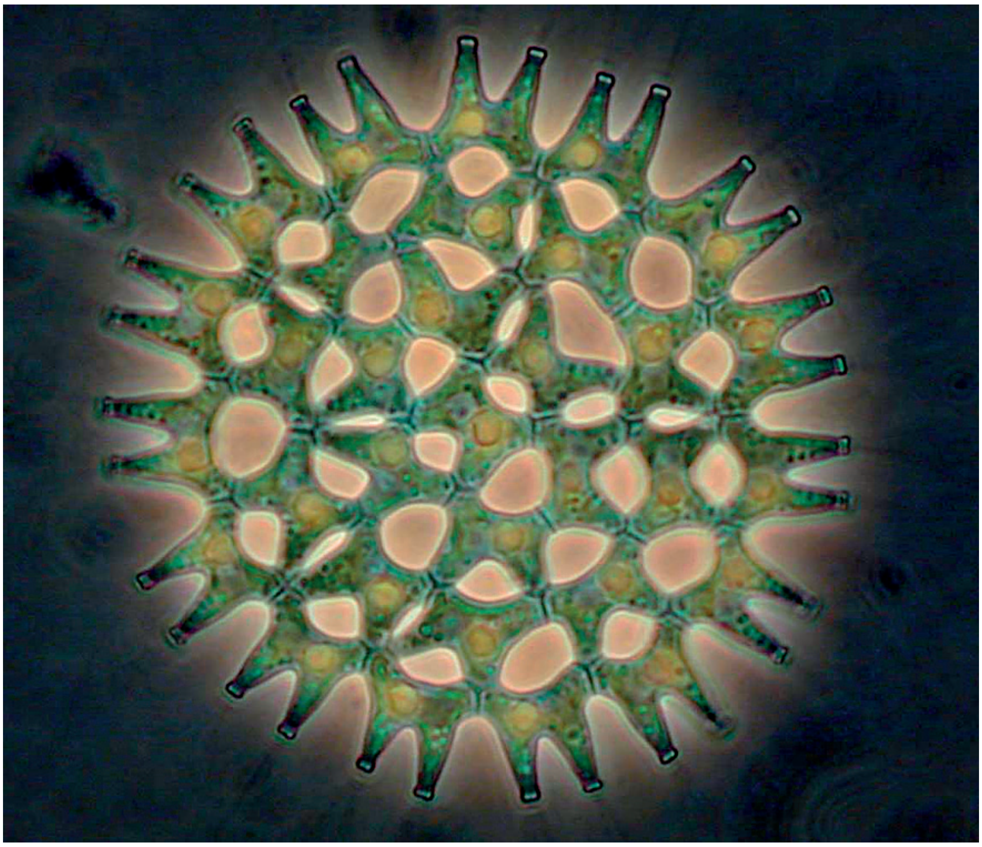
JYU DISSERTATIONS 117

Anita Mäki

Development of genetic phytoplankton monitoring



UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF MATHEMATICS
AND SCIENCE

JYU DISSERTATIONS 117

Anita Mäki

Development of Genetic Phytoplankton Monitoring

Esitetään Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Ambiotica-rakennuksen luentosalissa YAA303
syyskuun 13. päivänä 2019 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Mathematics and Science of the University of Jyväskylä,
in building Ambiotica, hall YAA303, on September 13, 2019 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2019

Editors
Varpu Marjomäki
Department of Biological and Environmental Science, University of Jyväskylä
Timo Hautala
Open Science Centre, University of Jyväskylä

Cover picture: Peidiastrum
Image: Reija Jokipii

This is a printout of the original online publication.
Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7822-8

ISBN 978-951-39-7822-8 (PDF)
URN:ISBN:978-951-39-7822-8
ISSN 2489-9003 Jyväskylä University Printing House, Jyväskylä 2019

ABSTRACT

Mäki, Anita
Development of genetic phytoplankton monitoring
Jyväskylä: University of Jyväskylä, 2019, 56 p.
(JYU Dissertations,
ISSN 2489-9003; 117)
ISBN 978-951-39-7822-8
Yhteenveto: Geneettisen kasviplanktonseurannan kehittäminen
Diss.

Unicellular microorganisms are the most abundant and diverse lifeforms, being the basis for the existence of other organisms. However, microscopic analysis of microorganisms is a slow process, and it is often impossible for the tiny creatures. To uncover microbial diversity via molecular tools, two library preparation techniques for high-throughput sequencing (HTS) of ribosomal RNA (rRNA) genes and rRNA were developed using phytoplankton as a model target group. Bioinformatics pipelines for data trimming and operational taxonomic unit (OTU) clustering were incorporated into the study.

At first, a cost- and time-efficient workflow for fluent barcoding and size trimming of sequencing templates was established. By targeting the sequencing to the same region of the template molecule, data could be optimally used for OTU-based algorithms. The lack of an extensive validation of molecular tools for phytoplankton diversity analysis gave reason to compare DNA and RNA preservation, extraction and HTS data trimming methods. All processes impacted the HTS results of a mock community, with known biomasses and carbon content for each species. In the rRNA gene-based community profiling, species richness was accurately revealed, but the relative abundances of species were biased due to the high species-specific variation in the rRNA gene copy numbers. However, rRNA-based analysis reflected the relative biomasses more closely and was recommended for studying the diversity of eukaryotes.

Even in the RNA based analysis, PCR amplification with gene-specific primers can bias the outcomes, as real universal primers for amplifying eukaryotic and prokaryotic ribosomal genes are not available. Therefore, a primer-independent rRNA workflow was developed to allow directional 5′-end HTS. A significant advantage of this method is that it allows simultaneous sequencing of all domains of life. The new workflow was applied to analyse water samples from 83 Finnish lakes. Results uncovered an incredibly high diversity of organisms and partly agreed with, and partly complemented, phytoplankton microscopy results.

Keywords: High throughput sequencing; phytoplankton; ribosomal RNA.

Anita Mäki, University of Jyväskylä, Department of Biological and Environmental Science, P.O. Box 35, FI-40014 University of Jyväskylä, Finland

TIIVISTELMÄ

Mäki, Anita
Development of genetic phytoplankton monitoring
Jyväskylä: University of Jyväskylä, 2019, 56 p.
(JYU Dissertations,
ISSN 2489-9003; 117)
ISBN 978-951-39-7822-8
Yhteenveto: Geneettisen kasviplanktonseurannan kehittäminen
Diss.

Yksisoluiset mikro-organismit ovat lukuisin eliömuoto, ja niiden olemassaolo on elintärkeää monisoluisille organismeille. Pienten mikro-organismien mikroskooppinen analyysi on hidasta ja usein mahdotonta. Tässä molekyylibiologisessa työssä kehitettiin DNA- ja RNA-perusteiset templaattien valmistelumenetelmät HTS-tekniikalla (high-throughput sequencing) tehtävään mikrobiyhteisöjen tutkimiseen käyttäen kasviplanktonia malliorganismina. Bioinformatiikka sekvensointidatan trimmaukseen ja OTU-klusterointiin (operational taxonomic unit) sisältyi tutkimukseen. Ensiksi kehitettiin kustannuksia ja aikaa säästävä, DNA-perusteinen HTS-templaattien indeksointi ja koon trimmausprosessi. Kohdistamalla sekvensointi templaattimolekyylien samaan kohtaan OTU-algoritmit toimivat optimaalisesti data-analyyseissä. Koska kasviplanktontutkimuksessa on ollut puute kattavasta molekylaaristen menetelmien testaamisesta, tässä työssä verrattiin solujen säilytysmenetelmien, eri DNA:n ja RNA:n eristysmenetelmien sekä HTS-datan trimmauksen vaikutuksia lopullisiin HTS-tuloksiin. Realistinen yhteisörakenne määritettiin lajien suhteellisista biomassa- ja hiilipitoisuuksista, joihin eri prosessien HTS-tuloksia verrattiin. DNA-perusteisessa yhteisömäärityksessä lajirikkaus näyttäytyi tarkasti, mutta lajien suhteelliset määrät vääristyivät suurten lajikohtaisten geenikopiomäärien vaihteluiden vuoksi. Sitä vastoin RNA-perusteinen analyysi muistutti lajien suhteellisia biomassoja paremmin ja RNA-perusteista analyysiä suositeltiin eukaryoottisten mikrobien yhteisöanalyyseihin. Myös RNA-perusteisessa analyysissä PCR-monistaminen voi vääristää tuloksia käytettäessä ribosomaalisen RNA:n geenispesifisiä alukkeita. Koska eukaryooteille ja prokaryooteille ei ole saatavilla universaalisia alukkeita, kehitettiin RNA-perusteinen, alukkeista riippumaton, suunnattu 5′-pään HTS, jolla voidaan sekvensoida kaikkien domeenien eliöt samanaikaisesti. Uutta menetelmää sovellettiin 83 Suomen järven vesinäytteiden analysointiin. Tulokset paljastivat uskomattoman suuren organismien monimuotoisuuden ja osittain vastasivat ja osittain täydensivät kasviplanktonin mikroskopointituloksia.

Avainsanat: HTS-tekniikka; kasviplankton; ribosomaalinen RNA.

Anita Mäki, University of Jyväskylä, Department of Biological and Environmental Science, P.O. Box 35, FI-40014 University of Jyväskylä, Finland

Author's address Anita Mäki
Department of Biological and Environmental Science
P.O. Box 35
FI-40014 University of Jyväskylä
Finland
anita.maki@jyu.fi

Supervisors Professor Marja Tiirola
Department of Biological and Environmental Science
P.O. Box 35
FI-40014 University of Jyväskylä
Finland

Docent Kristiina Vuorio
Finnish Environment Institute (SYKE)
Latokartanonkaari 11
00790 Helsinki
Finland

Reviewers Professor Agneta Andersson
Department of Ecology and Environmental Sciences
Umeå University
Linnaeus väg 6
901 87 Umeå
Sweden

Doctor Markus Majaneva
Norwegian University of Science and Technology
Department of Natural History
NTNU University Museum
Erling Skakkes gate 47A
NO-7491 Trondheim
Norway

Opponent Professor Kaarina Sivonen
Department of Microbiology
University of Helsinki
Viikinkaari 5
00790 Helsinki
Finland

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The thesis is based on the following original papers, which will be referred to in the text by their Roman numerals I–IV. I was responsible for the molecular biology studies for all papers and for the bioinformatics in Papers II–IV.

The initial plan for Paper I came from Marja Tiirola (MT), and all authors edited the plan. I performed the HTS for Paper I, and Anu Mikkonen assisted with bioinformatics. I designed the initial plan for Paper II, and MT, along with Pauliina Salmi (PS), edited the plan. For Paper II, PS counted phytoplankton samples and carried out dry mass and carbon content analyses, and I performed microscope imaging and HTS, while Anu Mikkonen assisted with bioinformatics and was responsible for statistical analyses. Anke Kremp provided the laboratory strains of the phytoplankton mock community and offered important knowledge about phytoplankton for Paper II. For Paper III, MT and I planned the study together. Kristiina Vuorio (KV), MT, and I designed Paper IV. KV performed phytoplankton microscopy and made comparisons between microscopy and HTS, while PS carried out the picoplankton analysis. Sanni Aalto was responsible for the statistical analysis in Paper IV.

I wrote Papers I-III, and KV and I wrote Paper IV together. All the papers were edited by all co-authors.

I Mäki A., Rissanen A.J. & Tiirola M. 2016. A practical method for barcoding and size-trimming PCR templates for amplicon sequencing. *BioTechniques* 60: 88–90.

II Mäki A., Salmi P., Mikkonen A., Kremp A. & Tiirola M. 2017. Sample preservation, DNA or RNA extraction and data analysis for high-throughput phytoplankton community sequencing. *Frontiers in Microbiology* 8: 1848.

III Mäki A. & Tiirola M. 2018. Directional high-throughput sequencing of RNAs without gene-specific primers. *BioTechniques*. 60: 219–223.

IV Vuorio K., Mäki A., Aalto S.L. & Tiirola M. 2019. Consistency of targeted metatranscriptomics and morphological characterization of phytoplankton communities. Manuscript.

ABBREVIATIONS

AGE	agarose gel electrophoresis
bp	base pair
cDNA	complementary DNA
DNA	deoxyribonucleic acid
EM	electron microscopy
emPCR	emulsion polymerase chain reaction
ER	endoplasmic reticulum
GCN	gene copy number
HABs	harmful algal blooms
HTS	high throughput sequencing
ISP	Ion sphere particles
nt	nucleotide
OTU	operational taxonomic unit
PGM	Personal Genome Machine
PCR	polymerase chain reaction
Pol I	RNA polymerase I
Pol II	RNA polymerase II
qPCR	quantitative polymerase chain reaction
RA	relative abundance
RNA	ribonucleic acid
RNase	ribonuclease
rrn	ribosomal RNA operon
rRNA	ribosomal RNA
SBS	sequencing by synthesis
TdT	terminal nucleotidyl transferase

1 INTRODUCTION

1.1 Microorganisms and phytoplankton

As unicellular microorganisms colonise our bodies, food, homes and almost every habitat of the environment, an improved understanding of the microbial world is essential. The estimated number of microbial species on Earth is proposed to be up to one trillion (10^{12}), and only a small fraction of those (less than 10^5 species) is represented by classified sequences (Locey and Lennon 2016). The challenge of identifying microbes is enormous, and effective methods are necessary for a better understanding of microbial life.

All cells of the two prokaryote domains, Bacteria and Archaea, are microbes, whereas, from the third domain, Eukarya, only unicellular species containing a membrane-bound nucleus and cell organelles are defined as microbes (Bauman *et al.* 2004). A wide range of microbes, including protozoa, fungi, archaea, acellular viruses (not considered microorganisms), bacteria and algae, has been observed over the past centuries. Advancements in light, fluorescence, confocal and electron microscopy techniques, along with culturing, staining and molecular methods, have enabled observing and acquiring knowledge about the structure, physiology, biochemistry and genetics of microorganisms (Madigan and Martinko 2006). One main advance in modern microbiology is the observation that ribosomal genes provide information about the phylogenetic relationships of microbes and offer the possibility to identify microbes (Woese and Fox 1977, Brooks 2013). Over the past few decades, rapid, automated, highly sensitive and specific molecular techniques have opened new abilities for microbial identification studies.

However, research of eukaryotic microorganisms has been lagging behind other microbial studies, and the need for new, suitable methods for studying micro– and nanoplankton and the dynamics of food webs has been emphasised (Caron *et al.* 2008, Calbet 2008).

1.1.1 Phytoplankton

Microscopic, drifting, photosynthetic, primary producers of the water systems are called phytoplankton. Plankton consists of organisms that have adapted to live part or all their lifetimes suspended in open water, and plankton's power of motility is too low to overcome the velocity and direction of water movements (Reynolds 2006). 'Phyto' is derived from Ancient Greek and means 'plant'. These single-celled microalgae may be prokaryotic photosynthetic cyanobacteria or eukaryotic cells, all of which provide organic matter for the other organisms, such as zooplankton and bacteria, living in waters (Vargas *et al.* 2006, Falkowski 2012). Algae are morphologically simple organisms without any specialist parts, such as roots, stems or leaves. Physiologically, they are autotrophic and capable of photosynthesis. If algae have become secondarily heterotrophic, they still have preserved genetic relatedness with their photosynthetic relatives (Bellinger 2015). Algae may exist as substrate-associated, benthic organisms, but studies in this thesis focus on planktonic, freely drifting algae.

TABLE 1 Major phyla of Algae and life form characteristics. (Modified from Bellinger 2015).

Algal phylum	Division	Life form
Cyanophyta	blue-green algae	unicellular, filamentous, or colonial
Chlorophyta	green algae	unicellular, filamentous, or colonial
Euglenophyta	euglenoids	unicellular
Xanthophyta	yellow-green algae	unicellular or filamentous
Dinophyta	dinoflagellates	unicellular
Cryptophyta	cryptomonads	unicellular
Chrysophyta	chrysophytes	unicellular or colonial
Bacillariophyta	diatoms	unicellular of filamentous

Freshwater, planktonic algae belong to eight major phyla (Table 1), determined by their microscopical appearances, biochemical compositions, and cytological features. Phytoplankton cell sizes vary: <2 µm in picoplankton; 2–20 µm in nanoplankton; >20 µm in microplankton; >200 µm in macroplankton (Sieburth *et al.* 1978, Bellinger 2015). By morphology (Fig. 1), such as size, shape and colour, phytoplankton is an extremely diverse group of species, which may exist in unicellular forms, filamentous (linear) colonies or other colonial types (Table 1). Therefore, the abovementioned phenotypic characteristics, together with traits such as the presence of chloroplasts and the possibility for motility (usually flagella occurrence, length, and number), are important factors in microscopic classification (Bellinger 2015). In addition, biochemical features are relevant identification markers, including pigmentation, food reserves, external covering (e.g., cellulose cell walls in green algae, opaline silica frustule in diatoms or peptidoglycan matrices or cell walls in blue-green algae) and distinctive combinations of pigments (e.g., chlorophylls and carotenes). Total biomass of algae can be estimated from chlorophyll-*a* because it is found in all pigmented algae (Bellinger 2015).

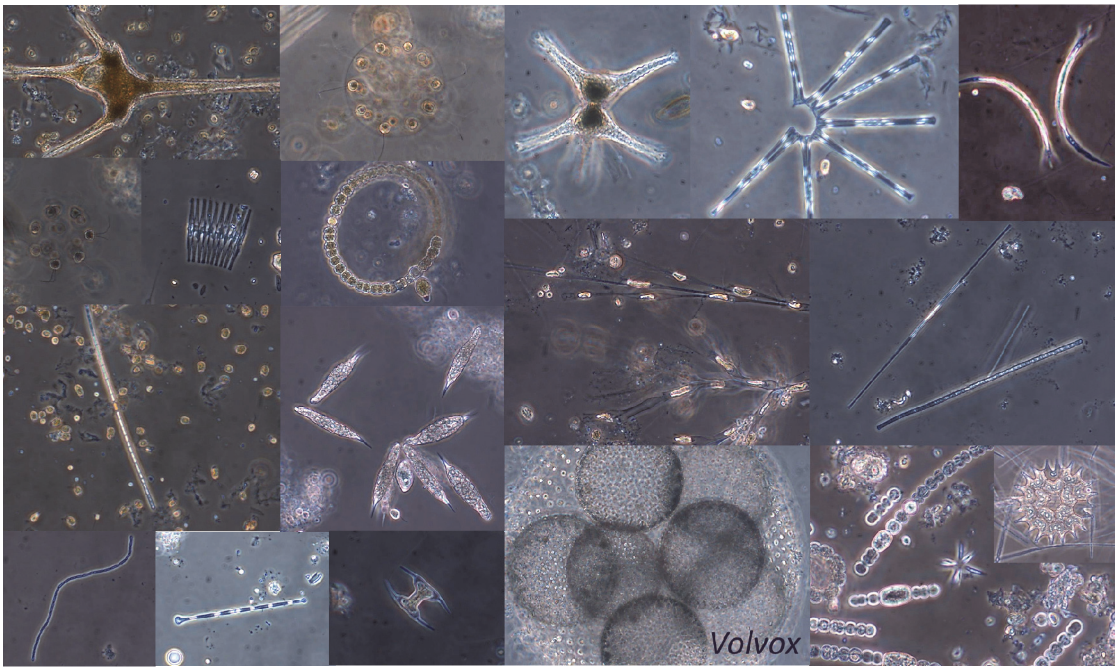


FIGURE 1 Photo collage representing the great diversity of cell shapes and the beauty of phytoplankton. Many colonies containing *Volvox* may be comprised of more than 1,000 cells. (Cells in the collage are not on the same scale.)

An exceptional structure of *Volvox* (Fig. 1) contains many colonies and thousands of cells. Cells of the same genotype have some levels of morphological differentiation and task distribution (Beardall *et al.* 2009, Bellinger 2015). However, the colonial forms of phytoplankton lack specialised cells and tissue forming capacity (Fenchel 2013).

Phytoplankton, together with green plants, are capable of photosynthesis (Catling and Zahnle 2002), and they have a fundamental role in Earth’s oxygen production and in the fixation of carbon dioxide from Earth's atmosphere. Knowledge about changes in phytoplankton community composition in response to Earth’s climate conditions is essential (Schulz *et al.* 2017, Basu and Mackey 2018). Environmental changes can disrupt phytoplankton photosynthesis and rates of oxygen production, and, therefore, understanding microbial responses is highly important (Sekerci and Petrovskii 2018).

In contrast to the redeeming features of phytoplankton, harmful algal blooms (HABs) can cause health issues for humans, domestic animals and wildlife. To identify HABs, the same methods used in microscopic identification can be employed, e.g., identification based on cell morphology and the optical properties of photopigments. Certain cellular target molecules can also be used in identification, such as cell surface moieties (polysaccharides, proteins and lipopolysaccharides) and nucleic acids (Sellner *et al.* 2003, Shumway *et al.* 2018). In both freshwater and marine environments, the frequency and extent of HABs are predicted to continue increasing due to impacts from human-induced nutrient input, urban runoff, global warming and drought events (Berdalet *et al.* 2015, Peacock *et al.* 2018).

Over the last 150 years, phytoplankton classification has been based on the agreements of microscopic and biochemical analyses, but subdivisions within classes, orders and families has become partly controversial (Reynolds 2006). Recently, revisions to the classification, nomenclature and diversity of eukaryotes has been published putting an emphasis on protists (Adl *et al*. 2019). Although standard microscopic analysis techniques, such as the Utermöhl method for quantitative phytoplankton analysis, have been applied for phytoplankton microscopic identification, variations among professional analysts have been high (Vuorio *et al.* 2007, Karlson *et al.* 2010).

Since the 1980s, robust molecular methods, such as gene sequencing and statistical matching of results to the closest species, have changed the taxonomic listing significantly. Entering the 21st century, the availability of new sequencing techniques and the further expansion of reference databases have justified the development of new, efficient, molecular methods for phytoplankton characterisation. Molecular data has been increasingly exploited in algal systematics, and sequencing data has provided new insights for identifying and classifying species and making nomenclatural decisions (Oliveira *et al.* 2018). Sequencing-based molecular methods are superior to microscopic techniques when the diversity of small picoplankton is investigated and when clear morphological characteristics are not available (Lepere *et al.* 2010, Bellinger 2015, Duan *et al.* 2018).

Since molecular methods have entered the field of microbial diversity studies, the development of universal primers has been one of the biggest challenges to constructing the sequencing library (Suzuki and Giovannoni 1996, Hong *et al.* 2009, Takahashi *et al.* 2014, Karst *et al.* 2018). Because phytoplankton comprises prokaryotic and eukaryotic cells, the challenge is even more complicated (Hadziavdic *et al.* 2014). Also, high variation in the gene copy numbers (GCNs) of marker genes for many eukaryotic species can bias DNA-based sequencing results (Zhu *et al.* 2005, Godhe *et al.* 2008, Gong *et al.* 2013, Wang *et al.* 2017, Needham *et al.* 2018).

1.2 Genetic markers in microbe characterisation

DNA sequence fragments from one organism varying from those of other organisms, can act genetic, molecular markers. DNA barcoding refers to a system, in which the standardised DNA marker genes of an organism can be used to identify the species in question (Hebert *et al.* 2003). A phylogenetic DNA or RNA marker segment (a molecular chronometer) is a highly conserved nucleic acid fragment that has the capacity to be chronometer and to behave like a molecular clock to estimate the degree of phylogenetic relationship between organisms. Variable regions in chronometers are essential to distinguishing closely related organisms. Some other important features in a chronometer are: 1. It should be a single-copy, rather than a multiple-copy, gene. 2. The alignment of the gene should be easy. 3. The information gained form the gene should be sufficient. 4. Amplification primers should be available (Cruickshank 2002). With a good chronometer, the ideal primer-pair could amplify the same region in different taxa in a single PCR run (Casiraghi *et al.* 2010).

More than three decades ago, in his review entitled 'Bacterial Evolution', Woese (1987) described the nature of molecular chronometers. One ultimate feature of the chronometer sequence is the ability to change randomly in time. To be useful for phylogenic measurement, a molecule should keep functional fidelity and should occur in all organisms. As ribosomal RNA (rRNA) meets these and other important expectations, such as having conserved and variable regions, it can be considered the most useful, and the most often used, molecular chronometer (Woese 1987, Vinje *et al.* 2014). Because of features such as ubiquity, size and low evolutionary rate, small subunit ribosomal RNA (SSU rRNA) gene sequences have also been used extensively for evolutionary studies, and collection of sequences has continued more than three decades (Van de Peer and De Wachter 1997).

Other DNA sequences, such as mitochondrial gene cytochrome c oxidase I (COI), have also been used as DNA barcodes (i.e., genetic taxon 'barcodes') in identification studies, especially at the beginning of the 21st century (Hebert *et al.* 2003). Other frequently used molecular markers for species-level biosystematics of eukaryotes are the mitochondrial 16S rRNA gene, the cytochrome b (CytB) gene, the NADH dehydrogenase subunit 5 (ND5) gene, the internal transcribed spacers (ITS1 and ITS2) of nuclear-encoded rRNA genes, and the ribulose bisphosphate carboxylase large chain (*rbcL*) gene of plastids (Pichard *et al.* 1997, Vences *et al.* 2005, Hajibabaei *et al.* 2007, Low *et al.* 2014).

1.2.1 Ribosomes and their 16S or 18S SSU rRNAs

Ribosomes are the site of protein synthesis in cells. The structures of the ribosomal proteins (r-protein) and rRNAs complexes have been studied intensively since the 1970s using electron microscopy (EM), X-ray crystallography and cryo-EM (Brown and Shao 2018, Denny and Greenleaf 2018). The detailed structural knowledge of ribosomes gathered during the last decade has increased the functional understanding of ribosomes (Chandramouli *et al.* 2008). A prokaryotic cell contains approximately 20,000 ribosomes, and actively growing mammalian cells contain from five million to ten million ribosomes, although the number varies largely during cell cycles (Cooper and Hausman 2007, Phillips *et al.* 2012, Raveh *et al.* 2016). The molecular mass of a ribosome varies from 2 to 4.5 MDa (megadaltons) depending on the organism (Jobe *et al.* 2018). Of ribosome mass, approximately two-thirds is RNA, and approximately one-third is protein (Lodish *et al.* 2000).

In eukaryotic cells, free ribosomes are found in cytosol, whereas membrane-bound ribosomes (Fig. 2) are attached to the cytosolic side of the rough endoplasmic reticulum (ER) and to the outer nuclear membrane on its cytosolic face (Alberts 2002). Free and membrane-bound ribosomes are structurally and functionally similar, but, if the protein being synthesised on the ribosome has an N-terminal ER signal sequence, the signal directs the ribosome to enter the ER (Alberts 2002).

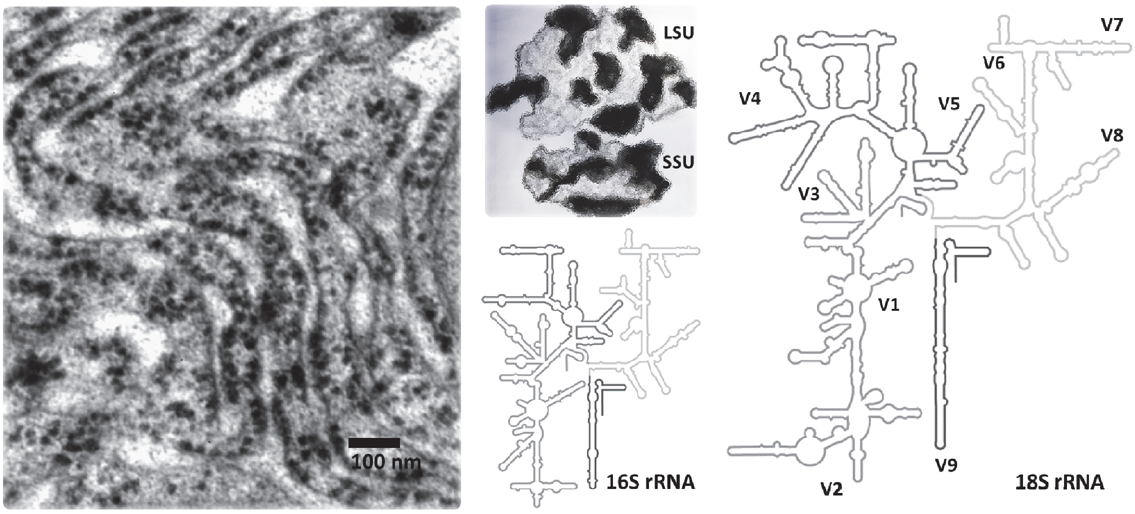


FIGURE 2 On the left is a transmission electron microscopy (TEM) image of ribosomes attached to the rough ER in a eukaryotic cell. The ribosomes, about 25–30 nm in diameter, look like black dots on the cytosolic face of the ER. In the middle is a schematic drawing of prokaryotic LSU and SSU of the ribosome, in which p-proteins are represented in black and rRNAs in grey; below the schematic is an illustration of the prokaryotic 16S rRNA secondary structure. On the right is a schematic of the secondary structure of eukaryotic 18S rRNA with variable regions.

Cellular organelles, mitochondria and chloroplasts, have their own ribosomes since, according to the endosymbiotic theory, they are organelles of endosymbiotic origin. Mitochondrial ribosomes have changed importantly when compared to the ribosomes of their alphaproteobacterial ancestors (Gray *et al.* 1999, Desmond *et al.* 2011). The significant responsibility of mitochondrial ribosomes is synthesising proteins for the oxidative phosphorylation system (Bieri *et al.* 2018). As for chloroplasts, their evolution can be traced to the endosymbiotic, photosynthetic prokaryotes, cyanobacteria (Bonen and Doolittle 1975, McFadden 2001, McFadden 2014). Chloroplasts have their own genome and the ability to synthesise protein, but the regulation of different processes is provided by many factors from the nucleus (Lyska *et al.* 2013, Bieri *et al.* 2017).

The ribosomes consist of an SSU (30S in prokaryotes and 40S in eukaryotes) and a large subunit (LSU, 50S in prokaryotes and 60S in eukaryotes). The prokaryotic SSU contains 21 r-proteins and 16S rRNA, whereas the LSU contains 34 r-proteins and 5S and 23S rRNAs, with some variations between species (Fig. 2). The eukaryotic SSU is composed of 32 r-proteins and 18S rRNA (Fig. 2), whereas the LSU is composed of 46 r-proteins and 5S, 5.8S and 25S/28S rRNAs, with some variation between species (Lafontaine and Tollervey 2001).

All rRNAs form secondary structures (Fig. 2) with base pairs, double-helices, loops, bulges and single-strands (Petrov *et al.* 2014). In the growing mammalian cell, only a small percentage of the dry weight is RNA, but about 80% of the total RNA pool is rRNA (Lodish *et al.* 2000, Alberts 2002). Unlike DNAs, single-stranded RNAs are prone to degradation. Intra- and extracellular ribonuclease (RNase) enzymes, with endo- or exonuclease activities, are enduring enzymes and are found ubiquitously to cleave RNAs (Yang 2011). Also, the structural features of an RNA molecule, such as single-stranded form and a reactive hydroxyl group on C_2 of the ribose sugar component, make RNA chemically labile (Lodish *et al.* 2000). The length of 18S rRNAs varies from about 1.5 kb to over 4.5 kb, while the average length of prokaryotic 16S rRNAs is about 1.5 kb (Neefs *et al.* 1990, Xie *et al.* 2011).

Classification of bacteria by 16S rRNA has been in practice for over four decades, and characterisation of eukaryotic 18S rRNAs has also continued for decades (Fox *et al.* 1977, Neefs *et al.* 1990, Wallner *et al.* 1993). In addition to identifying species by rRNAs, the gathered structural information and the knowledge of molecular mechanisms about ribosomes has broadened current understanding of translation dynamics in protein synthesis and knowledge of diseases caused by ribosome malfunction (Freed *et al.* 2010, Zhou *et al.* 2015, Henras *et al.* 2015, Jobe *et al.* 2018, Parks *et al.* 2018).

1.2.2 Post-transcriptional processing of eukaryotic rRNAs

Before 18S rRNA becomes a mature component of a functioning ribosome in a eukaryotic cell, it has undergone substantial processing (Fig. 3). This process requires over 200 conserved assemble factors, such as diverse RNA-binding proteins, endo- and exonucleases, RNA helicases, GTPases and ATPases (Pena *et al.* 2017). In bacterial cells, all RNAs are transcribed by the same RNA polymerase, but eukaryotic cells have specialised RNA polymerase I (Pol I) for producing rRNAs (except for 5S rRNA). Noncoding transcripts produced by Pol I, are neither capped nor polyadenylated (Alberts 2002). In a large 35S precursor rRNA transcript (Fig. 3), rRNAs are edged and separated by external transcribed spacers (ETS) and internal transcribed spacers (ITS), respectively. Chemical modification, cleavage of precursor to 20S/6S/25S rRNAs, and assembly of both large (60S) subunit and small (40S) subunit, occur in the nucleolus. The nucleolus is a non-membraned structure of the nucleus and the area for rRNA genes, rRNA processing enzymes, r-proteins made in the cytoplasm and the area for rRNA precursors synthesis.

Final assembly of 40S and 60S subunits into functional ribosomes happens outside the nucleus after each unit is individually transported through the nuclear pores to the cytoplasm (Alberts 2002, Pena *et al.* 2017). Departing from the other rRNAs maturation pathways, 5S rRNA is transcribed by RNA polymerase III. The gene is generally not located in the nucleolus, nor does the synthesis of the 5S rRNA molecule usually occur there (Ciganda and Williams 2011, Gibbons *et al.* 2015).

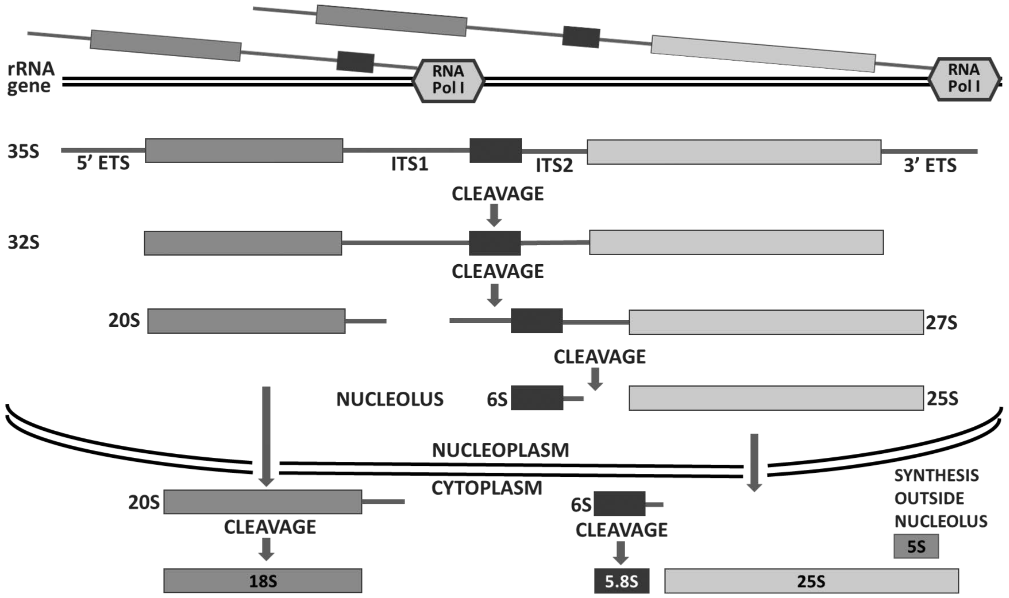


FIGURE 3 In a eukaryotic cell, the primary transcription of rRNAs by Pol I and initial maturation take place in a subnuclear structure, the nucleolus. In this figure, r-proteins are not shown, but the 40S pre-subunit, which is exported from the nucleus, consists of proteins and 20S pre-rRNA. In the cytoplasm, the pre-subunit matures to 18S rRNA containing SSU. Also, 25S and 5.8S rRNAs of the 60S subunit undergo sequential rRNAs processing before maturing to the LSU with proteins. Transcription and synthesis of 5S rRNA occur outside the nucleolus (not shown).

1.2.3 The bacterial 16S and eukaryotic 18S rRNA gene sequences

Because eukaryotic cells need approximately ten million ribosomes, the GCN of rRNA must be adequate. In human cells, about 200 rRNA gene copies per haploid genome ensure that ribosome synthesis can continue, whereas, in *E. coli*, only seven rRNA copies are necessary. In eukaryotic cells, rRNA genes, except the 5S gene, are located in the nucleolus and form humorous, fir-like tandem repeats when genes are transcribed (Condon *et al.* 1995, Alberts 2002). In bacterial cells, a single ribosomal RNA operon (*rrn*) contains a cluster of 16S, 23S and 5S rRNA genes with the internal transcribed spacer (ITS) region and tRNA genes. The GCNs of rRNA vary from one to 15 copies, and over 80% of sequenced bacterial genomes have multiple operons, which can be in chromosomes and plasmids. To avoid divergence, all 16S rRNA genes of bacterial cells should evolve in concert, despite the changes generated by mutation or horizontal gene transfer. However, in bacteria containing more than one *rrn*, polymorphism between intragenomic 16S rRNA genes is a frequent phenomenon, and, consequently, bacterial diversity may be overestimated from HTS library results (Klappenbach *et al.* 2001, Espejo and Plaza 2018). A publicly available, curated database (*rrn*DB) of *rrn* copy number variations is accessible for bacteria and archaea (Hein *et al.* 2014).

In eukaryotic cells, rRNA genes are organised into tandemly repeated units. From the schematic illustration of 35S precursor rRNA transcription (Fig. 3), the gene order can be inferred. GCN variations of rRNA genes between eukaryotic species are at a different level than those between prokaryotic species. When the quantitative polymerase chain reaction (qPCR) was used to determine the GCN variations of phytoplankton, GCNs varied from one to more than 1.2x10^4 among 18 strains (Zhu *et al.* 2005). It has also been shown that the number of rRNA gene copies is related to the genome and cell size and that copy number can vary from ten to over 10^4 in animals and plants (Godhe *et al.* 2008). When rRNA GCN variations of ciliate species were studied using qPCR, cloning and the sequencing of multiple clones, rRNA GCNs per cell were as high as 3.1x10^5, having much higher copy numbers than other protists and fungi. This can lead to an overestimation of the RA of ciliates in environmental samples based on sequencing results (Gong *et al.* 2013). When data of human and mouse genomes were analysed, GCNs varied widely across individuals, and intra- and interindividual nucleotide variations in rRNA genes were also found (Parks *et al.* 2018). Because intraspecific and intragenomic polymorphisms have been demonstrated among other species as well, non-concerted evolution of rRNAs may complicate diversity studies and phylogenetic reconstruction (Pereira and Baldwin 2016).

Sequencing of 16S and 18S rRNA genes has been an established practice for identifying microbes, but logical deduction of results among different studies is difficult. Although a consensus on the usage of 16S and 18S rRNA genes as molecular markers has been reached, it is still unclear which variable regions of the genes provide the best results for diversity studies. Nevertheless, recommendations have been published (Wu *et al.* 2015, Bradley *et al.* 2016). When the variable regions V1–3, V4–5 and V7–9 were compared using *in silico* PCR studies, the V1–3 region was recommended for HTS-based monitoring of planktonic eukaryote communities (Tanabe *et al.* 2016). When a mock community of 12 microalgal species was studied, the V8-V9 region was recommended (Bradley *et al.* 2016).

When the primer pair choice for diversity studies is in question, in addition to the initial *in silico* prediction of primer coverage, it would be beneficial to evaluate the effectiveness of primers using mock communities and to continue testing with environmental samples (Parada *et al.* 2016). Despite considering the notes above, the lack of broad-range primers for simultaneous amplification of 16S and 18S rRNA genes hinders comprehensive microbial diversity studies (Hadziavdic *et al.* 2014).

Additionally, e.g., in eukaryotic phytoplankton, 18S rRNA reference databases are still incomplete, and species level identification from HTS data is complicated (Le Bescot *et al.* 2016, Tragin *et al.* 2018). Although, since 1992, over two million SSU reference sequences of organisms (bacteria 1,861,569, Archaea 67,364 and Eukaryota 163,916) have been collected into the SILVA SSU Ref. v. 132 database, the sequences of the database represent only a small portion of all microbes (Quast *et al.* 2013).

1.3 HTS, molecular method for identifying microbes

Because most microbial species are unculturable in the lab, and because of the enormous species richness, only an incredibly small fraction of microbes has been characterised. Using an ecological theory of biodiversity, the estimated number of Earth's microbial species is 10^{11}–10^{12}, from which about 10^4 have been cultured, and less than 10^5 species are typified by classified sequences (Locey and Lennon 2016). The development of HTS method began a new era in microbiology, bypassing many conventional microbiological techniques, such as culture- and immunological-based methods or length polymorphism in PCR-amplified gene-segments based studies (Floyd *et al.* 2002, Boughner and Singh 2016). HTS methods allow fast, easy, sensitive and wide-ranging DNA and RNA analyses of samples consisting of cells from different species and complex, microbial communities, like human intestinal microbes. Environmental microbiological studies have also taken great advantage of HTS techniques. Completing genotypes revealing molecular data with analysis of phenotypes, knowledge of diversity and ecological roles of microbes is achieved even more efficiently. When protists, which could be identified according to their structures, were studied by comparing results from HTS with morphological data, the results partly agreed with and complemented each other (Santoferrara *et al.* 2016). If morphological variation between species is not distinguishable, HTS methods can bring real benefits to characterising microbes.

1.3.1 Ion Torrent and other sequencing methods

Sanger sequencing, developed in the 1970s, was considered the proverbial 'gold standard' for DNA sequencing before HTS techniques were developed. So-called 'first generation sequencing', Sanger sequencing is named after its developer, Frederick Sanger (Sanger *et al.* 1977). The technique is based on incorporating labelled dideoxynucleotides with the elongated fragment in a replication reaction. The chain termination system allows orderly detection of labelled nucleotides (A, T, C or G bases) according to the size of the fragments. The method is still applied, although its use has many limitations. Only a few hundred sequences are brought out by one Sanger sequencing run, and samples must contain clonal templates that make it impossible to obtain sequences directly from multi-species, environmental samples. The technique is time-consuming requiring many steps, such as microbe culturing and, in many cases, cloning. Although drawbacks exist in Sanger sequencing, Sanger results are used in clinical diagnostics as HTS results are prone to difficult region errors, such as homopolymer stretches and pseudogenes (Weiss *et al.* 2013, Xue *et al.* 2014, Mu *et al.* 2016, Roy *et al.* 2018). Sanger sequencing reads are also longer than HTS reads, 600–800 bp and 100–400 bp, respectively (Mardis 2017).

HTS techniques, i.e. next generation sequencing (NGS) or 'second generation sequencing', enable simultaneous, parallel sequencing from samples comprised of various species. Sequencing by synthesis (SBS) refers to DNA-polymerase-dependent methods that contain single-nucleotide addition (SNA) principles. Incorporation of a certain nucleotide can be detected, either by a fluorophore or by the ionic change. In the sequencing by ligation (SBL) method, a probe sequence with a fluorophore, which indicates a certain base, hybridises to a target template and, after imaging, identifies one of the four bases (Goodwin *et al.* 2016).

Pyrosequencing is an SBS method based on the enzymatic reaction during elongation of replication (Nyren and Lundin 1985, Nyren *et al.* 1993, Ambardar *et al.* 2016, Heather and Chain 2016, Alesheikh *et al.* 2018). In this technique, the DNA sequence is obtained from the release of the inorganic pyrophosphate during the elongation of a nascent DNA fragment and the subsequent enzymatic luminometric observation. After pyrosequencing, numerous HTS sequencing techniques, which are also based on SBS, have been developed.

Commonly used HTS platforms are Illumina, ABI SOLiD, Qiagen, 454 Life Sciences / Roche 454, Ion Torrent Personal Genome Machine (PGM) / Ion Proton and Pacific Biosciences (Liu *et al.* 2012, Reuter *et al.* 2015, Goodwin *et al.* 2016). Roche's 454 GS FLX+ HTS system increased the read length up to 1000 bp before this 454 pioneer of HTS systems exited the stage. Roche's services were discontinued between 2013 and 2016 (El-Metwally *et al.* 2014, Del Vecchio *et al.* 2017).

In 2010, Ion Torrent introduced a new sequencer, which has no expensive optics and in which real-time sequencing is based on ion detection (Rusk 2010, Merriman *et al.* 2012). All HTS studies in this thesis have been conducted using an alternative technology for pyrosequencing, the non-optical sequencing of Ion Torrent, which is based on SBS and SNA principles (Sakurai and Husimi 1992, Hizawa *et al.* 2006, Rothberg *et al.* 2011, Quail *et al.* 2012). Templates for PGM runs are prepared using the Torrent One Touch 2 system, which consists of emulsion PCR (emPCR) and enrichment of templated beads (ER) machineries. The new Ion Torrent template preparation instrument, Ion Chef, and the Ion S5 sequencer allow even more automated workflows.

In Ion Torrent HTS, DNA fragments to be sequenced (HTS library) are clonally amplified to the Ion sphere particles (ISP) using emPCR, and ISPs, coated with the template, are deposited into their own microwells in a small (2.5 cm x 2.5 cm size) sequencing chip. The number of wells in the Ion PGM sequencer chip depends on the chip choice: one million, six million or 11 million wells. During the PGM run, a flow of nucleotides, A, C, G and T, is streamed over the wells sequentially, one at a time. When a deoxyribonucleoside triphosphate (dNTP) is incorporated into a growing DNA strand, hydrogen ions are released. The pH of the surrounding solution in the well is changed after the hydrogen release, and the pH shift is detected by the sensor on the bottom of each well. The pH shift is converted into a voltage, and, in a subsequent signal-processing software, raw, digitised voltages are changed to the base calls (Rothberg *et al.* 2011).

According to the description of HTS models from Alesheikh *et al.* (2018), the exact definitions dividing 'second and third generation sequencing' technologies have not been underlined, albeit single molecule sequencing seems to be the current practice in the latest sequencing advancements. Nanopore sequencing has provided a new approach to the field without the necessity for template amplification, without the SBS method, and with real-time detection of the bases that pass through the nanopore (Traversi *et al.* 2013, Deamer *et al.* 2016, Parker *et al.* 2017, Jain *et al.* 2018, Mojarro *et al.* 2018). Nanopore can be either biological, with a limited lifetime, or solid-state, with more robustness and stability but with the analysis complicated by the ultrafast DNA fragment passing through the nanopore (Goto *et al.* 2016, Lepoitevin *et al.* 2017). The portable instrument, long sequencing reads, and low cost are some advantages of nanopore sequencing, while high error rates compared to earlier HTS methods have limited the use of nanopore sequencing techniques (Rang *et al.* 2018).

As sequencing technologies continue to develop, and the lengths of fragments to be sequenced are increasing up to hundreds of kilobases (kbp), particularly in nanopore sequencing, the nature of diversity studies is changing (Jain *et al.* 2018, Minei *et al.* 2018). Targeted, marker gene-based identification of species is changing so that the whole marker gene (Wurzbacher *et al.* 2019), or even entire genomes, are sequenced, and, accordingly, the resolution of identification can rise to the next level.

1.3.2 Construction of library for HTS

Compared to Sanger sequencing, basic HTS library preparation can be done surprisingly effortlessly. After DNA or RNA is isolated from samples, along with complementary DNA (cDNA) synthesis from RNA samples, only a few steps are needed to complete the basic library construction, and a library is usually ready for sequencing within two days (Mardis 2017). Albeit relatively easy to conduct, preparing a quality library is a critical step in achieving reliable HTS results, and sample-specific requirements can complicate processes substantially. Commercial kits for library preparation are available, but they are usually costly and occasionally impractical.

All HTS methods have one significant advantage: multiplexing, which means that barcode-tagged DNA templates of various samples can be pooled before sequencing (Wong *et al.* 2013, Oliveira *et al.* 2018). Sequencing efficiency and affordability per sample depend on the number of barcodes in use. The full advantages of HTS can be exploited more competently when up to 100 samples with barcodes are incorporated in a single HTS run (Smith *et al.* 2010). Barcodes are unique, short sequences (i.e., Ion Torrent barcodes are approximately 10 nt long), which are tagged to each DNA sample before the samples are pooled. The terminology here is confusing, and it is important to distinguish sequencing barcodes from DNA barcodes, which were discussed earlier in this research. Barcoding (i.e., indexing) can be done through commercial indexing kits or by specific barcodes especially designed for an individual study. The sequencing results, reads, can be demultiplexed using bioinformatics tools, and, consequently, each sample is identified (Mir *et al.* 2013).

After DNA or RNA extraction, basic HTS library preparation steps include: enzymatic, physical or chemical fragmentation of DNA or RNA, incorporating sequencing adapters and barcodes into the DNA fragments by ligation or using fusion primers in PCR, cDNA synthesis, size selection, primer dimers fragment removal and library amplification (Head *et al.* 2014). Detailed protocols vary, depending on the HTS platform (e.g., commonly used Illumina or Ion Torrent) and the objects to be sequenced. Library quality controls between each library preparation step are crucial, especially when new library construction methods are evaluated or developed. Frequently used, time-saving and versatile instruments for quality testing are microfluidic capillary gel electrophoresis instruments (e.g. Agilent Bioanalyzer and TapeStation).

A basic, library construction workflow for Ion Torrent sequencing (Fig. 4) includes the abovementioned steps. Additionally, template preparation for Ion Torrent sequencing by PGM sequencer includes clonal amplification of library templates into the ISPs using the emPCR machine, Ion OneTouch. The surfaces of ISPs are coated with immobilised, oligonucleotide fragments. The oligos prime the amplification as the sequence of the oligos is complementary to the sequence incorporated into the library templates. After emPCR, templates are physically attached to the ISP. The concentration of templates, i.e., the nano molarity of library sequences in emPCR, is critical to gaining clonally amplified ISPs and minimising the proportion of polyclonal reads in the final sequencing reads. After emPCR, templated ISPs are collected using a magnetic, particle-based enrichment machine, OneTouch ES (Enrichment System).

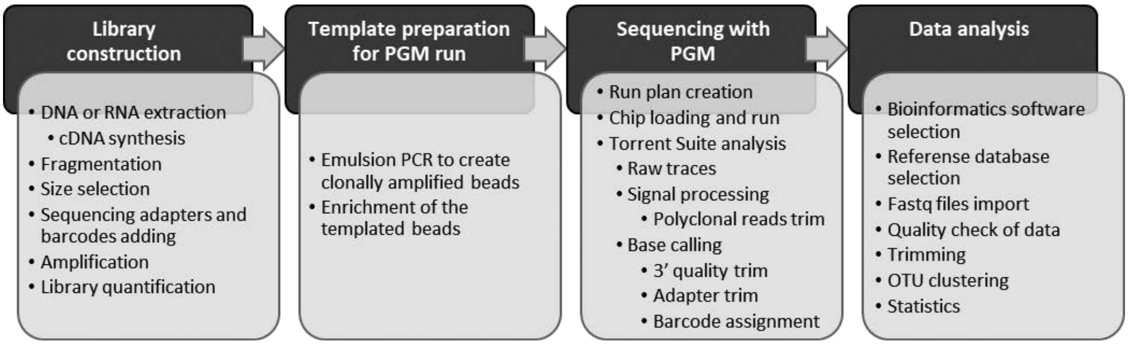


FIGURE 4 An overview of the HTS workflow for the Ion Torrent platform. Rearranging the library preparation steps order offers alternatives to standard library construction methods.

Since HTS methods came into the market, multiple variations of library preparation methods have been put forth by both researchers and commercial producers, and multiple, comparative studies of protocols have been published (Quail *et al.* 2012, Song *et al.* 2015, Bowers *et al.* 2015, Ng *et al.* 2018). As HTS methods are commonly used to study different types of organisms, many variations in library construction processes are necessary. However, a disadvantage to varying protocols is that comparing results obtained from different studies might be difficult. From nucleic acid isolation to data analysis, all steps may have their own specific impacts on HTS results and are potential causes of bias (Poptsova *et al.* 2014, Brandariz-Fontes *et al.* 2015, Ali *et al.* 2017).

1.4 HTS data analysis of microbes

1.4.1 Metagenomics and metatranscriptomics

Analysis of all HTS data requires computational tools and computing strategies. Bioinformatics softwares have an essential role in converting sequencing results to logical conclusions. Because millions of sequences can be obtained from a single HTS run, the enormous data volume has increased the requirements for computing power and skills. Rapid advances in sequencing methods and accumulation of the vast amount of sequencing data have made bioinformatics a significant field in biology.

HTS of microbes can be based on two different metagenomics approaches, shotgun metagenomics and targeted metagenomics (Siegwald *et al.* 2017). In untargeted, shotgun metagenomics, usually the whole genomic content of the extracted DNA is sequenced, and the results unveil the taxon of microbes, as well as the functional diversity of microbes. Shotgun sequencing can be used to recover whole genome sequences (Quince *et al.* 2017). The expense, huge data size and data complexity hinder shotgun metagenomics analysis. Whereas targeted, amplicon-based metagenomics refers to a taxonomically informative marker gene usage that significantly reduces data size, costs and computing power requirements (Siegwald *et al.* 2017). Yet metagenomics expression (in targeted metagenomics) can be considered a misnomer because the entire genomic content of the sample is not sequenced (Quince *et al.* 2017). All sequencings in this thesis have been targeted towards certain marker gene fragments, which originate either from isolated DNA or RNA, and, accordingly, data analysis methods are focused on the HTS results for the marker fragments.

Metatranscriptomic analysis covers the study of rRNA and mRNA of the microbial community from an environmental sample. Although studies of metatranscriptomics usually include removal of rRNA from isolated RNA to enrich protein coding sequences, focusing on rRNA would provide information on the actively synthesising organisms (Tveit *et al.* 2014, Petrova *et al.* 2017).

1.4.2 Quality of the HTS data and FASTQ files

The quality of sequencing reads (sequencing results) forms the basis of data analysis (Brockman *et al.* 2008, Loman *et al.* 2012). For SBS technologies, base-by-base error predictions, Phred quality scores and identifying high-quality bases have been developed, resulting in accurate quality scoring methods for SBS technologies (Ewing and Green 1998, Ewing *et al.* 1998, Brockman *et al.* 2008, Cock *et al.* 2010). In the 1990s, the original software, Phred ('Phil's read editor'), was developed by Phil Green for automatic reading of the fluorescent sequence chromatograms from Sanger sequencing (Rifai *et al.* 2018). The quality value (Q-score) of the error probability of a base call comes from the formula: $q = -10 \times \log_{10}(p)$, where q = quality value, and p = estimated error for a base call. Increasing Q-score value means a higher probability of the correct call. For example, for the base of Q-score 20 (Q20) the incorrect call probability (*p*) is 1/100, and, for Q30, *p* is 1/1000. According to the technical notes for Ion Torrent, the per-base quality score determination of an Ion Torrent read is based on a Phred-like method, which predicts the probability of right base call. The quality of the base incorporation signal is the prediction basis when base calls are generated.

After a PGM run, Torrent Suite software operates quality controls on sequencing reads before they are suitable for export. Polyclonality filtering is performed for sequencing results to exclude reads originating from polyclonal ISPs. In addition, 3′-end, low-quality regions of the reads are trimmed off, because the highest quality base calls are at the beginning of the reads and deteriorate towards the 3′-end of the reads (Loman *et al.* 2012). Alternatively, the reads can be exported without trimming or with adjusted trimming demands in the Torrent Suite software. After initial data trimming by Torrent Suite, the sequencing reads can be downloaded from the PMG server in the FASTQ file format, a text output file with the quality scores for each base. The FASTQ file format is an extended form of the FASTA file, which was originally developed for the FASTA suite tools by Bill Pearson in the late 1980s (Pearson and Lipman 1988, Cock *et al.* 2010).

1.4.3 Bioinformatics from FASTQ files to a community composition

To convert information from FASTQ files to logical conclusions, bioinformatics tools are needed for demultiplexing, quality trimming, denoising, chimera filtering, operational taxonomic unit (OTU) clustering and taxonomic assignment. Many commercial and open-source bioinformatics software are available, as well as reference databases for classification studies, but the choice of a tool and a reference database may be difficult (Nilakanta *et al.* 2014). Comparing of commonly used open-source bioinformatics tools, Qiime and Mothur, indicates that OTU clustering algorithms and algorithms for taxonomic classification differ from each other, along with chimera detection (Schloss *et al.* 2009, Caporaso *et al.* 2010, Lopez-Garcia *et al.* 2018). The results from these two tools highly agreed when the most abundant genera of 16S rRNA amplicon sequences of rumen microbiota composition were analysed using either SILVA or GreenGenes databases. The important differences came from the analysis of less frequent microbes when GreenGenes was applied as a reference database. These and other findings suggest that the choice of bioinformatics tools and databases can have a relevant impact on HTS inferences (Lindgreen *et al.* 2016, Lopez-Garcia *et al.* 2018).

Quality trimming is an important step for filtering reads that do not contain correct primer sequences, eradicating amplification primers from sequences, eradicating reads that are too short, defining maximal homopolymer length and eradicating low-quality bases. However, excessive quality-based trimming might bias the sequencing results, and moderate trimming is suggested to be optimal for many studies (Macmanes 2014, Williams *et al.* 2016).

Constructing OTUs is a common bioinformatics practice, and the earlier definition, molecular operational taxonomic unit (MOTU), has changed to classify closely related individuals (Blaxter 2004, Callahan *et al.* 2017). Although taxonomic classification could be done for trimmed reads with a sequence search algorithm BLAST (Basic Local Alignment Search Tool) and without OTU clustering, the data volume without OTU clustering is enormous for aligning all sequences (Edgar 2010). In 16S rRNA HTS data, OTU clustering identity levels can be assumed so that 97% defines species level, >95% defines genus level, and >80% defines phylum level, although these presumptions are controversial (Schloss and Handelsman 2005). Nevertheless, the commonly used 3% dissimilarity threshold should be carefully considered to avoid underestimation of OTU numbers (Chen *et al.* 2013). As evolution rates of bacterial lineages differ, evaluating evolutionary relationships only by sequence similarity can lead to non-monophyletic OTUs (Koeppel and Wu 2013). To assign sequences to OTUs, sequence trimming, denoising and chimera removal should be done beforehand to decrease difficulties in OTU clustering and decrease the number of artificial OTUs created (Schloss 2012). Increased sequencing error rates at the ends of reads can artificially cause too many OTUs. Result inconsistency can also originate from using an unsuitable reference database or an unsuitable OTU clustering algorithm and parameters (Bracciali *et al.* 2017, Golob *et al.* 2017). After obtaining a reference-based OTU clustering table, RAs of species can be inferred from the RAs of the reads. An open-reference OTU clustering refers to reference-based OTU picking, followed by de novo OTU picking from those sequences that differ too much from the references. A closed-reference clustering indicates reference-based OTU assignment without de novo OTU picking (Westcott and Schloss 2015).

2 AIMS OF THE STUDY

The main aims of this research were to evaluate and produce molecular methods for better characterisation of microbes, especially phytoplankton, and to apply these new methods to environmental samples. As the library preparation is a crucial step in HTS success, the studies focused on developing practical workflows for DNA- and RNA-based HTS library preparations. Bioinformatics workflows were planned along with the studies.

Four general questions triggered the studies:

1. What kind of DNA-based, HTS library preparation workflow could promote a cost-effective and practical sequencing of any gene, regardless of gene size, focusing the sequencing to the 5′-end of the genes?
2. Can time-consuming, microscopic identification of phytoplankton be replaced with rapid, automated, molecular methods, and what are the best choices for preparing HTS libraries for phytoplankton?
3. What kind of RNA-based, HTS library preparation workflow could promote primer-independent characterisations of microbes from all domains of life?
4. How equally do morphological (microscopic) and molecular (HTS) identification methods reveal phytoplankton community compositions in Finnish lakes?

To answer the first general question, a workflow for DNA-based library preparation was planned for archaeal 16S rRNA gene fragments from an environmental sample. The specific aim of the method setup was to achieve fluent barcoding and size trimming of HTS templates using gene-specific primers (I). An evaluation of the new method was planned so that equimolar concentrations of HTS libraries from four different template preparations were compared according to HTS results. One additional test targeted the standard, Ion Torrent adapter ligation protocol to compare it to our new method using 18S rRNA genes from phytoplankton. This method aimed to target the beginning of the sequencing to the same region so that the OTU clustering could be done without complications.

The second question justified a large-scale, comparative research design of present sample preservations, DNA and RNA extractions and HTS data analysis methods for phytoplankton (II). The sub-question was: 'How do different procedures affect the results of the HTS analysis?' The study design included forming a mock community pool of six phytoplankton strains, with variation in nucleus size and cell wall hardness. The HTS results of community composition could be compared to biomass and carbon content values so that the biasing molecular methods were traceable. Bioinformatics pipelines were also under evaluation.

To answer the third question, and to consider the RNA feasibility in the community composition identification studies, RNA-based, primer independent, library construction technique was planned for SSU rRNA (III). A phytoplankton mock community was used as a model target group for method validation. The aim was to achieve RAs of all living microbes from the mock community pool, simultaneously from the eukaryotic phytoplankton as well as from prokaryotic cells, in the same HTS library and run.

To determine an answer for the fourth question, phytoplankton community compositions in Finnish lake water samples were the objectives of the microscopic and RNA-based genetic identification studies (IV). The results of microscopic and genetic studies were subjects in a method-comparison. When the new, RNA-based, gene-specific, primer independent library preparation method was applied for HTS, all microbes from lake samples were simultaneously under investigation. This simultaneous sequencing of all domains of life was an ambitious goal for molecular identification.

3 MATERIALS AND METHODS

3.1 Summary of methods

The summary of the methods applied in this study is presented in Table 2. Detailed descriptions of reagents required for the methods can be found in the original publications.

TABLE 2 Methods applied in this thesis. A detailed account of the materials and methods used in the studies are described in the original publications.

Method	Publication
Microscopy	II, IV
Preservation	II, IV
Dry mass and carbon content determination	II
DNA and RNA extraction	I, II, III, IV
RNA concentration	II
PCR, qPCR and emulsion PCR	I, II, III, IV
cDNA synthesis	II, III, IV
AGE	I, II, III, IV
Fragments purification from the gel	III, IV
Fragmentation	I, II
Size selection	I, II, III, IV
Ligation	I, II, III, IV
Concentration measurement	I, II, III, IV
Purification of DNA and RNA fragments	I, II, III, IV
Sanger sequencing	II
Cloning	II
Ion Torrent HTS	I, II, III, IV
Bioinformatics and statistics	I, II, III, IV

3.2 OTU clustering principles

In this thesis, the de novo OTU picking in Mothur v.1.36.1 was accomplished via the average neighbour algorithm and, in CLC software (www.qiagenbioinformatics.com) software, by the distance-based greedy algorithm UCLUST (Schloss and Handelsman 2005, Schloss *et al.* 2009, Edgar 2010). When CLC software was applied, the de novo OTU picking method and reference-based OTU clustering method were used. In de novo OTU picking, reads clustering was based on sequence similarities, and taxonomic assignment was done for de novo OTUs afterwards via BLAST. In the reference-based method, reads were clustered against a reference sequence in the selected reference database, and, for reads that could not find a reference within the assigned similarly percentage, chimera crossover was checked. The reads that could not hit reference sequences were clustered with each other to detect de novo OTUs; 'allow the creation of new OTUs' was chosen in settings.

3.3 Description of samples

Detailed descriptions of samples and cells used in this study can be found in the original publications, shown in Table 3. Table 4 lists the phytoplankton mock community species used in Papers II and III.

TABLE 3 The original publications in which a detailed account of the cell cultures or environmental samples is described.

Samples	Publication
Environmental samples	I, IV
Cell cultures	II, III

TABLE 4 Cultured, mock community species used in the studies for Papers II and III.

Phytoplankton species	Division
Diatoma tenuis	diatom
Melosira arctica	diatom
Apocalathium malmogiense	dinoflagellate
Kryptoperidinium foliaceum	dinoflagellate
Monoraphidium sp.	green alga
Chlorella pyrenoidosa	green alga

4 RESULTS AND DISCUSSION

4.1 A new, DNA-based method for HTS library construction (I)

Prior to HTS, fragments to be sequenced should have the correct size, indexing barcode, and forward and reverse sequencing adapters. Many commercial library preparation protocols are available, but new, adaptable and low-priced alternatives offer freedom of choice (Bowers *et al.* 2015, Ng *et al.* 2018). This DNA-based, practical barcoding and size trimming method allowed sequencing of the of 5′-end regions of amplified genes when the gene was longer than can be utilised in the sequencing approach. A two-step PCR approach, with the help of gene-specific fusion primers (with universal M13 sequence overhang), allowed the use of M13-containing barcodes for countless genes.

The results verified that applying the newly-developed method offered a clear advantage for phylogenetic analysis. When the beginning of sequencing was targeted to the same 5′-end region of the amplicons, subsequent OTU clustering of the fragments could be done with maximised efficiency. The results showed that this new library preparation method selected fragments with full 5′-ends and, therefore, led to more accurate taxonomic classifications of OTUs than the standard method. In addition, amplicons from different samples could be pooled after barcoding PCR. This enabled the fragmentation, adapter ligation, size optimisation and purification steps to all be performed at once, in 'one tube principle'. The reduced library preparation costs are seen in Table 5 (Prices are from the date Paper I was written). Time and labour saving come as a bonus when the pooled samples are trimmed and purified simultaneously.

In HTS, read length and quality are important factors when comparing the results of library methods (Loman *et al.* 2012). The results of fragmented amplicons showed that the average read-length of the archaeal 16S rRNA gene and the archaeal methyl-coenzyme M reductase gene (*mcrA*) reads was correct before Torrent Suite quality trimming (I, Fig. 2A and B) and after trimming (I, Fig. 2E). Sequencing oversized (>500 bp) archaeal *mcrA* gene fragments yielded the lowest number of good quality ≥Q20 bases and revealed that fragmentation is needed for long amplicons. The highest average percentage of the bases, whose quality scores were ≥Q20, was obtained applying the new method and fragments from the archaeal *mcrA* gene (I, Fig. 2F).

TABLE 5 An example of the cost of reagents for common library construction for 50 individual reactions compared to the new 'one tube principle' method. *Kit contains Ion Plus Fragment Library Kit and Ion Shear Plus Reagents Kit. **Each barcode kit is enough for preparing ≤10 libraries per barcode for 100ng DNA input. The price for one sample has been calculated as follows: 7473€/96/10 = 7.78 €. ***The price of one sample is calculated as follow: 372€/7/15 = 3.54 €

Product name	Package / reactions	Price €	Price € for 50 reactions	Price € for one reaction
*Ion Xpress Plus Fragment Library Kit	10 reactions	868	4,340	87
**Ion Xpress Barcode Adapters 1–96 Kit	96 barcodes and P1-adapter	7,473	389	barcoded M13 <1 €
Agencourt AMPure XP	60 ml (606 reactions)	969	80 € x 2	2 € x 2
Pippin Prep	10 x 4 wells (one marker well)	613	766	15
***TapeStation HS D1000Screen Tapes and reagents	7 screen tapes 16 wells/tape (one marker well)	tapes: 278 reagents: 94	177	4
Total cost			5,832	110

The proper length of the sequencing reads tells that the quality of the 3′-end bases was good enough to produce, on average, 350 bp fragments (I, Fig. 2E). Without trimming in Torrent Suite software, the size of the fragments exceeded the favoured length of 400 bp (I, Fig. 2A and B). In Ion Torrent sequencing, the signal originating from the incorporation of dNTP into a growing DNA strand and release of hydrogen ions, deteriorates towards the 3′-end (Loman *et al.* 2012). By default, Torrent Suite software detects the low-quality bases from the 3′-end and trims them off, but the 3′-end trimming is optional.

When library construction amplification was done with long fusion primers, the polyclonal signal detection required an adjustment in the check region. If all fragments in an ISP are amplified from a single template (monoclonal), coherent, base-incorporating signals are obtained. The technical notes for Ion Torrent describe that positive signals come from about 44% of all nucleotide flows. So, in a well with clonal ISP, more than half of the flows yield zero signals, and the rest yield positive signals. Positive signals cluster around integer values. If an ISP is covered with two distinct populations (polyclonality), only half of the templates yield signals, and one nucleotide incorporation yields a signal value of about 0.5 instead of an integer value of 1.0. The zero signal flows are less frequent in polyclonal cases than in the clonal bead. The Torrent Suite algorithms calculate scores for each well, scores based on the percentage of non-zero signals and scores based on the degree to which signals have integer values. In the Ion Torrent software (Torrent Suite 4.2.1), polyclonality is, by default, checked from flows 12–70 because the detection of integer values is most clearly seen in the earliest flows. However, using extended primer lengths, filtering should be based on later flows, e.g., flows 120–160. In long fusion primer cases, all templates have equal sequences at the very beginning of the fragment, and detection of ambiguous, polyclonal bead signals works from the later region.

This new HTS library construction method (I, Fig. 1), in which the M13 sequence was exploited and the order of procedure steps was rearranged compared to common procedures, worked fluently. The new library construction method offers a cost-effective and time-saving protocol for any gene, regardless of gene size, focusing the sequencing to the 5′-end of amplicons. The method allows practical HTS library construction using earlier established and tested primer pairs, regardless of fragment length, since the protocol allows practical size trimming.

4.2 Validation of molecular methods for phytoplankton community (II)

Constant validation of HTS library preparation methods is needed to ensure the quality of sequencing information. A mock community with defined RAs of biomasses, carbon content and GCNs provided a useful baseline for evaluating how different preserving, DNA and RNA isolation and data analysis procedures (II, Fig. 1) affected HTS analysis results. A mock community of six phytoplankton species consisted of *Diatoma tenuis*, *Melosira arctica*, *Apocalathium malmogiense*, *Kryptoperidinium foliaceum*, *Monoraphidium* sp. and *Chlorella pyrenoidosa*, species with variation in cell wall hardness, cell size and nucleus size (II, Fig. 2). The ultimate question was, can microscopic identification of phytoplankton be replaced with molecular methods, and what are the preferred methods for preparing an HTS library for phytoplankton.

The results indicated that the primer pair Euk1A F/Euk516 R was suitable for amplification of all six phytoplankton species, whereas the V8 F/1510 R primer failed to amplify *Chlorella pyrenoidosa* (II, Fig. 3A). When environmental samples are investigated, designing proper primer pairs for non-biased amplification of all species might be an impossible task. In this comparative research, it was enough that all mock community species were successfully amplified with the Euk1A F/Euk516R primer pair, so it was chosen for the study.

The results showed that species-specific variations in 18S rRNA GCN biased the RAs of species in the final HTS results. Among the mock species, an over 10^4-fold variation in GCNs per cell was found (II, Fig. 3D). The GCN of *Apocalathium malmogiense* was the highest, at 3.3×10^4 per cell, and it dominated the profiles of RAs of DNA-based HTS analysis, although RAs of masses (dry and wet) and carbon content of *Apocalathium malmogiense* were less than 10% of the corresponding values of all species (II, Fig. 6A and C). To exclude the bias caused by DNA isolation procedures, particularly the cell lysing efficiency, 4 ng of separately isolated DNA from each species was pooled and sequenced. In this sample, too, the dominance of *Apocalathium malmogiense* was superior, as well as in the theoretical template relationships (TTR) sample, in which the 18S rRNA GCN of each species was specified separately from the same volume of isolated DNA (II, Fig. 6D).

The results indicated that a frequently occurring sequence in HTS results only represented a high GCN of the taxon and not the real, lesser RA of that taxon. Even among prokaryotic cells, which have only a few 16S rRNA GCNs, community structure inferences can be biased if sequence abundances are thought to represent the real organismal abundance. Better correlation between true organismal abundances versus 16S rRNA gene abundances was achieved by software estimating the genomic 16S rRNA GCN of taxa using phylogenetic reference data of sequenced genomes and, accordingly, defining the final estimations of community compositions more precisely (Kembel *et al.* 2012, Angly *et al.* 2014). However, recent studies by Louca *et al.* (2018) suggest that default correction of the 16S GCNs in microbiome inspection should be neglected, unless OTUs are closely enough related to sequenced genomes or unless a requirement for corrected OTU proportions justifies the additional noise produced in the data. At present, the imperfection of reference data still rules out the possibility of recovering data using knowledge of GCNs for eukaryotic 18S rRNA genes. After completion of reference databases, it is expected that tools that correct GCN bias will be frequently used in diversity studies, and the accuracy of amplicon-based RA estimations will be improved (Angly *et al.* 2014).

When HTS was based on RNA-based sequencing, the bias caused by GCN variation among species was avoided, and the proportions of mock species in the final HTS analysis were more realistic, when biomasses and carbon contents were used as indications of realistic RAs (II, Fig 6B and C). This result was logical because the number of ribosomes scales with cell size (Marguerat and Bahler 2012). Based on a non-metric multidimensional scaling (NMDS) ordination, RNA-based sequencing and Direct-zol extraction kit usage yielded the most realistic RAs of species (II, Fig. 8A). Accordingly, RNA-based HTS is recommended for phytoplankton 18S rRNA identification studies.

The HTS data trimming pipeline was optimised using model data, in which each mock community species was amplified separately with unique barcodes, and an equal number of amplicons from each species was pooled and sequenced. The HTS results of the model data showed equal RAs of mock community species after Torrent Suite 3′-end trimming and without 3′-end trimming (II, Fig. 4A). Too tight quality thresholds in Mothur software led to twisting in the RAs, but, by lightening the demands, more equal RAs of species were gained (II, Fig. 4B). Reads without 3′-end trimming in Torrent Suite and moderate trimming using CLC software gave the most equal distribution of the model data, and, therefore, this pipeline is recommended for phytoplankton bioinformatics (II, Fig. 4B).

In quality trimming, CLC software uses the modified-Mott trimming algorithm, where the quality score (Q) of each base is converted to an error probability (*p*), so low values are high-quality bases. Next, every base will receive a new value, which is a quality limit value (by default 0.05) minus p_{error}. For low-quality bases, this value is negative. CLC software calculates the running sum of this value from base to base, and zero is the lowest value. The valid fragment starts at the last zero value before the highest score and ends at the highest value of the running sum. So, an occasionally occurring moderate drop in base quality does not necessarily trim off the downstream part of the fragment, if the quality of the next bases continues to increase. This was the crucial quality check in the CLC trimming algorithm, which successfully retained the original distributions of model data reads (II, Fig. 4B). Using the MOTHUR trim.seqs command, quality trimming with too low qwindowsize parameters and too high qwindowaverage parameters (over a window) may erase fragments that exhibit an occasional drop in the quality score. The rRNA fragments are prone to form secondary structures, such as loops, which may disturb the exactly synchronous function of DNA polymerase in the PGM, SBS system. This asynchronous replication event may cause a slight delay in the signal and a casual quality drop in the base call. In some sequences of FASTQ files, an occasional quality drop was seen in the base-paired loop region (at around nt 150 position). A casual occurrence of homopolymers can also lead species-specific erasing from the data if too tight quality thresholds are assigned.

Whether microscopic identification of phytoplankton can be replaced with HTS is still an open question, and further method validation is needed to ensure unbiased HTS results. The results in this study verified that sample preservation, genetic material choice, extraction kit choice and bioinformatics pipeline choice all impact HTS results (II, Fig. 6). Recommendation of RNA-based HTS, rather than DNA-based HTS, and suggestions for the data trimming pipeline have been brought forward for phytoplankton HTS. However, the choice of methods in sample processing and bioinformatics pipelines are case-specific, according to the study species and software tools. Because even small changes in processes can critically affect results, detailed pipelines, including OTU generating strategy, clustering algorithm and reference database in use, should be specified in the method section of all diversity studies to avoid the illusion that standard procedure settings are optimal (Golob *et al.* 2017).

4.3 RNA-based, primer independent HTS (III)

As RNA-based HTS gave the most realistic RAs of the mock community, and since the real universal primers for simultaneous HTS of prokaryotes and eukaryotes are not available, a new, primer independent HTS library method was developed (III, Fig. 1). For the development and evaluation of this new method, the phytoplankton mock community species *Diatoma tenuis*, *Melosira arctica*, *Apocalathium malmogiense*, *Kryptoperidinium foliaceum*, *Monoraphidium* sp. and *Chlorella pyrenoidosa* were utilised, since the RAs of biomass and carbon content values for the mock community were determined earlier (II), although symbiotic prokaryotes were still unidentified (Singer *et al.* 2016).

Noncoding mature 16S and 18S rRNAs are not known to have a 3′-end poly-A tail, although polyadenylation of certain rRNA intermediates during post-transcriptional processing has been reported (Kuai *et al.* 2004, Slomovic *et al.* 2006, Hang *et al.* 2018, Fleischmann *et al.* 2019). However, amplification plan could not be based on a common oligo(dT) primer annealing to the poly-A tail. Thus, our first strategy rested on terminal nucleotidyl transferase (TdT) enzyme usage in poly(A) tailing, which was done to gel-extracted and reverse transcribed (random priming with P1 overhang) cDNA fragments of 16S and 18S rRNAs. Despite the optimisation of reaction conditions and anchored oligo(dT) primer usage, amplification frequently yielded truncated fragments through internal poly(A) priming. In addition, an unresolved challenge related to the length of the poly(A) tail generated by the TdT enzyme was encountered. So, the next strategy was based on a ligating reaction, and this strategy led to a working method. The library construction was based on ligation of the M13-RNA oligo to the 5′-end of the gel-extracted rRNA fragments purified from E-gel. cDNA synthesis was primed using a random hexamer primer with a sequencing adapter P1 overhang. After amplification, using barcoded Ion Torrent (with M13 sequence) and P1 primers, amplicons were purified and size-selected with magnetic bead purification in a one-step procedure (III, Fig. 1). The HTS results of a phytoplankton mock community showed a very similar profile to the RAs of biomass and carbon content values, and prokaryotic cells in cultures were identified simultaneously (III, Fig. 2).

The M13-RNA oligo concentration was critical in the ligation reaction to avoid self-ligation on rRNA fragments (III, Supplementary Fig. S1). At first, self-ligation was avoided by blocking the free OH-group of rRNAs with dideoxyadenosine triphosphate (ddATP) using the TdT enzyme. However, because the study’s main idea was to develop a method with the minimum workload, and because the TdT reaction was an extra step in the procedure, the self-ligation prevention of rRNAs was also tested with an excess of M13-RNA adapters in the ligation reaction. This method proved successful when the M13-RNA concentration was increased enough (III, Supplementary Fig. S1).

When cDNA synthesis is done with random primers, proper randomisation and continuous priming are important. In this study, the fragment length distribution of random primed cDNAs demonstrated a biased primer annealing by strong peaks formation (III, Supplementary Fig. S3). Also, when only particular size fraction, e.g., 200–300 nt fraction, was selected for the final data, RAs of species were biased (III, Supplementary Fig. S2). The primary or secondary structures of the rRNA in cDNA synthesis could result in biased annealing of random primers. To fix this problem, first, the number of degenerate (N) bases in the primer was increased or decreased, and, secondly, an organic solvent, 10% dimethyl sulfoxide (DMSO), was added to the cDNA synthesis reaction to decrease the secondary structure formation of rRNAs. Both ways failed to resolve the problem. Strong peak formation appeared also when the functional gene was transcribed with random primers, demonstrating that secondary structures of rRNAs are not necessarily the main reason for noncontinuous, random priming. The first solution for the random priming problem was found in bioinformatics. The bias was minimised, and realistic RAs were restored, when size distribution was kept wide-ranging in the size trimming of DNA fragments during data trimming. Secondly, when hand mixed, degenerate bases were used in cDNA synthesis, the size distribution of the cDNA fragments appeared to be more even. Nonetheless, more studies are needed to resolve the random priming problem.

So-called primer bias has hindered amplicon sequencing studies continuously, and primer tests have been repeatedly published (Klindworth *et al.* 2013, Bradley *et al.* 2016, Wear *et al.* 2018). For phytoplankton, comprising prokaryotic and eukaryotic cells, the issue has remained unresolved for decades. New, primer independent library construction methods are needed for non-biased, affordable, adjustable and simultaneous sequencing of all phytoplankton. According to our mock community studies, this method provides one opportunity to meet the foregoing requirements, as the prokaryotic cells of the algae cultures were revealed simultaneously with eukaryotic cells, and gene-specific primers were not used.

4.4 Microscopic and genetic phytoplankton identification (IV)

Applying the new primer independent method (III), HTS results of phytoplankton community compositions from the water samples of 83 Finnish lakes were compared with results of traditional light microscopy. In the HTS identification studies, the sequencing library consisted of 16S and 18S rRNA fragments of microbes. Zooplankton species were excluded from the study because the volume of filtered lake water was not adequate for comparison. In the HTS results, the RAs of phytoplankton class level results were consistent with microscopic, morphological identification results, but, because the SILVA reference database is incomplete, the list of taxa had low correspondence at the genus level. Using HTS, higher diversity at the genus level was obtained, but microscopy was able to differentiate more species (IV, Fig. 7). The HTS identification was superior compared to microscopy-based studies in discovering pico- and nano-sized taxa and phytoplankton with cryptic morphological characteristics. Detailed, comparative results are described in Paper IV, as this discussion focuses on evaluating the functionality of the primer independent method (III) in the environmental studies.

One drawback here was that the quality of extracted RNA was decreased. Unstable storage temperatures during the sampling and problems with the freezer during the filter storage might have impaired the integrity of the rRNAs. The cultured cells, used in earlier studies (III), were not exposed to temperature changes during storage, and their RNA extracts could be used as a baseline in comparison (Fig. 5). Otherwise, HTS library construction for the lake water and cultured samples was done in the same manner. In typical RNA extracts of cultured cells, eukaryotic rRNAs (18S/28S rRNA peaks) dominated the size-distribution profile of rRNAs, and RNA demonstrated good quality (Fig. 5). The integrity of RNA could be estimated from the length of 23S or 28S rRNA in relation to the length of 16S or 18S rRNA, respectively, as longer rRNA usually degrades earlier (Schroeder *et al.* 2006). In a typical lake water sample, prokaryotic cells (16S/23S peaks) were the majority, and the quality of rRNAs was decreased, showing notably smaller 23S peaks than 16S rRNA peaks in the Tape Station size distribution analysis (Fig. 5). Altogether, when RNA-based studies are conducted, good planning for sampling and special care in laboratory work is needed in every sample handing step.

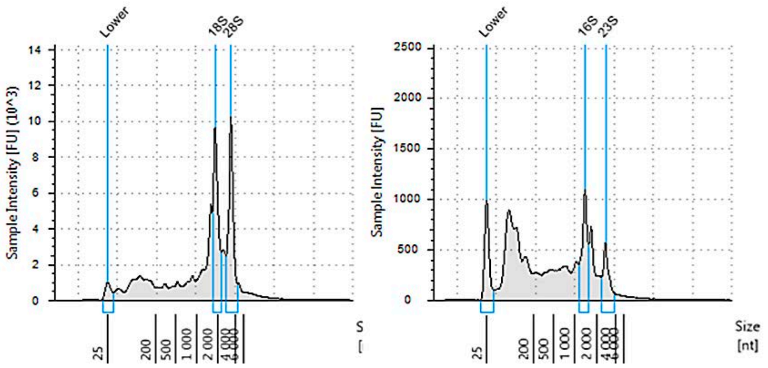


FIGURE 5 A typical example of fragments size-distribution of total RNA extracts from a filtered sample of laboratory cell cultures and from a filtered lake water sample. On the left, eukaryotic cells (18S/28S peaks) are in the majority in cell culture RNA extracts, whereas prokaryotic cells (16S/23S peaks) dominate RNA extracts from the lake water sample (on the right). The integrity of RNAs from cultured cells was good, but RNA from lake water samples typically showed some level of degradation. 25 nt peak is a lower marker.

Another trouble in the HTS results was related to the considerably low frequency of eukaryotic phytoplankton compared to cyanobacterial species. HTS results did not agree with microscopic results, where eukaryotic phytoplankton was represented much more frequently. The missing 18S rRNA gene sequences, and much higher numbers of 16S rRNA sequences, in the SILVA SSU rRNA database might be the main reason for low RAs of eukaryotic rRNAs in the HTS analysis, though one explanation can be found in ligation reactions.

In rRNA and M13-RNA ligation reactions, 16S and 18S rRNAs must have a 5′-end phosphate group, whereas M13-RNA oligos must have a free 3′-end OH-group, and the 5′-end phosphate group should be erased from the oligos. The ends of the oligos were suitable for ligation reactions, but the 5′-end phosphorylation state of the rRNAs was uncertain. Although ribosome maturation is a widely studied area (Lafontaine and Tollervey 2001, Schäfer *et al.* 2006, Granneman *et al.* 2010, Lafontaine 2015, An *et al.* 2018), only a few observations came up from search results concerning 5′-end phosphorylation states of mature rRNAs. In their preprint article, Fleischmann *et al.* (2019) describe that normally processed rRNA molecules have a single phosphate on their 5′-end. The 5′-end structure of Pol I transcribed precursor rRNAs have been studied using the DDRLACE method (the differential display of RNA ligase-mediated amplification of cDNA ends). Results revealed that about 20% of Pol I produced, putative rRNA precursors contained a 5′ tri- or diphosphate group, and about 80% of rRNA precursors were dephosphorylated (Bruderer *et al.* 2003). According to the studies, rRNAs undergo serial dephosphorylation and phosphorylation events during their maturation, but the question "what is the phosphorylation state of mature rRNAs′ 5′-ends?" is still waiting for an exact answer. The phosphorylation of rRNA fragments with T4 polynucleotide kinase before ligation reactions could be one solution to increasing ligation reaction efficiency and the number of 18S rRNA fragments in final HTS results. More studies are needed to resolve the 5′-end phosphorylating state of mature 16S and 18S rRNAs and the question concerning whether mature rRNAs have unknown mechanisms to protect their 5′-end integrity.

A loss of newly synthesised rRNA could be one reason, though to a lesser extent, for the relatively low frequency of eukaryotic phytoplankton in HTS results. Against earlier knowledge, recent findings in *Candida albicans* studies have shown that, during nutritional depletion, rRNAs are 5′-end capped, which protects rRNAs from 5′-end, phosphate-dependent, exonuclease digestion (Fleischmann *et al.* 2019). In our study, the time lag between sampling and filtering could have caused altered water conditions and stress responses in eukaryotic phytoplankton. This might have triggered a cellular signalling pathway, in which Pol I, which, in normal conditions, is responsible for rRNA synthesis, was subsided, and RNA polymerase II (Pol II) was activated, producing 5′-end caps for newly synthesised rRNAs. If this 5′-end capping by Pol II occurred in rRNA synthesis, the newly synthesised rRNAs were not ligated with M13-RNA, and they were excluded from the HTS results. However, to our knowledge, most of the rRNAs were in mature ribosomes, in which the rRNAs are presumably uncapped, and, in normal conditions, contained a 5´-end phosphate group (Fleischmann *et al.* 2019).

In this data, the alignment against reference sequences in OTU clustering was complicated, because reads did not have the same 5′-end starting position. Because rRNAs were more or less degraded, M13 adapter ligation to the 5′-end of the rRNAs did not yield sequences with full 5′-ends, and positional homology of the fragment was lost. In subsequent, reference-based OTU clustering, this method could yield long, centroid sequences, assigned as OTUs, which could be up to the full length of 18S or 16S rRNA sequences (Fig. 6). When the reference sequence was covered by reads from end to end, the full reference sequence was the centroid and was assigned as one OTU. Even if the reads covered only both ends of the same reference sequence, with the coverage containing a gap, the full reference was the centroid and was assigned as one OTU. If only a part of the reference had reads mapped to it, then the assigned centroid sequence was only a part of the reference sequence. Consequently, this caused trouble, as several OTUs could represent the same identity. The OTUs representing the same taxon were grouped before determining the RAs of taxa.

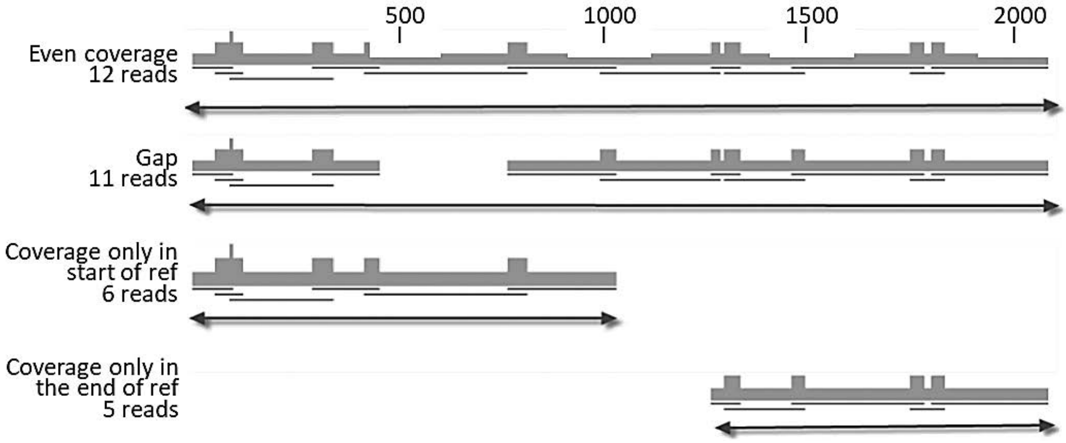


FIGURE 6 A simplified example of OTU picking, where reads were clustered against reference sequences (double-headed arrows). If the reference sequence was covered by reads against the full reference sequence, that was assigned to the centroid and OTU (upper arrow). When the reads covered both ends of the reference sequence, only one OTU was assigned even if there was a gap between the coverage of reads (second arrow). If reads covered only the 5′-end or the 3′-end of the reference sequence, two different OTUs were assigned.

The last aim of this thesis was to determine how equally both morphological (microscopic) and molecular (HTS) identification methods reveal phytoplankton community compositions in Finnish lakes. This new, RNA-based, primer independent method requires precise and aseptic working; but the biases due to CGN variations were avoided, and prokaryotic and eukaryotic phytoplankton were analysed simultaneously without so-called primer bias. As the proportion of prokaryotic phytoplankton order level results and eukaryotic class level results significantly corresponded, this method showed potential for simultaneous HTS of all domains of life without gene-specific primers. After reference databases are completed, molecular methods will have better correspondence with microscopic analysis, and, at some point, can probably be used independently without morphological analysis. Currently, morphological analysis and molecular HTS methods complement each other, and both methods have their own advantages and disadvantages.

5 CONCLUSIONS

Phytoplankton has a fundamental role in Earth's oxygen production and in the fixation of carbon dioxide gas through photosynthesis. From a different perspective, phytoplankton reflects the ecological status of waters. New, efficient genetic methods are needed to monitor phytoplankton because, for small phytoplankton, the resolution of the traditional microscopic method is not adequate and because the microscopic method is time-consuming, requires high-level specialists and cannot be standardised enough. This study focuses on the genetic methods used in phytoplankton monitoring but the results are applicable to other microbes, as well. In this thesis, two HTS library construction methods, DNA-based and RNA-based, were developed, one comparative study of existing molecular methods was provided and, finally, 83 Finnish lake water samples were analysed using microscopic and HTS methods, whose results were compared.

In the new, DNA-based HTS library construction method, time and cost were saved. Because indexing and pooling of samples were done first, fragmentation, sequencing adapter ligation, size-selection and purification could be done for all samples simultaneously, in a one-tube principle. As the 5'-end of all the gene fragments was selected and preserved, the use of OTU algorithms was optimised. Indexing barcodes with M13 overhand can be used for countless of genes. Because commercial kits and reagents are expensive and sometimes unpractical, affordable, adaptable and time-saving alternatives for commercial kits are necessary. The need for updated and usable HTS library preparation methods is evident in changing laboratory practices.

The validation of HTS methods and the production of common guidelines (with warning about general causes of distortions) are important in environmental studies. All steps in HTS, sample treatment, library construction, platform choice and data analysis methods, have their own challenges for achieving realistic sequencing results. When methods used in phytoplankton HTS library construction steps were validated, RNA-based sequencing was found to be a better choice than DNA-based sequencing, when phytoplankton community structure was determined using 18S rRNA gene fragments. In

DNA-based sequencing, huge, interspecific GCN variation biased the results, whereas, in RNA-based HTS, RAs of species showed realistic results when compared to RAs of biomasses and carbon contents. The results were consistent, as the number of ribosomes reflect the size of cells and, accordingly, reflect the biomass and carbon content of cells.

The development of a new, RNA-based HTS library construction method was a logical step after gaining the results of method validation studies because, in RNA-based HTS, the problems caused by interspecific GCN variations were avoided. This new, directional 5′-end sequencing method was primer independent, and all domains of life could be detected from the same HTS sample. HTS results showed potential for identifying prokaryotes and eukaryotes from the same sample, because RAs of eukaryotic phytoplankton were realistic, according to the RAs of biomasses and carbon contents of phytoplankton species, and prokaryotic bacterial cells were determined simultaneously.

The new, RNA-based HTS library construction method was applied for a comparative analysis of HTS results and microscopic-based results gained from environmental samples. Eukaryotic phytoplankton class level RA results, as well as cyanobacterial order level RAs, agreed when the results of different methods were compared. This gives encouraging views for the possibilities of replacing microscopic methods with molecular tools in future monitoring studies. Also, since small picoplankton and filamentous cyanobacteria could be differentiated with HTS, results were promising. However, the RAs of prokaryotic and eukaryotic phytoplankton did not agree with each other. Dinoflagellates and diatoms were well represented in the SILVA database, but others (e.g., chlorophytes) were poorly represented. Due to the lack of sequences in the reference database, more species were identified with microscopy analysis. When the primer independent method was used to analyse SSU rRNA diversity without primer bias, HTS results of RNA-based sequencing revealed adventurous and unexpected richness of microbes in the aquatic environment.

One of the most urgent needs for phytoplankton diversity studies is the completion of reference databases, especially relating to eukaryotic microbes. If the whole length of 16S and 18S rRNAs were sequenced, it would provide high-resolution for identification. No selection of variable regions would be needed and species with highly similar sequences could be distinguishable from each other at the species level. However, especially for phytoplankton, reference databases have limitations, such as incomplete coverage of rare species sequences at the species level and controversial taxonomic naming.

Acknowledgements

This study was carried out at the Department of Biological and Environmental Science, University of Jyväskylä. The study was financed by grants from the Academy of Finland and the European Research Council (to Prof. Marja Tiirola).

I am grateful for my supervisor, Marja, who gave me the opportunity to be part of her research group and carry out these exciting studies. I want to thank you for the way you guided me during these four years. You gave me a chance to work independently, but you were available when I needed guidance and support. I enjoyed the intelligent discussions with you and admire your creative scientific ideas and personality.

And Kristiina, my supervisor, too, what would be the right words about you to express my gratitude? You guided me to the world of phytoplankton and opened a new scientific perspective for me. You were always patient when I, again and again, asked the same questions about the taxonomy of phytoplankton. I was privileged to have a supervisor with such excellent knowledge of the field. You enlightened me in the mysteries of the water world and trusted my opinions about cell and molecular ideas, although my thoughts were sometimes too incoherent.

I have been privileged to have so many wonderful colleagues around me during the last four years. My office roommates, Lara and Pauliina: you really are something! We did have a humour of our own, and I miss that. We have had so many joyful moments, but we also shared troubles and disappointments with each other. This means that the office was like another home to me. Thank you for your beautiful words when I was down and troubled. I also want to thank all former and present members of our group. Thank you for working together for a common goal. We shared memorable moments in the lab. While pipetting we had deep conversations about life—of course, without causing pipetting errors! I think I may have talked too much, so thank you all for your patience and kindness.

Finally, my warmest thanks go to my beloved husband, Lauri, for his forbearance, support and endless love. Without your lovable and admirable sentiments and the atmosphere you create around us, I would not have been able to manage.

YHTEENVETO (RÉSUMÉ IN FINNISH)

Geneettisen kasviplanktonseurannan kehittäminen

Yksisoluiset mikro-organismit ovat määrältään ja monimuotoisuudeltaan runsain elämänmuoto maan päällä, mutta vain pieni osa mikrobeista on tunnistettu. Muiden eliöiden olemassaolo on mikrobeista riippuvaista, joten näiden organismien tunnistaminen on erittäin tärkeää. Päähuomio tutkimuksessa oli mikroskooppisilla levillä, kasviplanktonilla, joka on vesistöjen perustuottaja ja joka suurelta osin vastaa ilmakehän hapen tuotannosta. Mikroskooppisen pienillä, kelluvilla levillä on elintärkeä rooli maapallon ilmastolle, koska kasviplankton absorboi hiilidioksidia valtamerten ja muiden vesistöjen pinnalla. Perinteisesti kasviplanktonanalyysi on perustunut lajien morfologiseen tunnistamiseen mikroskopoimalla, mutta tämä on hidas ja pienten solujen tunnistamisessa jopa mahdoton menetelmä. Uudet korkean tason sekvensointitekniikat (HTS, high-throughput sequencing) tarjoavat tehokkaan keinon mikrobien monimuotoisuuden kartoittamiseen, sillä jopa usean sadan ympäristönäytteen mikrobit voidaan analysoida yhdellä sekvensointikerralla, josta saadaan miljoonia DNA-sekvenssejä. Tutkimuksessa kehitettiin sekä DNA-fragmenttien että RNA-fragmenttien prosessointimenetelmä HTS-tekniikalla tehtävään mikrobien monimuotoisuuden tunnistamiseen. Lisäksi tutkimuksessa vertailtiin nykyisiä molekulaarisia menetelmiä ja tulosten perusteella määriteltiin metodit, joilla kasviplanktonin yhteisörakenne saadaan parhaiten selville HTS-perusteisessa analyysissä. Sekvensointidatan käsittely, bioinformatiikka, sisältyi tutkimukseen. Lopuksi, käyttäen uutta RNA-perusteista metodia, 83 suomalaisen järven vesinäytteistä tunnistettiin yhtäaikaisesti kaikki mikrobit ja tuloksia verrattiin mikroskopoimalla saatuihin määrityksiin.

Uusi DNA-perusteinen, HTS-tekniikkaan kehitetty DNA-fragmenttien valmisteluprosessi on aikaa ja kustannuksia säästävä menetelmä. Menetelmässä organismeista eristetyn DNA:n merkkigeenin fragmentit indeksoitiin tunnistamiseen tarvittavalla barkoodisekvenssillä näytekohtaisesti, jonka jälkeen jopa 100 ympäristönäytettä voitiin yhdistää. DNA-fragmenttien katkaisu oikean kokoisiksi, sekvensointiadapterin sidonta, fragmenttien kokovalinta ja puhdistus voitiin tehdä yhtäaikaisesti jopa 100:lle näytteelle. Uusi menetelmä sopii käytettäväksi mille geenille tahansa ja DNA-fragmenttien indeksointiin käytettäviä barkoodeja on mahdollista hyödyntää lukemattomille geeneille. Menetelmää käyttäen myös bioinformatiikan algoritmit toimivat tehokkaasti, koska kaikki merkkigeenin DNA-fragmentit lähtevät samasta kohdasta geeniä.

Tutkimuksessa tehtiin myös kattava vertailututkimus nykyisistä molekulaarisista ja data-analyysi metodeista. Vertailussa käytettiin malliorganismeina kuutta kasviplanktonlajia, joiden omaisuudet, kuten solun koko, tuman koko, soluseinän vahvuus ja rakenne, poikkesivat toisistaan ja joiden lajikohtaiset kuiva- ja märkämassat sekä hiilipitoisuudet toimivat indikaattoreina realistisesta yhteisörakenteesta. HTS-analyysissä merkkigeeninä käytettiin eukaryoottisten solujen ribosomaalisen RNA:n (rRNA) geenifragmenttia, joka monistettiin geenispesifisiä alukkeita käyttäen. Kaikilla prosesseilla, joita käytettiin sekvensointinäytteiden valmistamisessa, oli vaikutusta HTS-perusteisiin yhteisön rakennemäärityksiin. Eniten DNA-perusteisia tuloksia vääristivät tutkitun rRNA-geenin lajikohtaiset geenikopiomäärien vaihtelut, jotka solutasolla olivat jopa kymmentuhatkertaisia. DNA-perusteisten sekvensointitulosten pohjalta tehdyt arviot yhteisön rakenteesta eivät vastanneet ennalta tiedetyn kasviplanktonyhteisön rakennetta, kun taas RNA-pohjaiset HTS-tulokset antoivat realistisen analyysin yhteisörakenteesta. Tulosten perusteella eukaryoottisten solujen monimuotoisuuden tutkimiseen suositeltiin RNA-pohjaista HTS-analyysiä.

Koska RNA-pohjainen mikrobiyhteisön kartoittaminen toimi hyvin ja koska kaikkia lajeja tasapuolisesti monistavia geenispesifisiä PCR-alukkeita ei ole saatavilla, tutkimuksessa kehitettiin RNA-perusteinen geenispesifisistä alukkeista riippumaton HTS-näytteen valmistusmenetelmä. Uuden menetelmän etuna on yhtäaikainen, kaikkien mikrobien tutkiminen samasta näytteestä. Menetelmän kehittämisessä ja arvioimisessa käytettiin samaa kuuden kasviplanktonlajin poolia kuin edellä menetelmien vertailututkimuksessa, koska kasviplanktonin lajikohtaiset kuiva- ja märkämassat sekä hiilipitoisuudet olivat etukäteen määriteltyinä. HTS-tulokset osoittivat uuden menetelmän toimivuuden, koska kasviplanktonin yhteisökuvaus vastasi realistista yhteisörakennetta ja samasta näytteestä pystyttiin määrittämään sekä prokaryoottiset että eukaryoottiset solut. Menetelmä tarjoaa näin ollen hyvän mahdollisuuden myös lajien välisille vuorovaikutustutkimuksille, kuten symbioosi- ja loistutkimuksille.

Lopuksi 83 suomalaisen järvivesinäytteen kasviplanktonlajit määritettiin morfologisella, mikroskopointiin perustuvalla menetelmällä ja kaikki mikrobit määritettiin uudella, RNA-perusteisella, geenispesifisistä PCR-alukkeista riippumattomalla menetelmällä. Tarkoituksena oli selvittää, voidaanko aikaa ja erikoisosaamista vaativat mikroskooppiset kasviplanktonmääritykset korvata HTS-tekniikoilla. Kun molempien menetelmien kasviplanktonia koskevia tuloksia vertailtiin keskenään, lahkotasolla syanobakteerien tulokset ja luokkatasolla eukaryoottisten kasviplanktonin tulokset osoittivat merkittävää vastaavuutta.

Vaikka molekylaarinen tunnistus tarjoaa jo lupaavia tuloksia kasviplanktonseurantaan, se ei vielä yksin riitä tunnistusmenetelmäksi. Tällä hetkellä kasviplanktonin lajitason referenssikirjastot, joiden avulla HTS-tuloksista luokitellaan taksonomisesti, ovat liian keskeneräisiä varsinkin eukaryoottisten mikrobien osalta. Toistaiseksi morfologiaan perustuvat ja molekulaariset menetelmät täydentävät toisiaan. Uuden, RNA-perusteisen HTS-näytteen valmistusmenetelmän etuna oli se, että sekä prokaryoottiset että eukaryoottiset solut voitiin analysoida samanaikaisesti huolimatta lajikohtaisista eroista sekvensseissä ja lajikohtaiset geenikopioiden määrät eivät vääristäneet tuloksia. Huomattavana hyötynä oli myös se, että pienet ja morfologisesti vaikeasti tunnistettavat kasviplanktonlajit voitiin tunnistaa. HTS-tulokset osoittivat uskomattoman suuren lajien monimuotoisuuden järviemme vesinäytteistä, vaikkakaan eläinplanktonin ja ei-planktonisten lajien määritykset eivät sisältyneet tähän tutkimukseen.

REFERENCES

Adl S.M., Bass D., Lane C.E., Lukeš J., Schoch C.L., Smirnov A., Agatha S., Berney C., Brown M.W., Burki F., Cárdenas P., Čepička I., Chistyakova L., Campo J., Dunthorn M., Edvardsen B., Eglit, Y., Guillou L., Hampl V., Heiss A.A., Hoppenrath M., James T.Y., Karnkowska A., Karpov S., Kim E., Kolisko M., Kudryavtsev A., Lahr D.J., Lara E., Le Gall L., Lynn D.H., Mann D.G., Massana R., Mitchell E.A., Morrow C., Park J.S., Pawlowski J.W., Powell M.J., Richter D.J., Rueckert S., Shadwick L., Shimano S., Spiegel F.W., Torruella G., Youssef N., Zlatogursky V. & Zhang Q. 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J. Eukaryot. Microbiol.* 66: 4-119.

Alberts B. 2002. *Molecular biology of the cell.* Garland Science, New York NY.

Alesheikh S., Shahtahmassebi N., Roknabadi M.R. & Pilevar Shahri R. 2018. Silicene nanoribbon as a new DNA sequencing device. *Physics Letters A* 382: 595–600.

Ali N., Rampazzo R.C.P., Costa A.D.T. & Krieger M.A. 2017. Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics. *Biomed. Res. Int.* 2017: 9306564.

Ambardar S., Gupta R., Trakroo D., Lal R. & Vakhlu J. 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J. Microbiol.* 56: 394–404.

An W., Du Y. & Ye K. 2018. Structural and functional analysis of Utp24, an endonuclease for processing 18S ribosomal RNA. *PLoS One* 13: e0195723.

Angly F.E., Dennis P.G., Skarshewski A., Vanwonterghem I., Hugenholtz P. & Tyson G.W. 2014. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* 2: 11-2618-2-11. eCollection 2014.

Basu S. & Mackey R.K. 2018. Phytoplankton as Key Mediators of the Biological Carbon Pump: Their Responses to a Changing Climate. *Sustainability* 10.

Bauman R.W., Machunis-Masuoka E. & Tizard I.R. 2004. *Microbiology.* Pearson/Benjamin Cummings.

Beardall J., Allen D., Bragg J., Finkel Z.V., Flynn K.J., Quigg A., Rees T.A.V., Richardson A. & Raven J.A. 2009. Allometry and stoichiometry of unicellular, colonial and multicellular phytoplankton. *New Phytol.* 181: 295–309.

Bellinger E.G., author. 2015. *Freshwater algae: identification, enumeration and use as bioindicators.* 2nd edition. Chichester, West Sussex; Hoboken, NJ: John Wiley & Sons Inc., 2015.

Berdalet E., Fleming L.E., Gowen R., Davidson K., Hess P., Backer L.C., Moore S.K., Hoagland P. & Enevoldsen H. 2015. Marine harmful algal blooms, human health and wellbeing: challenges and opportunities in the 21st century. *J. Mar. Biol. Assoc. UK* 2015: 10.1017/S0025315415001733. Epub 2015 Nov 20.

Bieri P., Greber B.J. & Ban N. 2018. High-resolution structures of mitochondrial ribosomes and their functional implications. *Curr. Opin. Struct. Biol.* 49: 44–53.

Bieri P., Leibundgut M., Saurer M., Boehringer D. & Ban N. 2017. The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. *EMBO J.* 36: 475–486.

Blaxter M.L. 2004. The promise of a DNA taxonomy. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 359: 669–679.

Bonen L. & Doolittle W.F. 1975. On the prokaryotic nature of red algal chloroplasts. *Proc. Natl. Acad. Sci. U. S. A.* 72: 2310–2314.

Boughner L.A. & Singh P. 2016. Microbial Ecology: Where are we now? *Postdoc J.* 4: 3–17.

Bowers R.M., Clum A., Tice H., Lim J., Singh K., Ciobanu D., Ngan C.Y., Cheng J.F., Tringe S.G. & Woyke T. 2015. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 16: 856-015-2063-6.

Bracciali A., Caravagna G., Gilbert D. & Tagliaferri R. 2017. *Computational Intelligence Methods for Bioinformatics and Biostatistics: 13th International Meeting, CIBB 2016, Stirling, UK, September 1–3, 2016, Revised Selected Papers.* Springer International Publishing.

Bradley I.M., Pinto A.J. & Guest J.S. 2016. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. *Appl. Environ. Microbiol.* 82: 5878–5891.

Brandariz-Fontes C., Camacho-Sanchez M., Vila C., Vega-Pla J.L., Rico C. & Leonard J.A. 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Sci. Rep.* 5: 8056.

Brockman W., Alvarez P., Young S., Garber M., Giannoukos G., Lee W.L., Russ C., Lander E.S., Nusbaum C. & Jaffe D.B. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* 18: 763–770.

Brooks H.J. 2013. Modern microbiology – a quiet revolution with many benefits. *Australas Med. J.* 6: 378–381.

Brown A. & Shao S. 2018. Ribosomes and cryo-EM: a duet. *Curr. Opin. Struct. Biol.* 52: 1–7.

Bruderer T., Tu L. & Lee M.G. 2003. The 5' end structure of transcripts derived from the rRNA gene and the RNA polymerase I transcribed protein coding genes in Trypanosoma brucei. *Mol. Biochem. Parasitol.* 129: 69–77.

Calbet A. 2008. The trophic roles of microzooplankton in marine systems. *ICES J. Mar. Sci.* 65: 325–331.

Callahan B.J., McMurdie P.J. & Holmes S.P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11: 2639–2643.

Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E.K., Fierer N., Pena A.G., Goodrich J.K., Gordon J.I., Huttley G.A., Kelley S.T., Knights D., Koenig J.E., Ley R.E., Lozupone C.A., McDonald D., Muegge B.D., Pirrung M., Reeder J., Sevinsky J.R., Turnbaugh P.J., Walters

W.A., Widmann J., Yatsunenko T., Zaneveld J. & Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7: 335–336.

Caron D.A., Worden A.Z., Countway P.D., Demir E. & Heidelberg K.B. 2008. Protists are microbes too: a perspective. *ISME J.* 3: 4.

Casiraghi M., Labra M., Ferri E., Galimberti A. & De Mattia F. 2010. DNA barcoding: a six-question tour to improve users' awareness about the method. *Brief Bioinform.* 11: 440–453.

Catling D.C. & Zahnle K. 2002. Evolution of Atmospheric Oxygen. In: Holton J.R., Pyle J. & Curry J.A. (eds.), *Encyclopedia of Atmospheric Sciences*, Academic Press, pp. 754–761.

Chandramouli P., Topf M., Menetret J.F., Eswar N., Cannone J.J., Gutell R.R., Sali A. & Akey C.W. 2008. Structure of the mammalian 80S ribosome at 8.7 A resolution. *Structure* 16: 535–548.

Chen W., Zhang C.K., Cheng Y., Zhang S. & Zhao H. 2013. A comparison of methods for clustering 16S rRNA sequences into OTUs. *PloS one* 8: e70837; e70837–e70837.

Ciganda M. & Williams N. 2011. Eukaryotic 5S rRNA biogenesis. *Wiley Interdiscip. Rev. RNA* 2: 523–533.

Cock P.J., Fields C.J., Goto N., Heuer M.L. & Rice P.M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38: 1767–1771.

Condon C., Liveris D., Squires C., Schwartz I. & Squires C.L. 1995. rRNA operon multiplicity in Escherichia coli and the physiological implications of rrn inactivation. *J. Bacteriol.* 177: 4152–4156.

Cooper G.M. & Hausman R.E. 2007. *The Cell: A Molecular Approach.* ASM Press.

Cruickshank R.H. 2002. Molecular markers for the phylogenetics of mites and ticks. *Systematic and Applied Acarology Vol 7: 31 Jul.2002*.

Deamer D., Akeson M. & Branton D. 2016. Three decades of nanopore sequencing. *Nat. Biotechnol.* 34: 518–524.

Del Vecchio F., Mastroiaco V., Di Marco A., Compagnoni C., Capece D., Zazzeroni F., Capalbo C., Alesse E. & Tessitore A. 2017. Next-generation sequencing: recent applications to the analysis of colorectal cancer. *J. Transl. Med.* 15: 246-017-1353-y.

Denny S.K. & Greenleaf W.J. 2018. Linking RNA Sequence, Structure, and Function on Massively Parallel High-Throughput Sequencers. *Cold Spring Harb. Perspect. Biol.*

Desmond E., Brochier-Armanet C., Forterre P. & Gribaldo S. 2011. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. *Res. Microbiol.* 162: 53–70.

Duan Y., Xie N., Song Z., Ward C.S., Yung C., Hunt D.E., Johnson Z.I. & Wang G. 2018. A High-Resolution Time Series Reveals Distinct Seasonal Patterns of Planktonic Fungi at a Temperate Coastal Ocean Site (Beaufort, North Carolina, USA). *Appl. Environ. Microbiol.* 84: e00967–18.

Edgar R.C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

El-Metwally S., Ouda O.M. & Helmy M. 2014. Next-Generation Sequencing Platforms. In: El-Metwally S., Ouda O.M. & Helmy M. (eds.), *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*, Springer New York, New York, NY, pp. 37–44.

Espejo R.T. & Plaza N. 2018. Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA. *Front. Microbiol.* 9: 1232.

Ewing B. & Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194.

Ewing B., Hillier L., Wendl M.C. & Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8: 175–185.

Falkowski P. 2012. Ocean Science: The power of plankton. *Nature* 483: S17–20.

Fenchel T. 2013. *Ecology of Protozoa: The Biology of Free-living Phagotropic Protists.* Springer Berlin Heidelberg.

Fleischmann J., Rocha M.A. & Hauser P.V. 2019. RNA Polymerase II is involved in 18S and 25S ribosomal RNA transcription, in Candida albicans. *bioRxiv*: 510156.

Floyd R., Abebe E., Papert A. & Blaxter M. 2002. Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11: 839–850.

Fox G.E., Magrum L.J., Balch W.E., Wolfe R.S. & Woese C.R. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc. Natl. Acad. Sci. U. S. A.* 74: 4537–4541.

Freed E.F., Bleichert F., Dutca L.M. & Baserga S.J. 2010. When ribosomes go bad: diseases of ribosome biogenesis. *Mol. Biosyst.* 6: 481–493.

Gibbons J.G., Branco A.T., Godinho S.A., Yu S. & Lemos B. 2015. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc. Natl. Acad. Sci. U. S. A.* 112: 2485–2490.

Godhe A., Asplund M.E., Harnstrom K., Saravanan V., Tyagi A. & Karunasagar I. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* 74: 7174–7182.

Golob J.L., Margolis E., Hoffman N.G. & Fredricks D.N. 2017. Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics* 18: 283-017-1690-0.

Gong J., Dong J., Liu X. & Massana R. 2013. Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* 164: 369–379.

Goodwin S., McPherson J.D. & McCombie W.R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17: 333–351.

Goto Y., Yanagi I., Matsui K., Yokoi T. & Takeda K. 2016. Integrated solid-state nanopore platform for nanopore fabrication via dielectric breakdown, DNA-speed deceleration and noise reduction. *Sci. Rep.* 6: 31324.

Granneman S., Petfalski E., Swiatkowska A. & Tollervey D. 2010. Cracking pre-40S ribosomal subunit structure by systematic analyses of RNA-protein cross-linking. *EMBO J.* 29: 2026.

Gray M.W., Burger G. & Lang B.F. 1999. Mitochondrial Evolution. *Science* 283: 1476.

Hadziavdic K., Lekang K., Lanzen A., Jonassen I., Thompson E.M. & Troedsson C. 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One* 9: e87624.

Hajibabaei M., Singer G.A.C., Hebert P.D.N. & Hickey D.A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* 23: 167–172.

Hang R., Wang Z., Deng X., Liu C., Yan B., Yang C., Song X., Mo B. & Cao X. 2018. Ribosomal RNA Biogenesis and Its Response to Chilling Stress in Oryza sativa. *Plant Physiol.* 177: 381–397.

Head S.R., Komori H.K., LaMere S.A., Whisenant T., Van Nieuwerburgh F., Salomon D.R. & Ordoukhanian P. 2014. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* 56: 61–4, 66, 68, passim.

Heather J.M. & Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* 107: 1–8.

Hebert P.D., Cywinska A., Ball S.L. & deWaard J.R. 2003. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270: 313–321.

Hein R., Stoddard S.F., Schmidt T.M., Roller B.R.K. & Smith B.J. 2014. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Asids Res.* 43: D593–D598.

Henras A.K., Plisson-Chastang C., O'Donohue M.F., Chakraborty A. & Gleizes P.E. 2015. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip. Rev. RNA* 6: 225–242.

Hizawa T., Sawada K., Takao H. & Ishida M. 2006. Fabrication of a two-dimensional pH image sensor using a charge transfer technique. *Sens. Actuators, B: Chem.* 117: 509–515.

Hong S., Bunge J., Leslin C., Jeon S. & Epstein S.S. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3: 1365–1373.

Jain M., Koren S., Miga K.H., Quick J., Rand A.C., Sasani T.A., Tyson J.R., Beggs A.D., Dilthey A.T., Fiddes I.T., Malla S., Marriott H., Nieto T., O'Grady J., Olsen H.E., Pedersen B.S., Rhie A., Richardson H., Quinlan A.R., Snutch T.P., Tee L., Paten B., Phillippy A.M., Simpson J.T., Loman N.J. & Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36: 338–345.

Jobe A., Liu Z., Gutierrez-Vargas C. & Frank J. 2018. New Insights into Ribosome Structure and Function. *Cold Spring Harbor Perspect. Biol.*

Karlson B., Cusack C. & Bresnan E. 2010. *Microscopic and molecular methods for quantitative phytoplankton analysis.* UNESCO, 110pp. (Intergovernmental Oceanographic Commission Manuals and Guides;55), Paris, France.

Karst S.M., Dueholm M.S., McIlroy S.J., Kirkegaard R.H., Nielsen P.H. & Albertsen M. 2018. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* 36: 190–195.

Kembel S.W., Wu M., Eisen J.A. & Green J.L. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* 8: e1002743.

Klappenbach J.A., Saxman P.R., Cole J.R. & Schmidt T.M. 2001. rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res.* 29: 181–184.

Klindworth A., Pruesse E., Schweer T., Peplies J., Quast C., Horn M. & Glockner F.O. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41: e1.

Koeppel A.F. & Wu M. 2013. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res.* 41: 5175–5188.

Kuai L., Fang F., Butler J.S. & Sherman F. 2004. Polyadenylation of rRNA in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* 101: 8581–8586.

Lafontaine D.L. 2015. Noncoding RNAs in eukaryotic ribosome biogenesis and function. *Nat. Struct. Mol. Biol.* 22: 11–19.

Lafontaine D.L.J. & Tollervey D. 2001. The function and synthesis of ribosomes. *Nature Reviews Molecular Cell Biology* 2: 514.

Le Bescot N., Mahe F., Audic S., Dimier C., Garet M.J., Poulain J., Wincker P., de Vargas C. & Siano R. 2016. Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environ. Microbiol.* 18: 609–626.

Lepere C., Masquelier S., Mangot J.F., Debroas D. & Domaizon I. 2010. Vertical structure of small eukaryotes in three lakes that differ by their trophic status: a quantitative approach. *ISME J.* 4: 1509–1519.

Lepoitevin M., Ma T., Bechelany M., Janot J.M. & Balme S. 2017. Functionalization of single solid state nanopores to mimic biological ion channels: A review. *Adv. Colloid Interface Sci.* 250: 195–213.

Lindgreen S., Adair K.L. & Gardner P.P. 2016. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* 6: 19233.

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L. & Law M. 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* 2012: 11.

Locey K.J. & Lennon J.T. 2016. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* 113: 5970–5975.

Lodish H., Berk A. & Zipursky S.e.a. 2000. *Molecular Cell Biology. 4th edition.* New York: W. H. Freeman.

Loman N.J., Misra R.V., Dallman T.J., Constantinidou C., Gharbia S.E., Wain J. & Pallen M.J. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30: 434–439.

Lopez-Garcia A., Pineda-Quiroga C., Atxaerandio R., Perez A., Hernandez I., Garcia-Rodriguez A. & Gonzalez-Recio O. 2018. Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences. *Front. Microbiol.* 9: 3010.

Louca S., Doebeli M. & Parfrey L.W. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* 6: 41-018-0420-9.

Low V.L., Tan T.K., Lim P.E., Domingues L.N., Tay S.T., Lim Y.A., Goh T.G., Panchadcharam C., Bathmanaban P. & Sofian-Azirun M. 2014. Use of COI, CytB and ND5 genes for intra- and inter-specific differentiation of Haematobia irritans and Haematobia exigua. *Vet. Parasitol.* 204: 439–442.

Lyska D., Meierhoff K. & Westhoff P. 2013. How to build functional thylakoid membranes: from plastid transcription to protein complex assembly. *Planta* 237: 413–428.

Macmanes M.D. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Front. Genet.* 5: 13.

Madigan M.T. & Martinko J.M. 2006. *Brock Biology of Microorganisms.* Pearson Prentice Hall.

Mardis E.R. 2017. DNA sequencing technologies: 2006–2016. *Nature Protocols* 12: 213.

Marguerat S. & Bahler J. 2012. Coordinating genome expression with cell size. *Trends Genet.* 28: 560–565.

McFadden G.I. 2014. Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harb Perspect. Biol.* 6: a016105.

McFadden G.I. 2001. Chloroplast origin and integration. *Plant Physiol.* 125: 50–53.

Merriman B., Ion Torrent R&D Team & Rothberg J.M. 2012. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* 33: 3397–3417.

Minei R., Hoshina R. & Ogura A. 2018. De novo assembly of middle-sized genome using MinION and Illumina sequencers. *BMC Genomics* 19: 700-018-5067-1.

Mir K., Neuhaus K., Bossert M. & Schober S. 2013. Short barcodes for next generation sequencing. *PLoS One* 8: e82933.

Mojarro A., Hachey J., Ruvkun G., Zuber M.T. & Carr C.E. 2018. CarrierSeq: a sequence analysis workflow for low-input nanopore sequencing. *BMC Bioinformatics* 19: 108-018-2124-3.

Mu W., Lu H.M., Chen J., Li S. & Elliott A.M. 2016. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *J. Mol. Diagn.* 18: 923–932.

Needham D.M., Fichot E.B., Wang E., Berdjeb L., Cram J.A., Fichot C.G. & Fuhrman J.A. 2018. Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *ISME J.* 12: 2417–2432.

Neefs J.M., Van d.P., Hendriks L. & De Wachter R. 1990. Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.* 18 Suppl: 2237–2317.

Ng S.H., Braxton C., Eloit M., Feng S.F., Fragnoud R., Mallet L., Mee E.T., Sathiamoorthy S., Vandeputte O. & Khan A.S. 2018. Current Perspectives on High-Throughput Sequencing (HTS) for Adventitious Virus Detection: Upstream Sample Processing and Library Preparation. *Viruses* 10: 10.3390/v10100566.

Nilakanta H., Drews K.L., Firrell S., Foulkes M.A. & Jablonski K.A. 2014. A review of software for analyzing molecular sequences. *BMC Res. Notes* 7: 830-0500-7-830.

Nyren P. & Lundin A. 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal. Biochem.* 151: 504–509.

Nyren P., Pettersson B. & Uhlen M. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* 208: 171–175.

Oliveira M.C., Repetti S.I., Iha C., Jackson C.J., DÃaz-Tapia P., Lubiana K.M.F., Cassano V., Costa J.F., Cremen M.C.M., Marcelino V.R. & Verbruggen H. 2018. High-throughput sequencing for algal systematics. *Eur. J. Phycol.* 53: 256–272.

Parada A.E., Needham D.M. & Fuhrman J.A. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* 18: 1403–1414.

Parker J., Helmstetter A.J., Devey D., Wilkinson T. & Papadopulos A.S.T. 2017. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci. Rep.* 7: 8345-017-08461-5.

Parks M.M., Kurylo C.M., Dass R.A., Bojmar L., Lyden D., Vincent C.T. & Blanchard S.C. 2018. Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci. Adv.* 4: eaao0665.

Peacock M.B., Gibble C.M., Senn D.B., Cloern J.E. & Kudela R.M. 2018. Blurred lines: Multiple freshwater and marine algal toxins at the land-sea interface of San Francisco Bay, California. *Harmful Algae* 73: 138–147.

Pearson W.R. & Lipman D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85: 2444.

Pena C., Hurt E. & Panse V.G. 2017. Eukaryotic ribosome assembly, transport and quality control. *Nat. Struct. Mol. Biol.* 24: 689–699.

Pereira T.J. & Baldwin J.G. 2016. Contrasting evolutionary patterns of 28S and ITS rRNA genes reveal high intragenomic variation in Cephalenchus (Nematoda): Implications for species delimitation. *Mol. Phylogenet. Evol.* 98: 244–260.

Petrov A.S., Bernier C.R., Gulen B., Waterbury C.C., Hershkovits E., Hsiao C., Harvey S.C., Hud N.V., Fox G.E., Wartell R.M. & Williams L.D. 2014. Secondary structures of rRNAs from all three domains of life. *PloS one* 9: e88222; e88222–e88222.

Petrova O.E., Garcia-Alcalde F., Zampaloni C. & Sauer K. 2017. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Sci. Rep.* 7: 41114.

Phillips R., Kondev J., Theriot J. & Garcia H. 2012. *Physical Biology of the Cell.* CRC Press.

Pichard S.L., Campbell L. & Paul J.H. 1997. Diversity of the ribulose bisphosphate carboxylase/oxygenase form I gene (rbcL) in natural phytoplankton communities. *Appl. Environ. Microbiol.* 63: 3600.

Poptsova M.S., Il'icheva I.A., Nechipurenko D.Y., Panchenko L.A., Khodikov M.V., Oparina N.Y., Polozov R.V., Nechipurenko Y.D. & Grokhovsky S.L. 2014. Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* 4: 4532.

Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P. & Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341-2164-13-341.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J. & Glockner F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41: D590–6.

Quince C., Walker A.W., Simpson J.T., Loman N.J. & Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35: 833–844.

Rang F.J., Kloosterman W.P. & de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19: 90-018-1462-9.

Raveh A., Margaliot M., Sontag E.D. & Tuller T. 2016. A model for competition for ribosomes in the cell. *Journal of the Royal Society, Interface* 13: 20151062.

Reuter J., Spacek D.V. & Snyder M. 2015. High-Throughput Sequencing Technologies. *Mol. Cell* 58: 586–597.

Reynolds C.S. 2006. *The Ecology of Phytoplankton.* Cambridge University Press, Cambridge.

Rifai N., Horvath A.R., Wittwer C.T. & Park J. 2018. *Principles and Applications of Molecular Diagnostics.* Elsevier Science.

Rothberg J.M., Hinz W., Rearick T.M., Schultz J., Mileski W., Davey M., Leamon J.H., Johnson K., Milgrew M.J., Edwards M., Hoon J., Simons J.F., Marran D., Myers J.W., Davidson J.F., Branting A., Nobile J.R., Puc B.P., Light D., Clark T.A., Huber M., Branciforte J.T., Stoner I.B., Cawley S.E., Lyons M., Fu Y., Homer N., Sedova M., Miao X., Reed B., Sabina J., Feierstein E., Schorn M., Alanjary M., Dimalanta E., Dressman D., Kasinskas R., Sokolsky T., Fidanza J.A., Namsaraev E., McKernan K.J., Williams A., Roth G.T. & Bustillo J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475: 348–352.

Roy S., Coldren C., Karunamurthy A., Kip N.S., Klee E.W., Lincoln S.E., Leon A., Pullambhatla M., Temple-Smolkin R.L., Voelkerding K.V., Wang C. & Carter A.B. 2018. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* 20: 4–27.

Rusk N. 2010. Torrents of sequence. *Nature Methods* 8: 44.

Sakurai T. & Husimi Y. 1992. Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. *Anal. Chem.* 64: 1996–1997.

Sanger F., Nicklen S. & Coulson A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74: 5463–5467.

Santoferrara L.F., Grattepanche J.D., Katz L.A. & McManus G.B. 2016. Patterns and processes in microbial biogeography: do molecules and morphologies give the same answers? *ISME J.* 10: 1779–1790.

Schäfer T., Maco B., Petfalski E., Tollervey D., Böttcher B., Aebi U. & Hurt E. 2006. Hrr25-dependent phosphorylation state regulates organization of the pre-40S subunit. *Nature* 441: 651–655.

Schloss P.D. & Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71: 1501–1506.

Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J. & Weber C.F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75: 7537–7541.

Schloss P.D. 2012. Secondary structure improves OTU assignments of 16S rRNA gene sequences. *ISME J.* 7: 457.

Schroeder A., Mueller O., Stocker S., Salowsky R., Leiber M., Gassmann M., Lightfoot S., Menzel W., Granzow M. & Ragg T. 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7: 3-2199-7-3.

Schulz K.G., Bach L.T., Bellerby R.G.J., Bermúdez R., Büdenbender J., Boxhammer T., Czerny J., Engel A., Ludwig A., Meyerhöfer M., Larsen A., Paul A.J., Sswat M. & Riebesell U. 2017. Phytoplankton Blooms at Increasing Levels of Atmospheric Carbon Dioxide: Experimental Evidence for Negative Effects on Prymnesiophytes and Positive on Small Picoeukaryotes. *Frontiers in Marine Science* 4: 64.

Sekerci Y. & Petrovskii S. 2018. Global Warming Can Lead to Depletion of Oxygen by Disrupting Phytoplankton Photosynthesis: A Mathematical Modelling Approach. *Geosciences* 8.

Sellner K.G., Doucette G.J. & Kirkpatrick G.J. 2003. Harmful algal blooms: causes, impacts and detection. *J. Ind. Microbiol. Biotechnol.* 30: 383–406.

Shumway S.E., Burkholder J.A.M. & Morton S.L. 2018. *Harmful Algal Blooms: A Compendium Desk Reference.* Wiley.

Sieburth J.M., Smetacek V. & Lenz J. 1978. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. *Limnol. Oceanogr.* 23: 1256–1263.

Siegwald L., Touzet H., Lemoine Y., Hot D., Audebert C. & Caboche S. 2017. Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. *PLoS One* 12: e0169563.

Singer E., Andreopoulos B., Bowers R.M., Lee J., Deshpande S., Chiniquy J., Ciobanu D., Klenk H.P., Zane M., Daum C., Clum A., Cheng J.F., Copeland

A. & Woyke T. 2016. Next generation sequencing data of a defined microbial mock community. *Sci. Data* 3: 160081.

Slomovic S., Laufer D., Geiger D. & Schuster G. 2006. Polyadenylation of ribosomal RNA in human cells. *Nucleic Acids Res.* 34: 2966–2975.

Smith A.M., Heisler L.E., St Onge R.P., Farias-Hesson E., Wallace I.M., Bodeau J., Harris A.N., Perry K.M., Giaever G., Pourmand N. & Nislow C. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* 38: e142.

Song Z., Schlatter D., Kennedy P., Kinkel L.L., Kistler H.C., Nguyen N. & Bates S.T. 2015. Effort versus Reward: Preparing Samples for Fungal Community Characterization in High-Throughput Sequencing Surveys of Soils. *PLoS One* 10: e0127234.

Suzuki M.T. & Giovannoni S.J. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* 62: 625–630.

Takahashi S., Tomita J., Nishioka K., Hisada T. & Nishijima M. 2014. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS One* 9: e105592.

Tanabe A.S., Nagai S., Hida K., Yasuike M., Fujiwara A., Nakamura Y., Takano Y. & Katakura S. 2016. Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel sequencing-based monitoring of the planktonic eukaryote community. *Mol. Ecol. Resour* 16: 402–414.

Tragin M., Zingone A. & Vaulot D. 2018. Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. *Environ. Microbiol.* 20: 506–520.

Traversi F., Raillon C., Benameur S.M., Liu K., Khlybov S., Tosun M., Krasnozhon D., Kis A. & Radenovic A. 2013. Detecting the translocation of DNA through a nanopore using graphene nanoribbons. *Nature Nanotechnology* 8: 939.

Tveit A.T., Urich T. & Svenning M.M. 2014. Metatranscriptomic analysis of arctic peat soil microbiota. *Appl. Environ. Microbiol.* 80: 5761–5772.

Van de Peer Y. & De Wachter R. 1997. Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *J. Mol. Evol.* 45: 619–630.

Vargas C.A., Escribano R. & Poulet S. 2006. Phytoplankton food quality determines time windows for successful zooplankton reproductive pulses. *Ecology* 87: 2992–2999.

Vences M., Thomas M., van der Meijden A., Chiari Y. & Vieites D.R. 2005. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Front. Zool.* 2: 5-9994-2-5.

Vinje H., Almoy T., Liland K.H. & Snipen L. 2014. A systematic search for discriminating sites in the 16S ribosomal RNA gene. *Microb. Inform. Exp.* 4: 2-5783-4-2.

Vuorio K., Lepistö L. & Holopainen A. 2007. Intercalibrations of freshwater phytoplankton analyses. *Boreal Environment Research* 12: 561–569.

Wallner G., Amann R. & Beisker W. 1993. Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. *Cytometry* 14: 136–143.

Wang C., Zhang T., Wang Y., Katz L.A., Gao F. & Song W. 2017. Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. *Proc. Biol. Sci.* 284: 10.1098/rspb.2017.0425.

Wear E.K., Wilbanks E.G., Nelson C.E. & Carlson C.A. 2018. Primer selection impacts specific population abundances but not community dynamics in a monthly time-series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton. *Environ. Microbiol.* 20: 2709–2726.

Weiss M.M., Van der Zwaag B., Jongbloed J.D.H., Vogel M.J., Brüggenwirth H.T., Lekanne Deprez R.H., Mook O., Ruivenkamp C.A.L., van Slegtenhorst M.A., van d.W., Waisfisz Q., Nelen M.R. & van d.S. 2013. Best Practice Guidelines for the Use of Next-Generation Sequencing Applications in Genome Diagnostics: A National Collaborative Study of Dutch Genome Diagnostic Laboratories. *Human Mutation* 34: 1313–1321.

Westcott S.L. & Schloss P.D. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3: e1487.

Williams C.R., Baccarella A., Parrish J.Z. & Kim C.C. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* 17: 103-016-0956-2.

Woese C.R. 1987. Bacterial evolution. *Microbiol Rev.* 51: 221.

Woese C.R. & Fox G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74: 5088–5090.

Wong K.H., Jin Y. & Moqtaderi Z. 2013. Multiplex Illumina sequencing using DNA barcoding. *Curr. Protoc. Mol. Biol.* Chapter 7: Unit 7.11.

Wu S., Xiong J. & Yu Y. 2015. Taxonomic resolutions based on 18S rRNA genes: a case study of subclass copepoda. *PLoS One* 10: e0131498.

Wurzbacher C., Larsson E., Bengtsson-Palme J., Van den Wyngaert S., Svantesson S., Kristiansson E., Kagami M. & Nilsson R.H. 2019. Introducing ribosomal tandem repeat barcoding for fungi. *Mol. Ecol. Resour.* 19: 118–127.

Xie Q., Lin J., Qin Y., Zhou J. & Bu W. 2011. Structural diversity of eukaryotic 18S rRNA and its impact on alignment and phylogenetic reconstruction. *Protein Cell* 2: 161–170.

Xue Y., Ankala A., Wilcox W.R. & Hegde M.R. 2014. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genetics in Medicine* 17: 444.

Yang W. 2011. Nucleases: diversity of structure, function and mechanism. *Q. Rev. Biophys.* 44: 1–93.

Zhou X., Liao W.J., Liao J.M., Liao P. & Lu H. 2015. Ribosomal proteins: functions beyond the ribosome. *J. Mol. Cell Biol.* 7: 92–104.

Zhu F., Massana R., Not F., Marie D. & Vaulot D. 2005. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* 52: 79–92.

ORIGINAL PAPERS

I

A PRACTICAL METHOD FOR BARCODING AND SIZE-TRIMMING PCR TEMPLATES FOR AMPLICON SEQUENCING

by

Mäki, A., Rissanen, A.J. & Tiirola, M. 2016

BioTechniques 60: 88–90

https://doi.org/10.2144/000114380

A practical method for barcoding and size-trimming PCR templates for amplicon sequencing

Anita Mäki, Antti J. Rissanen, and Marja Tiirola
Department of Biological and Environmental Science, University of Jyväskylä, Finland

BioTechniques 60:88-90 (February 2016) doi 10.2144/000114380
Keywords: next-generation sequencing; PCR; ligation; primer; polyclonality

Supplementary material for this article is available at www.BioTechniques.com/article/114380.

Sample barcoding facilitates the analysis of tens or even hundreds of samples in a single next-generation sequencing (NGS) run, but more efficient methods are needed for high-throughput barcoding and size-trimming of long PCR products. Here we present a two-step PCR approach for barcoding followed by pool shearing, adapter ligation, and 5´ end selection for trimming sets of DNA templates of any size. Our new trimming method offers clear benefits for phylogenetic studies, since targeting exactly the same region maximizes the alignment and enables the use of operational taxonomic unit (OTU)-based algorithms.

The efficiency of next-generation sequencing (NGS) of PCR amplicons has increased via sample barcoding (1), which facilitates multiplex sequencing of numerous samples and genes (such as ribosomal RNA or protein-coding genes) in the same run. Barcodes can be added to the PCR amplicons either by ligation or by performing the PCR amplification with fusion primers, which include both the barcodes and sequencing adapters. When analyzing the sequence diversity of 10 genes in 100 samples, barcoding with fusion primers would require 1000 barcoded primers (10 genes × 100 samples), making the task both laborious and expensive without dual indexing, which is only available on the Illumina MiSeq platform. Another challenge when amplifying target genes with well-established universal primer pairs is related to fragment size optimization. Two commonly used NGS platforms, Illumina and Ion Torrent, recommend maximum fragment lengths of about 300 and 400 nucleotides, respectively (2). Many previously established PCR primer pairs produce much longer amplicons, which must be cut for optimal NGS sequencing. Here we show how library preparation can be simplified with a two-step PCR protocol with M13-tagged primers and how the sample pool can be cut to a certain length all at once instead of performing shearing, adapter ligation, and size selection for each sample separately. Our protocol was validated by sequencing archaeal *16S rRNA* genes from environmental samples using the Ion Torrent (Life Technologies Corporation, Carlsbad, CA) chemistry (with sequencing adapters IonA and P1), but the same template preparation principles are also valid for the other NGS chemistries.

Our cost-efficient, labor-reducing method begins with amplification of each gene library by two-step PCR using barcoded primers, followed by pooling the libraries together (Figure 1). Shearing, ligation with a sequencing 3´ end adapter (P1 on the Ion Torrent platform), and size selection of the amplicons takes place in a single tube, and, very importantly, the process produces sequencing templates with full 5´ ends. Amplicons must be phosphorylated and blunt-ended to effectively ligate adapter P1. In this reaction, ends that are not sheared enzymatically (e.g., with Life Technologies' Ion Shear Plus reagent kit) are not phosphorylated, which prevents ligation of P1 to the 5´ end of the IonA adapter. Two overhanging deoxythymidine nucleotides in the P1 adapter (Supplementary Table S1) prevent the adapter from ligating in a false orientation, and phosphorothioate backbone modification protects the two overhanging nucleotides from exonuclease activity. Fragments that are also sheared from the 5´ side (the IonA side) having P1 on both ends are not efficiently amplified in the subsequent PCR and are not selected during the bead enrichment step, which selects IonA-positive beads. Thus, this method facilitates complete selection of sequences with full 5´ ends.

An M13 linker has been used in nested PCR to reduce the need to invest in fluorescent primers for microsatellite genotyping (3) and for the sequencing of amplicons from different exons of the human epidermal growth factor receptor (EGFR) gene using 454 chemistry (Roche) (4). Recently, barcoding with a similar two-step PCR approach with a 16-bp head-sequence has been designed for the Illumina platform (5) using templates of different sizes. Here, pooling barcoded libraries before shearing and final adapter ligation allows size optimization all

METHOD SUMMARY

Here we present a new protocol combining PCR and adapter ligation for next-generation sequencing (NGS) template preparation that greatly improves sample multiplexing. During library construction, the 5´ sequencing adapter is incorporated in a two-step PCR with universal barcoded M13-tailed primers, and the 3´ adapter is ligated to the pooled and sheared PCR fragments in a single tube, steps that assist in the focused sequencing of the 5´ ends of long PCR fragments. When using long fusion primers for template preparation, selection of the area where polyclonality is detected is required.

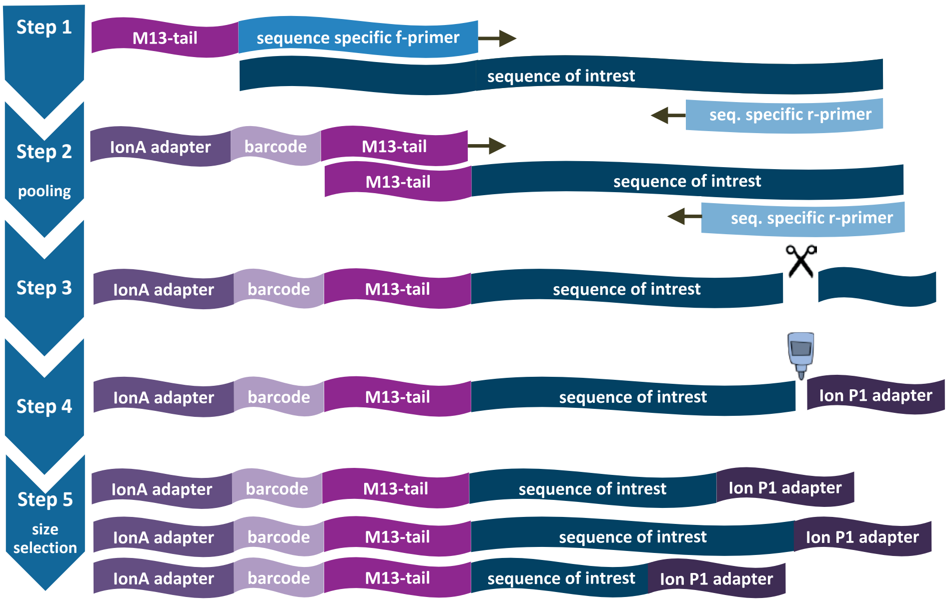


Figure 1. Combining PCR and ligation techniques for barcoding and trimming of long PCR fragments for next-generation sequencing (NGS) library preparation. In Step 1, an M13-tail is incorporated into the PCR products. In Step 2, the 5′ sequencing primer (IonA) and barcodes are incorporated by exploiting the M13 tail. Barcoded samples are then pooled together, and shearing, ligation of the P1 adapter (3′ sequencing primer), and size selection (Steps 3–5) are performed in a single tube.

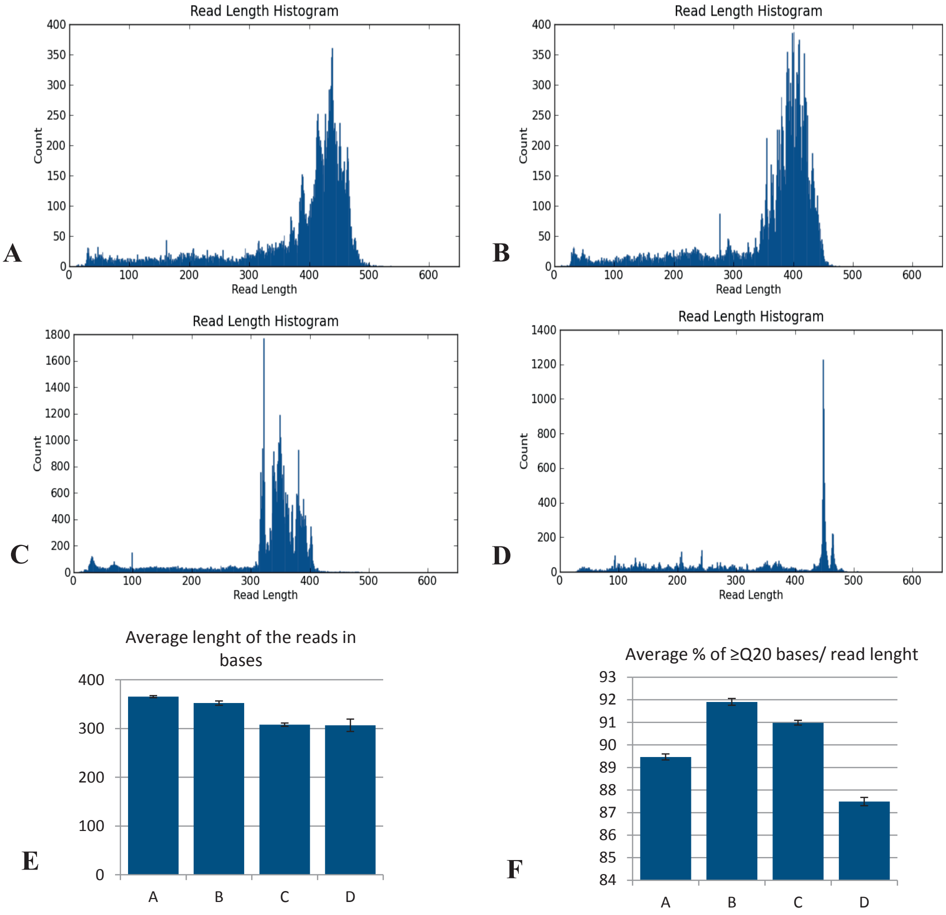


Figure 2. Comparison of read-length histograms and quality bar graphs for the same sequencing run on the Ion Torrent platform containing amplicons from various template preparation methods. The libraries for sequencing were prepared by (A) our method using DNA fragments of the archaeal *16S rRNA* gene; (B) our method using DNA fragments of the gene encoding the α subunit of the archaeal methyl-coenzyme M reductase (*mcrA*); (C) the fusion method (no fragmentation, M13-usage, or ligation) using DNA fragments of the bacterial *16S rRNA gene; or (D) our method with M13-tail usage at the 5′ end but without shearing of the DNA fragments (oversized fragments >500 bp with 3′ P1 adapter incorporated by PCR) of the archaeal mcrA* gene. Accordingly, the bar graphs show the average lengths (±SEM, $n = 10$) of the reads (in bases) (E) and the average read-specific percentage (±SEM, $n = 10$) of the bases with quality scores ≥Q20 (F) in these treatments. The average read lengths and the quality of reads were significantly higher for our method for the *mcrA* gene (B) than in the method where DNA fragments were left oversized (D) (read length: $t = 9.18$, $P < 0.001$; quality: $t = 18.8$, $P < 0.001$). All libraries were added in equimolar concentrations to the emulsion PCR before the main sequencing.

at once, as the Ion Torrent platform is sensitive for long template sizes.

To demonstrate our method using Ion Torrent chemistry, we sequenced fragments of the archaeal *16S rRNA* gene from environmental samples (see protocol in the Supplementary Material). Lyophilized slurries of lake sediment samples were extracted with a Power Soil DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, CA). Partial archaeal *16S rRNA* genes were amplified with forward primer M13–340F and reverse primer 1000R (Supplementary Table S2). Barcodes were added to each amplified sample with another six cycles of PCR where M13-tailed forward primer IonA_bc_M13 was annealed to the M13 sequence of the first PCR products. Amplicon size and yield were checked via agarose gel electrophoresis, and the purification of the PCR products was performed with the Agencourt AMPure XP purification system (Beckman Coulter, Brea, CA). DNA yield was determined with a Qubit 2.0 Fluorometer and a dsDNA HS Assay Kit (Thermo Fisher, Cambridge, UK), and the samples were pooled together. The pooled sample was further purified with AMPureXP, fragmented all at once using an Ion Shear Plus reagent kit (Life Technologies), and, with the same all at once principle, the P1 adapter (Supplementary Table S1) was ligated into fragmented DNA products using the Ion Plus Fragment Library kit (Life Technologies). DNA fragments were size-selected with the Pippin Prep system (Sage Science, Beverly, MA). Amplification of the size-selected fragments was performed using the Platinum PCR SuperMix High Fidelity kit (Life Technologies). Quantitation and size control were performed with the Ion Library TaqMan Quantitation kit (Life Technologies) and with the Agilent Bioanalyzer 2100 (Agilent Technologies, Stockport, UK) using the Agilent High-Sensitivity dsDNA kit. Emulsion PCR with the Ion OneTouch system and Ion OT2 400 kit (Life Technologies) (quality control included), templated bead enrichment, and sequencing with the Ion Personal Genome Machine (PGM) with an Ion PGM Sequencing 400 Kit and Ion 314 chip (Life Technologies) were performed in accordance with the manufacturer's instructions.

A comparative sequencing test was performed on the Ion Torrent platform using equimolar concentrations of libraries representing 4 template preparations (Figure 2): (*A*) our proposed method using DNA fragments of the archaeal *16S rRNA* gene; (*B*) our method using DNA fragments of the gene that encodes the α subunit of the archaeal methyl-coenzyme M reductase (*mcrA*); (*C*) the fusion method (no fragmentation, M13-usage, or ligation) using DNA fragments of the bacterial *16S rRNA* gene; and (*D*) our method with M13-tail usage at the 5´ end but without shearing of the DNA fragments (oversized fragments >500 bp with 3´ P1 adapter incorporated by PCR) of the archaeal *mcrA* gene. Comparatively good average read lengths were achieved using our library preparation method (Figure 2E). The average percentage of the bases per reads whose quality scores were ≥Q20 dropped to the lowest number using oversized *mcrA* libraries (Figure 2F).

We also compared our method with the standard Ion Torrent adapter ligation protocol (Supplementary Figure S1) to study *18S rRNA* genes from phytoplankton samples. The data were analyzed using Mothur (6). The standard method did not yield intact 5´ ends, as forward adapter ligation needs sheared ends. This severely reduced the length of the overlapping area of DNA fragments in subsequent sequence alignments. In contrast, the proposed method, which retains the 5´ ends, maximized the alignment length (data not shown). Thus, the information content utilized in the operational taxonomic unit (OTU) (e.g., at the standard $\mathrm{OTU}_{0.97}$ level) clustering and taxonomic classification of OTUs was higher with our method. Our approach, therefore, leads to more accurate identification and more efficient taxonomic classification of OTUs of marker genes (e.g., rRNA or functional genes) than the standard method.

After sequencing the amplicons with long fusion primers, filtering of the polyclonal sequences required adjustment changes in the check region. Using Ion Torrent software (Torrent Suite 4.2.1), polyclonality is, by default, checked during flows 12–70. With extended primer lengths, filtering has to be based on the later region: flows 120–160 (see protocol in the Supplementary Material).

Although NGS is increasingly becoming automated, preparation of multi-sample templates still requires many steps and much manual work. Here, we have shown that exploiting a universal head-sequence (such as M13) and rearranging the order of the steps in the template preparation offers a practical alternative to standard barcoding methods.

Author contributions

A.M. performed the molecular biology study and wrote the paper. M.T. and A.J.R edited the manuscript and offered valuable advice and intellectual contributions. A.M., A.J.R., and M.T. all contributed to the study design.

Acknowledgments

We thank Elina Virtanen for her dedicated technical assistance with PGM, and Anu Mikkonen and Veli-Mikko Puupponen for helping with the bioinformatics. A.M. was supported by Academy of Finland Project 260797 and ERC Micro-RIP 615146 awarded to M.T., and A.J.R was supported by Academy of Finland Project 140964.

Competing interests

The authors declare no competing interests.

References

1. **Parameswaran, P., R. Jalili, L. Tao, S. Shokralla, B. Gharizadeh, M. Ronaghi, and A.Z. Fire.** 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res. *35*:e130.
2. **van Dijk, E.L., H. Auger, Y. Jaszczyszyn, and C. Thermes.** 2014. Ten years of next-generation sequencing technology. Trends Genet. *30*:418-426.
3. **Schuelke, M.** 2000. An economic method for the fluorescent labeling of PCR fragments. Nat. Biotechnol. *18*:233-234.
4. **Daigle, D., B.B. Simen, and P. Pochart.** 2011. High-throughput sequencing of PCR products tagged with universal primers using 454 life sciences systems. *Curr Protoc Mol Biol. Chapter 7*:Unit7.5.
5. **Herbold, C.W., C. Pelikan, O. Kuzyk, B. Hausmann, R. Angel, D. Berry, and A. Loy.** 2015. A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. Front Microbiol. *6*:731.
6. **Schloss, P.D., S.L. Westcott, T. Ryabin, J.R. Hall, M. Hartmann, E.B. Hollister, R.A. Lesniewski, B.B. Oakley, et al.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*:7537-7541.

Received 18 May 2015; accepted 02 November 2015.

Address correspondence to Anita Mäki, Department of Biological and Environmental Science, PO Box 35, FI-40014 University of Jyväskylä, Finland. E-mail: anita.maki@jyu.fi

Supplementary material for:

A practical method for barcoding and size-trimming PCR templates for amplicon sequencing

Anita Mäki[1], Antti J. Rissanen[1], and Marja Tiirola[1]

[1]The Department of Biological and Environmental Science, University of Jyväskylä, Finland

Supplementary Table S1: P1 Ion Torrent sequencing adapter used in this study

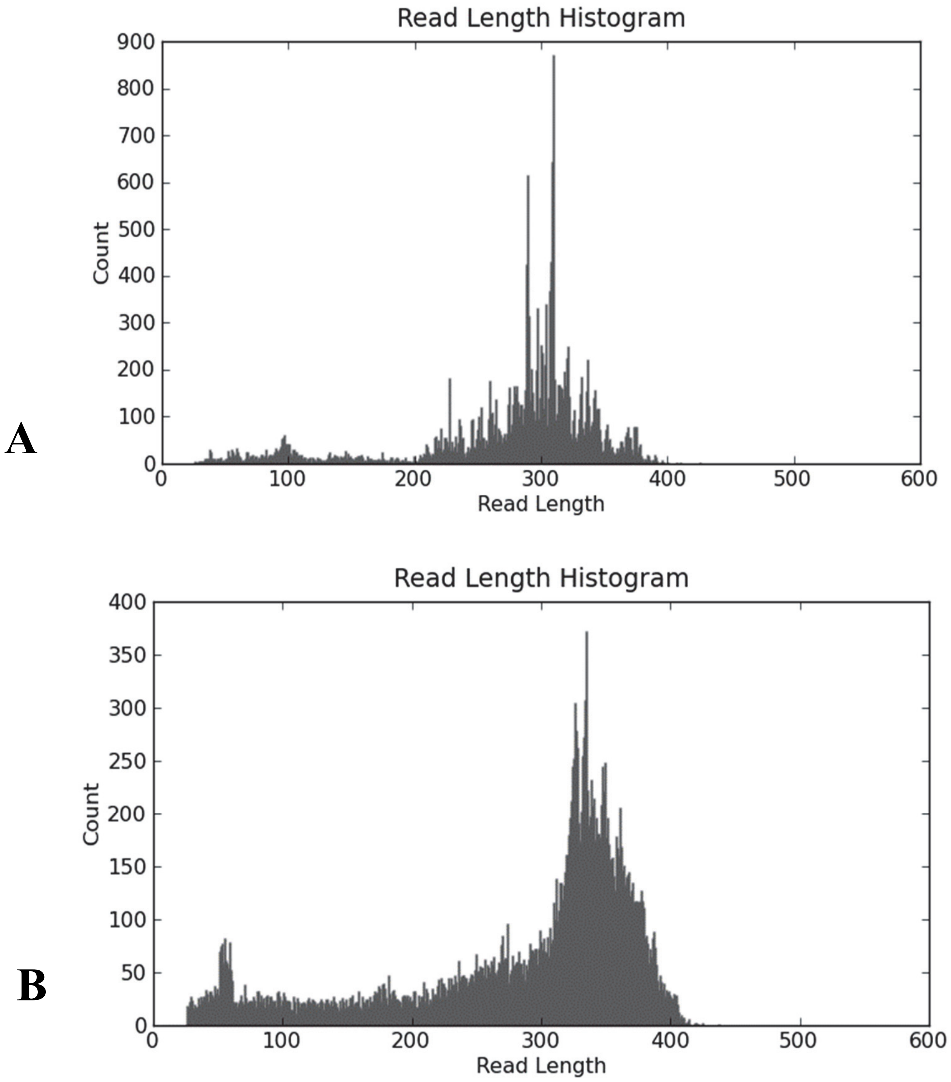
<table>
<tr><th>P1 Ion torrent sequencing adapter</th></tr>
<tr><td>5’- CCACTACGCCTCCG C T T TCC TCTCTA TGGGCAGTCGGTGAT - 3’
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
3’- T*T*GGTGATGCGGAGGCGAAAGGAGAGATACCCGTCAGCCACTA - 5’</td></tr>
<tr><td>* phosphorothioate bond</td></tr>
</table>

Supplementary Table S2: Oligos

Oligo’s Name	Sequence 5' to 3'
M13	TGTAAAACGACGGCCAGT
M13_Arch 340f	TGTAAAACGACGGCCAGTCCCTAYGGGGYGCASCAG
Arch 1000r	GGCCATGCACYWCYTCTC
IonA _bc1_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CTAAGGTAAC**TGTAAAACGACGGCCAGT**
IonA _bc2_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TAAGGAGAAC**TGTAAAACGACGGCCAGT**
IonA _bc3_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**AAGAGGATTC**TGTAAAACGACGGCCAGT**
IonA _bc4_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TACCAAGATC**TGTAAAACGACGGCCAGT**
IonA _bc5_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CAGAAGGAAC**TGTAAAACGACGGCCAGT**
IonA _bc6_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CTGCAAGTTC**TGTAAAACGACGGCCAGT**
IonA _bc7_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTCGTGATTC**TGTAAAACGACGGCCAGT**
IonA _bc8_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTCCGATAAC**TGTAAAACGACGGCCAGT**
IonA _bc9_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TGAGCGGAAC**TGTAAAACGACGGCCAGT**
IonA _bc10_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CTGACCGAAC**TGTAAAACGACGGCCAGT**
IonA _bc11_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCCTCGAATC**TGTAAAACGACGGCCAGT**
IonA _bc12_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TAGGTGGTTC**TGTAAAACGACGGCCAGT**
IonA _bc13_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCTAACGGAC**TGTAAAACGACGGCCAGT**
IonA _bc14_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTGGAGTGTC**TGTAAAACGACGGCCAGT**
IonA _bc15_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCTAGAGGTC**TGTAAAACGACGGCCAGT**
IonA _bc16_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCTGGATGAC**TGTAAAACGACGGCCAGT**
IonA _bc17_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCTATTCGTC**TGTAAAACGACGGCCAGT**
IonA _bc18_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**AGGCAATTGC**TGTAAAACGACGGCCAGT**
IonA _bc19_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTAGTCGGAC**TGTAAAACGACGGCCAGT**
IonA _bc20_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CAGATCCATC**TGTAAAACGACGGCCAGT**
IonA _bc21_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCGCAATTAC**TGTAAAACGACGGCCAGT**
IonA _bc22_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTCGAGACGC**TGTAAAACGACGGCCAGT**

IonA _bc23_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TGCCACGAAC**TGTAAAACGACGGCCAGT**
IonA _bc24_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**AACCTCATTC**TGTAAAACGACGGCCAGT**
IonA _bc25_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CCTGAGATAC**TGTAAAACGACGGCCAGT**
IonA _bc26_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTACAACCTC**TGTAAAACGACGGCCAGT**
IonA _bc27_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**AACCATCCGC**TGTAAAACGACGGCCAGT**
IonA _bc28_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**ATCCGGAATC**TGTAAAACGACGGCCAGT**
IonA _bc29_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCGACCACTC**TGTAAAACGACGGCCAGT**
IonA _bc30_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CGAGGTTATC**TGTAAAACGACGGCCAGT**
IonA _bc31_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCCAAGCTGC**TGTAAAACGACGGCCAGT**
IonA _bc32_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCTTACACAC**TGTAAAACGACGGCCAGT**
IonA _bc33_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTCTCATTGAAC**TGTAAAACGACGGCCAGT**
IonA _bc34_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCGCATCGTTC**TGTAAAACGACGGCCAGT**
IonA _bc35_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TAAGCCATTGTC**TGTAAAACGACGGCCAGT**
IonA _bc36_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**AAGGAATCGTC**TGTAAAACGACGGCCAGT**
IonA _bc37_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CTTGAGAATGTC**TGTAAAACGACGGCCAGT**
IonA _bc38_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TGGAGGACGGAC**TGTAAAACGACGGCCAGT**
IonA _bc39_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TAACAATCGGC**TGTAAAACGACGGCCAGT**
IonA _bc40_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CTGACATAATC**TGTAAAACGACGGCCAGT**
IonA _bc41_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTCCACTTCGC**TGTAAAACGACGGCCAGT**
IonA _bc42_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**AGCACGAATC**TGTAAAACGACGGCCAGT**
IonA _bc43_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**CTTGACACCGC**TGTAAAACGACGGCCAGT**
IonA _bc44_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTGGAGGCCAGC**TGTAAAACGACGGCCAGT**
IonA _bc45_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TGGAGCTTCCTC**TGTAAAACGACGGCCAGT**
IonA _bc46_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TCAGTCCGAAC**TGTAAAACGACGGCCAGT**
IonA _bc47_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TAAGGCAACCAC**TGTAAAACGACGGCCAGT**
IonA _bc48_ M13	CCATCTCATCCCTGCGTGTCTCCGAC**TCAG**TTCTAAGAGAC**TGTAAAACGACGGCCAGT**

The M13-tailed, barcoded IonA sequencing primers are composed of the following segments: The first 26 letters indicate IonA sequencing adapter bases, bold TCAG letters indicate library key bases, the next letters indicate barcode bases, and the last 18 bold letters indicate M13-tail bases. Note: this barcode set lacks GAT bases (barcode adapter sequence) at the end of the barcode sequence. This barcoded, M13-tailed IonA sequencing adapter usage can be considered a Fusion Method in which the GAT sequence can be omitted. All barcodes have a C as the last base, and there must be either an A, T or G after a barcode's last C (to avoid two-mer incorporation bias).



Supplementary Figure 1: Comparison of the sequencing reads lengths when using proposed trimming method and standard Ion Torrent ligation method for Ion Torrent sequencing. Proposed method: The 5’ sequencing adapter was incorporated in the PCR reaction and 3’adapter was ligated after fragmentation step (A). Standard method: Both sequencing adapters, 5’end and 3’end, were ligated to library amplicons after fragmentation (B). Sequencing was targeted to the ribosomal 18S rRNA gene.

Protocol for:

A practical method for barcoding and size-trimming PCR templates for amplicon sequencing

Anita Mäki[1], Antti J. Rissanen[1], and Marja Tiirola[1]

[1]The Department of Biological and Environmental Science, University of Jyväskylä, Finland

Library preparation and Ion Torrent sequencing of fragments of archaeal 16S rRNA gene using M13-Arch340forward/1000reverse -primer pair

Reagents:

- Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific)
- 10 mM dNTPs mix
- Agencourt AMPure XP beads (Beckman Coulter)
- Qubit Fluorometer with high-sensitive dsDNA kit (Life Technologies)
- Ion Shear Plus Reagents Kit (Life Technologies)
- Ion Plus Fragment Library kit (Life Technologies)
- 2% agarose cassettes, external markers, and reagents (100-600 bp) for the Pippin Prep instrument (Sage Science)
- Platinum PCR Super Mix High-Fidelity from Ion Plus Fragment Library kit (Life Technologies)
- P1 adapter from the Ion Xpress Barcode Adapters 1-16 Kit (Life Technologies)
- Ion PGM Template OT2 400 Kit (Life Technologies)
- Ion Sphere Quality Control Kit (Life Technologies)
- Ion PGM Sequencing 400 Kit (Life Technologies)

Oligos:

- 10 μM Forward Primer with M13 tail, now M13-Arch340
- 10 μM Reverse Primer, now Arch 1000
- 10 μM Forward Primer: IonA-barcode with M13 tail

I PCR

Components	40 μl reaction	Final concentration
5x Phusion HF Buffer	8 μl	1x
10 mM dNTPs mix	1 μl	250 μM
10 μM forward primer with M13-tail, M13-Arch340	1 μl	0.25 μM
10 μM reverse primer Arch 1000	1 μl	0.25 μM
Template DNA 2 ng/μl	3 μl	0.15 ng/μl
Phusion Hot Start II High-Fidelity DNA Polymerase (2 U/μL) (Thermo Scientific)	0.5 μl	0.025 U/μl
Nuclease-free water	25.5 μl	

Thermocycling conditions for PCR:

Cycle step	Temperature	Time	Cycles
Initial denaturation	98 °C	30 s	1
Denaturation	98 °C	10 s	
Annealing	55 °C	30 s	35
Extension	72 °C	60 s	
Final extension	72 °C	10 min	1

II PCR

Components	20 µl reaction	Final concentration
5x Phusion HF Buffer	4 µl	1x
10 mM dNTPs mix	0.5 µl	250 µM
10 µM forward primer: IonA-barcode with M13-tail	0.5 µl	0.25 µM
10 µM reverse primer Arch 1000	0.5 µl	0.25 µM
Template DNA from I PCR product	0.5 µl	0.15 ng/µl
Phusion Hot Start II High-Fidelity DNA Polymerase (2 U/µL) (Thermo Scientific)	0.25 µl	0.025 U/µl
Nuclease-free water	13.75 µl	

- Now, the forward primer is barcoded M13-tailed IonA adapter (10 µM), and the reverse primer is Arch1000 (10 µM).
- The master mix without the forward primer is aliquoted into separate PCR tubes at a volume of 19 µl.
- The forward primer and template from I PCR are added.
- Thermocycling conditions are same as in I PCR, but now, there are only six cycles.

Agarose Gel Electrophoresis

- Size- and quantity-checking are performed for both I PCR and II PCR-products.

Purification of II PCR products

- PCR products are purified with the Agencourt AMPure XP beads (Beckman Coulter) in accordance with the manufacturer's instructions.
- 1.5 x sample volume: 20 µl II PCR product and 30 µl Ampure.
- Elution to 20 µl H_2O.

Concentration measurement of II PCR products, sample pooling, purification of the pooled sample, and final concentration measurement of the pool

- Concentration measurement assay performed for amplicons of II PCR.
- Equal amounts of DNA from each sample (now 40 ng) are added to the pooled sample.
- Note that the pool’s final concentration should be sufficient for the shearing reaction; we need to have 100 ng of DNA for 50 µl of shearing reaction.
- AMPure XP purification (1.5 x) of the pooled sample is performed; note that one can elute the pooled sample in a proper H_2O volume in the AMPure XP final step.
- A Qubit 2.0 fluorometer with a high-sensitivity dsDNA assay kit is used on the purified pool.

Shearing and AMPure XP purification of the sheared pool

Components	50 µl reaction
10x Ion Shear Plus buffer	5 µl
Pooled sample DNA (8 ng/µl)	13 µl
Nuclease-free water	22 µl
Ion Shear Plus Enzyme Mix II (Life Technologies)	10 µl

- Chemical shearing is performed using 100 ng of DNA from the pooled sample in a 50 µl reaction volume with Ion Shear Plus Reagents Kit (Life Technologies) components.
- The reaction mixture is incubated at 37 °C for about 10 min (time depends on the demands of the fragment size), and reaction is ended on ice.
- The sheared DNA is purified using AMPure XP purification beads

Ligation of the P1 sequencing adapter

Components	50 µl reaction
AMPure purified template from previous step	37 µl
10× ligase buffer	5 µl
P1 sequencing adapter	1 µl
dNTPs mix	1 µl
DNA ligase (Life Technologies)	2 µl
Nick repair polymerase	4 µl

- Ion Plus Fragment Library kit (Life Technologies).
- P1 adapter from the Ion Xpress Barcode Adapters 1-16 Kit (Life Technologies).
- The reaction mix is placed in a thermal cycler, and the following program is used: the first cycle is 15 min at 25 °C, the second cycle is 5 min at 72 °C, and final hold is performed at 4 °C.

Pippin Prep size selection, size-selected product amplification, and AMPure XP purification

- 2% agarose cassettes, external markers, and reagents (100-600 bp) for the Pippin Prep instrument (Sage Science).
- The entire sample from previous step, with 16 µl of Pippin Prep 4x buffer, is added to the well.
- The “Tight” programming mode with 500 bp is used.
- Library amplification is usually required, particularly if the input to the shearing reaction is <100 ng.
- Platinum PCR Super Mix High-Fidelity from the Ion Plus Fragment Library kit (Life Technologies) is used for amplification in accordance with the manufacturer’s instructions.
- Note: avoid over-amplification, minimize the number of cycles (now nine cycles).

TaqMan, Bioanalyzer, or TapeStation

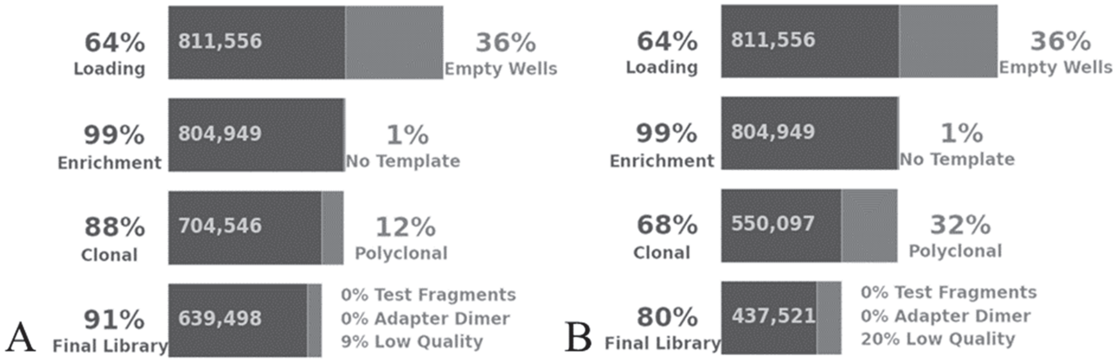
- Concentration and size determination for the Ion Torrent emulsion PCR.

Emulsion PCR (Ion sphere quality control included), bead washing, bead enrichment, and Ion Torrent sequencing with PGM

- This is performed in accordance with the manufacturer's instructions using Life Technologies reagents.

Reanalysing the data for better polyclonality detection

- A low polyclonality percentage may indicate that the polyclonal filtering must be better focused.
- The default settings used to check polyclonal spheres are during flows 12–70.
- The flow number is not comparable with the base pair number because during each flow, the incorporation of the base does not occur.
- Polyclonality filtering must be changed, for example, for flows 120–160, when long internal adaptors are used.
- Data reanalysis begins with signal processing, and a new command (in Torrent Suite 4.2.1 software) is written into Analysis args command's line: Analysis --from-beadfind --use-alternative-etbR-equation --mixed-first-flow=120 --mixed-last-flow=160.



Polyclonality detection: Reanalysing the sequencing data for better polyclonality detection. Polyclonal sphere detection was largely improved via reanalysing with the corrected settings. (A) With the default settings (12–70), a considerable proportion of polyclonal spheres remain undetected. (B) When the settings were adjusted to flows 120–160, better detection of polyclonal spheres is achieved.

II

SAMPLE PRESERVATION, DNA OR RNA EXTRACTION AND DATA ANALYSIS FOR HIGH-THROUGHPUT PHYTOPLANKTON COMMUNITY SEQUENCING

by

Mäki, A., Salmi, P., Mikkonen, A., Kremp, A. & Tiirola, M. 2017

Frontiers in Microbiology 8: 1848.

https://doi.org/10.3389/fmicb.2017.01848 frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 September 2017
doi: 10.3389/fmicb.2017.01848

Sample Preservation, DNA or RNA Extraction and Data Analysis for High-Throughput Phytoplankton Community Sequencing

Anita Mäki[1]*, *Pauliina Salmi*[1], *Anu Mikkonen*[1], *Anke Kremp*[2] *and Marja Tiirola*[1]

[1] Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland, [2] Marine Research Centre, Finnish Environment Institute, Helsinki, Finland

Phytoplankton is the basis for aquatic food webs and mirrors the water quality. Conventionally, phytoplankton analysis has been done using time consuming and partly subjective microscopic observations, but next generation sequencing (NGS) technologies provide promising potential for rapid automated examination of environmental samples. Because many phytoplankton species have tough cell walls, methods for cell lysis and DNA or RNA isolation need to be efficient to allow unbiased nucleic acid retrieval. Here, we analyzed how two phytoplankton preservation methods, three commercial DNA extraction kits and their improvements, three RNA extraction methods, and two data analysis procedures affected the results of the NGS analysis. A mock community was pooled from phytoplankton species with variation in nucleus size and cell wall hardness. Although the study showed potential for studying Lugol-preserved sample collections, it demonstrated critical challenges in the DNA-based phytoplankton analysis in overall. The 18S rRNA gene sequencing output was highly affected by the variation in the rRNA gene copy numbers per cell, while sample preservation and nucleic acid extraction methods formed another source of variation. At the top, sequence-specific variation in the data quality introduced unexpected bioinformatics bias when the sliding-window method was used for the quality trimming of the Ion Torrent data. While DNA-based analyses did not correlate with biomasses or cell numbers of the mock community, rRNA-based analyses were less affected by different RNA extraction procedures and had better match with the biomasses, dry weight and carbon contents, and are therefore recommended for quantitative phytoplankton analyses.

Keywords: next generation sequencing, phytoplankton, cell lysis, operational taxonomic units, Lugol

OPEN ACCESS

Edited by:
Télesphore Sime-Ngando,
Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Nico Salmaso,
Fondazione Edmund Mach, Italy
Jaak Truu,
University of Tartu, Estonia

****Correspondence:***
Anita Mäki
anita.maki@jyu.fi

Specialty section:
This article was submitted to Aquatic Microbiology, a section of the journal Frontiers in Microbiology

Received: *25 March 2017*
Accepted: *11 September 2017*
Published: *26 September 2017*

Citation:
Mäki A, Salmi P, Mikkonen A, Kremp A and Tiirola M (2017) Sample Preservation, DNA or RNA Extraction and Data Analysis for High-Throughput Phytoplankton Community Sequencing. Front. Microbiol. 8:1848. doi: 10.3389/fmicb.2017.01848

INTRODUCTION

Phytoplankton is often used to monitor the status of aquatic ecosystems, and effective methods for the characterization of phytoplankton samples are needed. Traditionally, phytoplankton community compositions have been studied using microscopic techniques and observing morphological characteristics. When applying microscopic identification methods, specific

Abbreviation: TTR, theoretical template relationship.

professional skills are needed and results can depend on the subjective interpretations. Small nano- and picoplanktonic cells are also difficult, if not impossible, to identify to species level (Eiler et al., 2013). These drawbacks are, for the most part, avoidable applying molecular methods for identification.

Next generation sequencing methods (NGS) enable DNA- and RNA-based analyses of uncultured species and, with exploiting the data cumulating in the data banks, biodiversity evaluation of phytoplankton can be renewed. Strong positive correlation between rRNA gene copy numbers and genome size (Prokopowich et al., 2003) or cell length in cultured algal strains (Godhe et al., 2008) gives promises for developing molecular monitoring of phytoplankton biovolumes to support and substitute microscopying. Although highly attractive, sequencing of phytoplankton samples has several challenges, which hinder the application of the tool. For phytoplankton, it is difficult to find broad-range PCR primers, and therefore primer bias can skew the actual diversity scene of microbes in community studies (Hong et al., 2009; Hadziavdic et al., 2014; Hugerth et al., 2014; Bradley et al., 2016). Another obstacle for molecular phytoplankton analysis arises from the lack of the classified sequences in the databases (Abad et al., 2016). Although several reference databases exist for rRNA genes of prokaryotes (SILVA, Greengenes, RDP) and for plastidial rRNA genes (Decelle et al., 2015) for photosynthetic eukaryotes, overall taxonomic resolution for phytoplankton is poor and scattered. As the NGS and single-cell technologies mature, we can expect expanding libraries and increasing lengths and qualities of reads, which will increase the taxonomic resolution of molecular phytoplankton analysis.

One challenge involves DNA/RNA extraction from the cells, as many comparative studies have described differences in isolation efficiencies (Stach et al., 2001; Hong et al., 2009; Simonelli et al., 2009; Rosic and Hoegh-Guldberg, 2010; Eland et al., 2012; Koid et al., 2012). Sample preservation in Lugol or by freezing, cell lysis and nucleic acid extraction without degradation are critical steps that can complicate the isolation of DNA and RNA from phytoplankton cells. Environmental samples contain cells with diverse cell properties, varying in cell size and firmness of cell walls, which may favor certain cells when using particular extraction procedures. Various physical, chemical and enzymatic cell lysis protocols are used in commercial kits, but bead-beating has become a gold standard. Yuan et al. (2015) found that bead-beating method can double the DNA yield of some phytoplankton species in comparison with the enzymatic non-bead-beating method. Eland et al. (2012) has suggested that additional freeze-thaw lysis might influence the effectiveness of beat beating. Although NGS enables molecular assessment of the diversity of microbial eukaryotic communities (Lie et al., 2014), factors like the primer bias and differences in DNA or RNA isolation efficiencies can mask the actual phytoplankton diversity and skew the results of environmental samples.

To study how sample preservation and the nucleic acid extraction methods affect NGS analysis of phytoplankton communities, we made a comprehensive experiment with a mock community comprising three algal classes (diatoms, dinoflagellates and green algae), two strains per each class. Sequencing results were compared against microscopic observations, dry masses and carbon contents of the mock cell pool. When finding that DNA-based analysis did not follow the biomass estimates we evaluated the variation in the rRNA gene copy number per DNA by using qPCR-based approaches on the separately extracted mock strains. Bioinformatics was optimized by performing the NGS sequencing for individually barcoded mock strain samples and evaluating the distribution of sequences in this model data during the steps of the trimming pipeline.

MATERIALS AND METHODS

Study Strategy

In this comparative study, NGS results of the mock community pool of six phytoplankton strains were analyzed according to used sample preservation and nucleic acid extraction methods (**Figure 1**). To interpret the NGS results, DNA samples were isolated from separate strains, and reference library of the 18S rRNA gene sequences was created applying Sanger sequencing. The NGS results were compared with original cell numbers, biomass and carbon content values in the mock pool. To evaluate the reliability of the nucleic acid isolation methods, several tests were done for separate strains. The match of the selected eukaryotic primer pair was tested *in silico* against the database and *in vitro* using quantitative PCR (qPCR) with an independent primer pair. The 18S rRNA gene copy numbers per extracted DNA and per cell were determined for each strain. TTR of rRNA genes in the original cell pool was calculated using gene copy numbers from equal volumes of extracted DNA (Power Biofilm extraction) of each strain. In the other test, separately extracted DNAs (Power Biofilm extraction) were combined in equal DNA amounts, and NGS was performed as in the original protocol. This test was done to reveal the potential bias due to preferential amplification of certain ribosomal sequence types during amplification. Therefore TTR-analysis avoided competition of primers, and "pooled DNA" analysis showed theoretical results if the DNA yields (in ng) of all mock cell cultures would have been equal. For optimizing bioinformatics pipelines, the effects of trimming procedures were evaluated with separately barcoded data of mock strains.

For the nucleic acid extraction experiments, cells of the mock community were pooled and stored in Lugol or by deep-freezing, and DNA or RNA extracts were isolated using different methods. From the extracted DNA and random primed cDNA, 18S rRNA genes were amplified using eukaryotic primers. After NGS and clustering sequences into OTUs, results were aligned to the reference sequences obtained by Sanger sequencing, and strain-specific proportions of sequences, after different cell-restoring and nucleic-acid extraction methods, were compared.

Microscopic Analysis of Phytoplankton Mock Community Strains

Non-axenic strains of 6 phytoplankton species isolated from the Baltic Sea (Hällfors and Hällfors, 1992) included *Diatoma tenuis,*

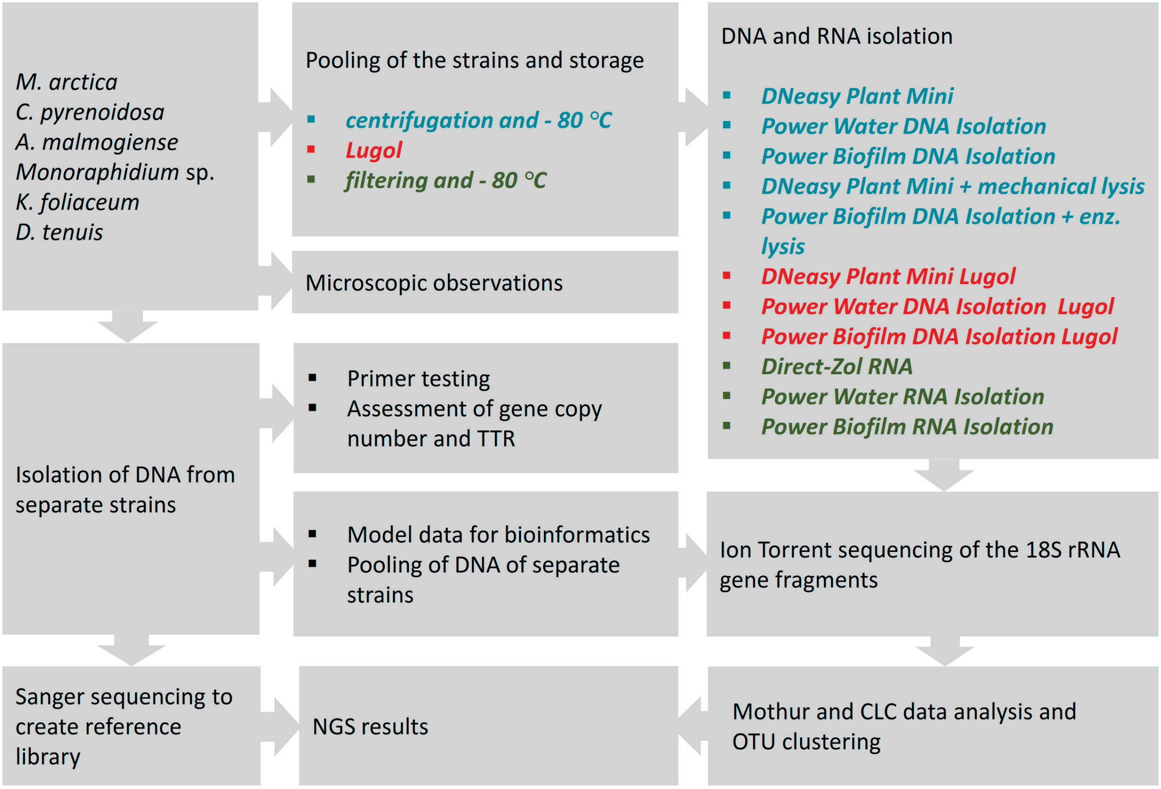

M. arctica
C. pyrenoidosa
A. malmogiense
Monoraphidium sp.
K. foliaceum
D. tenuis
Pooling of the strains and storage
centrifugation and - 80 °C
Lugol
filtering and - 80 °C
Microscopic observations
DNA and RNA isolation
DNeasy Plant Mini
Power Water DNA Isolation
Power Biofilm DNA Isolation
DNeasy Plant Mini + mechanical lysis
Power Biofilm DNA Isolation + enz. lysis
DNeasy Plant Mini Lugol
Power Water DNA Isolation Lugol
Power Biofilm DNA Isolation Lugol
Direct-Zol RNA
Power Water RNA Isolation
Power Biofilm RNA Isolation
Isolation of DNA from separate strains
Primer testing
Assessment of gene copy number and TTR
Model data for bioinformatics
Pooling of DNA of separate strains
Ion Torrent sequencing of the 18S rRNA gene fragments
Sanger sequencing to create reference library
NGS results
Mothur and CLC data analysis and OTU clustering

FIGURE 1 | The study strategy for the NGS analysis of a mock community, comprising of six phytoplankton strains, to compare cell-preserving and nucleic acids extraction methods. For nucleic acid isolation, three DNA extraction kits with their enzymatic or mechanical cell lysing modifications, and three RNA extraction methods were applied to study cell pools, preserved at –80°C or in acidic Lugol's solution. TTR refers to theoretical template relationship, OTU refers to operational taxonomic unit. Full strain names are presented in Section "Materials and Methods."

Melosira arctica, Apocalathium malmogiense, Kryptoperidinium foliaceum, Monoraphidium sp., and *Chlorella pyrenoidosa*, which were obtained from the Culture Collection of the Marine Research Centre, Finnish Environment Institute (SYKE MRC)/Tvärminne Zoological Station, University of Helsinki (Supplementary Table 1).

Phytoplankton cells were stained and mounted in ProLong Diamond Antifade Mountant with DAPI (Thermo Fisher Scientific, United States). Poly-L-Lysine (Sigma–Aldrich) was used to coat the coverslips to attach the cells. Imaging of the cells was performed using Zeiss Cell Observer HS wide-field microscope, Colibri LED light source at 365 nm wavelength for DAPI, Plan-Apochromat 63x (NA = 1.4) Ph3 oil immersion and Plan-Apochromat 100x (NA = 1.46) objectives and filter set 49 (excitation 365 nm and emission 445/50 nm).

Wet volume (biomass) and cell numbers of the mock community samples were assessed using Zeiss Axio Vert.A1 epifluorescence microscope applying counting strategy described by Salmi and Salonen (2016). Dry mass and carbon content was analyzed from the deep-frozen cell pellets (next chapter), dried in tin cups for 20 h at 65°C. The dry weight was determined using Sartorius M2P and Sartorius CP2P and the carbon mass in the dry weight sample was analyzed using the Thermo Delta V stable isotope mass spectrometer.

Preservation of the Phytoplankton Cells, Nucleic Acids Extraction and cDNA Synthesis

Before starting the mock community study, freshly grown 2 mL cell culture of each species was harvested by centrifuging at 3500 *g* for 10 min, and supernatant was removed leaving cell pellet and 100 μL of culture medium in the tubes. DNA was extracted separately from these cell pellets using Power Biofilm DNA isolation Kit according to manufacturer's instructions (MoBio Laboratories, Inc., Carlsbad, CA, United States) to test suitability of the primers, to inspect gene copy numbers per extracted DNA and per cell, to construct reference library applying Sanger sequencing, to pool equal DNA quantities for the control DNA pool, and to produce separately barcoded model data for optimizing the trimming pipeline (details in Section "Amplification of 18S rRNA Gene Fragments and Sequencing".)

For comparative analysis of cell preservation and DNA isolation methods, equal volumes of fresh, in the active cell growth phase growing cultures of the mock community were pooled and divided into 2 mL aliquots, which were centrifuged at 3500 *g* for 10 min to obtain 100 μL of cell-suspension, which was kept frozen at −80°C for 2 weeks. To test if storing cells in Lugol affects sequencing results, part of the pooled sample was stored in 1% acidic Lugol's solution (final concentration) at +4°C, and 2 mL aliquots were centrifuged at 3500 *g* for 10 min to obtain 100 μL of cell-suspension before DNA extraction. Cellular DNA was extracted from frozen and Lugol preserved cells using DNeasy Plant Mini Kit (Qiagen, United States), Power Water DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, CA, United States), and Power Biofilm DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, CA, United States). To determine, if addition of mechanical cell destruction would improve the cell lysis and consequently DNA yield, *DNeasy Plant Mini Kit DNA + mech.* extraction was done using manufacturer's instruction with additional mechanical treatments. Cells were exposed to extra freeze/thaw cycle by dipping them into the liquid nitrogen and disrupting cells by beat-beating at maximum vortex speed for 10 min in 0.1 mm Glass Beads Tubes (MoBio Laboratories, United States) in AP1 buffer (Qiagen). *Power Biofilm DNA Kit + enz.* DNA isolation was extended with additional enzymatic treatment, starting with inactivation of DNases by incubating cells at 75°C for 10 min (Wiame et al., 2000) and continuing incubation in Viscozyme enzyme solution (60 mg/ml) (Sigma–Aldrich) at 50°C for 1 h and after that in Proteinase K (0.5 mg/ml) (Thermo Scientific, United States) enzyme/TE-buffer (pH 8)/SDS (0.5 %) solution at 50°C for 1 h. After these additional mechanical or enzymatic treatments, isolation continued according to manufacturer's instructions. Three replicates were performed from all isolation methods and their variations. DNA concentration was checked using Qubit 2.0 Fluorometer and dsDNA High Sensitivity Assay Kit (Life Technologies, United States).

To perform RNA based sequencing analyses, 2 mL of fresh, pooled sample (from same pool as used in DNA extractions) was filtered through 25 mm diameter and 0.22 μm pore size polyethersulfone Millipore Express PLUS Membrane Filters (GPWP02500, Millipore, United States) using 25 mm Swinnex Filter Holders (SX0002500, Millipore, United States). After filtration the membranes were directly inserted into the MoBio Glass Beads Tubes before freezing to prevent RNA degradation when starting RNA isolation, so that lysis buffer could be added to the frozen cells. Samples were frozen at −80°C without delay and kept in freezer for 3 weeks before Direct-Zol RNA Micro Prep isolation (Zymo Research, Irvine, CA, United States) and for 2 months before MoBio Power Water and Power Biofilm RNA isolations (MoBio Laboratories, Inc., Carlsbad, CA, United States). Lysis buffer was added into the bead tubes before melting the sample tubes. Procedures of Power Water RNA Isolation Kit and Power Biofilm RNA Isolation Kit followed manufacturer's specialized instructions to co-extract small RNA fractions. Direct-Zol kit consists of spin column purification of RNA from TRIzol, which was added into bead tubes containing frozen mock sample filters. Bead tubes were vortexed at maximum speed for 1 min and centrifuged at 12000 × *g* for 1 min before supernatant collection. Because of the low RNA yield with MoBio kits, GeneJET RNA Cleanup and Concentration Micro Kit (Thermo Scientific, United States) was used to concentrate the RNA samples. RNA integrity and concentration was determined using TapeStation 2200 applying the High Sensitivity RNA ScreenTape system (Agilent Technologies, United States) and Qubit 2.0 Fluorometer applying the RNA Assay Kit (Life Technologies, United States).

The cDNA was synthesized by reverse transcription applying RevertAid First Strand cDNA Synthesis Kit's (Thermo Scientific, United States) using random priming from 50 ng (Power Water RNA), 5 ng (Power Biofilm RNA) and 60 ng (Direct-Zol RNA) of total RNA.

Amplification of 18S rRNA Gene Fragments and Sequencing

Two sets of 18S rRNA gene primers were tested *in silico* with the program SILVA TestPrime and *in vitro* using quantitative PCR (qPCR) to analyze whether primer pairs, Euk1A (5′-CTGGTTGATCCTGCCAG-3′) and reverse Euk516R (5′-ACCAGACTTGCCCTCC-3′) (Díez et al., 2001; Eland et al., 2012), and V8F (5′-AT AACAGGTCTGTGATGCCCT-3′) (Bradley et al., 2016) and reverse 1510R (5′-CCTTCYGCAGGTTCACCTAC-3′) (Amaral-Zettler et al., 2009) would anneal and amplify DNA sequences of mock community species. Primer pairs targeted different variable (V) regions, V1 to V3 and V8 to V9, respectively. Equal 4 ng amount of DNA extract from each strain was used as a template in separate reactions, using Bio-Rad CFX96 real time thermal cycler (Bio-Rad Laboratories) and Maxima SYBR Green/Fluorescein qPCR Master Mix (Thermo Scientific, United States) in a 25 μl reaction mixture with 0.4 μM of primers. The qPCR procedure started with an initial denaturation step at 94°C for 3 min and continued with 35 cycles of amplification (94°C for 30 s, 52°C for 1 min and 72°C for 1 min) with final extension at 72°C for 5 min. Since the M13-tail (5′-TGTAAAACGACGGCCAGT-3′) in the 5′-end of the Euk1A-forward primer was needed for sample barcoding for NGS sequencing, qPCR amplification reactions was also done with M13-Euk1A/Euk516R primer pairs to test if the tail would interfere the amplification. The Euk1A/Euk516R primers were used to prepare templates of individual strains for Sanger sequencing (Sanger et al., 1977). However, for *K. foliaceum* direct Sanger sequencing of the 18S rRNA fragment was only successful after isolating RNA, cDNA synthesis and cloning using the CloneJet PCR Cloning Kit (Thermo Scientific, United States). Sanger sequences of mock community strains were deposited in the European Nucleotide Archive (ENA) under study accession number PRJEB22147.

To study the effect of DNA extraction methods on the sequencing results, 3 ng of extracted DNA template and primer pair M13-Euk1A/Euk516R were used and the same PCR procedure was applied as above, except that PCR amplification was limited to 30 cycles. For the cDNA samples derived from the reverse transcription reaction, 2, 3, or 3 μL of cDNA of Direct-zol RNA isolation, Power Water RNA isolation, and Power Biofilm RNA isolation, respectively, was used as a template, and the amplification followed the same procedure as for DNA samples.

First PCR amplification was followed by the eight cycles of second PCR to add the barcoded sequencing adaptor IonA-M13. Barcoding of amplicons, size-trimming of the products and final Ion Torrent sequencing was done using the Ion Torrent Personal

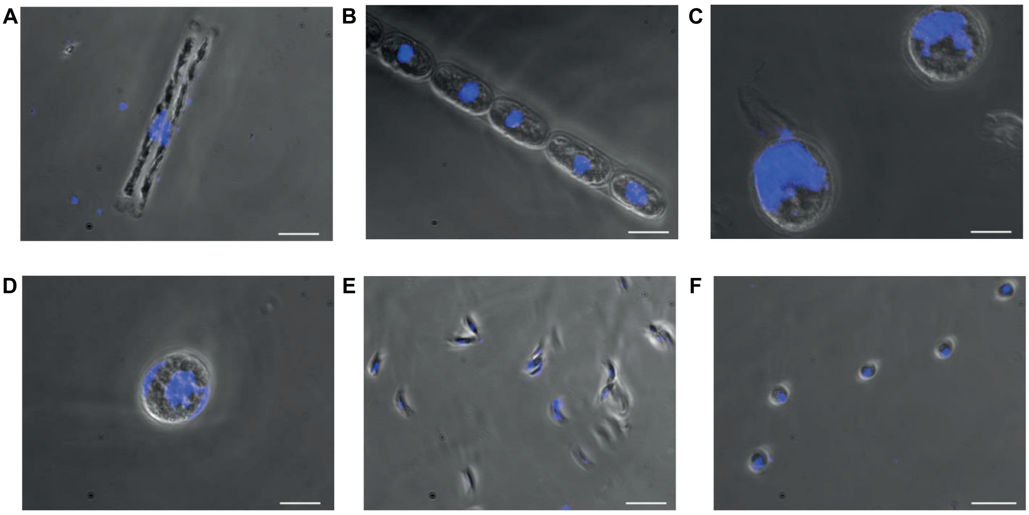


FIGURE 2 | Nucleus and cell sizes of phytoplankton cells **(A)** *D. tenuis,* **(B)** *M. arctica,* **(C)** *A. malmogiense,* **(D)** *K. foliaceum,* **(E)** *Monoraphidium* sp., and **(F)** *C. pyrenoidosa* were compared using microscope for visualization. DNA of the cells was stained with DAPI. Scale bars, 20 μM.

Genome Machine (PGM) as described by Mäki et al. (2016), except using the Hi-Q and Hi-Q View OT2 Kit, Hi-Q and Hi-Q View Sequencing Kit, and Ion 316v2 chip (Life Technologies).

The copy number of 18S rRNA gene was determined for each strain separately from 2 μL volume of DNA extracts (i.e., representing equal volumes of original cultures when pooled) to predict the theoretical template relationships in the mock pool (TTR). DNA extracts were used as a template and Euk1A/Euk516R as primers in the qPCR reaction, and copy numbers were determined with duplicate 5-point standard series of mock community member PCR products ranging from 1.5×10^4 to 1.5×10^8 (amplification efficiency 85%, y-intercept 41 cycles). For creating model data for optimizing the NGS data trimming pipeline, each strain was amplified separately with unique barcodes, applying the same procedures as above. When an equal number (pM) of the barcoded amplicons from each strain was used in subsequent sequencing, any observed biases in abundances can be assumed to have resulted from post-PCR steps: sequencing and/or sequence analysis. To check the effect of primer bias, variation in the gene copy numbers and theoretical results, if the DNA yields of all mock cell cultures would have been equal, 4 ng of isolated DNA from each strain was pooled, amplified, barcoded, and sequenced with the same reagents and procedures as above. Amplification products were analyzed using Agilent 2200 TapeStation system with the High Sensitivity D1000 ScreenTape (Agilent, United States) and in the agarose gel electrophoresis prior to sequencing. All control analyses were done in triplicate.

Data Analysis

The model data was utilized for evaluating and optimizing the trimming procedure. PGM sequencing data was initially trimmed with Torrent Suite 5.0.4 software including default adapter removal and adjusted polyclonality filtering (command: "--mixed-first-flow = 120 --mixed-last-flow = 160") because of the long internal adaptors. Default 3′ end quality trimming of Torrent Suite can be turned off (command: "--trim-qual-cutoff 100"), so both 3′ end trimmed reads and reads without 3′ end trimming were imported into Mothur v.1.36.1 (Schloss et al., 2009) and CLC Genomics Workbench 9.5.1 software[1]. The trimming workflows of Mothur and CLC software were evaluated using the separately barcoded 18S rRNA gene data of mock strains, pooled in equimolar concentrations (see the trimming parameter in results, Optimizing the bioinformatics pipeline). The de novo OTU clustering in Mothur v.1.36.1 was performed using average neighbor algorithm and in CLC using distance-based greedy algorithm UCLUST (Supplementary Tables 4, 5).

The trimming pipeline that best preserved the original relationships among barcode bins was chosen for further analyses. In this protocol PGM reads were first processed using Torrent Suite 5.0.4 software without 3′ end quality trimming. After initial processing of the reads, fastq files were imported into CLC software where the quality trimming was performed according the parameters gained from the model data analysis at OTU 97% identity clustering level (Supplementary Table 4).

Relative abundances of strains were square-root transformed before calculation of Bray-Curtis similarity matrix, based on which non-metric multidimensional scaling (NMDS) was calculated with 1000 repeats in PRIMER v. 6.1.12 and PERMANOVA+ v.1.0.2 (PRIMER-E/Quest Research Limited, Albany, New Zealand).

RESULTS

Phytoplankton Cells

In this study, cultures of *D. tenuis* and *M. arctica* (Diatomophyceae), *A. malmogiense* and *K. foliaceum*

[1] www.qiagenbioinformatics.com

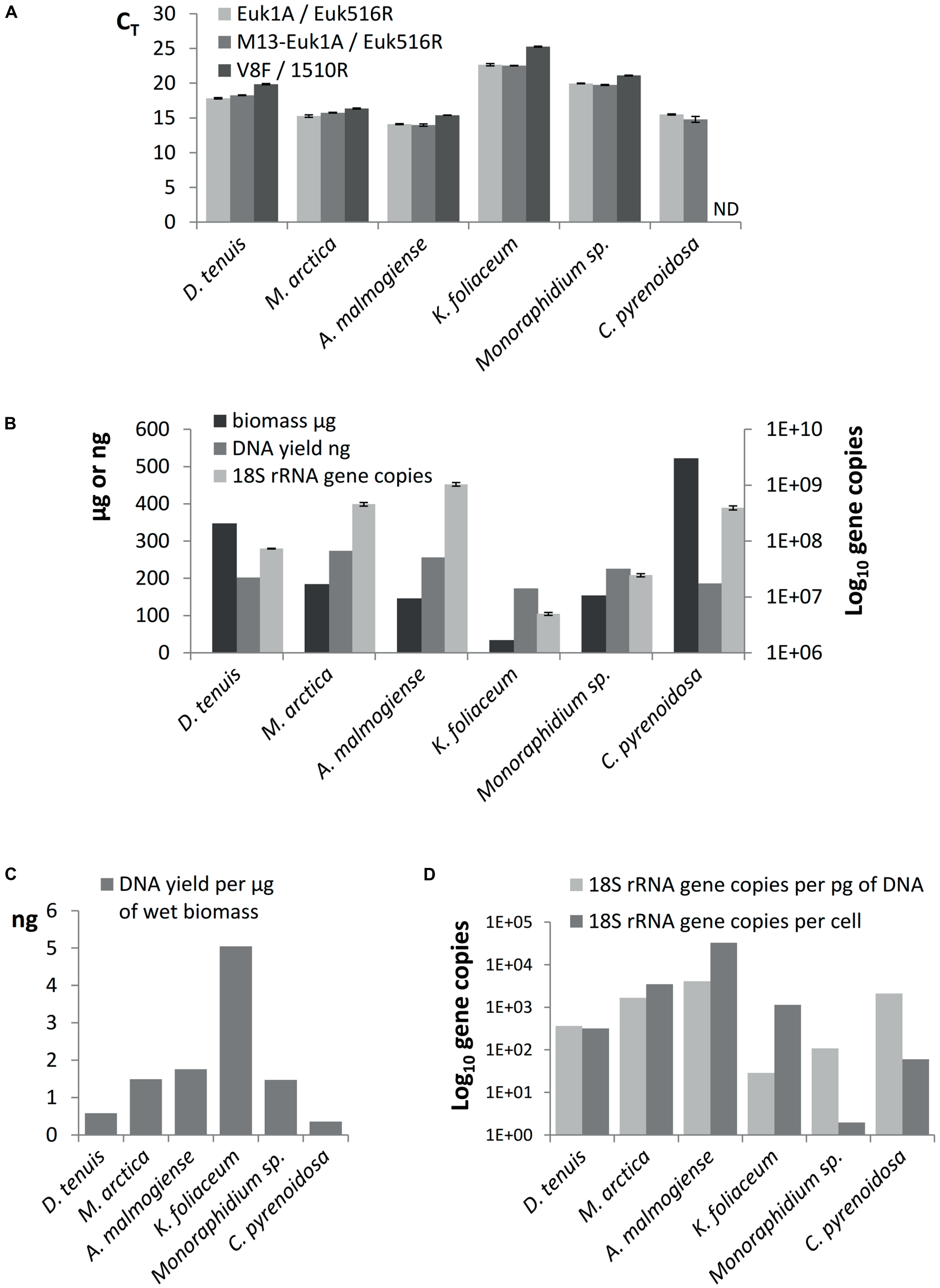


FIGURE 3 | (A) Comparison of threshold cycles (C_T) of qPCR when amplification of 18S rRNA genes was performed using Euk1A/Euk516R, M13-Euk1A / Euk516R and V8F / 1510r primer pairs from equal quantity of genomic DNA (4 ng). **(B)** Total wet biomasses, DNA yields, and 18S rRNA gene copy numbers per Power Biofilm extraction. qPCR results are presented as mean values of triplicates with standard errors. ND indicates the sample that was not possible to amplify with the V8F/1510R primer pair. **(C)** DNA yields per ng of wet biomass. **(D)** 18S rRNA gene copy numbers per cell and per pg of extracted DNA.

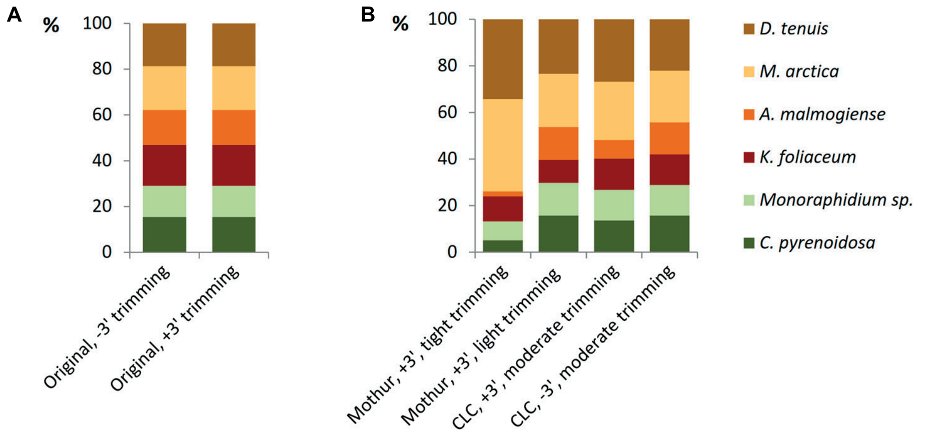


FIGURE 4 | The effect of trimming criteria on the NGS results of the model data of separately barcoded sequences of mock strains. **(A)** Torrent Suite 3′ end trimming of reads (+3′) retained the original relationship of barcoded reads as well as without trimming (–3′) imported reads. **(B)** When 3′ end trimmed (+3′, in Torrent Suite) reads were imported into the Mothur software and tight trimming procedures were applied, proportions of the strains became distorted. Resolving trimming lighter, proportions appeared more equal. The pipeline without 3′ end trimming in Torrent Suite and moderate trimming in CLC was selected for further analyses.

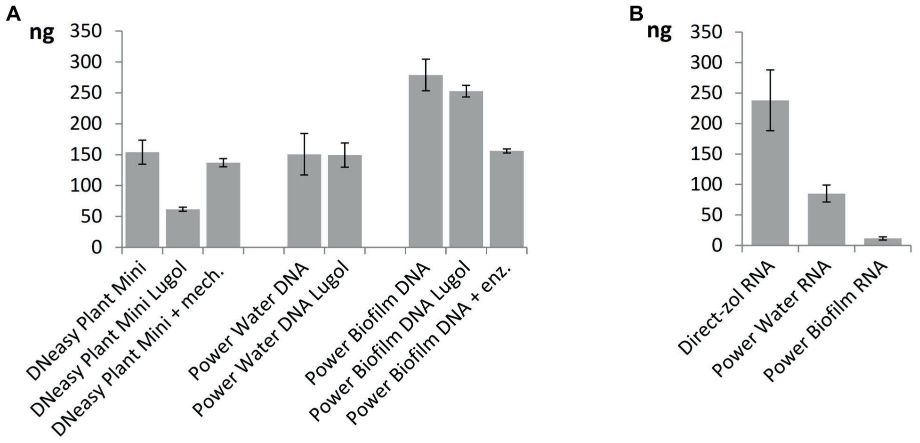


FIGURE 5 | Yields of nucleic acids from the mock phytoplankton community using **(A)** three DNA extraction kits with or without Lugol preservation or extra enzymatic or mechanical cell lysis steps and **(B)** three RNA extraction kits. All but Lugol stored samples were preserved at −80°C. Total extracted DNA or RNA yields are presented as mean values of three isolations with standard errors.

(Dinophyceae), *Monoraphidium* sp. and *C. pyrenoidosa* (Chlorophyceae) phytoplankton cells (Supplementary Table 1) were observed using a light microscope to determine the biomass, cell number (Supplementary Table 2), and the location and size of nucleus. Nuclei sizes varied between 2 and 28 μm among the species, being largest in *A. malmogiense* and *K. foliaceum* cells (**Figure 2**). A second nucleus of a diatom endosymbiont was visible in the dinoflagellate *K. foliaceum* (**Figure 2D**). NGS results confirmed the purity of the cultures and specificity of the primers used in the study, as 98% (variation 93–100%) of the NGS sequences could be classified to the six target strains when strains were sequenced separately or in the mock community pool (Supplementary Figure 1A). Only the data of dinoflagellate *K. foliaceum* contained 14% non-target sequences, which were derived from the known endosymbiont nucleus of diatom origin (Figueroa et al., 2009).

Amplification and Sequencing of the Partial 18S rRNA Gene

The results of SILVA TestPrime test (Supplementary Table 3) indicated that the Euk1A F / Euk516R primer pair was appropriate for amplification of fragments of the 18S rRNA gene of all six phytoplankton species. The threshold cycles (C_T) of qPCR of separately extracted mock strain DNAs confirmed that Euk1A/Euk516R primer pair amplified 18S rRNA gene of all species and M13-adapter part in forward primer did not affect the amplification efficiency (**Figure 3A**). Although the qPCR results with V8F/1510r primer pair mostly corresponded to C_T values of the other primer pair, this pair only amplified 5 of the 6 study species (not *C. pyrenoidosa*). 18S rRNA gene copy numbers in the extracted DNA were determined from the qPCR performed with the Euk1A / Euk516R primer pair (**Figure 3B**). The results

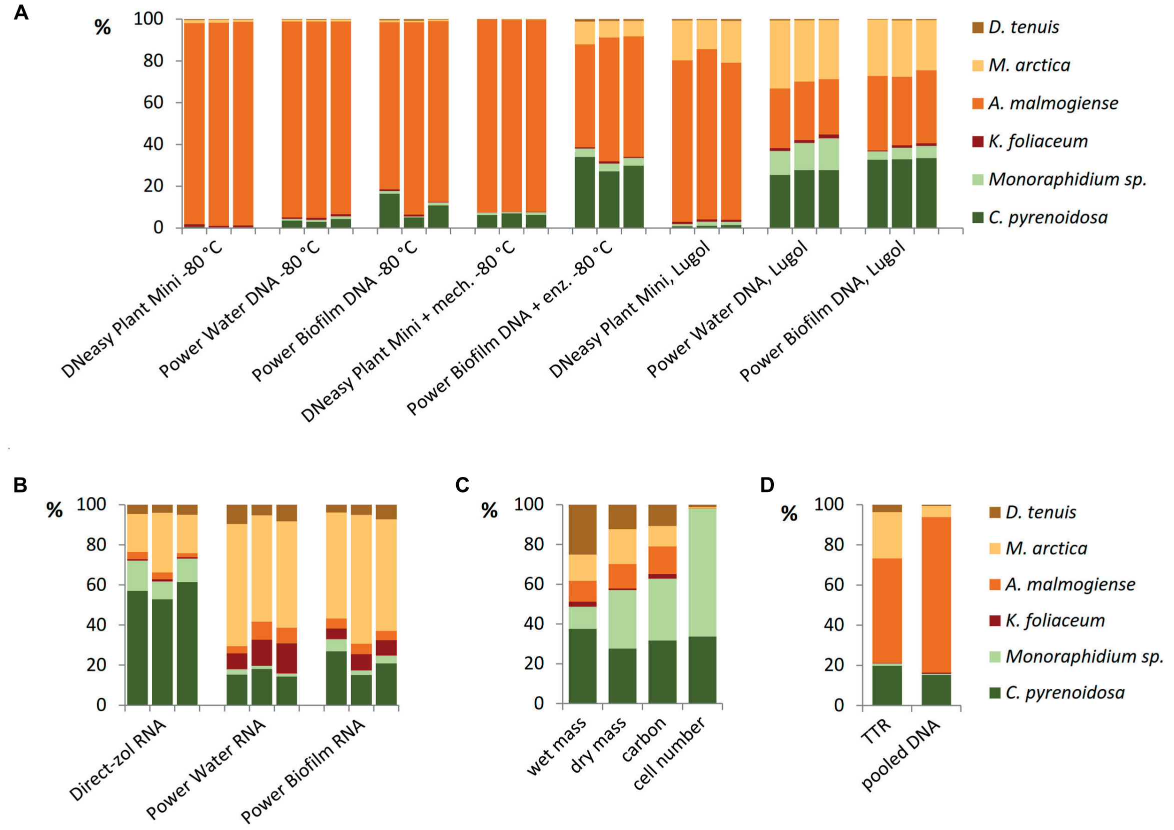


FIGURE 6 | Relationships of mock community strains based on **(A)** 18S rRNA gene analyses of DNA extractions and **(B)** 18S rRNA analyses of RNA extractions, **(C)** microscopically determined wet biomasses and cell numbers and dry mass and carbon content determinations, and **(D)** theoretical results based on separately extracted DNA's of mock strains using Power Biofilm kit. TTR refers to theoretical template relationship, when DNA extracts of strains were isolated and amplified separately counting the results based on individual 18S qPCR results, and pooled DNA refers to a sample containing equal amounts (weight) of DNA of mock strains.

showed 100-fold differences in the rRNA gene numbers in the mock strain DNAs, without correlation to original biomasses or DNA yields. *K. foliaceum* had the highest DNA yield per biomass (**Figure 3C**). Calculated ribosomal copy numbers per cell varied between 2 in *Monoraphidium* sp. to 33 000 in *A. malmogiense* (**Figure 3D**), which means over 10^4 variation in the rRNA operons per cell among the study strains.

Optimizing the Bioinformatics Pipeline

The model data of separately barcoded mock strain sequences was collected to optimize the quality trimming pipeline. 3′ end trimmed reads and reads without 3′ end trimming were imported from Torrent Suite 5.0.4 software to Mothur or CLC software and, before further trimming, strain-specific proportions of reads were equal in both the data sets (**Figure 4A**). When using the Mothur software for 3′ end trimmed reads and imposing tight quality requirements, such as minimum length of 180 bases and minimum quality average of 20 over a sliding window of 10 nucleotides (Supplementary Table 5), considerable number of *A. malmogiense* sequences were trimmed off and excessive increase of *M. arctica* sequences was observed (**Figure 4B**). When trimming requirements in Mothur were relaxed, with minimum length of 150 bases and no sliding-window quality check, proportions of sequences were less biased. The other trimming processes in both cases were kept similar, including in maximum two allowed mismatches in the primer region, one mismatch in the barcode region, and the maximum homopolymer length of eight. In the CLC pipeline, when reads without 3′ end trimming were imported from Torrent Suite and minimum length of read was imposed to 150, proportions of sequences followed the original distribution better than when starting with 3′ end trimmed reads. More careful examination of trimming revealed that the site that induced temporary decrease in the quality values was a loop in the rRNA gene structure. In the CLC program the default modified-Mott quality algorithm was used for end-trimming with error probability limit of 0.05. In both software programs, $\text{OTU}_{0.97}$ clustering was applied to identify similar sequences and the OTUs were then classified to species level against the reference library (Supplementary Tables 4, 5). When all OTU sequences, gained from nucleic acids extraction methods and from studies of separately sequenced strains, were aligned against the reference sequences of the mock community, target sequences were the most prevalent of all OTUs (Supplementary Figure 1 and Table 6). Although 98% of the $\text{OTU}_{0.97}$ sequences could be classified to right phytoplankton strains, comparison of rarefaction curves of the whole data at $OTU_{0.97}$ and $OTU_{0.99}$ levels showed that RNA extractions yielded less small sequencing errors than DNA extractions, which may be attributed to less PCR cycles needed for amplification of rRNA genes (Supplementary Figures 1B,C). CLC trimming settings retained the original distribution of sequences and therefore that pipeline was used for further comparison of data from the nucleic acid extraction methods.

Nucleic Acid Yield and NGS Results of the Mock Pool

When comparing nucleic acid extraction methods, highest DNA yields of the mock community was gained with Power Biofilm DNA isolation kit (**Figure 5A**). Additional mechanical lysis steps, here freeze/thaw cycle and beat-beating, or additional enzymatic lysis method, here incubation in Viscozyme/Proteinase K, did not increase the DNA yield when DNeasy Plant Mini Kit or Power Biofilm DNA isolation kit, respectively, were used and when compared to standard protocols recommended by the manufacturer. Lugol preservation decreased the DNA yield of the DNeasy extraction, but did not significantly affect the yield of the other kits. When comparing RNA extraction procedures the highest overall yield and preservation of small RNAs was gained using TRIzol-based Direct-zol RNA isolation method (**Figure 5B**). Size distribution histograms of final RNA extracts illustrate the integrity of RNA after extraction methods, showing best performance by the Direct-zol kit (Supplementary Figure 2).

Different DNA extraction and RNA extraction methods strongly affected the NGS results of the mock pool (**Figures 6A,B**). When Power Water and Biofilm kits, based on mechanical cell lysis (bead-beating), were used, Lugol preservation brought out green algae species better than when cells were preserved at −80°C (**Figure 6A**). Results of microscopic biomass counting (**Figure 6C**) showed that although small green algae species were numerically dominating in the mock pool, biomass values appeared quite evenly distributed, except *K. foliaceum*. (**Figure 6C** and Supplementary Table 2). Wet and dry biomass, cell carbon content, TTR values and NGS results of separately extracted and equally pooled DNA (**Figures 6C,D**) were used as indicators to evaluate different nucleic acid extraction methods. Strain specific copy numbers of 18S rRNA gene, determined using the qPCR, were used to calculate the theoretical template relationships in the original pool (TTR).

All the DNA isolation methods, except samples with additional enzymatic cell lysis, demonstrated that the *A. malmogiense* sequences strongly dominated the data if DNA extraction was done from frozen cells (**Figure 6A**). Preserving cells in Lugol, and usage of additional enzymatic lysis step favored green algae and *M. arctica* species decreasing *A. malmogiense* sequences, when Power kits were in use. Sequences of *K. foliaceum* were very weakly amplified from DNA isolations compared to RNA based sequencing (**Figures 6A,B**). Since *A. malmogiense* was overrepresented in all at −80°C stored DNA samples, and proportion of *M. arctica* increased in the Lugol preserved samples, we did additional DNA extractions using Power Biofilm isolation kit. We wanted to see if storing conditions and additional enzymatic lysis would affect the DNA yield of these two species. In this test additional enzymatic lysis steps did not affect the yield of *M. arctica* or *A. malmogiense* DNA extractions from fresh cells, but promoted a tenfold increase in DNA yields of *A. malmogiense* when the samples were stored at −80°C or in Lugol (**Figure 7**) before DNA extraction.

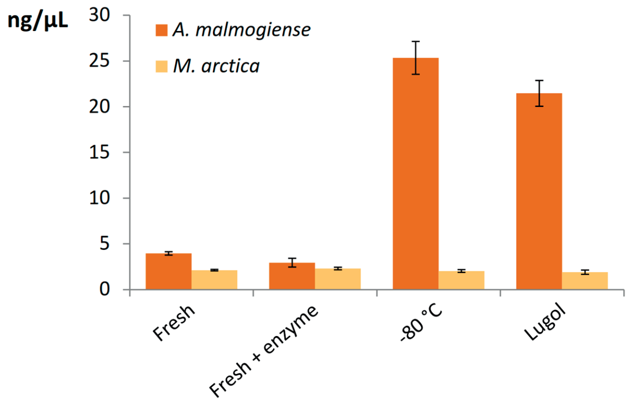


FIGURE 7 | The effect of storage and additional enzymatic lysis steps on the DNA yields of *A. malmogiense* and *M. arctica*. DNA was extracted using Power Biofilm kit from fresh cells with or without additional enzymatic step or from cells stored at –80°C or in Lugol.

Based on the NMDS ordination, DNA-based analyses (samples stored at −80°C), Lugol-preserved DNA-based analyses and RNA-based analyses were separated on the primary (horizontal) axis (**Figure 8A**). In this set, RNA-based analyses (especially Direct-zol extraction) most closely resembled the biomass (**Figures 8A,D**), dry mass and carbon content proportions of the mock cell pool. DNA-based NGS results were mostly affected by the high gene copy numbers of *A. malmogiense*, which overpowered the abundance of other species (**Figures 8A,B**). When the DNA samples were preserved with Lugol, this effect was not as massive, and the data resembled more RNA results, as well as biomass, dry mass and carbon content results (**Figures 8A,C**). The difference between deep-frozen and Lugol-preserved samples was large, even if the concentrations of the DNA extractions were in similar level. Sequence abundances of RNA extraction samples was the best indicator of biomass and, using Direct-Zol RNA isolation method, the presence of both green algae species was remarkable, as also in the biomass calculation (**Figure 8D**).

DISCUSSION

Molecular methods, especially high-throughput sequencing, have shown their effectiveness in the study of diversity and ecology of phytoplankton, potentially replacing traditional microscopic

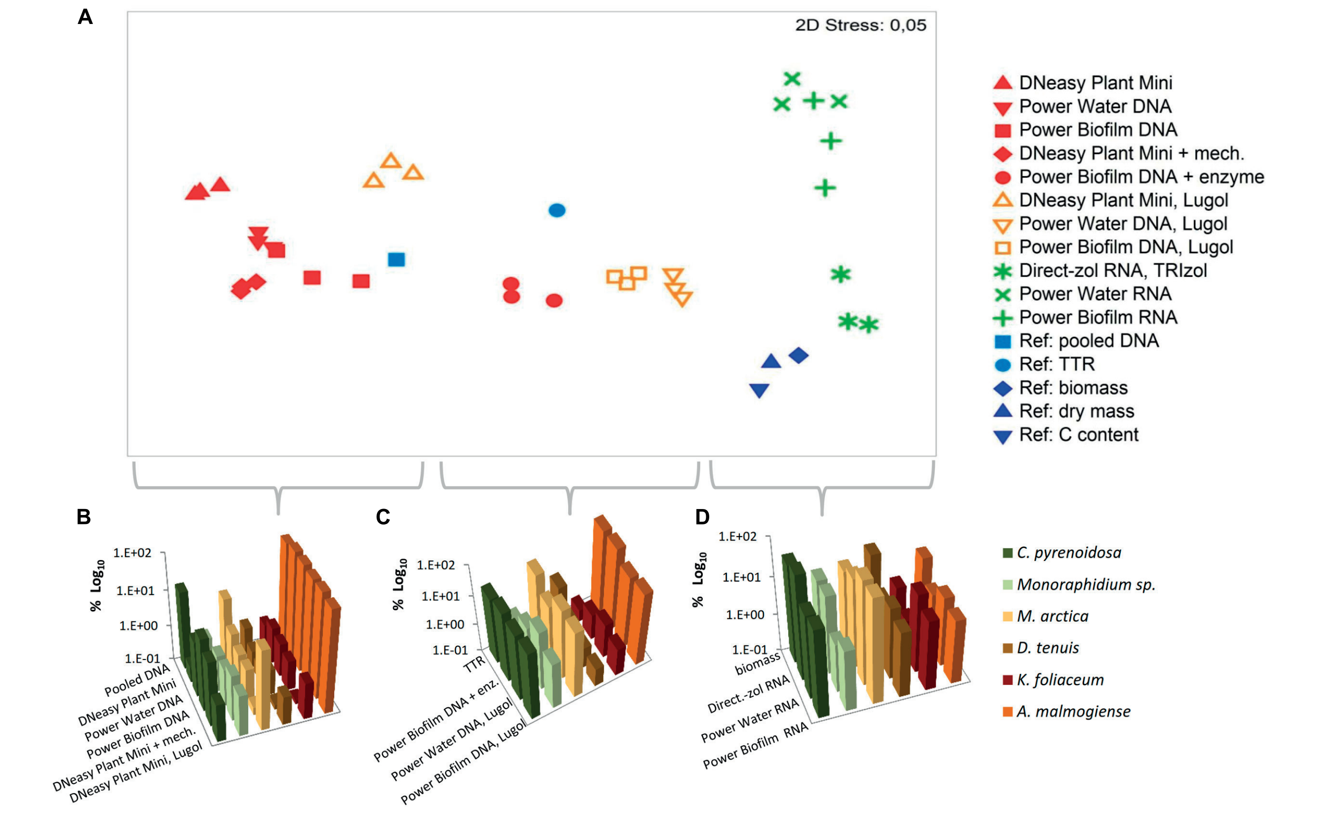


FIGURE 8 | Comparison of numerical relationships of mock species based on the NGS results and reference parameters. **(A)** Non-metric multidimensional scaling (NMDS) ordination of the data. **(B–D)** Illustration of the data in subsets separated by the NMDS ordination. TTR refers to theoretical template relationship. Note that the percentages in the y-axis are shown in logarithmic scale.

identification and quantification methods. Recently, the massive Tara Oceans voyage surveyed 210 ecosystems at global scale applying NGS methods and collecting environmental data (de Vargas et al., 2015; Pesant et al., 2015). The data of the expedition provided profound knowledge on eukaryotic plankton species revealing that their diversity was wider than earlier expected. Using NGS methods, indeed, it is possible to detect rare taxa when other identification techniques might miss these (Yu et al., 2015). Wang et al. (2014) have defined two critical genetic factors, which affect the results of molecular- and OTU-based characterization studies. At first, genetic polymorphisms of eukaryotic microscopic organisms are still unknown, which makes a point of defining OTUs at an optimal dissimilarity level. Another factor is that while bacterial genomes have only from one to several 16S rRNA gene copies, eukaryotic genomes may have thousands of 18S rRNA genes. Proper interpreting of rRNA gene-based abundances has crucial role in molecular characterization of protists, whose rRNA gene copy numbers can vary from a few in small species to 100s of 1000s in large species like dinoflagellates and ciliates (Fu and Gong, 2017), and actually in this study the variation spanned from 2 to 33 000 rRNA operons per cell. When interpreting NGS results of environmental samples, species with small nucleus and low gene copy numbers may be hidden in cases when a sample is rich with high gene copy number species, as here by *A. malmogiense*. Our study showed that small nucleated diatoms, even if having a high total biomass when compared to other species, displayed only minor occurrence in the final NGS results. This was due to lower rRNA gene copy numbers per DNA, as differences within the primer match was excluded by determining the gene copy numbers using two qPCR primer pairs targeting independent conservative areas of the ribosomal RNA gene.

The DNA yields from the mock community strains were on average 0.2% of the wet biomass when isolated using the Power Biofilm DNA isolation kit. When testing the other DNA extraction kits and their modifications on the mock pool, highest overall DNA yield was gained with Power Biofilm DNA isolation kit, and the yield was not improved by additional enzymatic lysis steps. Highest RNA yields (and also best small RNA yields) were obtained by the Direct-zol extraction system.

The overall yield of nucleic acids does not necessary indicate the best extraction method, if the quantitative diversity of species is subject of the study. Here, for example, some methods gave similar yields, but NGS results appeared different. This was especially clear between samples preserved at −80°C and in Lugol. Lugol preservation tended to decrease the dominance of the dinoflagellate *A. malmogiense* that was most efficiently extracted from deep-frozen samples, actually much better than from fresh samples even with the Power Biofilm kit. When adding an additional enzymatic lysis step for the Power Biofilm kit, DNA yields did not change, but the proportion of *A. malmogiense* decreased, giving way to diatom and green algae sequences.

To find out how should the most realistic mock community NGS data look out and evaluate studied nucleic acid extraction methods the results were portrayed against biomass and dry mass values, as well as qPCR-based analyses of the TTR and pooled DNA sample. Our results showed some correlation between theoretical prediction (TTR) and with Lugol preserved Power Water and Biofilm DNA isolation values or with Power Biofilm DNA isolation with additional enzymatic cell lysis method values. Wet biomass was considered to be the most natural indicator to which sequencing results could be compared, since it is the method that has been in use in traditional limnology and oceanology. Our results demonstrate that while DNA-based methods were mostly affected by the rRNA copy number variation, results based on random primed cDNA as a starting material yielded the most realistic measures of the biomass values.

As described in reviews published by Robasky et al. (2014) and van Dijk et al. (2014), NGS data can be easily biased in many phases over the procedures during the library preparation. Also our results demonstrated that many different factors influence the NGS results, but furthermore, data trimming can cause additional bias when certain sequences are discriminated. NGS data quality trimming must be customized to suit the study and sequencing platform. For example, gentle trimming of NGS data of low PHRED scores have been suggested (MacManes, 2014) and for RNA-seq trimming, justification of caution exist (Williams et al., 2016). Evaluation of bioinformatics steps can also be done using *in silico* sequence libraries, although they do not replace the real sequencing data (Hardwick et al., 2017). We suggest that a control sample of few known species, relevant to the study in question, should be included into the NGS, and the effects of the data trimming should be followed through the pipeline. In this study it was convenient to evaluate trimming effects using separately barcoded sequences. In the Ion Torrent sequencing, it is possible that secondary structures (loops) of the certain rRNA genes structures may have delayed the sequencing signal, thus temporarily decreasing the quality value, which later increased to the normal level. Whether this can be possible when sequencing with other platforms is not known by us.

Although many challenges still exist in molecular level identification of phytoplankton species like sequence data analyzing issues, primer biases and imperfection of DNA and RNA extraction methods, the advantages of molecular methods go beneath the surface. One remarkable benefit is that the data obtained from studies can be utilized in the long term when tools and capacity for bioinformatics data continue to develop. Considering phytoplankton molecular identification tools, one obstacle is the lack, limitedness or inaccuracy of reference libraries. The data collected beforehand can be reanalyzed and completed when libraries have been extended.

Deeper characterization of community structure of phytoplankton has advanced through new NGS techniques and tools for data interpretation are continuously improving (Johnson and Martiny, 2015) but evaluation of methods is still needed. Even though rare species may be revealed from the data, quantitative assessment of data may turn out excessively demanding. Microscopic observation, flow cytometry studies and other tools of identification and quantification of phytoplankton cells have proved their utility values in the past and are important tools to validate NGS results. This study showed that RNA-based data better correlated with biomass parameters and, as it indicates active protein synthesizing capacity of the community, avoids the problems of possible relic DNA (Carini et al., 2016).

CONCLUSION

We present thus far one of the most complete comparison of microscopic and molecular analysis of phytoplankton communities with real biomass and carbon values, especially focusing on the effects on the selection of nucleic acid extraction methods. This study demonstrated that DNA-based phytoplankton analysis was principally affected by the huge rRNA gene copy number variation among phytoplankton species, which makes quantitative NGS studies of phytoplankton very challenging to interpret. In the light of this study, it is possible and even favorable to preserve phytoplankton samples deep-frozen before extraction procedures. Preserving the samples in acidic Lugol's solution resulted in equal DNA yields and PCR performance, but affected community profiles. When comparing traditional biomass values and sequencing results, none of the DNA-based extraction methods resulted in coherent data, but RNA-based methods yielded more realistic relationship of organisms. Finally, the study demonstrated that bioinformatics can form a post-laboratory bias, if sequences are cut with narrow sliding-window algorithms, since the data quality has sequence-specific variation in the sites of secondary structures.

AUTHOR CONTRIBUTIONS

Study was designed by AM, PS, and MT. PS counted phytoplankton samples and carried out dry mass and carbon content analysis. AM performed microscope imaging, molecular biology experiments, NGS and data analysis. AMi assisted with bioinformatics and performed statistical analysis. Manuscript was prepared by AM. All authors contributed to the discussion of the results, reviewed and edited the manuscript. AK offered important intellectual knowledge about phytoplankton cells.

FUNDING

The study was supported by the funding from the Academy of Finland (grant 260797) and European Research Council (ERC) under the European Union's Seventh Framework Programme (FP/2007-2013, grant agreement No. 615146) both awarded to MT, and Academy of Finland grant 251564 supported the contribution of AK.

ACKNOWLEDGMENTS

We thank Elina Virtanen for her dedicated technical assistance with PGM. We also thank Veli-Mikko Puupponen for helping with bioinformatics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01848/full#supplementary-material

REFERENCES

Abad, D., Albaina, A., and Aguirre, M. (2016). Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. *Mar. Biol.* 163:149. doi: 10.1007/s00227-016-2920-0

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., and Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLOS ONE* 4:e6372. doi: 10.1371/journal.pone.0006372

Bradley, I. M., Pinto, A. J., and Guest, J. S. (2016). Design and evaluation of illumina miseq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl. Environ. Microbiol.* 82, 5878–5891. doi: 10.1128/AEM.01630-16

Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S., and Fierer, N. (2016). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* 2:16242. doi: 10.1038/nmicrobiol.2016.242

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348:1261605. doi: 10.1126/science.1261605

Decelle, J., Romac, S., Stern, R. F., Bendif el, M., Zingone, A., Audic, S., et al. (2015). PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol. Ecol. Resour.* 15, 1435–1445. doi: 10.1111/1755-0998.12401

Díez, B., Pedrós-Alió, C., Marsh, T. L., and Massana, R. (2001). Application of denaturing gradient gel electrophoresis (DGGE) to study the diversity of marine picoeukaryotic assemblages and comparison of DGGE with other molecular techniques. *Appl. Environ. Microbiol.* 67, 2942–2951. doi: 10.1128/AEM.67.7.2942-2951.2001

Eiler, A., Drakare, S., Bertilsson, S., Pernthaler, J., Peura, S., Rofner, C., et al. (2013). Unveiling distribution patterns of freshwater phytoplankton by a next generation sequencing based approach. *PLOS ONE* 8:e53516. doi: 10.1371/journal.pone.0053516

Eland, L. E., Davenport, R., and Mota, C. R. (2012). Evaluation of DNA extraction methods for freshwater eukaryotic microalgae. *Water Res.* 46, 5355–5364. doi: 10.1016/j.watres.2012.07.023

Figueroa, R. I., Bravo, I., Fraga, S., Garcés, E., and Llaveria, G. (2009). The life history and cell cycle of *Kryptoperidinium foliaceum*, a dinoflagellate with two eukaryotic nuclei. *Protist* 160, 285–300. doi: 10.1016/j.protis.2008.12.003

Fu, R., and Gong, J. (2017). Single cell analysis linking ribosomal (r)DNA and rRNA copy numbers to cell size and growth rate provides insights into molecular protistan ecology. *J. Eukaryot. Microbiol.* doi: 10.1111/jeu.12425 [Epub ahead of print].

Godhe, A., Asplund, M. E., Härnström, K., Saravanan, V., Tyagi, A., and Karunasagar, I. (2008). Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* 74, 7174–7182. doi: 10.1128/AEM.01298-08

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E. M., and Troedsson, C. (2014). Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLOS ONE* 9:e87624. doi: 10.1371/journal.pone.0087624

Hällfors, G., and Hällfors, S. (1992). The Tvärminne collection of algal cultures. *Tvärminne Stud.* 5, 15–19.

Hardwick, S. A., Deveson, I. W., and Mercer, T. R. (2017). Reference standards for next-generation sequencing. *Nat. Rev. Genet.* 18, 473–484. doi: 10.1038/nrg.2017.44

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3, 1365–1373. doi: 10.1038/ismej.2009.89

Hugerth, L. W., Muller, E. E., Hu, Y. O., Lebrun, L. A., Roume, H., Lundin, D., et al. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLOS ONE* 9:e95567. doi: 10.1371/journal.pone.0095567

Johnson, Z. I., and Martiny, A. C. (2015). Techniques for quantifying phytoplankton biodiversity. *Annu. Rev. Mar. Sci.* 7, 299–324. doi: 10.1146/annurev-marine-010814-015902

Koid, A., Nelson, W. C., Mraz, A., and Heidelberg, K. B. (2012). Comparative analysis of eukaryotic marine microbial assemblages from 18S rRNA gene and gene transcript clone libraries by using different methods of extraction. *Appl. Environ. Microbiol.* 78, 3958–3965. doi: 10.1128/AEM.06941-11

Lie, A. A., Liu, Z., Hu, S. K., Jones, A. C., Kim, D. Y., Countway, P. D., et al. (2014). Investigating microbial eukaryotic diversity from a global census: insights from a comparison of pyrotag and full-length sequences of 18S rRNA genes. *Appl. Environ. Microbiol.* 80, 4363–4373. doi: 10.1128/AEM.00057-14

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Front. Genet.* 5:13. doi: 10.3389/fgene.2014.00013

Mäki, A., Rissanen, A. J., and Tiirola, M. (2016). A practical method for barcoding and size-trimming PCR templates for amplicon sequencing. *Biotechniques* 60, 88–90. doi: 10.2144/000114380

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le, Bescot N, Gorsky, G., et al. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* 2:150023. doi: 10.1038/sdata.2015.23

Prokopowich, C. D., Gregory, T. R., and Crease, T. J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 46, 48–50. doi: 10.1139/g02-103

Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62. doi: 10.1038/nrg3655

Rosic, N. N., and Hoegh-Guldberg, O. (2010). A method for extracting a high-quality RNA from *Symbiodinium* sp. *J. Appl. Phycol.* 22, 139–146. doi: 10.1007/s10811-009-9433-x

Salmi, P., and Salonen, K. (2016). Regular build-up of the spring phytoplankton maximum before ice-break in a boreal lake. *Limnol. Oceanogr.* 61, 240–253. doi: 10.1002/lno.10214

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467. doi: 10.1073/pnas.74.12.5463

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Simonelli, P., Troedsson, C., and Nejstgaard, J. C. (2009). Evaluation of DNA extraction and handling procedures for PCR-based copepod feeding studies. *J. Plankton Res.* 31, 1465–1474. doi: 10.1093/plankt/fbp087

Stach, J. E., Bathe, S., Clapp, J. P., and Burns, R. G. (2001). PCR-SSCP comparison of 16S rDNA sequence diversity in soil DNA obtained using different isolation and purification methods. *FEMS Microbiol. Ecol.* 36, 139–151. doi: 10.1111/j.1574-6941.2001.tb00834.x

van Dijk, E. L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* 322, 12–20. doi: 10.1016/j.yexcr.2014.01.008

Wang, Y., Tian, R. M., Gao, Z. M., Bougouffa, S., and Qian, P. Y. (2014). Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PLOS ONE* 9:e90053. doi: 10.1371/journal.pone.0090053

Wiame, I., Remy, S., Swennen, R., and Sági, L. (2000). Irreversible heat inactivation of DNase I without RNA degradation. *Biotechniques.* 29, 252–256.

Williams, C. R., Baccarella, A., Parrish, J. Z., and Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* 17:103. doi: 10.1186/s12859-016-0956-2

Yu, L., Zhang, W., Liu, L., and Yang, J. (2015). Determining microeukaryotic plankton community around Xiamen Island, Southeast China, Using Illumina MiSeq and PCR-DGGE techniques. *PLOS ONE* 10:e0127721. doi: 10.1371/journal.pone.0127721

Yuan, J., Li, M., and Lin, S. (2015). An improved DNA extraction method for efficient and quantitative recovery of phytoplankton diversity in natural assemblages. *PLOS ONE* 10:e0133060. doi: 10.1371/journal.pone.0133060

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Material

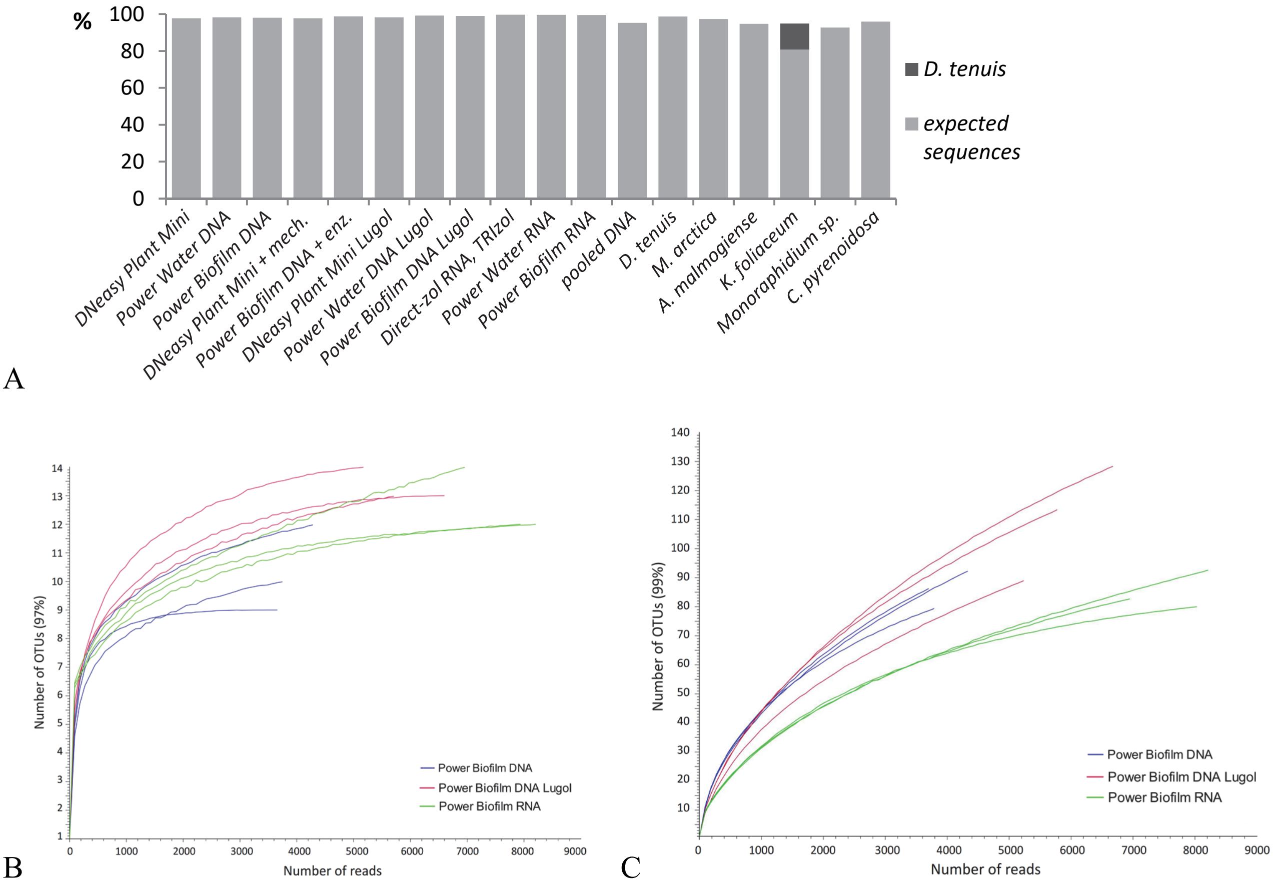
Sample preservation, DNA or RNA extraction and data analysis for high-throughput phytoplankton community sequencing

Anita Mäki[1*,] Pauliina Salmi[1], Anu Mikkonen[1], Anke Kremp[2] and Marja Tiirola[1]

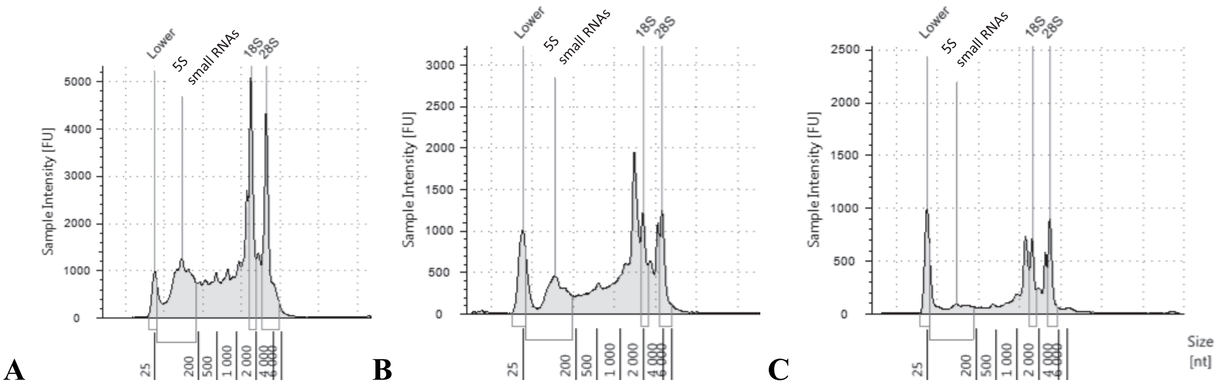
[1]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, [2]Marine Research Centre, Finnish Environment Institute, Helsinki, Finland

*** Correspondence:**
Anita Mäki
anita.maki@jyu.fi

Supplementary Figures



Supplementary Figure 1. Percentage of the sequences in the six main OTUs (similarity 0.97), that represent the six target phytoplankton strains and examples of rarefaction curves. (A) The average covery of these six main OTUs was 98 % (93–100 %) of sequences. Among the separately tested species was an exception, *K. foliaceum*, which is a binucleate cell having additional nucleus and 18S rRNA gene of diatom origin. (B) Example rarefaction curves of NGS results using $OTU_{0.97}$ clustering and (C) using $OTU_{0.99}$ clustering.



Supplementary Figure 2. Size distribution histograms of RNA extractions using (A) Direct-zol RNA extraction, (B) Power Water RNA isolation, and (C) Power Biofilm RNA isolation. The extracts were analyzed using TapeStation 2200 and the High Sensitivity RNA ScreenTape. Lower marker designates the 25 nt peak size.

Supplementary Tables

Supplementary Table 1. Algal strains of which the mock community pool was comprised for DNA and RNA isolation.

Strain ID	Taxon	Location of isolation	Time of isolation	Isolated by
SHTV-1	*Apocalathium malmogiense*	Tvärminne/Storfjärden	2002	Anke Kremp
KFF-1001	*Kryptoperidinium foliaceum*	Åland/Föglö	2010	Päivi Hakanen
DTTV-1401	*Diatoma tenuis*	Tvärminne/Storfjärden	2014	Päivi Hakanen
MATV-1402	*Melosira arctica*	Tvärminne/Längden	2014	Johanna Oja
TV70 *)	*Monoraphidium* sp.	*)	*)	*)
TV216 *)	*Chlorella pyrenoidosa*	*)	*)	*)

*) (Hällfors G, and S Hällfors, 1992)

Supplementary Table 2. Light microscopy data of the phytoplankton cell cultures before pooling the cells for nucleic acids isolation.

	Diatoma tenuis DTTV-1401	*Melosira arctica* MATV-1402	*Apocalathium malmogiense* SHTV-1	*Kryptoperidinium foliaceum* KFF-1001	*Monoraphidium* sp. TV 70	*Chlorella pyrenoidosa* TV216
medium	6 psu f/2 +Si	6 psu f/2 +Si	6 psu f/2 +Si	6 psu f/2 +Si	6 psu f/2 +Si	6 psu f/2 +Si
growth temperature (°C)	4	4	4	16	16	16
inoculation date	3.11.2015	3.11.2015	3.11.2015	3.11.2015	3.11.2015	3.11.2015
Preparation of samples for microscopy	1 mL culture + 2.5 mL PBS + 7.5 µL Lugol	1 mL culture + 2.5 mL PBS + 7.5 µL Lugol	1 mL culture + 2.5 mL PBS + 7.5 µL Lugol	1 mL culture + 2.5 mL PBS + 7.5 µL Lugol	250 µL culture + 3.25 mL PBS + 7.5 µL Lugol	250 µL culture + 3.25 mL PBS + 7.5 µL Lugol
Abundance in culture (cells L^{-1})	115197805	65894078	15821197	2185918	6264498945	3278874126
Abundance in culture (cells m^{-3})	1.15198E+11	65894078000	15821197000	2185918000	6.2645E+12	3.27887E+12
Abundance in culture (cells mL^{-1})	1.15E+05	6.59E+04	1.58E+04	2.19E+03	6.26E+06	3.28E+06
Biomass (mg m^{-3})	173805.4575	92110.6282	72934.4289	17145.8941	76877.9311	261099.6395
Biomass (µg mL^{-1})	173.8054575	92.1106282	72.9344289	17.1458941	76.8779311	261.0996395
CFL95% for biomass	29	29	29	34	4	18
Parallel fields counted	16	20	12	50	10	10
Average cell mass (mg)	1.51E-06	1.40E-06	4.61E-06	7.84E-06	1.23E-08	7.96E-08
Average cell volume (µL)	1.51E-06	1.40E-06	4.61E-06	7.84E-06	1.23E-08	7.96E-08
Average cell volume (1 fL = 1 1 $µm^3$)	1509	1398	4610	7844	12	80
Cell shape	cylinder	circular cylinder	flattened ellipsoid	flattened ellipsoid	double cone	sphere
Average cell lenght (µm)	60.5	16.3	25.3675	29.87	7.5	-
Lenght range (µm)	57.5-75	12.5-37.5	20.0-30.0	15.0-37.5	-	-
Average cell width (µm)	5.0	10.3	19.9225	23.93	2.5	5.23125
Width range (µm)	-	10.0-15.0	15.0-25.0	12.5-30.0	-	5.0-7.5
Cell depth/height (µm)	5.0	-	17.1375	18.4938	-	-
depth range (µm)	-	-	12.5-20.0	10.0-25.0	-	-

Supplementary Table 3. The match of the Euk1A F / Euk516 R primer pair for mock community taxons tested using the TestPrime tool against the non-redundant version of the SILVA SSU Ref database allowing one mismatch occurrence.

taxonomy	coverage	specificity	accessions	eligible	match	mis-match	no-data
Eukaryota;Archaeplastida;Chloroplastida;Chlorophyta; Chlorophyceae;Sphaeropleales;Monoraphidium;	100	92.5	15	5	5	0	10
Eukaryota;Archaeplastida;Chloroplastida;Chlorophyta; Trebouxiophyceae;Chlorellales;Chlorella;	88.9	92.5	16	9	8	1	7
Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae; Peridiniphycidae;Peridiniales;Kryptoperidinium;	100	92.5	2	2	2	0	0
Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae; Peridiniphycidae;Thoracosphaeraceae;*)Scrippsiella;	88.9	92.5	90	9	8	1	81
Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea; Bacillariophytina;Bacillariophyceae;Diatoma;	100	92.5	10	6	6	0	4
Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea; Coscinodiscophytina;Melosirids;Melosira;	66.7	92.5	6	3	2	1	3

*) *Scrippsiella hangoei* is an earlier synonym of *A. malmogiense,* see Craveiro, S. C., Daugbjerg, N., Moestrup, Ø., & Calado, A. J. (2017). Studies on *Peridinium aciculiferum* and *Peridinium malmogiense* (=*Scrippsiella hangoei*): comparison with *Chimonodinium lomnickii* and description of *Apocalathium gen. nov*. (Dinophyceae). *Phycologia*, *56*(1), 21-35. DOI: 10.2216/16-20.1.

Supplementary Table 4. Final trimming and OTU picking parameters for the NGS data using CLC Genomics Workbench 9.5.1 software. For the comparative testing of nucleic acid extraction methods a total of 362,728 sequences were processed, of which 136,778 sequences were removed during the trimming.

Trimming and OTU picking parameters	
Trim adapter list	M13_Euk1A
Quality trim	Yes
Quality limit	0.05
Minimum number of nucleotides in reads	150
OTU picking	De novo OTU clustering
Similarity percentage	97 %
Minimum occurrences	10 (2 in “posit. control”)
Fuzzy match duplicates	No
Find best match	Yes
Chimera crossover cost	3
Kmer size	6

Supplementary Table 5. Tested pipeline for the model data analysis using Mothur v.1.36.1 bioinformatics platform. Applying quality criteria for trimming, such as minimum length of 180 bases and minimum quality average over a window of 20, resulted in biased proportional sequence abundances of model data sample and was not used for final data analysis. The commands used and the number of usage of CPUs are presented according the order they were assigned. The dataset was named as "phyto" and "summary.seqs" command was given frequently to follow the processing.

A brief comment of the function of trimming	Command in Mothur software
Extract sequences reads from a .sff file	sffinfo(sff=phyto.sff)
Preprocess features needed to screen and sort sequences	trim.seqs(fasta=phyto.fasta, oligos=phyto3.oligos, qfile=phyto.qual, pdiffs=2, bdiffs=1, maxambig=0, maxhomop=8, qwindowaverage=20, qwindowsize=10, minlength=180, processors=16)
Unique (re-replicate) identical sequences to save time in processing	unique.seqs(fasta=phyto.trim.fasta)
Align a fasta-formatted sequences against Silva database	align.seqs(fasta=phyto.trim.unique.fasta, reference=silva.nr_v123.align, flip=T, processors=8
Summarize the quality of sequences (e.g. check the start and end points for the next command)	summary.seqs(fasta=phyto.trim.unique.align, name=phyto.trim.names)
Fulfill or cull defined criteria	screen.seqs(fasta=phyto.trim.unique.align, name=phyto.trim.names, group=phyto.groups, start=1046, optimize=end, criteria=95, processors=8)
Remove columns from alignments based on a defined criteria	filter.seqs(fasta=phyto.trim.unique.good.align, vertical=T, trump=., processors=8)
Unique identical sequences	unique.seqs(fasta=phyto.trim.unique.good.filter.fasta, name=phyto.trim.good.names)
Remove sequences for sequencing error mitigation	pre.cluster(fasta=phyto.trim.unique.good.filter.unique.fasta, name=phyto.trim.unique.good.filter.names, group=phyto.good.groups, diffs=2)
Search for chimeric sequences	chimera.uchime(fasta=phyto.trim.unique.good.filter.unique.precluster.fasta, name=phyto.trim.unique.good.filter.unique.precluster.names, group=phyto.good.groups, processors=16)
Remove chimeric sequences	remove.seqs(accnos=phyto.trim.unique.good.filter.unique.precluster.denovo.uchime.accnos, fasta=phyto.trim.unique.good.filter.unique.precluster.fasta, name=phyto.trim.unique.good.filter.unique.precluster.names, group=phyto.good.groups, dups=T)
Classify sequences taxonomically against database using defined criteria	classify.seqs(fasta=phyto.trim.unique.good.filter.unique.precluster.pick.fasta, name=phyto.trim.unique.good.filter.unique.precluster.pick.names, group=phyto.good.pick.groups, template=silva.nr_v123.align, taxonomy=silva.nr_v123.tax, cutoff=80, iters=1000, processors=16)
Generate a new file that contains sequences of defined taxon (excluding removed)	remove.lineage(fasta=phyto.trim.unique.good.filter.unique.precluster.pick.fasta, name=phyto.trim.unique.good.filter.unique.precluster.pick.names, group=phyto.good.pick.groups, taxonomy=phyto.trim.unique.good.filter.unique.precluster.pick.nr_v123.wang.taxonomy, taxon=unknown)
Rename all the filenames into a simplified format (example: only fasta-file)	system(cp phyto.trim.unique.good.filter.unique.precluster.pick.pick.fasta phyto.final.fasta)
Calculate pairwise distances between aligned DNA sequences so OTU clustering can be done accordingly	dist.seqs(fasta=phyto.final.fasta, cutoff=0.15, processors=16)

Assign sequences to OTUs (default clustering algorithm: average neighbor)	cluster(column=phyto.final.dist, name=phyto.final.names)
Create a OTU-file, OTUs occurrence per barcode	make.shared(list=phyto.final.an.list, group=phyto.final.groups, label=0.03)
Classification of OTUs	classify.otu(list=phyto.final.an.list, name=phyto.final.names, taxonomy=phyto.final.taxonomy, label=0.03)
Generates a fasta-file containing only a representative sequence for each OTU	get.oturep(column=phyto.final.dist, list=phyto.final.an.list, name=phyto.final.names, fasta=phyto.final.fasta, method=abundance, weighted=true)

Supplementary Table 6. Clustering of the NGS reads to target phytoplankton strains ($OTU_{0.97}$ level) of the mock community pool and separate cell cultures. *K. foliaceum*, which is known to be a binucleate containing nucleus of diatom origin and accordingly had two dominant sequences of 18S rRNA gene.

Extraction method or strains	All reads	Target sequences	Target sequences % of all reads
DNeasy Plant Mini	10655	10421	97.8
Power Water DNA	10112	9939	98.3
Power Biofilm DNA	11891	11654	98.0
DNeasy Plant Mini + mech.	11344	11090	97.8
Power Biofilm DNA + enz.	14764	14589	98.8
DNeasy Plant Mini Lugol	11676	11479	98.3
Power Water DNA Lugol	16527	16395	99.2
Power Biofilm DNA Lugol	17652	17468	99.0
Direct-zol RNA, TRIzol	23112	23016	99.6
Power Water RNA	20921	20828	99.6
Power Biofilm RNA	23192	23074	99.5
pooled DNA	21657	20627	95.2
D. tenuis	26595	26250	98.7
M. arctica	27189	26455	97.3
A. malmogiense	17321	16403	94.7
K. foliaceum *(D. tenuis)*	19345	15652 (2683)	80.9 (13.9)
Monoraphidium sp.	16834	15604	92.7
C. pyrenoidosa	19609	18800	95.9

III

DIRECTIONAL HIGH-THROUGHPUT SEQUENCING OF RNAS WITHOUT GENE-SPECIFIC PRIMERS

by

Mäki, A. & Tiirola, M. 2018

BioTechniques 60: 219–223

https://doi.org/10.2144/btn-2018-0082

Directional high-throughput sequencing of RNAs without gene-specific primers

Anita Mäki*,[1] & Marja Tiirola[1]

[1]*Department of Biological and Environmental Science, Nanoscience Center, University of Jyväskylä, PO Box 35, FI-40014 Finland*

BioTechniques 65: 219-223 (October 2018) 10.2144/btn-2018-0082

Keywords: algae • ligation • next-generation sequencing • phytoplankton • primer bias • random priming • ribosomal RNA

Ribosomal RNA analysis is a useful tool for characterization of microbial communities. However, the lack of broad-range primers has hampered the simultaneous analysis of eukaryotic and prokaryotic members by amplicon sequencing. We present a complete workflow for directional, primer-independent sequencing of size-selected small subunit ribosomal RNA fragments. The library preparation protocol includes gel extraction of the target RNA, ligation of an RNA oligo to the 5′-end of the target, and cDNA synthesis with a tailed random-hexamer primer and further barcoding. The sequencing results of a phytoplankton mock community showed a highly similar profile to the biomass indicators. This method has universal potential for microbiome studies, and is compatible for the 5′-end sequencing of other RNA types with minimum library preparation costs.

The phylogenic characterization of microbial communities rests mainly on the small subunit ribosomal RNA (SSU rRNA) genes, 16S rRNA in prokaryotes and its counterpart 18S rRNA in eukaryotes; however, the lack of good universal primer sequences and incomplete reference databases complicates the analysis of eukaryotic species. Our previous phytoplankton mock community study showed that RNA-based community profiling correlated better with the biomass measures than DNA-based sequencing, in which the variation in the rRNA gene copy numbers per algae cell can be over 100-fold [1]. Moreover, profiling microbial communities using 16S/18S rRNA instead of rRNA genes especially brings out living organisms with active protein synthesis and is free of dissolved or relic DNA [2].

To amplify RNA templates without the natural poly(A)-tail, the use of gene-specific primers can be avoided by 3′- and 5′-end poly(A) tailing of RNA or by ligation of adaptor sequences as reviewed by van Dijk *et al.* [3]. When primer-independent methods have been applied, our knowledge on the rRNA molecules and the diversity of life have been expanded [4–6]. Commercial kits have been fabricated for nontargeted amplification of RNA. In small RNA library construction of NEBNext products of New England Biolabs (MA, USA), the workflow is based on a ligation of 3′ and 5′ adaptors, whereas in mRNA library construction, cDNA synthesis is based on RNA fragmentation and random priming. Diagenode's (Seraing, Belgium) CATS RNA-seq kit is based on poly(A) tailing and the ligation-free 'Capture and Amplification by Tailing and Switching' method. In the common Illumina (CA, USA) library preparation workflow, random fragmentation of the cDNA sample is followed by 5′ and 3′ adapter ligation. Machida and Lin [7] describe in their article other commercial library preparation kits, such as SMART cDNA library construction kit of Clontech and Exact START Eukaryotic mRNA 5′- & 3′- RACE kit of Epicenter (WI, USA), for which the procedures are easy to repeat. However, the kits available are either designed to cover the entire length of RNAs, and therefore not efficient for rRNA sequencing where the full alignment of the fragments would be beneficial, or are rather expensive to use for environmental studies. When we searched a suitable method to perform rRNA 5′-end sequencing of phytoplankton samples, amplification of poly(A)-tailed cDNA products with an oligo(dT) primer preferentially resulted in internal poly(A) priming and truncated amplification fragments, as was warned by Nam *et al.* [8]. Therefore, we recognized a lack of a directional 5′-end sequencing method that could be readily used for preparing barcoded RNA sequencing libraries without the need for gene-specific primers or oligo(dT) priming.

Our target was to construct a library preparation workflow to study the 5′-end of the SSU rRNAs of all organisms without the bias caused by gene-specific primers or poly(A) tailing. We validated the workflow (Figure 1) by exploiting a eukaryotic cell pool consisting of six phytoplankton species: *Diatoma tenuis*, *Melosira arctica*, *Apocalathium malmogiense*, *Kryptoperidinium foliaceum*, *Monoraphidium*

METHOD SUMMARY

The primer-independent RNA library construction workflow includes ligation of the RNA oligo (M13) to the 5´-end of the gel-extracted small subunit ribosomal RNA, and subsequent reverse transcription of the template using a random hexamer primer with a sequencing adapter overhang. Following amplification of the RNA fragments with barcoded Ion Torrent primers, the amplicons are size-selected with magnetic bead purification.

sp. and *Chlorella pyrenoidosa*. The biomass, dry mass, carbon content and cell number values of each species of the pool were determined earlier [1], but symbiotic prokaryotic cells were unidentified. Here, total RNA was extracted from the mock cells (a detailed protocol and list of the reagents is available in the Supplementary Protocol) using a Direct-Zol RNA MicroPrep isolation kit (Zymo Research, CA, USA). 16S/18S rRNA fragments were cut from a precast 1% agarose E-Gel EX gel (Invitrogen, MA, USA) and purified using a Zymoclean Gel RNA Recovery Kit (Zymo Research). The universal M13-RNA forward adapter sequence (5′-UGUAAAACGACGGCCAGU-3′) was ligated to the purified 16S/18S rRNA fragments with T4 RNA ligase (Promega, WI, USA) applying the manufacturer's instructions. After purification of the ligation product, Ion Torrent sequencing adapter P1 was exploited as an overhang in the random primed (5′-CCTCTCTATGGGCAGTCGGTGATNNNNNN-3′) cDNA synthesis. Purified cDNA was amplified with a barcoded Ion Torrent sequencing adapter (IonA) with M13-sequence in the 3′-end and P1 as the reverse primer. For one-step size-selection and purification of amplicons, the dual size selection procedure of the ProNex Size-Selective Purification System (Promega) was applied, targeting the selection between 300 and 550 bp. Sequencing was performed with the Ion Torrent Personal Genome Machine (PGM) as described by Mäki *et al.* [1]. The data was analyzed using CLC Genomics Workbench 11 software (www.qiagenbioinformatics.com) using a phytoplankton reference database comprising mock community sequences [1] or Silva v128 16S reference rRNA gene database for prokaryotic analysis and using OTU 97% identity clustering level (Supplementary Figure 2).

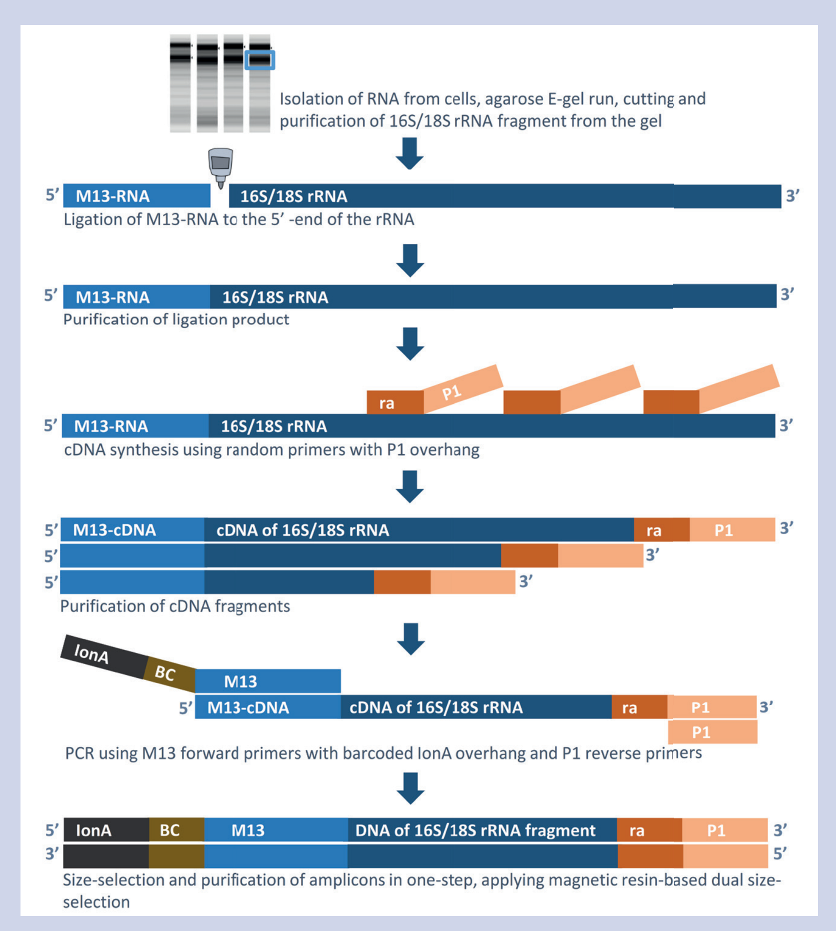

Isolation of RNA from cells, agarose E-gel run, cutting and purification of 16S/18S rRNA fragment from the gel
5′
M13-RNA
16S/18S rRNA
3′
Ligation of M13-RNA to the 5′ -end of the rRNA
5′
M13-RNA
16S/18S rRNA
3′
Purification of ligation product
ra
P1
5′
M13-RNA
16S/18S rRNA
3′
cDNA synthesis using random primers with P1 overhang
5′
M13-cDNA
cDNA of 16S/18S rRNA
ra
P1
3′
5′
3′
5′
3′
Purification of cDNA fragments
IonA
BC
M13
5′
M13-cDNA
cDNA of 16S/18S rRNA
ra
P1
3′
P1
PCR using M13 forward primers with barcoded IonA overhang and P1 reverse primers
5′
IonA
BC
M13
DNA of 16S/18S rRNA fragment
ra
P1
3′
3′
5′
Size-selection and purification of amplicons in one-step, applying magnetic resin-based dual size-selection

Figure 1. Construction of rRNA sequencing library without gene-specific primers. Ribosomal 16S/18S RNA fragments are extracted from total RNA, M13 RNA oligo is ligated to the purified product, and cDNA synthesis is primed with a random primer containing P1-adapter overhang for the Ion Torrent sequencing. Finally, barcoding and amplification of the construct is performed using the M13 and P1 sites to create the sequencing library.

Relative abundances of strains were used to calculate Bray-Curtis similarity matrix, based on which nonmetric multidimensional scaling (NMDS) was calculated with 1000 repeats in PRIMER v. 6.1.12 and PERMANOVA+ v.1.0.2 (PRIMER-E/Quest Research Limited, Albany, New Zealand).

Two steps were critical for the workflow. First, when the RNA template is ligated to the M13 RNA oligo, self-ligation of fragments may occur. To avoid this, we tested blocking the free OH-group of the original RNA with ddATP using the TdT enzyme, but this caused extra steps in the process. The other method was to block self-ligation by introducing an excess of M13-RNA adapters during the ligation reaction. When the concentration of the M13-RNA adapter in the ligation reaction was 10,000 times greater than the rRNA concentration, self-ligation of 16S/18S rRNAs was efficiently prevented (Supplementary Figure 1), and the latter procedure was selected.

The second critical step was related to random priming, which was not fully continuous throughout the sequences, but the specific priming site was affected by the primary or secondary structure of the RNA in the first strand cDNA synthesis, as already recognized in previous studies [9,10]. Nonuniform random priming caused noncontinuous length distribution in the final sequencing products. When the final sequencing data were sorted to 150–200 bp, 200–300 bp, 300–500 bp and 150–500 bp size fractions, it was found that the selection of size fraction affected the outcome, the relative abundance of species (Supplementary Figure 2). Strain-specific distribution of sequence lengths was also noticeable when random primed cDNAs of *A. malmogiense*, *Monoraphidium* sp. and *M. arctica* were amplified with the eukaryotic rRNA-specific forward primer Euk1A (5′-CTGGTTGATCCTGCCAG-3′) (Supplementary Figure 3A). Adjusting the number on degenerate nucleotides in the random hexamer did not improve the uneven priming pattern, since five, six and seven degenerate bases (N) resulted in a similar patterning, but eight N-bases generated even stronger peak formation, when the RNA of *Monoraphidium* sp. was studied (Supplementary Figure 3B).

To make the random priming more continuous, we tested addition of the organic solvent dimethyl sulfoxide (DMSO), which should improve the efficiency of cDNA synthesis by decreasing the secondary structures of RNA [11]. However, 10% DMSO did not reduce the patterning in the random priming (Supplementary Figure 3C). We also tested whether the non-continuous random priming was only related to rRNA, which has strong secondary structures, or universal. We used our random priming procedure for the protein-coding RNA of the firefly luciferase (*Fluc*) gene. The RNA was transcribed from control template DNA of the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs Inc.). The size distribution of amplified cDNA fragments of *Fluc* was not better than that of rRNA (Supplementary Figure 3D), showing that the secondary structures are not the only reason for the noncontinuous random priming pattern of the rRNA. When the tailed random primer P1–6N was manufactured using hand-mixing of the degenerate bases (Integrated DNA Technologies, Freising, Germany) to guarantee that all four bases will be equally represented, size distribution of the PCR-amplified cDNA fragments was more even and spread out to more peaks (Supplementary Figure 3E), suggesting that special handmixing of (tailed) random oligos is recommended and would solve part of the noncontinuous nature of the random priming.

In rRNA analysis every sequence matters, and species-specific differences in sequence lengths may affect the outcome of the analysis if quality trimming selects the data. The effect of the noncontinuous random priming was further avoided by maintaining a large size distribution during the data trimming and analysis. This resulted in a realistic relationship between the mock community species, when sequencing results were compared with defined biomass indicators (Figure 2A & B). In the NMDS ordination of data, our workflow closely resembled dry mass proportions of the cells of the mock pool (Figure 2B), and also prokaryotic partners of the mock community were analyzed at the same sequencing run (Figure 2C).

As even half of the microbial diversity may remain unrevealed using gene-specific primers [12], efficient primer-free techniques are needed to study environmental microbiomes. We predict that the method here described may provide an affordable alternative for commercial kits to study the real diversity of the rRNA world. However, special care has to be given for ensuring that noncontinuous random priming does not affect the results.

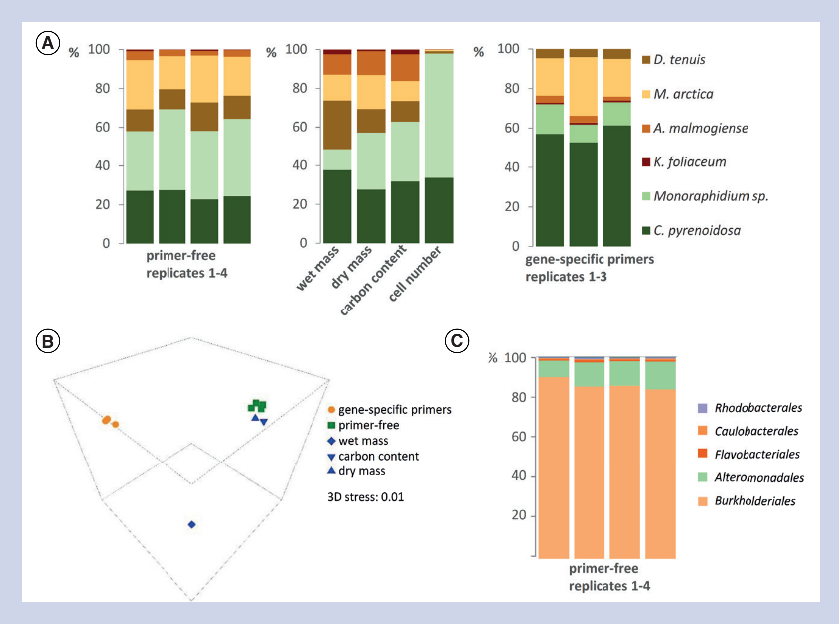


Figure 2. Comparison of the primer-free sequencing of rRNA and biomass indicators of the phytoplankton mock community. (A) Primer-free and gene-specific [1] sequencing of rRNA yielded different proportional relationships between species, where the results of the former better resembled the dry biomass and carbon content of the pool. Wet biomass [1] was determined microscopically. **(B)** Nonmetric multidimensional scaling ordination showed a close relationship between the primer-free sequencing results and biomass relationships. **(C)** Primer-free sequencing revealed simultaneously prokaryotic rRNAs of organisms that were grown in symbiosis with the algae cultures (main orders shown).

Acknowledgments

We thank Elina Virtanen for performing the Ion Torrent sequencing.

Author contributions

AM performed the molecular studies and data analysis and wrote the manuscript. MT reviewed and edited the manuscript and offered instructions and intellectual discussion during the study. AM and MT contributed to the study design.

Financial & competing interests disclosure

AM was supported by the European Research Council (ERC) through a grant awarded to MT under the European Union's Seventh Framework Programme (FP/2007-2013, grant 615146). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Supplementary data

To view the supplementary material that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2018-0082

Open access

References

1. Mäki A, Salmi P, Mikkonen A, Kremp A, Tiirola M. Sample preservation, DNA or RNA extraction and data analysis for high-throughput phytoplankton community sequencing. *Front. Microbiol.* 8, 1848 (2017).
2. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* 2(3), 16242 (2016).
3. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* 322(1), 12–20 (2014).
4. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a million high-quality, full-length microbial 16S and

18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* 36(2), 190–195 (2018).

5. Hoshino T, Inagaki F. A comparative study of microbial diversity and community structure in marine sediments using poly(A) tailing and reverse transcription-PCR. *Front. Microbiol.* 4, 160 (2013).
6. Turchinovich A, Surowy H, Serva A, Zapatka M, Lichter P, Burwinkel B. Capture and amplification by tailing and switching (CATS). An ultrasensitive ligation-independent method for generation of DNA libraries for deep sequencing from picogram amounts of DNA and RNA. *RNA Biol.* 11(7), 817–828 (2014).
7. Machida RJ, Lin Y. Four methods of preparing mRNA 5′ end libraries using the Illumina sequencing platform. *PLoS One* 9(7), e101812 (2014).
8. Nam DK, Lee S, Zhou G *et al*. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl Acad. Sci. USA* 99(9), 6152–6156 (2002).
9. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38(12), e131 (2010).
10. van Gurp TP, McIntyre LM, Verhoeven KJ. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One* 8(12), e85583 (2013).
11. Yasukawa K, Konishi A, Inouye K. Effects of organic solvents on the reverse transcription reaction catalyzed by reverse transcriptases from avian myeloblastosis virus and Moloney murine leukemia virus. *Biosci. Biotechnol. Biochem.* 74(9), 1925–1930 (2010).
12. Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3(12), 1365–1373 (2009).

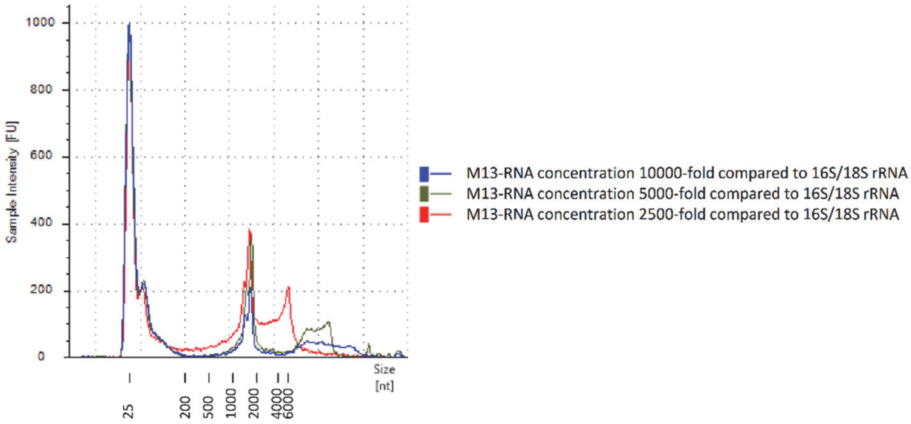
First draft submitted: 13 June 2018; Accepted for publication: 6 August 2018

Address correspondence to: Anita Mäki; Department of Biological and Environmental Science, Nanoscience Center, PO Box 35, FI-40014 University of Jyväskylä, Finland; E-mail: anita.maki@jyu.fi

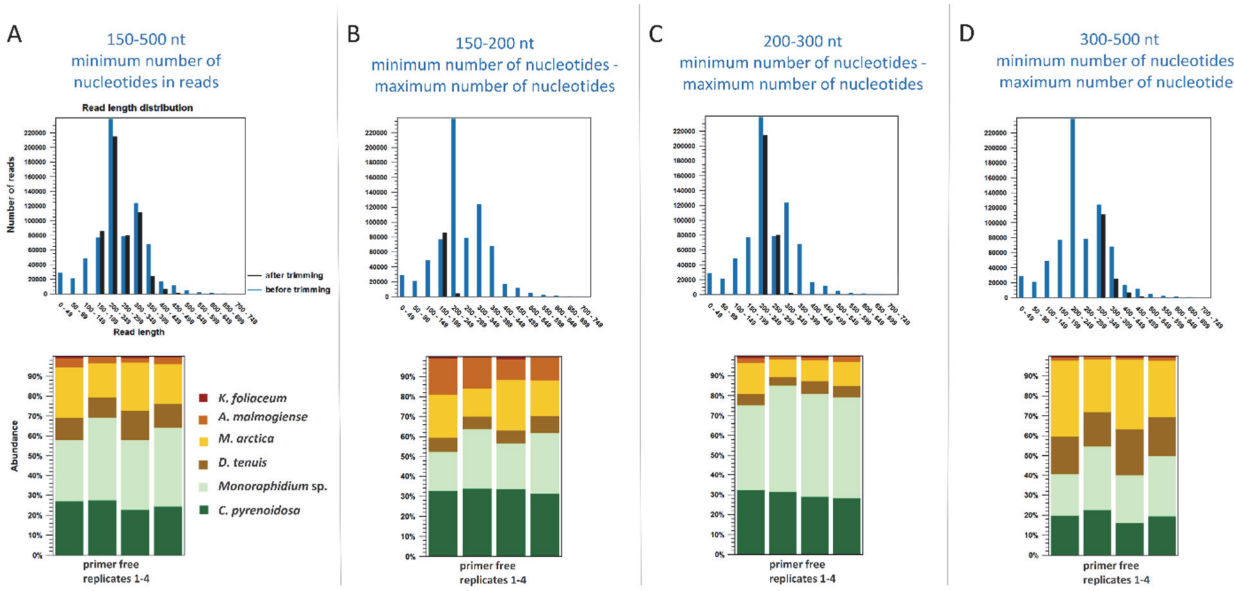
Benchmark

Supplementary material for:

Directional high-throughput sequencing of RNAs without gene-specific primers



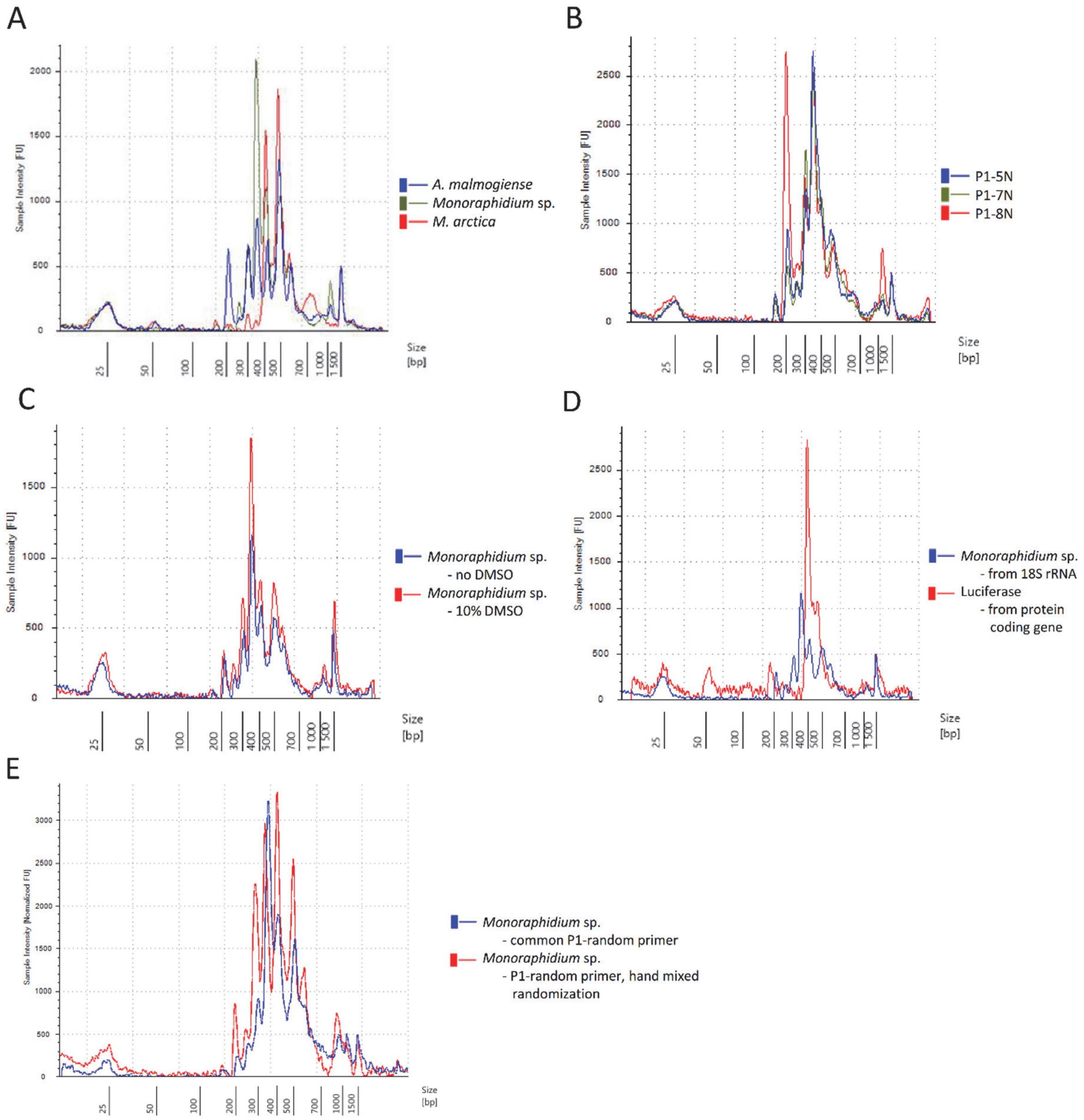
Supplementary Figure S1. M13-RNA concentration changes in ligation reaction had effect on profiles of ligation product. Tape Station HS RNA comparison analysis shows size distribution profiles of ligation products. M13-RNA concentration in ligation reaction was adjusted to 2 500-, 5 000- or 10 000-folded compared to 16S/18S rRNA concentration. Increasing M13-RNA concentration decreased self-ligation of the original 16S/18S rRNA sample. 25 nt peak refers to lower marker.



Trimming parameters: Quality limit = 0.05; Trim adapter list = M13_adapter; Remove 5' terminal nucleotides = No; Remove 3' terminal nucleotides = No

OTU picking parameters: Reference based OTU clustering; Taxonomy similarity percentage = 99; OTU similarity percentage = 97; Minimum occurrences = 1; Chimera crossover cost = 3; Kmer size = 6; Mismatch cost = 1; Minimum score = 40; Gap cost = 4; Maximum unaligned end mismatches = 5

Supplementary Figure S2. Comparison of abundances of species when analysis comprises only proportion of the sequencing reads. When random priming was used in the cDNA synthesis, abundances of species fluctuated depending on the size fraction that was picked from the reads, due to non-continuous random priming. To avoid a possible bias, it is recommended to select a wide size fraction of the sequencing reads. Data analysis was performed using CLC Genomics Workbench 11 software (www.qiagenbioinformatics.com).



Supplementary Figure S3. Analysis of the length distribution of the random primed cDNAs. All the cDNA products were amplified with Euk1A SSU rRNA gene specific forward primer and P1 reverse primer, and analysed using Tape Station HS D1000 (Agilent) agarose gel electrophoresis. (A) Amplification of cDNAs of *Apocalathium malmogiense*, *Monoraphidium* sp., and *Melosira arctica* shows non-continuous random priming in the reverse transcription for all the species. (B) Increasing or decreasing the number of degenerate (N) bases in the 3'-end of the P1 oligo did not prevent the pattern of non-continuous random priming in the reverse transcription, when *Monoraphidium* sp. cDNA was further amplified using the gene specific forward primer Euk1A and P1. (C) 10 % DMSO in the cDNA synthesis reactions of 18S rRNA of *Monoraphidium* sp. did not prevent the non-continuous random priming. (D) Comparison of random primed 18S rRNA (secondary structures) and luciferase protein coding RNA (linear) shows that random priming was non-continuous for both the rRNA and protein coding gene transcripts. (E) When using random oligos that were manufactured with hand-mixing (Integrated DNA Technologies, Germany), size distribution of the random amplified fragments was spread out more equally than using standard random oligos. 25 bp peak refers to the lower marker and 1500 bp peaks to the upper marker peak.

Protocol for:

Directional high-throughput sequencing of RNAs without gene-specific primers

Reagents:

- Direct-zol RNA MicroPrep isolation kit (Zymo Research, Irvine, CA, USA)
- High Sensitivity RNA ScreenTape Assay (Agilent Technologies, Germany)
- 1% agarose E-Gel EX gel (Invitrogen, USA)
- Zymoclean Gel RNA Recovery Kit (Zymo Research, Irvine, CA, USA)
- T4 RNA ligase (Promega, USA)
- Recombinant RNasin Ribonuclease Inhibitor (Promega, USA)
- PEG 8000, Molecular Biology Grade Polyethylene Glycol 8000 (Promega, USA)
- Agencourt RNAClean XP (Beckman Coulter)
- RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific, USA)
- Maxima SYBR Green/Fluorescein qPCR Master Mix (Thermo Scientific, USA)
- ProNex Size-Selective Purification System (Promega, USA)
- High Sensitivity D1000 ScreenTape System (Agilent Technologies, Germany)
- Qubit Fluorometer with high-sensitive dsDNA kit (Thermo Fisher Scientific, USA)
- Ion PGM Hi-Q View OT2 400 Kit (Thermo Fisher Scientific, USA)
- Ion Sphere Quality Control Kit (Thermo Fisher Scientific, USA)
- Ion PGM Hi-Q View Sequencing Kit (Thermo Fisher Scientific, USA)

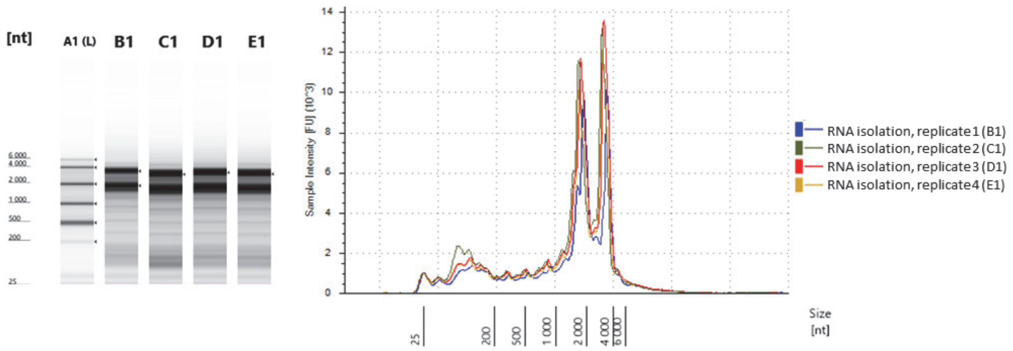
Oligos:

- 100 µM M13-RNA oligo (5'-UGUAAAACGACGGCCAGU-3')
- 100 µM P1-6N tailed random primer (5'-CCTCTCTATGGGCAGTCGGTGATNNNNNN-3')
- a set of 10 µM forward primers IonA-barcode with M13 tail (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGX^{10}TGTAAAACGACGGCCAGT-3'), where X^{10} refers to Ion Torrent barcode sequences
- 10 µM reverse primer P1 (5'-CCTCTCTATGGGCAGTCGGTGAT-3')

Protocol:

1. RNA isolation from the frozen cells following Tough-to-Lyse instructions of Direct-zol RNA Micro Prep kit and proper aseptic RNA handling techniques
    - add 300-500 µL of TRI Reagent to the frozen cells before they have thawed
    - beat beating in 2 mL tubes with 0.1 mm Glass Beads (MoBio Laboratories, USA) using Power Lyse 24 homogenizer at 3400 RPM for 40 seconds
    - proceed with Direct-zol instructions for sample purification

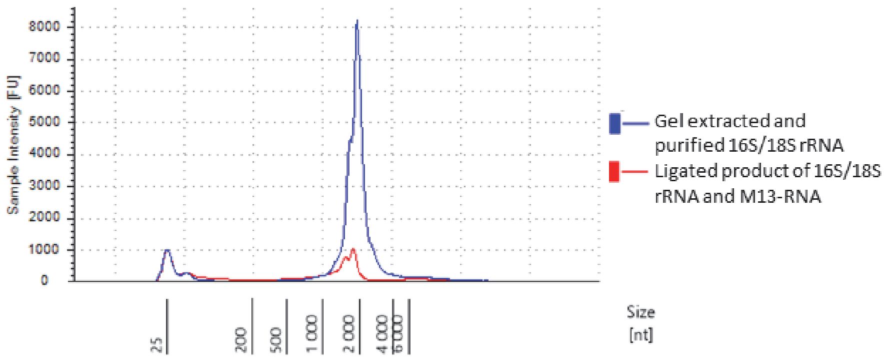
2. High Sensitivity RNA ScreenTape Assay to check RNA isolation



3. Cutting and purification of rRNA 16S/18S fragments from precast 1% agarose E-Gel EX gel
   - purification using instructions of Zymoclean Gel RNA Recovery Kit
4. High Sensitivity RNA ScreenTape Assay from purified 16S/18S rRNA fragments
   - see an example figure in point 7.
5. Ligation of M13-RNA to purified 16S/18S rRNA fragments with Promega’s T4 RNA ligase
   - prepare 40 % PEG solution in advance
   - now 20 000-fold concentration of M13-RNA compared to rRNA
     - an example of ligation ingredients in the table below: 5.9 nM rRNA sample and 100 µM M13-RNA adapter
   - incubate the reaction at 37 °C for 40 minutes in thermocycler
   - no **heat activation** after the reaction, but apply directly for purification

Components	**20.4 µL reaction**
0.02 pmol 16S/18S rRNA	3.4 µL
400 pmol M13-RNA	4 µL
T4 RNA Ligase 10X buffer	2 µL
RNasin Ribonuclease Inhibitor (40u/µL)	0.5 µL
PEG, 40 %	10 µL
T4 RNA Ligase	0.5 µL

6. Agencourt RNAClean XP cleaning of the ligation products
   - add 20 µL nuclease-free water to the sample to dilute the viscous solution before purification
   - purification in accordance with the manufacturer's instructions
   - 1.8 X sample volume: 40 µL ligation product (20 µL ligation solution + 20 µL H_2O) and 72 µL RNAClean XP solution
   - elution to 20 µL of nuclease-free water
7. High Sensitivity RNA ScreenTape Assay from ligation products

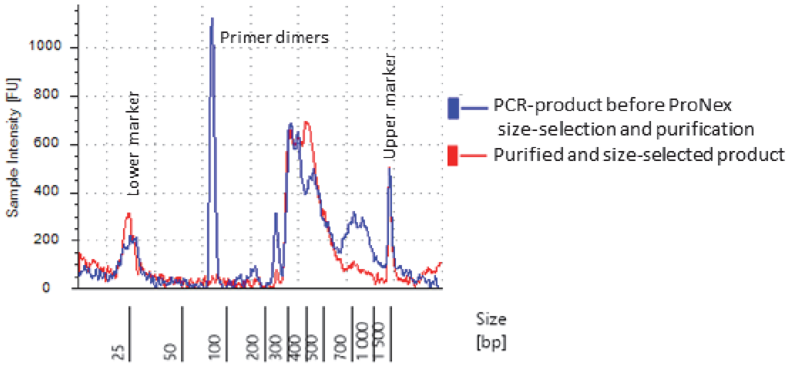


8. cDNA synthesis using RevertAid First Strand cDNA Synthesis Kit
    - use 100 pmol of P1-random primers in 20 µL reaction
    - 8 µL of purified ligation product as a template
    - incubation at 25 °C for 5 min, 45 °C for 60 min, and termination of reaction by heating at 70 °C for 5 min
9. Agencourt RNAClean XP cleaning of the cDNA products
    - purification in accordance with the manufacturer's instructions
    - 1.6 X sample volume
    - elution to 18 µL of nuclease-free water
10. Amplification of the cDNA using Maxima SYBR Green/Fluorescein qPCR Master Mix

Components	37.5 µl reaction	Final concentration
Maxima SYBR Green Master Mix	18.75 µL	
10 µM forward primer: IonA-barcode with M13-tail	1.5 µL	0.4 µM
10 µM reverse primer: P1	1.5 µL	0.4 µM
Template cDNA	6 µL	
Nuclease-free water	9.75 µL	

Cycle step	Temperature	Time	Cycles
Initial denaturation	95 °C	5 min	1
Denaturation	95 °C	15 s	
Annealing	52 °C	30 s	30
Extension	72 °C	30 s	
Final extension	72 °C	5 min	1

11. Purification, dual size-selection and concentration of PCR products using ProNex Size-Selective Purification System and dual size-selection instructions
    - to eliminate too long fragments: mix 1:1 (v/v ratio) of PCR product and ProNext (here 35 µL + 35 µL) and after placing sample on a magnetic stand, transfer the supernatant to a clean tube (too long fragments stay to the beads)
    - to eliminate too short fragment: mix additional 0.28:1 (v/v) ratio of ProNex (here 9.8 µL) into the supernatant and after placing sample on a magnetic stand short fragments are in the supernatant and desired fragment are bound to the resin
    - continue following the washing and elution steps
    - now elution to 18 µL of elution buffer to concentrate the sample
12. High Sensitivity D1000 ScreenTape System Assay from purified products



13. Concentration measurement of purified products, pooling of samples (and purification of the pooled sample if needed), and final concentration measurement of the pool for OT2 emulsion PCR of Ion Torrent sequencing
    - concentration measurement of each sample using Tape Station system or Qubit Fluorometer and pooling equal amounts of DNA (now 20 ng)
14. OT2 emulsion PCR (Ion sphere quality control included), bead washing, bead enrichment, and Ion Torrent sequencing with PGM
    - performed in accordance with the manufacturer's instructions using Life Technologies reagents
15. Data analysis (see supplementary material)

IV

CONSISTENCY OF TARGETED METATRANSCRIPTOMICS AND MORPHOLOGICAL CHARACTERIZATION OF PHYTOPLANKTON COMMUNITIES

by

Vuorio, K., Mäki, A., Aalto, S.L. & Tiirola, M. 2019

Manuscript

Request paper from the author