates that nanopore sequencing is the best option and the taxonomic identifications based on the nanopore sequencing data are likely to be more accurate.

Assemblies

It was not possible to produce complete MAGs from the available nanopore sequencing data. The use of metagenomic assembly on the full in depth station 5 sample did reduce the number of reads and increase read length. While it may not be feasible to fully assemble genomes from relatively small quantities of nanopore sequencing data, it may be useful to work with longer reads, and better quality BLAST-nt hits. After assembly and filtering there were around 7800 BLAST-nt hits of longer than 1000 bp and greater than 90% identity matching to 61 known and 39 uncultured species. These improved reads could be useful for analysis. The single species assemblies produced were low coverage and incomplete when compared to the reference genome. The '*Candidatus* Pelagibacter' genome is one of the smallest ever discovered so it is unlikely that any other assemblies would be successful where this was not. Therefore, it is likely that significantly more than 12 Gbp yield is required for nanopore sequencing to produce a successful MAG from similar samples, possibly improved by higher molecular weight DNA or high-throughput long-read sequencing such as PacBio.

Work on eukaryotic MAGs from ocean metagenomic sequencing data has been based around Illumina short-read datasets of over 500 Gb from large sampling efforts (Delmont et al. 2022; Duncan et al. 2020). Future work could include high yield, high molecular weight DNA extraction and high volume sequencing. New MinION flowcells can produce up to 20 Gbp from 400 ng of DNA input, with researchers recently finding that even lower inputs can still provide good yields (Heavens et al. 2021), and the PromethION can produce over 200 Gbp per run. This could allow for the production of large datasets to produce MAGs from nanopore data, which would bring the benefits of long reads to improve accuracy, and improve assembly quality. Ship-based DNA extraction yields were comfortably above 400 ng, although it should be noted that that is generally in the best case scenario with clean, single species, amplified DNA libraries. Amplication-free metagenomic samples containing organisms such as diatoms and coccolithophores, from which it is not straightforward to extract high quality DNA, may be limited in yield per flow cell.

Assembly-free functional annotation

Assembly-free functional analysis was used to provide insights into the most prevalent genes identified in the nanopore sequencing data. Functional analysis is a key part of genomics studies, as it provides context beyond taxonomic identification, giving us information on organisms' cellular processes, evolutionary history, and ability to adapt to changing conditions, but it is also challenging, particularly for understudied organisms, and has advanced less rapidly than other aspects of genomics analysis (Crécy-Lagard et al. 2022). Around 15% of the identified genes were of an unknown function. This is in line with expectations given the paucity of ocean microbiome sequencing data (Abreu et al. 2022), but it reduces the utility of such analyses. Recently a large number of new genes have been identified, through analysis of data from the Tara Oceans project, more work such as this is required to harness the potential of functional analysis. Comparative analysis between bacterial and eukaryotic classified reads agreed with the taxonomic classification, finding bacterial genes in bacterial reads, and eukaryotic genes in eukaryotic reads. One of the most identified eukaryotic genes was associated with photosynthesis, which agrees with the known presence of phytoplankton. It was not possible to identify any genes which were correlated to with specific genera or species classifications, or with any of the sample locations.

This absence of correlation between genes and species could be an indication of functional redundancy, whereby genes are found to be present throughout the microbiome rather than associated with only certain species, meaning that multiple species undertake similar roles in the ecosystem, such as photosynthesis(Zhong et al. 2020). Functional redundancy has been found to increase the resilience of populations to changing conditions, such as climate change (Hoppe et al. 2017).

With larger, more accurate nanopore sequencing datasets it may be that gene identification would be improved. This could allow for greater insights, for example there were genes identified which are related to nutrient uptake and psychrotolerance, but it was not possible to establish whether these were found more at certain sampling points or in the presence of certain taxas from the available data. From large datasets it would likely be possible to produce MAGs which can then be functionally annotated as assemblies. This would improve the accuracy of functional annotation, and provide useful context for where genes came from, which could help to understand the means by which phytoplankton adapt to their environment.

3.4.3 Case study: Environmental data

Measurements of environmental variables were collected as part of experiments by other researchers, and from the CTD used for sampling during the DY098 cruise 3.3.3 and this was used to investigate whether it could provide context for the nanopore sequencing data outputs. This was not particularly successful, either with taxonomic classifications or Pfam gene identifications, although weak (statistically insignificant) correlations were found between diatom read numbers and lower temperatures and higher silica measurements. It is possible that under a wider range of conditions a stronger correlation would become clear. A small proportion of the sequenced reads were classified, and as such the read numbers corresponding to many species of interest were very low, and only a fraction of the identified genes were found in reads which had a taxonomic classification. We know that the interactions between phytoplankton and the surrounding ecosystem is highly complex, and there is a pressing need to improve our understanding in order develop more effective models and understand the effects changing conditions are likely to have. Future work on this topic would involve collecting samples over a wider geographic range, hopefully providing a larger number of classified reads given increased sequencing yields and accuracy, and recent additions to the available sequencing data for phytoplankton. With this, it should be possible to investigate more effectively the correlations between phytoplankton, their genes, and environmental variables.

3.4.4 Conclusions

In conclusion, the proof-of-concept study for real-time analysis of *in situ* DNA sequencing of polar ocean samples onboard a research vessel was successful. This was not something that had previously been achieved and the results showed that it could provide useful results. As well as *in situ* sequencing, more in depth analysis was performed which gave some insights into the contributing factors for phytoplankton distribution in polar oceans. This work gives an indication of what is possible with MinION sequencing, and highlights potential areas for improvement. Future work including a second research cruise to test an improved protocol and increase the sample size would allow for firmer conclusions to be drawn as to correlations between organisms of interest such as diatoms, and to understand the effects of environmental conditions on diatom populations.

References

Abreu, A., E. Bourgois, A. Gristwood, R. Troublé, D. Arendt, J. Bilic, R. Finn, E. Heard, B. Rouse, and J. Vamathevan (2022). "Priorities for ocean microbiome research". In: *Nature Microbiology* 7.7, pp. 937–947.

Andrews, S. (2010). *FASTQC. A quality control tool for high throughput sequence data*.

Armbrust, E. V. (May 2009). "The life of diatoms in the world's oceans". In: *Nature* 459.7244, pp. 185–192. ISSN: 0028-0836. DOI: 10.1038/nature08057.

Banse, K. (1996). "Low Seasonality of Low Concentrations of Surface Chlorophyll in the Subantarctic Water Ring: Underwater Irradiance, Iron, Or Grazing?" In: *Progress in Oceanography* 37.3-4, pp. 241–291. DOI: 10.1016/s0079-6611(96)00006-7.

Bar-On, Y. M. and R. Milo (2019). "The biomass composition of the oceans: a blueprint of our blue planet". In: *Cell* 179.7, pp. 1451–1454.

Bindoff, N. L., J. Willebrand, V. Artale, A. Cazenave, J. M. Gregory, S. Gulev, K. Hanawa, C. Le Quere, S. Levitus, Y. Nojiri, et al. (2007). "Observations: oceanic climate change and sea level". In.

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a Flexible Trimmer for Illumina Sequence Data". In: *Bioinformatics* 30.15, pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bork, P., C. Bowler, C. De Vargas, G. Gorsky, E. Karsenti, and P. Wincker (2015). "Tara Oceans studies plankton at Planetary scale". In: *Science* 348.6237, p. 873. ISSN: 10959203. DOI: 10.1126/science.aac5605.

Borrione, I. and R. Schlitzer (2013). "Distribution and recurrence of phytoplankton blooms around South Georgia, Southern Ocean". In: *Biogeosciences* 10.1, pp. 217–231. DOI: 10.5194/bg-10-217-2013.

Boyd, P. W. (Oct. 2002). "Review of environmental factors controlling phytoplankton processes in the Southern Ocean". In: *Journal of Phycology* 38.October 2001, pp. 844–861. ISSN: 0022-3646. DOI: 10.1046/j.1529-8817.2002.t01-1-01203.x.

Constable, A. J., J. Melbourne-Thomas, S. P. Corney, K. R. Arrigo, C. Barbraud, D. K. A. Barnes, N. L. Bindoff, P. W. Boyd, A. Brandt, D. P. Costa, A. T. Davidson, H. W. Ducklow, L. Emmerson, M. Fukuchi, J. Gutt, M. A. Hindell, E. E. Hofmann, G. W. Hosie, T. Iida, S. Jacob, N. M. Johnston, S. Kawaguchi, N. Kokubun, P. Koubbi, M.-A. Lea, A. Makhado, R. A. Massom, K. Meiners, M. P. Meredith, E. J. Murphy, S. Nicol, K. Reid, K. Richerson, M. J. Riddle, S. R. Rintoul, W. O. Smith, C. Southwell, J. S. Stark, M. Sumner, K. M. Swadling, K. T. Takahashi, P. N. Trathan, D. C. Welsford, H. Weimerskirch, K. J. Westwood, B. C. Wienecke, D. Wolf-Gladrow, S. W. Wright, J. C. Xavier, and P. Ziegler (2014). "Climate

Change and Southern Ocean Ecosystems I: How Changes in Physical Habitats Directly Affect Marine Biota". In: *Global Change Biology* 20.10, pp. 3004–3025. DOI: 10.1111/gcb.12623.

Cota, G. F. (May 1985). "Photoadaptation of high Arctic ice algae". In: *Nature* 315.6016, pp. 556–557. ISSN: 0028-0836. DOI: 10.1038/316507a0.

Crécy-Lagard, V. de, R. Amorin de Hegedus, C. Arighi, J. Babor, A. Bateman, I. Blaby, C. Blaby-Haas, A. J. Bridge, S. K. Burley, S. Cleveland, et al. (2022). "A roadmap for the functional annotation of protein families: a community perspective". In: *Database* 2022.

De Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukes, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, and E. Karsenti (2015). "Eukaryotic plankton diversity in the sunlit ocean". In: *Science* 348.6237, pp. 1261605–1/11. ISSN: 0717-6163. DOI: 10.1007/s13398-014-0173-7.2. arXiv: 9809069v1 [arXiv:gr-qc].

Delmont, T. O., M. Gaia, D. D. Hinsinger, P. Frémont, C. Vanni, A. Fernandez-Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. D. Silva, M. Wessner, B. Noel, J.-M. Aury, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, S. Sunagawa, S. G. Acinas, P. Bork, E. Karsenti, C. Bowler, C. Sardet, L. Stemmann, C. de Vargas, P. Wincker, M. Lescot, M. Babin, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, O. Jaillon, S. Kandels, D. Iudicone, H. Ogata, S. Pesant, M. B. Sullivan, F. Not, K.-B. Lee, E. Boss, G. Cochrane, M. Follows, N. Poulton, J. Raes, M. Sieracki, and S. Speich (2022). "Functional Repertoire Convergence of Distantly Related Eukaryotic Plankton Lineages Abundant in the Sunlit Ocean". In: *Cell Genomics* 2.5, p. 100123. DOI: 10.1016/j.xgen.2022.100123.

Duncan, A., K. Barry, C. Daum, E. Eloe-Fadrosh, S. Roux, S. G. Tringe, K. Schmidt, K. U. Valentin, N. Varghese, I. V. Grigoriev, R. Leggett, V. Moulton, and T. Mock (2020). *Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle*. DOI: 10.1101/2020.06.16.154583.

Fernández-González, C., G. A. Tarran, N. Schuback, E. M. S. Woodward, J. Arístegui, and E. Marañón (2022). "Phytoplankton responses to changing temperature and nutrient availability are consistent across the tropical and subtropical Atlantic". In: *Communications Biology* 5.1, p. 1035.

Fiala, M. and L. Oriol (Oct. 1990). “Ligbt-Temperature Interactions on tbe Growth of Antarctic Diatoms”. In: *Polar Biology* 10.8, pp. 629–636. ISSN: 0722-4060. DOI: 10.1007/BF00239374.

Grant, S., A. Constable, B. Raymond, and S. Doust (2006). “Bioregionalisation of the Southern Ocean: report of experts workshop, Hobart, September 2006”. In: *WWF-Australia and ACE CRC*.

Heavens, D., D. Chooneea, M. Giolai, P. Cuber, P. Aanstad, S. Martin, M. Alston, R. Misra, M. D. Clark, and R. M. Leggett (Oct. 2021). “How low can you go? Driving down the DNA input requirements for nanopore sequencing”. In: DOI: 10.1101/2021.10.15.464554.

Hinz, D., A. Poulton, M. Nielsdóttir, S. Steigenberger, R. Korb, E. Achterberg, and T. Bibby (2012). “Comparative Seasonal Biogeography of Mineralising Nannoplankton in the Scotia Sea: Emiliania Huxleyi, Fragilariopsis Spp. and Tetraparma Pelagica”. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 59-60.nil, pp. 57–66. DOI: 10.1016/j.dsr2.2011.09.002.

Hoegh-Guldberg, O. and J. F. Bruno (June 2010). “The Impact of Climate Change on the World’s Marine Ecosystems”. In: *Science* 328.5985, pp. 1523–1528. DOI: 10.1126/science.1189930.

Hoppe, C. J., N. Schuback, D. M. Semeniuk, M. T. Maldonado, and B. Rost (2017). “Functional redundancy facilitates resilience of subarctic phytoplankton assemblages toward ocean acidification and high irradiance”. In: *Frontiers in Marine Science* 4, p. 229.

Hunt, B. P. and G. W. Hosie (2005). “Zonal Structure of Zooplankton Communities in the Southern Ocean South of Australia: Results From a 2150km Continuous Plankton Recorder Transect”. In: *Deep Sea Research Part I: Oceanographic Research Papers* 52.7, pp. 1241–1271. DOI: 10.1016/j.dsr.2004.11.019.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster (Feb. 2007). “MEGAN analysis of metagenomic data”. In: *Genome Research* 17.3, pp. 377–386. ISSN: 1088-9051. DOI: 10.1101/gr.5969107.

Jin, P. and S. Agustí (2018). “Fast adaptation of tropical diatoms to increased warming with trade-offs”. In: *Scientific reports* 8.1, p. 17771.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter (2014). “Interproscan 5: Genome-Scale Protein Function Classification”. In: *Bioinformatics* 30.9, pp. 1236–1240. DOI: 10.1093/bioinformatics/btu031.

Katz, M. E., J. D. Wright, K. G. Miller, B. S. Cramer, K. Fennel, and P. G. Falkowski (June 2005). “Biological overprint of the geological carbon cycle”. In: *Marine Geology* 217.3-4, pp. 323–338. DOI: 10.1016/j.margeo.2004.08.005.

Kolmogorov, M., D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. L. Smith, and P. A. Pevzner (2020).

"Metaflye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs". In: *Nature Methods* 17.11, pp. 1103–1110. DOI: 10.1038/s41592-020-00971-x.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner (2019). "Assembly of Long, Error-Prone Reads Using Repeat Graphs". In: *Nature Biotechnology* 37.5, pp. 540–546. DOI: 10.1038/s41587-019-0072-8.

Laufkötter, C., M. Vogt, N. Gruber, M. Aita-Noguchi, O. Aumont, L. Bopp, E. Buitenhuis, S. C. Doney, J. Dunne, T. Hashioka, J. Hauck, T. Hirata, J. John, C. L. Quéré, I. D. Lima, H. Nakano, R. Seferian, I. Totterdell, M. Vichi, and C. Völker (Dec. 2015). "Drivers and uncertainties of future global marine primary production in marine ecosystem models". In: *Biogeosciences* 12.23, pp. 6955–6984. DOI: 10.5194/bg-12-6955-2015.

Leggett, R. M., D. Heavens, M. Caccamo, M. D. Clark, and R. P. Davey (Sept. 2015). "NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles". In: *Bioinformatics* 32.1, btv540. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv540.

Li, H. (2018). "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* 34.18, pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

Magoc, T. and S. L. Salzberg (2011). "Flash: Fast Length Adjustment of Short Reads To Improve Genome Assemblies". In: *Bioinformatics* 27.21, pp. 2957–2963. DOI: 10.1093/bioinformatics/btr507.

Martin, S. and R. M. Leggett (2021). "Alvis: a Tool for Contig and Read Alignment Visualisation and Chimera Detection". In: *BMC Bioinformatics* 22.1, p. 124. DOI: 10.1186/s12859-021-04056-0.

Mills, E. L. (2005). "From Discovery to discovery: the hydrology of the Southern Ocean, 1885–1937". In: *Archives of natural history* 32.2, pp. 246–264.

Mock, T. and A. Kirkham (2012). "What can we learn from genomics approaches in marine ecology? From sequences to eco-systems biology!" In: *Marine Ecology* 33.2, pp. 131–148. ISSN: 01739565. DOI: 10.1111/j.1439-0485.2011.00479.x.

Mock, T. and B. M. Kroon (Sept. 2002). "Photosynthetic energy conversion under extreme conditions—I: important role of lipids as structural modulators and energy sink under N-limited growth in Antarctic sea ice diatoms". In: *Phytochemistry* 61.1, pp. 41–51. ISSN: 0031-9422. DOI: 10.1016/S0031-9422(02)00216-9.

Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland, B. J. Ward, A. E. Allen, C. L. Dupont, S. Frickenhaus, F. Maumus, A. Veluchamy, T. Wu, K. W. Barry, A. Falciatore, M. I. Ferrante, A. E. Fortunato, G. Glöckner, A. Gruber, R. Hipkin, M. G. Janech, P. G. Kroth, F. Leese, E. A. Lindquist, B. R. Lyon, J. Martin, C. Mayer, M. Parker, H. Quesneville, J. A. Raymond, C. Uhlig, R. E. Valas, K. U. Valentin, A. Z.

Worden, E. V. Armbrust, M. D. Clark, C. Bowler, B. R. Green, V. Moulton, C. van Oosterhout, and I. V. Grigoriev (Jan. 2017). "Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus". In: *Nature* 541.7638, pp. 536–540. ISSN: 0028-0836. DOI: 10.1038/nature20803.

Mock, T. and K. Valentin (July 2004). "Photosynthesis and cold acclimation: molecular evidence from a polar diatom". In: *Journal of Phycology* 40.4, pp. 732–741. ISSN: 00223646. DOI: 10.1111/j.1529-8817.2004.03224.x.

Morgan-Kiss, R. M., A. G. Ivanov, T. Pocock, M. Krol, L. Gudynaite-Savitch, and N. P. A. Huner (Aug. 2005). "The Antarctic Psychrophile, Chlamydomonas Raudensis Ettl (Uwo241) (Chlorophyceae, Chlorophyta), Exhibits a Limited Capacity To Photoacclimate To Red Light1". In: *Journal of Phycology* 41.4, pp. 791–800. ISSN: 0022-3646. DOI: 10.1111/j.1529-8817.2005.04174.x.

Nurk, S., D. Meleshko, A. Korobeynikov, and P. A. Pevzner (2017). "Metaspades: a New Versatile Metagenomic Assembler". In: *Genome Research* 27.5, pp. 824–834. DOI: 10.1101/gr.213959.116.

Pauli, N.-C., C. M. Flintrop, C. Konrad, E. A. Pakhomov, S. Swoboda, F. Koch, X.-L. Wang, J.-C. Zhang, A. S. Brierley, M. Bernasconi, et al. (2021). "Krill and salp faecal pellets contribute equally to the carbon flux at the Antarctic Peninsula". In: *Nature Communications* 12.1, pp. 1–12.

Pearman, W. S., N. E. Freed, and O. K. Silander (2020). "Testing the advantages and disadvantages of short-and long-read eukaryotic metagenomics using simulated reads". In: *BMC bioinformatics* 21.1, pp. 1–15.

Rodrigues, D. F. and J. M. Tiedje (Jan. 2008). "Coping with Our Cold Planet". In: *Applied and Environmental Microbiology* 74.6, pp. 1677–1686. DOI: 10.1128/aem.02000-07.

Rogers, S. O. and A. J. Bendich (1994). "Extraction of total cellular DNA from plants, algae and fungi". In: *Plant Molecular Biology Manual*. Ed. by S. B. Gelvin and R. A. Schilperoort. Plant Molecular Biology Manual. Dordrecht: Springer Netherlands, pp. 183–190. ISBN: 978-94-011-0511-8. DOI: 10.1007/978-94-011-0511-8_12.

Ryan, K. G., P. Ralph, and A. McMinn (Oct. 2004). "Acclimation of Antarctic bottom-ice algal communities to lowered salinities during melting". In: *Polar Biology* 27.11, pp. 679–686. ISSN: 07224060. DOI: 10.1007/s00300-004-0636-y.

Schlosser, C., K. Schmidt, A. Aquilina, W. B. Homoky, M. Castrillejo, R. A. Mills, M. D. Patey, S. Fielding, A. Atkinson, and E. P. Achterberg (2018). "Mechanisms of Dissolved and Labile Particulate Iron Supply To Shelf Waters and Phytoplankton Blooms Off South Georgia, Southern Ocean". In: *Biogeosciences* 15.16, pp. 4973–4993. DOI: 10.5194/bg-15-4973-2018.

Sokolov, S. and S. R. Rintoul (2009). "Circumpolar Structure and Distribution of the Antarctic Circumpolar Current Fronts: 1. Mean Circumpolar Paths".

In: *Journal of Geophysical Research* 114.C11, p. C11018. DOI: 10.1029/2008jc005108.

Stefan, C. P., A. T. Hall, A. S. Graham, and T. D. Minogue (2022). "Comparison of Illumina and Oxford Nanopore Sequencing Technologies for Pathogen Detection From Clinical Matrices Using Molecular Inversion Probes". In: *The Journal of Molecular Diagnostics* 24.4, pp. 395–405. DOI: 10.1016/j.jmoldx.2021.12.005.

Tehei, M. and G. Zaccai (Aug. 2005). "Adaptation to extreme environments: Macromolecular dynamics in complex systems". In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1724.3, pp. 404–410. DOI: 10.1016/j.bbagen.2005.05.007.

Tirichine, L., A. Rastogi, and C. Bowler (Apr. 2017). "Recent progress in diatom genomics and epigenomics". In: *Current Opinion in Plant Biology* 36, pp. 46–55. ISSN: 13695266. DOI: 10.1016/j.pbi.2017.02.001.

Tréguer, P., C. Bowler, B. Moriceau, S. Dutkiewicz, M. Gehlen, O. Aumont, L. Bittner, R. Dugdale, Z. Finkel, D. Iudicone, O. Jahn, L. Guidi, M. Lasbleiz, K. Leblanc, M. Levy, and P. Pondaven (Jan. 2018). "Influence of diatom diversity on the ocean biological carbon pump". In: *Nature Geoscience* 11.1, pp. 27–37. ISSN: 1752-0894. DOI: 10.1038/s41561-017-0028-x.

Wan, W., G. M. Gadd, D. He, W. Liu, X. Xiong, L. Ye, Y. Cheng, and Y. Yang (2023). "Abundance and diversity of eukaryotic rather than bacterial community relate closely to the trophic level of urban lakes". In: *Environmental Microbiology* 25.3, pp. 661–674.

Whitehouse, M., A. Atkinson, R. Korb, H. Venables, D. Pond, and M. Gordon (2012). "Substantial Primary Production in the Land-Remote Region of the Central and Northern Scotia Sea". In: *Deep Sea Research Part II: Topical Studies in Oceanography* 59-60.nil, pp. 47–56. DOI: 10.1016/j.dsr2.2011.05.010.

Zhong, D., L. Listmann, M.-E. Santelia, and C.-E. Schaum (2020). "Functional redundancy in natural pico-phytoplankton communities depends on temperature and biogeography". In: *Biology Letters* 16.8, p. 20200330.

4

Pier-Seq: From boats to buckets, developing an improved workflow for *in situ* nanopore sequencing of ocean microbiomes

4.1 Introduction

This chapter presents work carried out to develop an improved workflow for *in situ* nanopore sequencing and real-time analysis, and to test it in the field. A time-course experiment was also carried out to investigate the utility of nanopore sequencing for monitoring population flux. I performed all of the DNA extraction experiments, and the live sequencing experiment on Cromer Pier, supported by Richard Leggett, Darren Heavens, and Ned Peel from Earlham Institute. The DNA extraction and laboratory-based sequencing was carried out by Darren Heavens due to occupancy restrictions as a result of covid-19. I performed all of the analysis of the seqeuncing data.

4.1.1 Overview

One of the most exciting opportunities associated with nanopore sequencing is the potential of using cheap, user friendly protocols to support wide-scale high spatio-temporal resolution citizen science projects. This chapter covers the development of an adapted, simpler and safer protocol for use on British beaches, perhaps by several citizen science or school groups either in a single location over a period of several weeks or on one occasion by groups located around the entire coastline. A simplified protocol with less reliance on highly toxic chemicals could also be used with little training by general research staff onboard research cruises, potentially allowing for the use of filtration and sequencing as a standard data collection protocol.

It was hoped that this refined protocol would streamline the pipeline making planning future trips more straightforward, more reactive, and ultimately de-skill the process enabling training of non molecular biology educated staff to perform the experimentation and analysis safely and accurately. This would be particularly useful for citizen science and engagement efforts, as well as data gathering expeditions staffed by general research scientists.

4.1.2 Sample collection and DNA extraction

DNA yields obtained in the field in 3 were impressive, comfortably exceeding the 1 µg required for nanopore sequencing. Developments in flowcells mean that the requirement for DNA in a standard MinION kit is 400 ng, with rapid kits for use in field conditions requiring only 100 ng, with yields of >7.5 Gbp achieved from 110 ng (Heavens et al. 2021).

Given the long collection and filtration times and employing a complex DNA extraction protocol using toxic chemicals, however, the DNA molecule length as indicated by the N50 of 1-2 kbp were somewhat disappointing. This indicated that achieving the desired metagenomic assemblies would be more difficult than expected with low throughput sequencing, and that diagnostic classification might be more appropriate. The workflow therefore allows for a separation between *in situ* and laboratory-based sequencing experiments, aiming to use the best tool for the job at hand. Portable sequencing is best suited to short runs which can be used to determine what is present and inform further sampling, or to give an overview of a wide area. This works well with rapid, low-toxicity DNA extraction protocols using bead-beating which are easy to transport and can be used quickly and easily in the field. Where in depth analysis or the production of MAGs is required, laboratory-based nanopore sequencing is better suited. This allows for full 72 hour runs to maximise yield, which may not be practical in the field, non-portable GridION and PromethION platforms offer high-throughput nanopore sequencing, and the super high accuracy basecaller, which is not available on the MinION Mk1C, can be used for improved accuracy. In this case, a longer extraction protocol requiring toxic chemicals to maximise molecular weight may be more acceptable. Therefore, depending on experimental goal, the appropriate DNA extraction method and sequencing technologies can be selected.

DNA extraction protocols can be grouped into three methods- those based on chemical lysis, those using enzymatic lysis and physical lysis. The gold standard approaches for taxonomic assignment of complex metagenomic samples involve bead beating as the physical action of beads colliding and trapping cells between

ensures cell lysis of even the most difficult samples. Advantages of this approach include reduced equipment requirements, fewer steps resulting in a faster processing time, no toxic chemicals and improved DNA yields although some consider a downside to be shorter DNA lengths. With the improved DNA yields also comes the potential to reduce the volume of seawater that needs to be filtered to obtain sufficient material for sequencing.

4.1.3 Advances in Nanopore sequencing and analysis

Since 2019 there have been improvements both in Nanopore sequencing and analysis software. There has been a focus on making sequencing cold chain free and this has resulted in MinION flowcells now being able to be stored at an ambient temperature for up to one month with no significant degradation of pores and the availability of lyophilised reagents for library construction. Sequencing accuracy has also improved from 93 to 99% and flowcell yields of over 20 Gbp being achieved in the laboratory. An added advantage of the increase in accuracy is that classification of shorter reads will also be improved. In terms of software NanoOK has morphed into MARTi (https://marti.cyverseuk.org) (Leggett et al. 2018) which improves the user experience, with a simplified graphical user interface, and a wider variety of analyses available to perform in real time.

ONT have released the Flongle flowcell which offers the ability to produce smaller nanopore sequencing datasets using cheaper flowcells with lower yield, up to 2.8 Gbp at a cost of around £60 per sequencing run. At approximately 1/10th of the cost of a MinION flowcell, this reduces the cost barrier for involving citizen science groups and increases the time-range or number of sample sites that can be covered, at a lower yield presenting a snapshot of the coastal microbiome. ONT are also developing a smartphone-based sequencing platform run from an app, called the SmidgION, with further reduced sequencing yield but which could be used in conjunction with quantitative analysis such as qPCR to identify harmful microbes.

Together with sample collection and DNA extraction developments, this brings us to a point where anyone could run the experiments with minimal training, raising the possibility of it being used by citizen science groups or schools outreach events, to capture a snapshots of the marine microbiome at specific locations or at around the coastline at one time. Another potential use is for scientists who are already present on research cruises or at research stations, allowing for an increase in the areas sampled and opens the possibility of it being added to

the standard roster of analyses and being used to direct research cruises and sampling.

4.1.4 Cromer Pier

It was decided to evaluate the improved workflow and improved nanopore sequencing capabilities locally, through a longitudinal experiment carried out on the Norfolk coast in the summer of 2021, in accordance with national and local lockdown restrictions at the time.

The Norfolk coast is part of the North Sea which is a shelf sea of the Atlantic Ocean covering 570,000 square kilometres between Great Britain, Norway, Denmark, Germany, the Netherlands, Belgium, and France, as can be seen in figure 4.1.1. It meets the Atlantic Ocean from the English channel in the southwest, and the Norwegian Sea in the north. Shelf seas are more productive than open oceans and support over 80% of fisheries worldwide (Pauly et al. 2002); the North Sea supports economically important fishing and boating industries, as well as crude oil extraction and coastal tourism and leisure for the countries around it, especially the UK.

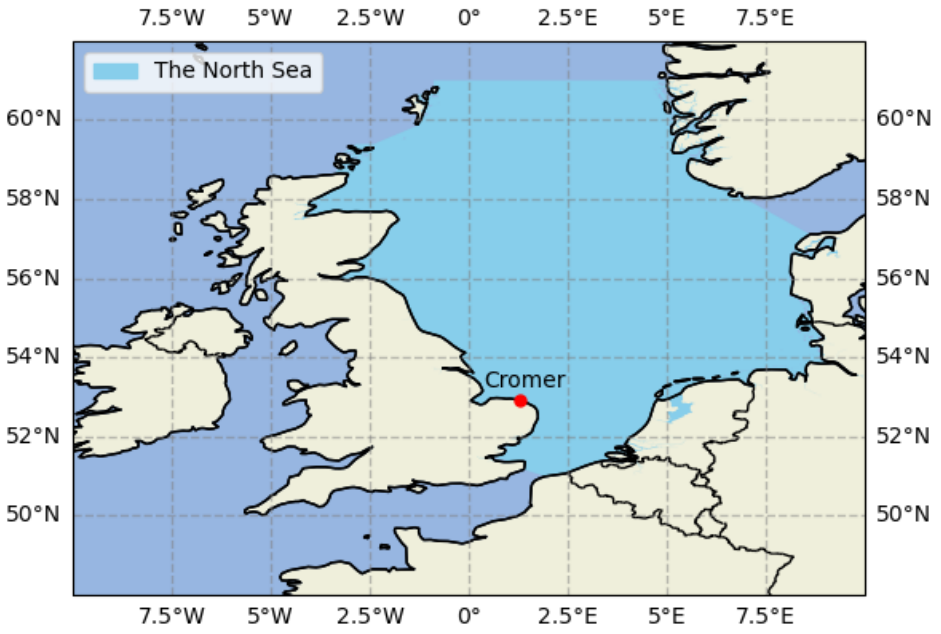


Figure 4.1.1: Map showing the sampling location of Cromer, and the North Sea boundaries as defined by the International Hydrographic Organisation.

Particular concerns in this area include the effects of climate change. There are regular blooms of phytoplankton off the Norfolk coast, including of Emiliania huxleyi which is important in the global carbon pump (Dassow et al. 2009). The coastline is extensively monitored with satellite imagery and sampling by Plymouth Marine Laboratory (PML) which hosts the NERC Earth Observation Data Acquisition and Analysis Service (NEODAAS) providing satellite data to researchers studying the area. This imagery shows when large blooms occur, and can indicate the levels of phytoplankton in the water, as well giving information on temperature and other metadata.

This area has not been the subject of previous DNA sequencing investigations, and there may be interesting insights into the microbial populations on the Norfolk coast. A 2017 study (Capuzzo et al. 2018) found that primary production in the North Sea has reduced between 1988 and 2013 based on chlorophyll and underwater light measurements. This change is thought to be due to increased sea surface temperatures, and reduced inflow of nutrients from rivers. Reduced primary production over this time period was found to correlate with reductions in zooplankton and a number of commercially important fish species. Projections of the effects of climate change on the North Sea reflect a complex balance of drivers of primary productivity. Climate change will increasingly affect marine life. Temperature increases, and increased acidity due to increased CO_2 concentrations, along with changes to precipitation patterns affecting riverine outflow, and changing wind patterns altering sea states and upwellings, will all contribute to shifts in presence and abundance of marine species at all trophic levels (Holt et al. 2016). It is not known whether the Norfolk coast has been affected in the same way. DNA sequencing of microbe populations could allow us to find out and to monitor future changes.

Understanding the coastal microbiome and monitoring for changes as a result of climate change could be useful across a range of fields including scientific research into the effects of climate change on biodiversity and community structure, assessing the health of the ecosystem as a whole, and the potential impacts on the coastal economy through fishing and tourism. Coastal microbiomes are particularly affected by human activity due to proximity, with higher concentrations of persistent chemicals, dissolved organic carbon, sewage discharges, and other pollutants all of which affect the marine microbiome (Cerro-Gálvez et al. 2021). To date there is no coastal micriobiome monitoring project in the U.K, although the Western Channel Observatory, which collects time-series oceanographic and biodiversity data in the Western English Channel, has started to incorporate molecular approaches to further investigate

biodiversity and function which is not captured by traditional methods (*Western Channel Observatory 2022* 2022).

The collection and analysis of biodiversity samples and DNA sequencing data is time consuming, and can be costly, but there are numerous examples globally of citizen science programmes producing high quality ecological data which increases the immediate understanding of the ecosystem and can be used by researchers for more complex analysis, including CALeDNA programme where volunteers collected environmental DNA samples (Meyer et al. 2021). There is currently a great deal of public interest in the quality of coastal waters, the effects of pollutants on the marine ecosystem, and the presence harmful microbes around the British coast, due to highly publicised increases in the discharge of untreated sewage into the sea since 2019, with sequencing providing new insights into the impacts on water quality and public health risks (Zan et al. 2023).

Given this high level of public interest, there is the potential for a citizen science programme monitoring the coastal microbiome through sampling and nanopore sequencing, reporting on microbiome composition and diversity, and potentially monitoring water quality with regard to human health by checking for the presence of toxin producing phytoplankton and pathogenic bacteria and viruses. The main human health concerns in UK waters are toxin producing phytoplankton, especially where they occur in high concentrations forming harmful algal blooms (HABs), and *Vibrio* bacteria and noroviruses (NoV). The prevalence of these harmful organisms is likely to increase due to climate change, as HABs are more likely to occur in warmer, less well mixed waters, and more frequent and intense storms result in increased sewage overflow events, which lead to discharge of untreated sewage into the sea (Bresnan et al. 2020; Karlson et al. 2021). MinION sequencing has been used to assess water quality in the USA and India with some success, although not all toxin genes were identified (Hamner et al. 2019; Acharya et al. 2019).

A further potential use for nanopore sequencing is outreach, for example with schools and young people, as well as programmes aimed at the general public. Outreach is crucial to developing future research scientists and improving understanding within the public about the potential impacts of climate change on our oceans, and the importance of the ocean ecosystems and microbe communities, which is important where political action is required to reduce climate change (Barberán et al. 2016). Programmes such as the Darwin Tree of Life Project are developing outreach programmes to engage the public in the aim of capturing the full biodiversity of the British Isles through DNA sequencing (Lewin et al. 2018), with portable nanopore sequencing a potentially useful outreach tool.

4.1.5 Pier-Seq

It is important to stress test any protocol destined for in field use, particularly where it will be used by those with little training, to ensure that it is robust and fit for purpose. To test the new and improved protocols a pier sequencing pipeline (Pier-Seq) was devised which would act as a model for potential future citizen science sequencing efforts. Sampling at high tide from the end of a pier would guarantee access to seawater without needing a boat, giving flexibility for collection times, while the use of lithium ion battery power would allow DNA extraction, sample processing and data analysis to be performed on the pier itself without the need to take samples back to the laboratory. If this proved successful it would be a good indicator as to the suitability of the revised pipeline for on ship deployment.

The aim of this study was to develop a streamlined, low toxicity, user-friendly nanopore sequencing workflow which can be used with minimal training by citizen scientists to investigate coastal microbe populations. ONT MinION sequencing was used to investigate the phytoplankton populations in the North Sea at Cromer Pier in the summer of 2021, alongside satellite imagery to build a picture of which organisms were present and investigate whether nanopore sequencing, for example through using Flongle flowcells or the forthcoming ONT SmidgION platform, would be an effective tool to monitor changing populations and alert researchers to the presence of harmful or pathogenic microbes.

4.2 Methods

In order to establish this improved workflow we carried out the following processes:

- Improving DNA extraction methods
- Time-course sampling carried out at Cromer Pier
- Live sequencing experiment carried out at Cromer Pier
- Laboratory sequencing of time-course samples
- Analysis of all sequencing samples

4.2.1 Improving DNA extraction methods

DNA extraction and sequencing

Prior to sampling at Cromer Pier, to test improved protocols, DNA extractions were carried out on samples collected during the DY098 cruise but not sequenced on board. A range of extraction techniques were tested do determine the best method for extracting DNA from ocean metagenomic samples. Sample 1 from Ship-Seq, see 3 was selected for this experiment, with 1/8 of the filter used for each protocol.

Methods tested:

- Modified CTAB
- CTAB with beads
- MagAttract
- Qiagen power soil

Modified CTAB was carried out as with DY098 (see Chapter 3) using chloroform kept at -18 ℃ instead of PCIA and without 2-mercaptoethanol

CTAB with DNeasy beads - as with modified CTAB but with DNeasy beads added to the CTAB mixture and 5 minutes bead beating at 1500 rpm prior to incubation.

MagAttract + DNeasy beads - as in the MagAttract manual but with DNeasy beads added to the lysis buffer and 5 minutes bead beating at 1500 rpm prior to step 5.

Qiagen Power Soil kit - as in the Qiagen Power Soil manual but with DNeasy beads added to the lysis buffer and 5 minutes bead beating at 1500 rpm prior to step 3.

Extracted DNA from each of these methods was measured on a Qubit 2.0 to determine DNA recovery and an Agilent TapeStation to determine DNA molecular weight, and sequenced using the SQK-LSK109, EXP-NBD104, and EXP-NBD114, with barcoding on one flowcell and demultiplexed into the original samples using guppy.

Analysis

The extraction methods were compared based on N50, yield, and ease of extraction onboard a research vessel.

The sequencing results were analysed using the BLAST-nt database and MEGAN (Huson et al. 2007), using a minimum lowest common ancestor minimum support percentage of 0.1%. With this cutoff, taxonomic nodes containing less than 0.1% of the reads would be merged with their parent node.

The BLAST outputs from Chapter 3 - in depth filtering of Sample 6 - were analysed using MEGAN as above, using the first 20,000, 40,000, 60,000, 80,000, 100,000, 200,000, 300,000, 400,000, 500,000, and 600,000 reads. The log of the number of reads was plotted against the log of the number of genera matches for each fraction and a linear model applied in R with 95% confidence intervals.

The sequencing output from Chapter 3 - in depth filtering of Sample 6 - was sorted by length and filtered to remove reads below 100bp. The resulting FASTA file was split into 12 fraction each containing 267,998 reads in order of increasing length. The lengths are shown in table 4.1. This was split into chunks of 100 reads. 500 chunks from each section were randomly subsampled for BLAST and MEGAN analysis as above. Fraction 1 contained insufficient BLAST-nt hits for MEGAN analysis to proceed. The log of the median length of reads in each fraction was plotted against the log of the number of genera matches for each fraction and a linear model applied in R with 95% confidence intervals. Pearson's correlation coefficient was also calculated.

4.2.2 Sampling and filtration

Samples were collected weekly from the same location at Cromer Pier, at high tide over 6 weeks in from the 22nd of July to the 2nd of September 2021 (excluding the week commencing 09/08/2021) with a live sequencing experiment on the 9th of September. A bucket was lowered from the pier to the surface of the sea and filled before being raised again. The water was left to settle if particularly silty and then a syringe was used to filter 20 mL of sea water through two 0.22 μm Swinney filters. The filters were placed on dry ice and transported back to storage at -80 ℃ before sequencing after all samples were collected. Sea temperature, date, and time were recorded, along with general weather conditions. Satellite imagery showing chlorophyll levels was provided by PML NEODAAS.

Table 4.1: Read length fractions of sample 6

Fraction	Min length	Max length
1	100	379
2	379	574
3	574	820
4	820	1128
5	1128	1515
6	1515	1993
7	1993	2602
8	2602	3436
9	3436	4648
10	4648	6639
11	6639	10532
12	10532	121004

This protocol was repeated for the live sequencing experiment, although part of the sample was processed in entirety through to sequencing and analysis on the pier.

4.2.3 DNA Extraction and library preparation

Laboratory extractions and library preparation

Magnetic bead based DNA extraction adapted from (Heavens et al. 2021).

20 mL of sea water was filtered using a syringe and a Millipore 13 mm swinny filter with a 0.2 µm filter fitted. The filter containing the biomass was placed in a 2 mL tube containing 0.25 gs of PowerSoil Pro zirconium beads and 125 µL of CD1 buffer and this in turn placed in the SuperFastPrep-2 and beaten for 20 seconds at a speed code of 20. The tube was then spun for 30 seconds at 10,000 rcf in an Eppendorf 5415R centrifuge (Eppendorf, Stevenage, UK) and the supernatant transferred to a fresh 1.5 mL Lobind Eppendorf tube (Eppendorf).

An equal volume of Kapa pure beads (Roche) was added to the lysed and bead beaten cells, vortexed and then incubated at room temperature for 5 minutes. The tube was then pulse spun in a microfuge then placed in a magnetic particle concentrator and the beads allowed to concentrate. The supernatant was

discarded and the beads were washed twice with fresh 70% ethanol. Care was taken to remove all the ethanol and the tube removed from the MPC and the beads resuspended in 10 µL of Qiagen CD6 buffer and incubated at room temperature for 2 minutes. The tube was then pulse spun in a microfuge then placed in a magnetic particle concentrator and the beads allowed to concentrate. The supernatant containing the DNA was then transferred to a fresh 1.5 mL Lobind Eppendorf tube.

A 1 µL aliquot of DNA was used to determine concentration using the Qubit BR assay (Life Technologies, Loughborough, UK) and a second 1 µL aliquot was used to determine molecule length with an Agilent Genomic Tape (Agilent, Cheadle, UK) on an Agilent TapeStation (Agilent).

DNA quantity was measured using a Qubit and processed for sequencing using the Nanopore native barcoding kit 104. Sequencing was performed on a GridION using the highly accurate basecaller settings.

Field extraction and library preparation

Figure 4.2.1 shows the equipment before and during live sequencing at Cromer Pier.

The same extraction process was used for the live Pier-Seq extraction and sequencing, using the Nanopore rapid barcoding kit and the built-in basecaller settings for the MinION Mk1C. For further analysis away from the pier, the sample was re-basecalled using guppy's Super Accuracy basecaller model.

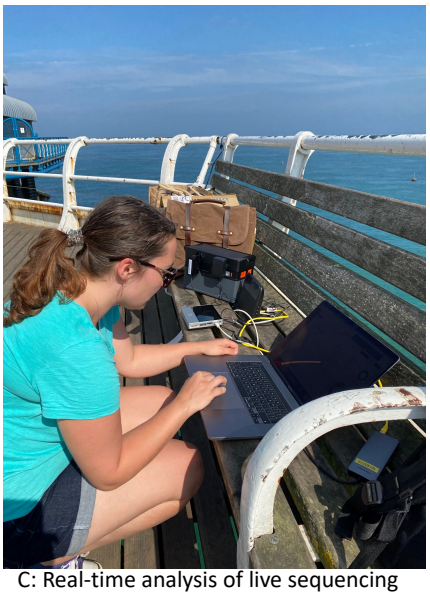
4.2.4 DNA sequencing and analysis

MARTi

Sequencing results were analysed using MARTi (http://marti.cyverseuk.org) (Leggett et al. 2018) which was configured with a minumum read length of 100 bp and minimum read quality score of 8 based on MinKnow output, BLAST-nt settings were maxiumum hits per read 20, minimum percent identity 60, and minimum read length 100 bp. These limits were also used to filter BLAST output for analysis outside MARTi. MARTi was used to produce plots at various taxa levels as well as to produce taxa accumulation curves to assess data completeness.

A: Field equipment packed in a 50l box. MinION Mk1C visible on top.

B: DNA extraction and sequencing kit on a bench at Cromer Pier: Magnetic test-tube rack, pipette, RapidPrep, Qubit, spare MinION Mk1C, battery power source, microfuge, MinION flowcell, running MinION Mk1C



C: Real-time analysis of live sequencing data, showing laptop, MinION Mk1C, router, and battery power source

Real-time analysis: Laptop running MARTi showing output

Figure 4.2.1: Photographs of the equipment before and during live sequencing at Cromer Pier. Panel A: Field equipment packed in a 50l box; Panel B: DNA extraction and sequencing kit on a bench at Cromer Pier; Panel C: Real-time analysis of live sequencing data; Panel D: Laptop running MARTi with output

Basecalled data from the time-course experiment was exported to MARTi for analysis after sequencing.

Basecalled data from the live sequencing experiment was transferred in real-time from the MinION to a laptop using rsync. On the laptop, MARTi performed BLAST-based analysis using a subset of the nt database. MARTi makes use of a BLAST option which allows the search space to be restricted to specified taxa IDs, resulting in much quicker analysis. In this case, we restricted the search to the set of taxon IDs representing Chlorophyta, SAR, Haptista, Viridiplantae, Bacteroidetes, and Proteobacteria.

MEGAN

Samples were subsequently analysed using MEGAN (Huson et al. 2007), with a lowest common ancestor minimum support percentage of 0.1%.

4.3 Results

4.3.1 DNA extraction and sequencing improvements

Yield and sequencing results

DNA yields are shown in table 4.2. CTAB extraction had the highest yield, followed by Qiagen, CTAB with beads and then MagAttract.

Table 4.2: DNA recovery in ng for each extraction method caption of the table

Extraction method	DNA recovery (ng)
CTAB	38.4
CTAB beads	23.0
Qiagen	28.6
MagAttract	21.0

TapeStation outputs for the different extractions are shown in figure 4.3.1

The extraction methods all show a peak at around 100 bp except for CTAB without beads. MagAttract had very little HMW DNA with a low peak of 100 FU at 33 kbp, Qiagen PowerSoil had a broad peak at around 1.3 kbp, CTAB without beads had a low peak at around 4 kbp, and CTAB with beads had a sharp peak of nearly 700 FU at around 10 kbp. TapeStation outputs give an indication of DNA quality prior to library preparation which can affect DNA quality, and can differ from sequencing results as shorter reads are preferentially sequenced during nanopore sequencing.

The sequencing yield, N50, total number of reads, and reads longer than 20 Kbp are shown in table 4.3.

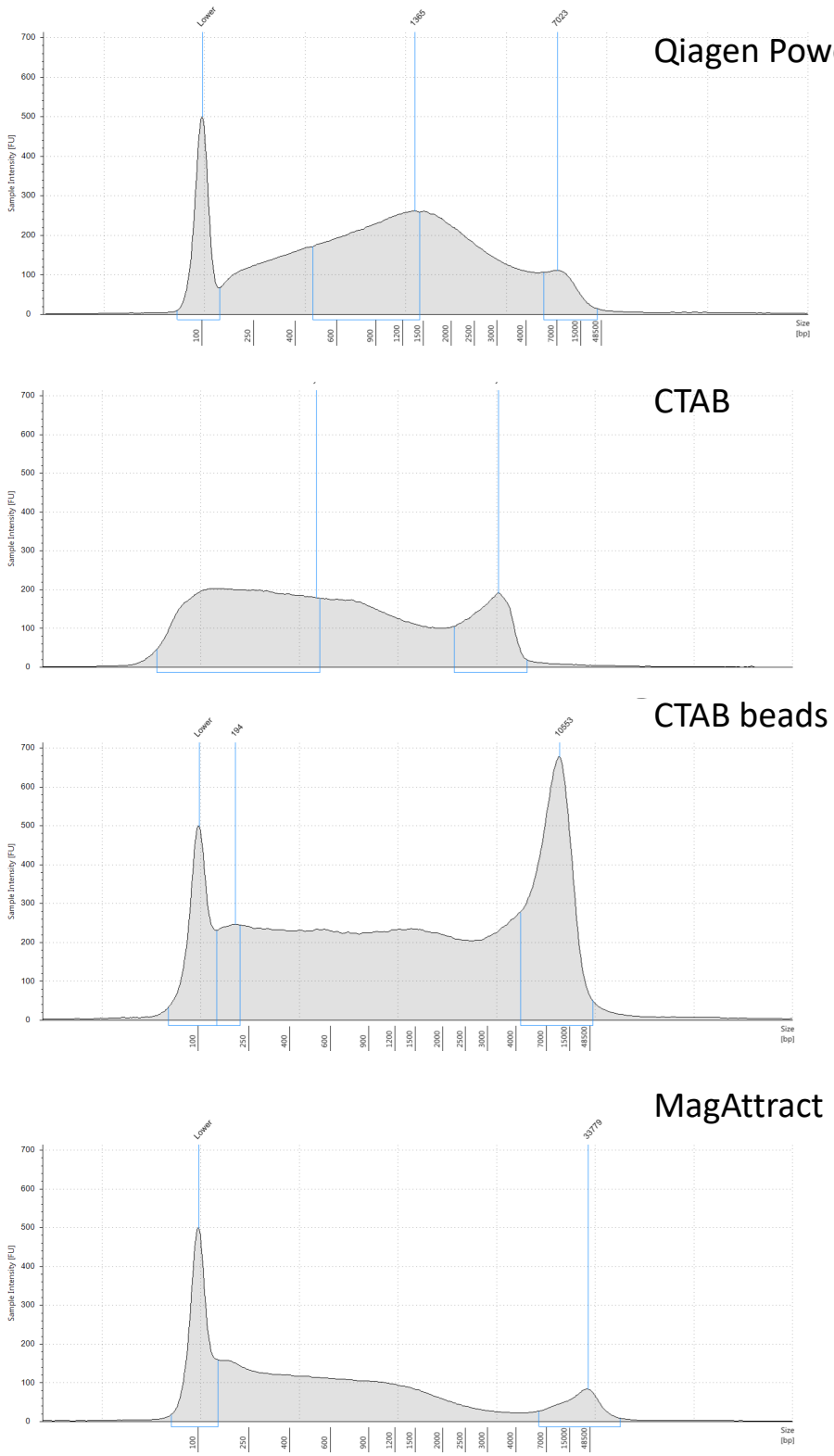


Figure 4.3.1: Tape Station output for Ship-Seq sample 1 extraction

Table 4.3: Yield, N50, total number of reads, and number of reads longer than 20 Kbp for each extraction method

Extraction	Yield (Gbp)	N50 (bp)	Reads	Reads > 20 kbp
CTAB	1.301	753	2036183	200
CTAB beads	1.338	935	1835564	61
Qiagen	0.960	652	1703742	197
MagAttract	1.363	592	2553612	270

Identification results

Sequencing data from each of these experiments was analysed using MEGAN to determine whether the DNA extraction method affected the identification of organisms. As can be seen in figure 4.3.2, the extraction methods produced broadly similar results across the top 30 genera.

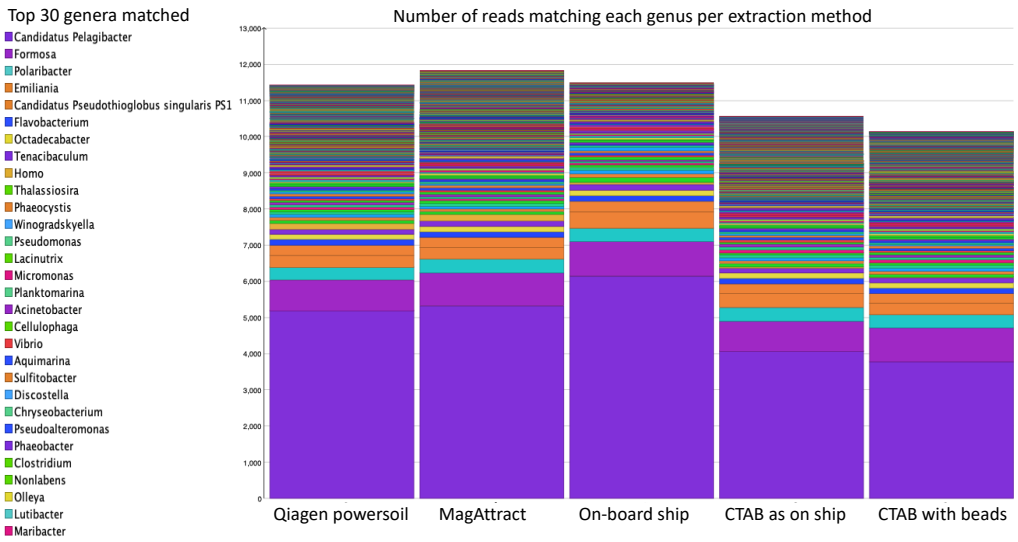


Figure 4.3.2: Stacked bar chart showing the number of reads in each sample to each genera, with the top 30 genera shown in the legend.

The effect of read number on taxonomic identification was investigated using a linear model as shown in figure 4.3.3. It can be seen that there is a strong linear relationship between read number and genera matches. This indicates that the more reads analysed will result in more genera found. The Pearson's correlation coefficient for this relationship was 0.95 indicating a strongly linear relationship between read number and the number of genera matched which is statistically significant, p-value 1.944e-05.

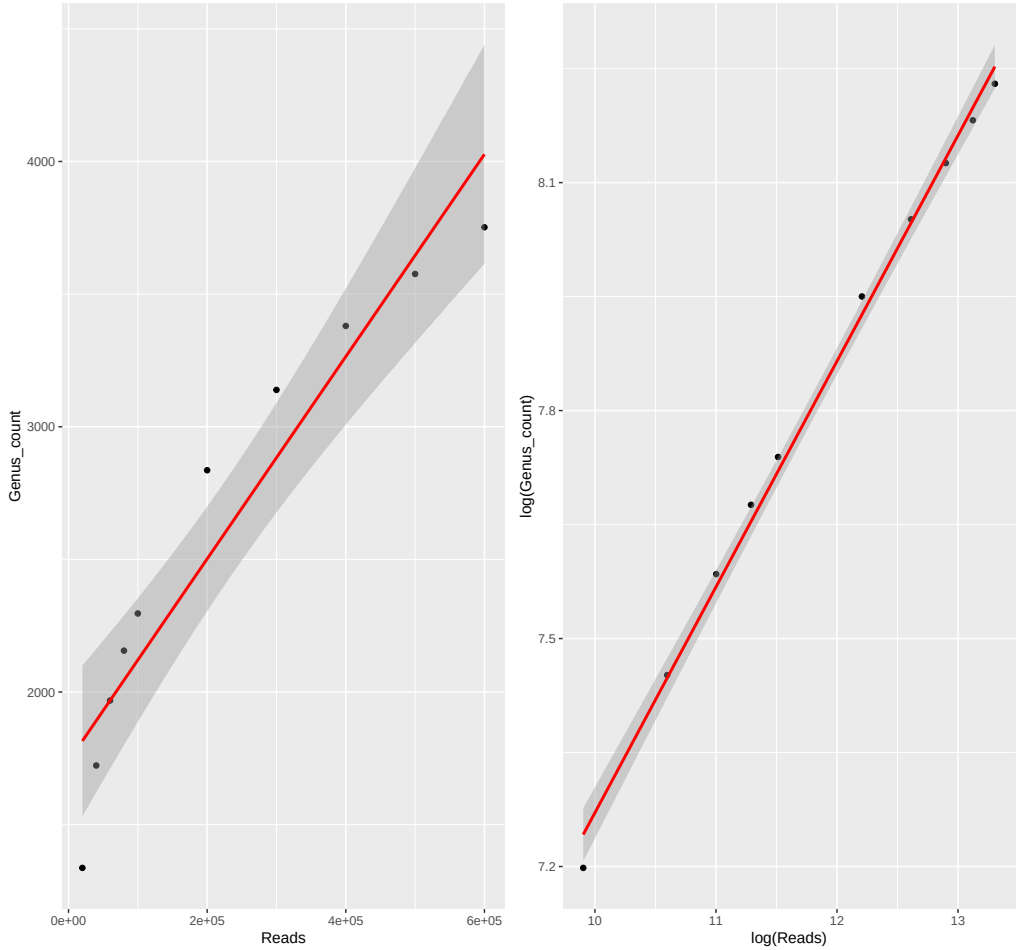


Figure 4.3.3: Linear model showing the relationship between read number and genus count. Panel A read number vs genus count, Panel B log of read number vs log of genus count. Both panels show a linear model showing 95% confidence intervals

To further determine the effect of the number of reads sequenced on taxonomic identification, nanopore sequencing data from Ship-Seq in-depth sequencing of sample 6 was analysed at the genus level in fractions containing the first 200 reads up to 6000 reads. The MEGAN analysis of this experiment is shown in figure 4.3.4. It can be seen that the number of reads analyses has very little effect on the percentage of matches per sample for the first approximately 50 matches. These matches constitute 60% of the matches in each sample. After the first 50 matches there is a clear effect on the genera represented in each sample and they can be seen to diverge.

To investigate the effect of read length on taxonomic identification, the same linear model was carried out as for the read number analysis above. As can be seen in figure 4.3.5, genus count increases with increasing read length up to 5 kbp, above which the genus count remains the same, indicating that read lengths greater than 5 kbp do not increase taxonomic classfication at the genus level.

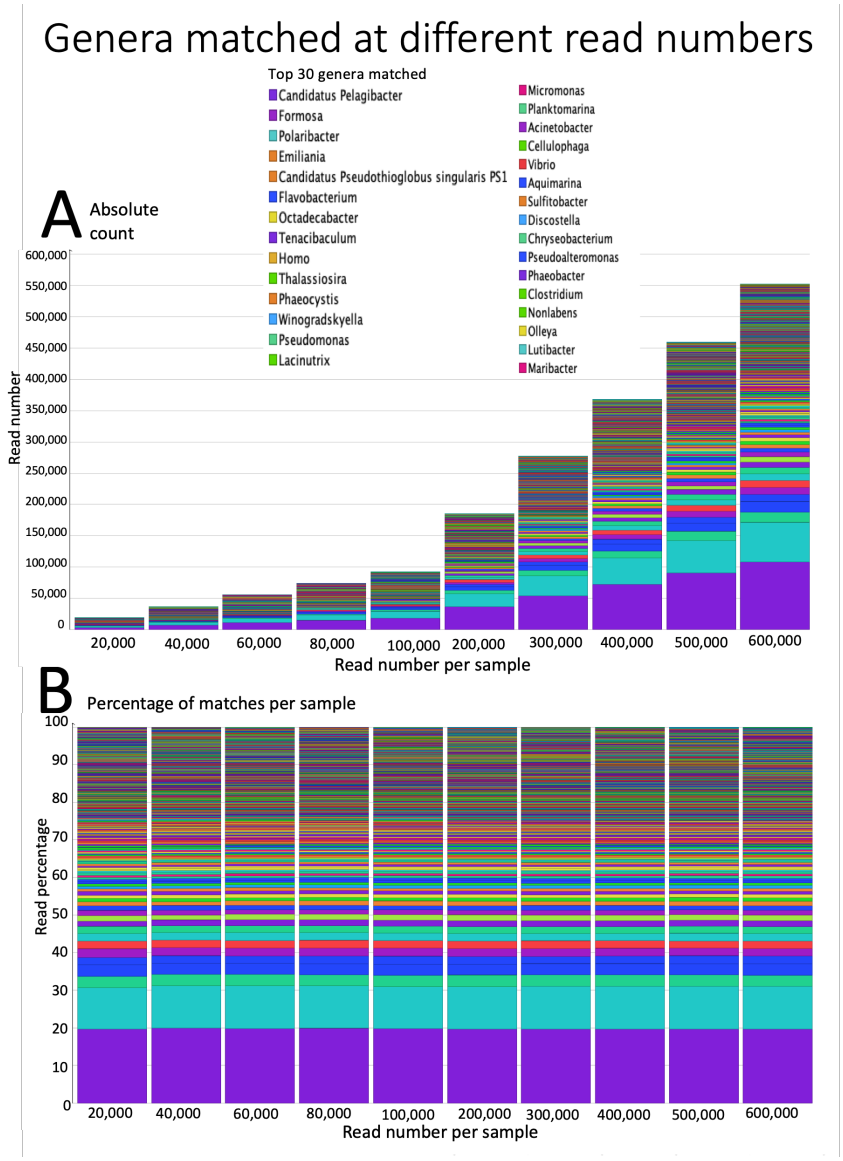


Figure 4.3.4: Stacked bar charts showing the effect of read number on genera matched in MEGAN. Each sample contains an increasing number of reads. Panel A shows matches as an absolute count, panel B shows matches as a percentage of counts per sample.

To further investigate the the effect of read length on taxonomic identification, data from Ship-Seq in depth sequencing of sample 6 was analysed at genus level in fractions of increasing read length. The MEGAN analysis of this experiment is shown in figure 4.3.6 showing the 11 fractions which had sufficient matches for analysis and the unfractionated original sample. There are clearly visible differences in the stacked bar chart for each fraction. The order of appearance is the same in each fraction for top 25 genera matched, although the proportions are different, and after the top 25 matches the order diverges. These genera are representative of 55-65% of the matched hits. The fraction most similar to the

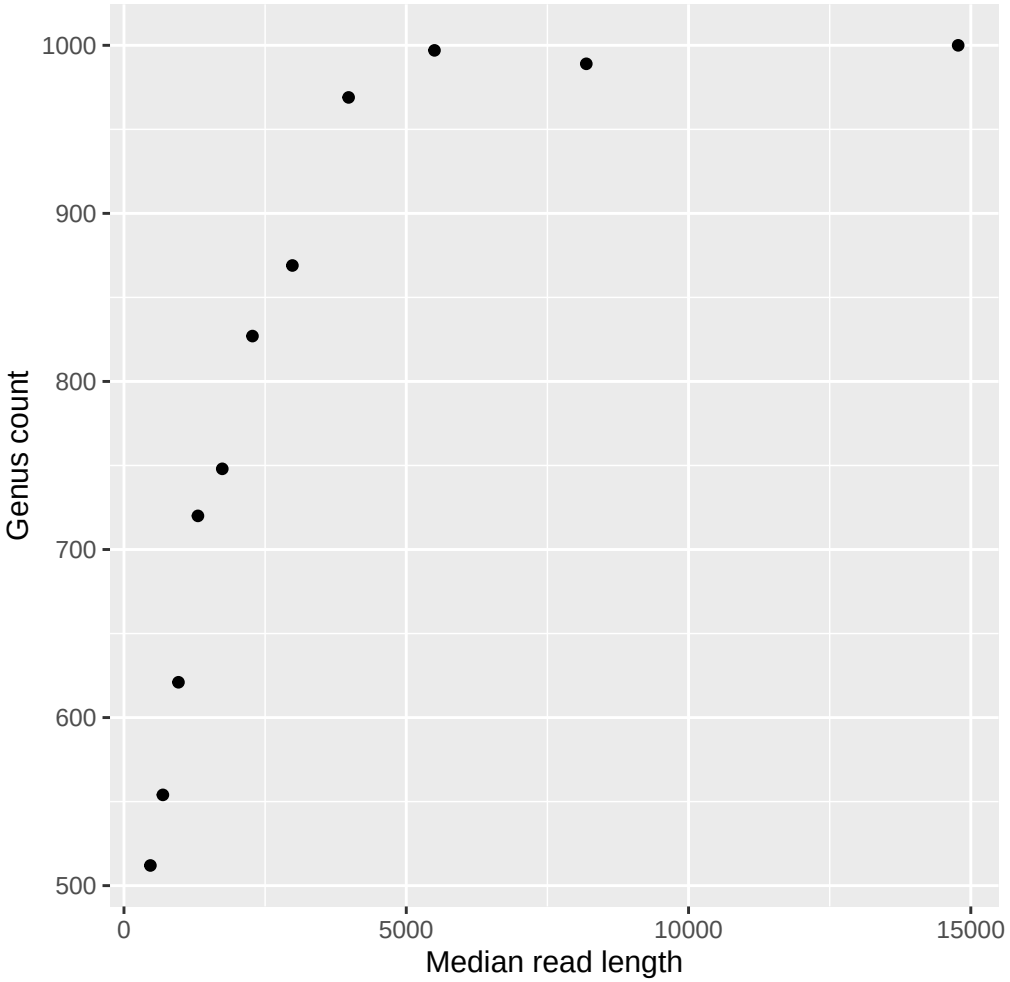


Figure 4.3.5: Scatter plot showing the relationship between read length and genus count. Read length increases with genus count up to 5 kbp, after which genus count does not increase.

unfractionated sample is fraction 8 which has a length range of 2602 to 3436 basepairs.

These investigations indicate that for the most frequent matches constituting the bulk of the genera matched in each sample, there is little effect from read length and number. Investigations focussing on less frequent matches could be skewed by insufficient read numbers or reads which are particularly long or short.

4.3.2 Pier-Seq DNA extraction and sequencing results

The time-course samples were run on a GridION for 72 hours, returning a total of 1.4 Gbp of sequence data. The N50 of the time-course samples combined

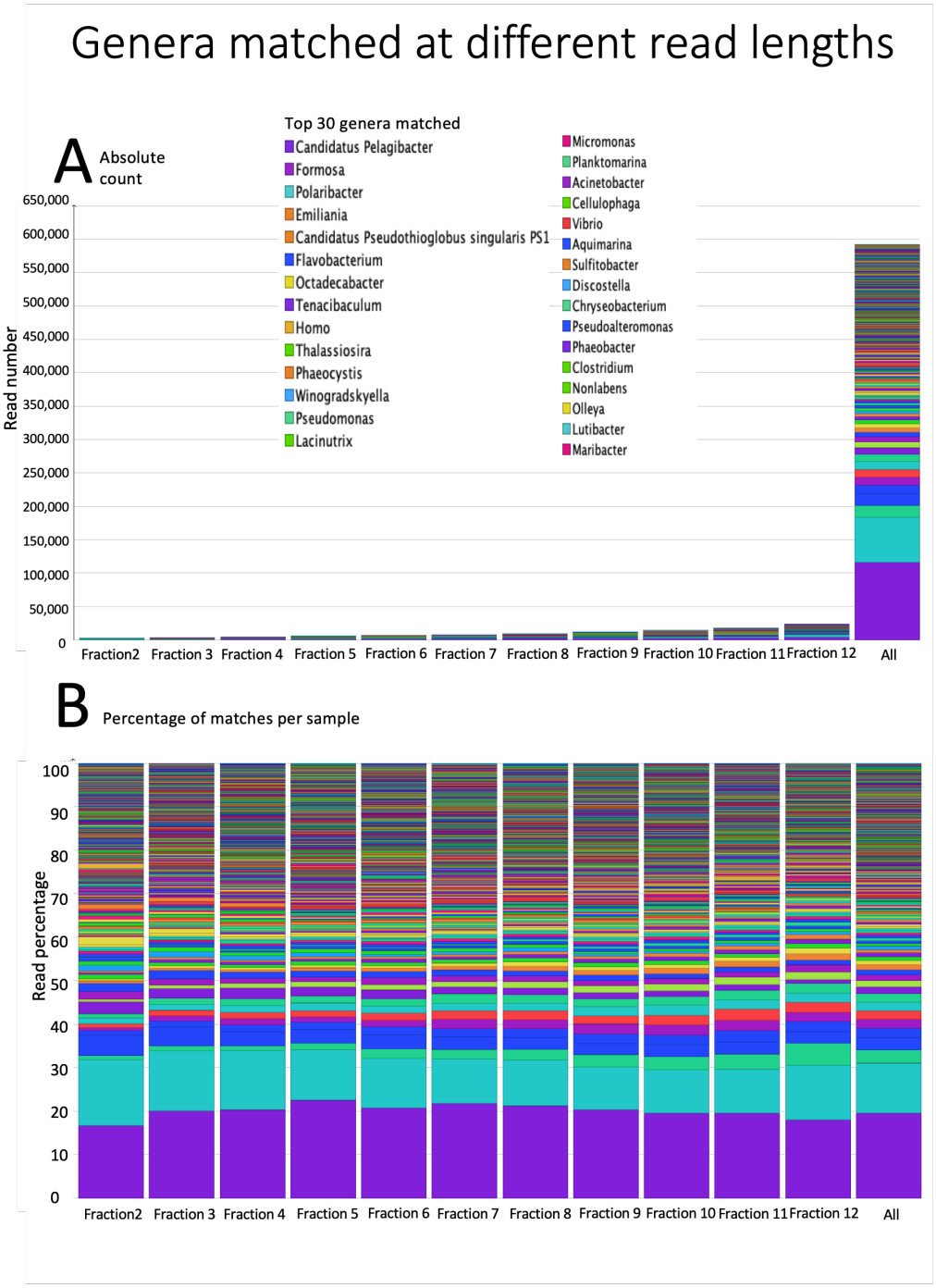


Figure 4.3.6: Stacked bar charts showing the effect of read length on genera matched in MEGAN. Each sample fraction contains reads of increasing length. Panel A shows matches as an absolute count, panel B shows matches as a percentage of counts per sample.

Table 4.4: DNA extraction yield, sequencing yield, total number of reads, N50, and number of reads longer than 10 Kbp for each extraction method

Sample	DNA (ng)	Total (Mbp)	Reads	N50	Mean (bp)	>10000
1	42.5	118.2	163089	916	724	64
2	41.7	619.2	571123	1427	1084	330
3	51.1	437.8	426468	1346	1027	211
4	49.1	139.0	121628	1530	1143	93
5	9.1	0.9	989	1341	952	0
6	16.8	3.4	3454	1430	994	2
7	53.2	190.0	158648	1605	1197	314

was approximately 1330 bp, as can be seen in figure 4.3.7. Table 4.4 shows the sequencing results for the time-course and live sequencing experiment.

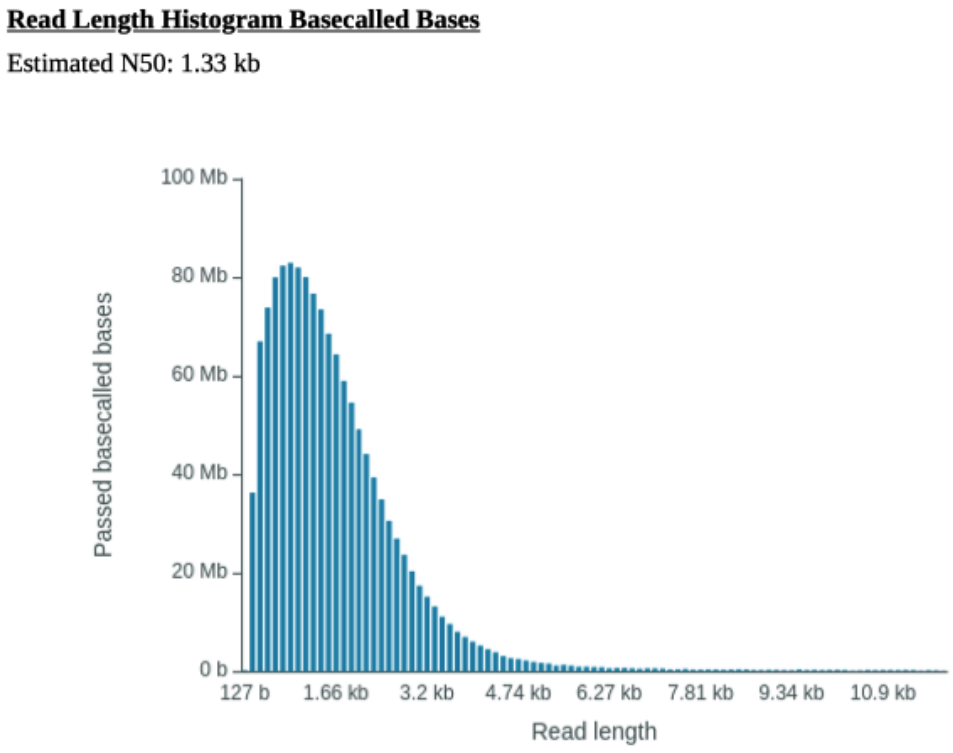


Figure 4.3.7: Histogram produced as part of the Nanpore sequencing summary report showing the read length and N50 for the reads which passed quality checks in of the time-course sequencing experiment.

The live sample was run on a MinION Mk1C for 4 hours *in situ* before transport back to the lab to finish the run. The run returned a total of 191 Mbp and had an N50 of approximately 1580 bp as can be seen in figure 4.3.8.

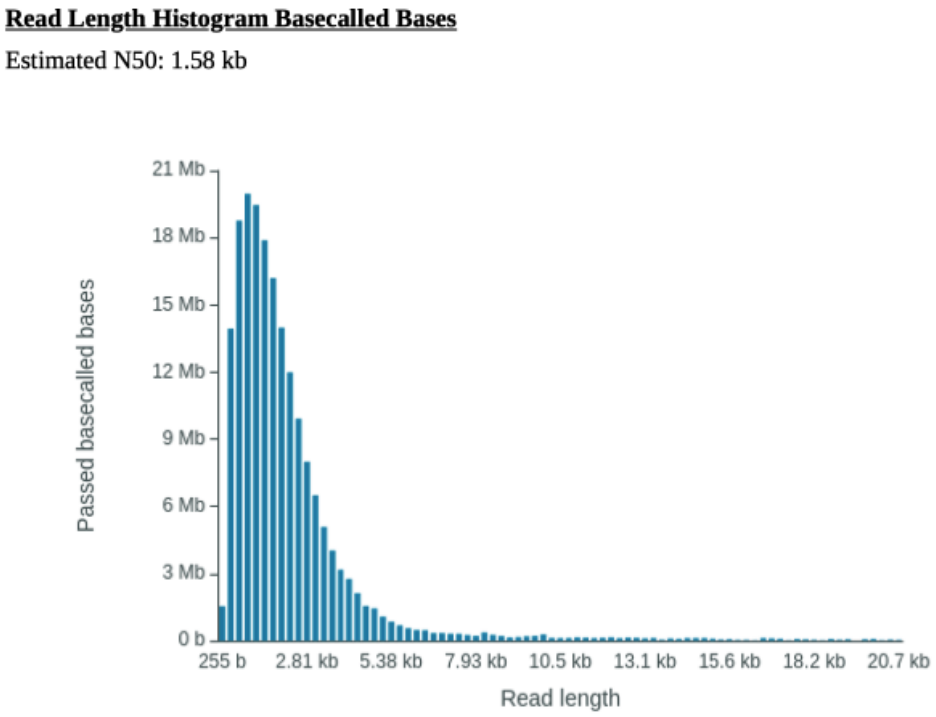


Figure 4.3.8: Histogram produced as part of the Nanpore sequencing summary report showing the read length and N50 for the reads which passed quality checks in of the live sequencing experiment.

Most samples produced 40-50 ng of DNA, with samples 5 and 6 producing far less, at 9.1 ng and 16.8 ng respectively. These samples also resulted in a very low number of reads and total sequencing data compared to the rest of the samples. The mean read length of these two samples was not out of line with those of the other samples, although there were far fewer reads.

The live sequencing produced results within an hour of sampling, with the most abundant genera identified within the 4 hour live sequencing period.

4.3.3 Revised workflow

The revised workflow can be seen in figure 4.3.9 with improved elements highlighted

Improvements are as follows:

A reduced volume of seawater is required for filtration. Filtration volume reduced from 50-100 l to 20 mL. Rapid DNA extraction can be carried out with bead beating and low volumes of reagents which are non-toxic, and can be

Figure 4.3.9: Workflow for the Pier-Seq protocol. Improvements are highlighted in pale blue, with processes which have not been changed in dark blue.

transported at ambient temperature. Rapid library preparation is carried out using a reduced number of reagents and sequencing is done with flow cells which can be transported at ambient temperature, and provide a significantly greater yield.

In situ DNA sequencing is performed using a MinION Mk1C which performs basecalling natively with a standard accuracy basecaller. A high accuracy base caller can be used to re-basecall the data once returned to the lab.

Analysis in the field is carried out using MARTi, a more user friendly, powerful tool compared to NanoOK-RT. This offers the ability to see which organisms are present at various different taxa levels, alongside distributions and rarefaction curves.

4.3.4 Analysis

Establishing the required input and sequencing yield for taxonomic classification

Taxa accumulation curves were produced at species level for each sample, see figure 4.3.10. At species level, none of the samples have plateaued, indicating that more data could yield greater numbers. At genus level, there is some indication of plateauing in samples 2 and 3 which had the most reads, and this is further seen at family level where these samples are beginning to flatten off while the others are not (see Appendices C and G).

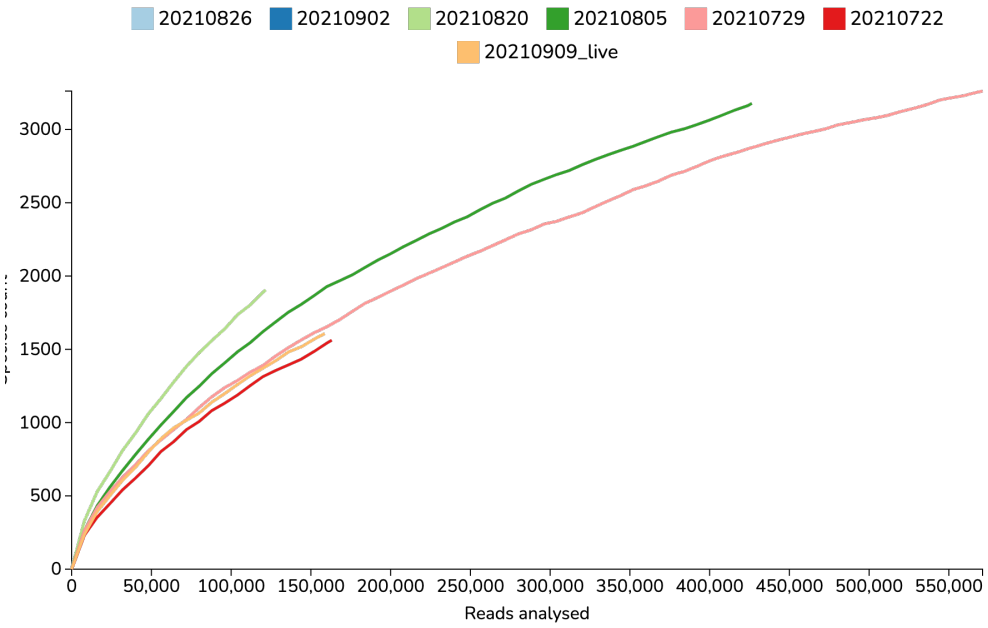


Figure 4.3.10: Taxa accumulation curve at species level, showing species found against reads analysed. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

Table 4.5: Number of taxonomic classifications for archaea, bacteria, and eukaryotes before length and identity filtration.

Sample	Archaea	Bacteria	Eukaryota	Viruses
1	381	2676229	116390	162907
2	2429	9363691	495742	485134
3	2914	7314428	275890	472762
4	1250	2067872	109535	146078
5	3	28225	597	1893
6	9	82334	2329	3542
7	1361	1314709	236507	33254

Table 4.6: Number of taxonomic classifications for archaea, bacteria, and eukaryotes after length and identity filtration.

Sample	Archaea	Bacteria	Eukaryota	Viruses
1	293	120321	58380	24301
2	1646	494432	275960	38329
3	2062	436073	125798	51330
4	993	136126	67724	13794
5	3	652	146	118
6	8	2999	741	224
7	1069	106741	141960	10137

Taxonomic classification of ocean metagenomic samples

The number of reads and BLAST hits before and after filtration based on length of greater than 100 bp and identity greater than 60% are shown in figure 4.3.11.

The number of reads per sample generally correlated with the number of blast hits. Pre- and post-filtration alignments for each superkingdom can be seen in tables 4.5.

At superkingdom level, between around 60% and 75% of the hits of each sample were for bacteria, with 20-30% matching eukaryotes, 5-10% matching viruses, and between 0 and 5% matching archaea see figure 4.3.12. The distribution between superkingdoms does not appear to be related to DNA yield, or read number.

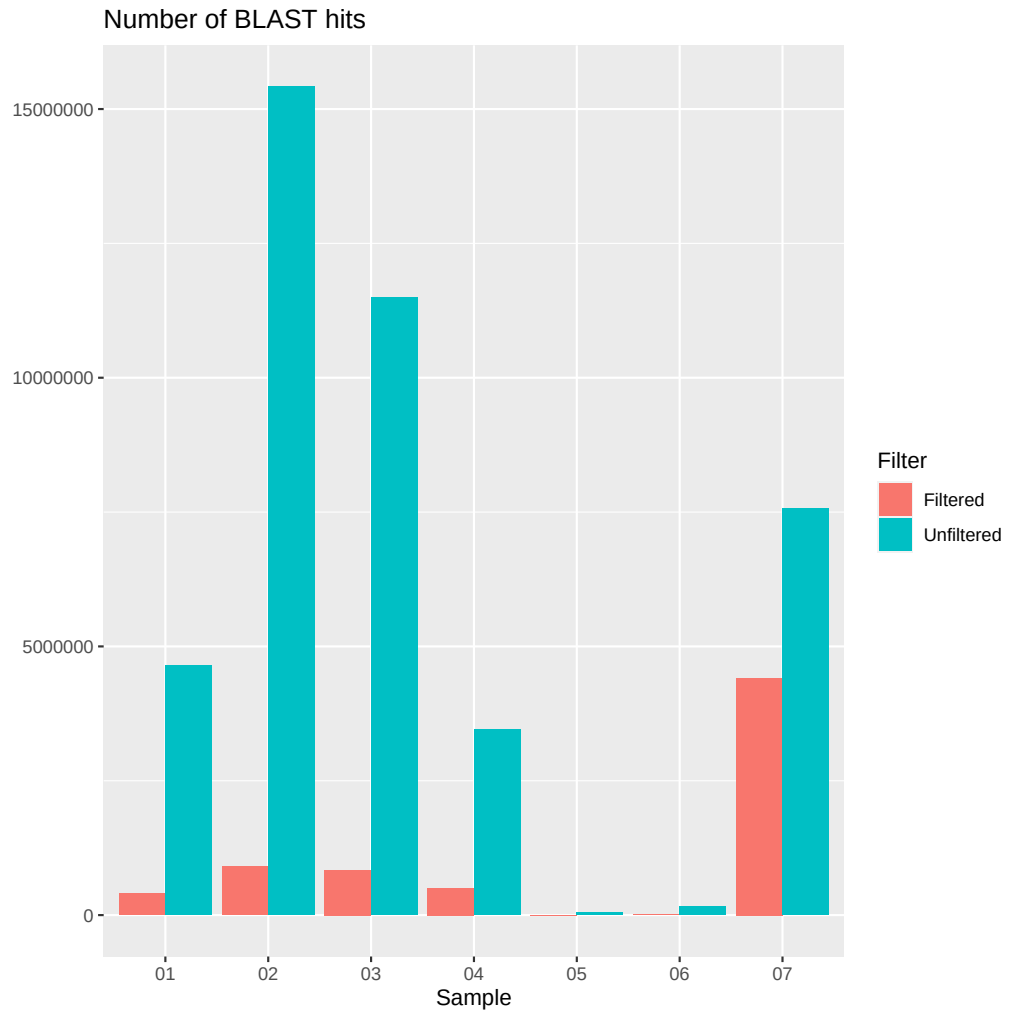


Figure 4.3.11: Bar chart showing number of reads and number of BLAST hits per sample before and after filtration.

The length of hits in each sample for each superkingdom can be seen in figure 4.3.13. The median lengths are similar across all superkingdoms and samples. The median, 25th, and 75th centiles were broadly similar across samples and superkingdoms, with bacteria slightly higher at 445 compared to 376, 364, and 351 for viruses, eukaryota, and archaea respectively. Longer outliers are mainly seen in bacterial and eukaryotic hits with all hits over 15,000 bp matching bacteria alongside most over 5000 bp with a small number matching eukaryota and viruses, and one archaea. Between 2500 and 5000 bp all superkingdoms are present, although hits from samples 5 and 6, which produced little sequencing data, are almost exclusively below 2500 bp for all superkingdoms. It appears from this that length of hit correlates to the amount sequencing data produced, with higher sequencing yields producing more long reads across samples and superkingdoms.

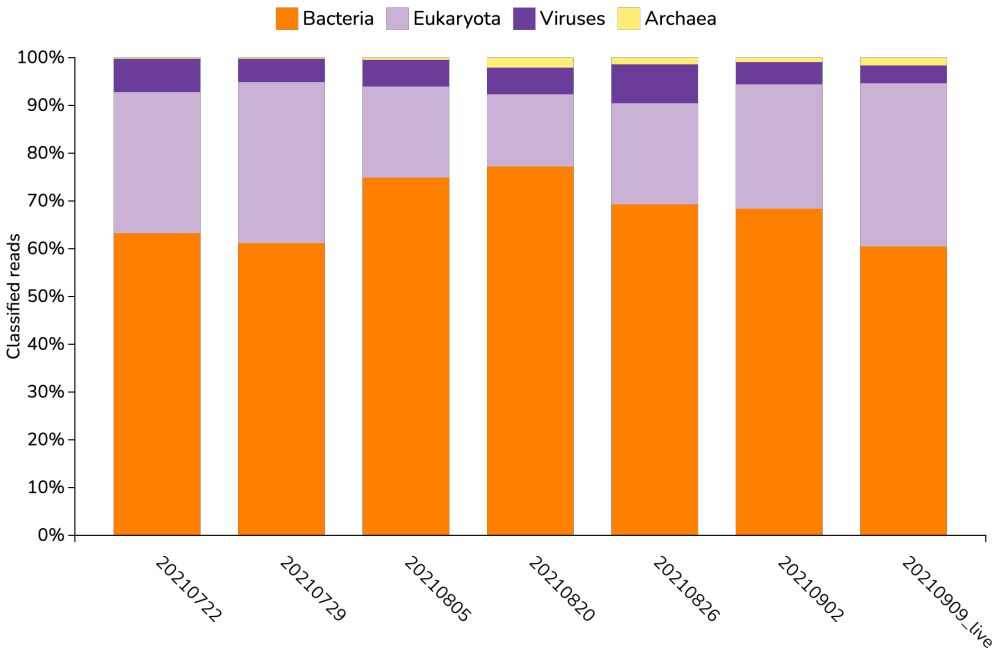


Figure 4.3.12: Stacked bar chart showing superkingdom level matches for each sample. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

As can be seen in figure 4.3.14, the most frequent matches at family level are Pelagibacteraceae, Bathycoccaceae, and Rhodobacteraceae. Pelagibacteraceae are free-living proteobacteria found in marine environments, and Rhodobacteraceae are also proteobacteria and occur in a range of environments although they are commonly found in water. Bathycoccaceae are green algae in the chlorophyta phylym, in the class Mamiellophyceae and the order Mamiellales. Mamiellaceae, also a Mamiellophyte was also found in all samples. Phylum, genus, and species level stacked bar charts can be seen in Appendix H.

Treemaps were produced at genus level for each sample after separating at kingdom level into Eukaryota, Prokaryota, and Viruses, to provide an overview of the taxonomic distribution within each kingdom. These can be seen in figures 4.3.15, 4.3.16, and 4.3.17 with larger individual versions available in Appendices D to F.

Figure D shows that Chlorophyta are by far the most abundant eukaryotic matches, followed by Bacillariophyta and Chordata. *Ostreococcus lucimarinus*, a globally abundant small celled green alga, is the largest genus in all samples excluding 4, where *Micromonas commoda*, a green alga in the Mamiellacaeae family, and

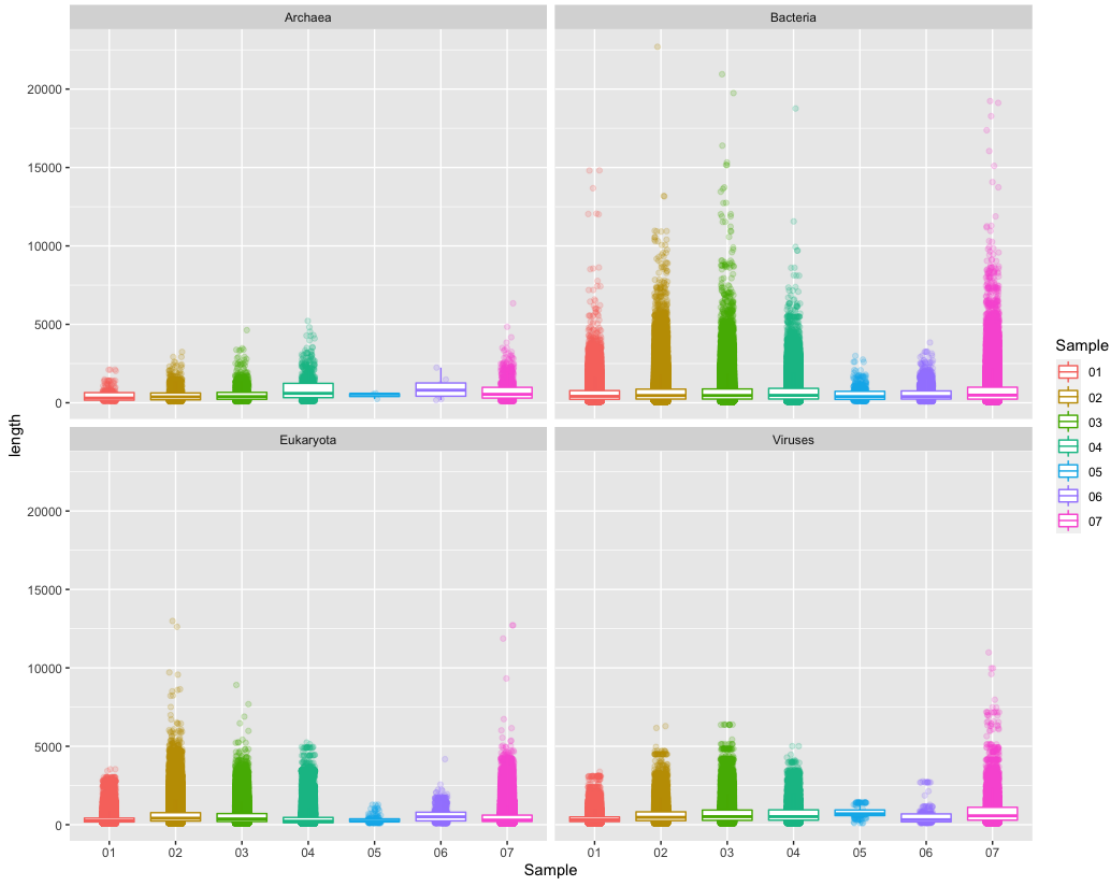


Figure 4.3.13: Combined Box and scatter plots showing length distribution of matches in filtered blast hits. Boxplot shows median, 25th centile, and 75th centile. Scatter shows all hit lengths.

Bathycoccus prasinos which is a green picoalga, were the most abundant. The samples show similar distributions, with increasing differences visible in less abundant taxa.

As can be seen in figure 4.3.16, Proteobacteria and Bacteroidetes are the largest prokaryotic matches at the Group level followed by Cyanobacteria and Actinobacteria in every sample, with the largest genera being *Planktomarina temperata* and Ca. *Pelagibacter*. Candidatus denotes a prokaryotic taxa which have been characterised but have not been cultured and so have not been officially named, this is common in the case of organsims which have been sequenced as part of metagenomic studies, or 16S RNA gene sequencing and have not, or cannot, be grown in the laboratory (Stackebrandt et al. 2002; Rappé et al. 2002). There is clear variation between samples, while the overall picture is constant between them, with the exception of sample 5:20210826 which had very low yield compared to the other samples, and shows correspondingly fewer genera.

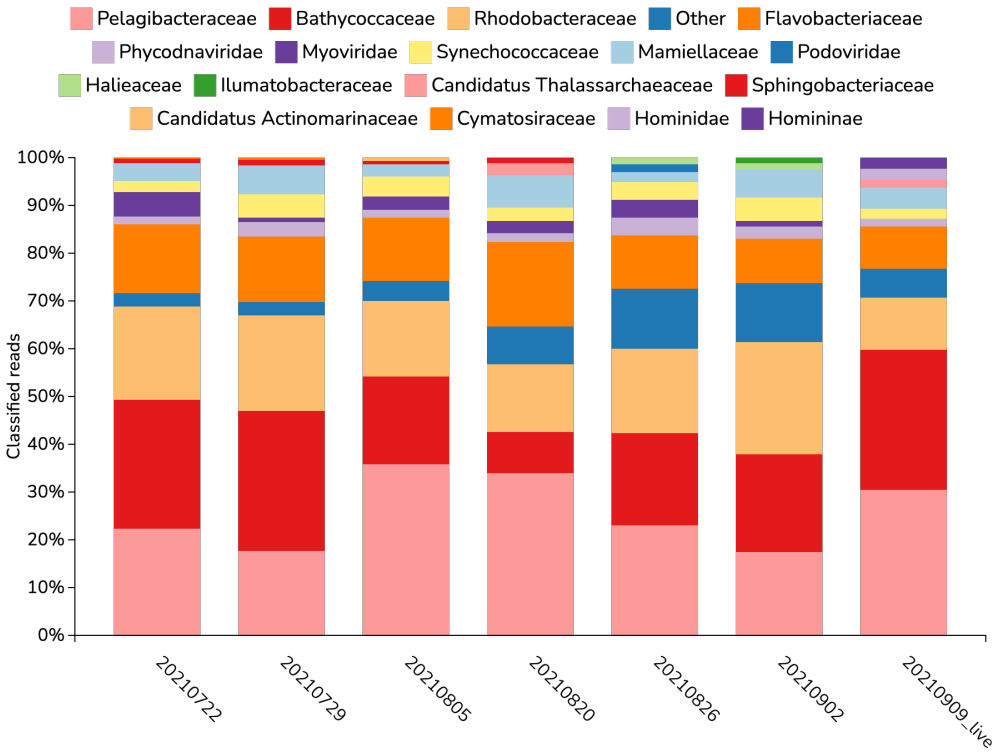


Figure 4.3.14: Stacked bar chart showing family level matches for each sample. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

Figure 4.3.17 shows that the genera found are all part of the Uroviricota and Nucleocytoviricota, with Uroviricota dominating in samples 1, 3, 4, 5, and 7, while Nucleocytoviricota are the most abundant in samples 2 and 6. There is clear variation between samples, both in the most abundant genera and the overall number identified. At the genus and species level, between 7 and 35% of BLAST hits were unassigned to any taxonomic and are shown as other.

There were a total of 326,379 species-level blast hits in the time-course including the live sequencing, with species number identified per sample between 1562 and 3178 for samples 1, 2, 3, 4, and 7, with samples 5 and 6 having much lower species numbers, commensurate with much lower DNA and sequencing yields. Table 4.7 shows the number of species and blast hits per sample. *Thalassiosira* and *Skeletonema* species were found in all of the samples excluding sample 5. Given the low read number for sample 5 non-detection does not necessarily indicate absence. The number of reads in each sample matching each *Thalassiosira* species are shown in Appendix I. Most *Thalassiosira* species had fewer than 100 reads. *Thalassiosira oceanica*, *Thalassiosira pseudonana*, and

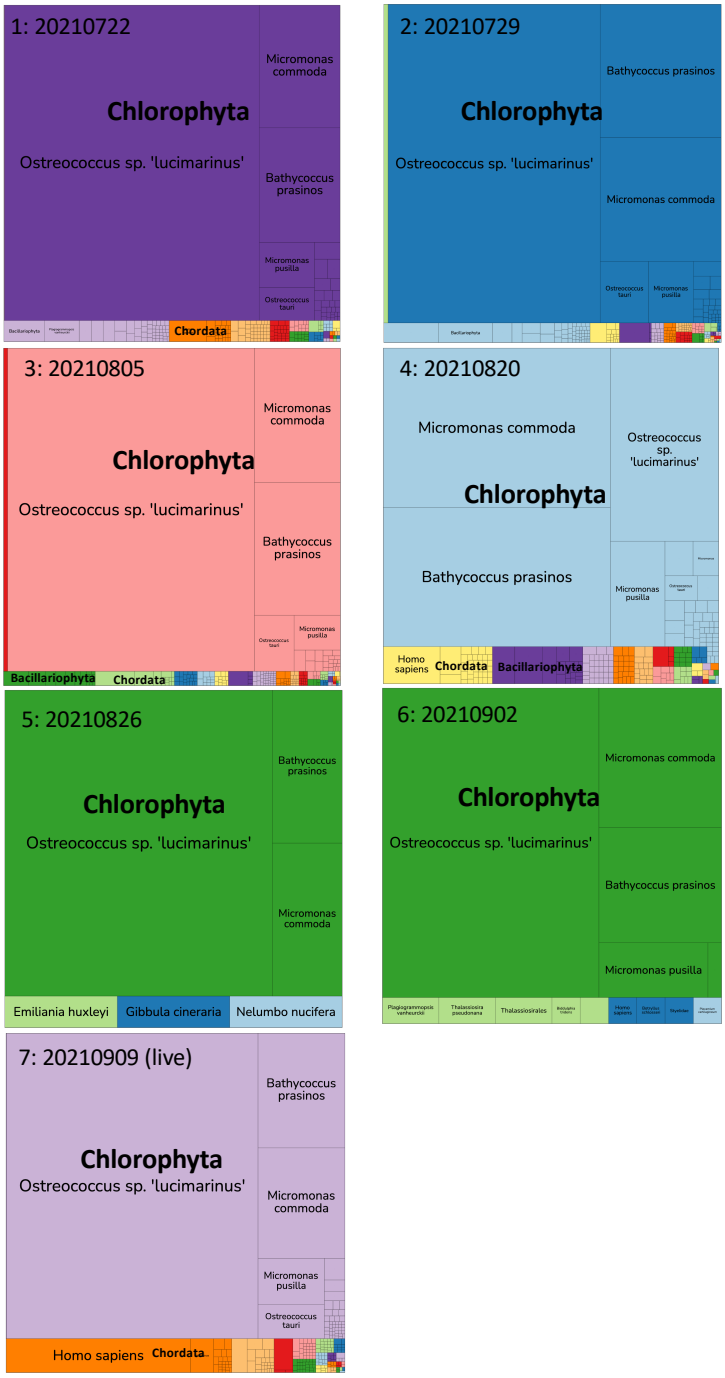


Figure 4.3.15: Treemaps showing genus level matches within Eukaryota for each sample, with group labels added. Sample labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

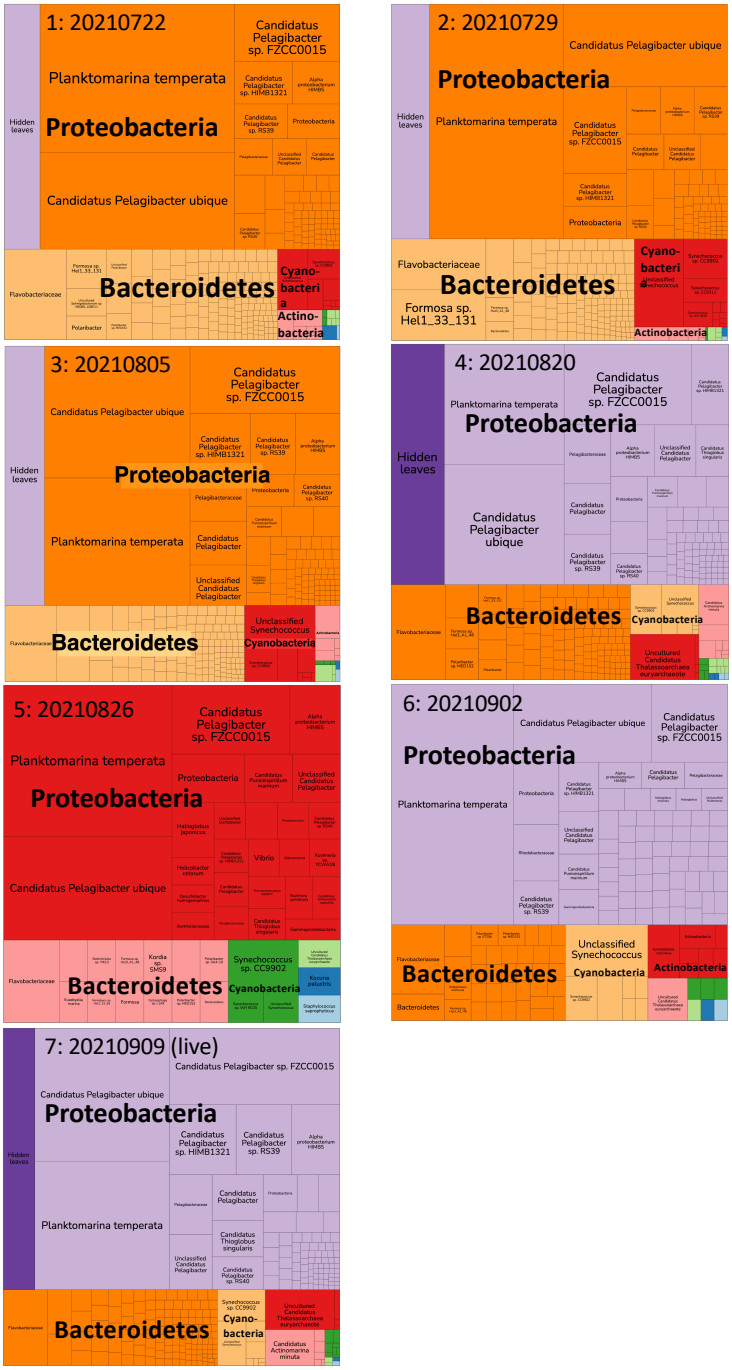


Figure 4.3.16: Treemaps showing genus level matches within Prokaryota for each sample, with group labels added. Sample labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

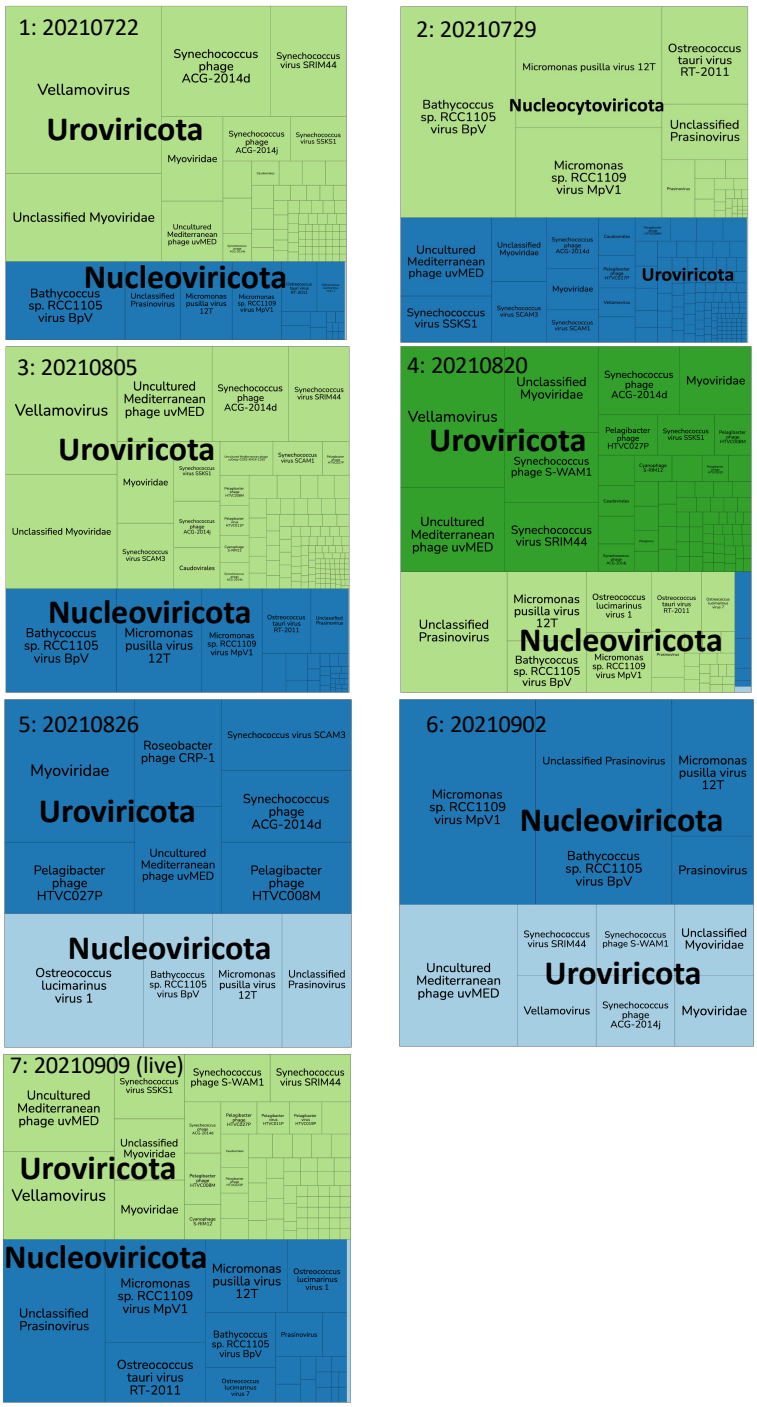


Figure 4.3.17: Treemaps showing genus level within Viruses for each sample, with group labels added. Sample labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

Table 4.7: Number of individual species identified in each sample, and the total number of BLAST-nt matches per sample.

Sample	Species	Hits
1	1562	30321
2	3262	131069
3	3178	80936
4	1905	19806
5	58	137
6	138	581
7	1654	63529

Thalassiosira weissflogii were the only species to have more than 1000 reads across all samples with 1041, 2428, and 1094 reads respectively. The number of reads matching each *Skeletonema* species are shown in Appendix J. Most *Skeletonema* species had fewer than 250 reads across all samples and the only species to have more than 1000 reads across all the samples was *Skeletonema pseudocostatum* with 1277 reads. *Teleaulax amphioexia* was also found at low levels across samples 1-4 and sample 7.

Emiliania huxleyi was found in all samples, mostly at less than 500 reads per sample with more in samples 3, and 7, and over 3000 in sample 2, as can be seen in figure 4.3.18. Most of the reads aligned with *E, huxleyi* were a few hundred base pairs long, with a small number over 1000 bp, mainly in sample 2, which produced most of the *Emiliania huxleyi* classified reads, and sample 7 which produced the next highest number. The combined total of hits was 1.22 Mbp of which 0.76 Mbp was in sample 2.

Vibrio species were identified in all 7 Pier-Seq data samples, particularly samples 2 and 3, as can be seen in Appendix K. There were 16,211 reads altogether which matched to *Vibrio*. The most abundant species with more than 1000 reads were *Vibrio alginolyticus*, *Vibrio cholerae*, and *Vibrio parahaemolyticus*. The read lengths varied from 100 to 1672 bp, although the vast majority were below 200 bp.

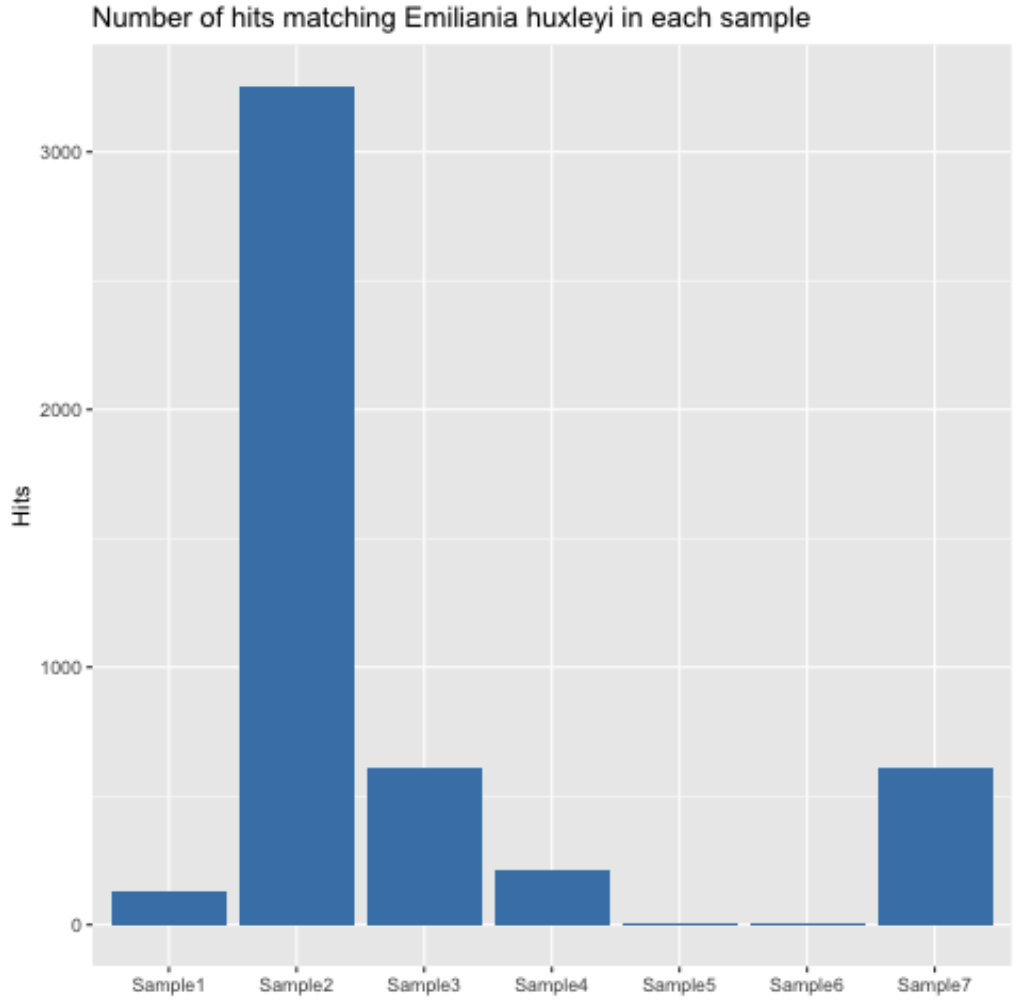


Figure 4.3.18: Bar chart showing the number of reads matching to *Emiliania huxleyi* in each sample.

4.4 Discussion

4.4.1 Improved workflow for nanopore sequencing of ocean microbiomes

It was decided to use a low volume sample of 20 mL for filtration, and to use the CTAB with beads protocol for further analysis of Ship-Seq samples in the lab - see Chapter 3 - and to use the Qiagen PowerSoil protocol for fieldwork. This was based on the high sequencing yields and N50 provided by the CTAB with beads protocol balanced against the time required and toxicity of the reagents which made it less suitable for fieldwork. It was decided that the rapidity, simplicity, and safety of the Qiagen PowerSoil protocol, coupled with its good DNA recovery, sequencing yield, and N50 made it a better choice for the Pier-Seq experiments.

In future, further experiments could be carried out to test other protocols, and adaptations to those already trialed. The taxonomic identification results were similar for all the protocols.

From the investigation into the effect of read number on genus identification, it was established that the read number was unlikely to affect the overview of what was present at the genus level, particularly the most frequent matches, indicating that as long as the samples returned more than 20,000 reads the main conclusions of taxonomic identification were likely to be correct. Future work could include metabarcoding to further verify the results and allow for a comparison between metabarcoding and nanopore sequencing for precision, costs, and the utility of further analysis. The total number of genera found, however, was very closely linked to the number of reads sequenced, indicating that for a full understanding of the composition of organisms with low abundance, large sample sizes are required. As the aim of the Pier-Seq experiment was to get an overview of the organisms present, examining those which are most abundant, and their change over the time-course, this meant that relatively low data volumes could be accepted, allowing for low sampling volumes. For in depth analysis of less abundant organisms higher sample volumes would be required to extract sufficient DNA for high-yield sequencing. These experiments were based on the analysis of a 13 Gbp sequencing dataset from a single sample, which has a flattening although not entirely plateaued rarefaction curve at the genus level which indicates that sequencing yields greater than 13 Gbp may be required for full taxonomic identification of ocean samples.

The read length experiment was less straightforward to draw conclusions from, not least because removal of short reads can mean that information is lost as organisms with short genomes, or which are particularly prone to shearing in DNA extraction, which is a particular problem where GC content is high as in many diatoms and other phytoplankton, are disproportionately removed. As a result, it was decided not to use a short read elimination kit which removes DNA <25000 bp so as to avoid information loss, and the use of an extraction method which did not return extremely HMW DNA was accepted. The experiment showed a limited effect on the big picture of genera present, affecting mainly those with lower numbers of matches.

The MARTi filtering step is stricter than the filtering used for MEGAN analysis, and for Pier-Seq was set to show the top 10 genera per sample. With this limit removed, however, the total genera found ranged from 39 for sample 6 with only 989 reads to 1418 for sample 2 with over 570,000 reads. None of the samples had reached saturation at genus level, indicating that the genera found in the MARTi analysis represent the most abundant and strongly evidenced matches

and are unlikely to be affected by read number or read length, as less abundant genera are unlikely to be represented. More in depth analysis, looking into less abundant organisms, or to determine differences between similar species or strains - for example, to identify the genetic basis for differences in ability to withstand infection or colonise niches as seen between the haploid and diploid *E. huxleyi* strains in 2 - would likely be benefited by higher read lengths. From these experiments, however, it clear that read length is not the main driver for genus-level identification, meaning that the protocol could be used by citizen scientists to give a big picture overview of the microbiome, providing data which could be used to aid researchers in planning more in depth sampling.

Improvements to the DNA extraction and sequencing workflow as covered in figure 4.3.9 have resulted in a protocol which requires low volumes of sea water input, and less than 1 hour from sampling to first results. It does not rely on transport and use of toxic reagents, requires only a small number of reagents to be kept frozen, and uses flowcells which can be transported at ambient temperature. DNA sequencing and analysis is carried out using one laptop and one sequencing machine. Compared to the Ship-Seq protocol, which required up to 100 l of water to be sampled and took several hours from sampling to first results due to slow filtration and long incubation times for DNA extraction; relied on the transport and use of phenol-chloroform - a highly toxic chemical which poses risk to human and animal life and requires extensive safety arrangements; required a large proportion of reagents and flowcells to be kept chilled or frozen; and needed 2 laptops to be transported for sequencing and analysis, the process has been streamlined and simplified. This increases its utility for citizen science and outreach, as well as making it easier to implement as part of a roster of standard experiments on research cruises.

Focussing on taxonomic identification allowed the acceptance of a lower quality of DNA extraction with lower molecular weight. Identification of species requires less HMW data than genome assembly or gene analsis. Coupled with improved accuracy in basecallers, especially the high and super high accuracy base callers available in the lab reduced our need for HMW DNA and the associated extraction techniques. This allowed the removal of toxic chemicals from the DNA extraction protocol, coupled with the use bead beating to increase DNA yield at a slightly lower molecular weight. This allowed for a significantly reduced volume of water to be filtered, from tens of litres in the 2019 cruise down to 20 mL per filter in 2022. The Southern Ocean is particularly clear so more than 20 mL would likely be required there but we would still expect to require significantly reduced volumes. This would make it easier to fit sequencing experiments into a busy CTD schedule on board a ship, and take far less time to collect manually from land and would

also be filtered and put on ice more quickly - in a matter of 2-3 minutes from collection compared to up to 8 hours, preserving DNA quality more effectively. It also reduces storage requirements in the freezers on board research vessels or in ice boxes for sampling from land. It would take only a few minutes to filter three samples, one to be extracted and sequenced *in situ* with bead beating and rapid ONT extraction kits, and two to be stored on ice and sequenced back in the lab with techniques more suited to extracting HMW DNA or whatever technique was indicated by the results of the *in situ* experiment. This has particular benefits for use with citizen science sampling experiments where sequencing could be carried out at a range of times or locations at low depth, giving a snapshot, or for outreach and engagement activities.

Reduced sampling volume and time requirements mean that DNA sequencing can be used in a wider range of experiments and be used in more innovative ways. A low sample volume requirement opens doors to experiments in multiple locations, for example within an hour of high tide multiple locations could be sampled from a single bucket or niskin bottle at each location. This would allow for comparisons between locations across a time series by one research team. Low sample volumes also make sampling easier in bad weather, and for small teams. Shorter filtration periods reduce the time between sampling and DNA extraction or storage, improving the quality of the sample. A shorter time period from sample to results could allow for the course of a research ship to be directed by sequencing data when pin pointing optimal sampling locations. Combined with read-until sequencing, where a sample is DNA sequenced until a certain organism is found, using rarefaction curves to determine a cut off point, *in situ* sequencing could be used to determine whether larger scale sampling is indicated at a certain time and location.

The reduction in volume required from samples was achieved by focussing on taxonomic identification as opposed to in depth analysis and assembly. This was decided based on the Ship-Seq pilot study where, after sequencing and analysis, it became clear that this is what relatively low yield, short run, *in situ* sequencing experiments are best suited for. For in depth genomic analysis, high-yield, long run lab-based DNA sequencing is a better fit. In these instances an increased sample volume may prove beneficial, alongside high-throughput DNA sequencing technology, such as the ONT PromethION or Illumina platforms. These could provide the data required for the production of metagenomic genome assemblies, or functional annotation to investigate the genetic basis for differences within and between sampled populations. Future studies could investigate how 20 ml sample volume compares to increased volumes, to determine whether increases in sample volume, up to 100 ml using syringe filters and up to 10 litres with

pump-aided filtration, depending on sample location suitability, would provide improved resolution while remaining sufficiently fast and simple for citizen science projects.

A bead-beating stage was added to the protocol because it allows the rapid increase of DNA extraction yield by pulverising cells. This increases the volume of DNA available during the extraction process but can also result in some damage to DNA strands, reducing the molecular weight. As such, bead-beating can result in some some loss of information, especially when using Nanopore sequencing which has traditionally had lower accuracy rates and relied on long DNA strands - although increasing accuracy in Nanopore basecalling is rapidly improving the results achieved on short strand sequencing. In the comparison experiment, however, 5 minutes of bead-beating did not appear to have an adverse effect on the molecular weight as the CTAB and bead-beating protocol returning a higher N50 than the CTAB without bead-beating protocol. For experiments where HMW DNA is required, especially those carried out in a laboratory as opposed to in the field, a protocol such as the CTAB with phenol-chloroform may give better results. If high sequencing yields are required, increased sample volumes for filtration may also be beneficial.

A further consideration in this new workflow was the development of a protocol which does not rely on toxic chemicals. This was desirable for several reasons. First, there is training required for anyone working with, or in the same room as phenol-chloroform. International transport requires extensive paperwork and safety arrangements and ensuring that regulations such as temperature stability and protection can be met at all stages of the journey. The use of such an extraction protocol limits the number of people who can carry out sequencing experiments to those who have been specifically trained, reducing the number of research cruises on which DNA sequencing is practical. A further consideration was the length of time the DNA protocol takes. The CTAB extraction protocol used for Ship-Seq involved an incubation step of 4 hours. This limited its use in the field, especially if working without a power source. The Qiagen PowerSoil protocol can be performed rapidly, taking no more than 1 hour including the rapid Nanopore library preparation protocol and requires no toxic chemicals or long incubation steps.

The alterations made by ONT to their flow cells, making them more robust and able to withstand ambient storage and transport as opposed to requiring constant refrigeration simplifies the process of international transportation for fieldwork, as well as widening access to countries where cold chain transport is difficult to guarantee. Transportation of the equipment required for sequencing has been made easier with the introduction of the MinION Mk1C which is a portable

sequencing machine with built in analysis software and a touch screen. This means that all that is required for *in situ* sequencing is the Mk1C and a laptop for real time organism identification. During the Pier-Seq experiments we also tested a portable power pack and solar power source which allowed us to run the experiment for several hours without an electricity supply.

The use of MARTi offers a simple, user-friendly way to observe the results of sequencing experiments in real time. A successor to NanoOK-RT which was used in Chapter 3, it is both more powerful and easier for researchers to use. For the live sequencing experiment a filtered BLAST database was used which included only taxa containing sea-living organisms. This reduced the time taken to search the database. While this was beneficial in making it possible to carry out real time organism identification in the field, it has the potential to affect the organisms found and it would be sensible to check against the full database afterwards.

These alterations to the workflow combine to give a protocol which could be used by anyone, anywhere. Many research cruises have scientists on board who carry out general experiments and this new protocol could be used by them with a small amount of training. It would take up a small amount of their time, which is split between a range of activities, and if carried out on a large number of research cruises as part of the standard experimental roster it could provide a broad and extremely useful resource. Increasing the number of experiments carried out would give us a better overview of the current microbial populations and allow us to monitor them over time. One of the key benefits of portable DNA sequencing is the democratisation of science. Simpler, easier to use workflows, which can be carried out end-to-end by researchers rather than relying on sending off samples, will increase the number of researchers who can perform DNA sequencing in the lab or in the field for non-model organisms, giving us more information about the ecosystems we rely on.

4.4.2 Analysis

None of the samples had a plateau on the rarefaction curves, indicating that 550,000 reads was insufficient to cover all taxonomic identification present. More sequencing data would be beneficial for determining which organisms are present, absence from these results cannot be concluded to mean that the organism is not present in the sample location. The species level rarefaction curve shows no levelling off, while the family level is beginning to flatten, indicating that there were fewer families still to be sequenced than species. This is to be expected,

as there is a greater number of lower taxonomic ranks than higher, and species saturation relies on high DNA volume, with a riverine study finding that species level saturation was not achieved at 10 Gbp of nanopore MinION sequencing (Reddington et al. 2020).

The read numbers for each sample were ranged from 989 to >500,000. Samples 5 and 6 produced very little sequencing data, 989 and 3454 reads respectively. They showed broadly similar distributions at all taxonomic levels to the other samples which ranged from 120,000 to 511,000 reads but conclusions based on those samples should be treated with caution in the absence of more sequencing data. 100,000 reads is a widely used threshold for metagenomic sequencing analyses, and this was cleared for 5 of the 7 samples. As was seen from the comparison between nanopore and Illumina data in figures 3.3.11 and 3.3.10, identification of the most abundant taxa does not appear to be affected by low saturation levels. From the investigation into the effects of read number on taxonomic identification, it is unlikely that the overview of genus presence is affected by low read number in samples 1-4 and sample 7, although the composition of organisms with a lower abundance will likely have been missed. Sample 2 and 3 have far more reads than any other samples, with 571,123 and 426,468 reads respectively and they also have the highest number of species matches. In-depth sequencing of samples, using an increased sampling volume and several filters, with a yield of more than 7 Gbp could be used to increase the likelihood of capturing most or all of the organisms present at a lower taxonomic level. Based on these results, the Pier-Seq samples can give information on the presence of an organism but it is not possible to draw conclusions as to its absence from this data.

The mean read length for each sample ranged from 724 to 1127 bp, with 20-50% of each sample containing reads longer than 1000 bp. The read length experiment was based on samples of 20,000 reads which again excludes samples 5 and 6 from comparison, and further sequencing would be advisable before carrying out more analysis on those samples. Sample 1 has a mean read length of 724 and 20% of reads greater than 1000 bp while samples 2-4 and sample 7 had mean read lengths of 1027-1197 with >30% of reads greater than 1000 bp. This indicates that while the overview of genus presence is likely to be unaffected, differences in proportions of genera found between sample 1 and samples 2-4 and 7 may be due to differences in read length.

Filtration removed the vast majority of hits in the MARTi dataset. In sample 1, for example, 74% of reads were unclassified. This increases the likelihood that matches shown in the MARTi data are genuinely present. MEGAN analysis, as was used for Ship-Seq and the read length and read number experiments had no

minimum read length match, meaning results may be less reliable but that there is more potential to match short reads of less abundant organisms, provided they are above the threshold of 0.1%. For live sequencing, and where am overview of the big picture is desired, MARTi and reasonably strict criteria for determining presence is ideal. When carrying out more in depth analysis, and where there is the opportunity to consider the results and whether they are plausible, it may be beneficial to use the more lax criteria as used in the MEGAN analysis to avoid information loss.

Overall, around 60% of filtered BLAST-nt hits were for bacteria, with 30% for eukaryotes, 7% for viruses and the remaining 3% archaea. In the unfiltered BLAST-nt hits, bacteria make up 89% of hits and eukaryotes around 5%, which indicates that shorter reads have been filtered out, particularly those matching to bacterial vectors such as *E. coli*.

Read length distribution was compared between samples and superkingdoms. There were no statistically significant differences in length between superkingdoms or samples. The longest reads across all samples were classified as bacteria, followed by eukaryote, viruses and archaea. As reads classified as bacteria constitute the majority in every sample it is likely that the differences in read length are due to increased read numbers being analysed. This makes sense given that short DNA strands are sequenced preferentially in nanopore sequencing, so the lengths are likely to increase with yield. A size selection step could be used to remove reads below a certain threshold if desired, although the results of the read length fractionation experiment indicated that read length did not have an appreciable effect on the overall taxonomic classification. The length of reads sequenced in the Pier-Seq experiment were not sufficiently long for strands longer than a bacterial genome to be sequenced.

At each taxonomic level there was variation between the samples in the proportions of different taxa present. There were broad similarities in the identifications between each sample for the most abundant taxa which make up approximately 70-80% of the total classified reads but the taxa with fewer reads were different between samples at each level. Given the small sample size, and the limited DNA sequencing data available for these groups of reads, it is difficult to attribute these differences to genuine population changes over time, as opposed to sampling error. At the species level, a significant proportion of the reads are unclassified, or classified as uncultured, the genus level is therefore more useful for determining what is present. Increased sampling depth could help to alleviate this problem. Increasing the number of marine microorganisms present in the BLAST-nt database would also be useful, as sequencing depth cannot help to match against species which are not present in the database.

There is still relatively little genomic data available for marine microorganisms and phytoplankton, especially full assemblies.

The main aim of this experiment was to identify eukaryotic phytoplankton, with kingdom-separated analysis identifying the taxonomic distrubution within Eukaryota for each sample. The main eukaryotic phylum identified in the Pier-Seq samples was Chlorophyta. The distribution of Chlorophyta is poorly understood, partly because they are too small to easily distinguish accurately with microscopy, and DNA sequencing of ocean samples is giving an improved understanding of their ecology and distribution. Three of the most abundant genera shown in 4.3.15, Bathycoccus, Micromonas, and Ostreococcus, are globally distributed and known to be numerically important in coastal, nutrient-rich waters (Moreau et al. 2012; Vannier et al. 2016). The Ocean Sampling Day (OSD) project (Tragin and Vaulot 2018) which sampled surface waters for sequencing at coastal locations around the globe found that Chlorophyta made up 29% of the reads from photosynthetic organisms globally and around 6% of the photosynthetic reads in the North Sea. They were found to be most abundant close to the coast, although there was a reduced contribution from Chlorophyta from the equator to 10% of photosynthetic reads at 60 °N. This would indicate that at 52 °N, Cromer would not be expected to have a large contribution from Chlorophyta, which does not agree with the Pier-Seq findings of Chlorophyta making up up to 27% of all classified reads. In the Pier-Seq data set Chlorophyta are almost all made up of Mamiellales at the order level with Bathycoccaceae and Mamiellacae the only Chlorophyte families which can be seen at the family level. This agrees with the OSD finding that the most commonly found Chlorophyta in the North Sea were Mamiellacae, which in one site of the coast of Belgium accounted for 99% of Chlorophyta reads. The Mamiellaceae reads are made up of Ostreococcus (70% of Mamiellophyceae reads), Bathycoccus (16% of Mamiellophyceae reads), and Micromonas (13% Mamiellophyceae reads). This compares to 53%, 31%, and 14% respectively found by the OSD project.

Outside of Chlorophyta, bacteria and viruses dominated the outputs in Pier-Seq, as they did in Ship-Seq. Other eukaryotic phytoplankton identified include several species of diatoms *Thalassiosira* and *Skeletonema*, cryptophyte *Teleaulax amphioexia*, and coccolithohore *Emiliania*. The majority of these reads for all species came from sample 2. Sample 2 had the highest sequencing yield and the most reads of all of the Pier-Seq samples, along with one of the highest mean read length and the most reads over 1000 and 10,000 base pairs. This indicates that finding diatoms, coccolithophores, and cryptophytes is more likely when the sequencing yield is higher. The length of *Emiliania huxleyi* reads was

examined and the mean length was less than 250 bp across all samples, there were around 4500 reads matching *E. huxleyi* altogether, with the longest read just over 2500 bp. This was insufficient for further analysis to attempt genome assembly or alignment to identify the strain and compare against the *E. huxleyi* found in Chapter 3 or assembled in Chapter 2. In a recent study, 140 MAGs of Arctic phytoplankton were produced from 679 Gbp high-throughput Illumina sequencing data (Duncan et al. 2020) and it may be that high-throughput PacBio or nanopore PromethION sequencing are better suited to MAG production while still harnessing the benefits of long-read sequencing.

The bacteria and viruses found in Pier-Seq are generally temperate marine organisms as would be expected from the sampling location. Previous studies have found that (photo)heterotrophic bacterioplankton communities near the sea surface are largely populated by Proteobacteria, *Sphingobacteria*, and *Flavobacteria* (Giebel et al. 2010) which is in keeping with the Pier-Seq findings. At the genus level, '*Candidatus* Pelagibacter' was most abundant, with all of those reads resolving to '*Candidatus* Pelagibacter ubique' (Also known as SAR 11) at species level. This was also found in Ship-Seq, which is in accordance with its ubiquity in global oceans (Rappé et al. 2002).

The presence of *Vibrio* is not necessarily a cause for concern, without benchmarking it is unknown whether the amount found is unusual, or sufficient to cause disease and it is impossible to say whether the *Vibrio* bacteria were alive when they were sequenced. This experiment does indicate, however, that it would be possible to detect *Vibrio* and other pathogenic or harmful microorganisms. A PMA treatment could be used to determine microbial viability (Legrand et al. 2021) and the presence could be quantified using spike-in experiments, allowing researchers to determine the potential risks. The most abundant *Vibrio* species were *V. alginolyticus*, *V. cholerae*, and *V. parahaemolytica* all of which cause disease in humans. Given the short read lengths it is possible that these are random or incorrect alignments. *Vibrio* sequences are present in high numbers in genomics databases because they are disproportionately heavily studied and represented in genomic databases due to their effects on human health, which could contribute to the large numbers classified. Increased sequencing depth and read length, along with the use of *Vibrio* specific quantitative PCR would help to determine the true presence of pathogenic *Vibrio* species.

4.4.3 Future work

Future work would include a sampling and sequencing experiment with volunteer citizen scientists. Live sequencing experiments carried out by different groups around the country could provide a snapshot of coastal marine microbial diversity. Alternatively, a group of volunteers could sample and sequence once a week over a timecourse to capture changing diversity over time. Outreach events associated with these experiments could help to engage children and young people, as well as the wider public, widening understanding of ocean microbes.

As discussed above, it would be interesting to explore other protocols and sampling volumes, to determine the balance between quick and simple, and optimising sequencing depth.

ONT now produce lyophilised kits which can be stored at ambient temperature for up to 30 days and could be used when travelling without access to freezers. The cold-chain is a barrier to the use of nanopore sequencing technology in the field. New high-capture kits are being produced which will maximise the data produced from low volume inputs. This would allow for the continued use of low sample volumes. For *in situ* experiments, especially those being used to give a quick overview of microbial populations, smaller, cheaper Flongle flowcells capable of sequencing up to 2.8 Gbp can be used. This would reduce the cost of running multiple sequencing experiments, either with multiple groups at different locations, or at the same location over a time course.

The forthcoming ONT SmidgION could also be used to sequence at low depth in conjunction with PCR or qPCR, for example to identify and quantify harmful microbes, or a confirm the presence of a species of interest to researchers prior to deeper sequencing with more costly, higher yield flowcells. Another forthcoming ONT technology, a successor to the Mk1C called the Mk1D, which can run MinION flowcells from a tablet via an app could further increase the portability and user-friendliness of higher yield nanopore sequencing for citizen science and outreach. Finally, the ONT VolTRAX V2, which automates library preparation could be used to make the process of preparing the extracted metagenomic DNA for sequencing more portable, as it requires no extra laboratory equipment, as well as quicker, and more consistent.

The use of satellite imagery to assess chlorophyll levels and likely phytoplankton abundance before sequencing was trialled, involving the analysis of surface temperature, enhanced ocean colour, and chlorophyll were analysed. The intention was that this would allow for targeted sequencing during periods of interest such as in the event of a phytoplankton bloom. The length of the

sampling period, during which the weather was very stable, meant that there was little variation in the satellite outputs, however, and there were no blooms. A longer coastal time-course spanning several months or a year, perhaps at more than one location could be used to investigate the microbiome at the Norfolk Coast and establish the use of satellite imaging as a tool for choosing sampling time and location. Samples could also be taken for nutrient analysis. Sampling over a year would give a picture of population change as temperatures and weather conditions change. Sampling in multiple locations would allow for comparison of populations between sites, and correlation to temperature, chlorophyll, and nutrient availability metadata.

4.4.4 Summary and conclusion

An improved workflow has been produced for nanopore sequencing of ocean microbiomes. This streamlined workflow could be used by researchers with a little training, potentially increasing the number of sequencing experiments which can be carried out as it could be carried out without a DNA sequencing specialist present. The improvements have reduced the time and sample volume requirements, removed toxic reagents, and reduced the reliance on cold-chain transport.

Optimisation of the workflow was achieved through testing multiple alternative DNA extraction methods, and developing increased portability through the use of the RapidPrep2 for bead beating and the MinION Mk1C which has built in DNA sequencing and basecalling capabilities to carry out the live sequencing experiment. The workflow was tested at Cromer Pier on the Norfolk Coast, in accordance with lockdown restrictions at the time, over a series of several weeks with samples collected for multiplex sequencing and one live sequencing experiment carried out on the pier.

The workflow produced DNA recovery and sequencing yields in line with expectations, and sufficient for taxonomic identification of the majority of reads, although increased sequencing depth would increase the likelihood of capturing low-abundance organisms. This may be of particular relevance where the organisms of interest are likely to be less abundant, as is the case with diatoms and coccolithophores. *Thalassiosira*, *Skeletonema*, and *Emiliania* species were identified alongside Chlorophyta and'*Candidatus* Pelagibacter'. Taxonomic identification was in line with expectations based on previous studies of microbe populations in the North Sea. Pathogenic bacteria were identified, although more investigation would be required to draw conclusions as to the risks they pose.

Further work would include implementing the improved workflow on a research cruise. This would allow for testing the workflow further in the field and would provide a large dataset of polar ocean samples for anlysis, following on from Ship-Seq. A longer coastal study following on from Pier-Seq would give further opportunity for establishing the new workflow, and to determine whether monitoring satellite data would be a useful tool for researchers planning sampling.

In order to improve our understanding of ocean microbiomes, and to model the potential impacts of climate change, we need more information about the composition, distribition and flux in communities. This research has shown that nanopore sequencing from low sample volume, low DNA input from rapid DNA extraction can provide taxonomic classification of ocean microbiomes. This could allow researchers to build up a baseline picture of ocean microbial communities and monitor population change over time. The portability and real-time analysis capabilities mean that this would be particularly useful for monitoring ocean microbes onboard research ships. This is especially true for polar ocean microbial communities, where the sample to laboratory analysis timeline can be in the order of months and involve long transits. The low cost and simplified protocol which can implemented by non-specialists mean that *in situ* nanopore sequencing with real-time analysis could be widely used across ocean research projects, vastly increasing the potential for population monitoring.

References

Acharya, K., S. Khanal, K. Pantha, N. Amatya, R. J. Davenport, and D. Werner (2019). "A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality". In: *Scientific reports* 9.1, pp. 1–11. DOI: 10.25405/data.ncl.9693533.

Barberán, A., T. J. Hammer, A. A. Madden, and N. Fierer (2016). "Microbes should be central to ecological education and outreach". In: *Journal of microbiology & biology education* 17.1, pp. 23–28.

Bresnan, E., C. Baker-Austin, C. Campos, K. Davidson, M. Edwards, A. Hall, A. McKinney, and A. Turner (2020). "Impacts of climate change on human health, HABs and bathing waters, relevant to the coastal and marine environment around the UK". In: *MCCIP Science Review 2020*, pp. 521–545. DOI: 10.14465/2020.arc22.hhe.

Capuzzo, E., C. P. Lynam, J. Barry, D. Stephens, R. M. Forster, N. Greenwood, A. McQuatters-Gollop, T. Silva, S. M. van Leeuwen, and G. H. Engelhard (2018). "A decline in primary production in the North Sea over 25 years, associated with reductions in zooplankton abundance and fish stock recruitment". In: *Global change biology* 24.1, e352–e364. DOI: 10.1111/gcb.13916.

Cerro-Gálvez, E., J. Dachs, D. Lundin, M.-C. Fernández-Pinos, M. Sebastián, and M. Vila-Costa (2021). "Responses of coastal marine microbiomes exposed to anthropogenic dissolved organic carbon". In: *Environmental Science & Technology* 55.14, pp. 9609–9621.

Dassow, P. von, H. Ogata, I. Probert, P. Wincker, C. Da Silva, S. Audic, J.-M. Claverie, and C. de Vargas (Oct. 2009). "Transcriptome analysis of functional differentiation between haploid and diploid cells of Emiliania huxleyi, a globally significant photosynthetic calcifying cell". In: *Genome Biology* 10.10, R114. ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-10-r114.

Duncan, A., K. Barry, C. Daum, E. Eloe-Fadrosh, S. Roux, S. G. Tringe, K. Schmidt, K. U. Valentin, N. Varghese, I. V. Grigoriev, R. Leggett, V. Moulton, and T. Mock (2020). *Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle*. DOI: 10.1101/2020.06.16.154583.

Giebel, H.-A., D. Kalhoefer, A. Lemke, S. Thole, R. Gahl-Janssen, M. Simon, and T. Brinkhoff (2010). "Distribution of Roseobacter Rca and Sar11 Lineages in the North Sea and Characteristics of an Abundant Rca Isolate". In: *The ISME Journal* 5.1, pp. 8–19. DOI: 10.1038/ismej.2010.87.

Hamner, S., B. L. Brown, N. A. Hasan, M. J. Franklin, J. Doyle, M. J. Eggers, R. R. Colwell, and T. E. Ford (2019). "Metagenomic profiling of microbial pathogens in the Little Bighorn River, Montana". In: *International journal of environmental research and public health* 16.7, p. 1097. DOI: 10.3390/ijerph16071097.

Heavens, D., D. Chooneea, M. Giolai, P. Cuber, P. Aanstad, S. Martin, M. Alston, R. Misra, M. D. Clark, and R. M. Leggett (Oct. 2021). "How low can you go? Driving down the DNA input requirements for nanopore sequencing". In: DOI: 10.1101/2021.10.15.464554.

Holt, J., C. Schrum, H. Cannaby, U. Daewel, I. Allen, Y. Artioli, L. Bopp, M. Butenschon, B. A. Fach, J. Harle, et al. (2016). "Potential impacts of climate change on the primary production of regional seas: a comparative analysis of five European seas". In: *Progress in Oceanography* 140, pp. 91–115. DOI: 10.1016/j.pocean.2015.11.004.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster (Feb. 2007). "MEGAN analysis of metagenomic data". In: *Genome Research* 17.3, pp. 377–386. ISSN: 1088-9051. DOI: 10.1101/gr.5969107.

Karlson, B., P. Andersen, L. Arneborg, A. Cembella, W. Eikrem, U. John, J. J. West, K. Klemm, J. Kobos, S. Lehtinen, et al. (2021). "Harmful algal blooms and their effects in coastal seas of Northern Europe". In: *Harmful Algae* 102, p. 101989.

Leggett, R. M., C. Alcon-Giner, D. Heavens, S. Caim, T. C. Brook, M. Kujawska, S. Martin, L. Hoyles, P. Clarke, L. J. Hall, and M. D. Clark (2018). "Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics". In: *bioRxiv*. DOI: 10.1101/180406. eprint: https://www.biorxiv.org/content/early/2018/10/12/180406.full.pdf.

Legrand, T., M. Wos-Oxley, J. Wynne, L. Weyrich, and A. Oxley (2021). "Dead Or Alive: Microbial Viability Treatment Reveals Both Active and Inactive Bacterial Constituents in the Fish Gut Microbiota". In: *Journal of Applied Microbiology* 131.5, pp. 2528–2538. DOI: 10.1111/jam.15113.

Lewin, H. A., G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, and G. Zhang (2018). "Earth Biogenome Project: Sequencing Life for the Future of Life". In: *Proceedings of the National Academy of Sciences* 115.17, pp. 4325–4333. DOI: 10.1073/pnas.1720115115.

Meyer, R., M. Ramos, M. Lin, T. Schweizer, Z. Gold, D. Ramos, S. Shirazi, G. Kandlikar, W. Kwan, E. Curd, et al. (2021). "The CALeDNA program: Citizen scientists and researchers inventory California's biodiversity". In: *California Agriculture* 75.1, pp. 20–32.

Moreau, H., B. Verhelst, A. Couloux, E. Derelle, S. Rombauts, N. Grimsley, M. Van Bel, J. Poulain, M. Katinka, M. F. Hohmann-Marriott, et al. (2012). "Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage". In: *Genome Biology* 13, pp. 1–16.

Pauly, D., V. Christensen, S. Guénette, T. J. Pitcher, U. R. Sumaila, C. J. Walters, R. Watson, and D. Zeller (2002). "Towards sustainability in world fisheries". In: *Nature* 418.6898, pp. 689–695. DOI: 10.1038/nature01017.

Rappé, M. S., S. A. Connon, K. L. Vergin, and S. J. Giovannoni (2002). "Cultivation of the Ubiquitous Sar11 Marine Bacterioplankton Clade". In: *Nature* 418.6898, pp. 630–633. DOI: 10.1038/nature00917.

Reddington, K., D. Eccles, J. O'Grady, D. M. Drown, L. H. Hansen, T. K. Nielsen, A.-L. Ducluzeau, R. M. Leggett, D. Heavens, N. Peel, et al. (2020). "Metagenomic analysis of planktonic riverine microbial consortia using nanopore sequencing reveals insight into river microbe taxonomy and function". In: *GigaScience* 9.6, giaa053.

Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kämpfer, M. C. Maiden, X. Nesme, R. Rosselló-Mora, J. Swings, H. G. Trüper, et al. (2002). "Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology." In: *International journal of systematic and evolutionary microbiology* 52.3, pp. 1043–1047.

Tragin, M. and D. Vaulot (2018). "Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset". In: *Scientific Reports* 8.1, pp. 1–12. DOI: 10.1038/s41598-018-32338-w.

Vannier, T., J. Leconte, Y. Seeleuthner, S. Mondy, E. Pelletier, J.-M. Aury, C. de Vargas, M. Sieracki, D. Iudicone, D. Vaulot, et al. (2016). "Survey of the green picoalga Bathycoccus genomes in the global ocean". In: *Scientific reports* 6.1, p. 37900.

Western Channel Observatory 2022 (2022).

Zan, R., A. Blackburn, J. Plaimart, K. Acharya, C. Walsh, R. Stirling, C. G. Kilsby, and D. Werner (2023). "Environmental DNA clarifies impacts of combined sewer overflows on the bacteriology of an urban river and resulting risks to public health". In: *Science of The Total Environment* 889, p. 164282.

5

Discussion

5.1 Developments and advances in nanopore sequencing of ocean microbiomes

Nanopore sequencing is a new and rapidly advancing field. The first pilot release of nanopore sequencing devices, the Oxford Nanopore Technologies (ONT) MinION Access Programme (MAP), began in 2014. This revolutionised DNA and RNA sequencing, democratising the field to allow researchers to undertake their own DNA sequencing experiments relatively cheaply, costing around £600 per flowcell. This allowed a move away from DNA sequencing of established model organisms through specialised genomics laboratories, giving individual lab groups the opportunity to perform DNA sequencing of any organism or biome of interest. The portability of the MinION further altered the landscape of genomics-based research, as genomes and biomes could be investigated *in situ*, with real-time results. These advances resulted in a rapid development of tools and technologies to complement nanopore sequencing both in the lab and in the field.

One area which stands to benefit significantly from these developments is the study of ocean microbes, particularly eukaryotic phytoplankton. Ocean microbe communities are hugely important ecologically and play an essential role in the global climate through biogeochemical cycles, but populations are under threat due to anthropogenic climate change. Eukaryotic phytoplankton make up the majority of the biomass in the ocean microbiome (Bar-On and Milo 2019), but there is relatively little genomic information available from them as they are enormously under studied in comparison to their importance. There is an urgent need to build a clearer understanding of their diversity, influences on biogeochemical cycling and the processes through which they exert them, the likely effect climate change will have on them, and ways in which they can be protected (Cavicchioli et al. 2019). To achieve this, we need to characterise phytoplankton communities now, before there is significant change in response to climate change, and establish a programme of consistent monitoring, so that changes can be identified relative to the baseline (Ferguson et al. 2023).

Interactions within communities as well as with environmental variables, and larger organisms should be included in such monitoring, to gain a full understanding both of the effects phytoplankton have on their environment, and of the potential indirect effects of climate change on phytoplankton through environmental factors, or other organisms (Abreu et al. 2022).

Challenges associated with sequencing eukaryotic plankton include their large, complex genomes which make it difficult to produce high quality genome assemblies, difficulties growing many species in culture, and their extreme diversity (Obiol et al. 2020). As long-read DNA sequencing improves and throughput increases, large complex genomes are increasingly manageable. High throughput, highly accurate, long-read nanopore sequencing is allowing the production of good quality genome assemblies from metagenomic data (Duncan et al. 2020), removing the need for culturing samples prior to sequencing, while portable DNA sequencing using the MinION allows for the sequencing of samples *in situ* without requiring any storage at all. Projects such as Tara Oceans are providing sequencing information for a wide range of eukaryotic phytoplankton which is being used to produce assemblies and taxonomic analysis (Royo-Llonch et al. 2021; De Vargas et al. 2015), while the EBP aims to produce genome assemblies for all eukaryotes including phytoplankton (Lewin et al. 2022). Together, this increase in information and understanding can be used to produce improved models of ocean microbiomes, and provide a clearer picture of their interactions and impacts.

The aim of this project was to use nanopore sequencing to study ocean microbiomes, particularly eukaryotic phytoplankton. This was split into three main parts. Chapter 2 presented work on the production of a high quality *E. huxleyi* genome assembly from nanopore sequencing data, including optimisation of DNA extraction, nanopore sequencing, assembly, quality assessment, and removal of contaminants. Chapter 3 covered the *in situ* metagenomic nanopore sequencing of polar ocean samples onboard a research cruise in the Southern Ocean, followed by land-based nanopore sequencing of samples collected but not sequenced onboard and analyses including assembly-free functional annotation. Chapter 4,presented *in situ* and laboratory-based metagenomic nanopore sequencing of samples collected at Cromer Pier on the Norfolk Coast, to investigate the distribution of phytoplankton communities over a time-course, and develop and evaluate new, simplified and safer protocols for the use of citizen scientists and outreach efforts, and potentially for use onboard research cruises as part of the standard experiment roster, taking advantage of advances in nanopore sequencing technology since Chapter 3.

When this project began in 2017, nanopore flowcells were temperamental and required cold-chain transport and refrigerated storage. It was not uncommon for hundreds out of the 2048 nanopores to have deteriorated prior to beginning a sequencing experiment. There were baseline requirements of at least 1 µg of DNA to yield around 1 Gbp of data per flowcell, and the basecalling accuracy was around 95%. Today, by contrast, 400 ng of DNA is enough to reliably return 10 Gbp of sequencing data from flowcells which can be stored and transported at ambient temperatures, see figure 5.1.1 for a representation of increasing yield from MinION flowcells over time.

The newest generation of flowcells (10.4.1) feature improved nanopores, which are capable of detecting bases more accurately at higher speeds, increasing yield and accuracy. Basecalling accuracy has increased to around 99.5%, and there are constant improvements still being made. Barcoding kits allow for multiplexing of samples, meaning that researchers can sequence up to 95 samples on the same flowcell and easily analyse the resulting data. Flongle flow cells offer nanopore sequencing of up to 3 Gbp on a smaller flowcell costing around £60. The ONT GridION and PromethION allow for high throughput nanopore sequencing. The GridION is capable of running 5 MinION or flongle flowcells at once, while specialised PromethION flowcells can produce over 200 Gbp of sequencing data with up to 48 flowcells per machine - around 10,000 Gbp in total.

These improvements, along with other experimental and technological developments, mean that researchers can now produce genome assemblies which are both more accurate and more contiguous, alongside improved classification of metagenomic sequences. MAGs and improved functional annotation would allow for more comparative analysis between strains and species, and allow us to investigate the genes which allow them to adapt to different niches. With higher sequencing yields, this is now possible for samples of the type collected during fieldwork. Given these improvements, if I were starting this project now, as opposed to 5 years ago, I would approach much of the work differently, for example with different assembly strategies for *E. huxleyi* based on a larger sequencing yield of highly accurate reads, or a focus on production of MAGs from metagenomic nanopore data. The following sections will discuss the improvements and changes to protocols and technologies since the work in Chapters 2, 3, and 4 was carried out, and consider how these, and future improvements, can be incorporated into similar projects.

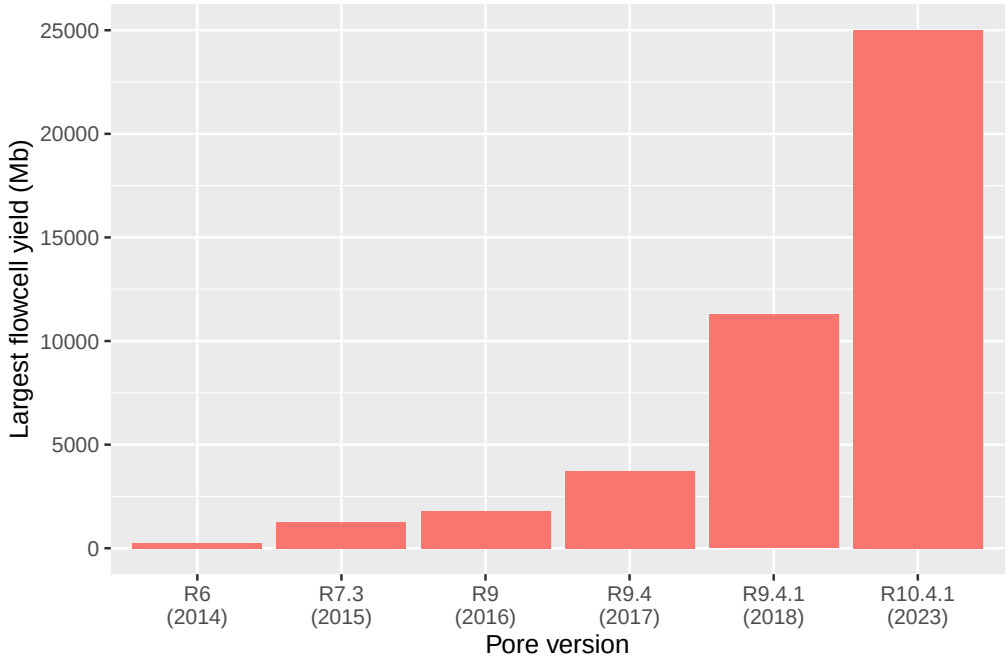


Figure 5.1.1: Bar chart showing the change in yield from a single MinION flowcell over time. Produced by Richard Leggett, 2023.

5.2 Producing a high quality assembly *E. huxleyi* RCC1217 genome assembly

5.2.1 Importance of an *E. huxleyi* genome assembly

E. huxleyi is a globally important coccolithophore which is heavily involved in biogeochemical cycling of carbon and calcium (Paasche 2001). *E. huxleyi* is known to be highly adaptable, having been identified in all global oceans, with recent poleward expansion into the Southern and Arctic oceans (Winter et al. 2014). The basis for this adaptability was illuminated in 2013 with analysis of the genome assembly of a diploid calcified strain CCMP1516, along with sequencing of several variants, which established that *E. huxleyi* has a core set of genes universally present in all strains and a subset of genes which is differentially present across strains, referred to as a pan-genome (Read et al. 2013). These findings improved our understanding of the adaptability to a wide range of environmental conditions exhibited by *E. huxleyi*, but there are still unanswered questions, particularly around the haploid phase of its life cycle.

Like other coccolithophores, *E. huxleyi* has a haplo-diplontic life cycle with differences in calcification, morphology, and cell size in the haploid and diploid phases (Dassow et al. 2009). This is believed to contribute to its dominance in

global oceans by expanding the range of habitats it can inhabit, and also means that it can have different ecological impacts depending on life cycle phase (Vries et al. 2021). While recent studies have identified differences between the haploid and diploid phases including metabolism, UV tolerance, nutrient limitation which allow them to withstand different conditions (Ruan et al. 2023; Rokitta et al. 2014; Dassow et al. 2009), the haploid phase is relatively poorly studied which is a limiting factor in fully understanding the complex ecological and climactic contributions and impacts. To improve this understanding, a haploid strain of *E. huxleyi* was sequenced and an assembly produced, Chapter 2.

Genome assemblies are a key part of genomics research, allowing the study of an organisms' evolutionary history, adaptations to environmental change, and diversity within populations (Rhie et al. 2021). Through functional annotation, genes can be identified, giving us information about the molecular mechanisms underlying traits, and how these change over time and within communities (Delmont et al. 2022). Future functional annotation of the RCC1217 assembly would provide researchers with a useful resource for taxonomic classification, structural analysis, and research into the haploid phase of the coccolithophore life cycle. Comparative analysis with the publicly available CCMP1516 assembly could provide insights into the genetic basis for differences between haploid and diploid cells, and their differing contributions and impacts to the global climate and ecology. A future genome assembly of the diploid RCC1216 strain could allow for a direct comparison of haploid and diploid phases.

The RCC1217 genome assembly discussed in Chapter 2 represents a significant advancement in genomics thanks to the introduction of nanopore sequencing and use of Hi-C long-range data to produce high quality, contiguous assemblies of complex genomes. Since the project was undertaken, there have been further advances in HMW DNA extraction, nanopore sequencing yield and accuracy, and in genome assembly algorithms. These changes have further increased the utility of nanopore data and will lead to assemblies which are more accurate, and more complete. Future projects using nanopore sequencing will benefit from the ability to produce high-quality genome assemblies for non-model organisms with complex, noisy genomes.

5.2.2 Coverage and read length

As discussed in Chapter 2, the production of an improved *E. huxleyi* RCC1217 assembly based on nanopore sequencing was undertaken in 2018. The DNA extraction method was carefully chosen following extensive testing to produce

high-purity HMW DNA, and produced reads with an N50 of around 25 kbp, which was well above that achieved with alternative protocols and in previous experiments. Continued research and testing since 2018, however, has identified HMW DNA extraction protocols developed for plant, fungal, and algal samples, including the Qiagen DNeasy PowerSoil kit (`https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/microbial-dna/dneasy-powersoil-pro-kit/?catno=47016`). This may offer better results as they are specially adapted to break down the polysaccharide cell walls which make it harder to lyse cells and which can contaminate the extracted DNA. If the production of an improved *E. huxleyi*, or other phytoplankton species with a complex genome, were to begin today, these alternative extraction protocols should be considered, as an increase in raw read N50 to 30 kbp and above would improve the quality of the resulting assembly.

A potential starting point would be the CTAB and Genomic-Tips combined protocol used for *M. acuminata*, a recent ground-breaking highly contiguous assembly of the *Musca acuminata* banana genome. Banana genomes are highly complex with varying ploidy, large repeat sequences, reciprocal translocations, and inversions, which had previously limited the success of genome assembly attempts. Using a dedicated HMW DNA extraction method, and long-read nanopore sequencing, researchers were able to resolve the majority of repeat regions and structural variation to produce gapless chromosomes. The DNA extraction protcol, recommended on the ONT Community in 2019, for HMW DNA extraction from plant leaves combined the CTAB and Genomic-Tips extraction techniques covered in Chapter 2. This method produced reads with an N50 of over 30 kbp. Nanopore sequencing was used to produce over 90 Gbp of sequences produced using a single PromethION flowcell, giving coverage of 177x which allowed researchers to produce a contiguous assembly. The assembly was then improved using polishing both with nanopore reads and PCR-free Illumina reads before validation using optical mapping (Belser et al. 2021). This provides insights into how a *de novo* genome assembly for RCC1217 might be approached today.

An alternative approach would be to use PacBio HiFi long read sequencing along with chromatin capture such as Omni-C, as has been used recently to produce highly contiguous genome assemblies for both *Perilla frutescens* (Tamura et al. 2023) and *Callipepla californica* (Benham et al. 2023). A 2020 study found that nanopore long-read sequencing on ONT platforms produced assemblies with fewer errrors associated with long repeats, assembled into higher contiguity (18 contigs, 10 of which were assembled into a single chromosome, as opposed to 394 contigs and 3 chromosome-level contigs with PacBio HiFi), likely due to

the ultralong read length. It also found, however, that PacBio HiFi assemblies resulted in fewer single nucleotide errors and small insertions and deletions. (Lang et al. 2020). This indicates that depending on the project PacBio HiFi sequencing may be more appropriate, where nucelotide errors are particularly critical, and that nanopore sequencing efforts are likely to require polishing with more accurate sequencing data such as from Illumina. Where resources are sufficient, a combination of PacBio HiFi and nanopore sequencing may yield the best results.

The depth of coverage of sequencing reads required to produce a high quality assembly is lower for long-read sequencing than for short-read sequencing, due to the longer reads being easier to assemble. In general coverage of at least 30-60x of each genome to be assembled allows for the production of genome assemblies which are contiguous and sufficiently accurate for single nucleotide variations to be analysed (Koren et al. 2017). The required coverage depth can be higher for highly complex genomes. In this project, the coverage was around 18x, and the assembly was of reasonable quality. As can be seen in 5.1.1, and considering the work done in the Leggett lab to compare current nanopore sequencing outputs to those from 2018 as referenced in 2, if the same sequencing protocol were to be repeated today with the new flowcells, the expected coverage would be around 150x which should be more than sufficient to produce a high quality, accurate genome assembly, even taking into account the complexity of the *E. huxleyi* genome. Combined with the use of Hi-C long range data, and polishing with highly accurate Illumina reads as with this project, it would likely be possible to produce a chromosome level assembly.

5.2.3 Long-read accuracy

The relatively high error rate of nanopore sequencing means that nanopore sequencing based eukaryotic genome assemblies are generally supplemented with long-range data, such as Hi-C or optical mapping. A recent telomere-to-telomere assembly of the human X chromosome was produced through a combination of nanopore sequencing, SMRT sequencing, linked-read sequencing, and optical mapping. Through manual curation, and significant financial and time investment, this allowed researchers to produce a highly contiguous genome but there are still gaps between contigs especially in complex regions (Miga et al. 2020). This is because while long-range technologies can help to arrange contigs, they cannot fill gaps. Gaps are generally found in complex, repetetive regions which are difficult to resolve without reads which span the gaps. Increasing accuracy, and hybrid assembly methods using Illumina sequences for polishing,

coupled with ultra long-read sequencing, could allow gaps such as these in genome assemblies to be filled (Belser et al. 2021). The accuracy of nanopore sequencing has increased significantly since 2018, from around 95% to over 99.5% with the newest flowcells (v10.4.1) and super-high accuracy basecalling software from ONT, see figure 5.2.1.

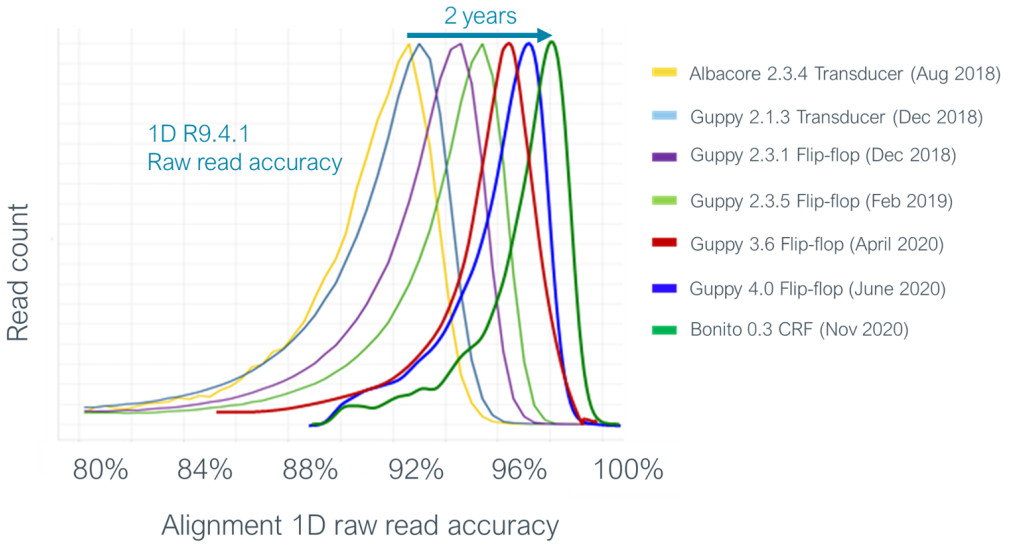


Figure 5.2.1: Line graph showing the increase in accuracy of nanopore sequencing over time from 2018-2020, reproduced from https://nanoporetech.com/how-it-works/basecalling.

5.2.4 Identifying and removing contaminants

Contamination of sequencing data and genome assemblies is a growing problem for genomics. As sequencing projects move away from highly studied culturable bacteria and lab-grown model organisms to sequencing from environmental samples or wild flora and fauna, there is increased risk of contamination from other organisms. Other points of contamination include growth of unwanted organisms in an axenic culture, or human contamination of samples during DNA extraction and processing. *In silico* contamination occurs during data processing, for example in the separation of metagenomic sequences, with chimeric sequences being produced through merging of similar sequences during metagenomic assembly, or through adapter sequences which have not been removed (Cornet and Baurain 2022). The presence of contaminants in genome assemblies causes a range of problems for research. Contamination of a genome assembly can lead researchers to draw erroneous conclusions, such as the report of extensive horizontal gene transfer found in a tardigrade genome which was later shown to be due to bacterial contamination of the assembly (Arakawa 2016).

Also, databases including genome assemblies are often used for identification when unknown organisms are sequenced, and contamination of a reference assembly can result in misidentification. Adapter sequences present in either a published genome or sequencing data searched against a database can result in incorrect taxonomic identification (Sturm, Schroeder, and Bauer 2016). This was seen in Chapter 3 when the Ship-Seq samples were analysed prior to adapter removal, with the results indicating a surprisingly large biomass of sugarcane aphids living in the Southern Ocean. Contamination has also been found to be a cause of problems for phylogenomic analyses, leading to incorrect conclusions about organisms' evolutionary history (Schierwater et al. 2009; Philippe et al. 2011). It is important, therefore, to guard against contamination of sequencing data and genome assemblies at every stage of the process, to ensure that genome assemblies and databases remain useful sources of information.

It is difficult to identify contamination in a *de novo* genome assembly, especially one produced from a range of different data types. Many methods for assessing the contamination of a genome assembly rely on mapping raw reads to the assembly to identify parts of the assembly which have different levels of read coverage which can help to identify contaminant reads. This is not workable with a hybrid genome assembly which has been polished, correcting the assembly so that it no longer matches the raw reads. For the RCC1217 assembly, the contamination was identified after assembly, scaffolding, and polishing so an alignment-based contaminant identification was used, based on BLAST-nt analysis to identify assembled contigs which matched to species other than *E. huxleyi*. This method has downsides, particularly for a member of a clade which is so underrepresented in genome assembly databases, in that it can be hard to determine whether all of the sequences which have reported alignments to non-*E. huxleyi* species are due to contamination. Where the length of blast hits is low, it is possible that the alignment is simply a random match, and where the percentage identity of a match is low it may indicate that there is not a good match for that sequence in the database, perhaps because there are few haptophyte assemblies available to check against. Alignment-based contamination identification is particularly challenging for species such as *E. huxleyi* or diatoms, which have a complex evolutionary history which includes multiple symbiosis events and potential for horizontal gene transfer. As such a conservative threshold for alignment was set, to attempt to balance the removal of contaminants against the retention of potentially real sequences. If this project were to be repeated, a key consideration would be to consider using tools such as KAT prior to polishing and sending the assembly to Dovetail to identify potential contamination before producing a hybrid assembly. A further consideration would be the use of axenic cultures, such as used in (Harvey et al.

2016), where the culture is cleared of bacteria, either through physical separation or use of antibiotics, although axenic phytoplankton cultures often suffer from reduced viability (Shishlyannikov et al. 2011).

5.2.5 Completeness and contiguity: Perfection versus progress

Contiguity is an important indicator of genome assembly quality, with N50 and L50 commonly used to describe it. These are useful for assessing and comparing the contiguity of a genome assembly, but they can be misrepresented in noisy datasets and an alternative measure, the U50 has been proposed which relies on comparison to a reference genome, so is unsuitable for use with *de novo* assemblies where there is no reference genome available (Castro and Ng 2017). As such, the N50 and L50 were used to assess the contiguity of the RCC1217 genome assembly. The contiguity of the various draft assemblies produced in the process were compared to decide on the best one to use, and the resulting assembly contiguity was compared to the publicly available *E. huxleyi* CCMP1516 assembly. The N50 of the RCC1217 assembly was an order of magnitude greater than that of CCMP1516, with the L50 an order of magnitude smaller, indicating far greater contiguity for the new RCC1217 genome assembly compared to the 2013 CCMP1516 assembly. Comparisons between the intermediate draft assemblies show that the biggest increases in contiguity came from the Hi-C/Hi-Rise scaffolding performed by Dovetail Genomics using long-range data (0.2 Mbp N50 to 1.08 Mbp), and from polishing the HiC-Canu intermediate assembly with Illumina reads (1.08 Mbp N50 to 5.2 Mbp).

A further indication of genome assembly quality is the BUSCO score, which estimates the completeness of a genome assembly based on the presence, absence, duplication, or fragmentation of a set of highly conserved genes from a given taxa (Manni et al. 2021). As with contiguity, BUSCO scores were compared between the intermediate draft assemblies and between the final assembly and the published CCMP1516 assembly. The BUSCO score for the haploid RCC1217 was 16% higher than for the CCMP1516, which indicates the RCC1217 assembly is more complete. As with contiguity the completeness increased the most (from around 13% to 58%) after the polishing stage. This indicates that nanopore sequencing accuracy was a limiting factor in the quality of the genome assembly, and that with the increased coverage afforded by greater flowcell yield and the improved accuracy from updated flowcells and the newer basecalling software since 2018, the contiguity of the nanopore-only assemblies would be vastly improved.

The question of what constitutes a finished genome assembly ready for publication is not necessarily simple to answer. A complete genome is currently defined by the NCBI assembly database (`https://www.ncbi.nlm.nih.gov/assembly/help/`) as an assembly where "all chromosomes are gapless and have no runs of 10 or more ambiguous bases (Ns), no unplaced or unlocalised scaffolds, and all the expected chromosomes are present". This standard is not an easy one to reach, with only 38,530 of the over 1.3 million assemblies in the NCBI assembly database marked as complete, 305 of which are eukaryotic, 567 are archaeal, and 37,658 are bacterial. Producing a complete genome assembly for eukaryotic species is extremely difficult, and requires a significant investment of time and computational resources. There are 30 complete protist genome assemblies all of which are parasites, mainly plasmodium, leishmania, and cryptosporidium species which cause disease in humans and livestock. These all have small haploid genomes, with relatively few regions which are hard to assemble compared to highly complex plant genomes, and due to the need to differentiate between strains to treat infection effectively, there has been a big effort to produce complete genomes.

Given the lack of complete genome assemblies, and the difficulty, time, and cost involved in producing them, researchers must decide at what point to sacrifice perfection for progress. While complete genome assemblies remain the ideal, incomplete assemblies are widely used in research and they are able to provide a significant amount of information about an organism, variation within and between species, and epigenetics. Chromosome-level assemblies are defined by the NCBI as containing "sequence for one or more chromosomes. This could be a completely sequenced chromosome without gaps or a chromosome containing scaffolds or contigs with gaps between them. There may also be unplaced or unlocalized scaffolds.". These form the basis for functional, population, and comparative genomics research for most eukaryotic species. The human genome project was declared complete in 2003 with 92% of the genome sequenced, and it was not until 2022 that the first complete telomere-to-telomere gapless human genome assembly was published (Nurk et al. 2022).

Long-read sequencing is bringing complete genome assemblies into reach for other organisms, as seen with the *M. acuminata* assembly which featured 5 gapless telomere-to-telomere chromosome assemblies (Belser et al. 2021) but this is not yet fully feasible. As such, it is sensible to take a pragmatic view and produce genome assemblies that are as complete as possible at the time they are produced and make improvements as scientific advances allow.

The highest feasible completeness is dependent on the organism being sequenced, as small simple genomes can be fully assembled relatively easily while complex genomes can take a great deal of time and resources to get to

chromosome level. This is the approach proposed by the EBP (Lewin et al. 2022), which aims to produce chromosome level assemblies for all of the approximately 1.65 million eukaryotic species on earth. This was decided to balance the time and resources required to produce a complete assembly against the need for high quality assemblies. It has also been decided that for certain organisms, such as uncultured eukaryotes or those with highly complex genomes, that the EBP will accept lower quality assemblies if the production of a chromosome-level assembly is not feasible in the timescale of the project (Lewin et al. 2022).

The RCC1217 assembly is at a scaffold level, which the NCBI defines as an assembly where "some contigs have been connected across into scaffolds but the scaffolds are all unplaced or unlocalised". There are six publicly available haptophyte genomes in the NCBI database, of which none are complete, one is at chromosome-level, with two at scaffold level and three at contig level. This indicates that the RCC1217 assembly is comparatively good for this clade and that the assembly is likely to be of benefit to researchers as it expands the limited resources available. Long-read sequencing is still a relatively new technology, with best practices still being established; currently there is no standard process for the production of a high quality *de novo* genome assembly based on long-read sequencing. Instead each one uses an *ad hoc* combination of tools and technologies. This has benefits as it allows researchers to be flexible and determine the best methods for the organism they are working on, but limits the production of genome assemblies based on long-read technology to a small subset of genomics scientists. Some of the key benefits of nanopore sequencing are the low upfront cost and small footprint, which opens up sequencing to a wider range of users. The development of a standardised set of protocols for genome assembly of a range of different organisms, with recommended data types, analysis tools, and quality checking, would allow more researchers to produce high quality genome assemblies to advance their research, and increase the number of publicly available genomes.

5.3 Metagenomic nanopore sequencing of ocean microbiomes

Polar ocean microbial communities support polar foodwebs, are heavily involved in nutrient cycling, and make an enormous contribution to CO_2 sequestration and O_2 production (Katz et al. 2004). Polar eukaryotic phytoplankton, such as diatoms, are especially important in biogeochemical cycles (Boyd 2002). These

populations are under disproportionate threat from climate change, due to rapid warming of polar oceans and the impacts of climate change on features such as the ACC in the Southern Ocean which is important for the creation of diverse habitats (Smetacek and Nicol 2005). Despite this, they are particularly under studied, due to their remote location and challenging sampling conditions, as well as challenges associated with sample storage and transport, which in the case of Southern Ocean research cruises can take months. Projects such as Tara Oceans (Brum et al. 2015), which collected and sequenced samples from over 200 sites across the global oceans, including polar oceans, have shown that our previous understanding of ocean microbiome composition and diversity from culture-based studies was flawed, particularly for eukaryotic phytoplankton (Malviya et al. 2016; Carradec et al. 2018; Obiol et al. 2020). Research cruises are undertaken each year by organisations such as the British Antarctic Survey in the Antarctic, undertaking sampling including of phytoplankton communities, and there are regular sampling efforts in the Arctic which provide samples for sequencing analysis. The portability, relatively low cost, and ease of use afforded by portable nanopore sequencing raises the possibility of non-specialist researchers using nanopore sequencing to perform *in situ* real-time analysis during field sampling, either as a stand-alone dataset, or to complement the collection of samples for storage and later analysis. This would allow researchers to get an immediate picture of the microbial community at a given sample location, evaluate sampling techniques, and help to optimise sampling location and volume, opening opportunities for targeted sampling of populations of interest. Using read-until techniques, it is possible to sequence only until a set threshold is reached, such as a set number of sequences classified as belonging to a species of interest, allowing for rapid production of results from metagenomic samples.

This was the basis for Ship-Seq, Chapter 3, which was planned as a proof-of-concept experiment to test the feasibility of sequencing and analysis onboard a research ship in the Southern Ocean, to produce taxonomic classification of polar ocean microbiomes, and to understand what alterations would be required for the protocol to be used by non-specialist scientists. Extra samples were collected and more in-depth analyses carried out in the laboratory to investigate polar ocean microbial communities, evaluate the sequencing results, and test the feasibility of functional annotation, correlation to metadata, and genome assembly. The study provided insights into the microbial communities in the Southern Ocean, and showed that MinION sequencing with real-time analysis during a research cruise is viable, and worthwhile. Further land-based analyses showed that nanopore sequencing data from polar ocean samples can be used to identify important eukaryotic phytoplankton such as

diatom and coccolithophore populations, indicating that nanopore sequencing could be a useful method for monitoring populations over time, particularly if carried out as standard during research cruises - collecting data regularly across wide areas to help establish a baseline and identify changes, for example as a result of climate change. Research into the effects of climate change on ocean microbiomes, and the knock-on effects on the biological carbon pump, have identified a lack of baseline information on ocean microbe communities as a significant barrier to improving existing models (Henson et al. 2021). Assembly-free functional annotation showed the most abundant genes, and comparative analysis between bacteria and eukaryotes identified divergent genes, and helped to confirm taxonomic classification, but there was insufficient data for production of MAGs, or for effective correlation of taxonomic classifications or genes to metadata, with experiments indicating that over 100 Gb of sequencing data, requiring at least 5 MinION flowcells based on what is currently achieved in normal laboratory conditions. This experiment was overall a success, a number of points for improvement were identified, mainly related to the complexity of the workflow, and limitations imposed by the technology available at the time. These included the complexity and hazardousness of the DNA extraction methods, the instability of flowcells in transit and the resulting low sequencing yields, and the relative complexity of the computational sequencing and analysis which required two computers, and use of the command-line interface.

Cheap, portable, easy to use sequencing technologies opens the possibility of increased citizen science contributions to ocean microbial ecology. Citizen science is an increasingly important part of improving our understanding of the composition and diversity of ecosystems, with many projects around the world where citizen scientists make important contributions, including CALeDNA (Meyer et al. 2021). Citizen science is one way of increasing public engagement, which is important for developing understanding of the importance of ocean micriobiomes, and the challenges facing them, and by extension us, as a result of climate change and the effects of human industry.

Pier-Seq, 4 tested a simplified protocol using a low sample volume, rapid sample preparation, and optionally *in situ* sequencing which could be used by anyone with little training at low cost. This would provide a snapshot of the most abundant microbes present at a given location, which could be combined with data from groups performing experiments either over time or at different locations, which could provide researchers with a starting point from which to plan deeper analyses, and provide an improved understanding of microbial diversity and abundance around the British coast. In-field and laboratory

samples were analysed to test the new protocols and to investigate the Norfolk coast marine microbiome. From Chapter 4 it was established that low-input sequencing is effective for *in situ* sequencing and taxonomic classification, with larger samples optionally collected for sequencing later in the laboratory for in depth analyses and the production of assemblies or functional annotations. The Norfolk coast marine microbiome, encompassing bacteria, viruses, and eukaryotic microbes, was analysed over a time-course to evaluate the ability of nanopore sequences to monitor change over time, but there was little variation, perhaps due to the timescale being too short.

Based on the workflow used in Chapter 4, and advances in nanopore sequencing technologies since 2021, the following sections consider the changes which would be required if nanopore sequencing were to be carried out *in situ* onboard a research ship in 2022-2023.

5.3.1 Travelling light - portability and streamlining for *in situ* sequencing

Some of the key technological improvements since the Ship-Seq project which would be beneficial for a research cruise are the increased portability and ease of transport for ONT sequencing equipment and consumables. The MinION Mk1C removes the need for a second computer, since it has built in compute capabilities and software to run the sequencing control and basecalling software. During Pier-Seq we tested many of these improvements. Flowcells can now be transported at ambient temperature instead of requiring cold-chain transport, and are also far more stable, which reduces the redundancy required for transport and maximises yield. The yields from flowcells have also improved hugely, regularly yielding 10 times as much data as was achieved in Ship-Seq. This would allow for the sequencing of more samples on board the ship through the use of multiplexing where multiple samples are sequenced on the same flowcell. Other portability improvements not tested during Pier-Seq include lyophilised reagents, in powder form which can be transported at ambient temperature and reconstituted with water before use, rather than requiring cold-chain transit. These have been used for fieldwork in adverse environments (Maestri et al. 2019) and could also be useful for researchers travelling for fieldwork or to join a research cruise which requires long-haul travel with limited access to cold storage. The relatively poor performance of the ship-based sequencing runs is likely to have been a result of degradations to flowcells and/or reagents. New flongle flowcells could be used for test runs or even, depending on yield requirements, single sample runs. They can produce up to around 3 Gbp of sequencing data, allowing researchers to save the

larger, more expensive full size MinION flowcells for more in depth sequencing or multiplexing.

The improved workflow trialled in Pier-Seq, see figure 4.3.9, included an adapted DNA extraction protocol to reduce the volume, and toxicity of reagents to be transported, in order to simplify transport and make the process more user-friendly. The revised protocol used the Qiagen DNeasy PowerSoil kit which removes the requirement for highly toxic reagents including β-mercaptoethanol. The use of a kit simplifies transport as the manufacturer produces safety data sheets for the kit as a whole, and the reagents are already sealed for transit, with researchers transporting DNA and RNA extraction kits, including Qiagen DNeasy as standard checked luggage for air travel (Maestri et al. 2019; Quick et al. 2016). A mini kit with 50 preparations would be more than sufficient for most research cruise applications, with a larger 500 preparation kit available for longer fieldwork stints. If reagents were transported in this manner it would increase flexibility for remote fieldwork, such as polar research cruises, which often have complex logistical arrangements requiring equipment and reagents to be sent several months in advance of a cruise if they are not being transported in checked luggage.

The revised DNA extraction protocol is faster, taking under an hour from sample to sequence, and it is easier for non-specialists, with no difficult steps such as carefully pipetting the top layer of a phase separation or special training required to handle the chemicals. This could allow nanopore sequencing to be added to the roster of general science onboard research cruises or at research stations. The improved protocol also has significantly reduced sample volume requirements which increases the flexibility of incorporating sequencing into the other scientific activities taking place onboard a research ship. The DNA yield from 20 mL sample volumes in the North Sea was low at around 5ng per sample, but this was sufficient for an overview of taxonomic identification in the immediate sample location. The sample volume required for equivalent taxonomic identification in the Southern Ocean as was achieved in the North Sea remains to be tested, however, so it is possible that the sampling requirements would need to be increased for areas with low concentration of phytoplankton.

5.3.2 Taxonomic classification of metagenomic sequences

One of the key benefits of portable nanopore sequencing is the ability to perform real-time analysis *in situ*. Taxonomic identification in real-time can be used to conserve the use of resources in the field where they may be at a premium: if

only aiming to confirm the presence of a species in the field, with further analysis performed later, sequencing could be stopped once it had been identified; or, using a rarefaction curve, sequencing could be stopped at a point where sufficient sequencing had taken place.

For Ship-Seq, real-time analysis was carried out using NanoOK RT (Leggett et al. 2018). NanoOK RT performs BLAST-based classification of reads for real-time analysis and comes with a companion tool, NanoOK Reporter, which provides visualisation of the community composition with a doughnut plot or a taxonomic tree. This worked well for Ship-Seq but it required the use of the command-line and may be challenging for non-specialists.

Subsequent to Ship-Seq, NanoOK RT was superceded by MARTi (Meteagenomic Analysis in Real-time), `https://github.com/richardmleggett/MARTi` (Leggett et al. 2018). This was used for Pier-Seq and provided a range of advantages over NanoOK-RT for *in situ* analysis. The user interface has been simplified, with a GUI interface to initiate analyses as opposed to command-line only, which would reduce the training required for non-specialists to use it. The range of visualisations has been increased, now encompassing interactive plots including stacked bar charts, doughnut plots, taxonomic tree, a tree map, and a rarefaction curve. The taxonomic level for analysis can be changed, and multiple samples can be compared. This reduces the post-sequencing analysis required as much of the analysis can be performed during the run. The results from MARTi were used to examine taxonomic identification at various levels, establish a likely sampling depth required, and to look for specific species of interest such as *E. huxley* and diatom species. The use of MARTi, combined with a simplified DNA extraction protocol open up the use of nanopore sequencing in the field to a broader range of researchers.

5.3.3 Metagenomic assembled genomes (MAGs)

Neither Ship-Seq nor Pier-Seq datasets were sufficient to yield MAGs for eukaryotic phytoplankton. Attempts to produce assemblies of known high-abundance genomes from the metagenomic data were not very successful, with one partial assembly produced for the bacterium '*Candidatus* Pelagibacter ubique', which has the smallest genome of any known free living organism at 1.3 Mbp. Higher coverage would be needed for MAG production to be viable as the the total sequencing yield of 28 Gbp, including 12 Gbp in depth sequencing one sampling station, provided sufficient data for taxonomic analysis but was insufficient for MAGs.

The production of genome assemblies from metagenomic sequencing data is a relatively new approach, and until recently there were very few eukaryotic MAGs which were largely produced from communities with low diversity (West et al. 2018; Joli et al. 2017). MAGs present an opportunity to study organisms which are not easy to grow in the lab in culture, and which may not have been cultured at all. Culture-dependent methods for phytoplankton research have limited our understanding of their diversity and interactions, because the majority of marine microbes are not culturable, and so their distributions, or existence, have not been visible to researchers. Work done by the Tara Oceans project (Bork et al. 2015) illuminated this knowledge gap, as it was shown that some widely reported phytoplankton with important roles in food webs, carbon cycling, and biogeochemical cycles, had previously been assumed to be minor contributors based on their presence in cultures.

The next step would be to investigate these newly discovered interactions and biogeochemical cycle contributions, but this has been difficult without genome assemblies which can offer a clue as to the functional differences between groups. Recently, three large groups of eukaryotic phytoplankton MAGs with functional annotations have been presented, one with 683 MAGs from the Tara Oceans project covering a wide range of sample locations including some polar stations, (Delmont et al. 2022), another with over 900 particle-associated eukaryotic MAGs from the Tara Oceans data (Alexander et al. 2021), and 143 from the North Atlantic and Arctic oceans (Duncan et al. 2020). These used high-throughput Illumina sequencing data, with an input of 679 Gbp used for the Arctic and North Atlantic MAGs.

A recent North Sea study has found that long-read sequencing technologies produce higher quality bacterial MAGs with similar composition at species level compared to Illumina sequencing, while Illumina sequencing recovers more MAGs and a greater species number, due to the sequencing depth being higher. (Orellana et al. 2023). A study carried out on the Californian coast found that prokaryotic MAGs were not particularly improved by long-read sequencing data, and they may be less reliable for reporting microbial community composition and function, but that for eukayotes, only long-read sequencing was capable of providing high quality MAGs (Patin and Goodwin 2022). As such, there are costs and benefits to both approaches, and it is likely that a combination of Illumina sequencing and long-read sequencing with PacBio or nanopore would be the optimal approach to capture a range of MAGs as well as community composition and function.

As with single species *de novo* assembly, the often large genome size and relative complexity means that production of assemblies for eukaryotic

phytoplankton lag behind that of bacteria. A good-quality MAG is defined as >90% complete with <5% contaminants (Bowers et al. 2017). The genomes of eukaryotic phytoplankton range in size from *Ostreococcus* at 12 Mbp to dinoflagellates with estimated genome sizes ranging from 1 Gbp to s over 250 Gbp (Lin 2011). Ditaom genomes range from 20 Mbp to around 200 Mbp and the *E. huxleyi* genome around 140 Mbp (Read et al. 2013). Given that the taxonomic identification for Ship-Seq and Pier-Seq indicated that the total proportion of a sample taken up by eukaryotes was below 30%, it is likely that producing MAGs for eukaryotes from a dataset such as Ship-Seq would likely require hundreds of Gbp of sequencing data.

In 2018, it was not feasible to create a nanopore sequencing dataset of this size, due to both prohibitive cost and low nanopore sequencing yields. Since then, however, there have been significant advances in nanopore sequencing yield which could allow for the production of large nanopore sequencing datasets. Were the Ship-Seq project to be repeated now, increased biomass could be collected by adapting sampling quantities, such as by using a filtration stand and pump as in Ship-Seq, or a larger alternative to the swinney filters used for Pier-Seq depending on location and resources available. A test sample could be analysed to establish the yield from a given sample volume. The extra samples could be collected at particular sampling points of interest, for example after identification of species of interest based on real-time analysis, or more widely depending on constraints on time and other factors, including sequencing cost. These samples could be sequenced using high-throughput nanopore platforms such as the GridION or PromethION. With new flow cells requiring an input of only 100 ng, it would be relatively easy to produce a long-read dataset of 200-300 Gbp, at a cost of around £5000, which is competitive with other sequencing technologies. This would provide an invaluable resource for researchers of polar phytoplankton, and allow for the production of MAGs with functional annotation which could be used to investigate important polar phytoplankton community interactions and adaptations.

5.3.4 Functional annotation

Functional annotation would allow us to investigate the underlying reasons for observations made from the taxnomic data. For example, functional analysis of the Arctic and North Atlantic MAGs by Duncan et al. 2020 helped to confirm taxonomic placement of the MAGs, and indicates that there may be differences between prokaryotic and eukaryotic evolutionary responses to selective pressures. Assembly-free functional annotation was carried out for Ship-Seq but it was

difficult to draw any conclusions from this data. This is likely to be due to a combination of factors including the small sample sizes, the relatively high error rate of nanopore reads, and also that and assembly-free functional annotation is less accurate than functional annotation of assemblies or contigs (Vázquez-Castellanos et al. 2014). Improvements in accuracy of nanopore sequencing since Ship-Seq, from 95% in 2018 to over 99% in 2022 would potentially allow for improved functional annotations of raw reads. Alternatively, production of MAGs which could be more easily annotated, with the added context provided by the assembly would allow for improved functional analysis. PacBio HiFi sequencing data would perhaps be useful for generating more accurate reads or even eukaryotic MAGs which could be better functionally annotated, although this would increase cost and limit read length unless used in combination.

5.4 Future Developments

5.4.1 The future of nanopore sequencing technology

The ONT vision is to enable anyone to sequence anything anywhere, with new technologies constantly under development to move closer toward this goal. Portable sequencing with the ONT MinION has opened doors to a wide variety of research opportunities since its release in 2014 with new technologies providing simplified processes and increased reliability, making *in situ* sequencing a reality.

Nanopore sequencing accuracy and yield continue to increase through the development of flowcell chemistry, including modifications to nanopores and other flowcell components, and through basecalling software improvements, through algorithm refinement and the incorporation of machine learning. Further alterations and improvements are under development, including a potential alternative flowcell chemistry which, similar to PacBio HiFi sequencing, allows the same strand to be sequenced multiple times through manipulation of the motor proteins and helicases to unzip the DNA and stop it escaping once it has finished passing through the pore. This could be used to increase accuracy by producing multiple copies of the same read, and as the technique calculates the length of the sequence on the first pass through the pore, this could be used selectively on long reads to produce highly accurate long reads for closing the gaps in genome assemblies.

The MinION is the only available portable DNA sequencing machine, with the newer Mk1C version including a built-in computer to run sequencing and real-time basecalling. The Mk1D is an upcoming MinION sequencing device

which will connect to a tablet computer so the sequencing and data processing can be run from an app. Similarly under development is the SmidgION, a miniature sequencing device which could run from a mobile phone. The SmidgION would see reduced sequencing output compared to the MinION, and could be particularly useful in the monitoring of disease or environmental contamination. An android app, named Genopo, has been developed to perform nanopore sequencing analysis on a mobile phone, which in combination with a SmidgION would allow an entirely smartphone-based sequencing and analysis process (Samarakoon et al. 2020).

The benefits of smartphone or tablet-based sequencing and analysis will not be fully realised, however, until DNA extraction and library preparation methods are similarly streamlined. Field-based DNA extraction and library preparation methods with reduced reagent, equipment, energy, and time requirements will be needed. To deliver automated library preparation, the VolTRAX V2 has been developed. This is a small device with a disposable sample cartridge which performs library preparation protocols from extracted DNA sample input to a prepared library which is ready to be loaded onto a flowcell. The VolTRAX reduces reagent requirements and waste, and ensures consistency across samples, and is capable of running PCR and PCR-free library preparations through the software-controlled liquid movement around the cartridge to perform different reactions, with built-in heating elements for incubation steps, and magnets for bead clean-up. The development of a portable automated DNA extraction device is more complex, due to different requirements for different sample types, organisms and experiments, but it would be transformative for *in situ* sequencing if successfully produced, removing the need for transport of expensive, heavy, and resource-intensive laboratory equipment and increasing the efficiency and consistency of DNA extractions. Coupled with an app-based simplified sequencing interface as discussed above, this could improve usability for researchers with little training in sequencing protocols, increasing the potentail use in citizen science projects.

Further developments for nanopore sequencing include the introduction of a solid state nanopore. Currently nanopores are made of protein which is delicate and subject to degradation due to a range of factors including: high temperatures; storage for longer than 30 days at ambient temperature, or 12 weeks under refrigeration; and use and re-use. This degradation limits sequencing yield and also limits the ability to wash and re-use flowcells, increasing costs and waste. Solid state nanopores, which are currently under development, (Goto et al. 2020), could be used to make flowcells more robust with the ability to withstand greater temperature changes and long term storage, and increase the number of times they can be washed and reused. This development would be of particular benefit

for *in situ* sequencing in remote locations where flowcells cannot be replaced. In the Ship-Seq experiment covered in Chapter 3, of 4 flowcells which were taken onboard, one had too few live pores to be usable, two had approximately three quarters of the original 2048 pores degraded at the start of sequencing, while the final flowcell had half of the original pores degraded. Flowcell stability has vastly improved since, but the introduction of solid state nanopores could improve it still further.

5.4.2 The future of nanopore sequencing of ocean microbiomes

There are thousands of eukaryotic phytoplankton for which there are no genome assemblies available, a problem which is particularly pronounced for polar communities (Abreu et al. 2022). Given the continued increases in yield and accuracy, combined with potential future developments, nanopore sequencing alone will likely be sufficient for the production of a high quality genome assembly of a from a single species sample (Liu et al. 2022). For species such as *E. huxleyi*, with genomes in the hundreds of megabasepairs, a single MinION flowcell yielding over 20 Gbp would provide high coverage, and highly accurate long reads could help to bridge gaps, resolving repeat regions and areas of low complexity for a cost of around £600. For organisms such as dinoflagellates, which can have genome sizes of over 250 Gbp and for which there are no chromosome level assemblies, continuing increases in yield of MinION and PromethION flowcells, and decreasing cost, might soon bring a fully assembled high quality genome within reach. These advances bring the Earth Biogenome Project's ambitious goal to produce a genome assembly for every eukaryotic species on earth closer to fruition, which would vastly increase the genomic resources available to researchers.

The field of metagenomic sequencing is rapidly advancing. Taxonomic classification, while still useful, is being augmented by the production of MAGs from classified reads. Recently, researchers using eukaryotic sequencing data from polar ocean samples have produced a huge number of new metagenomic assembled genomes (MAGs) and corresponding functional annotations (Duncan et al. 2020). This is providing insights into the genomes and genes of previously uncultured and un-sequenced species, by allowing wider comparison of genomes from different species and locations, improving our understanding of eukaryotic phytoplankton ecology and evolution. The quality of the taxonomic classification, MAGs, and functional annotation for eukaryotic metagenomes is limited by the use of short-read sequencing which is ineffective for resolving highly complex regions of a genome (Lapidus and Korobeynikov 2021).

Improved quality of MAGs, through the use of nanopore sequencing technology would allow researchers to move beyond cultivation-based methods for sequencing phytoplankton and instead focus on sequencing from environmental samples. This would have the added benefit of avoiding any genomic changes as a result of growth in culture being captured during sequencing, and allowing the analysis of intra- and interspecific interactions, as well as interactions with environmental variables.

Increasing flowcell stability, and the increasing yield achievable with portable sequencing devices raises the possibility that monitoring of polar phytoplankton communities could be carried out through sequencing *in situ* during research cruises, without the need for samples to be stored and analysed later. Increasingly portable sequencing devices, such as the MinION Mk1D and user-friendly lightweight analysis software such as MARTi, combined with simplified DNA extraction protocols, and easily transportable reagents, open up the possibility that anyone will be able to undertake DNA sequencing of samples with minimal training. Lyophilised reagents available from ONT for library preparations could allow nanopore sequencing to be carried out without needing to maintain a cold-chain in transit, with the potential for lyophilised DNA extraction kits from NEB in the future. Coupled with improvements to flowcell stability, possibly through the use of solid-state nanopores, this will make it much easier to carry out nanopore sequencing in remote locations. ONT's VolTRAX automates the process of library preparation, resulting in consistent libraries across experiments and streamlines the preparations required for nanopore sequencing. Future developments may include a DNA extraction step, which would reduce the training required to perform nanopore sequencing still further. Together these developments could allow the addition of DNA sequencing to any polar ocean research cruise, without the need for complex arrangements or a specialist researcher, which could result in a huge increase in DNA sequencing of polar ocean microbial communities.

5.5 Summary and Conclusion

Nanopore sequencing offers great potential for genomics research into non-model organisms, such as eukaryotic phytoplankton. The aims of this project were to produce a high quality genome assembly for the haptophyte *E. huxleyi* RCC1217 using nanopore long-read sequencing data to resolve complexities and improve understanding of the haploid phase of the *E. huxleyi* life cycle; to perform real-time *in situ* sequencing and analysis of ocean microbiomes; and to investigate

the potential of portable nanopore sequencing for citizen science and public engagement.

This has been successful, with a genome assembly which is of comparable quality to other recently published haptophyte genome assemblies, and shows significant improvement compared the only other publicly available *E. huxleyi* genome assembly, CCMP1516. *In situ* sequencing of ocean microbiomes with real-time analysis was successfully performed in 2019 onboard a research cruise in the Southern Ocean, for what is believed to be the first time. This provided an insight into the populations present in the location, and acted as a proof-of-concept for the sampling, DNA extraction, library preparation, and sequencing workflow in the field. A streamlined protocol optimised for use by citizen scientists and researchers with little training was developed and sucessfully used pilot study for ONT MinION sequencing at Cromer Pier investigating the ocean microbiome at the Norfolk coast. The resulting workflow describes a flexible approach to sampling, DNA extraction, and sequencing, allowing for different experimental goals, and is sufficiently simplified to allow a non-specialist researcher to perform *in situ* sequencing with minimal training, which could result in a wider use of nanopore sequencing in the field.

Nanopore sequencing is a relatively new and developing field with rapid improvements between the beginning of this project and its conclusion, which opens doors to exciting new developments and the prospect of real advances in phytoplankton genomics in the near future. Eukaryotic phytoplankton, particularly those in polar oceans are a hugely important part of the planet's climate and ecology, and are under significant threat from climate change, but they are currently poorly characterised and under studied. Advances in nanopore sequencing offer potential for increasingly high quality genome assemblies from eukaryotic phytoplankton, and improved metagenomic analysis of environmental samples. This will allow researchers to build improved models of ocean microbial communities, capturing their complex interactions and helping to advance understanding of their ecological and climactic contributions and impacts, as well as examine how they might be affected by climate change and how they might be protected.

References

Abreu, A., E. Bourgois, A. Gristwood, R. Troublé, D. Arendt, J. Bilic, R. Finn, E. Heard, B. Rouse, and J. Vamathevan (2022). “Priorities for ocean microbiome research”. In: *Nature Microbiology* 7.7, pp. 937–947.

Alexander, H., S. K. Hu, A. I. Krinos, M. Pachiadaki, B. J. Tully, C. J. Neely, and T. Reiter (2021). “Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton”. In: *bioRxiv*, pp. 2021–07.

Arakawa, K. (2016). “No Evidence for Extensive Horizontal Gene Transfer From the Draft Genome of a Tardigrade”. In: *Proceedings of the National Academy of Sciences* 113.22, nil. DOI: 10.1073/pnas.1602711113.

Bar-On, Y. M. and R. Milo (2019). “The biomass composition of the oceans: a blueprint of our blue planet”. In: *Cell* 179.7, pp. 1451–1454.

Belser, C., F.-C. Baurens, B. Noel, G. Martin, C. Cruaud, B. Istace, N. Yahiaoui, K. Labadie, E. Hřibová, J. Doležel, A. Lemainque, P. Wincker, A. D’Hont, and J.-M. Aury (2021). “Telomere-To-Telomere Gapless Chromosomes of Banana Using Nanopore Sequencing”. In: *Communications Biology* 4.1, p. 1047. DOI: 10.1038/s42003-021-02559-3.

Benham, P. M., C. Cicero, M. Escalona, E. Beraut, M. P. Marimuthu, O. Nguyen, M. W. Nachman, and R. C. Bowie (2023). “A highly contiguous genome assembly for the California quail (Callipepla californica)”. In: *Journal of Heredity* 114.4, pp. 418–427.

Bork, P., C. Bowler, C. De Vargas, G. Gorsky, E. Karsenti, and P. Wincker (2015). “Tara Oceans studies plankton at Planetary scale”. In: *Science* 348.6237, p. 873. ISSN: 10959203. DOI: 10.1126/science.aac5605.

Bowers, R. M., T. G. S. Consortium, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke (2017). “Minimum Information About a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea”. In: *Nature Biotechnology* 35.8, pp. 725–731. DOI: 10.1038/nbt.3893.

Boyd, P. W. (Oct. 2002). “Review of environmental factors controlling phytoplankton processes in the Southern Ocean”. In: *Journal of Phycology* 38.October 2001, pp. 844–861. ISSN: 0022-3646. DOI: 10.1046/j.1529-8817.2002.t01-1-01203.x.

Brum, J. R., J. C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. de Vargas, J. M. Gasol, G. Gorsky, A. C. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B. T. Poulos, S. M. Schwenck, S. Speich, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, T. O. Tara Oceans Coordinators, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, and M. B. Sullivan (May 2015). “Ocean plankton. Patterns and ecological drivers of ocean viral communities.” In: *Science* 348.6237, p. 1261498. ISSN: 1095-9203. DOI: 10.1126/science.1261498. arXiv: science.1261498 [10.1126].

Carradec, Q., E. Pelletier, C. Da Silva, A. Alberti, Y. Seeleuthner, R. Blanc-Mathieu, G. Lima-Mendez, F. Rocha, L. Tirichine, and K. Labadie (2018). “A global ocean atlas of eukaryotic genes”. In: *Nature communications* 9.1, pp. 1–13. ISSN: 2041-1723.

Castro, C. J. and T. F. F. Ng (2017). “U₅₀: a New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs”. In: *Journal of Computational Biology* 24.11, pp. 1071–1080. DOI: 10.1089/cmb.2017.0013.

Cavicchioli, R., W. J. Ripple, K. N. Timmis, F. Azam, L. R. Bakken, M. Baylis, M. J. Behrenfeld, A. Boetius, P. W. Boyd, A. T. Classen, et al. (2019). “Scientists’ warning to humanity: microorganisms and climate change”. In: *Nature Reviews Microbiology* 17.9, pp. 569–586.

Cornet, L. and D. Baurain (2022). “Contamination Detection in Genomic Data: More Is Not Enough”. In: *Genome Biology* 23.1, p. 60. DOI: 10.1186/s13059-022-02619-9.

Dassow, P. von, H. Ogata, I. Probert, P. Wincker, C. Da Silva, S. Audic, J.-M. Claverie, and C. de Vargas (Oct. 2009). “Transcriptome analysis of functional differentiation between haploid and diploid cells of Emiliania huxleyi, a globally significant photosynthetic calcifying cell”. In: *Genome Biology* 10.10, R114. ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-10-r114.

De Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukes, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, and E.

Karsenti (2015). “Eukaryotic plankton diversity in the sunlit ocean”. In: *Science* 348.6237, pp. 1261605–1/11. ISSN: 0717-6163. DOI: 10.1007/s13398-014-0173-7.2. arXiv: 9809069v1 [arXiv:gr-qc].

Delmont, T. O., M. Gaia, D. D. Hinsinger, P. Frémont, C. Vanni, A. Fernandez-Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. D. Silva, M. Wessner, B. Noel, J.-M. Aury, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, S. Sunagawa, S. G. Acinas, P. Bork, E. Karsenti, C. Bowler, C. Sardet, L. Stemmann, C. de Vargas, P. Wincker, M. Lescot, M. Babin, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, O. Jaillon, S. Kandels, D. Iudicone, H. Ogata, S. Pesant, M. B. Sullivan, F. Not, K.-B. Lee, E. Boss, G. Cochrane, M. Follows, N. Poulton, J. Raes, M. Sieracki, and S. Speich (2022). “Functional Repertoire Convergence of Distantly Related Eukaryotic Plankton Lineages Abundant in the Sunlit Ocean”. In: *Cell Genomics* 2.5, p. 100123. DOI: 10.1016/j.xgen.2022.100123.

Duncan, A., K. Barry, C. Daum, E. Eloe-Fadrosh, S. Roux, S. G. Tringe, K. Schmidt, K. U. Valentin, N. Varghese, I. V. Grigoriev, R. Leggett, V. Moulton, and T. Mock (2020). *Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle*. DOI: 10.1101/2020.06.16.154583.

Ferguson, D. K., C. Li, A. Chakraborty, D. A. Gittins, M. Fowler, J. Webb, C. Campbell, N. Morrison, A. MacDonald, and C. R. Hubert (2023). “Multi-year seabed environmental baseline in deep-sea offshore oil prospective areas established using microbial biodiversity”. In: *Marine Pollution Bulletin* 194, p. 115308.

Goto, Y., R. Akahori, I. Yanagi, and K.-i. Takeda (2020). “Solid-state nanopores towards single-molecule DNA sequencing”. In: *Journal of human genetics* 65.1, pp. 69–77.

Harvey, E. L., R. W. Deering, D. C. Rowley, A. El Gamal, M. Schorn, B. S. Moore, M. D. Johnson, T. J. Mincer, and K. E. Whalen (2016). “A bacterial quorum-sensing precursor induces mortality in the marine coccolithophore, Emiliania huxleyi”. In: *Frontiers in Microbiology* 7, p. 59.

Henson, S. A., B. Cael, S. R. Allen, and S. Dutkiewicz (2021). “Future phytoplankton diversity in a changing climate”. In: *Nature communications* 12.1, p. 5372.

Joli, N., A. Monier, R. Logares, and C. Lovejoy (2017). “Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome”. In: *The ISME Journal* 11.6, pp. 1372–1385.

Katz, M. E., Z. V. Finkel, D. Grzebyk, A. H. Knoll, and P. G. Falkowski (2004). “Evolutionary trajectories and biogeochemical impacts of marine eukaryotic phytoplankton”. In: *Annu. Rev. Ecol. Evol. Syst.* 35, pp. 523–556.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy (2017). “Canu: scalable and accurate long-read assembly via adaptive k-mer

weighting and repeat separation". In: *Genome research* 27.5, pp. 722–736. DOI: 10.1101/gr.215087.116.

Lang, D., S. Zhang, P. Ren, F. Liang, Z. Sun, G. Meng, Y. Tan, X. Li, Q. Lai, L. Han, et al. (2020). "Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore". In: *Gigascience* 9.12, giaa123.

Lapidus, A. L. and A. I. Korobeynikov (2021). "Metagenomic data assembly–the way of decoding unknown microorganisms". In: *Frontiers in Microbiology* 12, p. 613791.

Leggett, R. M., C. Alcon-Giner, D. Heavens, S. Caim, T. C. Brook, M. Kujawska, S. Martin, L. Hoyles, P. Clarke, L. J. Hall, and M. D. Clark (2018). "Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics". In: *bioRxiv*. DOI: 10.1101/180406. eprint: https://www.biorxiv.org/content/early/2018/10/12/180406.full.pdf.

Lewin, H. A., S. Richards, E. L. Aiden, M. L. Allende, J. M. Archibald, M. Bálint, K. B. Barker, B. Baumgartner, K. Belov, G. Bertorelle, M. L. Blaxter, J. Cai, N. D. Caperello, K. Carlson, J. C. Castilla-Rubio, S.-M. Chaw, L. Chen, A. K. Childers, J. A. Coddington, D. A. Conde, M. Corominas, K. A. Crandall, A. J. Crawford, F. DiPalma, R. Durbin, T. E. Ebenezer, S. V. Edwards, O. Fedrigo, P. Flicek, G. Formenti, R. A. Gibbs, M. T. P. Gilbert, M. M. Goldstein, J. M. Graves, H. T. Greely, I. V. Grigoriev, K. J. Hackett, N. Hall, D. Haussler, K. M. Helgen, C. J. Hogg, S. Isobe, K. S. Jakobsen, A. Janke, E. D. Jarvis, W. E. Johnson, S. J. M. Jones, E. K. Karlsson, P. J. Kersey, J.-H. Kim, W. J. Kress, S. Kuraku, M. K. N. Lawniczak, J. H. Leebens-Mack, X. Li, K. Lindblad-Toh, X. Liu, J. V. Lopez, T. Marques-Bonet, S. Mazard, J. A. K. Mazet, C. J. Mazzoni, E. W. Myers, R. J. O'Neill, S. Paez, H. Park, G. E. Robinson, C. Roquet, O. A. Ryder, J. S. M. Sabir, H. B. Shaffer, T. M. Shank, J. S. Sherkow, P. S. Soltis, B. Tang, L. Tedersoo, M. Uliano-Silva, K. Wang, X. Wei, R. Wetzer, J. L. Wilson, X. Xu, H. Yang, A. D. Yoder, and G. Zhang (2022). "The Earth Biogenome Project 2020: Starting the Clock". In: *Proceedings of the National Academy of Sciences* 119.4, nil. DOI: 10.1073/pnas.2115635118.

Lin, S. (2011). "Genomic understanding of dinoflagellates". In: *Research in microbiology* 162.6, pp. 551–569.

Liu, L., Y. Yang, Y. Deng, and T. Zhang (2022). "Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes". In: *Microbiome* 10.1, pp. 1–7.

Maestri, Cosentino, Paterno, Freitag, Garces, Marcolungo, Alfano, Njunjić, Schilthuizen, Slik, Menegon, Rossato, and Delledonne (2019). "A Rapid and Accurate Minion-Based Workflow for Tracking Species Biodiversity in the Field". In: *Genes* 10.6, p. 468. DOI: 10.3390/genes10060468.

Malviya, S., E. Scalco, S. Audic, F. Vincent, A. Veluchamy, J. Poulain, P. Wincker, D. Iudicone, C. de Vargas, L. Bittner, A. Zingone, and C. Bowler (Mar. 2016). "Insights into global diatom distribution and diversity in the world's ocean." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.11, E1516–25. ISSN: 1091-6490. DOI: 10.1073/pnas.1509523113.

Manni, M., M. R. Berkeley, M. Seppey, and E. M. Zdobnov (2021). "Busco: Assessing Genomic Data Quality and Beyond". In: *Current Protocols* 1.12, nil. DOI: 10.1002/cpz1.323.

Meyer, R., M. Ramos, M. Lin, T. Schweizer, Z. Gold, D. Ramos, S. Shirazi, G. Kandlikar, W. Kwan, E. Curd, et al. (2021). "The CALeDNA program: Citizen scientists and researchers inventory California's biodiversity". In: *California Agriculture* 75.1, pp. 20–32.

Miga, K. H., S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T. A. G. Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy (2020). "Telomere-To-Telomere Assembly of a Complete Human X Chromosome". In: *Nature* 585.7823, pp. 79–84. DOI: 10.1038/s41586-020-2547-7.

Nurk, S. et al. (2022). "The Complete Sequence of a Human Genome". In: *Science* 376.6588, pp. 44–53. DOI: 10.1126/science.abj6987.

Obiol, A., C. R. Giner, P. Sánchez, C. M. Duarte, S. G. Acinas, and R. Massana (2020). "A metagenomic assessment of microbial eukaryotic diversity in the global ocean". In: *Molecular Ecology Resources* n/a, pp. 1–14. DOI: 10.1111/1755-0998.13147. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13147.

Orellana, L. H., K. Krüger, C. Sidhu, and R. Amann (2023). "Comparing genomes recovered from time-series metagenomes using long-and short-read sequencing technologies". In: *Microbiome* 11.1, p. 105.

Paasche, E. (2001). "A Review of the Coccolithophorid Emiliania Huxleyi (Prymnesiophyceae), With Particular Reference To Growth, Coccolith Formation, and Calcification-Photosynthesis Interactions". In: *Phycologia* 40.6, pp. 503–529. DOI: 10.2216/i0031-8884-40-6-503.1.

Patin, N. and K. Goodwin (2022). "Long-Read Sequencing Improves Recovery of Picoeukaryotic Genomes and Zooplankton Marker Genes from Marine Metagenomes". In: *Msystems* 7.6, e00595–22.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain (2011). "Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough". In: *PLoS Biology* 9.3, e1000602. DOI: 10.1371/journal.pbio.1000602.

Quick, J., N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, et al. (2016). "Real-time, portable genome sequencing for Ebola surveillance". In: *Nature* 530.7589, p. 228. DOI: 10.1038/nature16996.

Read, B. A., E. huxleyi Annotation Consortium, J. Kegel, M. J. Klute, A. Kuo, S. C. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J.-M. Claverie, S. Frickenhaus, K. Gonzalez, E. K. Herman, Y.-C. Lin, J. Napier, H. Ogata, A. F. Sarno, J. Shmutz, D. Schroeder, C. de Vargas, F. Verret, P. von Dassow, K. Valentin, Y. V. de Peer, G. Wheeler, J. B. Dacks, C. F. Delwiche, S. T. Dyhrman, G. Glöckner, U. John, T. Richards, A. Z. Worden, X. Zhang, and I. V. Grigoriev (2013). "Pan Genome of the Phytoplankton Emiliania Underpins Its Global Distribution". In: *Nature* 499.7457, pp. 209–213. DOI: 10.1038/nature12221.

Rhie, A. et al. (2021). "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species". In: *Nature* 592.7856, pp. 737–746. DOI: 10.1038/s41586-021-03451-0.

Rokitta, S. D., P. Von Dassow, B. Rost, and U. John (2014). "Emiliania huxleyi endures N-limitation with an efficient metabolic budgeting and effective ATP synthesis". In: *BMC genomics* 15.1, pp. 1–14.

Royo-Llonch, M., P. Sánchez, C. Ruiz-González, G. Salazar, C. Pedrós-Alió, M. Sebastián, K. Labadie, L. Paoli, F. M. Ibarbalz, L. Zinger, B. Churcheward, M. Babin, P. Bork, E. Boss, G. Cochrane, C. de Vargas, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Setmmann, M. B. Sullivan, S. Chaffron, D. Eveillard, E. Karsenti, S. Sunagawa, P. Wincker, L. Karp-Boss, C. Bowler, S. G. Acinas, and T. O. Coordinators (2021). "Compendium of 530 Metagenome-Assembled Bacterial and Archaeal Genomes From the Polar Arctic Ocean". In: *Nature Microbiology* 6.12, pp. 1561–1574. DOI: 10.1038/s41564-021-00979-9.

Ruan, Z., M. Lu, H. Lin, S. Chen, P. Li, W. Chen, H. Xu, and D. Qiu (2023). "Different photosynthetic responses of haploid and diploid Emiliania huxleyi (Prymnesiophyceae) to high light and ultraviolet radiation". In: *Bioresources and Bioprocessing* 10.1, p. 40.

Samarakoon, H., S. Punchihewa, A. Senanayake, J. M. Hammond, I. Stevanovski, J. M. Ferguson, R. Ragel, H. Gamaarachchi, and I. W. Deveson (2020). "Genopo: a Nanopore Sequencing Analysis Toolkit for Portable

Android Devices". In: *Communications Biology* 3.1, p. 538. DOI: 10.1038/s42003-020-01270-z.

Schierwater, B., S.-O. Kolokotronis, M. Eitel, and R. DeSalle (2009). "The Diploblast-Bilateria Sister Hypothesis". In: *Communicative & Integrative Biology* 2.5, pp. 403–405. DOI: 10.4161/cib.2.5.8763.

Shishlyannikov, S. M., Y. R. Zakharova, N. A. Volokitina, I. S. Mikhailov, D. P. Petrova, and Y. V. Likhoshway (2011). "A Procedure for Establishing an Axenic Culture of the Diatomsynedra acussubsp.radians(Kütz.) Skabibitsch. From Lake Baikal". In: *Limnology and Oceanography: Methods* 9.10, pp. 478–484. DOI: 10.4319/lom.2011.9.478.

Smetacek, V. and S. Nicol (Sept. 2005). "Polar ocean ecosystems in a changing world". In: *Nature* 437.7057, pp. 362–368. ISSN: 1476-4687. DOI: 10.1038/nature04161.

Sturm, M., C. Schroeder, and P. Bauer (2016). "Seqpurge: Highly-Sensitive Adapter Trimming for Paired-End Ngs Data". In: *BMC Bioinformatics* 17.1, p. 208. DOI: 10.1186/s12859-016-1069-7.

Tamura, K., M. Sakamoto, Y. Tanizawa, T. Mochizuki, S. Matsushita, Y. Kato, T. Ishikawa, K. Okuhara, Y. Nakamura, and H. Bono (2023). "A highly contiguous genome assembly of red perilla (Perilla frutescens) domesticated in Japan". In: *DNA Research* 30.1, dsac044.

Vázquez-Castellanos, J. F., R. García-López, V. Pérez-Brocal, M. Pignatelli, and A. Moya (2014). "Comparison of Different Assembly and Annotation Tools on Analysis of Simulated Viral Metagenomic Communities in the Gut". In: *BMC Genomics* 15.1, p. 37. DOI: 10.1186/1471-2164-15-37.

Vries, J. de, F. Monteiro, G. Wheeler, A. Poulton, J. Godrijan, F. Cerino, E. Malinverno, G. Langer, and C. Brownlee (2021). "Haplo-diplontic life cycle expands coccolithophore niche". In: *Biogeosciences* 18.3, pp. 1161–1184.

West, P. T., A. J. Probst, I. V. Grigoriev, B. C. Thomas, and J. F. Banfield (2018). "Genome-reconstruction for eukaryotes from complex natural microbial communities". In: *Genome research* 28.4, pp. 569–580.

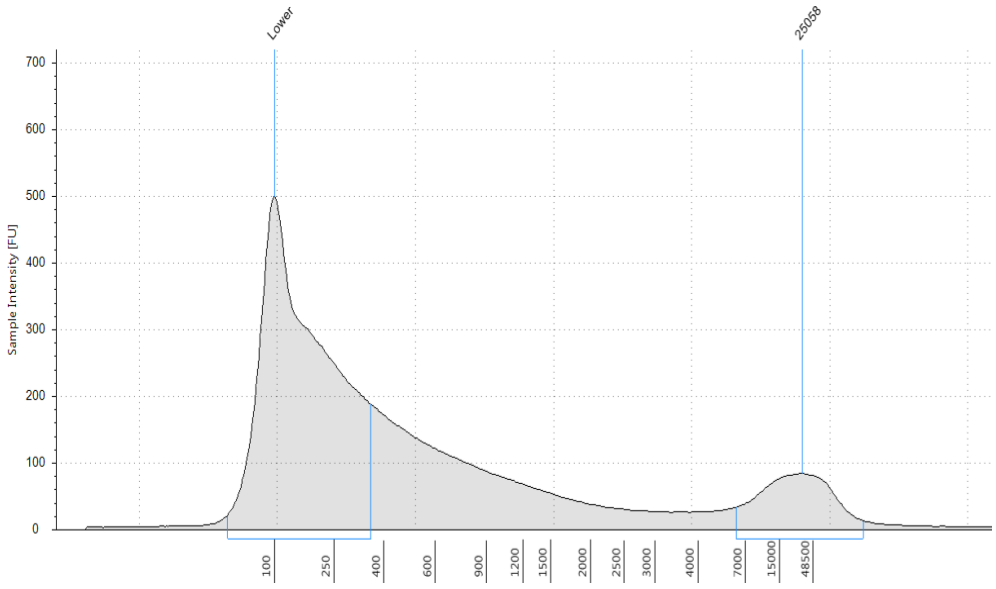
Winter, A., J. Henderiks, L. Beaufort, R. E. M. Rickaby, and C. W. Brown (Mar. 2014). "Poleward expansion of the coccolithophore Emiliania huxleyi". In: *Journal of Plankton Research* 36.2, pp. 316–325. ISSN: 14643774. DOI: 10.1093/plankt/fbt110.

A

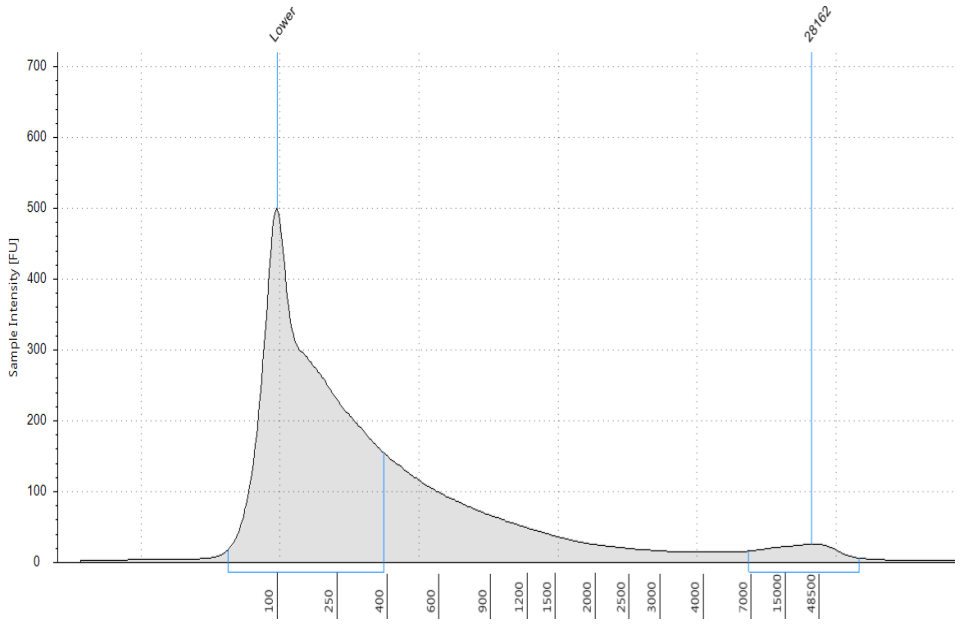
TapeStation output from DNA extractions

Agilent TapeStation output for station 5. Shows the molecular weight of the DNA against the sample intensity (FU), giving a visualisation of the molecular weight distribution of the DNA fragments in the sample.

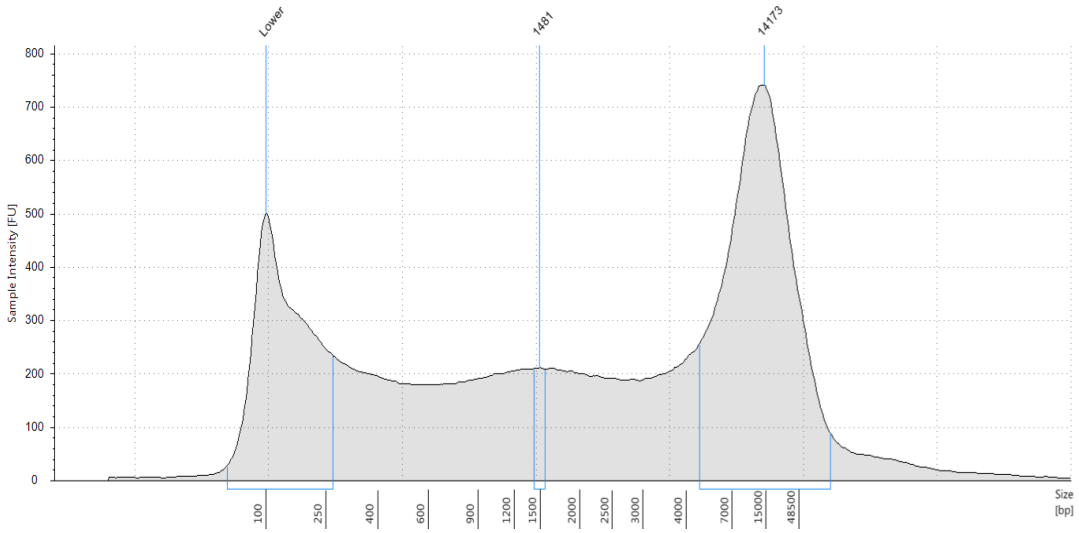
Sample 1



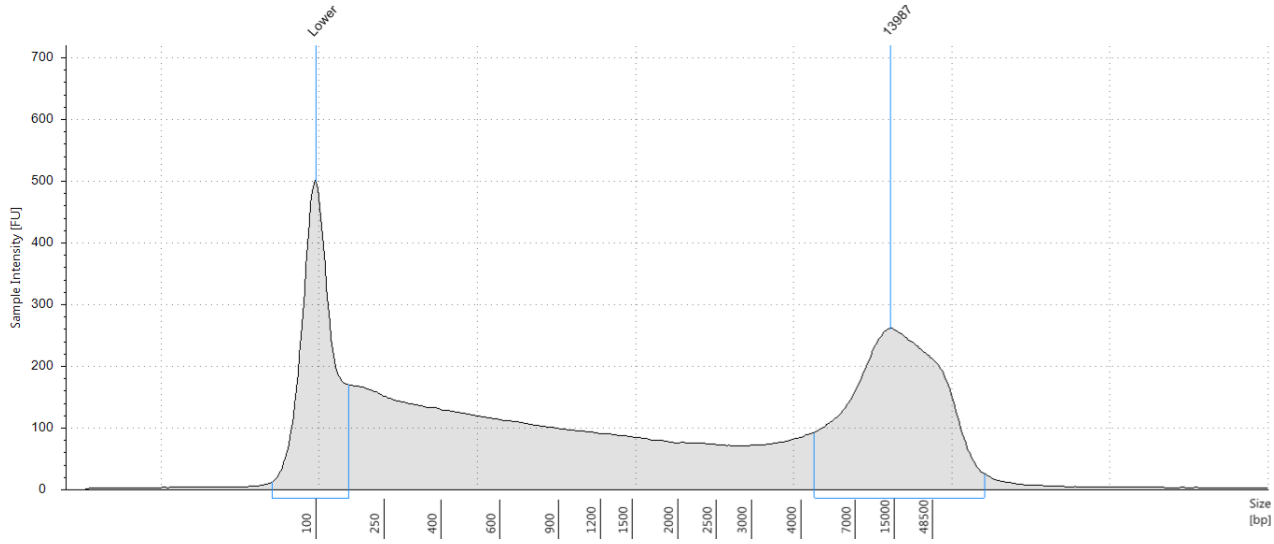
Sample 2



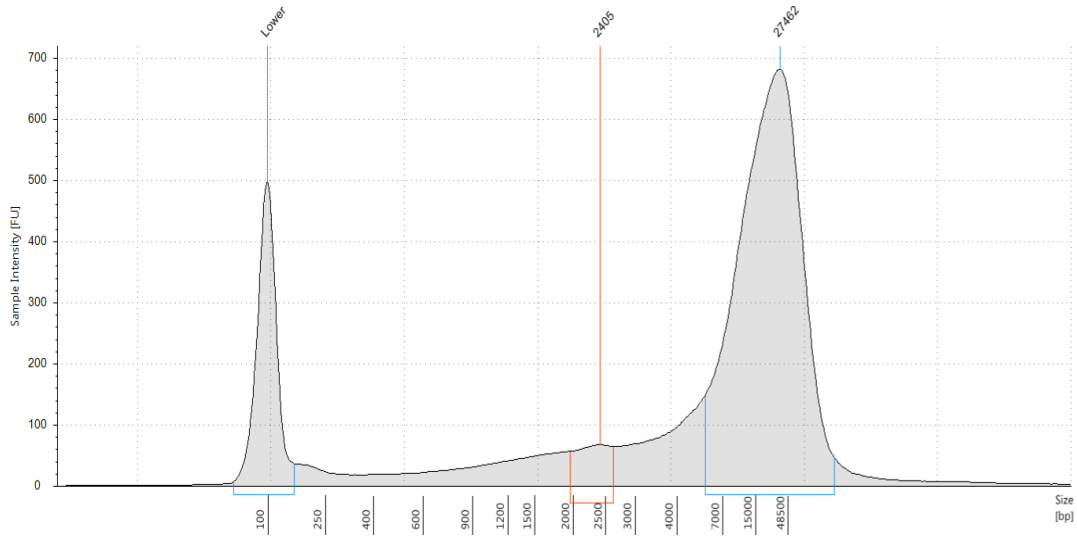
Sample 3



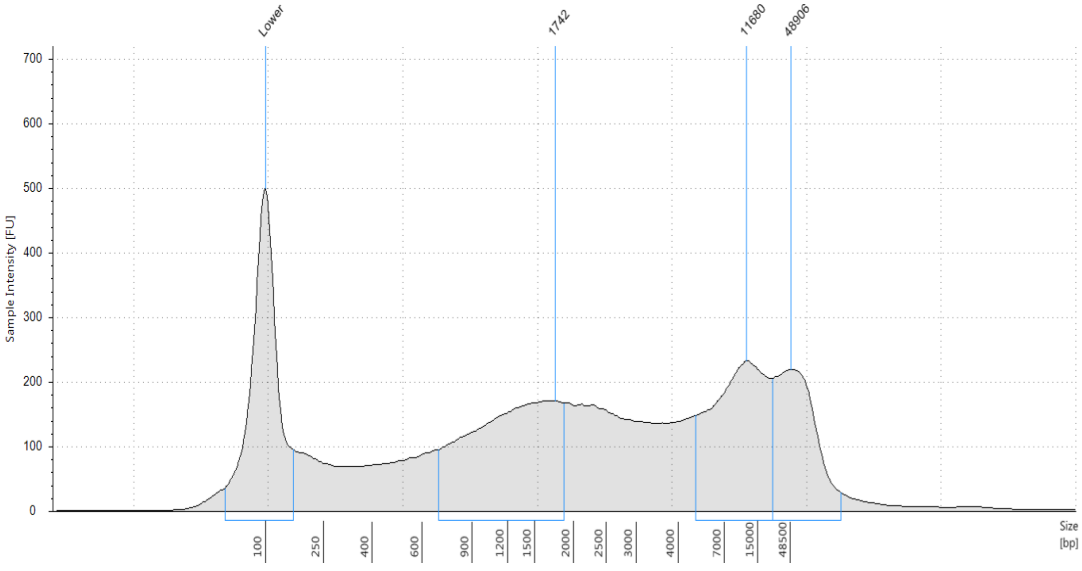
Sample 4



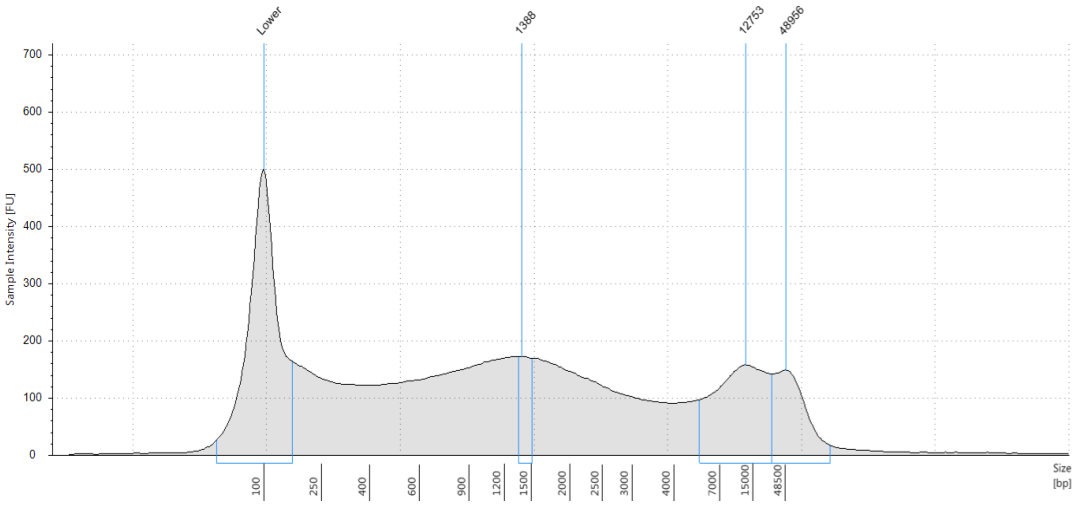
Sample 5



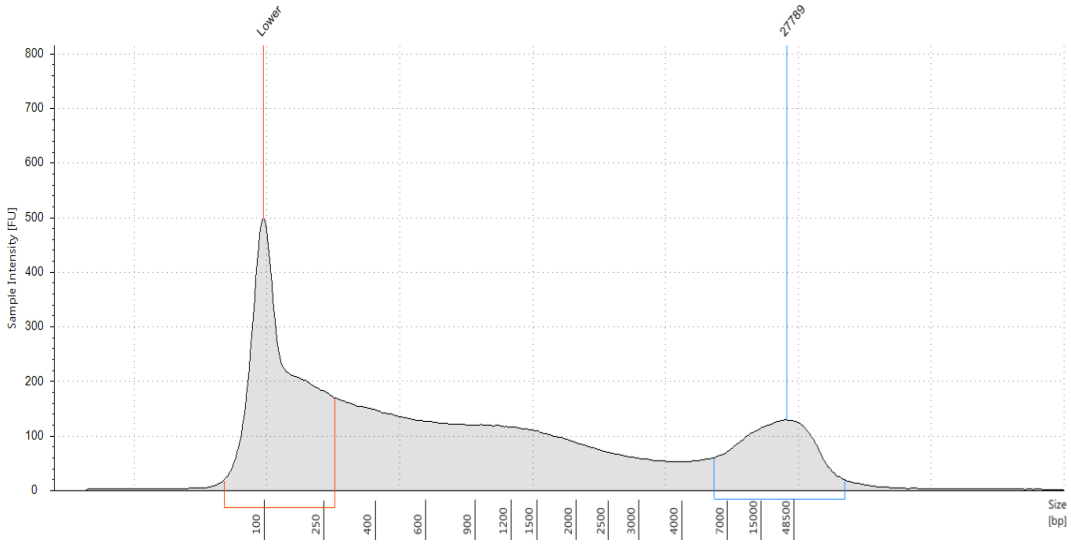
Sample 6



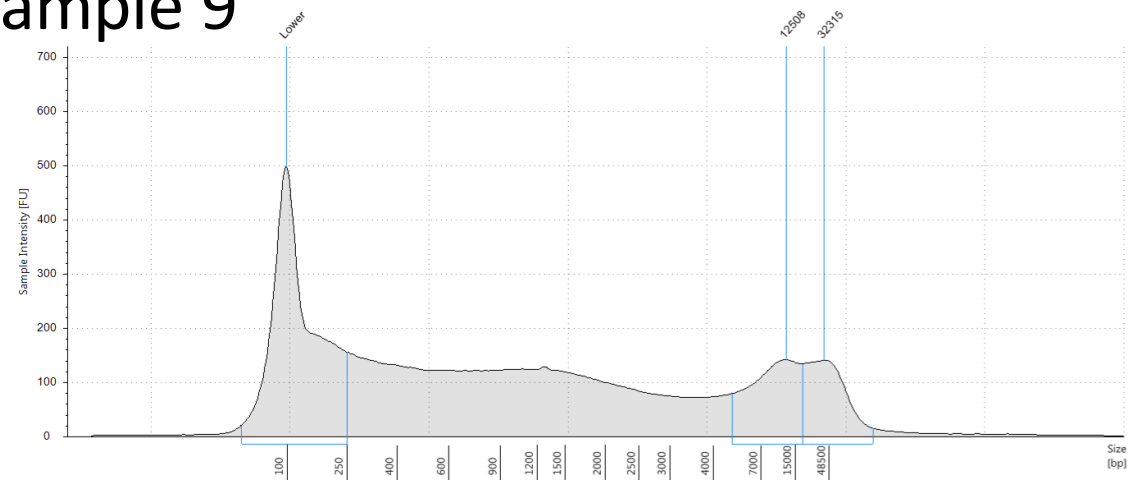
Sample 7



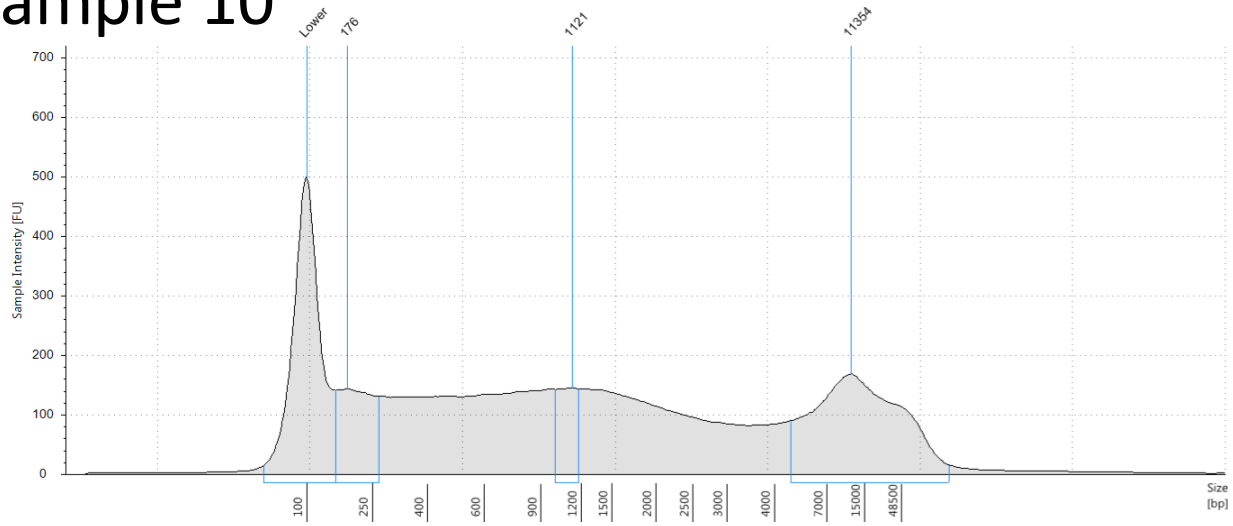
Sample 8



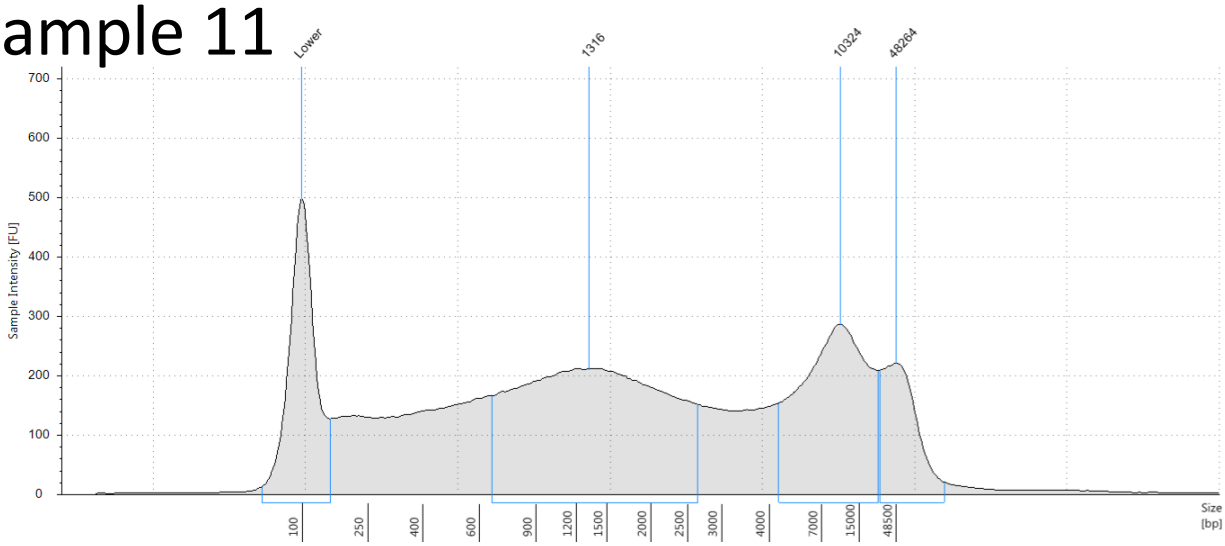
Sample 9



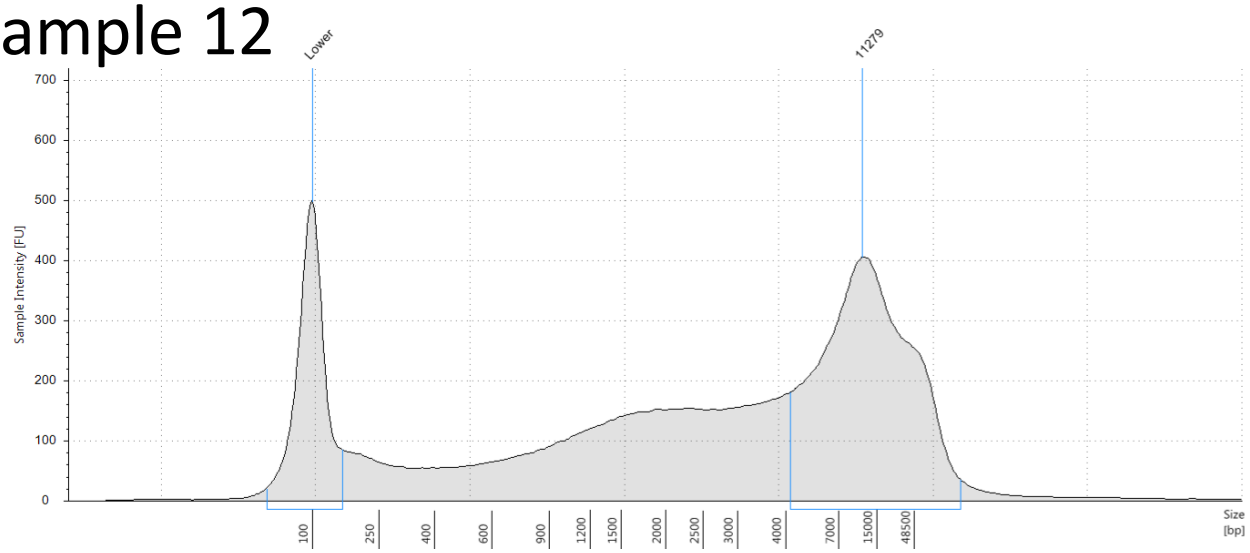
Sample 10



Sample 11



Sample 12



B

Metadata

CTD sensors: Temperature, Conductivity, Digiquartz Pressure, Dissolved O2 , Fluorimeter, Altimeter, UWIRR PAR, DWIRR PAR, Backscatter, Transmissometer, 20L water samplers, LADCP.

The particulate and dissolved organic matter and nutrient collection and analysis were carried out by Flavia Saccomandi and Cecilia Silvestri of ISPRA (instituto Italiano per Le Risorse ambientali) during the DY098 research cruise.

POC and TPN: For particulate organic carbon (POC) and total particulate nitrogen (TPN), 5L were filtered onto a pre-combusted 47-mm Whatman GF/F filters. After collection, filters were stored at -20 ℃ until analysing. In Italy filters were dried in an oven (60 ℃) for 24 h. Filters were then cut into two parts; one of this was treated by acidification with 1 N HCl until to completely remove carbonates and then re-placed in the oven for 2 h and finally POC determined by a CHN Elemental Analyzer Flash 2000 (Thermo Scientific). TPN was directly determined by CHN Elemental Analyzer Flash 2000 (Thermo Scientific) on the no acidified filter parts. Reference material BCSS (NRC, Canada) was used to assess the accuracy of analytical data.

DON: Sampled seawater was filtered using GF-F filters (0.7 μm), frozen immediately, and analysed in triplicate for dissolved inorganic nitrogen (ammonium, nitrate, nitrite) using a flow injection auto-analyser with a minimum nitrate detection limit of 0.2 μM/L-1.

Si: Si concentrations were analysed on a Thermo iCAP6300 Duo-ICP with a detection limit of 0.003 μM.

The results from these experiments were combined with measurements from the CTD sensors, provided by BAS Polar data Center, to produce a metadata table which was used for analysis of nanopore sequencing data in section 3.3.3- see table B.1

Table B.1: Table showing the metadata collected from the CTD sensors, and by Flavia Saccomandi and Cecilia Silvestri on the DY098 research cruise for each of the sampled stations. Station relates to table 3.1. Depth (m); Si (μM); PO4, NO3, NO2, NH4 μM/L-1; POC, DOC, TPN (μM); Salinity (g/L), Temperature ℃

Station	Depth	Si	PO4	NO3	NO2	NH4	POC	TPN	DOC	Sal	Temp
1	33.8	13.65	1.55	20.89	0.29	0.68	19.6	3.0	2.501	33.6	4.404
2	53.2	10.75	1.45	18.45	0.31	0.74	22.3	3.5	2.567	33.9	1.465
3	22.9	10.26	1.92	22.69	0.25	0.89	6.6	0.7	1.993	33.8	3.486
4	7.9	14.05	1.67	21.01	0.24	0.92	7.4	1.6	2.575	33.1	3.200
5	78.4	16.32	1.93	21.23	0.29	0.54	20.8	4.2	0	33.9	1.461
6	63.5	15.34	1.95	20.34	0.27	0.56	9.0	1.2	0	33.9	0.484
7	41.7	13.35	1.34	19.76	0.27	0.56	13.4	2.9	2.596	33.8	0.934
8	21.7	12.33	1.32	20.32	0.25	0.88	16.4	2.5	0	33.6	0.820
9	28.5	14.45	1.45	22.56	0.27	0.58	17.1	3.5	1.876	33.9	2.893
10	33.3	11.33	1.87	21.31	0.2	0.9	16.5	3.4	2.708	33.8	3.300
11	63.2	16.43	1.89	23.45	0.26	0.93	7.3	1.6	2.764	33.9	0.595
12	59.4	16.52	1.9	22.33	0.23	0.78	6.0	0.5	0	33.9	-0.105

C

Genus level rarefaction curve

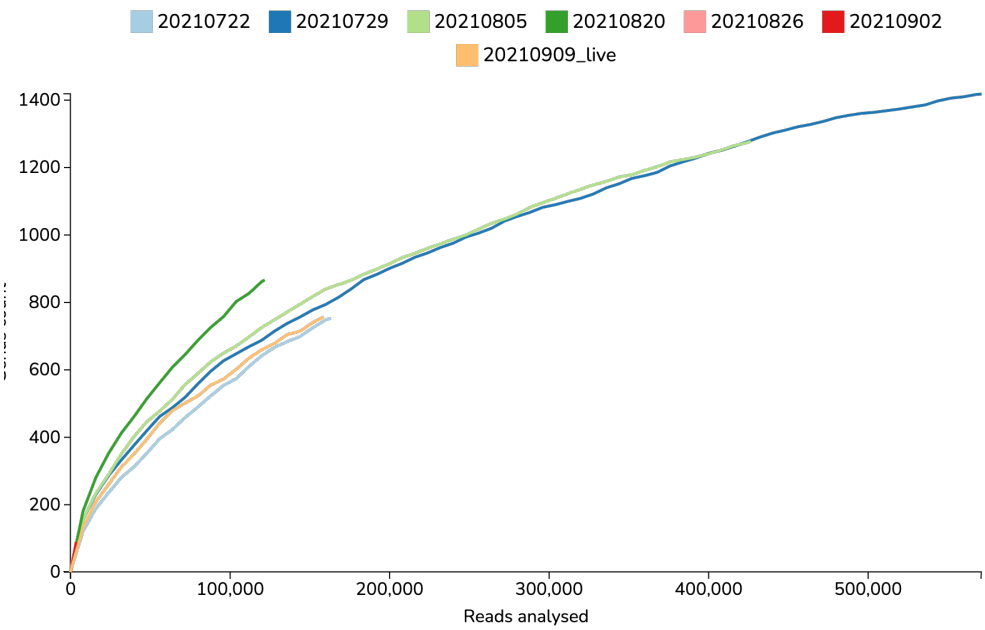


Figure C.0.1: Taxa accumulation curve at genus level, showing genera found against reads analysed. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

D

Eukaryotic treemaps

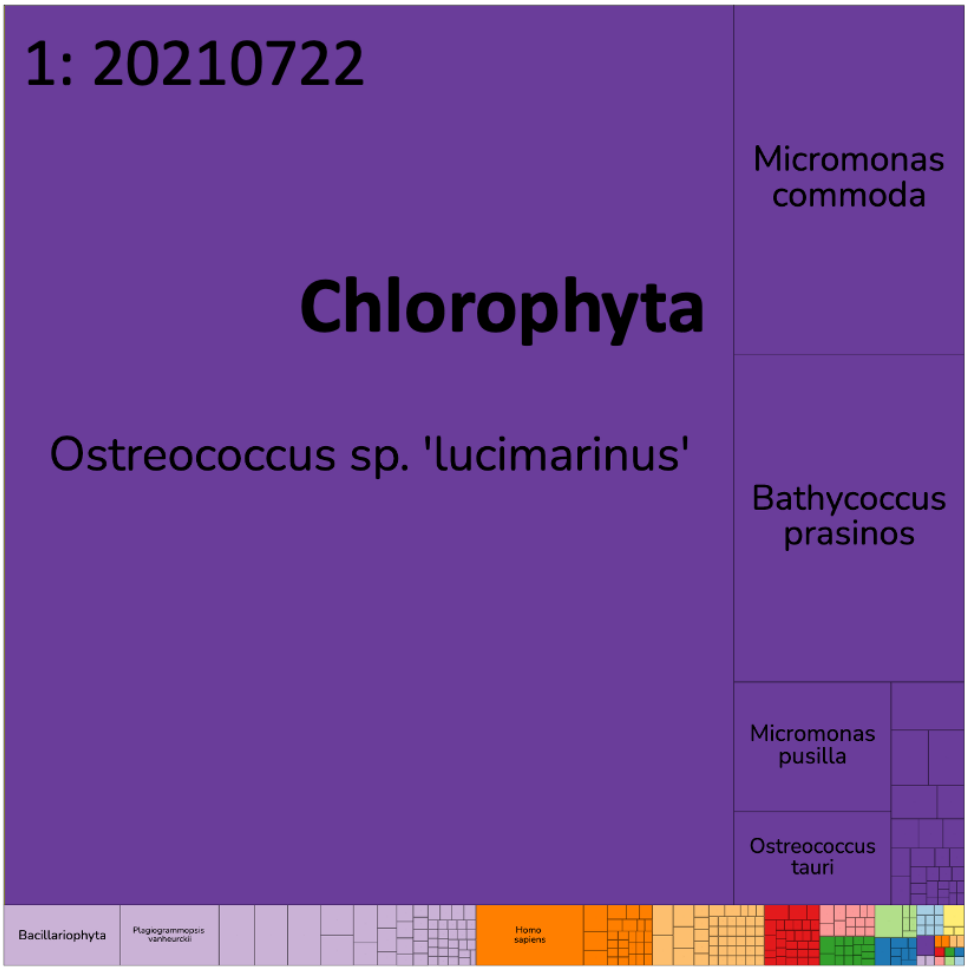


Figure D.0.1: Treemap at genus level showing eukaryotic genera identified in sample 1.

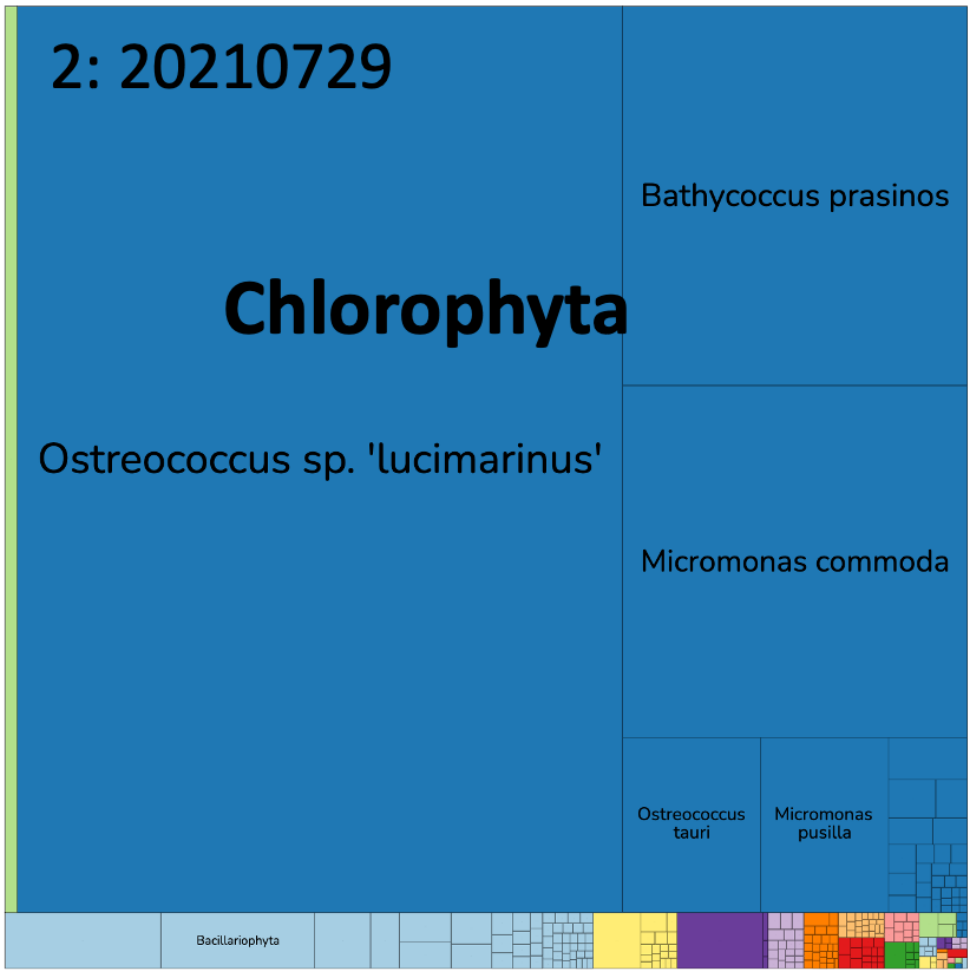


Figure D.0.2: Treemap at genus level showing eukaryotic genera identified in sample 2.

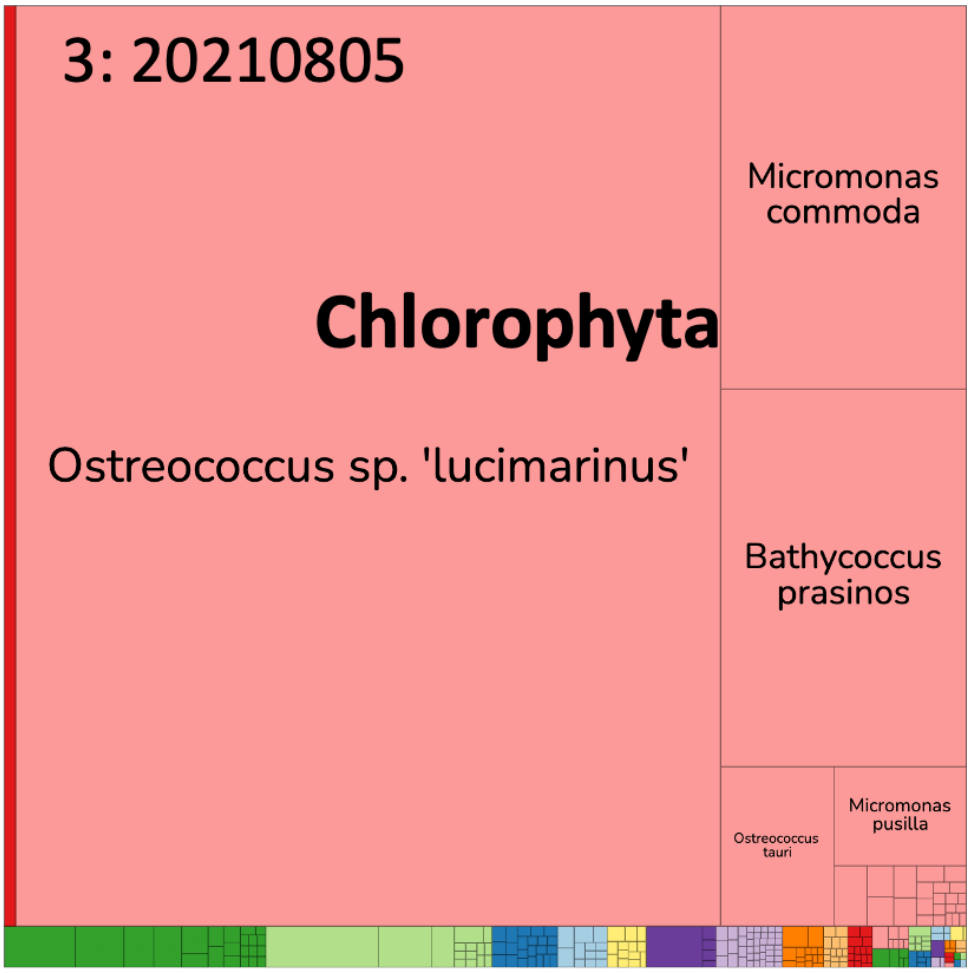


Figure D.0.3: Treemap at genus level showing eukaryotic genera identified in sample 3.

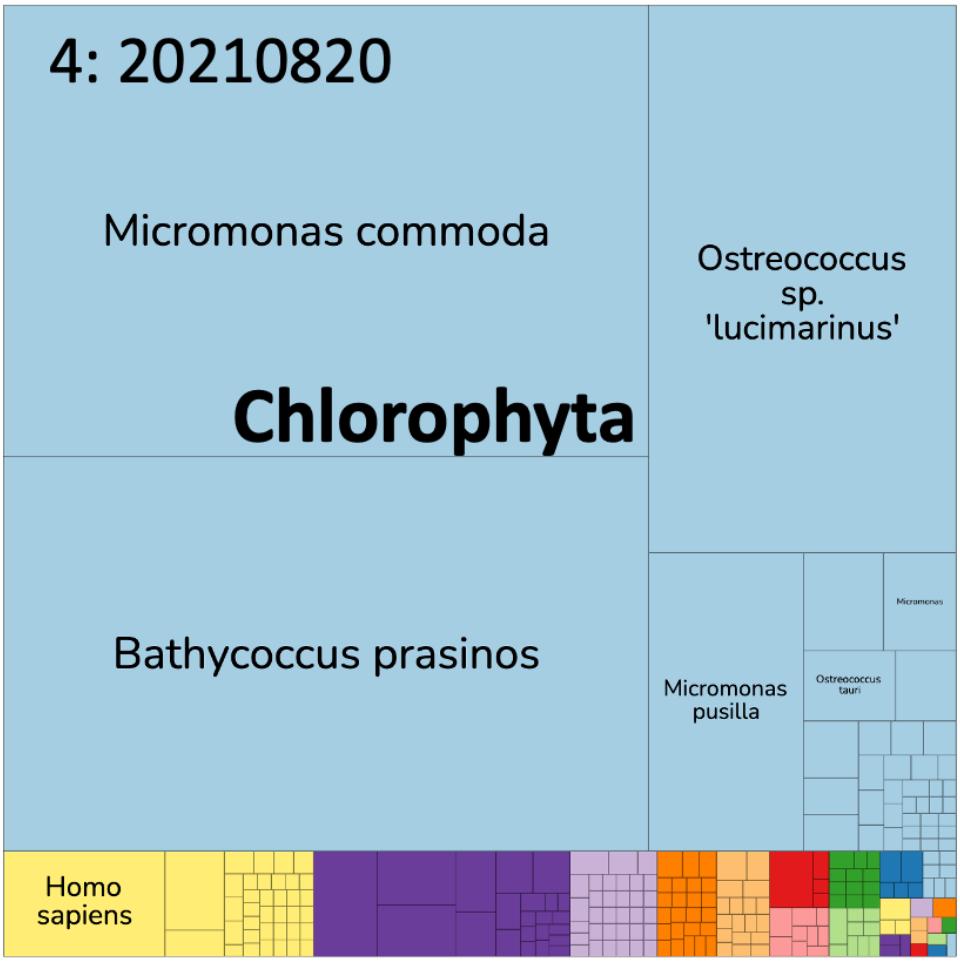


Figure D.0.4: Treemap at genus level showing eukaryotic genera identified in sample 4.

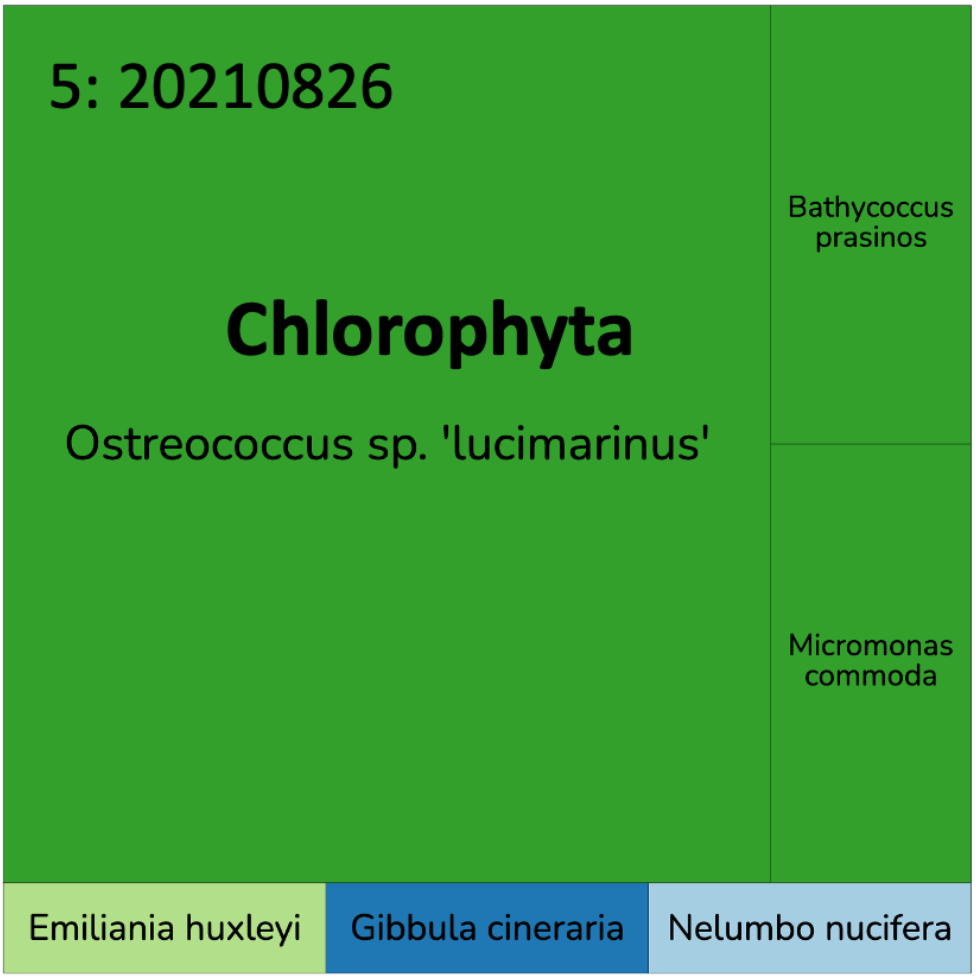


Figure D.0.5: Treemap at genus level showing eukaryotic genera identified in sample 5.

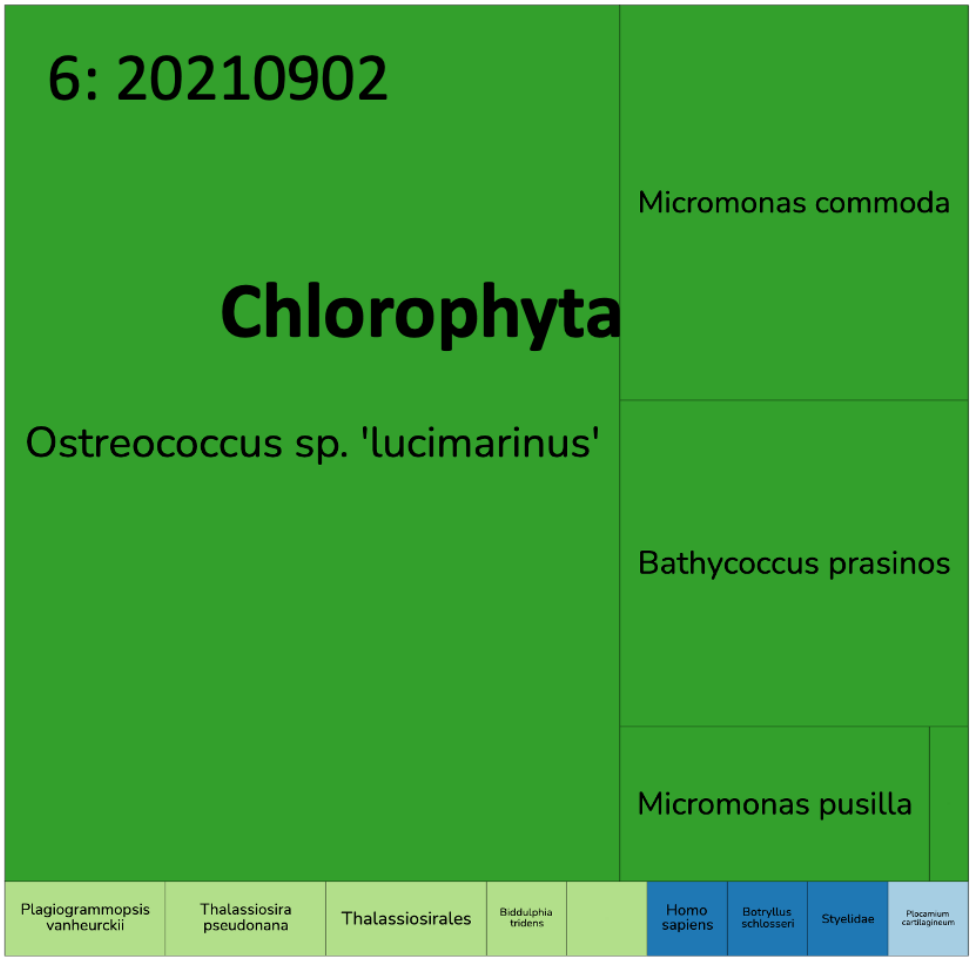


Figure D.0.6: Treemap at genus level showing eukaryotic genera identified in sample 6.

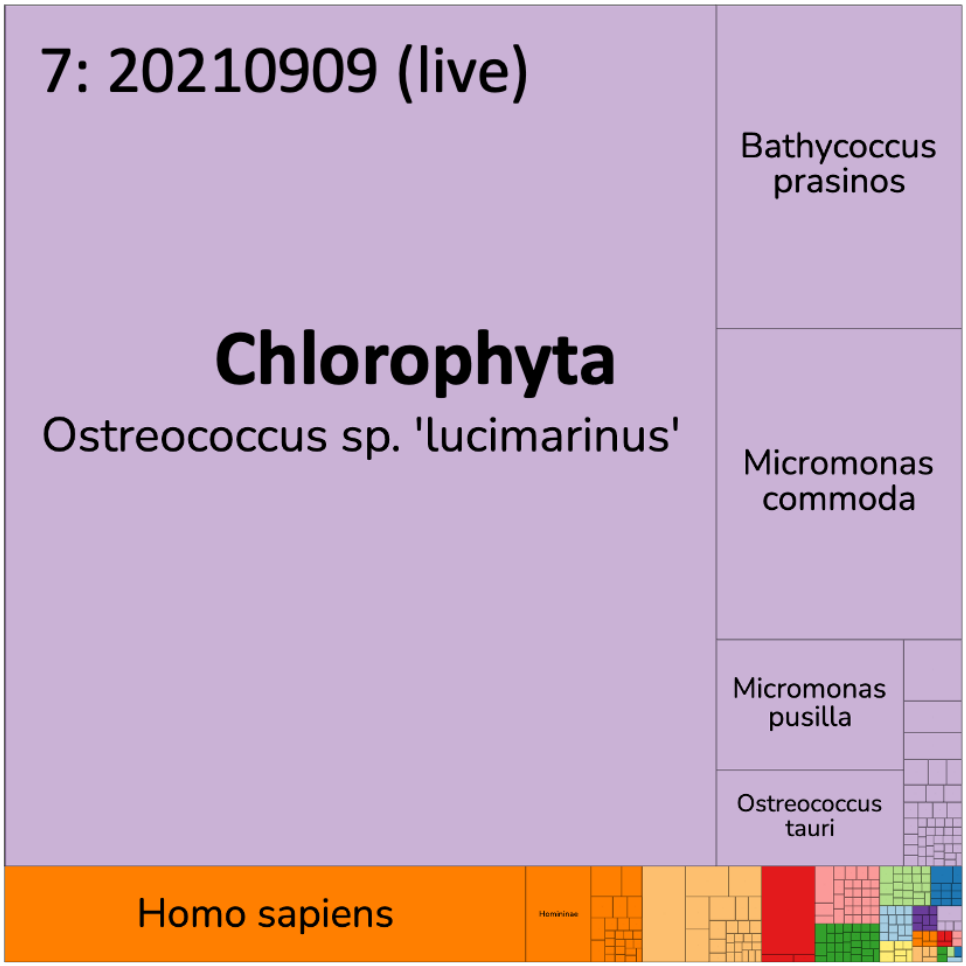


Figure D.0.7: Treemap at genus level showing eukaryotic genera identified in sample 7.

E

Prokaryotic treemaps

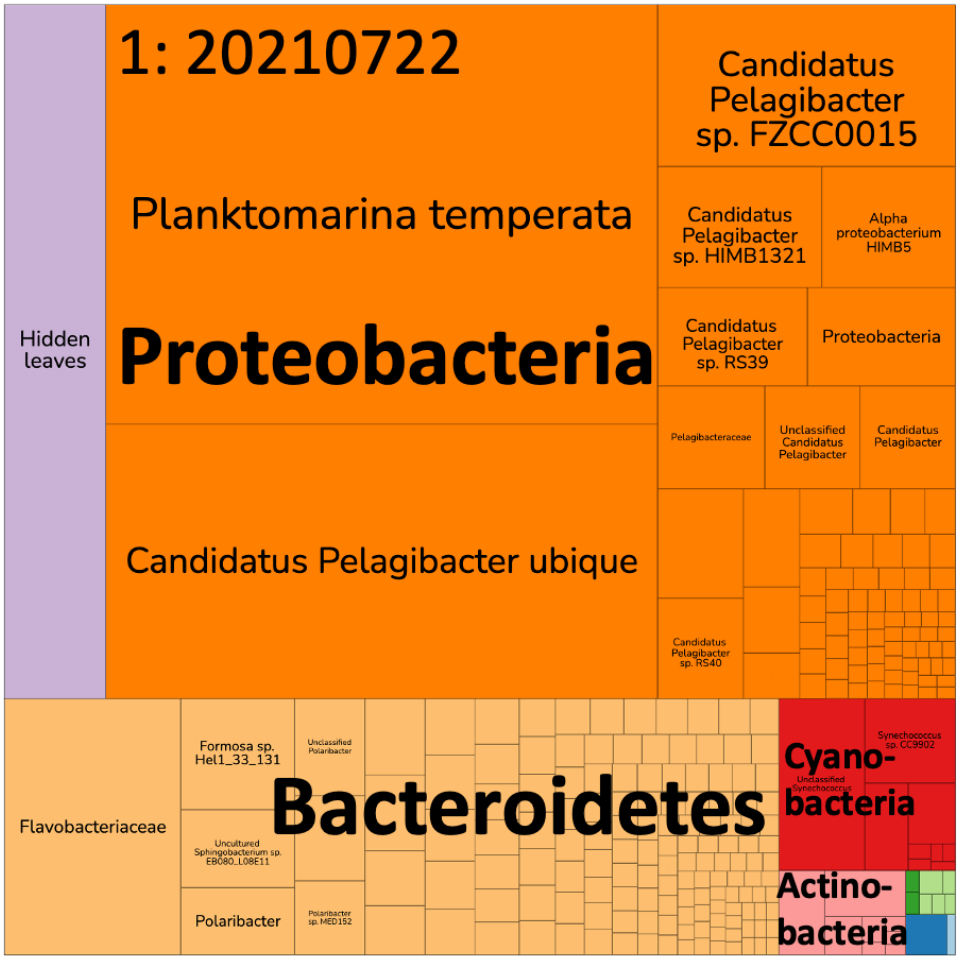


Figure E.0.1: Treemap at genus level showing prokaryotic genera identified in sample 1.

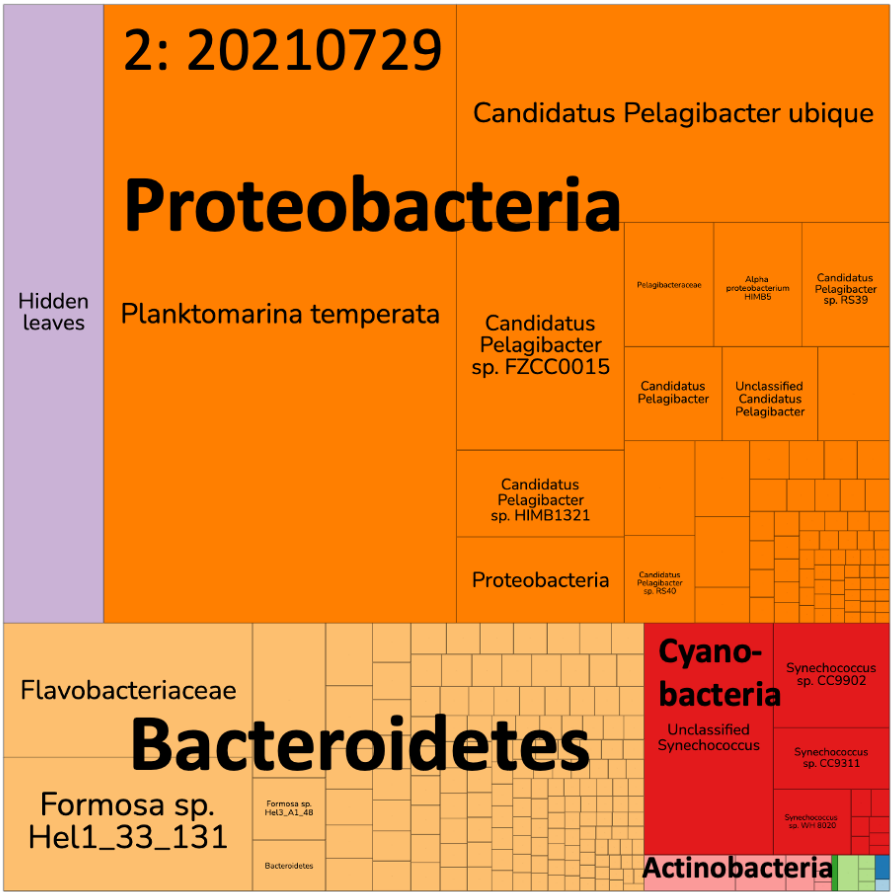


Figure E.0.2: Treemap at genus level showing prokaryotic genera identified in sample 2.

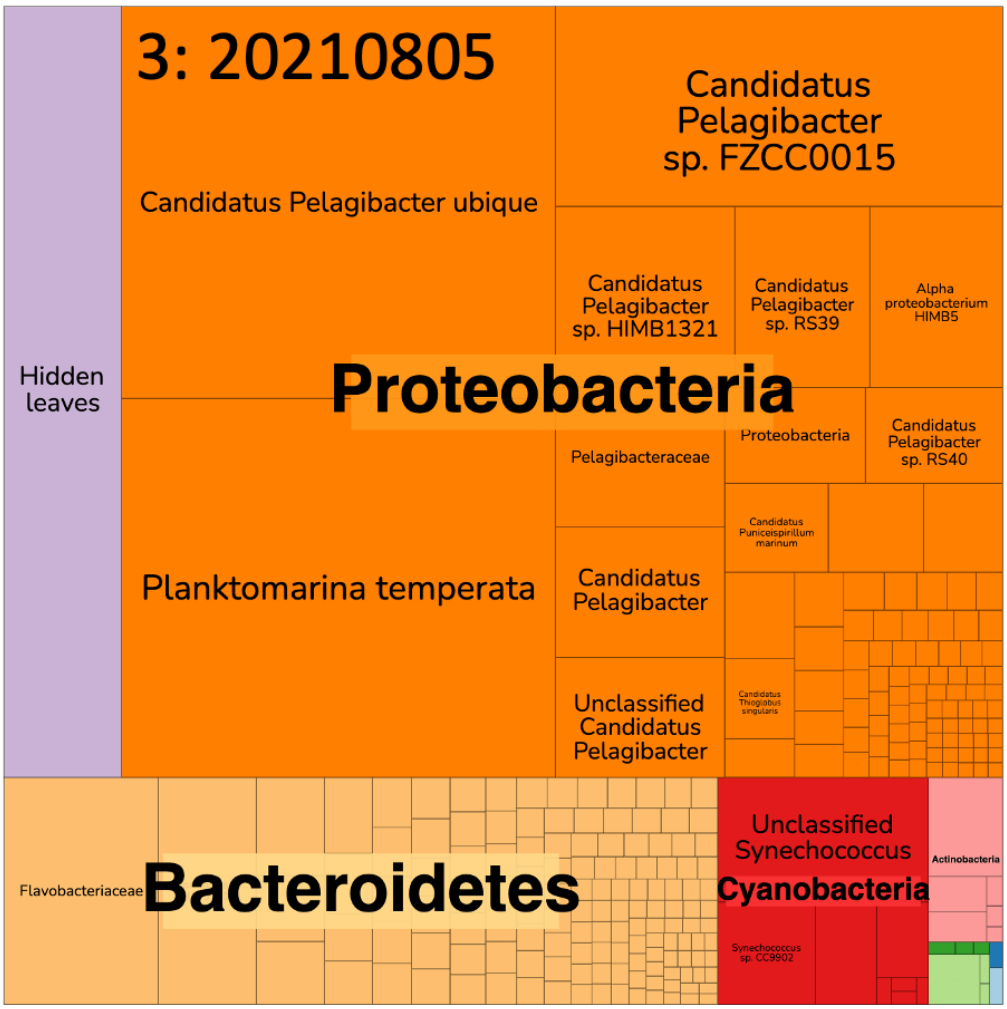


Figure E.0.3: Treemap at genus level showing prokaryotic genera identified in sample 3.

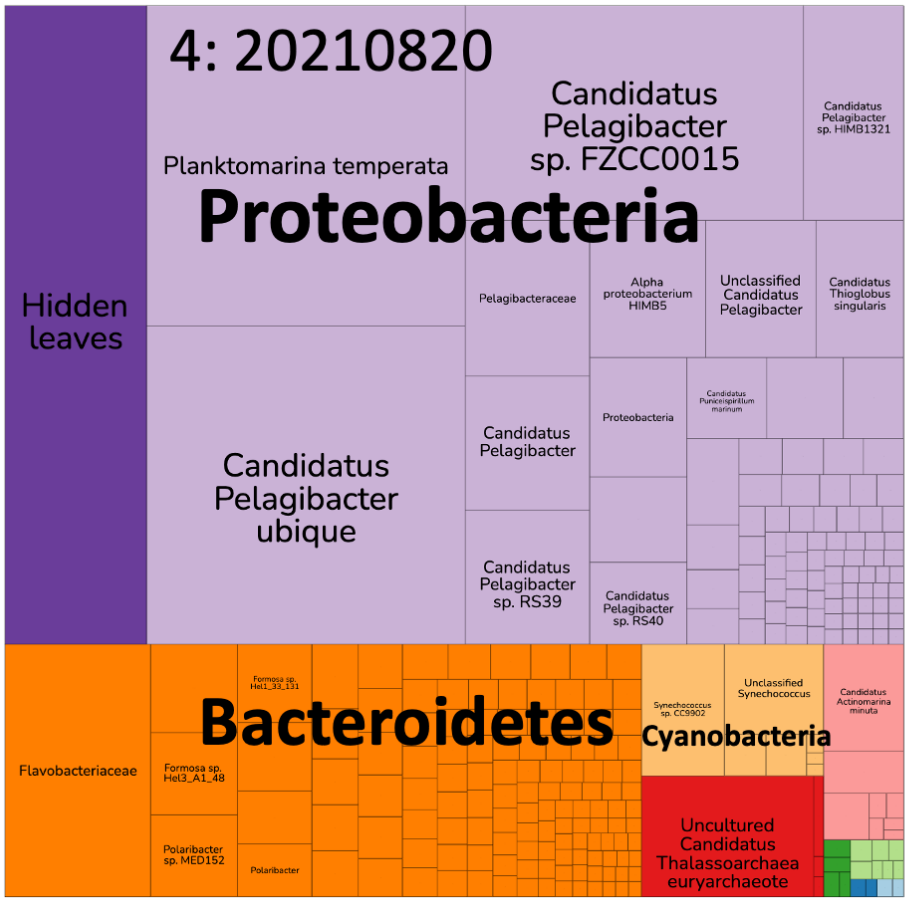


Figure E.0.4: Treemap at genus level showing prokaryotic genera identified in sample 4.

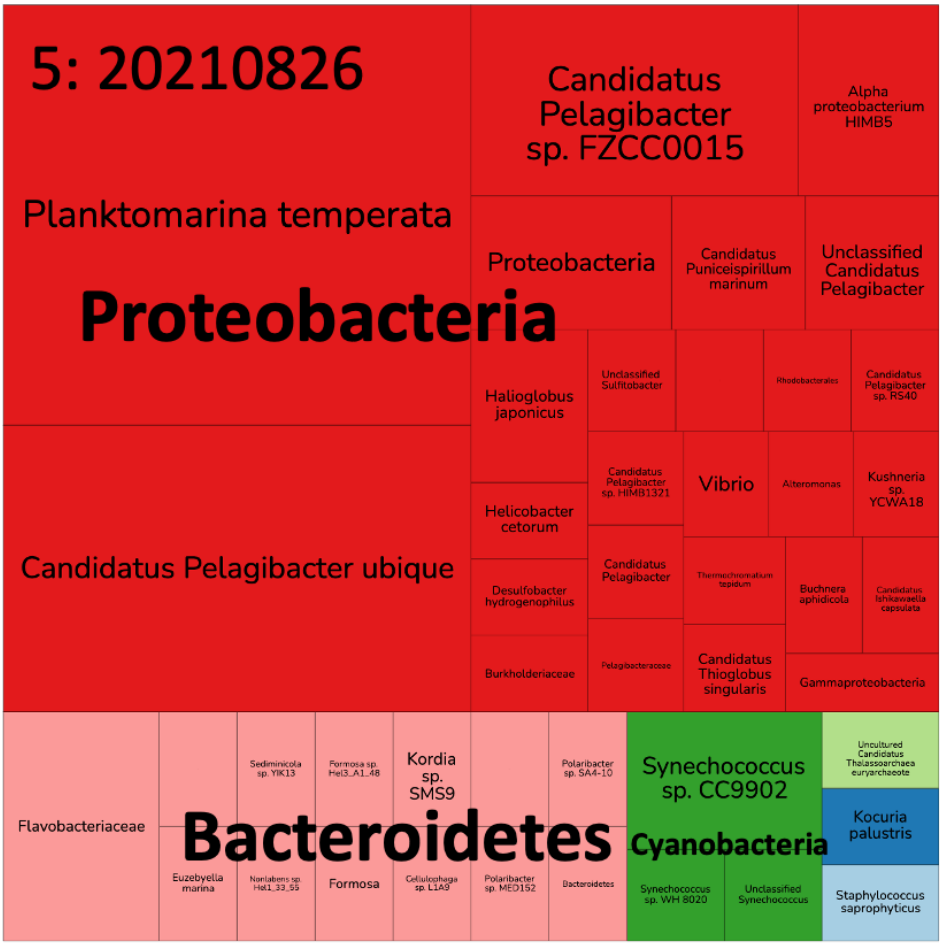


Figure E.0.5: Treemap at genus level showing prokaryotic genera identified in sample 5.

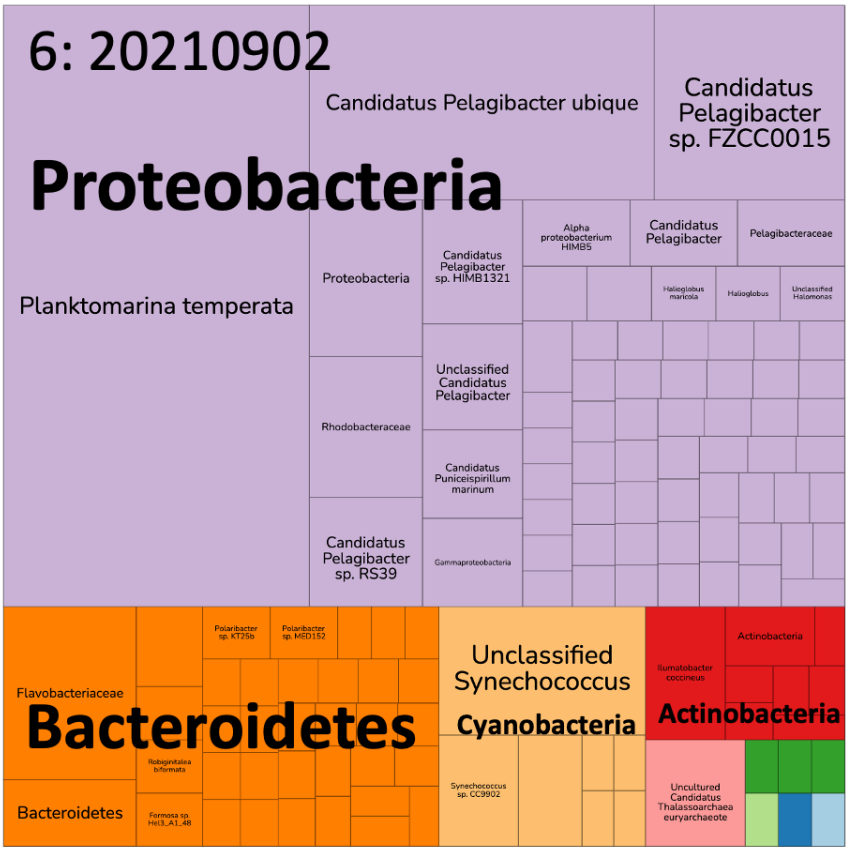


Figure E.0.6: Treemap at genus level showing prokaryotic genera identified in sample 6.

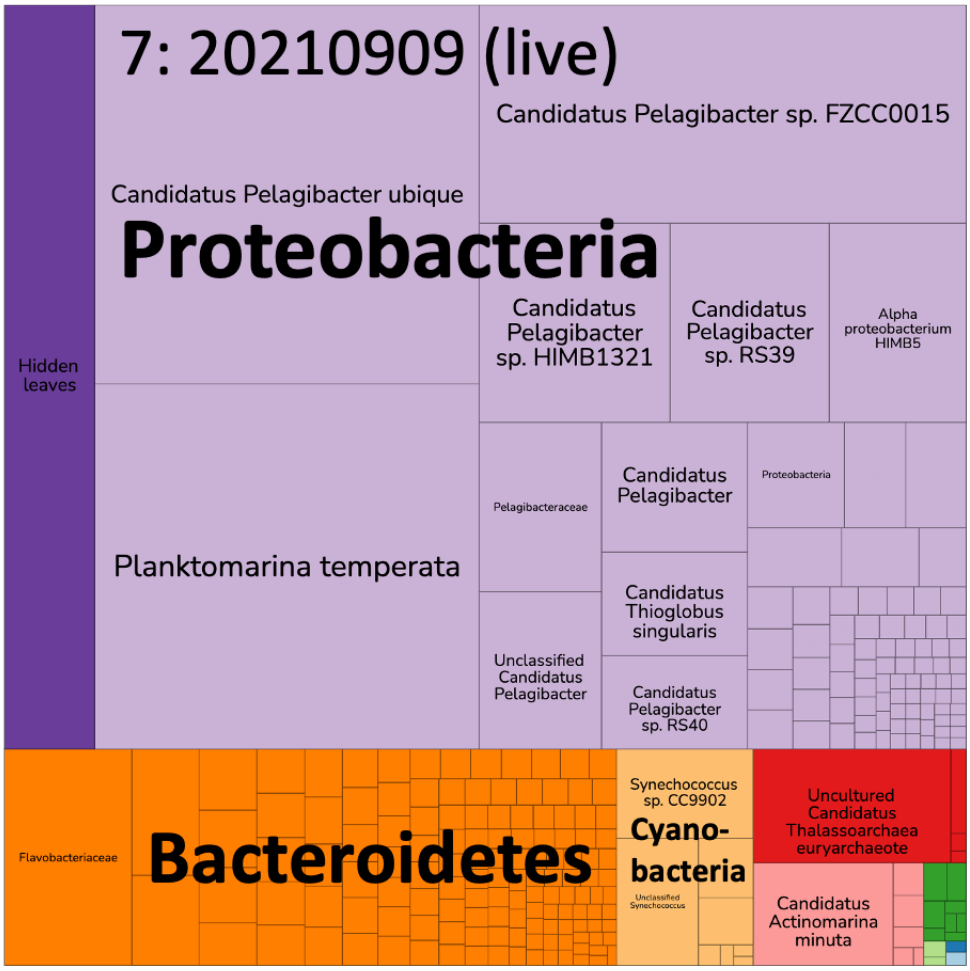


Figure E.0.7: Treemap at genus level showing prokaryotic genera identified in sample 7.

F

Viral treemaps

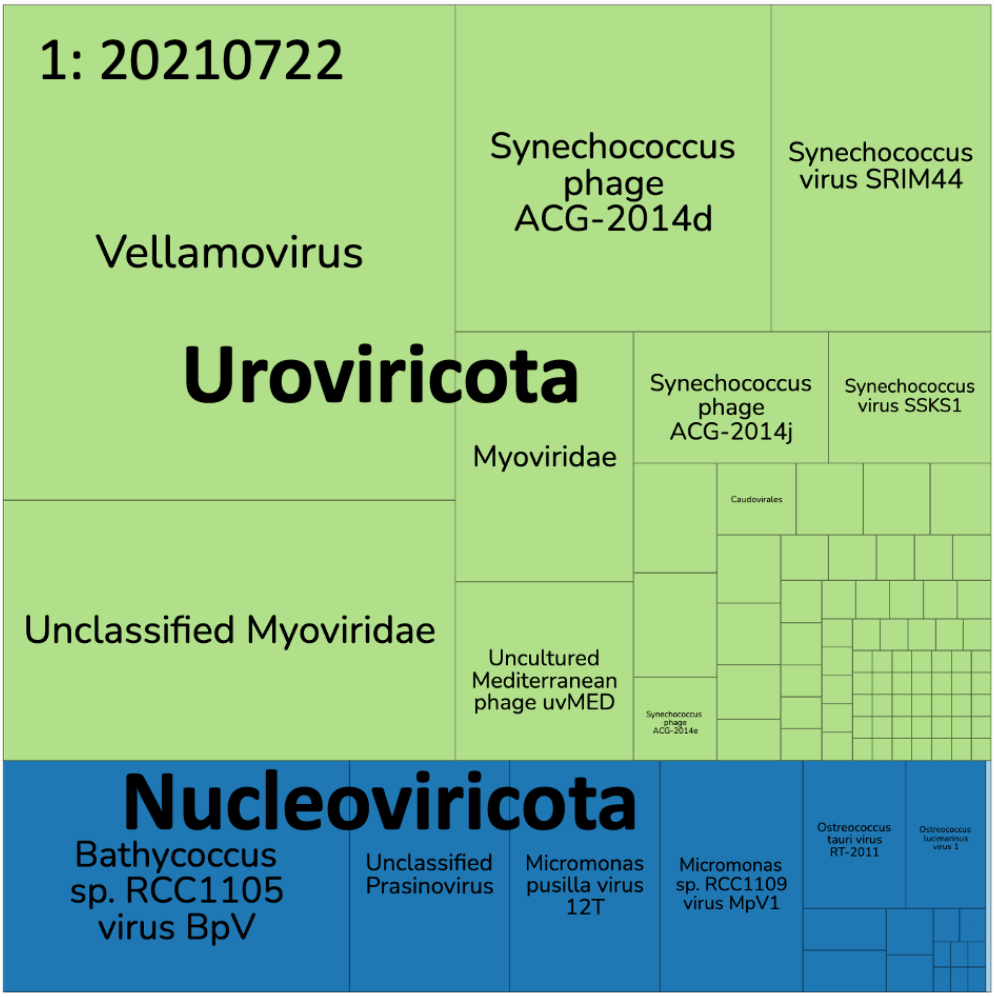


Figure F.0.1: Treemap at genus level showing viral genera identified in sample 1.

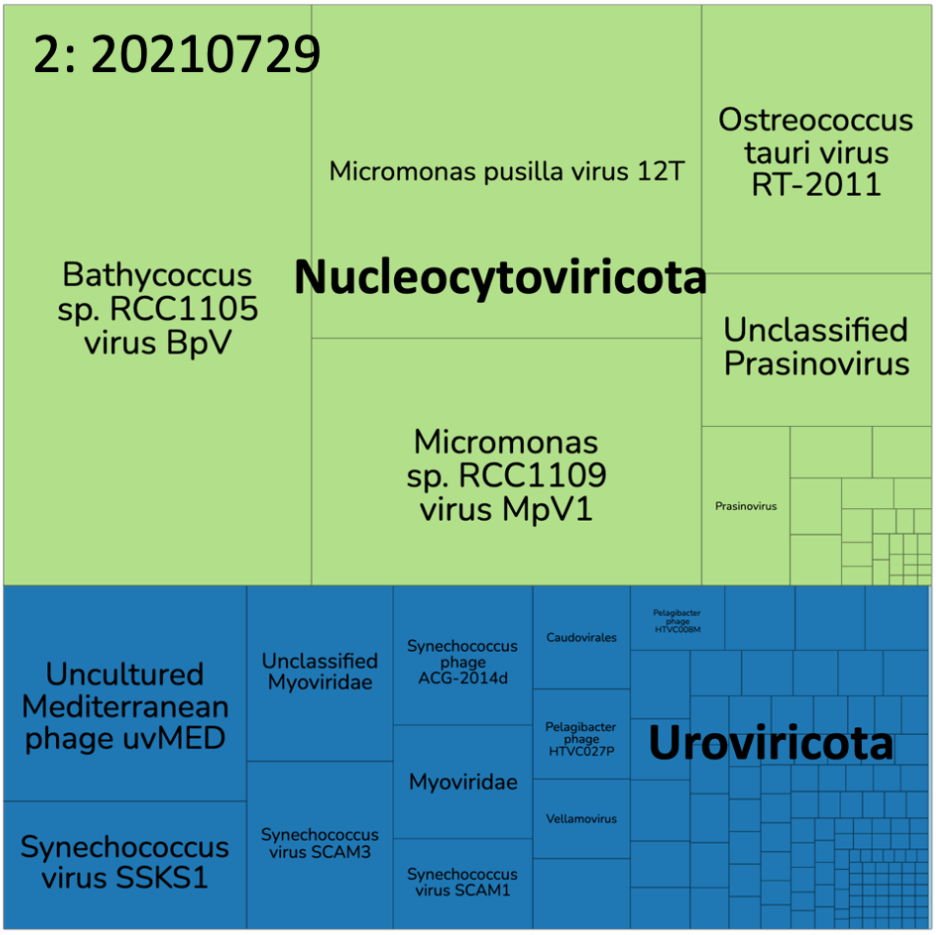


Figure F.0.2: Treemap at genus level showing viral genera identified in sample 2.

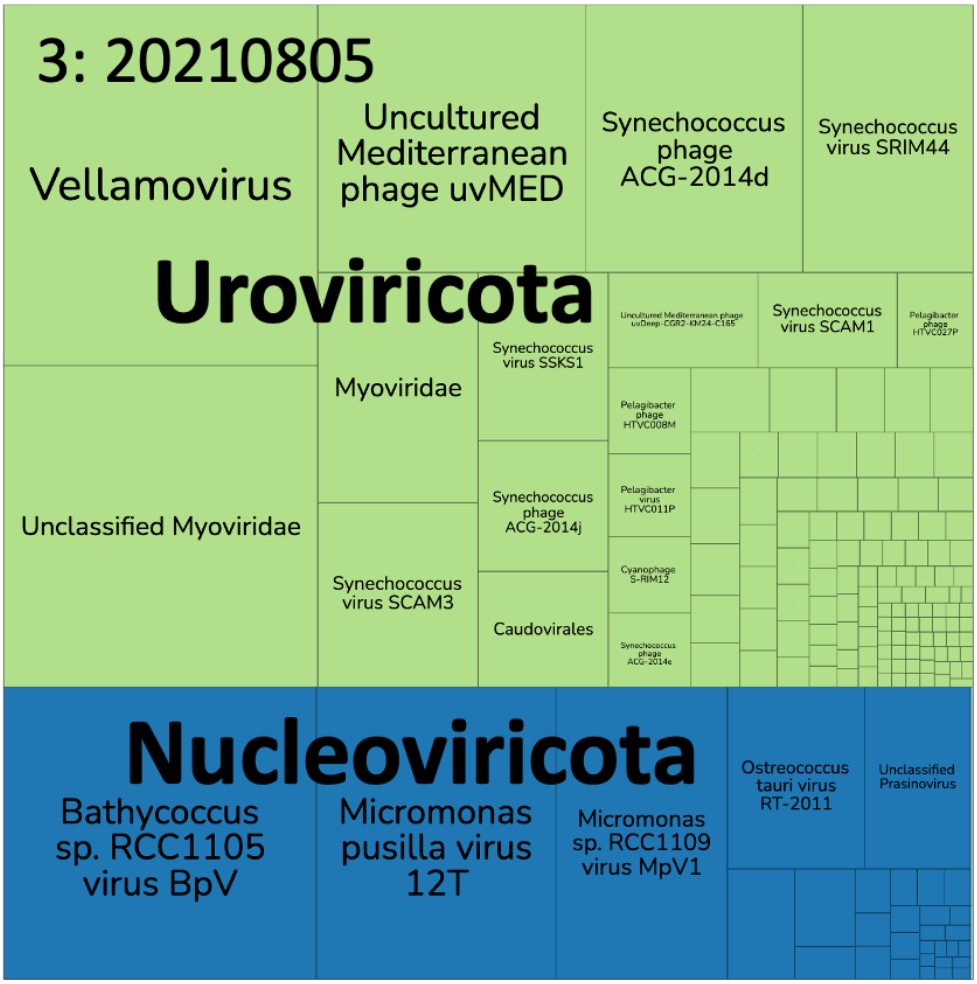


Figure F.0.3: Treemap at genus level showing viral genera identified in sample 3.

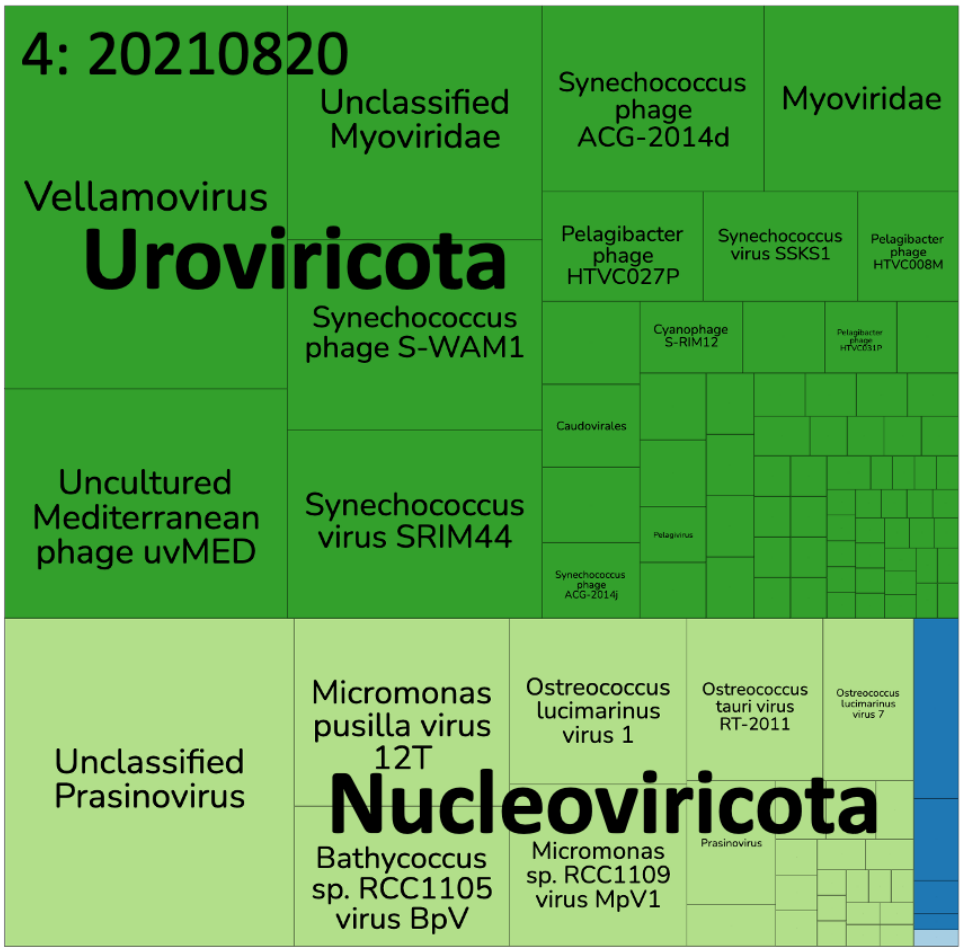


Figure F.0.4: Treemap at genus level showing viral genera identified in sample 4.

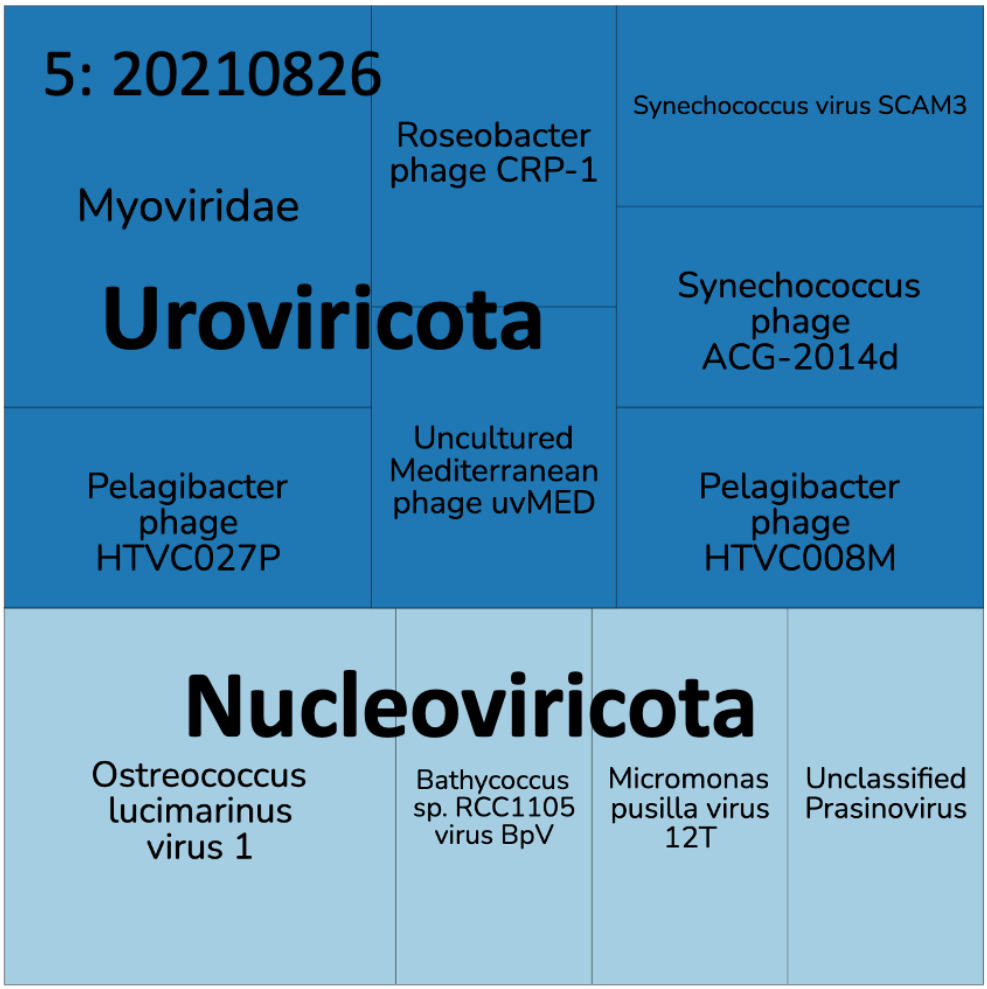


Figure F.0.5: Treemap at genus level showing viral genera identified in sample 5.

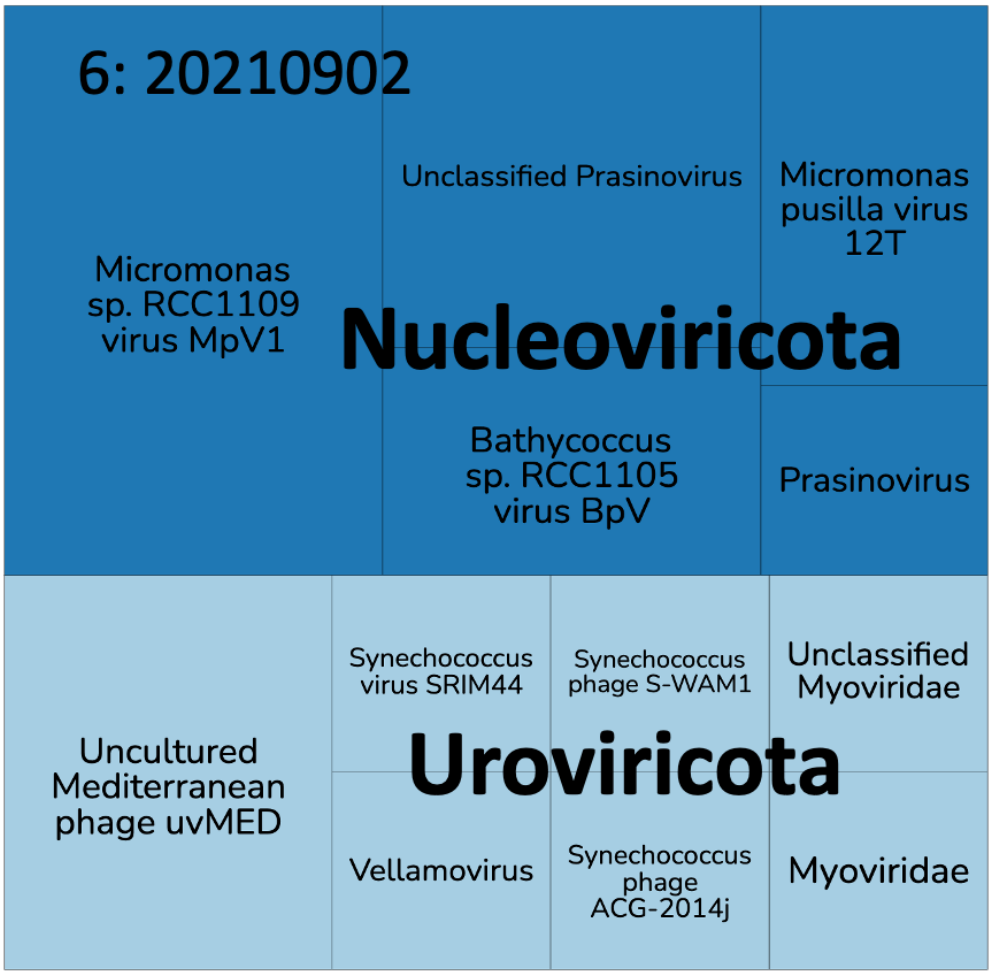


Figure F.0.6: Treemap at genus level showing viral genera identified in sample 6.

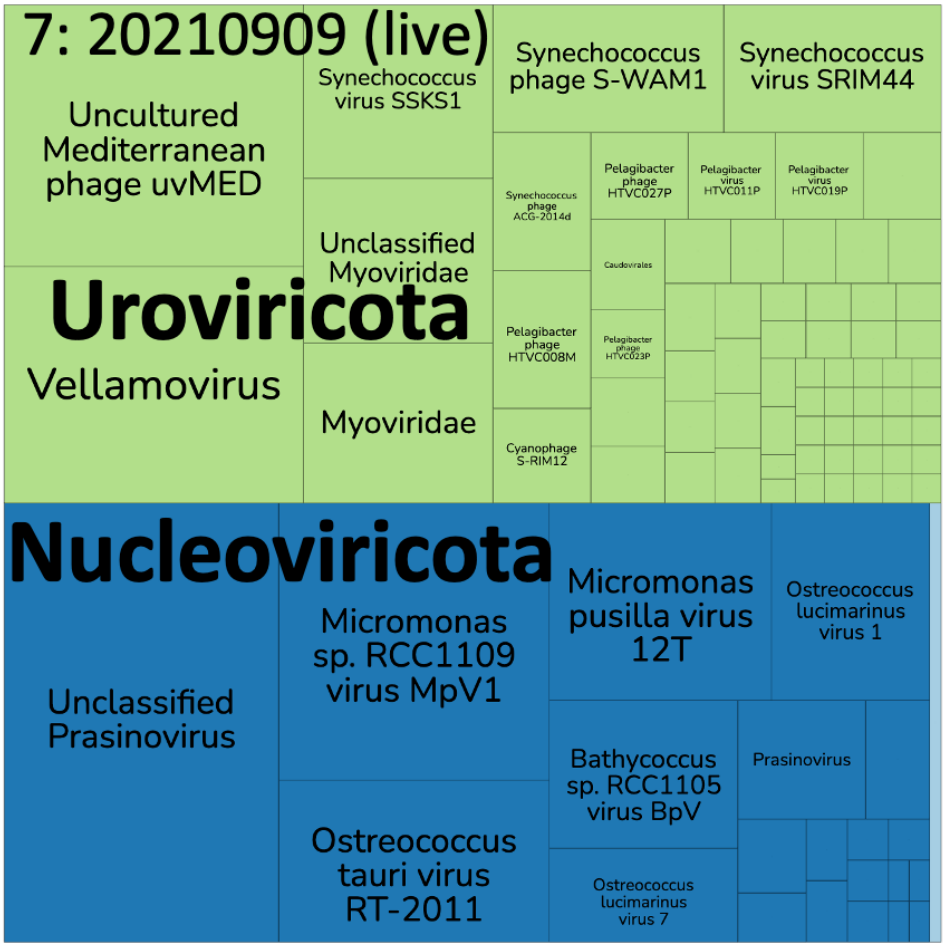


Figure F.0.7: Treemap at genus level showing viral genera identified in sample 7.

G

Family level rarefaction curve

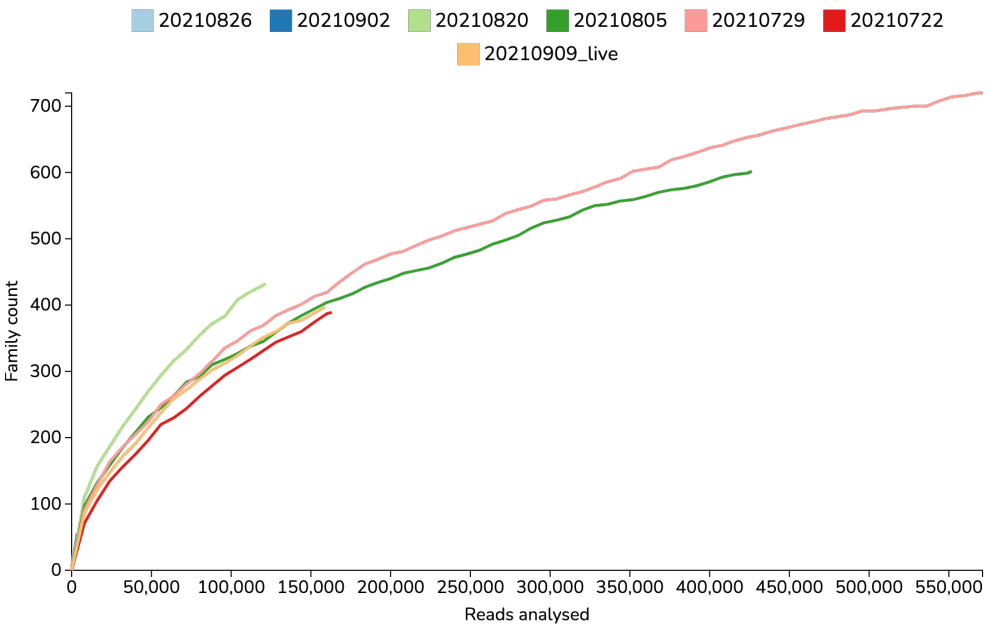


Figure G.0.1: Taxa accumulation curve at family level, showing families found against reads analysed. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

H

Phylum, genus, and species level stacked bar charts

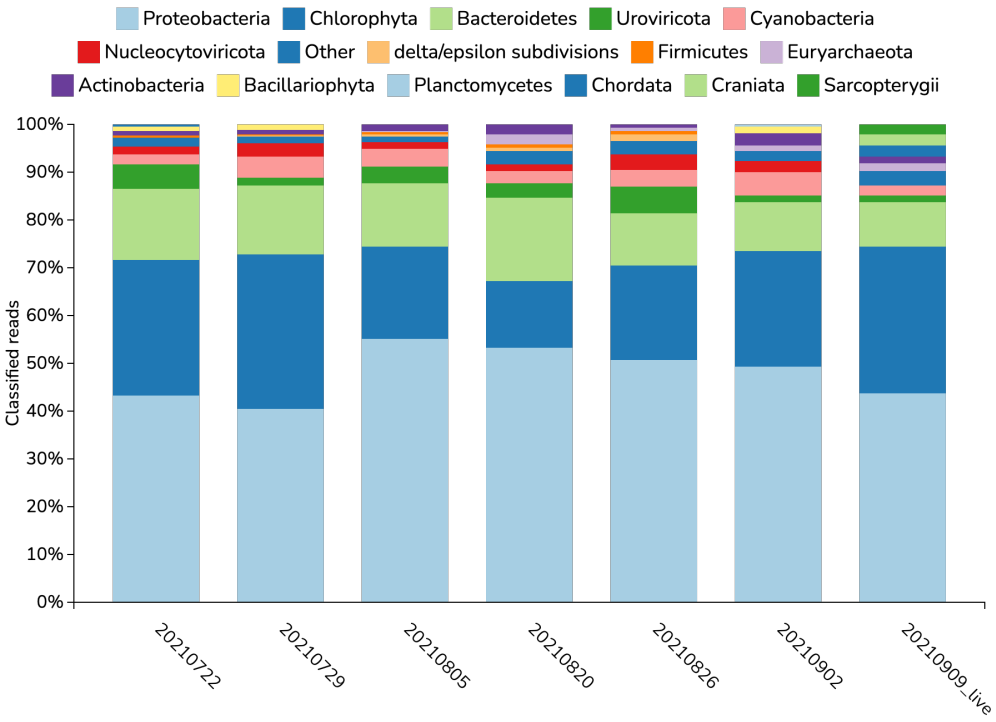


Figure H.0.1: Stacked bar chart showing phylum level matches for each sample. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

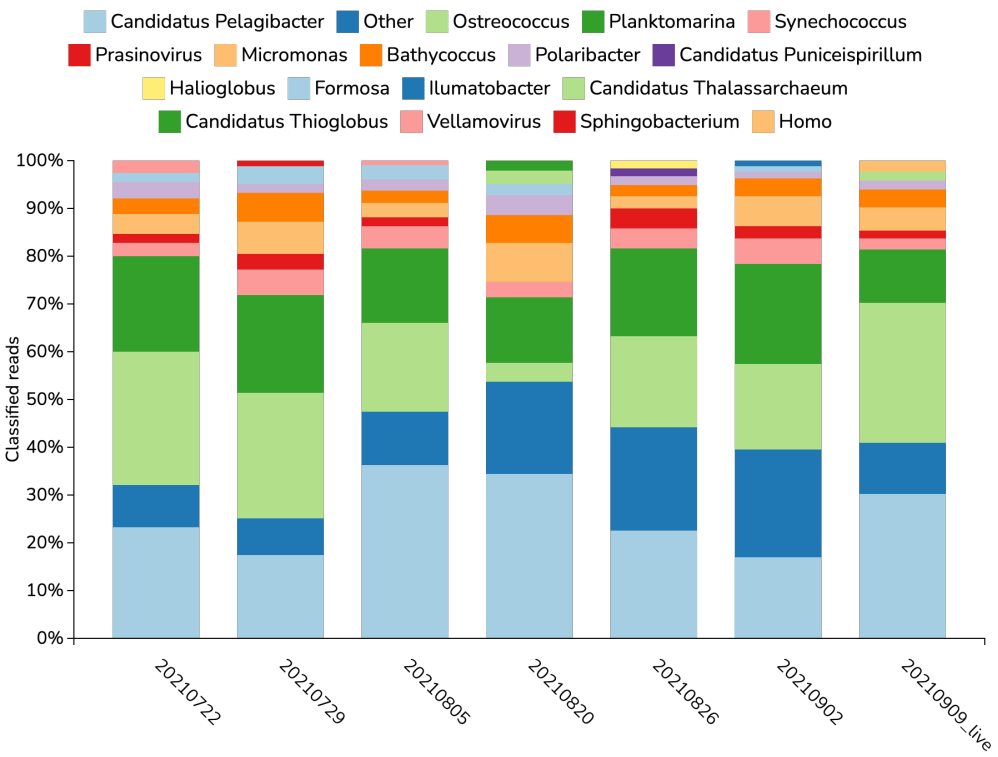


Figure H.0.2: Stacked bar chart showing genus level matches for each sample. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

Figure H.0.3: Stacked bar chart showing species level matches for each sample. Labels are dates sampled in chronological order. Sample 1: 20210722, Sample 2: 20210729, Sample 3: 20210805, Sample 4: 20210820, Sample 5: 20210826, Sample 6: 20210902, Sample 7 (live): 20210909 live.

I

Thalassiosira species read numbers

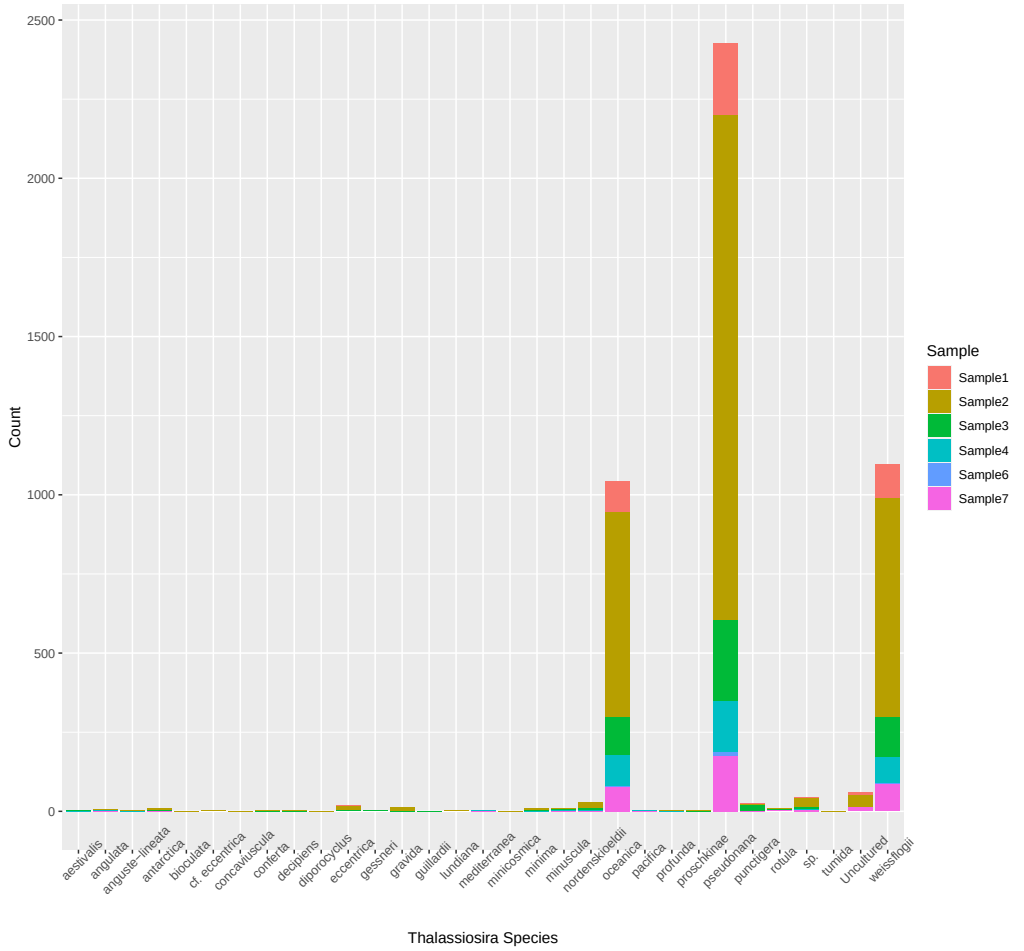


Figure I.0.1: Stacked bar chart showing the number of reads with a blast match to *Thalassiosira* species in each sample

J

Skeletonema species read numbers

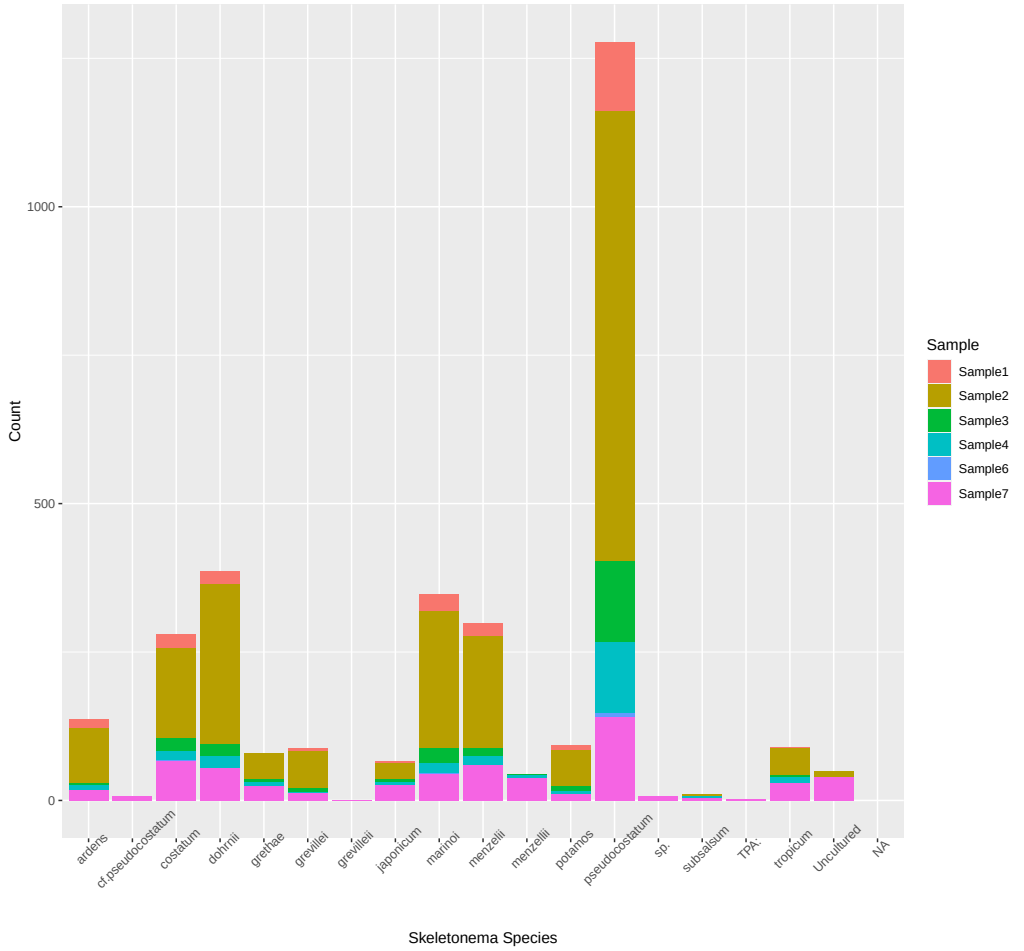


Figure J.0.1: Stacked bar chart showing the number of reads matching to *Skeletonema* species in each sample

K

Vibrio species read numbers

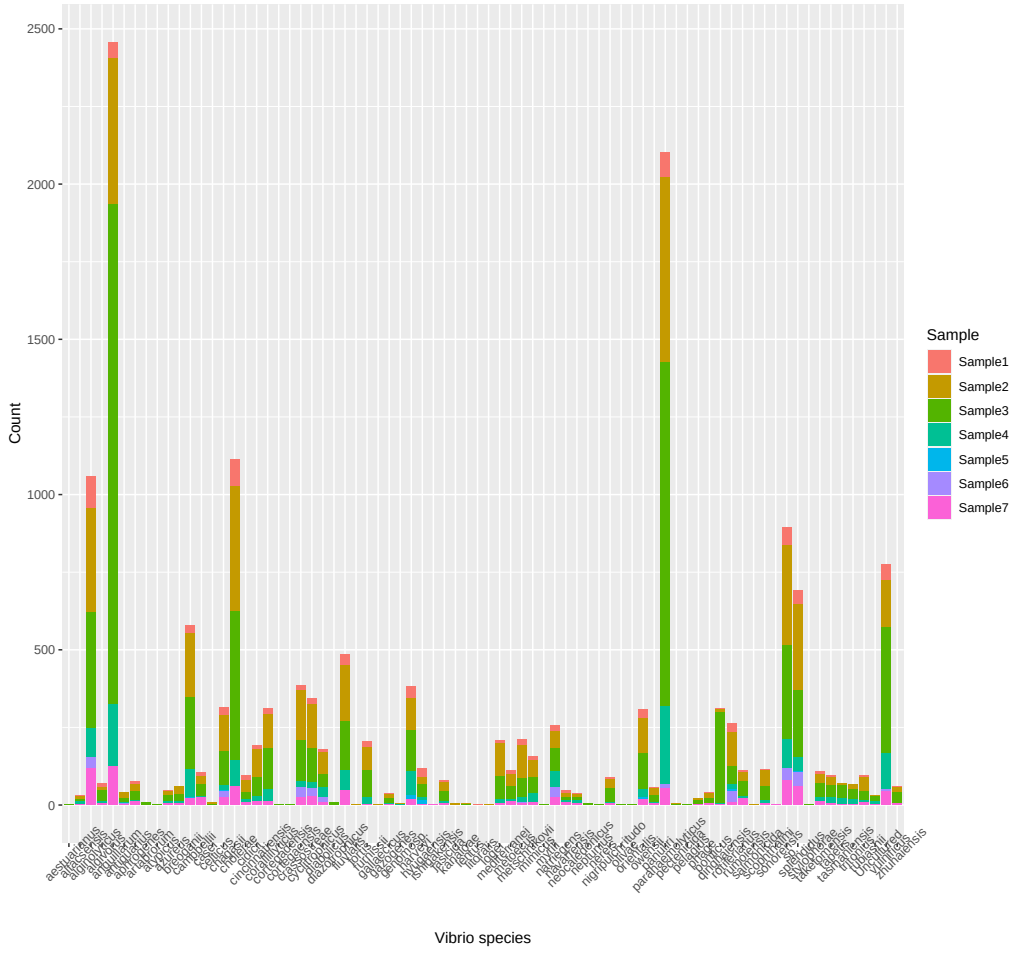


Figure K.0.1: Read number of *Vibrio* species in each sample