
# Metagenomic - and Metatranscriptomic investigation of three plankton communities in the north Atlantic and the Baltic Sea

**Master's Thesis**

A thesis submitted to the

**University of Applied Science Bremerhaven**

for the degree of

**Master of Science**

written by

Gerrit Rohner

Conducted at the

Alfred Wegener Institute

Helmholtz centre for Polar and Marine Research

First Examiner:

Prof. Dr. Stephan Frickenhaus
University of Applied Science Bremerhaven
Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research

Second Examiner:

Prof. Dr. Allan Cembella
University of Bremen
Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research

**Declaration of Authorship**

I hereby declare that I have developed and written the enclosed Master Thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked as such. The thesis was not used in the same or in a similar version to achieve an academic grading before.

Bremerhaven, 20 July 2016 _____

Gerrit Rohner

**Acknowledgements**

Many people have contributed to the completion of the thesis. Special thanks go to:

**Dr. Uwe John** for the excellent supervision of my thesis and his patience and advice.

**Prof. Dr. Stephan Frickenhaus** and **Prof. Dr. Allan Cembella** for agreeing to be my examiners, their continued interest in my thesis and their time as well as the possibility to work at the AWI,

**Dr. Sylke Wohlrab** and **Stephanie Westphal** for their valuable council and expertise.

The group of **Ecological Chemistry** for the companionship.

**Jörn Helge Martens** for the constructive proof reading.

And finally **my parents, my brother** and my **partner** for their unwavering support.

**Abstract**

The following study describes marine eukaryotic microorganisms of the North-Atlantic around the western coast of Greenland, the north Atlantic waters around the Norwegian Lofotes and in a fjord system of the Swedish west coast. DNA sequences and mRNA gathered from expeditions to the coasts of Greenland, Norway and Sweden were used as data set for a comparative meta-genomic study in order to characterize the compositions of the local marine eukaryotic microorganisms. Furthermore an investigation the impact of abiotic factors like temperature, salinity, depth of the water column and availability of key nutrients on the composition of the population was conducted. Parallel to the assessment of the biodiversity through amplicon sequencing of the 28S *ribosomal DNA (rDNA)* the biological activities for the three sampling locations was achieved by analysis of the expressed mRNA. As an often reoccurring problem of transcriptomic studies is that the currently available databases do not work reliably for the annotation of sequences from marine environments, a purpose build databank with reference data purely from marine microorganisms was used for the annotation of the sequences in order to achieve a higher yield of annotated sequences. By linking the metagenomic observations and the results from the transcriptomic analysis more in depth insights were achieved that would not possible by either approach on its own, such as interactions between species or the differences between the most abundant species and the most biological active species.

**Kurzdarstellung**

Die folgende Studie beschreibt eukaryotische marine Mikroorganismen von drei Standorten des Nord-Atlantik rund um die Westküste von Grönland, den Gewässern rund um den norwegischen Lofoten und in einem Fjordsystem der schwedischen Westküste.
DNA - und mRNA - Sequenzen von zwei separaten Expeditionen dienten als Grundlage sowohl für metagenomische Fragestellungen, um die Zusammensetzungen der lokalen Populationen zu charakterisieren als auch einer Genexpressionsanalyse. Ferner wurden die Probenstandorte hinsichtlich der lokalen abiotischen Faktoren wie Temperatur, Salzgehalt, die Tiefe der Wassersäule und die Verfügbarkeit der wichtigsten Nährstoffe charakterisiert und die Auswirkungen von auf die Zusammensetzung der Populationen untersucht.
Parallel zur Beurteilung der biologischen Vielfalt durch Amplikon-Sequenzierung der 28S hypervariablen D1/D2 Region der ribosomalen DNA (rDNA) sowohl mit als auch ohne anschließendem phylogenetischen Placement durch Phylogentische Bäume wurden die biologischen Aktivitäten der drei Probenahmestellen wurde durch Analyse der exprimierten mRNA bestimmt.
Ein häufig wiederkehrendes Problem von Transkriptom-Studien ist die derzeit spärliche Verfügbarkeit eukaryotischen Sequenzen in den Datenbanken. Durch Zusammenstellung einer eigenen Referenzdatenbank mit den verfügbaren eukaryotischen Sequenzen wurde eine höhere Ausbeute an annotierten Sequenzen erreicht.
Die Verknüpfung der metagenomischen Beobachtungen mit den Ergebnissen der Genexpressionsanalyse erlaubte weitere Untersuchungen.

## I. Used abbreviations

| | |
|---|---|
| BAC | bacterial artificial chromosome |
| Bp | base pair |
| CBD | Convention on Biological Diversity |
| cDNA | complementary DNA |
| CDS | coding sequence |
| DNA | Deoxyribonucleic acid |
| DOC | dissolved organic carbon |
| EST | Expressed sequence tag |
| EtOH | Ethanol |
| GO | Gene ontology |
| GOS | Global ocean sampling expedition |
| ITS | internal transcribed spacer |
| LSU | large subunit |
| Mbp | Megabasepairs |
| MID | multiplex identifier |
| Min | minutes |
| MMETSP | Marine Microbial Eukaryote Transcriptome Sequencing Project |
| OMZ | Marine oxygen minimum zones |
| OTU | operational taxonomic unit |
| PFAM | Pfam protein database |
| PSP | paralytic shellfish poisoning |
| Psu | practical salinity unit |
| rDNA | ribosomal DNA |
| RNA | Ribonucleic acid |
| rRNA | ribosomal RNA |
| RT | room temperature |

| | |
|---|---|
| Sec | seconds |
| SSU | small subunit |
| Station “…” | St”…” |
| WGC | West Greenland Current |

# Table of contents

# 1 Introduction

## 1.1 Determination of biodiversity

Biodiversity can be measured in a number of ways. The most commonly used index is the taxonomic richness of a defined area with knowledge of changes to the richness over time. The taxonomical richness can be expressed by several diversity indices such as the Shannon and the Simpson index. Biodiversity is also measured among others morphologically, genetically and functionally. This project uses two "meta-"approaches for determination of biodiversity. Utilising environmental samples of mRNA and rRNA allows the determination of Biodiversity by metabarcoding (rRNA) and functional metatranscriptomics (mRNA).

## 1.2 Eukaryotic molecular markers for metagenomics

Requirement for a function molecular marker is to contain enough variations to allow for determination of evolutionary changes as well as reliable and accurate species distinction. Depending on the study either DNA, RNA or protein molecules can serve as molecular marker.

Pioneer work on molecular markers was conducted by Woese et al. (Woese 1990) by their utilisation of the 16S rRNA for separations of the three taxonomical domains of Eukarya, Prokarya and Archea.

Most commonly used molecular marker in eukaryotes are the 18S *small subunit* (SSU) or parts of the 28S *large subunit* (LSU). The *ribosomal DNA* (rDNA) array in eukaryotes consists of the 18S and 28S subunit genes, which are separated by the 5.8S gene and *internal transcribed spacers (ITS)* up- and downstream of the 5.8 S gene. Both the 18S and the 28S are characterised by different evolutionary rates, the ITS regions are supposedly the most variable ones but it is therefore more difficult to align more evolutionary distinct taxa.

The *small subunit (SSU)* with the slower evolutionary rate (Hillis 1991) is featured in more publications as it fulfils a conform function in all organisms. The SSU is applied to differentiate large numbers of not closely related species. The 28S LSU contains the hyper variable D1/D2 domain. This domain is between 600 and 700 *base pairs (Bp)* in length and due to numerous known variations it allows for high resolution species discrimination. Studies have used the 28S LSU for differentiation of multiple protists such as Dinophyta (Medlin 1998, John, Fensome et al. 2003, Toebe, Joshi et al. 2012, Toebe, Alpermann et al. 2013, John 2015, Tillmann 2015), Bacillariophyta (Moniz 2009) and Haptophyta. Trade-off for the higher resolution in regards to species discrimination is that despite the observed advantages of the LSU fewer studies have been conducted (Peplies 2004).

For in depth differentiation and characterization of novel species by molecular markers often simultaneous sequencing of all three regions was utilised (Tillmann 2011).

## 1.3 Origen of meta-genomic and meta-transcriptomic studies

The term metagenomics first was used in a ground breaking publication of Handelsmann et al. (Handelsman 1998). The group coined the term Shotgun metagenomics which in turn was picked up by many following groups using different approaches. The differences to the metabarcoding approach where only a target molecular marker are amplified and sequenced is, that in metagenomics parts of the genome or even whole genomes from environmental samples are amplified and sequenced. This approach is due to the cell size and structural complexity differences only suitable for prokaryotes. The new term meta- means that instead of looking at a single, or a selected few target genes in a study the focus shift towards registering all genomic information contained in the sample in an attempt to find all members of the population living in the sampled environment and in addition to quantify the results to determine the composition of the population. While the metagenomics shotgun sequencing approach was widely used in prokaryotic studies (Beja 2000, Beja 2004, Venter 2004) deployment of meta- genomics for eukaryotes lagged behind.

## 1.4 Current state of knowledge in marine eukaryote metagenomics and metatranscriptomics

The lagging of eukaryotes compared to their prokaryote counterparts are manifold. Firstly it stands to reason that the current Tree of life is inaccurate regarding protists due to an overabundance of data of cultivatable lineages with simultaneous poor representation of data from uncultivated lineages (Roy 2014). Many globally occurring taxa cannot be successfully cultivated over longer periods (Cuvelier 2010, Jardillier 2010) or often stay undetected due to their small cell size (Gilbert 2011). And yet exactly this group of heterokonts has been postulated to contain the most ribosomal biodiversity of eukaryotic plankton (Pawlowski 2012, Vargas 2015) and are reported to have a tremendous impact on the global primary biomass production. Two recent publications credited commonly by PCR detected yet uncultured pico-prymnesiophytes with 25 % of the overall primary production (Cuvelier 2010, Jardillier 2010).

Additional issues regarding eukaryotes are foreign gene acquisition by either plastid endosymbiosis (Moustafa 2009, Curtis 2012) or horizontal gene transfer (Keeling 2008, Chan 2012, Bhattacharya 2013) as well as the generally increased DNA contend of eukaryote cells compared to prokaryotic cells. This is best shown with the examples of the Dinophyta. Dinophyta contain an average of 100 chromosomes. Each one of those 100 chromosomes contains at least as much DNA as a single yeast cell (Spector 1984). However only a small part of the DNA in eukaryotes codes for genes and can be used in metagenomic studies. Compared to prokaryotes eukaryotes therefore have a very low Gene to non-coding DNA ratio. Therefore the generation of sequences is not the problem but not many of those generated sequences will present a coding region which actually contains information. Non coding regions are also more prone to higher genetic variation which can make to matching of sequencing to sequences in a databank difficult.

Considering the difficulties of eukaryote marine organisms current methods allow only for partial recovery of a sample`s biodiversity with little success regarding sample replication (Edgcomb 2011).

Considering the difficulties with eukaryotes most transcriptional studies to date concentrated on lab grown clones of a singular species and the changes in the species transcriptome in reaction to biotic and abiotic factors. Abiotic factors contain nutrient availability and environmental properties of the sample environment while biotic factors describe mainly grazing and the impact of parasites. Until recently abiotic factors were projected to have a more extensive impact on local biodiversity compared to biotic factors (Verity 1996, Lima-Mendez 2015, Worden 2015).

Grazing is an extensively researched biotic influence often in context of toxin production of Dinophyta (Tillmann 2011, Yang 2011, Tillmann 2015, Waal 2015) as well as Haptophyta (John 2001). John et al. compared the effectiveness of the toxins prymnesiophyte flagellate *Chrysochromulina polylepis* against the grazing of the heterotrophic dinophyta *Oxyrrhis marina.* In a comparative approach two different clones of the prymnesiophyte, one clone toxic and the other non-toxic were grazed by *O. marina.* The toxin was observed to reduce the grazing considerably while not being able to kill the predator in any concentration.

A study of 16 toxic strains of the *Alexandrium* genus were tested for the toxic effects on the dinophyta *Oblea rotunda* and *Oxyrrhis marina* (Tillmann 2002). Some of the tested *Alexandrium* strains caused loss of motility and cell lysis. In addition it was observed that toxic effects occurred independently of paralytic shellfish poisoning (PSP) toxins. Induction of PSP toxins in presence of copepod species was studied with a transcriptomic model study for the dinoflagellate Alexandrium minutum (Yang 2011). This first transcriptomic study of grazer-induced induction of toxins found a limited set of 14 genes affected by copepod presence. The study was expanded upon with a functional genomic approach (Wohlrab 2010). *Alexandrium tamarense* was studied regarding species-specific increase in toxin content when exposed to three copepod species. The study demonstrated species specific response for another *Alexandrium* species, hinting at co-evolutionary processes of the phycotoxins.

Aside from grazing marine dinoflagellates are parasitized by other dinoflagellate species. Host-specific infections were observed in a study spanning 3 several years (Chambouvet 2008) with host species being infected by a single genetically distinct parasite year after year.

The molecular processes of an infection were studied for the parasitoid *Amoebophrya sp*. for two of its hosts *Alexandrium tamarense* (Lu 2014) and *Karlodinium veneficum (Bachvaroff 2009).* The study between interactions of *Karlodinium veneficum* and *Amoebophrya sp*. was the first transcriptomic study of host-parasite interactions generating 898 ESTs. The study of Lu et al. demonstrated similarities in the general mechanisms parasitoid infection that have remained stable throughout evolution within different phyla.

Most published studies under laboratory conditions up to date concentrating on observing transcriptomic changes during changing of abiotic factors such as temperature, light influx,

variations of nitrogen -, phosphate- or carbon concentration in the medium. This concept gives a general starting point as to what to look for as the field of metatranscriptomics in eukaryotes is still fairly new and there is a lack of available databases to compare and verify one's findings to.

One example for studies with EST libraries dealing with metabolic change in eukaryotic cells after changing of the environmental temperature is the study of Mock et al. (Mock 2005). The group constructed an EST library from the diatom *Fragilariopsis cylindrus* 5 days after shifting the incubation temperature from 5 °C to -1.8 ° C with the aim to gain insight into the cold adaptation of the diatom. The impact of temperature on the transcriptome of eukaryotes was further explored by the group of Toseland et al. (Toseland 2013) and the group of Uhlig and al. (Uhlig 2015). Toseland et al. (2013) found that the rate of protein synthesis increases with higher temperatures while the number of ribosomes and rRNA in the cells decrease. This leads the group to the conclusion that temperature has a greater impact on the transcriptome of cells than currently recognized. The group of Uhlig et al. (2015) studied the expression of ice binding proteins in several organisms of the Arctic Ocean, a field of study which was also worked on by other groups. A publication by Krell et al. (Krell 2008) used an EST approach to find a whole new class of salt-stress induced icebinding proteins in the diatom *Fragilariopsis cylindrus.*

Apart from the influence of temperature on the transcriptome other key factor are the available concentrations of nutrients such as phosphor and nitrogen and to a lesser degree other substrates such as sulphate or iron. Bender et al. (Bender 2014) compared transcriptional profiles of three phylogenetically distant diatoms *Thalassiosira pseudonana, Fragilariopsis cylindrus,* and *Pseudo-nitzschia multiseries* grown under transitioning conditions – from optimal condition to nitrate limitation. Aim of the study was to identify transcriptional patterns causes by substrate limitation that are uniform to all three species as well as differences in the metabolic response. Another study that investigates metabolic changes in response to environmental impulses especially in diatoms is the publication from Parker et al. (Parker 2005). The group exposed the diatom *Thalassiosira pseudonana* to simultaneous changes in regards to light, nitrogen substrate and temperature in order to find whether there were synergistic effects of the abiotic impulses on the expression levels of the metabolic pathways of nitrogen- and carbon cycling. A publication over the nitrogen- and phosphor metabolism of a haptophyte was conducted by Beszteri et al. (Claire 2005, Beszteri 2011). EST libraries of the haptophyte *Prymnesium parvum* under nitrogen- and phosphor-starvation were generated and compared with available transcriptome datasets of other haptophytes. The group of Morey et al. conducted a similar microarray experiment for nitrogen- and phosphor saturation with the Dinophyta *Karenia brevis.* A publication along the same lines of conducted research was made by the group of Maheswari et al. (Maheswari 2010). The group stresses that in order to deepen the knowledge of diatoms studies are needed that link specific metabolic functions to diatom specific genes. In the study the group grew and analysed the two diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* under different conditions regarding different sources of nitrogen, varying the saturation with carbon dioxide, silicate and iron, and furthermore abiotic stresses such as lowered temperatures and low salinity. In total 130,000 ESTs were constructed and the group was able to created functional

annotations of varies transcripts providing insights into expression patterns of genes involved in various metabolic and regulatory pathways was well as the discovery of novel genes with unknown functions.

The specific influence of iron on the transcriptomic profile of phytoplankton, especially diatoms, was investigated by the group of Marchetti et al.(Marchetti 2012). The group used comparative transcriptomics between diatoms and other eukaryotic plankton grown in iron-enrichment medium. The diatom transcriptome reacted strongly to the presence of iron with hundreds of genes differentially expressed in the iron-enriched community compared with the iron-limited community. The effected genes hail from different metabolic pathways such as chlorophyll components, nitrate assimilation and the urea cycle, as well as synthesis of carbohydrate storage compounds. The group reasoned with these findings that oceanic diatoms consistently dominate iron enrichments in high-nitrate, low-chlorophyll regions as the diatoms appear a continued dependence on iron-free photosynthetic proteins rather than substituting for iron-containing functional equivalents. This ability would allow diatoms to divert their newly acquired iron toward nitrate assimilation giving them an advantage other eukaryotic plankton.

So far only few publications are available in which metagenomic- and metatranscriptomic observations are combined in order to gain a deeper understanding of functional processes within naturally occurring planktonic communities. An example for Dinophyta was published by Jäckisch et al (Jaeckisch 2011). The group characterized *Alexandrium ostenfeldii* by function transcriptomics as well as by repeats and transponson-related sequences in the spliced leader sequences. The findings were compared with three other datasets of *Alexandrium* species and the generated data represents currently the largest genomic dataset for Dinophyta.

Another comparative study featuring metagenomic- and metatranscriptomic observations was conducted by Alexander et al. (Alexander 2015). A metatranscriptomic characterization of an algal bloom featured mainly *Prorocentrum minimum* as identified comparatively with clonal *P. minimum* sample by SSU 5.8S LSU operon sequences and phylogenic placement by mitochondrial dataset. Functional annotation of the bloom transcriptome was achieved by BLAST against *gene onthology (GO)* and *Pfam protein databank (PFAM)* databanks.

Available studies for diatoms were published by Alexander et al., Pearson et al.and Marchetti et al. (Marchetti 2012, Alexander 2015, Pearson 2015). Alexander et al. generated 5 metatranscriptomic assemblies and compared the contained expression profiles of *Skeletoma spp.* and *Thalassiosira rotula* against generated reference expression profiles through additional nitrogen- and phospherus incubation experiments. Hereby the group researched the niche partitioning of the two diatoms in regards to the plankton paradoxon. Marchetti et al. researched the influence of iron on the transcriptome of diatoms comparing metatranscriptomes of pacific environmental samples with iron-enriched clonal diatom datasets and was able to explain the rapid response of diatoms to changes in availability of iron. Datasets of comparative metatranscriptomics of Antarctic environmental samples were generated by Pearson et al..  Functional annotation was conducted by a custom reference

dataset from available transcriptomes. The group compared the transcriptomes of three locations and diatoms in particular and was able to show the influence of 4°C temperature difference between locations on the expression of proteins.

## 1.5 Objectives of the project

This project used a two forked approach for determination of biodiversity. By looking at the biodiversity of the three sampling locations we gain insight into the local community composition of microorganisms. This is foremost taking stock what species can be found in an area with a degree of getting an idea of the abundances of the local individual species. As there are multiple sampling locations for each station, we gain a deeper insight in the biodiversity of the sample location. The three different geographical regions featured in the project were specifically chosen as they were guaranteed to include numerous environmental differences to impact either the inter-regional species diversity and/or the active metabolically pathways. Several biochemically relevant factors like dissolves nitrogen, carbon and phosphate were recorded during collection of the samples. Those factors allow researching the influence of the environment on the biodiversity and the expression profiles. The detection of a particular species or an observation of a high or low relative abundance of the species at one of the stations could potentially be linked to a higher or lower concentration of nutrients. This might be a hint for a higher base need of the specific nutrient for the species. By comparing the three sets of data from the three sampling locations with each other we aim to gain information on a much larger scale than would be accessible from the individual sample sites on their own.

- What local species composition can be determined by the 28S rRNA LSU approach?
- Can the composition of a local population be linked to any environmental parameter of the given sample location?
- Are any trends of a change in local biodiversity visible?

The second line of approach in this study is to look at the transcriptome of the samples from the three locations. On the smallest scale we gain an insight on the composition of microorganisms as well as whom the most active organisms were at the time when the sample was taken. Looking at the differences in the transcriptomes of samples from one of the sampling locations, might show local changes in the biodiversity, which again in turn might be a sign of a local change in environmental conditions, for example a locally increased availability of nutrients.

Looking at yet a larger scale the project aims to couple the findings of transcriptome part of the study with the results of the previously described genomic studies. Firstly this allows an estimation of sequence relatedness or diversity between the two individual sample pools. As combining observations regarding biodiversity from metagenomics and metatranscriptomics is a relatively novel approach with few publications any insights are potentially valuable. As understanding the phytoplankton diversity and the influence of various environmental conditions on the phytoplankton composition is vital for the understanding of the marine

ecosystem, any insight gained from this study could be helpful to understand potential reactions of the marine ecosystem to for e.g. global change.

One of the most severe bottlenecks in current metatranscriptomic studies are the lack of useable databases. Matching one's sequencing results to an existing database can be difficult. It is reported that studies in the past have favoured groups of marine eukaryotes leading to those groups being overrepresented in the dataset while other less popular groups are underrepresented or not represented at all (Burki 2014, Campo 2014). Going a step further, after matching the sequenced amplicons to the database the functional annotation of the *coding sequence (CDS)* within the database is the next important step. For the databases currently in use the reliability and completeness of the annotation is a potential issue. In order to circumvent this potential shortcoming an objective of the project is the creating of a custom in house databank. This databank is to contain solely marine phytoplankton an through the much narrower scope of the databank both the functional annotation and identification of sequences is projected to be more accurate in comparison to the larger currently available databanks. A potential pitfall however of the usage of a narrowed down database could be the occurrence of sequences in the samples that have been omitted or not included in the database leading to an unsuccessful matching, which has been reported in other studies (DeLong 2014).

Conclusion of the major objectives of the project:

- Determination of potential interaction between environmental/chemical parameters and abundance of observed species
- To compare the biodiversity of the three samples locations in regards to
    - Differences in compositions of the local population between the stations of one of the three sampling region
    - Differences in the species composition between the three sampling regions

- To compare if usage of the custom database results in lower numbers of unassigned sequences compared to queries against conventional general purpose databases.

- To compare the determined metagenomic 28S rRNA biodiversity with observed species diversity from functional transcriptomic approach.

- To find the most active metabolic pathways in the transcriptomes and to find potential correlations to the observed chemical and physiological parameters.

# 2 Material and Methods

## 2.1 Sampling locations

The basis of this project were samples gathered during two cruises in 2013 and 2014. The sampling sites in question are the Disko Bay on the Greenlandic west coast, the Sortlandsound of the Norwegian Lofoten and a Fjord system east of the Island Orust of western Sweden. The three particular sites were choses under the premises that different environments will result in a different composition of the local microbial populations.

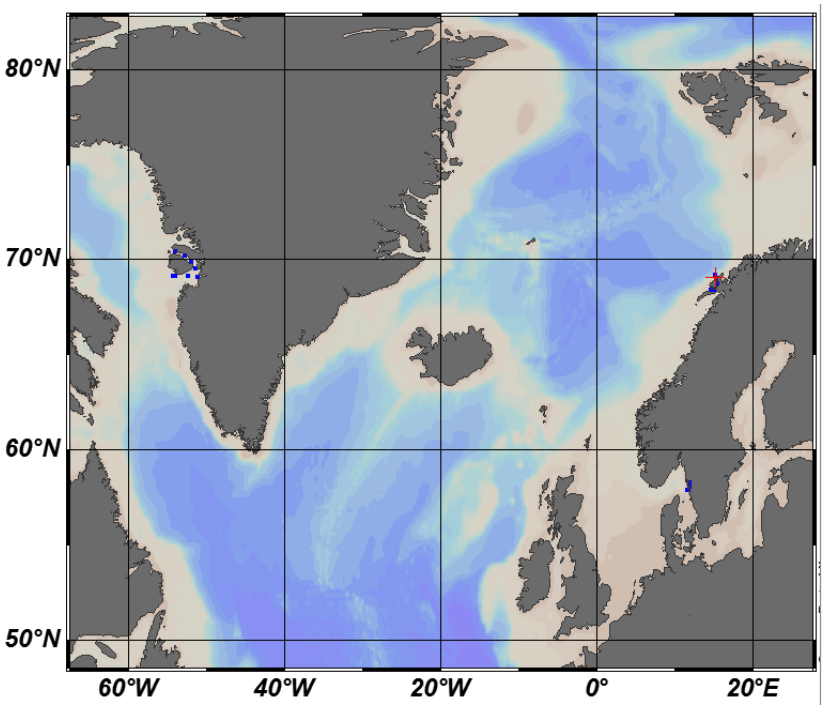



***Figure 1**: Overview of the three sampling locations featured in this study. Plot shows Greenlandic sites on the west coast of Greenland, the Sortlanssound in Norway and the fjord system on the western coast of Sweden.*

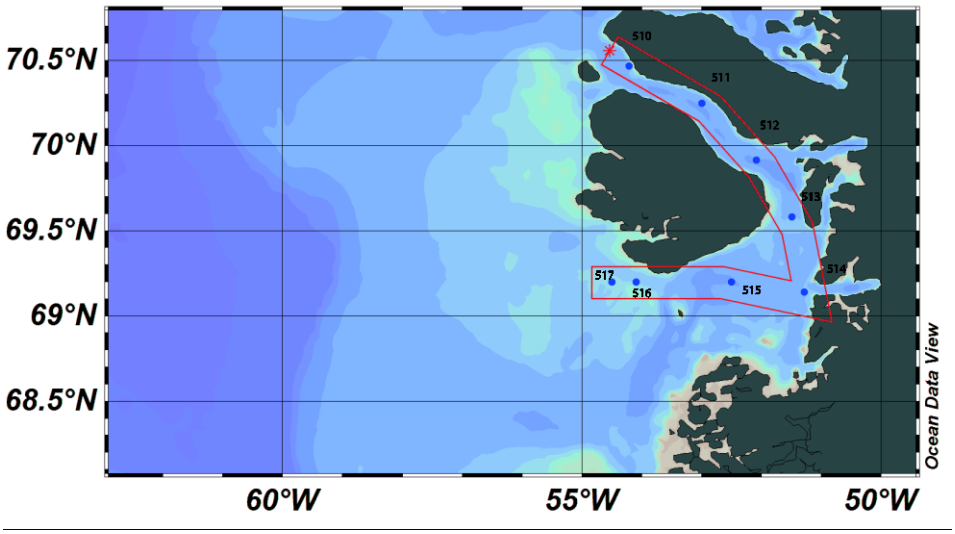



**Figure 2a**: *Location of the 5 selected Danish sites west of Greenland. Unlike for the two other regions the sites picked for this experiment are not sequential. The samples from the stations St510, St511, St515, St516, and St517 were chosen for the experiment. The sample sites follow the Sullorsuaq strait, which separates the Disko Island from the Nuussuaq peninsular, starting at the Baffin Bay in the north through the Vaigat strait into the Disko Bay in the south of Disko Island. However for the description of the physiological conditions of the regions the local stations St512, St513 and St514 were included as well. Compared to the other two regions the later following evaluation of the physiological conditions of the sample regions the region Disko Bay in Greenland consists of eight instead of five points of measurement.*

The waters of the west coast of Greenland are characterized by three water bodies: the surface water to a depth of 60 meter. The subsurface between a depth 60 and 250 meters and the *West Greenland Current (WGC)* waters below 250 meters.

The surface waters are influenced by freshwater from glaciers, icebergs, and sea ice melt and are warmer than subsurface waters in the summer months. The surface water temperature was recorded in coastal waters between 1 and 3 °C (at 5 m water depth) (Ribergaard 2013), along with minimum surface water salinity of 31.6 (formerly *practical salinity unit (psu)).* The salinity of the surface water increases with the distance to Greenland mainland with a salinity of 34.2 in the Qaumarujuk Fjord. Regarding temperate and salinity stratification of the parameters was observed during the summer months. The waters of the WGC are characterized by a maximum water temperature of 4.9 °C along with maximum salinity of 34.8 (Fyllas Banke at 400 m water depth). The maximum coverage of sea ice in the area occurs around the month of May while the minimal sea ice coverage takes place around August. (DMI 2007, Krawczyk 2014)

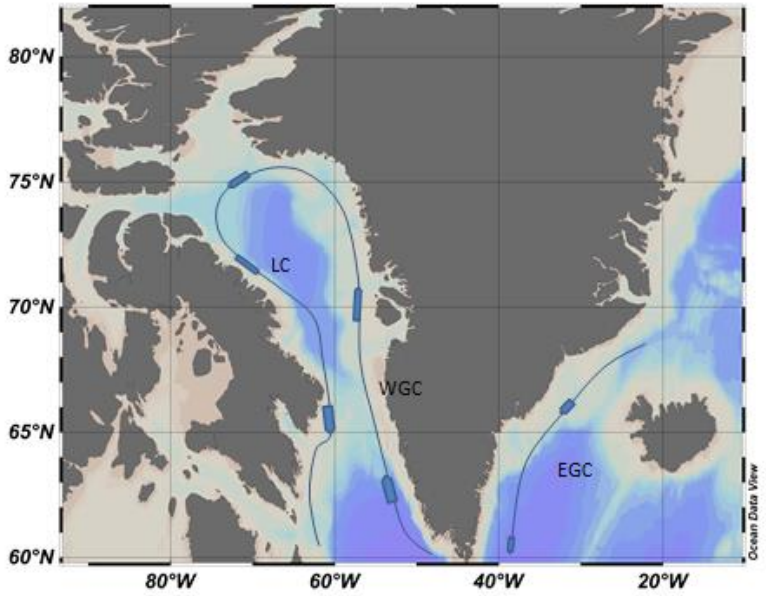



***Figure 2b*** *Sea currents of Greenland's western coast. Figure shows the major currents influencing the Greenlandic sample sites. Named currents are clockwise Eastern Greenland Current (EGC), Western Greenland Current (WGC) and the Labrador Current (LC).*

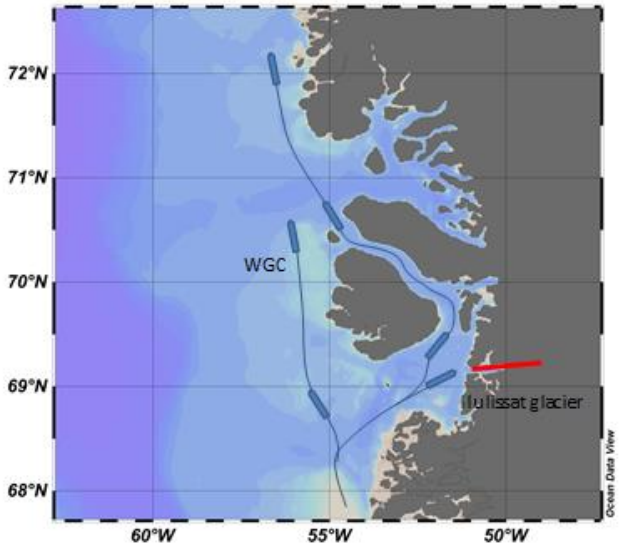



***Figure 2c:*** *More detailed overview sea currents of the Disko Bay. The West Greenland Current flows from the south around both sides of the Disko Island, on the western side passing through the Disco Bay and Vaigat strait further north. The Ilulissat Ice fjord containing the Jakobshavn Isbræ is marked in red.*

**Lofoten, Norway**

Oceanography

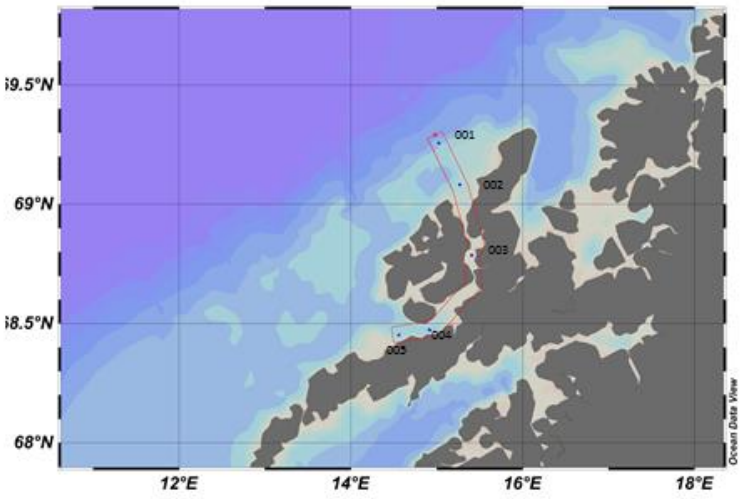



***Figure 3a*** *Location of the 5 selected Norwegian stations. The stations follow the Sortlandsound between the three isles of Andøya, Langøya and Hinnøya. The stations start with the station St001 north of the sound in the European polar sea, following the sound south-west until its end.*

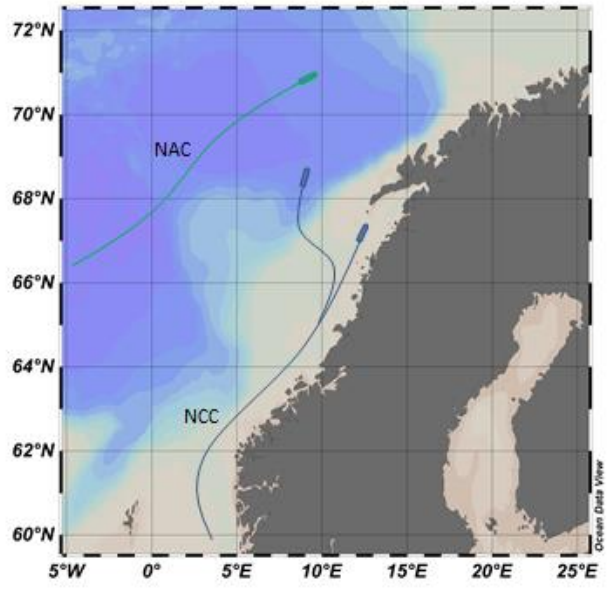



***Figure 3b:*** *Sea Currents of Norwegian coast. Figure shows the major currents influencing the Norwegian sample stations. Named currents are the Norwegian Coastal Current (NCC) and the North Atlantic Current (NAC).*

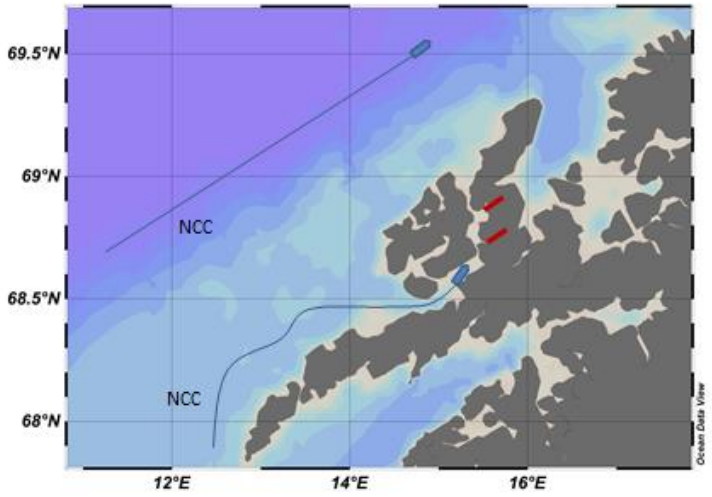



***Figure 3c:*** *Sea Currents of Norwegian coast. Figure shows the major currents influencing the Norwegian sample stations. Coastal regions around the Sortlandssound are mainly influenced by the Norwegian Coastal Current (NCC). The NCC flows through the sound from south to north. Minor influxes of freshwater are from rivers of the Forfjorden and the Langvatnet Lake to the north, both marked in red.*

The waters around the Lofotes are characterized by Artic water in the upper 500 m of depth. The salinity decreased from 35 near the islands to 37,9 over the span of 30 km. In the same gradient the temperature decreased from 3 °C to 0 °C with the exception of the immediate surface water. To the southeast of the Lofoten Atlantic water dominated the upper 600 – 700 m with a salinity around 35 and temperatures around 3 °C.

In the Greenland basin to the northwest of the Lofoten the water to a depth to 1200 m contains salinity below 34,9 and a temperature around 0 °C. This water also forms an intermediate layer in the Lofoten basin around 800 to 1000 m.

The deeper water layer originates from the Arctic ocean with a salinity around 34,9 around 2000 m in depth (Drange 2005).

**Fjord system east of Orust, Sweden**

Oceanography

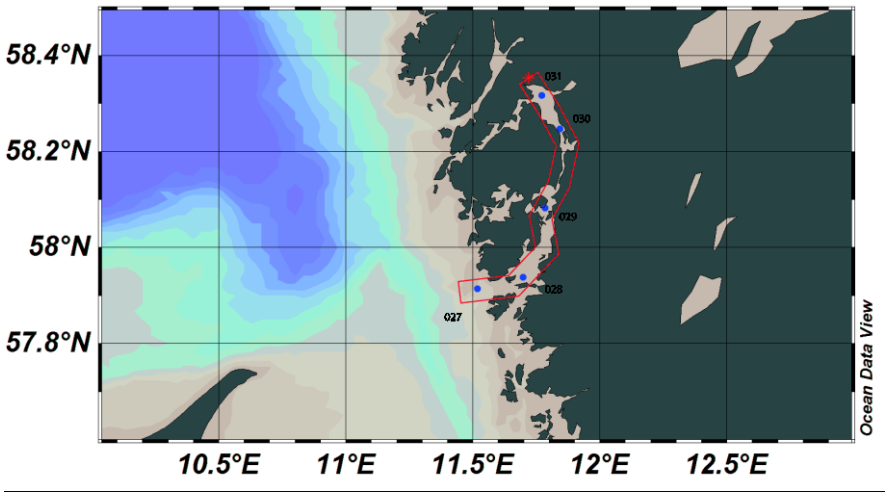



***Figure 4a:*** *Location of the 5 selected Swedish stations. The five stations follow the fjord system west of the Island Orust from the Skagerrak north into the Havstensfjorden-Svälte kile. The sampling stations are numbered consecutively from St027 at the mouth of the fjords to St031at the mouth of the Byford. Samples from Sweden were taken during the expedition "HE-431" in 2012.*

The sample stations in the fjord system near Goteborg lie between the Skagerrak at the mouth of the fjord and the Byfjord at the fjords head. The area is defined by the shallowness of the sea with a depth of only around 30 m. The water column is further defined by a larger difference in salinity due to the water exchange with the North Sea. The salinity of the surface water in the Kattegat is 25 - 30 and the temperature varies seasonally between 6 and 20 degrees. The deeper layer of water in the Kattegat, starting at a depth of 10 - 20 m, stems from the North Sea with salinity around 35 and an average temperature of 6 °C. The larger difference in salinity is maintained due to strong stratification of the water column. In the deeper water layers oxygen depletion occurs frequently caused by the lack of vertical mixing of the water column as well as different residence times of the top and bottom water layers. The top layer of brackish Baltic Sea water stays only few weeks in the areas while the bottom layer can stay up to months. For this sampling location more so than for the others there can be expected a distinct anthropic influence as the samples were taken from the bay of a larger city (Goteborg) and the Kattegat outside said fjord system are a major shipping route (SMHI 2013).

Sweden

The Danish and Swedish region of the Kattegat contains multiple currents influencing the physiological properties of the water in the sample regions. Low saline water from the Baltic

Sea mixed with more saline water from the North Sea. The water exits the Kattegat north along the Norwegian coast.

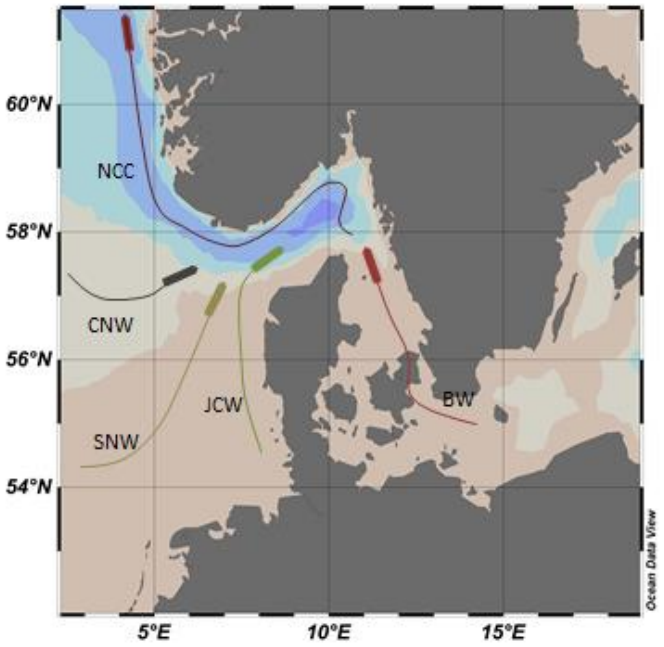



***Figure 4b:*** *Sea Currents of the Kattegat. Figure shows the major currents influencing the Swedish sample stations. Named currents are clockwise Baltic water (BW), Jutland coastal water (JCW), southern North Sea water (SNW), central North Sea water (CNW) and Norwegian Coastal current (NCC).*

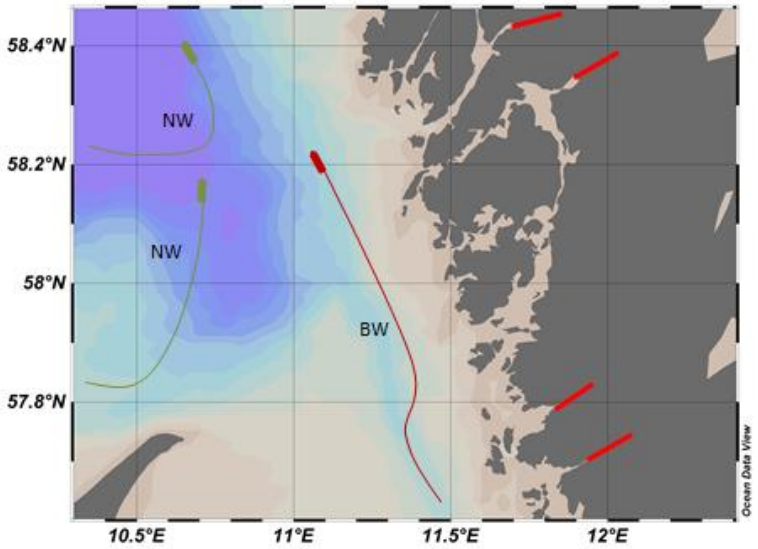



***Figure 4c:*** *More detailed overview sea currents of the Kattegat. Figure shows the major currents influencing the Swedish sample sites. Named currents are Baltic water (BW) and North seawater (NW). Major freshwater inlets from rivers marked read.*

## 2.2 Sampling procedure

Seawater was collected from depths of 10 meters using a pump and fractionated using a filter tower. The gained fraction with 200 µm, 50 µm and 20 µm cut-off value respectively were rinsed from the filters of the filter tower using filtrated seawater and transferred into Greiner tubes. The sample volume was increased to 45 ml with filtered seawater and afterwards the sample was aliquoted into four samples with 15 ml volume each. Both kinds of samples, for DNA- as well as for RNA - isolation were pelleted by centrifugation. The pellet for DNA isolations was resuspended in 400 µl AP1 Lysis buffer and transferred into a 2 ml Freezer vial. The pellet for the upcoming RNA isolation was resuspended in 1000 µl TRIzol and also transferred into a 2 ml Freezer vial. Both samples were stored at -80 °C until further use.

## 2.3 DNA isolation

DNA isolation was carried out by Quiagen DNeasy© Plant mini kit. Not more than 100 mg wet weight of sample were lysed. 400 µl buffer AP1 and 4 µl RNase A were added. Sample was vortexed and incubated at 65 °C for 10 min. 130 µl of buffer P3 was added an incubated 5 min on ice. Lysate was centrifuged at 20,000 x g for 5 min and pipetted into QIAshredder spin column. Spin column was placed into 2 ml collection tube and centrifuged at 20,000 x g. Flow-through was transferred into new tube and 1,5 volumes of buffer AW1 was added. 650 µl of mixture was added to DNeasy Mini spin column placed in 2 ml collection tube. Column and tube were centrifugated 1 min at 6000 x g. Flow-through was discarded and previous step was repeated with remaining sample. Spin column was transferred into new 2 ml collection tube and 500 µl AW2 was added. Column and collection tube were centrifuged at 6000 x g for 2 min. Spin column was transferred into 1,5 ml micro centrifuge tube. 100 µl AE buffer was added to column and incubated for 5 min at RT. Eluted DNA was removed from spin column by centrifugation for 1 min at 6000 x g. Elution step was repeated with remaining sample. DNA was stored at -20 °C until further use.

## 2.4 RNA isolation

RNA isolation was conducted according to an AWI in house RNA isolation protocol

Cell lysis was achieved by using a Magnalyser at speed setting “6.5” with two impulses 20 sec length. 200 µl Chloroform were added and vortexed 30 sec for. Mixture was transferred to Phase-lock-tube, inverted and incubated at RT for 5 min. After incubation mixture was centrifuged at 12.000 x g for 15 min while cooled at 4 °C. Transfer upper phase of supernatant transferred to 1.5 ml Eppendorf cup. 5µl of 5M linear acrylamide (10-20 µg/ml) and 1/10 volume 3M sodium acetate were added and well mixed. 1 volume of cooled isopropanol (100%) was added and vortexed for 15 sec. Mixture was incubated at – 20 °C for 90 min and centrifuged 12.000 x g for 20 min while cooled at 4 °C. Supernatant was

removed, 1 ml cold *Ethanol (EtOH)* (70%) was added and mixed by vortex. Mixture was centrifuged 12.000 x g for 10 min while cooled at 4 °C. Supernatant was removed and washing step was repeated with 1 ml cold EtOH (96%). Mixture was pelleted by centrifugation (max speed 10 min, 4 °C). Supernatant was removed and pellet resuspended by flicking the tube after adding 20 µl RNase-free water. 2 µl of RNA suspension was aliquoted for determination of the RNA concentration by Nanodrop. Isolated RNA was stored at -70 °C.

## 2.5 Gene Markers/ Generation of Sequences

### 2.5.1 DNA Sequencing

DNA as well as RNA sequences from the three sampling locations were treated with different methods for sequencing. Micro plankton fractions of the DNA samples taken during the expedition to Greenland were sequenced in house at the Alfred Wegener Institute in Bremerhaven, Germany using 454 sequencing. All DNA samples from the stations from Norway and Sweden gathered during the "Heinke" expedition in 2014 were sequenced by the Max Plank Institute.

Target for DNA sequencing were the D1-D2 sequences of the *Large Sub-Unit (LSU)* rDNA (Sonnenberg, Nolte et al. 2007).

### 2.5.2 Genmarker 454 Sequencing

The 454 Sequencing method was chosen for the generation of all sequences used in this study. The 454 approach allows a wide range of experimental designs as it allows with minor changes in the used protocol for uni- or bilateral sequencing as well as pooling of samples for larger number of query amplicons.

For identification of amplicons in pooled samples *multiplex identifiers (MID)* are added during planning and construction of the DNA library.

For the 454-amplicon approach used in this study the primers Dir-F and Dir2-CR, for the amplification of the 28S hypervariable D1/D2 region were used. The primers were modified with several sequences for the 454 sequencing:

- The above mentioned MID for identification of pooled samples. As the Lib-A protocol, offing both bidirectional and unidirectional sequencing, was used the MID sequences of the forward and reverse primer needed to vary, in order to tell apart the forward and reverse reads.
- An additional identification key consisting of the 4 nucleotides 5' – TCAG – 3'. This sequence allows for simultaneous sequencing of a number of samples in one sequencing run
- The sequence adaptor A (5' – CGTATCGCCTCCCTCGCGCAA – 3') from the LibA kit for the forward primer and the sequence adaptor B ( 5' – CTATGCGCCTTGCCAGCCCGC – 3') for the reverse primer.

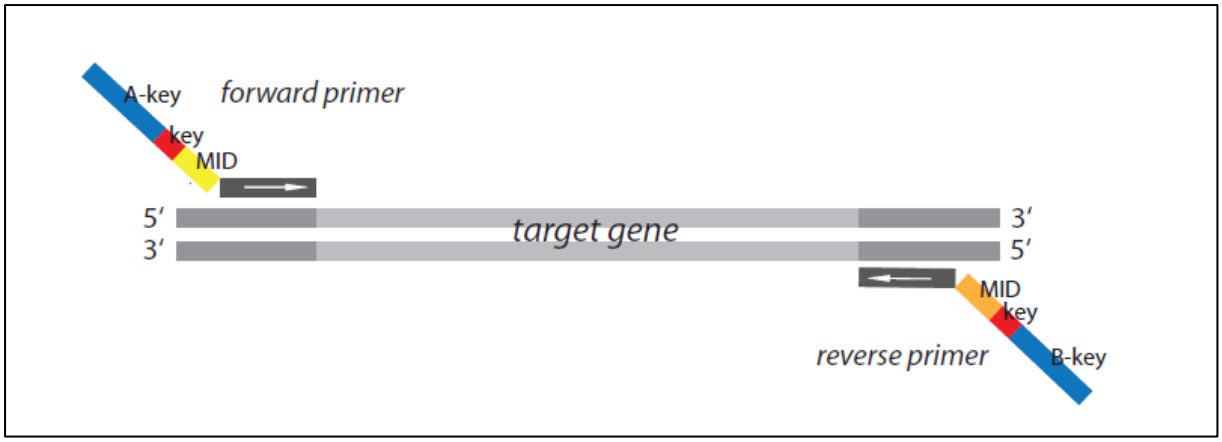



***Figure 5 :*** *Schematic representation of the primer constructs used in the 454-amplicon sequencing protocol (Westphal 2013). In this study the target gene was the hypervariable 28S D1/D2 region. Forward primer: 5' - sequence adaptor A - TCAG key - MID Adaptor - template specific primer - 3'.*

*Reverse primer: 5' - sequence adaptor B - TCAG key - MID Adaptor – template specific primer - 3'. MID sequences vary. For the Lib-L approach, the reverse primer does not include a MID adaptor. The A and B key from the Lib-A kit vary from the Lib-L kit.*

### 2.5.3 RNA sequencing

All RNA samples were sequenced at the Leibniz Institute in Jena, Germany using paired-end illumina sequencing.

### 2.5.4 Metatranscriptomic Reference database

For an optimal alignment of the cDNA sequences generated by the Leibiz Institute in Jena the construction of an in house custom reference database was proposed.

Said tree is going to be based on the following datasets:

- Selected data sets of the *Marine Microbial Eukaryote Transcriptome Sequencing Project* (MMETSP)(Keeling 2014)
  CDS regions of the following selection of "combined assemblies):
  [http://camera.calit2.net/mmetsp/combinedassemblies.php]
  As well as [: http://camera.calit2.net/mmetsp/listMetazoa.php]
  In early 2015 the data of the MMETSP project was moved to the iMicrobe project found at [: http://data.imicrobe.us/ ]
- Additional CDS from the RefSeq database of the NCBI [http://www.ncbi.nlm.nih.gov/gene/]
- Additional CDS from the short single read EST database of the NCBI [http://www.ncbi.nlm.nih.gov/nucest/]
- Furthermore CDS from multiple Haptophyta are needed with are lacking in the previous three mentioned sources. The species *Prymnesium*, *Pavlova*, *Phaeocystis* and *E. hux* were proposed.

Annotation of the coding sequencing is proposed to be conducted via either KEGG (Kanehisa 2013), COG (Tatusov 2000) and/or the Trinotate webserver (Grabherr 2013).

### 2.5.5 UCLUST

Generated sequences were clustered into functional groups by the UCLUST algorithm (Edgar 2010) found at:

http://drive5.com/usearch/manual/uclust_algo.html

## 2.6 Ocean Data Viewer

Visualisation of gathered environmental data was conducted with Ocean Data Viewer (ODV) (Schlitzer 2016). ODV allows for analysis and visualization of oceanographic data. Used were ODV versions ODV 4.7.4 and ODV 4.7.5. ODV software package was used to plot measured environmental data and nutrient distribution for the sample locations.

## 2.7 R-Studio and used packages

Program R-Studio (R Core Team 2013) was used for several inquiries of this study.

Clustering of stations was conducted by R packages “vegan” (Oksanen 2013) and “Deseq2” (Love 2016). The “metaMDS” function of the package was used for *Non metric dimensional scaling plot (NMDS)* (figure 16a). NMDS allows plotting of any distances matrixes in an ordination plot with reduced space. The method is robust enough that it can cope with missing values in used distance plots if enough values are left to determine relations of the objects in the plot.

Relative Abundance plots were conducted by R package “MANTA” (Marchetti 2016). *Relative Abundance (RA)* is an integer based derivate of a Bland-Altman plot. The Bland-Altman correlation for analysing agreement between two different assays. Along the same lines the RA plots allows visualization of two condition data. The plot is a rotated 45 °C scatterplot plotting unique data points in a distinctive arrow shape.

Test for similarity by Mantel was conducted by R package “ade4” (Dray 2016). The Mantel test is testing the correlation between two matrixes. Word cloud plots of results of the enrichment analysis were conducted by R package “wordcloud” (Fellows 2013). Word cloud or tag cloud diagrams is a method for visualisation of text data. Importance or abundance of data values is represented by increased front size. The method allows for fast identification of most prominent entries. The word cloud plotting was used for processing of results of the enrichment analysis.

## 2.8 "Quantitative Insights into Microbial Ecology" (QIIME)

QIIME is a bioinformatics pipeline (Caporaso 2010) allowing input of raw DNA sequences from sequence protocols such as illumina. QIIME includes automated demultiplexing, OTU picking and phylogenetic placement.

Work steps were conducted by an in-house pipeline (Stecher 2015).

## 2.9 Similarity Percentage (SIMPER) analysis

The SIMPER method (Clarke 1993) uses the Bray Curtis dissimilarity matrix for assaying the dissimilarity between two objects. Used in this study was the "Community analysis package 5" of the Pisces conservatorium:

(http://www.pisces-conservation.com/caphelp/similaritypercentages%28simper.html)

The Bray Curtis dissimilarity matrix allows determining the dissimilarity between two different input files based on the counts of each sites.

Bray Curtis dissimilarity is defined as follows

$$\boldsymbol{BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}}$$

with $S_i$ and $S_j$ representing the total number of specimens counted at each site and $C_{ij}$ the number of common specimens of two sets.

## 2.10 Enrichment Analysis

Enrichment analyses allows for identification of entries over-represented in a larger sample. In the study the analysis was conducted by statistical test of hypergeometrical distribution followed by correction of the received p-values with the Bonferroni correction.

The hypergeometrical distribution is used to determine if subpopulations are over- or under-represented in a sample.

Hypergeometrical distribution is defined as follows

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

with N representing the population size, K the number of successes in population, k the observed successes in n draws

# 3 Results

## 3.1 Environmental characterization of sample sites

Environmental data gathered from the sample sites of the three regions during the two expeditions were used to describe the three geographical regions in regards to their environmental conditions and the availability of nutrients. The following figures 1 to 9 each describe one environmental factor such as temperature, salinity, oxygen, fluorescence, silicate, phosphate, nitrate, nitrite and ammonium. Greenland contained eight *stations* (St.) (St510 trough 517) spanning a distance of 370 kilometers in the environmental charaterizations while the two other regions contained only the five stations that were used in the remainder of the study. Greenlandic region of the Disko Bay in the left column showed the greatest maximal depth with 570 *m.* (*meter*). The eight stations were indicated by the vertical lines from St510 on the left side of the column through the Vaigat Strait into the Disko Bay through to St517 on the right side of the column. The five stations of Norway (St001 to 005) covered a distance of 120 km. The sampled sites of the Norwegian region followed the Sortlandssoundet between the Islands Langøya and Hinnøya with St001 being located north of the soundet and St005 at the western end of the soundet. Maximum water depth in the region was located at St004 with a depth of 180 *m.* The sampled Swedish region was located in the fjord system west of the Island Orust with the St027 to 031 covering a distance of 70 km. First Swedish sample site St027 was located at the mouth of the fjord system near the Island Tjörn with the remaining stations being located along the fjord system northbound with the final St031 ending west of the Byfjord. Maximum water depth in the region was observed near the first site, outside the fjords exceeding 30 m of depth. This depth was not reached in the fjord system proper.

### 3.1.1 Temperature

For the most part the temperature in the waters in the Greenland region lay between 0 and 5 °C (figure 6). Notable exemptions were the influx of cold water from the Ilulissat glacier between St513 and St514 and warmer waters from the station St515 towards the station St516. In this area the temperature rose up to 10 °C to a depth to around 17 m, presumably influenced by warmer, southern currencies entering Disko Bay.

The temperature profile of the Norwegian Lofoten is displayed in the middle column. Sampled sites were displayed by the vertical lines with the first St001 north of the Sortlandssound on the left side of the column to the fifth St005 on the right side. The maximal water depth was reached at around 180 m while the shallowest depth measured around 40 m. The lowest measured temperature was 6 °C at a depth around 120 m between St004 and St005, while the highest temperature was measured in the surface waters of the same sampling sites with temperatures reaching well above 10 °C. This warm water was displayed in horizontal layer with a defined thermocline around 10 m of depth. Aside from

the two extremes the temperature in the lower depths was uniformly around 6 °C, while the waters in the first 30 meters showed a gradient from 6 °C towards the measured maximum.

Highest measured temperature in the Swedish region reached 20 °C within the upper 10 m. of the water column of the first station St027. The same station also contained the lowest recorded temperature around 6 °C, outside the fjord system, around a depth of 25 meters. The remaining four stations lay inside the fjord system west of the Island Orust resulting in shallower water between 15 and 26 meters and a uniform temperature throughout the entire water column around 18 °C.

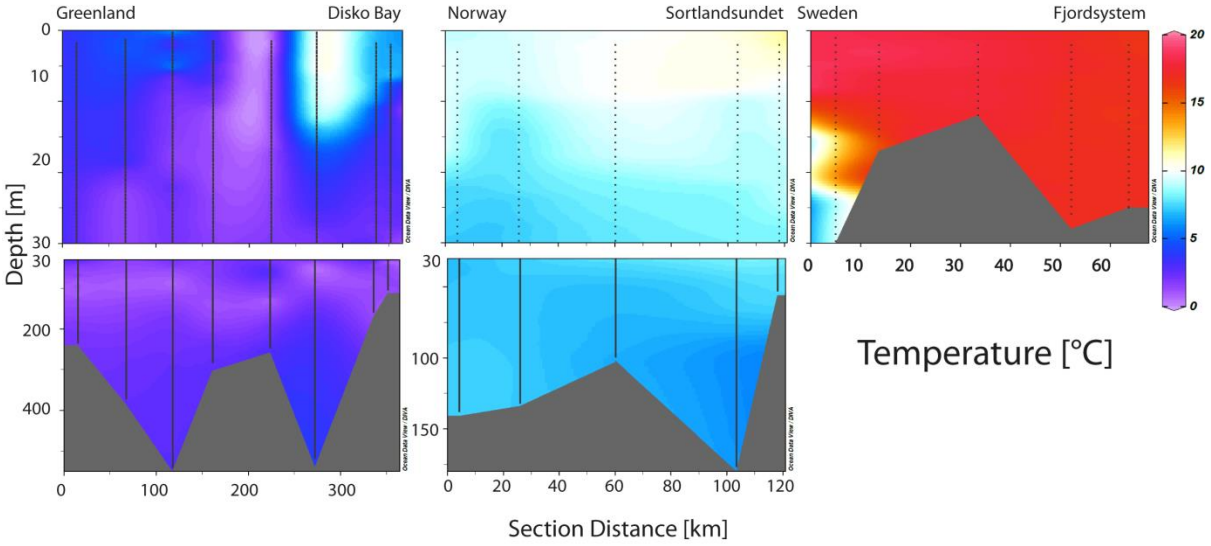



***Figure 6:*** *Temperature characterization of the three sample regions of the Greenlandic Disko Bay, the Norwegian Sortlandsound and the Swedish fjord system east of the Island Orust. The three plots in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden. Plots were created in Ocean Data Viewer (ODV). The figures themselves contain the three samples regions in individual columns from Greenland on the left, Norway in the center column and Sweden on the right.*

### 3.1.2 Salinity

The waters of the Disko Bay on the left column showed a major difference in salinity (figure 2) between the higher and lower water layers. Within the first 30 m the salinity varied strongly throughout water depth as well as laterally between sample sites, between 29- and upwards of 34. From 30 m downwards the salinity reached 33 stayed relatively consistent around that concentration.

The waters of the Sortlandssound displayed less fluctuation in regards to salinity than the waters of the Disko Bay. Throughout the shallow as well as the deeper waters the salinity stayed consistently around 34. Only the deeper waters around St001 showed a slightly higher salinity due to the influx of more water with higher salinity from the polar sea.

The highest variance in salinity yet was measured in the Swedish sampling region. The waters of the fjord system were distinguished by a distinctive stratification of a less saline

upper layer atop a layer of higher salinity. Outside the fjord system around St027 the salinity changed from 28 at a water depth of 19 m to 33 within only 6 m. The recorded salinity of 33 marked the highest measured salinity in the region. In the more upstream regions of the fjord the only occurrence of salinity greater than 28 was observed at a depth around 25 m in the final two sites.

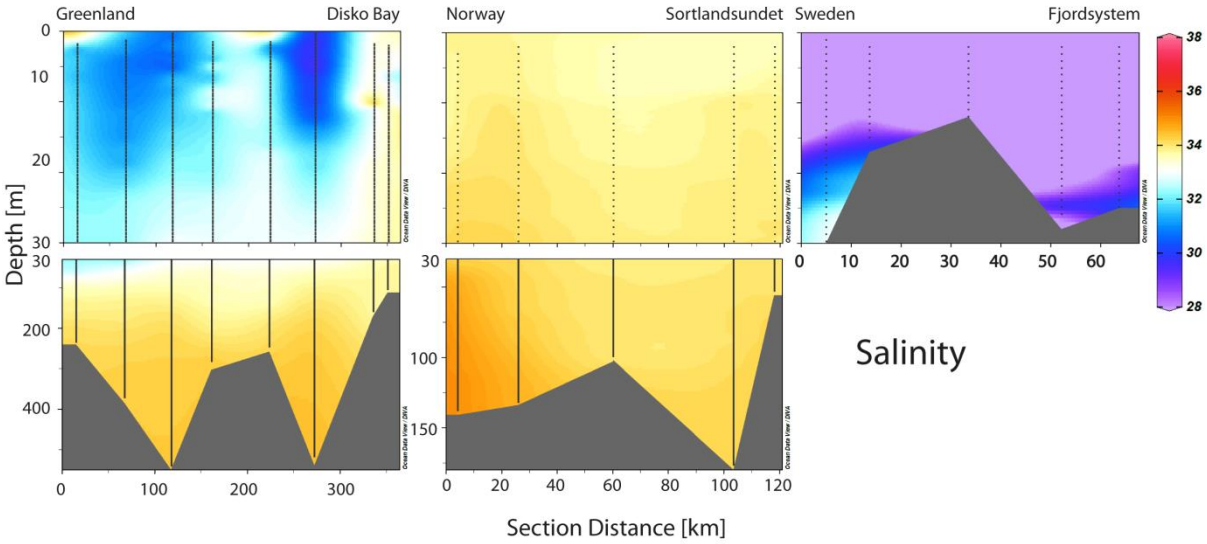



***Figure 7:*** *Salinity characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

### 3.1.3 Oxygen

The waters of Greenland showed in comparison with the other two regions the highest oxygen concentration (figure 3). In the surface waters (0-3m) to a depth to 30 m the measured dissolved oxygen reached 10 ml/l at serval sampling sites and at several water depths. The lowest concentration of oxygen in the surface water was measured at St515 and St516 with a locally limited decline in oxygen to 4 ml/l at a depth of 12-13 meters. In the deeper water of the region, in water depth between 30 and 50 meters the concentration of oxygen remained around 8 ml/l. The deeper water layers contained an oxygen concentration around 6 ml/l with a very small localized spot of lower oxygen concentration around 3ml/l at a depth of 310 m at sampling site St512.

For the waters around the Norwegian region the oxygen concentration was distributed fairly evenly throughout to upper and lower layers with a concentration around 6 ml/l. The only deviation was a local decline at the deepest point of the region at St004 where the oxygen concentration reached 5 ml/l.

In the Swedish region above 10 m water depth at St027 at the mouth of the fjord system the oxygen concentration was determined at 6 ml/l. Within the fjord system this oxygen concentration gradually dropped from 10 m at St029 to 20 m at St031 in a sharply defined stratified layer. The bottom of the fjord around St030 as well at the St028 contained a

concentration of 2 ml/l thereby being classified as hypoxic. Outside of the fjords in the Kattegat at St027 the oxygen concentration decreased even further to anoxic from a depth of 13 m downwards.

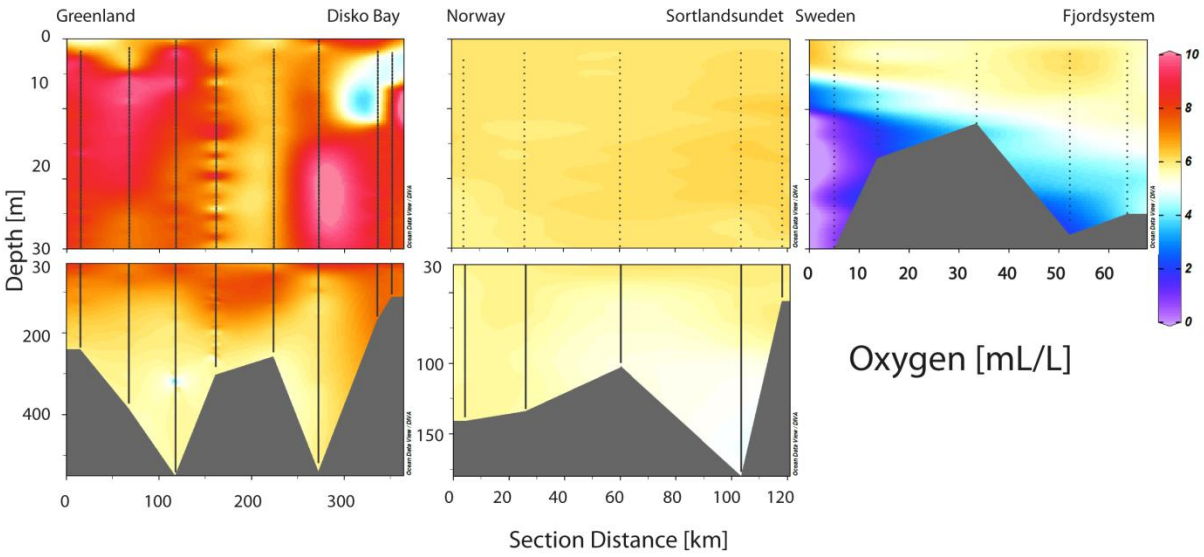



***Figure 8:*** *Oxygen characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

### 3.1.4 Chlorophyll a concentration and distribution

The characterisation of chlorophyll a content (figure 9) shows the highest occurrence in the Greenlandic Disko Bay. The shallower waters of Greenland featured great locally variating occurrence of fluorescence. Significant local maxima representing a subsurface chlorophyll maximal layer were located in the water columns of the St511 between 8 and 20 m and the St512 from the surface to a water depth of 10 m. In those areas the measured chlorophyll a varied between 4 - and 6 mg/m³ and locally reached even 7 mg/m³. These measured values stood out easily as the highest over all three regions. Another area of increased chlorophyll a content in Greenlandic waters was found waters between the stations St515 and St516 at a depth of 20 – 30 meters and around the St517 at a depth between 10 to 25 m. There the chlorophyll was determined between 1 and 2 mg/m³. Aside from the described two areas the shallower waters of Greenland did not display measurable chlorophyll.

The sampled surface waters around the Norwegian Sortlandssound displayed a diffuse background presence around or slightly below 1 mg/m³. Two localized maxima were determined in the waters of St001 were the chlorophyll a concentration reached 4 mg/m³ at a depth of 15 m and at St004 between 10 - and 22 m depth where the chlorophyll a reached 3 mg/m³.

Similarly to the Norwegian waters the tested Swedish sample region was shown to display general background fluorescence around 1 mg/m³. Two localized spots with increased

chlorophyll a concentration up to 4 - and respectively 3 mg/m³ could be found in the water column of first and fourth stations St027 and St030.

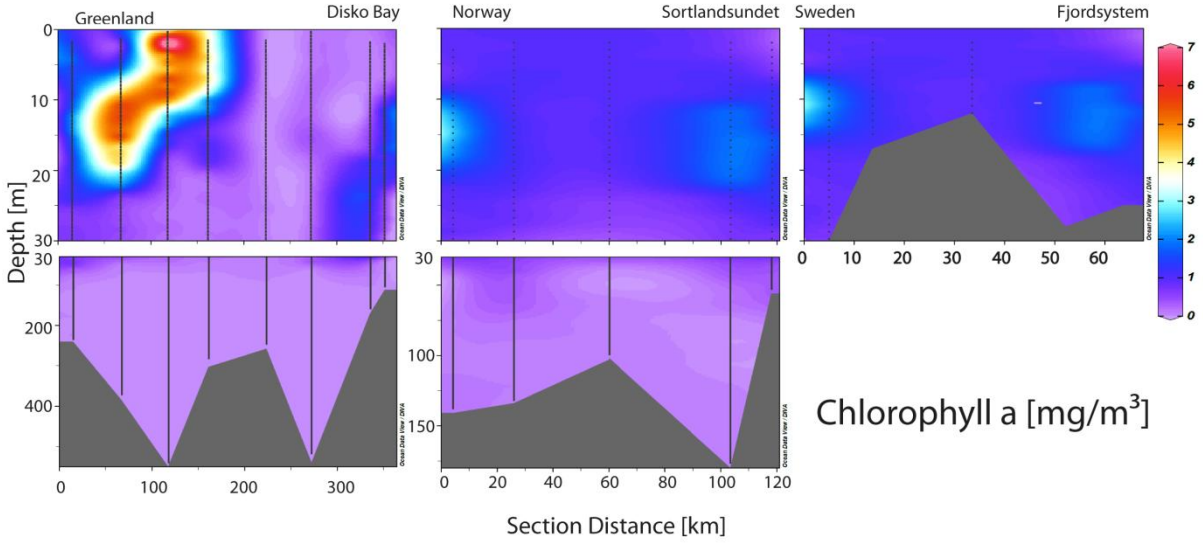



***Figure 9:*** *Chlorophyll a characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

### 3.1.5 Nutrients

**Silicate**

The waters of Greenland displayed virtually no dissolved silicate (figure 10) at all through the entire water column. The first 30 m of the Norwegian water were similarly nearly free from dissolved silicate. In the deeper layers the silicate concentration increased gradually to around 4 µM for all St. with the exception of the fourth St004. As St004 covered greater water depth up to 180 m the gradual increase of the silicate continued to a final concentration of 7 µM at 180 m of depth. Unlike the Norwegian waters the Swedish region did not display a vertical but a lateral increase in measured silicate. The highest determined concentration was outside the Byford at St027. Here the nutrient contends went from around 7 µM at the water surface to a concentration to 23 µM at 30 m. All sequential stations in the fjord proper displayed significantly less dissolved silicate. St030 was determined to contain a concentration of 16 µM which marked the second highest measured silicate concentration after St027. At the head of the fjord around St031 almost no silicate was detectable.

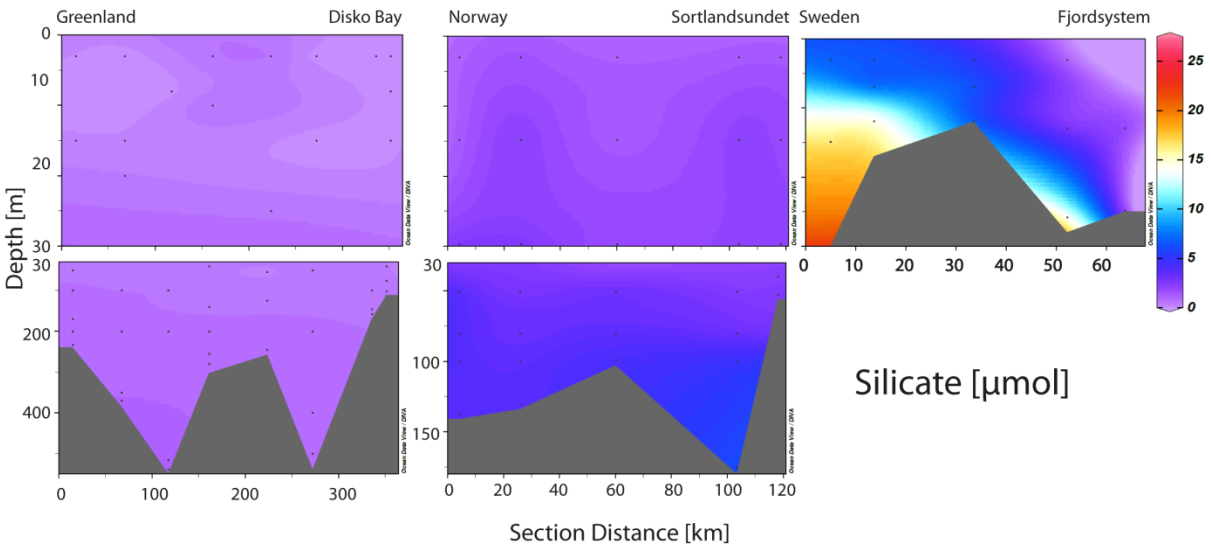



***Figure 10:*** *Silicate characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

## Phosphate

Neither of the entire Swedish and Norwegian regions did contain any measurable phosphate (figure 11). The water column of Greenland however displayed a range of different phosphate concentrations. The shallower surface water up to a depth of 20 m contained areas with no discernible nutrient concentrations with the exception of the water columns of stations St513 and St514. Here the phosphate concentration rose to around 10 µM due to influx of terrestrial freshwater. Around 25 meters of depth the phosphate concentration rose starkly up to 15µM in a clearly defined horizontal layer throughout all stations. For the deeper waters between 30 and 180 m the concentrations varied between 3 and up to 10 µM. From 180 m onward another sharp increase in phosphate concentration was observed over the entire course of the region. With rising water depth the phosphate contend rose steadily. In the two deeper trenches near the stations St513 and St516 the maximum nutrient was observed with 19 and 14 µM respectively.

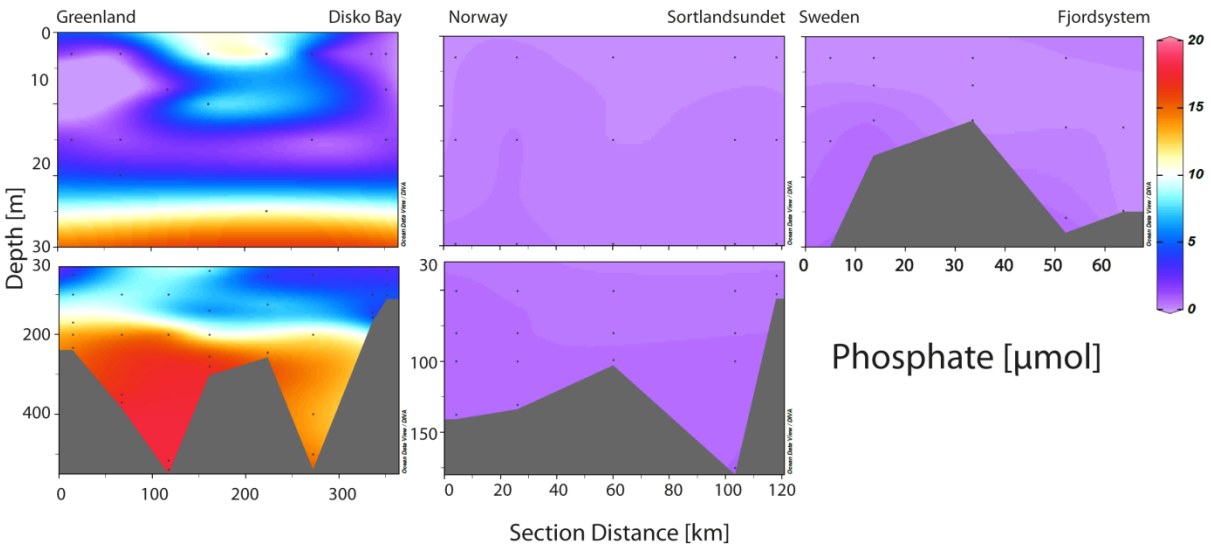



***Figure 11:*** *Phosphate characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

## Nitrate

The water column of Greenland showed a range of different nitrate concentrations (figure 12). The upper 30 m showed locally enriched nitrate concentrations of 12 µM at St513 mirroring the in figure 6 described influx of phosphate from the Ilulissat glacier. In the water between 10 and 20 m of depth most areas displayed little to no nitrate. Only the areas below the observed high nitrate concentration in the surface waters showed a measurable concentration between 5 and 8 µM of nitrate. With increasing water depth the amount also increased up to an evenly observed concentration around 15 µm from 200 m onwards. The transition from a lower, variable nitrate concentration to a higher and evenly distributed concentration of nitrate occurred along a sharp transect at a water depth between 150 and 180 m of water depth. In the surface waters around Norway only low concentrations of nitrate were measured in the first 30 m. The highest observed nitrate concentration was found in St002 at 30 m of depth with a concentration of 10 µM. Comparably to the observations made in the deeper waters of Greenland, the measured nitrate concentrations increased with rising depths. Highest observed concentrations were around 30 µM around 150 meters of depth. Unlike to the previous regions barely any nitrate was measured in Sweden as the highest concentrations did not exceed 2 µM of dissolved nitrate.

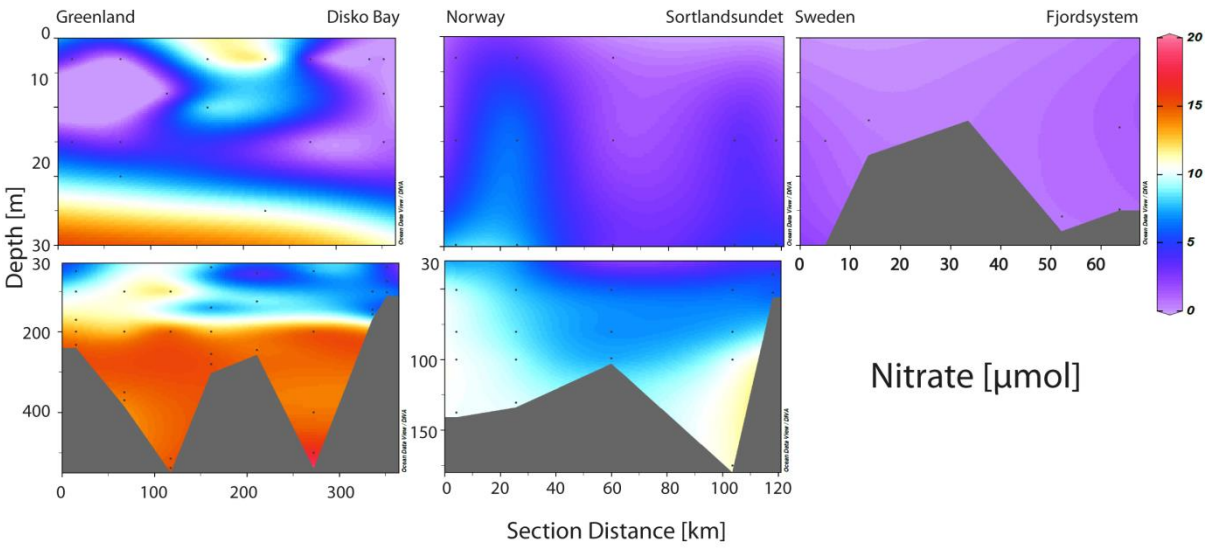



***Figure 12:*** *Nitrate characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

**Nitrite**

Unlike the overserved nitrate concentrations the waters around the Disko Bay showed only a small nitrite concentration (figure 13). In the entire water column the measured nitrite concentration did not exceed 0.25 µM. This maximal concentration occurred in a horizontal layer around 150 meters of depth throughout the sampled region. For the last three stations (St515, St516 and St517) the elevated nitrite level also expanded into shallower waters up to a depth of 30 meters. The waters around Norway in terms of nitrite concentrations were similar to the Greenlandic water, as for most of the water did not contain any measurable nitrite. The highest observed concentration was around 0.25 µM but unlike the waters of Greenland this concentration was measured in a vertical transect in the waters of the third station St003 downwards from 40 m of depth. The sampled region of Sweden showed the highest overserved concentrations of nitrite exceeding locally 1.5 µM of dissolved nitrite in the deepest shoals of the second and fifth sample sites (St028 and St031). From the observed maxima the concentrations decreased sharply to around 0.75 µM within 10 meters of depth. As no measurement of nitrite concentration was determined at the time of the expedition the displayed values above 15 m the displayed gradient is an extrapolation.

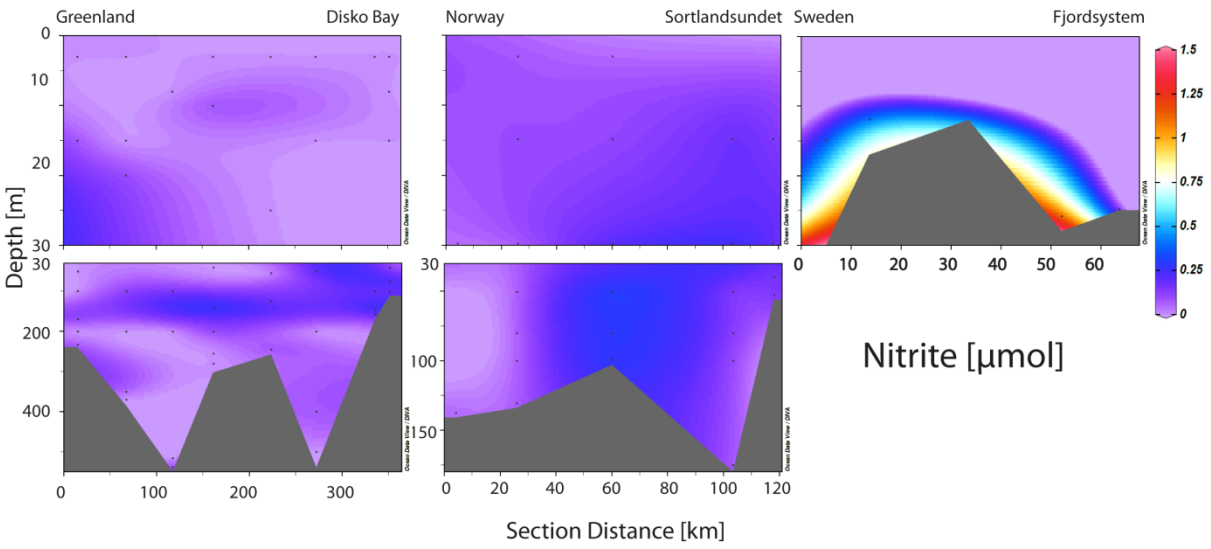



***Figure 13:*** *Nitrite characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

**Ammonium**

The observed ammonium content showed a wide range of different concentrations in the waters around the Disko Bay (figure 14). The highest measured concentration of 3 µM occurred in a localized area around 150 m of depth in the waters of St516. From that local maximum a narrow horizontal layer of increased ammonium concentration expanded through the water columns of St512 through to St516. Elevated levels of ammonium up to a concentration of 2 µM were also detected in the around the sites St510, St514, St515, St516 as well as in St517 in 30 - to 35 m. Beside the described locations of higher ammonium concentration wide areas of the sampled region did not show a measurable concentration at all. This applies especially for the deeper waters exceeding 200 m. The sampled Norwegian sound displayed an overall lower concentration of ammonium as well as lesser localized ammonium gradients. The highest measured ammonium concentration did not exceed 1.3 µM of dissolved ammonium. This peak was localized in the water column of St513 at a depth of 30 m. A narrow horizontal layer of increased ammonium content extended from the observed peak through the water column of St512 and St511. This layer however was only visible in the upper plot, showing in the first 30 meters of the water column. The remaining water column did not show any measurable ammonium content. The Swedish waters showed two very localized ammonium maxima reaching a concentration of over 2 µM of dissolved ammonium in the deepest trenches of St030 and St027 around 30 m. The maximum at the site St027 however was only an extrapolation as no ammonium content was experimentally determined at the time of the expedition beyond a water depth of 15 m.

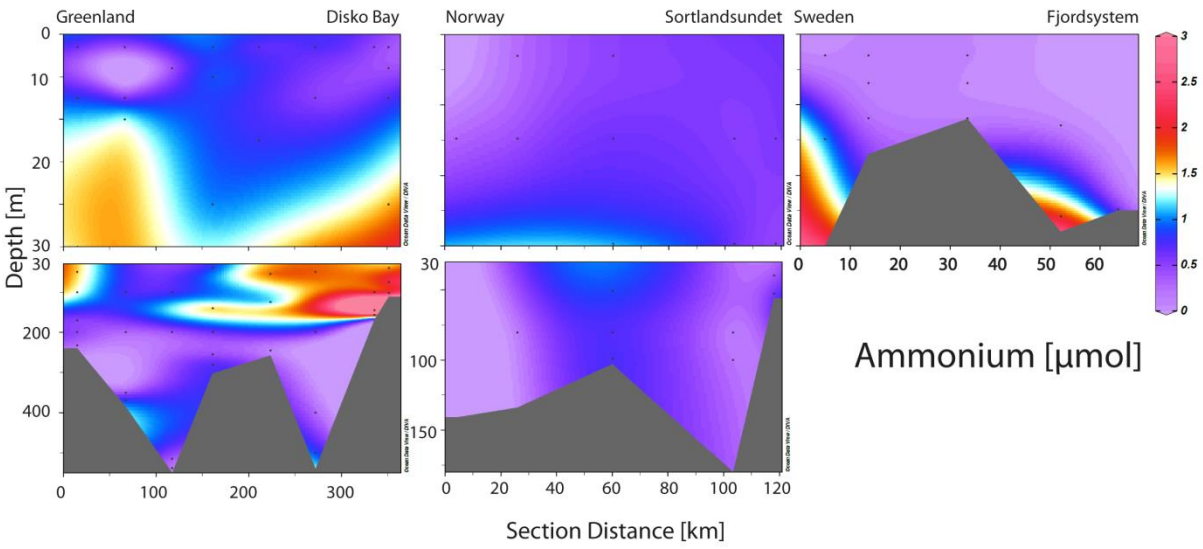



***Figure 14:*** *Ammonium characterization of the three regions and the contained sampling sites. The three pictures in the top row show the first 30 meters of the water column. The second row shows the rest of the water column depending on the maximal water depth of the sample site. Regions depicted from left to right are Greenland, Norway and Sweden.*

## 3.2 Sequence retrieval rates throughout the subsequent work steps of proceeding experiment

As the constructed cDNA libraries contained functional information the libraries were processed through Trinotate pipeline for functional annotation of the potential transcripts. In this metagenomics approach about 550,000 transcripts with *eukaryotic orthologues groups* (COG) identifier were received.

The overlap between the two partial results of assigning species to contigs and matching the transcripts of the cDNA to a metabolic pathway resulted in around 100,000 entries which contained both a through BLAST assigned species identity as well as a functional annotation following the COG system.

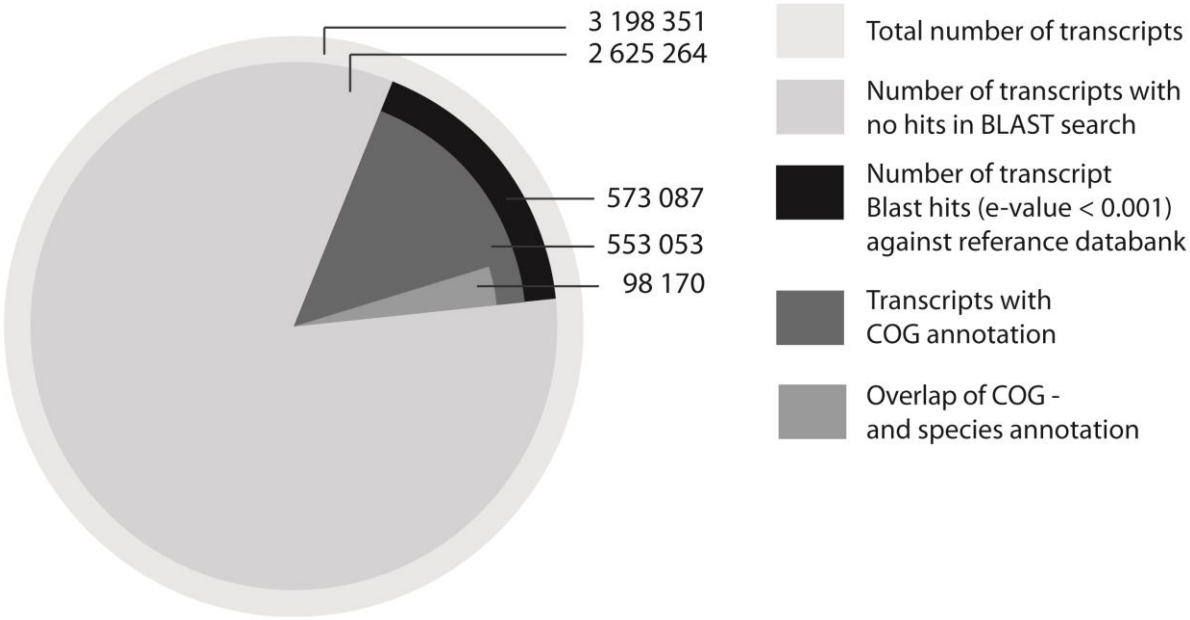



***Figure 15:*** *Sequence reduction along the study and retrieval rates. The starting assembly contained 3.198.351 contigs in total. After BLAST against Reference databank 573.087 hits remained (only the best BLAST hit per contig was picked). The BLAST searches were conducted for multiple e-values, depicted are the results for the e-value of* $e^{0.001}$*. The selected BLAST hits resulted in 553.053 transcripts. Of those transcripts 98.170 had COG annotations and were attributed to a phylum or species.*

## 3.3 Clustering of transcripts intra- and inter sample region

Similarities and dissimilarities between sample sites and on a larger level the three geographical regions, can foreshadow eventual findings of the biodiversity and metagenomic functional investigations. Clustering of transcripts was carried out twofold, once with *non-*

*metric multidimensional scaling (NMDS)* and secondly DESeq 2 package for R. The results were visualised with a heat map style plot.

In the NMDS Greenland displayed the most differences between its samples, resulting in a large assigned cluster. The Swedish region seemed to be contrary as the individual sample sites displayed very little dissimilarity resulting a in a tightly fitted cluster. The sampled stations of Norway lay in between as four of the five sites were clustered tightly together and the final fifth site, St001 outlying on its own and thereby influencing the displayed Norwegian cluster size. Environmental factors derived from the Environmental characterization of the regions (figures 6 – 14) were added as vectors to the NMDS in order to visualize the eventual influence of environmental factors on biodiversity.

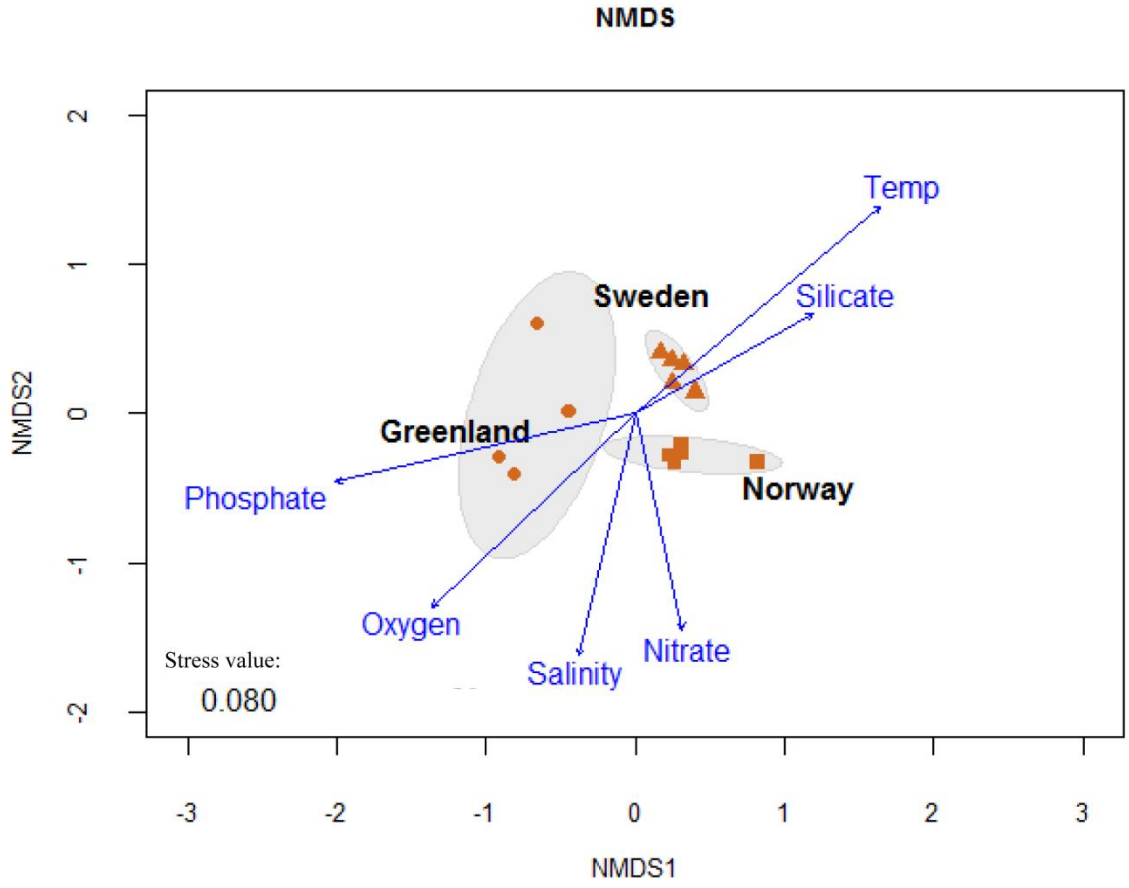



***Figure 16a:*** *Non metric multidimensional scaling plot attributed to the three regions. Environmental parameters are projected as vectors onto to coordinate system. Figure is presented in a nonmetric multidimensional system. The three regions form separate clusters and each cluster differed in circumference depending on the intra-region similarity between the samples. Greenlandic stations are represented by circles, Swedish stations by triangles and Norwegian stations by squares. Full report file of NMDS available in supplemental data.*

The heat map (figure 16b) confirmed the findings from the NMDS plot. The first station of the Norway (St001) showed increased dissimilarity from other Norwegian sampling sites as well as from other regions and formed its own branch. The remaining four Norwegian sites formed their own cluster. All five Swedish sites formed their own cluster in between the Norwegian and Greenlandic cluster. Sample sites from Greenland formed two separate clusters with were closely related. The sites Greenland St11 and St512 formed their own cluster apart from the remaining three Greenlandic samples.

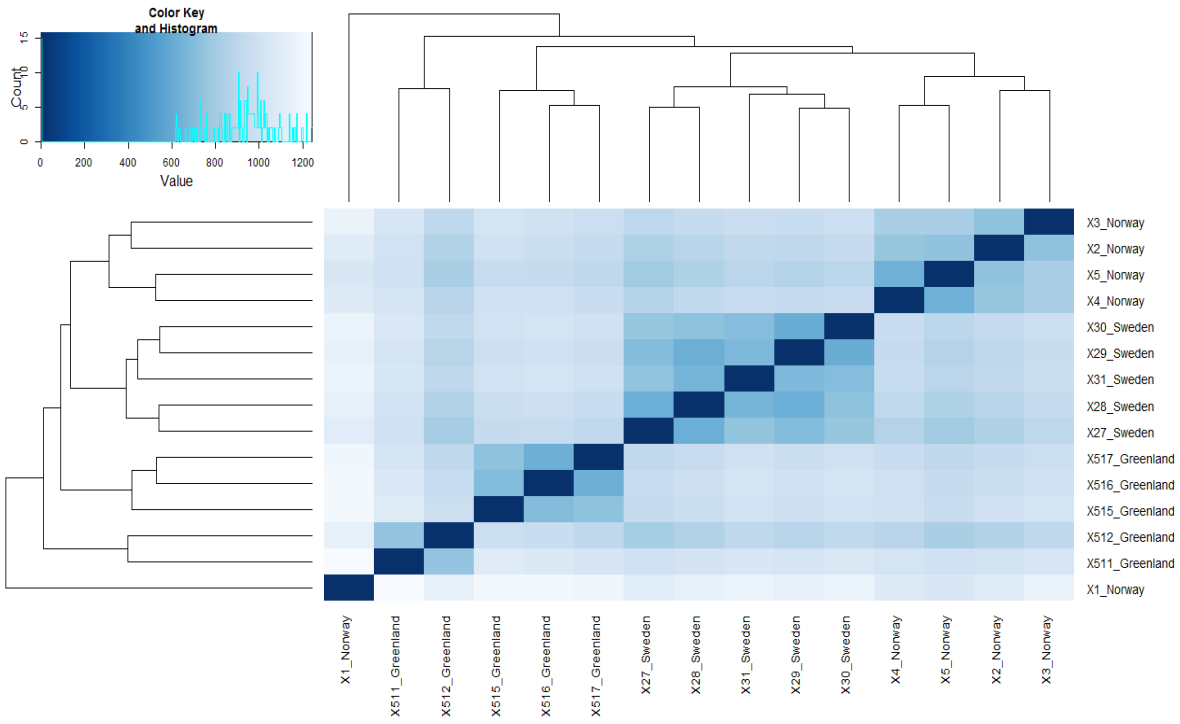



***Figure 16b:*** *Clustered heat map of similarity between transcript profiles based on observed contigs in the assembly. Increased similarity indicated by darker blue colour.*

## 3.4 Biodiversity

Biodiversity was determined in two separate ways. Firstly by QIIME software based on 28S rRNA LSU amplicon sequences obtained from the 15 sample stations and secondly by BLAST of the assembled sequences against the constructed reference databank.

## 3.5 28S rRNA LSU approach

28S rRNA amplicon sequences optained from the 15 sample sites spread acrous the three regions were analysed with the QIIME software. The optained *orthologus taxonomic clusters* (OTU) were refined by phylogenetic placement with reference trees and then depicted on the regional level as well as intra region on a sample site level.

***Regions overview***

The region of Greenland contained the most overall OTUs with 90,529, followed by Norway with 25,555 and Sweden with 20,249. Greenland displayed the least amount of unassigned OTUs around 5 % while in both Sweden and Norway around 20 % of OTUs were not assigned to a taxonomical clade. Dinophyta were represented the strongest in Greenland (~35 %) followed by Sweden with ~30 % and ~ 10 % in Norway. Around 45 % of all OTUs in Greenland finally were assigned to the phylum of Bacillariophyta. This phylum was represented in with 30 % Sweden and the strongest in Norway with a share of over 70 % of all OTUs.

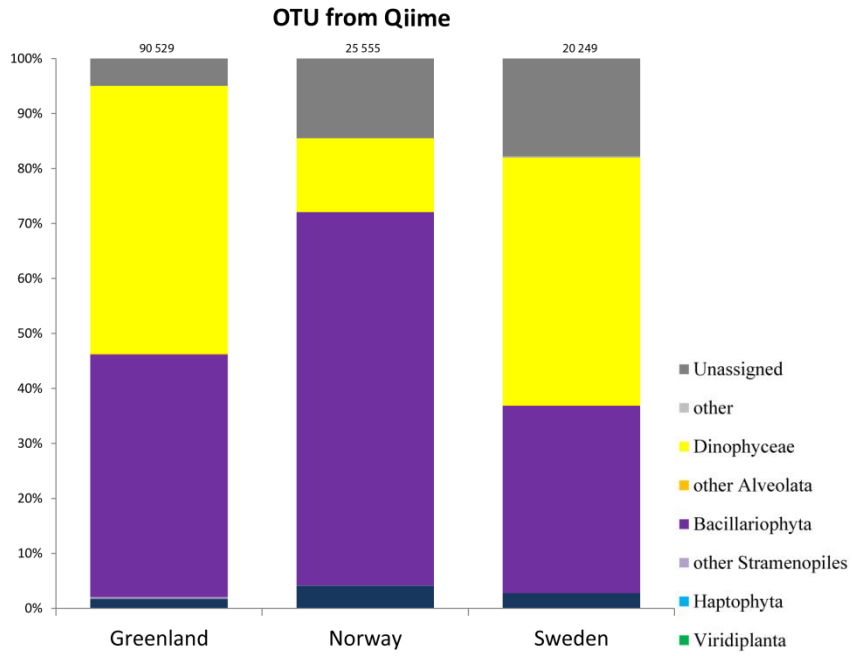



***Figure 17:*** *OTU abundance derived from phylogenetic placement. Displayed are abundances of assigned classes of the three regions. Regions displayed from left to right are Greenland left, Sweden in the middle and Norway on the right side. Number of transcripts assigned to region given on top of the stacks. Species assignments are given in percentage.*

### *Greenlandic region*

The five sites of Greenland contained a range of different transcript counts. St511 contained the fewest OTUs with 9,002 followed in ascending order by St512 with 11,427, St515 with 20,268, St517 with 20,502 and St516 contained the most OTUs with 29,530. The proportion of unassigned transcripts was less than 5 % for stations St511, S512 and St515 while the stations St516 and St517 contained around 10 % of unassigned transcripts. Transcripts which did not fall in the categories “unassigned”, “Alveolata”, “Stramenophiles”, “Haptophyta” or “Viriplanta” were grouped into the category “other”.

The phylum of Dinophyta were represented in St511 and St512 around 10 %. The other three stations contained larger fractions of Dinophyta around 30 % for St516 and St517 and around 80 % for St515. Besides Dinophyta only few other Alveolata were detected, St515 contained the only discernible fraction around 1 %. Bacillariophyta made up a major portion in St511 and St512 covering more than 85 % of all OTUs. In stations St516 and St517 around 40 % of all OTUs were allotted to Bacillariophyta and St515 contained the smallest amount with less than 10 %. Besides Bacillariophyta only few other Stramenophiles were detected, resulting in only small fractions around 1 % in St516 and St517. Throughout all stations neither Haptophyta nor Viridiplanta were detected.

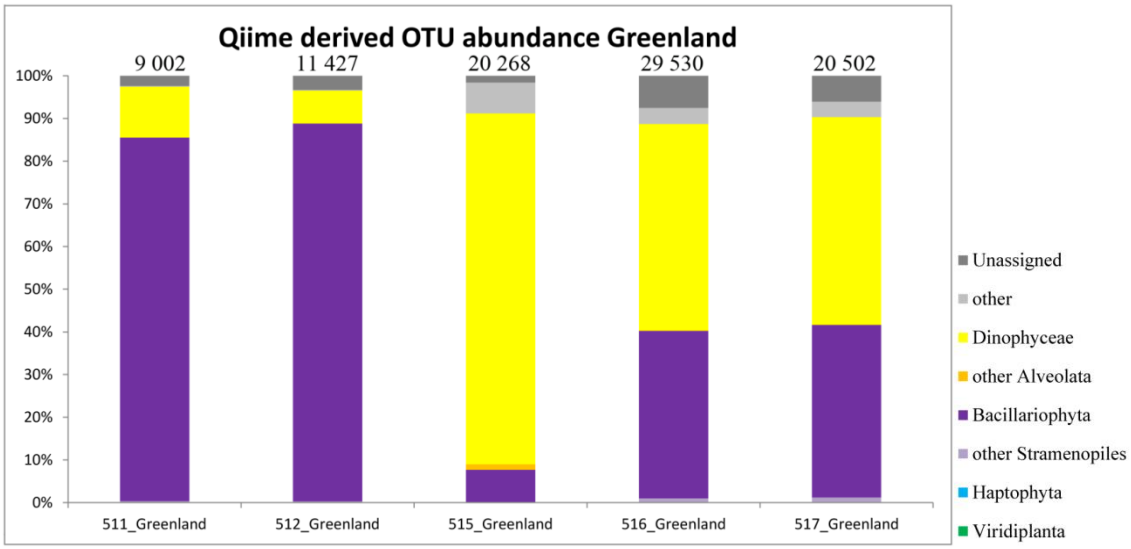



***Figure 18:*** *OTU abundance derived from 28S rRNA phylogenetic placement. Displayed are abundances of assigned classes of the five sampling site in the region of Greenland. Regions displayed from left to right are Greenlandic St511 to St517. Number of transcripts assigned to region given on top of the stacks. Species assignments are given in percentage.*

### *Norwegian region*

The Norwegian sites yielded fewer OTUs per sites in comparison to the Greenlandic sites. The first St001 contained only 830 OTUs followed by 4,483 OTUs in St002, 4,889 in St005. St004 was scored with 5, 992 OTUs and the third St003 contained the most OTUs with 9,361.

The proportion of unassigned transcripts was around 5 % for stations St002 and St003 while the remaining displayed a larger number of unassigned transcripts. Stations St004 and St001 contained proportions around 20 % and in St005 more than 30 % Sequences were assigned to the category of unassigned transcripts. The "other" category was represented the strongest in stations St001, St002 and St 005 with quotas above 10 % of all OTUs. The second and third stations St002 and St003 contained only marginal fractions of "other" sequences around 1%. The phylum of Dinophyta was represented in second and third stations around 5 %. The other three stations contained larger fractions of Dinophyta between 10 % for the first station St. 001 and more than 15 % for fourth and fifth stations St004 and St005. Besides Dinophyta only few other Alveolata were detected, no station contained a discernible fraction. Bacillariophyta made up a major portion in stations St001 to St004, allotting for half of all OTUs in the first and fourth station and reaching over 85 % coverage for the second and third station. The fifth station St005 contained the smallest fraction of Bacillariophyta which amounts to 30 % of OTUs. Similarly to the situation of other Alveolata besides Dinophyta no Stramenophiles besides Bacillariophyta were detected. Throughout all stations neither Haptophyta nor Viridiplanta were detected.

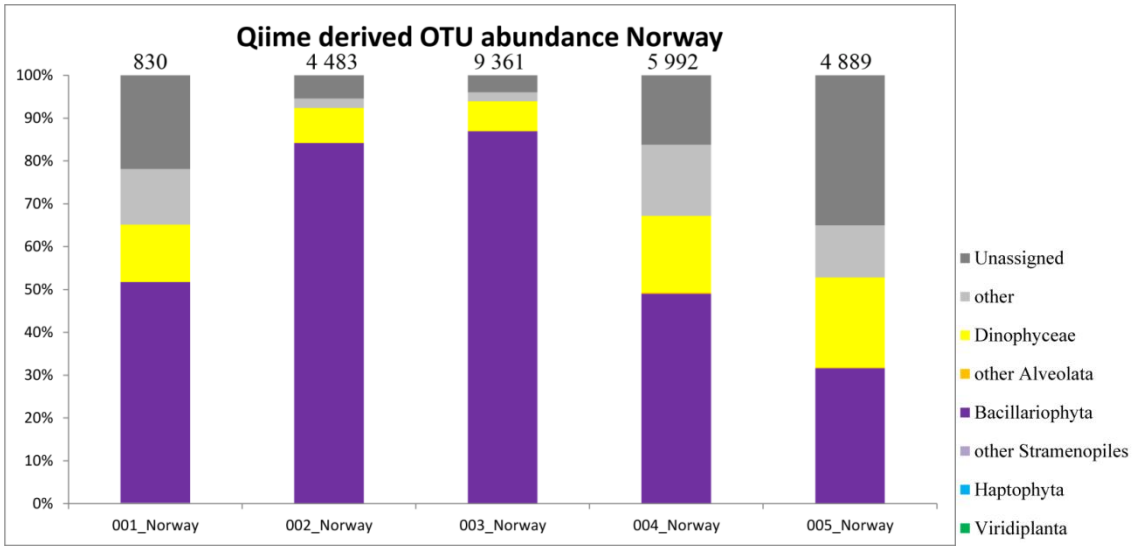



***Figure 19:*** *OTU abundance derived from 28S rRNA phylogenetic placement. Displayed are abundances of assigned classes of the five sampling site in the region of Norway. Regions displayed from left to right are Norwegian St001 to St005. Number of transcripts assigned to region given on top of the stacks. Species assignments are given in percentage.*

### *Swedish region*

The transcript count of the Swedish stations yielded in comparison fewer transcript to the Greenlandic stations and comparable read numbers to the Norwegian region. Second St028 contained the fewest OTUs with 2,985. Stations St027, St029 and St030 contained with 4,071, 3,942 and 4,070 respectively a comparable number of OTUs. Final station St. 031 contained the highest number of OTUs in the region with 5,181. The proportion of unassigned transcripts varied throughout the sample region from 30 % in the first two stations

St027 and St028, to 15 % for St029 and 10 % and less for the last two stations. The sequences allocated to category “other” followed along the same decline to an extent. While St027 and St028 contained fractions around 5 %, the portions were diminished to 2 and 3 % for the remaining St029, St030 and St031. The phylum of Dinophyta represented between 42 – and 45 % of the total OTU count in the stations St027, St028 and St030. The third station St029 contained a larger fraction of 69 % while the final station depicted a smaller fraction of less than 10 %. The transcript count allotted to Bacillariophyta varied between the five stations with content around 15 % for the first two stations, a slightly smaller portion of 7 % in the third station St029 to increased Bacillariophyta fraction in the last two stations of 24 % and 78 % respectably. Throughout all stations of the Swedish region no other Alveolate, other Stramenopiles, Haptophyta or Viridiplanta were detected.

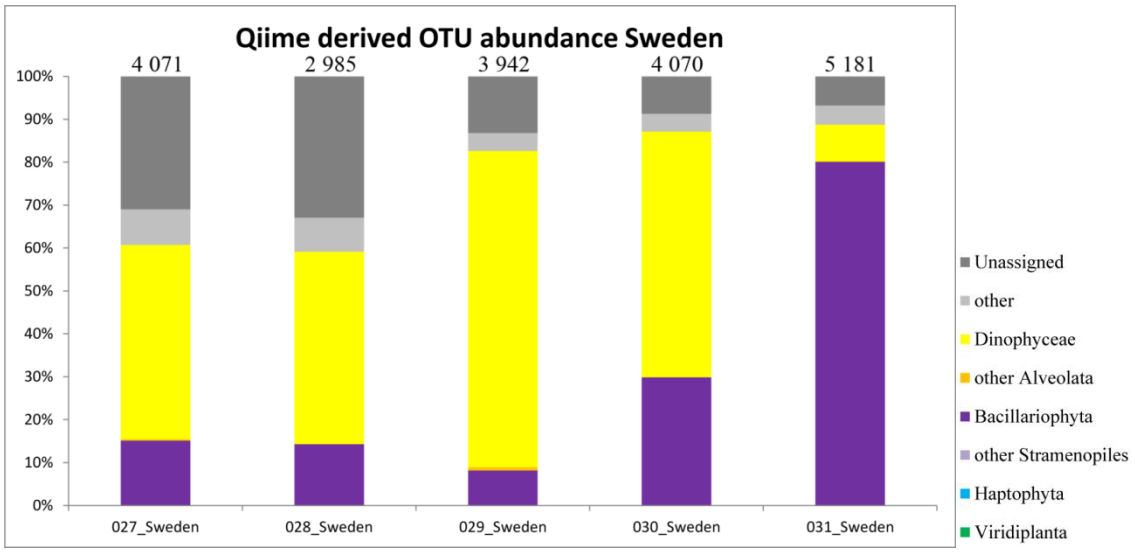



***Figure 20:*** *OTU abundance derived from 28S rRNA phylogenetic placement. Displayed are abundances of assigned classes of the five sampling site in the region of Sweden. Regions displayed from left to right are Swedish St027 to St031. Number of transcripts assigned to region given on top of the stacks. Species assignments are given in percentage.*

## 3.6 BLAST derived OTU abundance

The achieved 28S rRNA phylogenetic placement taxonomic classification served as a benchmark to judge the succes of the transcript derived taxonomy. Similarly to the procedure of 28S rRNA phylogenetic placement the BLAST derived transcripts were clustered in OTUs and plotted firstly inter-regional followed by intra-regional plots.

***Regions overview***

The region of Greenland contained the most overall OTUs with 21,179,013, followed by Norway with 16,219,474 and Sweden with 14,503,106 identified OTUs. The number of unassigned OTUs makes up a minor fraction of all regions with 0.1 % shares in Greenland and Norway and 0.2 % in Sweden. Share of "other" OTU lies at 4.6 % for Greenland. 0.8 % for Sweden and 1.4 % for Norway. Dinophyta were represented in Greenland with 15 %, followed by Sweden with 83 % and 29 % in Norway. Unlike with the 28S rRNA results the BLAST results allowed for an identification of Alevolata besides Dinophyta. The category amounted to 6.1 % for Greenland. 10.7 % for Sweden and 8.6 % for Norway. 64 % of all OTUs in Greenland were allotted to the class of Bacillariophyta. This class made up a minor faction of 4 % in Sweden and 67 % of all OTUs in Norway. Other Stramenophiles amounted to 5 % in Greenland, 1 % in Sweden and 2 % in Norway. Haptophyta were found in small amounts which resulted in fractions of 1 % in Greenland and less than 1% for both Sweden and a Norway. Viriplantae made up 4 % of all sequences in Greenland and around 1 % for both Sweden and Norway.

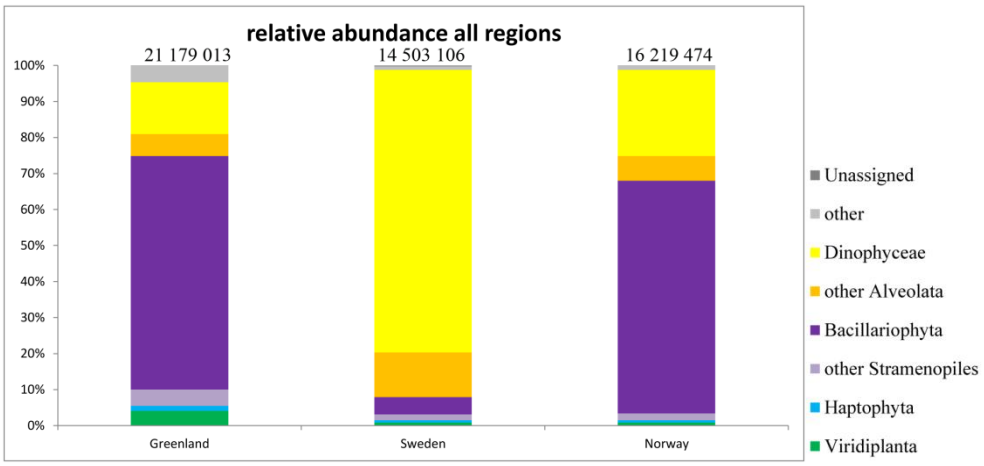



***Figure 21:*** *OTU abundance after BLAST against reference databank. Displayed are abundances of allocated classes of the three regions. Regions displayed from left to right are Greenland left, Sweden in the middle and Norway on the right side. Number of transcripts allotted to region given on top of the stacks. Species allotments are given in percentage.*

***Greenlandic region***

Between the five sites of Greenland we detected a caesura in terms of the detected OTUs. The first two samples St511 and St512 contained 6,557,174 and 6,773,169 OTUs

respectively, while the remaining three sites measured considerably less. The third site St515 contained the fewest OTUs of the region with 1.980.004 and was surpassed in terms of read counts by sites St516 with 2,897,175 and final station St517 with 2.971.491 OTUs.

The proportion of unassigned sequences was less than 1 % for all stations. The category "other" was scored with less than 1 % for the first two sites, 15 % for the third station St515 and around 10 % for the two remaining stations St516 and St517. The phylum of Dinophyta made up a minor fraction in first two stations St511 and St512 around 3 % and 9 %. In the remaining sites the Dinophyta were presented more strongly around 30 %. Other Alveolates made up 1.5 % of the OTUs in St511 and St512 and between 15 – and 12 % in the remaining stations of the region. The phylum Bacillariophyta made up a major portion in stations St511 and St512 covering more than 90 % of all transcripts. The remaining stations contained considerably less Bacillariophyta OTUs, for sites St516 and St517 around 30 % and site 515 only 15 %. Few other Stramenophiles were detected in stations St516 and St517 resulting in only small fractions around 1 %. This phylum was stronger represented in the remaining sites at around 10 %. Haptophyta were detected in small amounts only, amounting to around 1 % in the first two stations and around 2 % in the remaining three stations. Viridiplanta were allotted to less than 1 % for sites St511 and St512 while making up 13 % for the third station and 9 - and 10% for the fourth and fifth station respectively.

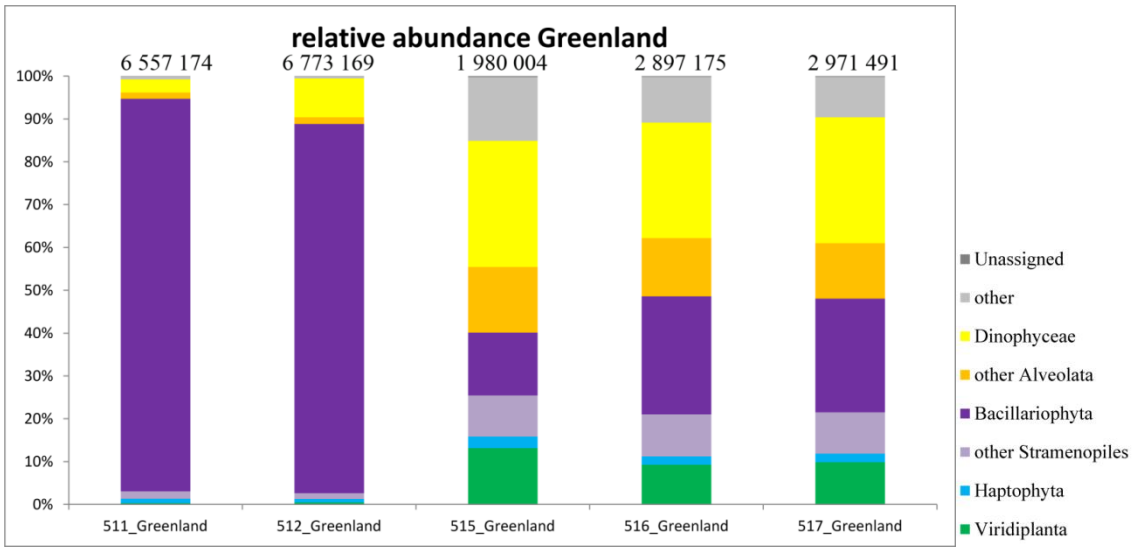



***Figure 22:*** *OTU abundance after BLAST against reference databank. Displayed are abundances of assigned classes of the five sampling site in the region of Greenland. Regions displayed from left to right are Greenlandic St511 to St517. Number of transcripts allotted to region given on top of the stacks. Species assignments are given in percentage.*

### *Norwegian region*

The Norwegian stations yielded in descending order 3,814,089 in the first stations, 3,432,302 in the second, 3,560,302 in the third and 2,792,605 and 2,620,116 OTUs for the fourth and fifth stations respectively. Unassigned transcripts amounted for less than 0.5 % throughout all

five sites. The category “other” was scored higher, coming above 1 % for the stations St512, St.516 and St517 and around 0.5 % for the first and third station. The content of determined Dinophyta varied throughout the region from 7 % for the first site to 24 % in the second, 12% for the third, 39 % for the fourth and 48 % for the final station St005. Other Alveolata besides Dinophyta followed the same distribution through the stations with ratios of 2 % for St001, 4.5 % for second station, 3% for third and sequentially 13- and 16 % for fourth and fifth.

The phylum Bacillariophyta represented the strongest fraction in the first three stations with ratios of 88 %, 67 % and 81 % respectively. The two remaining fourth and fifth stations St004 and St005 contained smaller Bacillariophyta fractions of 42 % in St004 and 29 % in St005. Other Bacillariophyta were detected only in small numbers between 1.5 – and 2 % for the first three stations and around 2 % for the final two. Both Haptophyta and Viridiplanta were scored even lower never exceeding 1.5 % throughout the region.

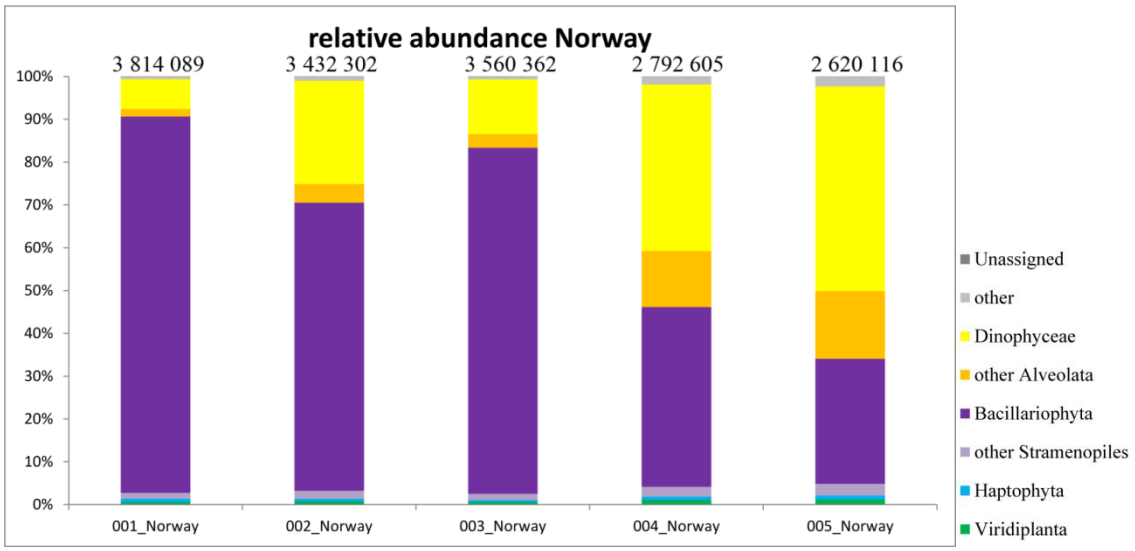



***Figure 23:*** *OTU abundance after BLAST against reference databank. Displayed are abundances of assigned classes of the five sampling site in the region of Norway. Regions displayed from left to right are Norwegian St001 to St005. Number of transcripts assigned to region given on top of the stacks. Species assignments are given in percentage.*

### *Swedish region*

The overall OTU count of the Swedish stations yielded in comparison fewer counts to the Greenlandic stations and comparable transcripts numbers to the Norwegian region. Station St027 contained the fewest transcripts with 2,065,476. From there a gap in OTU count followed to the second lowest count in station St030 with 2,897,433. St028 followed with 2,998,933 OTUs. Remaining two stations achieved over 3 million OTU count with St031 having 3.185.450 and St029 with 3, 355,814 transcripts. Hardly any OTUs were assigned to category “unassigned sequences” as only stations St004 and St005 showed fractions of 0.1 %. The group was represented higher with the first and third stations scoring 0.5 %, second stations scoring 1 % and stations St004 and St005 showing increased fractions of 1.8 – and

2.2 %. Phylum of Dinophyta represented the overwhelming fraction in all stations with fractions of 79 %, 80 %, 82 %, 79 % and 72 % in order. Other Alveolata were scored higher than in previous regions with 8 % for the first St. while the other four stations rated of 12 %. The transcript count assigned to phylum Bacillariophyta was highest in fifth station with 9.5 % followed by the first St. with 7%. Stations two to four contained between 3.5 – and 2 % Bacillariophyta. Other Stramenopiles were scored highest in the first station with 2.5 % while the other stations displayed fractions below 2 %. Haptophyta as well as Viridiplanta were detected only in traces throughout the region never exceeding rations above 1.3 %.

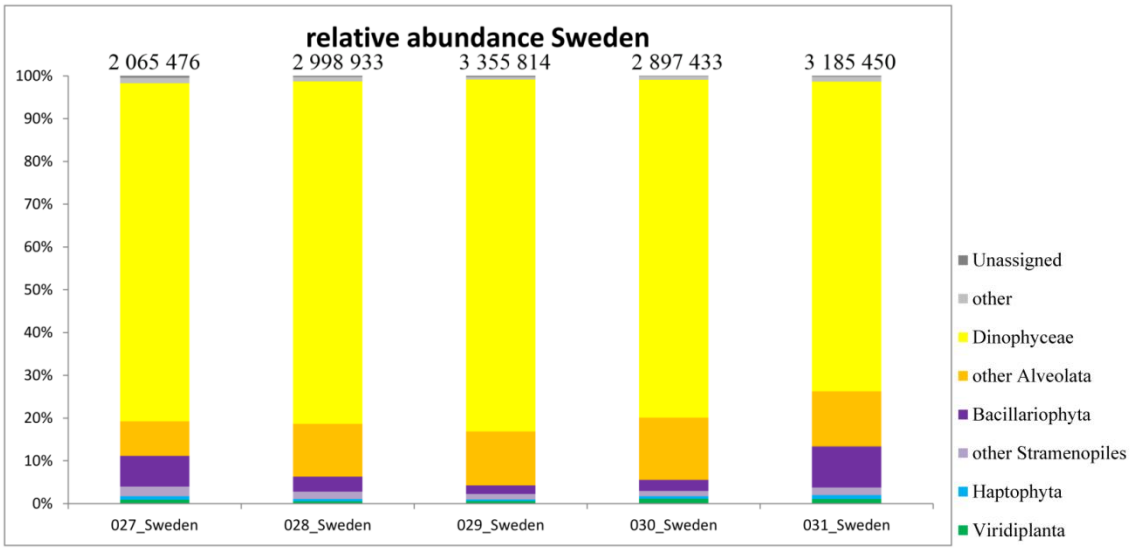



***Figure 24:*** *OTU abundance after BLAST against reference databank. Displayed are abundances of assigned classes of the five sampling stations in the region of Sweden. Regions displayed from left to right are Swedish St027 to St031. Number of transcripts assigned to region given on top of the stacks. Species assignments are given in percentage.*

## 3.7 Relative abundance plots

A method for visualization of the received abundances was the plotting of the transcript derived taxonomy data in a *relative abundance* (RA) plot. First set of RA plots with complete count data of the regions are found in supplemental data (Supplemental figure X). Plots of the first set were visibly over plotted which made interpretation difficult. In addition wide ranges of the plots were sequences not assigned to a taxon, resulting in a white pie chart.

The second set of plots was conceived from the ground up to avoid over plotting by dividing the original count data into the three major taxonomical groups Dinophyta, Bacillariophyta and Haptophyta and furthermore by limiting these subsets to the 10 most abundant species.

### 3.7.1 Relative abundance - plots of Bacillariophyta

First comparison between Greenland and Norway (figure 20) did show a fairly even distribution of the Bacillariophyta. The individual charts formed an approximate arrow shape indicative of an even distribution. In terms of taxa distribution a visibly number of pie charts in the Norwegian upper half of the plot showed large fractions in yellow colour indicative of a strong presence of the genus *Chaetoceros*. The lower Greenlandic half displayed large fractions in green and red, assigned to the genera *Thalassiosira* and *Chaetoceros*.

The second plot between Greenland and Sweden displayed a different pattern. Instead of the even arrow shape in the plot of Greenland and Norway the shape of the Greenland and Sweden was malformed. The upper Swedish half displayed a dense field of smaller pie charts indicative of sequences with a lower expression. It was not discernible from number of individual pie charts whether one of the compared regions contained more Bacillariophyta than the other. In terms of the colouring of the charts no distinguishable pattern was observed therefore the ten most abundance species seemed to be distributed evenly between the two geographical regions.

The third plot between Norway and Sweden displayed a similar pattern to the previously described plot. The amplitude of the charts allotted to the Swedish region seemed smaller in comparison to the other two regions resulting in the uneven arrow shape. In terms of colour the Norwegian region seemed to feature the previously observed elevated higher number of yellow fractions indicative of the genus *Chaetoceros*.

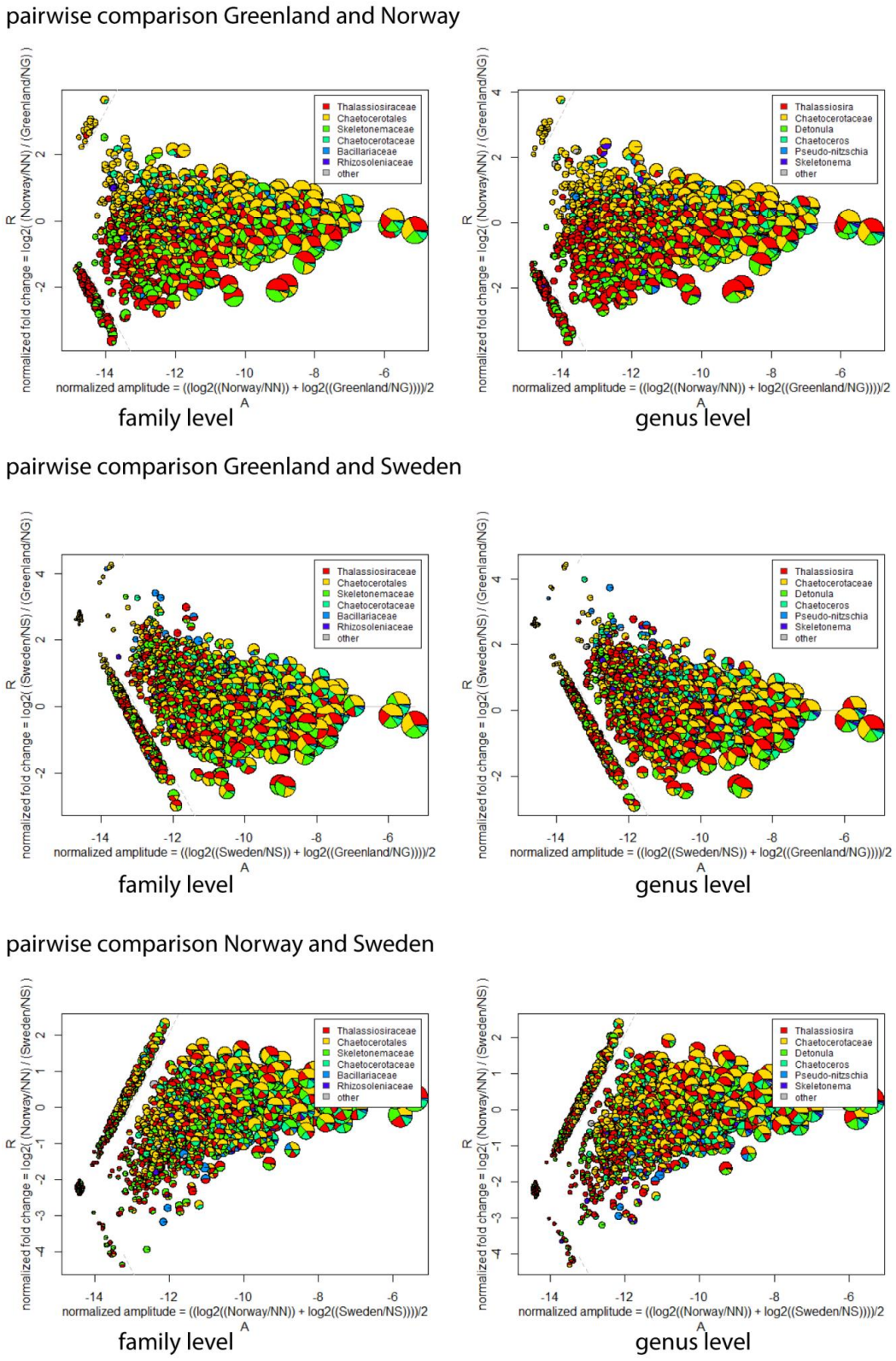



***Figure 25:*** *Pairwise comparison of 10 most abundant Bacillariophyta taxa. Plots of the left depict the "family" level, plots on the right the "genus" level. Paired regions for comparison are from top to bottom Greenland and Norway, Greenland and Sweden and Norway and Sweden. The 10 most abundant genera for the Bacillariophyta are* Thalassiosira, Chaetoceros, Corethron, Detonula, Skeletonema, Pseudo-nitzschia *and* Dactyliosola.

### 3.7.2 Relative abundance - plots of Dinophyta

First comparison between Greenland and Norway did not show an even distribution of the Dinophyta between the two regions, more Dinophyta were present in Greenland. A number of the pie charts in the upper area of the visible charts tended to increase in size and displayed a larger contingent of red. The increased size of the pie charts pointed toward a comparatively larger number of transcripts originating from Dinophyta in Greenland. The large portion of red in the pie charts was the result of a predominance of the genus *Lingulodinium*.

Second comparison between Greenland and Sweden displayed both previously described features except both were more pronounced. The distribution of charts in total seemed even more shifted towards Greenland and the dominance of *Lingulodinium* was more definitive.

The final comparison between Sweden and Norway deviated from the previously observed pattern not in terms of the arrow shape but in terms of species distribution. Unlike the two previous plots no distinguishable colour pattern was observed.

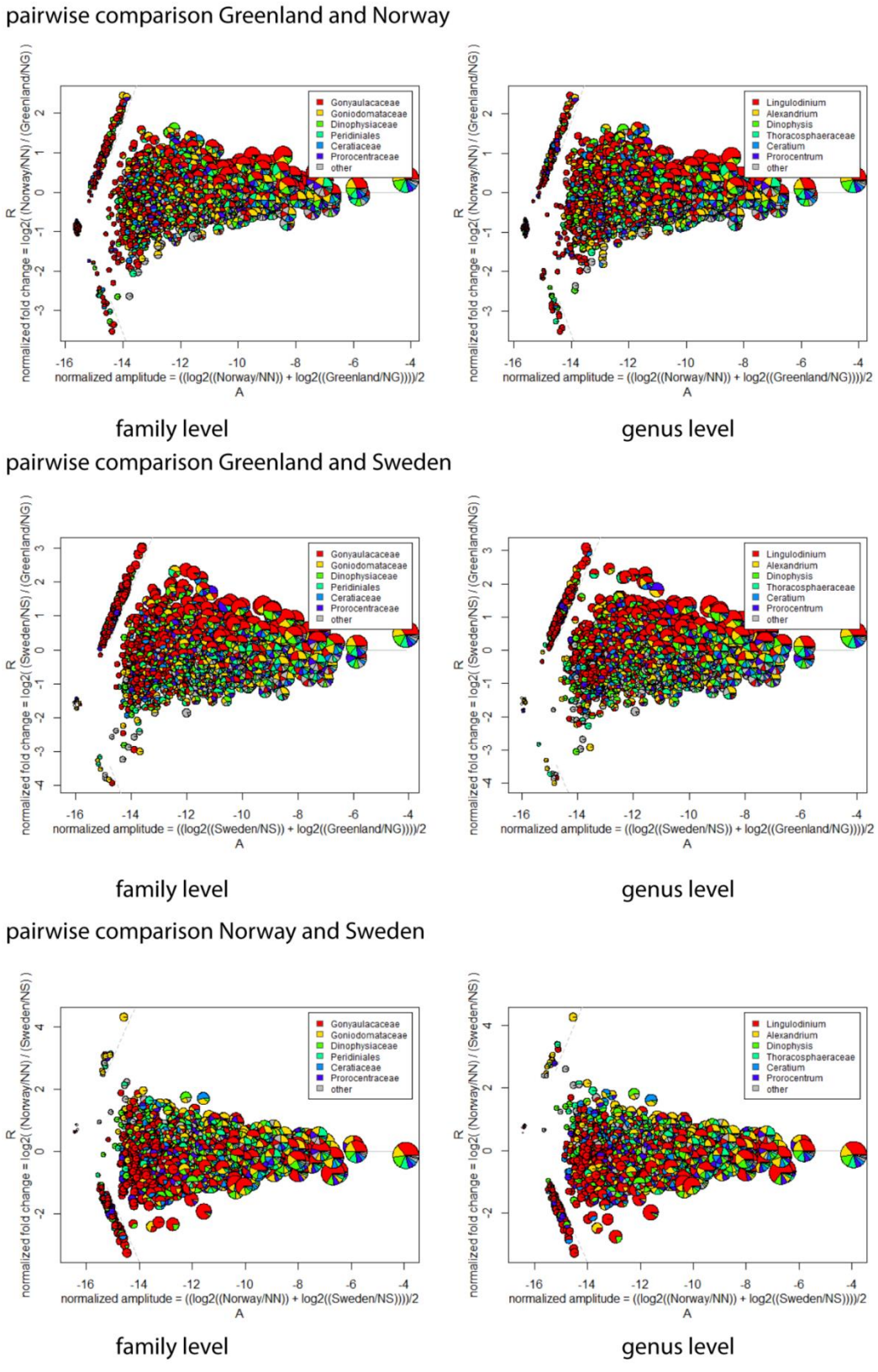



***Figure 26:*** *Pairwise comparison of 10 most abundant Dinophyta taxa. Plots of the left depict the "family" level, plots on the right the "genus" level. Paired regions for comparison are from top to bottom Greenland and Norway, Greenland and Sweden and Norway and Sweden. The 10 most abundant taxa for the Dinophyta are* Lingulodinium, Alexandrium, Scrippsiella, Symbiodinium, Prorocentrum, Dinophysis, Ceratinum, Heterocapsa *and* Glenodinium.

### 3.7.3 Relative abundance - plots of Haptophyta

The RA plots of the Haptophyta contained fewer transcripts compared to the other two taxonomical groups. All three of the Haptophyta plots appeared under plotted and it was not definitive whether Haptophyta were distributed evenly in the three comparative plots. None of the three plots displayed more than an approximated arrow shape indicative of an even distribution therefore no definitive conclusions were viable.

In addition due to the under plotting not patterns in regards to colour distribution within the contained pie charts emerged further limiting conclusions.

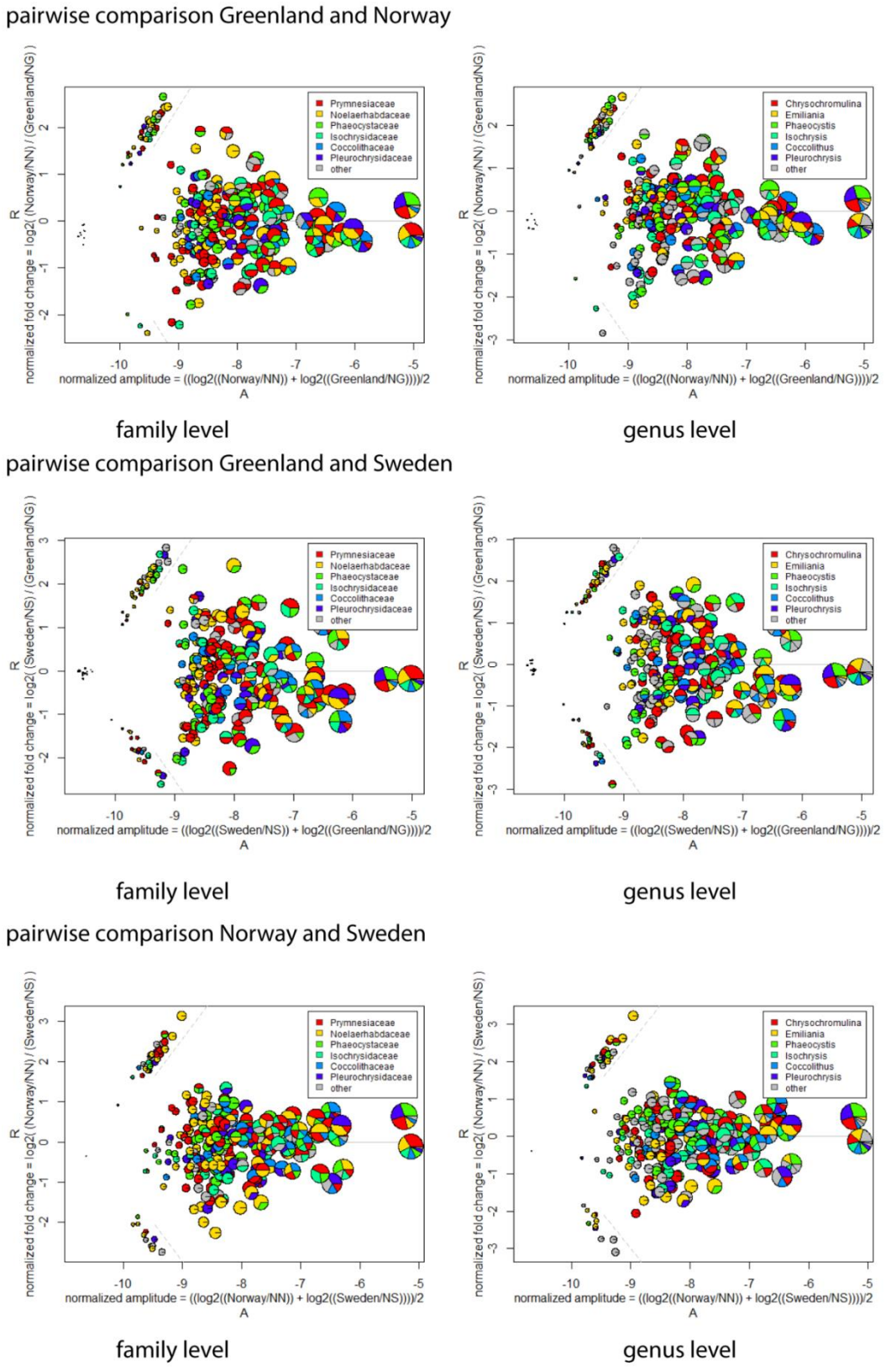



***Figure 27:*** *Pairwise comparison of 10 most abundant Haptophyta taxa. Plots of the left depict the "family" level, plots on the right the "genus" level. Paired regions for comparison are from top to bottom Greenland and Norway, Greenland and Sweden and Norway and Sweden. The 10 most abundant taxa for the Haptophyta are* Emiliana, Phaeocystis, Chrysochromulina, Isochrysis, Coccolithus, Pleurochrysis, Prymnesium, Gephyrocapsa, Calcidiscus *and* Chrysoculter.

## 3.8 Expression profiles / Metagenomics

***Table 1:*** *Heat-map of expression profiles of the three regions. COG categories are grouped into four groups based on their functions. In terms of total transcript count the region of Greenland shows the highest number of reads while the Swedish regions displays the fewest with about 1/3 as many reads. For all three regions about a third of all regions fall into the COG-category "unknown function". In the other three functional categories the group "information storage and processing" is represented highly in terms of contained reads.*

| | COG-category | Sweden | Norway | Greenland |
|---|---|---|---|---|
| information storage and processing | Translation | 932,887 | 1,488,010 | 2,422,525 |
| | Replication and repair | 213,863 | 256,703 | 311,114 |
| | Transcription | 206,572 | 253,245 | 344,494 |
| | Chromatin Structure and dynamics | 12,343 | 57,550 | 100,284 |
| | RNA processing and modification | 7,930 | 15,914 | 20,181 |
| cellular processes and signaling | Post-translational modification, protein turnover, chaperone functions | 410,414 | 795,249 | 1,047,478 |
| | Signal Transduction | 212,554 | 186,427 | 230,859 |
| | Cytoskeleton | 152,005 | 147,369 | 524,512 |
| | Cell wall/membrane/envelop biogenesis | 62,647 | 117,917 | 174,920 |
| | Intracellular trafficing and secretion | 47,174 | 124,126 | 115,203 |
| | Cell cycle control and mitosis | 31,787 | 38,700 | 34,921 |
| | Defense mechanisms | 11,476 | 22,225 | 47,770 |
| | Cell motility | 2,003 | 2,354 | 5,755 |
| | Nuclear structure | 717 | 1,576 | 2,365 |
| metabolism | Energy production and conversion | 427,290 | 598,931 | 547,420 |
| | Carbohydrate metabolism and transport | 323,335 | 415,171 | 487,950 |
| | Coenzyme metabolis | 205,588 | 244,968 | 230,264 |
| | Amino Acid metabolis and transport | 190,660 | 302,434 | 366,027 |
| | Inorganic ion transport and metabolism | 158,627 | 212,691 | 248,082 |
| | Nucleotide metabolism and transport | 127,334 | 111,876 | 99,852 |
| | Lipid metabolism | 82,410 | 150,638 | 171,726 |
| | Secondary Structure | 36,650 | 56,696 | 65,732 |
| Unknown or predicted | general function only predicted | 460,744 | 587,103 | 759,012 |
| | Function Unknown | 95,835 | 139,618 | 261,857 |
| | Sum of reads | 4,412,845 | 6,327,491 | 8,620,303 |

Within the functional group "information storage and processing" the first three cog-categories are in descending order "Translation", "Replication and repair" and "Transcription" for all three regions. "Replication and repair" was scored third in Greenland while scored second highest in both other regions. The functional group in total accounted for around 31 % of all transcripts in Sweden, 32 % in Norway and 37 % in Greenland.

The second most represented functional group is called "cellular processes and signalling". For Sweden and Norway are in descending order "post-translations modification, protein turnover and chaperone functions", "signal transduction" and "Cytoskeleton". For Greenland the last two named categories changed places. In total the group "cellular processes and

signalling" accounted for 21 % of all reads in Sweden, 22 % in Norway and 25 % in Greenland.

"Metabolism" functional group contained 8 COG categories. Three highest expressed categories were "Energy production and conversion", "Carbohydrate metabolism and transport" and "Coenzyme metabolism". The named three categories occurred in same sequence over all regions. The group "metabolism" contained 35 % of all reads in Sweden, 22 % in Norway and 25 % in Greenland.

Final function group contained the two COG categories of "predicted functions only" and "Unknown function". This group accounted for 12 % of the reads in Swedish region and 11 % in both Norway and Greenland.

## 3.9 Intra-region breakdown of expression

All Greenlandic stations showed a comparable number of allotted COG categories. Highest contingent of transcripts over all stations fell into “Translation” category. Other observed major protein category were “Cytoskeleton”, “Post-translational modification, protein turnover, chaperone functions”, “General functional prediction only prospects” and “energy production and conversion”. Minor differences in the classification patterns were observed for stations St511 and St512 in comparison to the other remaining stations. The named stations displayed considerably less protein of the “Cytoskeleton”, “Chromatin Structure and dynamics” and “Cell cycle control and mitosis” classification.

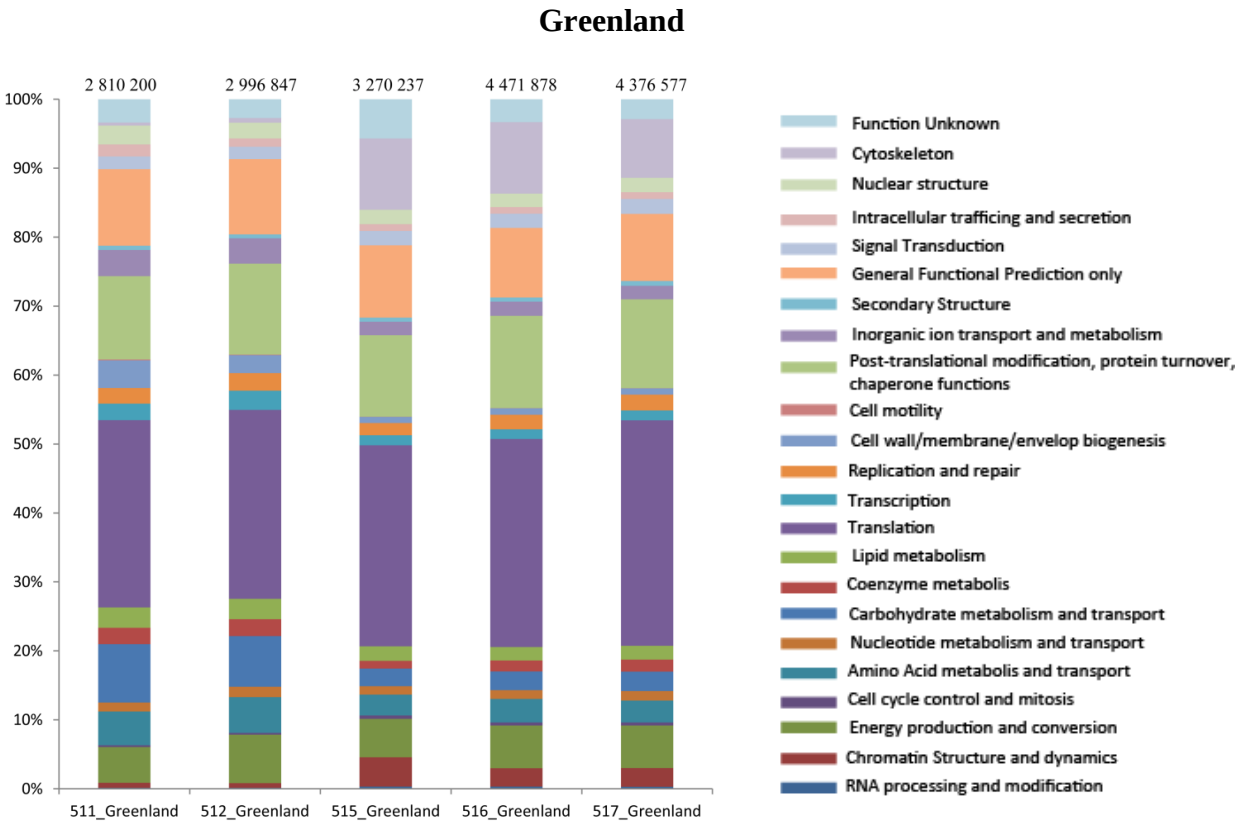



***Figure 28:*** *COG comparison of protein classifications of the Greenland sites. Assignments of classified transcripts given in percent, total number of transcripts per station given on top of columns.*

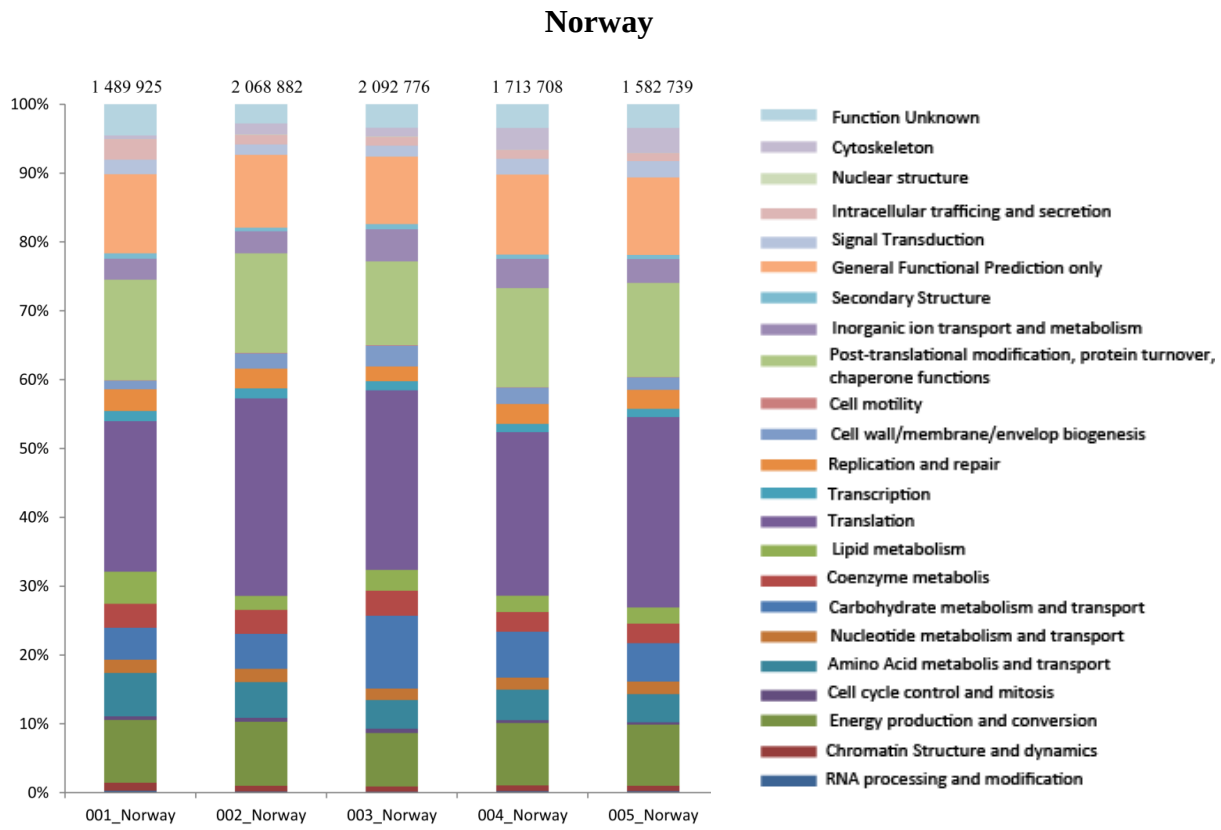



***Figure 29:*** *COG comparison of protein classifications of the Norwegian sites.*

The Norwegian region displayed some variation in regards to number of allotted reads was visible. While Norwegian stations St002 and St003 both had more than 2 million transcripts the other three fell between 300 – and 500.000 aligned transcripts short. Highest contingent of transcripts over all stations were allotted into the “Translation” classification. Other observed major protein categories were “Carbohydrate metabolism and transport”, “Post-translational modification, protein turnover, chaperone functions”, “General functional prediction only prospects” and “energy production and conversion”. Minor differences in the classification pattern were observed for stations St511 and St512 in comparison to the other remaining stations. The named stations displayed considerably less transcription of the “Cytoskeleton”, “Chromatin Structure and dynamics” and “Cell cycle control and mitosis” classification.

The Swedish stations displayed fairly comparable number of transcripts over the five stations between 1 million and 1.2 million. COG category “Translation” was again the most abundant COG category.

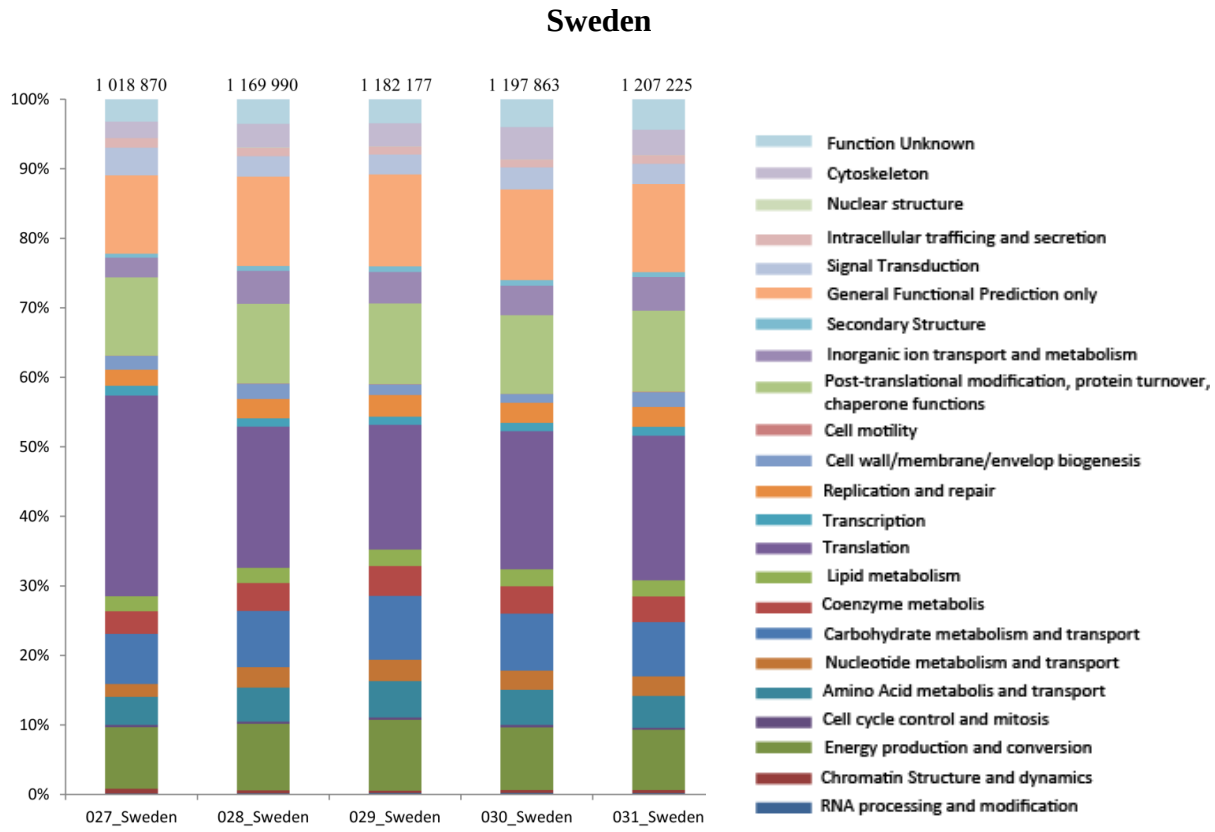



***Figure 30:*** *COG comparison of protein classifications of the Sweden sites.*

### 3.9.1 Processed intra-region comparison of expression profiles

In regards to total sum of transcripts the Swedish sample region contained the lowest number of transcripts while the Greenlandic displayed to highest number of transcripts.

The composition of the three profiles was consistent over large parts for all three regions. This does not only encompass the sequence of cog categories but also for the ratios of how many transcripts were assigned to the individual category. For all three regions most reads were allotted for proteins in the “Translation” cog category. The categories “PTM, protein turnover, ..”, “general functions prediction only” and “energy production and conversion” was scored highly for all three regions and are found under the top 5 categories. At the other end of the spectrum the categories “Cell motility” and “RNA processing and modification” are listed under the least expressed proteins.

Divergences between the regions were most apparent between Greenland and the other two regions: A number of categories such as “Cytoskeleton”, “Chromatin structures and dynamics” and “Nuclear structure” were higher expressed.

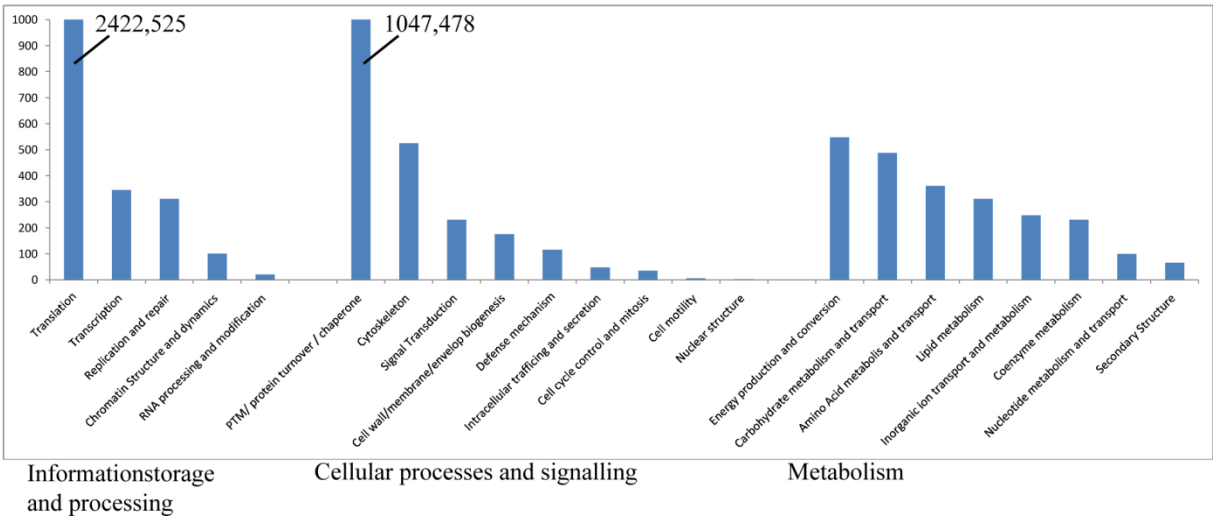



***Figure 31 A:*** *Expression profiles of the Greenlandic region. Displayed are the number of transcripts which could be allocated to the available COG categories in a thousand reads. Scale of Y-axis was adjusted to a maximal value of 1 million reads. Categories "Translation" and "PTM / protein turnover / chaperone" exceeded 1 million y-axis cap with 2.4 and 1.04 million reads respectively. The COG categories were grouped into three larger classes of "information storage and processing", "cellular processes and signalling" and "metabolism". Individual COG categories within said classes are sorted by expression levels in descending order.*

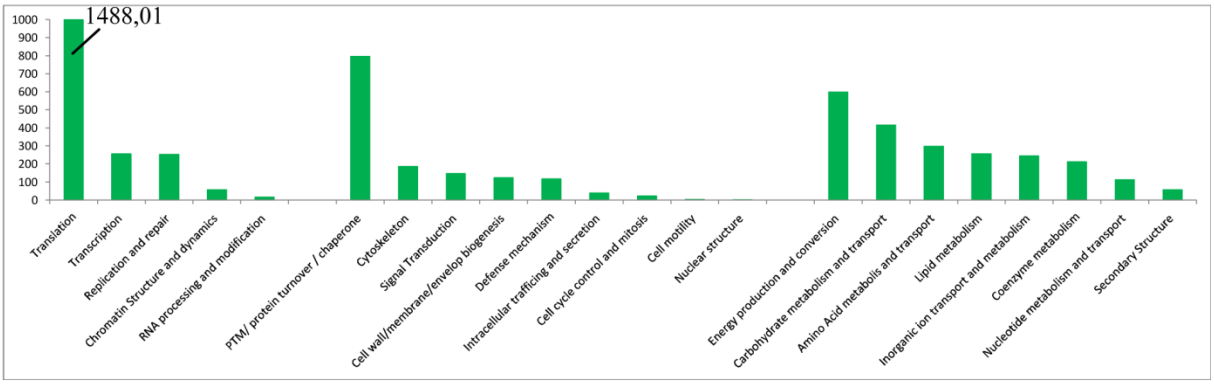



***Figure 31 B:*** *Expression profiles of the Norwegian region. Displayed are the number of transcripts which could be allocated to the available COG categories in a thousand reads. Scale of Y-axis was adjusted to a maximal value of 1 million reads. Category "Translation" exceeded 1 million y-axis cap with 1.4 million reads. The COG categories were grouped into three larger classes of "information storage and processing", "cellular processes and signalling" and "metabolism". Individual COG categories within said classes are sorted by expression levels in descending order.*

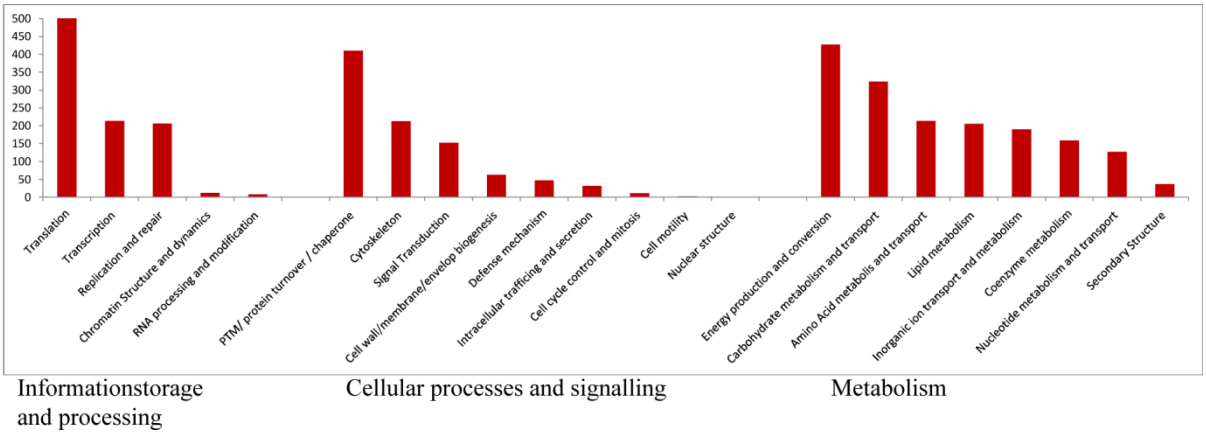



***Figure 31 C:*** *Expression profiles of the Swedish region. Displayed are the number of transcripts which could be allocated to the available COG categories in a thousand transcripts. Scale of Y-axis was adjusted to a maximal value of 500.000 transcripts. The COG categories were grouped into three larger classes of "information storage and processing", "cellular processes and signalling" and "metabolism". Individual COG categories within said classes are sorted by expression levels in descending order.*

### 3.9.2 Regional comparison of expression

Scored reads of individual COG categories were summed region wise and plotted in the three previously established function category.

The first functional group "Information storage and processing" contained the five COG categories "Translation", "Replication and repair", "Transcription", "Chromatin structure and dynamics" and "RNA processing and modification". As was observed previously differences in allotted transcripts occurred between the three regions as well as between COG categories. The later caused the need for multiple rescaling of the y-axis of the plots within each of the displayed functional groups.

The category "Translation" was highest expressed category warranting a scaled y-axis up to 2,5 million reads. "Replication and repair" and "Transcription" were scored 2$^{nd}$ and 3$^{rd}$ highest with y-axis scale of up to 35,000 reads. "Chromatin structure and dynamics" and "RNA processing and modification" were the two final COG categories within the functional group with to lowest rates of expression warranting a y-axis scale of 100,000 reads at most.

All five COG categories present the general scheme of Greenland being the highest expressed region followed by Norway with approximately ¾ to 2/3 of the Greenlandic expression levels. Swedish expression levels lay lower with 2/3 to ½ of the Greenlandic expression. One visible exemption it the fourth category "Chromatin structure and dynamics" were Sweden was scored visibly lower.

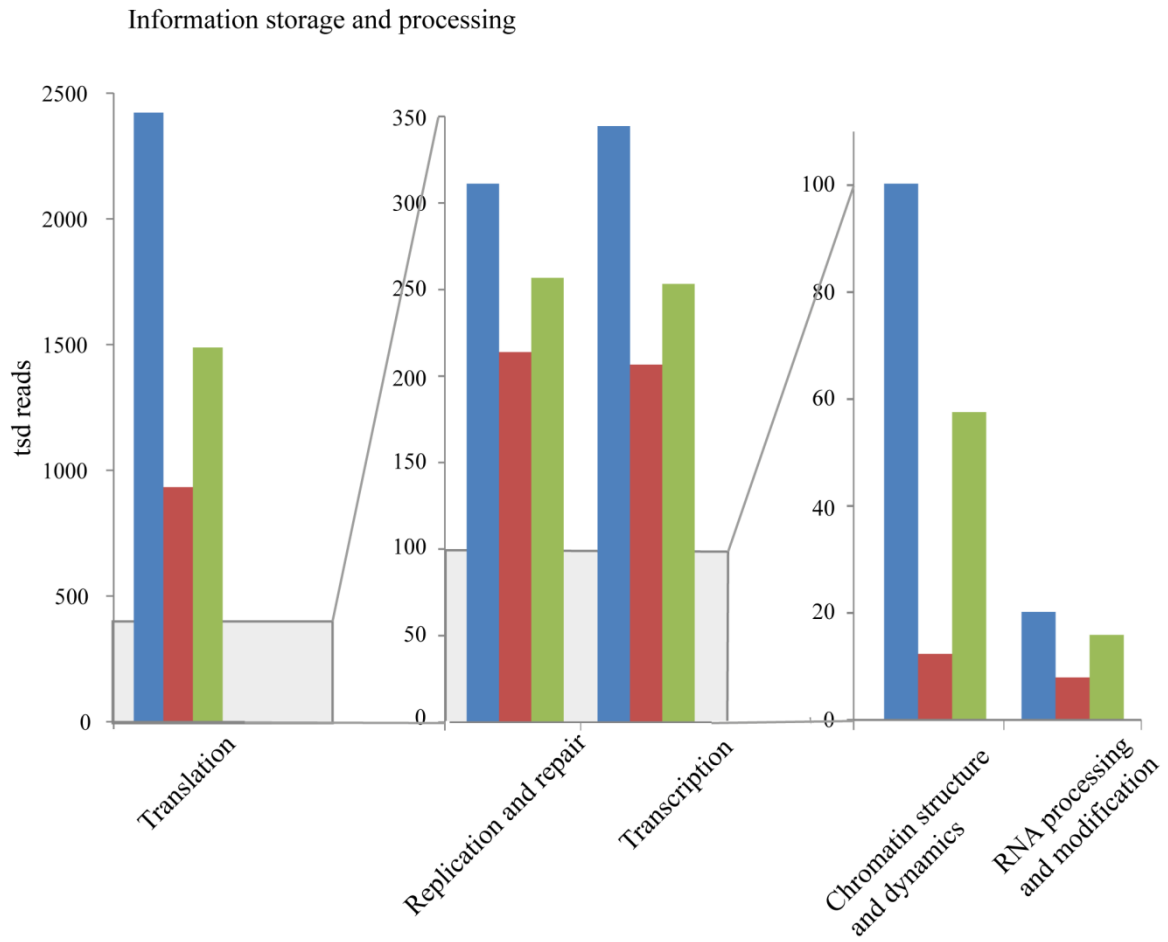



***Figure 32:*** *Regional comparison of the function COG grouping "information storage and processing". Said function COG grouping contained the COG categories of "Translation", "Replication and repair", "Transcription", "Chromatin structure and dynamics" and "RNA processing and modification". Transcripts are given in 1000 transcripts. Greenlandic region given in in blue, Sweden in red and Norway in green.*

The second functional group "Cellular processes and signalling" contained the nine COG categories. The differences in assigned transcripts caused the rescaling of the y-axis of the plots four times within the functional group.

The category "PTM and protein turnover" was highest expressed category with Greenland scoring more than 1 million reads. "Cytoskeleton" was scored 2$^{nd}$ highest with y-axis scaled up to half a million transcripts. The third grouping contained five COG categories and was the largest faction within "Cellular processes and signalling". The y-axis for that groping was set to 250,000 reads. The remaining two COG categories "Cell motility" and "Nuclear structure" were only observed in limited numbers not exceeding 6,000 transcripts.

The rations between the expression levels of the regions as described above did not apply extensively. Only the first COG category "PTM, protein turnover ..." , the third category "Cell wall/ membrane synthesis" and the sixth "Intracellular trafficking and secretion" displayed the established ratios. In the second category "Cytoskeleton" the Norwegian region appeared less prominent. Conversely in the category "Signal transduction" the

Norwegian sites appeared more strongly. The Swedish region deviated from the ratios described in the first functional group in the categories “Cytoskeleton” and “Signal transduction” where expression levels surpassed the Norwegian region as well as in categories “Cell cycle control and mitosis and “Cell motility” were the expression levels were scored comparatively with the Norwegian ones.

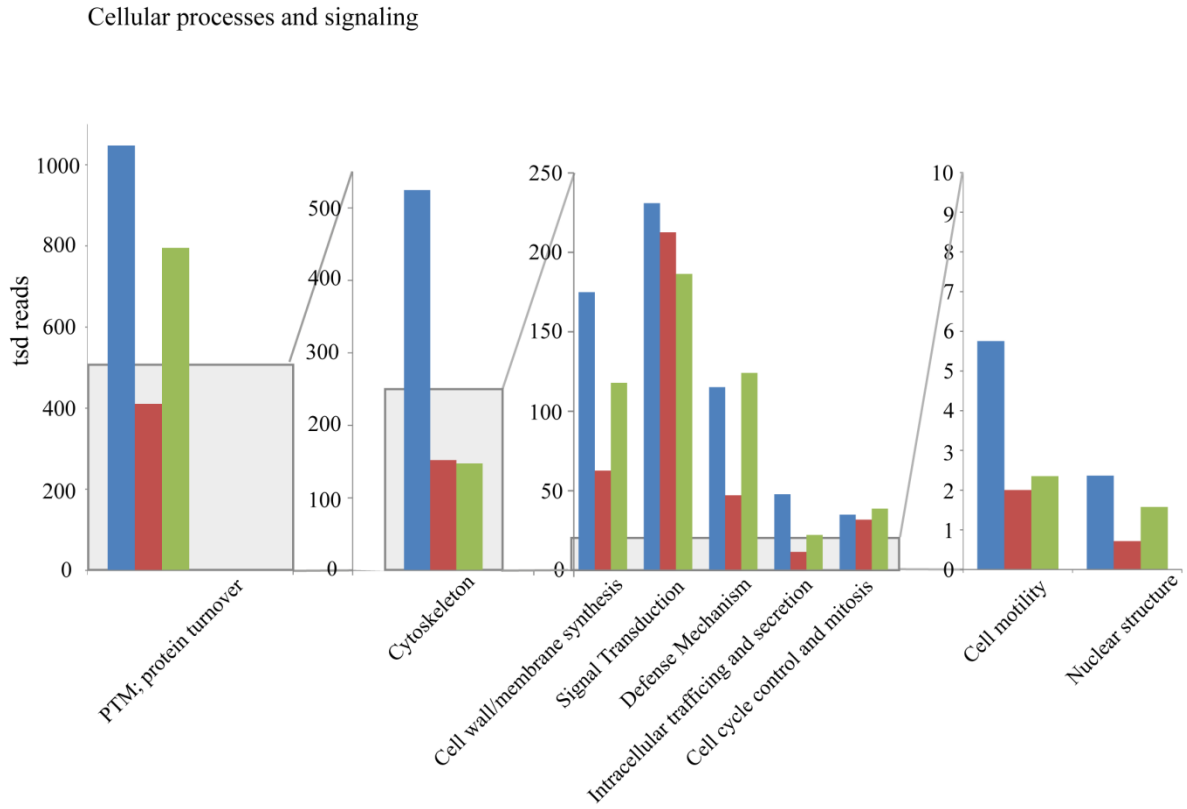



***Figure 33:*** *Regional comparison of the function COG grouping “cellular processes and signalling”. Said function COG grouping contained the COG categories of “PTM, protein turnover, chaperone proteins”, “Cytoskeleton”, “Cell wall/membrane/envelope synthesis”, “Signal transduction” , “Defence Mechanism”, “Intracellular trafficking and secretion”, “Cell cycle control and mitosis”, “Cell motility” and “Nuclear structure”. Transcripts are given in 1000 transcripts. Greenlandic region given in in blue, Sweden in red and Norway in green.*

The third functional group “Metabolism” contained the eight COG categories. The y-axis of the plots was rescaled two times within the functional group, the first time after the first two categories from 600,000 to 350,000 transcripts and the second time for the last two categories to 150,000 transcripts.

Despite the rescaling of the y-axis the COG categories contained in “Metabolism” displayed a narrower scope in comparison of the other two functional groups. Within the group all regions displayed expression levels of approximately comparative levels. Greenlandic region still was observed as the highest scored region with the exceptions of COG categories “Energy production and conversion” where Norway took the first place and “Nucleotide metabolism and transport” where Greenland was scored the lowest after both Sweden and Norway.

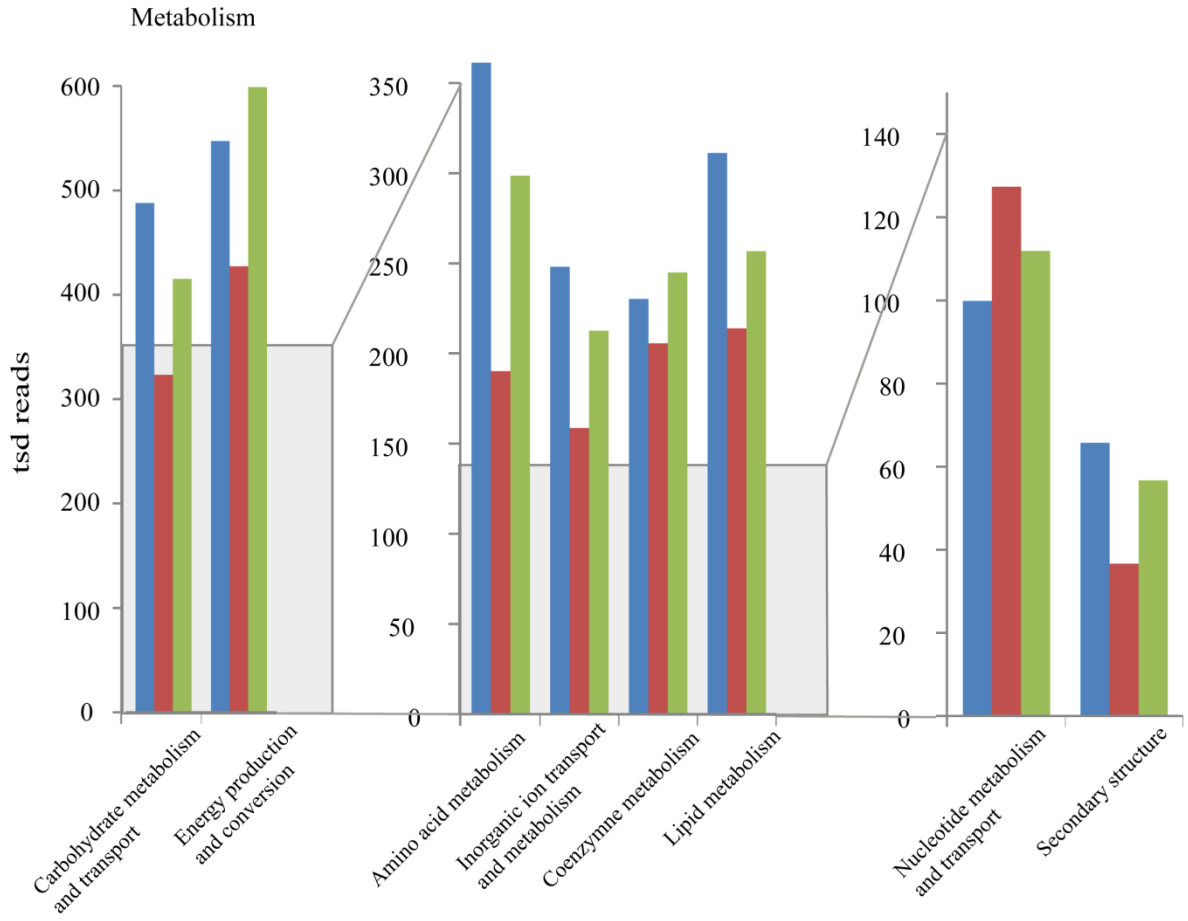



***Figure 34:*** *Regional comparison of the function COG grouping "metabolism". Said function COG grouping contained the COG categories of "Carbohydrate metabolism and transport", "Energy production and conversion", "Amino acid metabolism", "Inorganic ion transport and metabolism" , "Coenzyme metabolism", "Lipid metabolism", "Nucleotide metabolism and transport" and "Secondary structure". Transcripts are given in 1000 transcripts. Greenlandic region given in blue, Sweden in red and Norway in green.*

### 3.9.3 Linked expression profiles of individual species

Combination of oberservations made from previous diversity- and proteomics-aspects of the investigation was achieved by dividing the observed species into the three most prominent taxonomical taxa "Dinophyta", "Bacillariophyta" and "Haptophyceae" and subsequently plotting them with the individual COG categories.

Figures 30 through 33 display exemplary the COG category "Translation" for the three taxonomical sub sets. For the remaining figures of other COG categories see the Supplemental data.

The Dinoflagellata contained 27 species in total. For most depicted species expression of COG category "Translation" was strongest in the Swedish region. Species with highest

expression of all regions were *Lingulodinium, Alexandrium, Scrippsiella, Prorocentrum* and *Symbiodinium.*

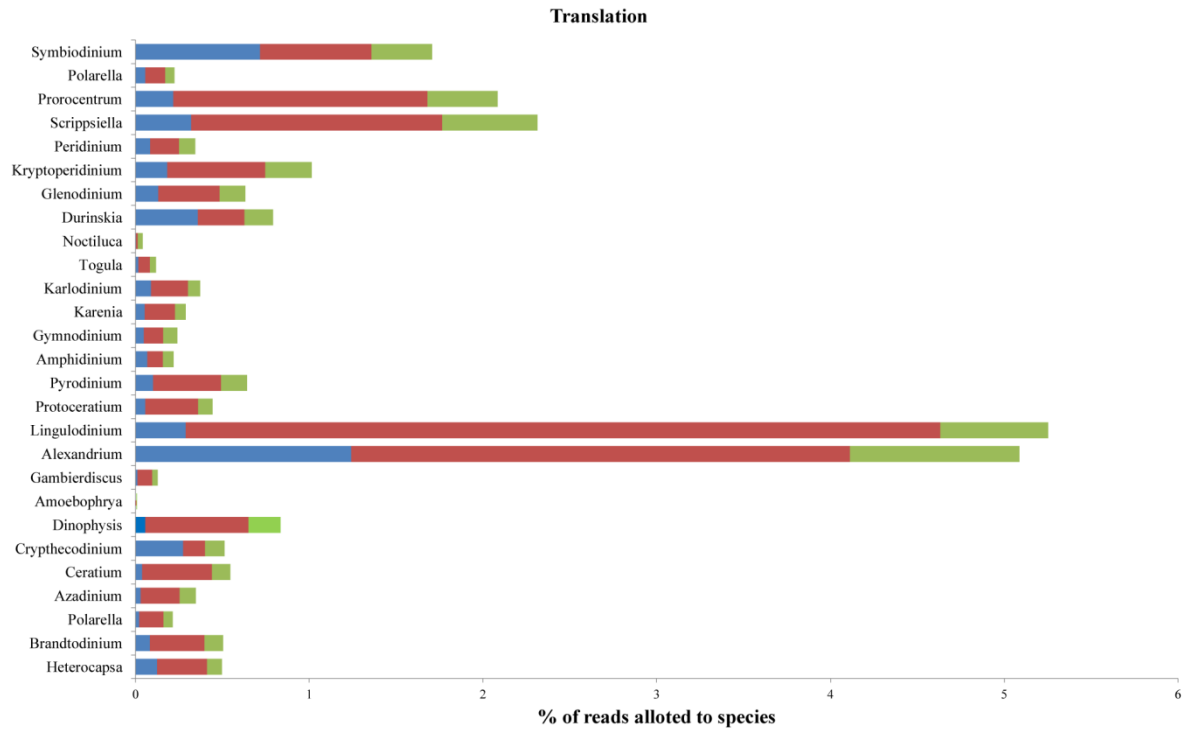



***Figure 35:*** *COG – category Translation for Dinophyta. Greenlandic region given in blue, Sweden in red and Norway in green.*

For the phylum Bacillariophyta 39 species were found to be of importance on expression of the different COG categories. Protein expression from Bacillariophyta were only found in low numbers in Swedish sample stations. Scored expression levels among the different species were distributed unevenly which expression levels of several species overshadowing others. For the Greenlandic region *Thalassiosira* and *Chaetoceros* and to a lesser extent *Detonula, Skeletonema* and *Pseudo-nitzschia* were dominant. Norwegian expression profiles were dominated by observed expression profiles of species *Chaetoceros* and *Corethron.*

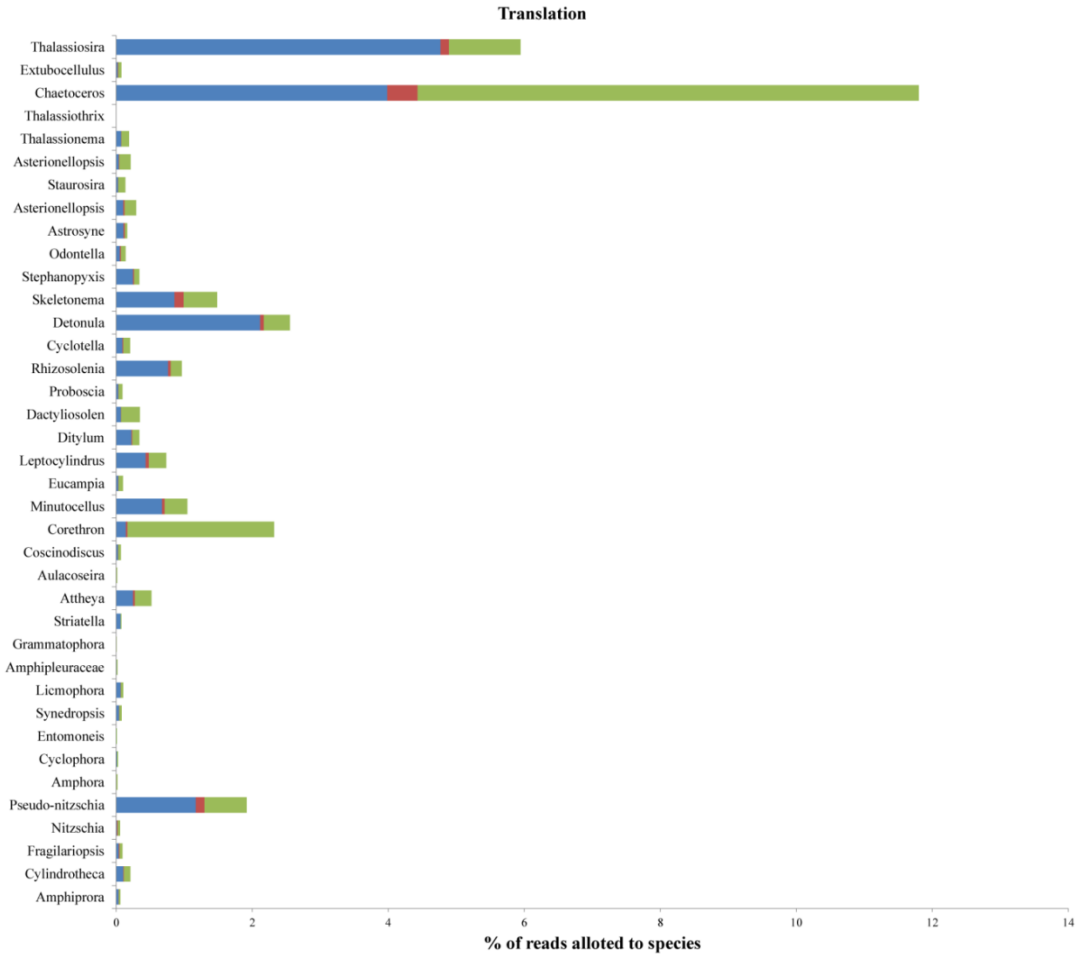



***Figure 36:*** *COG – category Translation for Bacillariophyta. Greenlandic region given in blue, Sweden in red and Norway in green.*

Transcripts found to be originating from various Haptophyta were found in smaller numbers in comparison to Dinophyta or Bacillariophyta. Twelve species were found to be of importance in regards to observed COG expressions. Haptophyta transcripts were found most numerous in Greenland with the species *Prymnesium* and *Pleurochrysis* being the most represented.

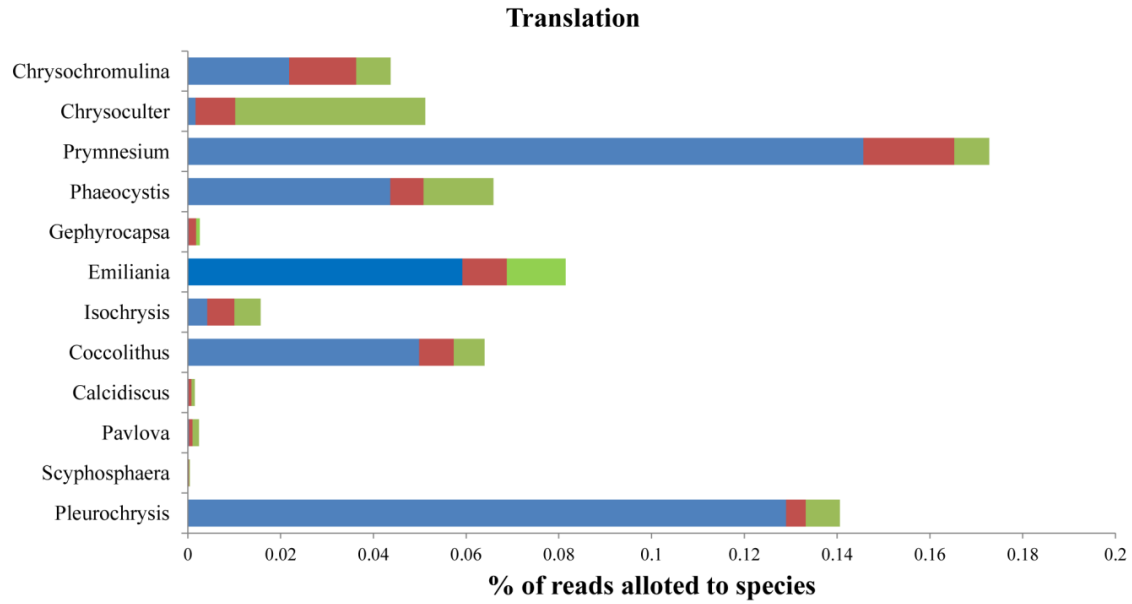



***Figure 37:*** *COG – category "Translation" for Haptophyta. Greenlandic region given in blue, Sweden in red and Norway in green.*

## 3.10 Correlation between regions

### 3.10.1 Test of correlation of dataset with Mantel test

The pairwise comparison of the count tables allowed the determination the grade of similarity between the data through different stage of the study. The count table of the pre – BLAST search data and the counts of the 28S rRNA LSU results were found to be significantly correlated. In contrast the later stages of the investigations, namely the comparisons of pre- and post-BLAST search count tables as well as the post-BLAST counts with the reduced set of BLAST hits with matched COG annotation were not significantly similar.

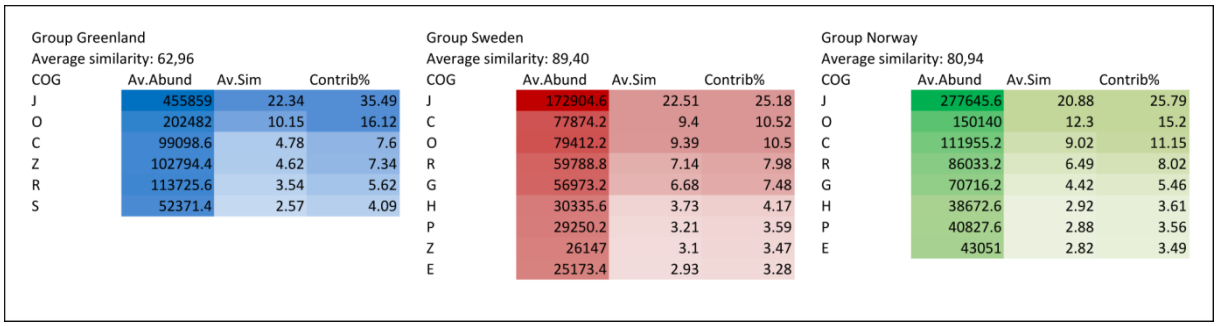
### 3.10.2 Similarity percentage (SIMPER) Analysis

The Similarity percentage (SIMPER) was conducted as extension of the previous Mantel test. Instead of a search for correlation between matrixes it allowed the determination of similarity between the three regions in regards to a chosen factor of an input file compiled for the analysis. To that end a set of three input files were analysed with the SIMPER analysis: Firstly the similarity between regions in the expression of COG categories, secondly the similarity in biodiversity and thirdly a combination of the two previous factors.

Regarding the similarities in protein expression the intra region the largest expressed categories mirrored previously observed expression levels. A major difference towards previous examinations it that the input file included the COG categories of "unknown (S)" and" predicted only (R)" which were excluded from previous examinations. Apart from those two categories the depicted COG categories were comparable to the previous analysed protein expression levels. Greenland displayed the lowest similarity in expression levels with

63 % followed by Norway with 81 % and Sweden with a similarity in protein expression of 89 %.

***Table 2:*** *Shortened results of COG category transcript count. Results were shortened to contain 75 % of all transcripts. Figure shows similarity in COG category expression within sample region. Columns give average abundance of transcripts assigned to category, average similarity as well as percentual contribution of COG category to overall expression.*



Group Greenland
Average similarity: 62,96

| COG | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| J | 455859 | 22.34 | 35.49 |
| O | 202482 | 10.15 | 16.12 |
| C | 99098.6 | 4.78 | 7.6 |
| Z | 102794.4 | 4.62 | 7.34 |
| R | 113725.6 | 3.54 | 5.62 |
| S | 52371.4 | 2.57 | 4.09 |

Group Sweden
Average similarity: 89,40

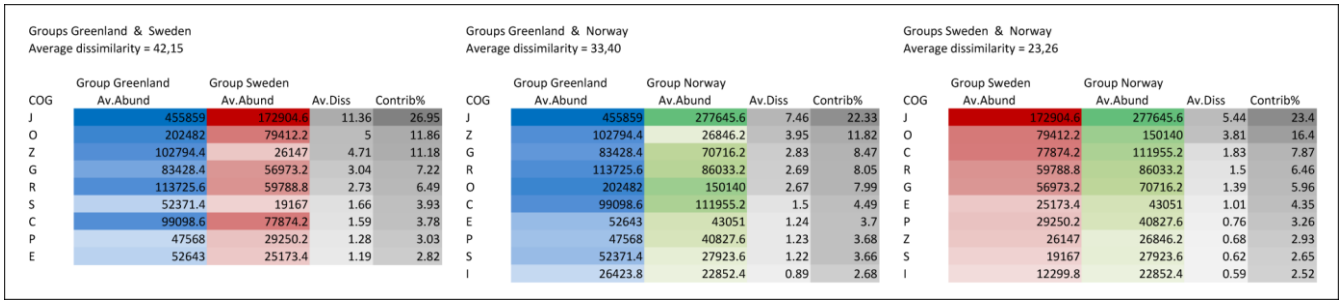
| COG | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| J | 172904.6 | 22.51 | 25.18 |
| C | 77874.2 | 9.4 | 10.52 |
| O | 79412.2 | 9.39 | 10.5 |
| R | 59788.8 | 7.14 | 7.98 |
| G | 56973.2 | 6.68 | 7.48 |
| H | 30335.6 | 3.73 | 4.17 |
| P | 29250.2 | 3.21 | 3.59 |
| Z | 26147 | 3.1 | 3.47 |
| E | 25173.4 | 2.93 | 3.28 |

Group Norway
Average similarity: 80,94

| COG | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| J | 277645.6 | 20.88 | 25.79 |
| O | 150140 | 12.3 | 15.2 |
| C | 111955.2 | 9.02 | 11.15 |
| R | 86033.2 | 6.49 | 8.02 |
| G | 70716.2 | 4.42 | 5.46 |
| H | 38672.6 | 2.92 | 3.61 |
| P | 40827.6 | 2.88 | 3.56 |
| E | 43051 | 2.82 | 3.49 |

By comparing the similarities in protein expression pairwise between two regions the degree of the region correlation was determined. On the three tested combinations Greenland and Sweden were the most dissimilar with an averaged dissimilarity of 42 % followed by Greenland and Sweden with 33 % while Sweden and Norway were scored the closed matched based on COG expression profiles with a dissimilarity of 23 %.

***Table 3:*** *Shortened results of the pairwise SIMPER analysis. Results were shortened to contain 75 % of all transcripts. Figure shows similarity in COG category expression in pairwise comparison of two regions. Columns give average abundance of transcripts allotted to category of a region, average dissimilarity as well as percentual contribution of cog category to overall expression. Compared pairs are from left to right are Greenland and Sweden, Greenland and Norway and Sweden and Norway.*



Groups Greenland & Sweden
Average dissimilarity = 42,15

| COG | Group Greenland Av.Abund | Group Sweden Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| J | 455859 | 172904.6 | 11.36 | 26.95 |
| O | 202482 | 79412.2 | 5 | 11.86 |
| Z | 102794.4 | 26147 | 4.71 | 11.18 |
| G | 83428.4 | 56973.2 | 3.04 | 7.22 |
| R | 113725.6 | 59788.8 | 2.73 | 6.49 |
| S | 52371.4 | 19167 | 1.66 | 3.93 |
| C | 99098.6 | 77874.2 | 1.59 | 3.78 |
| P | 47568 | 29250.2 | 1.28 | 3.03 |
| E | 52643 | 25173.4 | 1.19 | 2.82 |

Groups Greenland & Norway
Average dissimilarity = 33,40

| COG | Group Greenland Av.Abund | Group Norway Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| J | 455859 | 277645.6 | 7.46 | 22.33 |
| Z | 102794.4 | 26846.2 | 3.95 | 11.82 |
| G | 83428.4 | 70716.2 | 2.83 | 8.47 |
| R | 113725.6 | 86033.2 | 2.69 | 8.05 |
| O | 202482 | 150140 | 2.67 | 7.99 |
| C | 99098.6 | 111955.2 | 1.5 | 4.49 |
| E | 52643 | 43051 | 1.24 | 3.7 |
| P | 47568 | 40827.6 | 1.23 | 3.68 |
| S | 52371.4 | 27923.6 | 1.22 | 3.66 |
| I | 26423.8 | 22852.4 | 0.89 | 2.68 |

Groups Sweden & Norway
Average dissimilarity = 23,26

| COG | Group Sweden Av.Abund | Group Norway Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| J | 172904.6 | 277645.6 | 5.44 | 23.4 |
| O | 79412.2 | 150140 | 3.81 | 16.4 |
| C | 77874.2 | 111955.2 | 1.83 | 7.87 |
| R | 59788.8 | 86033.2 | 1.5 | 6.46 |
| G | 56973.2 | 70716.2 | 1.39 | 5.96 |
| E | 25173.4 | 43051 | 1.01 | 4.35 |
| P | 29250.2 | 40827.6 | 0.76 | 3.26 |
| Z | 26147 | 26846.2 | 0.68 | 2.93 |
| S | 19167 | 27923.6 | 0.62 | 2.65 |
| I | 12299.8 | 22852.4 | 0.59 | 2.52 |

Regarding the similarities in species abundance within regions the SIMPER brought out the variance in diversity very clearly. Greenland displayed the lowest similarity with 46 % and the depicted 17 species together amounted for 50 of the transcripts. The Swedish region lay at the other end of the spectrum with a similarity of 84 % and six species accounted for 50 of all transcripts. The Norwegian sites were scored similar to the Greenland with a similarity of 48 %.

***Table 4:*** *Shortened results of species abundance. Results were shortened to contain 50 % of all transcripts with can be allotted to a species. Columns give average abundance of transcripts allotted to species of a region, average similarity as well as percentual contribution of species to overall species diversity.*

Group Greenland
Average similarity: 45,84

| Species | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| Thalassiosira_sp. | 311294.4 | 4.63 | 10.1 |
| Chaetocerotaceae_sp. | 88156.4 | 3.23 | 7.04 |
| Detonula_confervacea | 149397.6 | 2.24 | 4.88 |
| Chaetoceros_cf. neogracile | 131956.2 | 2 | 4.36 |
| Symbiodinium_sp. | 34245.6 | 1.78 | 3.88 |
| Lingulodinium_sp. | 20170.2 | 1.06 | 2.32 |
| Alexandrium_sp. | 21005 | 1.04 | 2.27 |
| Chaetoceros_sp. | 33548 | 0.96 | 2.09 |
| Genus nov._species nov. | 18832.8 | 0.86 | 1.88 |
| Thoracosphaeraceae_Scrippsiella | 16633.2 | 0.78 | 1.7 |
| Pseudokeronopsis_sp. | 17119.2 | 0.76 | 1.65 |
| Favella_sp. | 15714.4 | 0.69 | 1.51 |
| Minutocellus_polymorphus | 20584 | 0.64 | 1.4 |
| Tiarina_fusus | 11484.6 | 0.64 | 1.4 |
| Paraphysomonas_vestita | 14706.2 | 0.62 | 1.36 |
| Blepharisma_japonicum | 12375.2 | 0.59 | 1.29 |
| Unknown_Unknown | 12624.6 | 0.58 | 1.27 |

Group Sweden
Average similarity: 84,69

| Species | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| Lingulodinium_sp. | 209115.2 | 22.16 | 26.17 |
| Alexandrium_sp. | 46744.8 | 5.84 | 6.9 |
| Thoracosphaeraceae_Scrippsiella | 45429 | 5.76 | 6.8 |
| Prorocentrum_sp. | 35749.8 | 4.07 | 4.8 |
| Tiarina_fusus | 37272.2 | 3.87 | 4.57 |
| Dinophysis_acuminata | 24288.2 | 2.98 | 3.52 |

Group Norway
Average similarity: 48,33

| Species | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| Chaetocerotaceae_sp. | 217225.8 | 7.37 | 15.25 |
| Chaetoceros_cf. neogracile | 59326 | 3.16 | 6.55 |
| Chaetoceros_sp. | 49280.8 | 2.31 | 4.78 |
| Lingulodinium_sp. | 37121 | 2.17 | 4.5 |
| Thalassiosira_sp. | 33306.2 | 1.92 | 3.97 |
| Thoracosphaeraceae_Scrippsiella | 27437.2 | 1.7 | 3.51 |
| Corethron_hystrix | 190918.4 | 1.63 | 3.37 |
| Alexandrium_sp. | 24902.8 | 1.46 | 3.01 |
| Symbiodinium_sp. | 16713.2 | 1.13 | 2.33 |
| Chaetoceros_brevis | 19931 | 1.02 | 2.11 |
| Dinophysis_acuminata | 12732.8 | 0.82 | 1.7 |

The inter-region comparison highlighted the differences in species abundance drastically. Greenland and Sweden were the most dissimilar with an average dissimilarity of 73 %. The species abundant in one region were a minority in the other region. This was especially explicit for the species *Thalassiosira* and *Lingulodinium* which occupied the two top spots in the table.

The remaining two comparisons Greenland and Norway as well as Sweden and Norway displayed a similar extent of average dissimilarity with 61 %.

***Table 5:*** *Shortened results of pairwise species abundance. Results were shortened to contain 50 % of all transcripts. Results were shortened to contain 50 % of all transcripts with can be assigned to a species. Columns give average abundance of transcripts allotted to species of a region, average dissimilarity as well as percentual contribution of species to overall species diversity. Compared pairs are from top to bottom are Greenland and Sweden, Greenland and Norway and Sweden and Norway.*

Groups Greenland & Sweden
Average dissimilarity = 72,85

| Species | Group Greenland Av.Abund | Group Sweden Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| Thalassiosira_sp. | 311294.4 | 3850.8 | 10.55 | 14.48 |
| Lingulodinium_sp. | 20170.2 | 209115.2 | 8.92 | 12.24 |
| Detonula_confervacea | 149397.6 | 1877.6 | 5.08 | 6.97 |
| Chaetoceros_cf. neogracile | 131956.2 | 4060 | 4.45 | 6.11 |
| Chaetocerotaceae_sp. | 88156.4 | 7458 | 3.19 | 4.38 |
| Thoracosphaeraceae_Scrippsiella | 16633.2 | 45429 | 1.4 | 1.92 |
| Prorocentrum_sp. | 7658 | 35749.8 | 1.34 | 1.84 |
| Tiarina_fusus | 11484.6 | 37272.2 | 1.19 | 1.64 |
| Chaetoceros_sp. | 33548 | 2202.6 | 1.18 | 1.62 |

Groups Greenland & Norway
Average dissimilarity = 61,56

| Species | Group Greenland Av.Abund | Group Norway Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| Thalassiosira_sp. | 311294.4 | 33306.2 | 9.01 | 14.63 |
| Corethron_hystrix | 9217.4 | 190918.4 | 7.25 | 11.77 |
| Chaetocerotaceae_sp. | 88156.4 | 217225.8 | 6.85 | 11.13 |
| Detonula_confervacea | 149397.6 | 12843.2 | 4.29 | 6.96 |
| Chaetoceros_cf. neogracile | 131956.2 | 59326 | 3.94 | 6.4 |

Groups Sweden & Norway
Average dissimilarity = 61,44

| Species | Group Sweden Av.Abund | Group Norway Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| Chaetocerotaceae_sp. | 7458 | 217225.8 | 10.67 | 17.36 |
| Corethron_hystrix | 1302.8 | 190918.4 | 9.99 | 16.26 |
| Lingulodinium_sp. | 209115.2 | 37121 | 9.19 | 14.97 |
| Chaetoceros_cf. neogracile | 4060 | 59326 | 2.92 | 4.75 |

The combination of species with expression of COG categories highlighted a different aspect of the investigations. Greenland displayed an average similarity of 41 % as a result of the elevated species diversity. Swedish region lay at the other end of the spectrum with a similarity of 80 % due to a much lower diversity. While not only fewer different species were listed in the Swedish sample region but also the species *Lingulodinium* appeared with six different COG categories dominating the region.

Norway in terms of similarity was scored closer to Greenland with 45 %. However among the most abundant species different sub species *Chaetocerotaceae* and *Chaetoceros* were listed with multiple COG categories.

***Table 6:*** *Shortened results of abundance of combination of species with COG category. Results were shortened to contain 25 % of all transcripts. Columns give average abundance of transcripts assigned to species of a region, average similarity as well as percentual contribution of species to overall transcript count.*

Group Greenland
Average similarity: 41,10

| Species | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| Thalassiosira_sp._J | 71428.4 | 1.08 | 2.62 |
| Chaetocerotaceae_sp._J | 26358.6 | 0.86 | 2.08 |
| Detonula_confervacea_J | 32176.8 | 0.73 | 1.77 |
| Symbiodinium_sp._J | 12286.2 | 0.7 | 1.69 |
| Pseudokeronopsis_sp._Z | 15535 | 0.69 | 1.68 |
| Thalassiosira_sp._O | 32848 | 0.5 | 1.21 |
| Paraphysomonas_vestita_Z | 11191 | 0.48 | 1.18 |
| Alexandrium_sp._J | 9078.8 | 0.47 | 1.14 |
| Thalassiosira_sp._R | 33642 | 0.46 | 1.13 |
| Chaetocerotaceae_sp._O | 11978.4 | 0.41 | 1 |
| Rhizosolenia_setigera_J | 13132.4 | 0.41 | 0.99 |
| Minutocellus_polymorphus_J | 11484.4 | 0.4 | 0.98 |
| Tiarina_fusus_J | 6158.4 | 0.38 | 0.91 |
| Chaetoceros_cf. neogracile_O | 28693.6 | 0.37 | 0.91 |
| Chaetoceros_cf. neogracile_J | 20285 | 0.37 | 0.91 |
| Symbiodinium_sp._Z | 7150 | 0.34 | 0.83 |
| Chaetocerotaceae_sp._R | 7951.6 | 0.33 | 0.79 |
| Blepharisma_japonicum_Z | 7367.4 | 0.28 | 0.69 |
| Aplanochrytrium_sp._Z | 6556.4 | 0.28 | 0.69 |
| Thalassiosira_sp._G | 22105.4 | 0.26 | 0.64 |
| Favella_sp._E | 6926.2 | 0.26 | 0.64 |
| Thalassiosira_sp._C | 16150 | 0.26 | 0.64 |

Group Sweden
Average similarity: 79,55

| Species | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| Lingulodinium_sp._J | 34859.2 | 4.07 | 5.12 |
| Lingulodinium_sp._R | 23643.4 | 2.41 | 3.03 |
| Lingulodinium_sp._O | 20477.4 | 2.18 | 2.74 |
| Lingulodinium_sp._C | 18347 | 1.9 | 2.39 |
| Lingulodinium_sp._G | 17655 | 1.72 | 2.16 |
| Alexandrium_sp._J | 12167.6 | 1.59 | 2 |
| Thoracosphaeraceae_Scrippsiella_J | 11857.2 | 1.47 | 1.84 |
| Tiarina_fusus_J | 12357.2 | 1.45 | 1.82 |
| Lingulodinium_sp._P | 12772.8 | 1.24 | 1.56 |
| Prorocentrum_sp._J | 9852.8 | 1.2 | 1.51 |
| Gonyaulax_spinifera_J | 8801.8 | 1.13 | 1.42 |

Group Norway
Average similarity: 45,05

| Species | Av.Abund | Av.Sim | Contrib% |
|---|---|---|---|
| Chaetocerotaceae_sp._J | 52341.6 | 1.89 | 4.2 |
| Chaetocerotaceae_sp._O | 26746.6 | 1.03 | 2.28 |
| Chaetoceros_cf. neogracile_J | 16766.8 | 0.91 | 2.03 |
| Chaetocerotaceae_sp._C | 19280.6 | 0.69 | 1.52 |
| Thalassiosira_sp._J | 12153.4 | 0.64 | 1.42 |
| Chaetocerotaceae_sp._R | 16946.6 | 0.54 | 1.19 |
| Chaetoceros_sp._J | 11429.4 | 0.52 | 1.16 |
| Chaetoceros_cf. neogracile_O | 8359.4 | 0.5 | 1.11 |
| Chaetoceros_cf. neogracile_C | 10397.8 | 0.49 | 1.1 |
| Chaetoceros_brevis_O | 7174 | 0.44 | 0.98 |
| Chaetocerotaceae_sp._P | 13620.8 | 0.43 | 0.95 |
| Strombidinopsis_acuminatum_J | 9177 | 0.41 | 0.91 |
| Lingulodinium_sp._J | 6842.4 | 0.39 | 0.86 |
| Chaetocerotaceae_sp._G | 16771.4 | 0.38 | 0.85 |
| Alexandrium_sp._J | 5858.4 | 0.38 | 0.84 |
| Thoracosphaeraceae_Scrippsiella_J | 6419.2 | 0.38 | 0.84 |
| Symbiodinium_sp._J | 4252.8 | 0.34 | 0.76 |
| Lingulodinium_sp._O | 4853.2 | 0.29 | 0.64 |
| Favella_sp._J | 6185.4 | 0.28 | 0.63 |
| Chaetoceros_sp._O | 5553.4 | 0.28 | 0.62 |
| Chaetocerotaceae_sp._E | 8084.8 | 0.28 | 0.62 |

The comparative plotting of two regions combination of species with expression of COG categories pronounced the differences between the regions. SIMPER rated average dissimilarity for all three pairwise comparisons above 65 % with the pair of Greenland and Sweden achieving a dissimilarity of 77 %.

***Table 7:*** *Shortened results of pairwise SIMPER analysis. Results were shortened to contain 25 % of all transcripts. Columns give average abundance of transcripts assigned to combined species/COG category of a region, average dissimilarity as well as percentual contribution of given species/cog category. Compared pairs are from left to right are Greenland and Sweden, Greenland and Norway and Sweden and Norway.*

Groups Greenland & Sweden
Average dissimilarity = 76,87

| Species | Group Greenland Av.Abund | Group Sweden Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| Thalassiosira_sp._J | 71428.4 | 974.4 | 2.42 | 3.15 |
| Lingulodinium_sp._J | 4349.6 | 34859.2 | 1.45 | 1.89 |
| Detonula_confervacea_J | 32176.8 | 430.6 | 1.15 | 1.5 |
| Thalassiosira_sp._R | 33642 | 247.2 | 1.15 | 1.49 |
| Thalassiosira_sp._O | 32848 | 468.2 | 1.12 | 1.45 |
| Lingulodinium_sp._R | 2617.2 | 23643.4 | 0.99 | 1.29 |
| Chaetoceros_cf._neogracile_O | 28693.6 | 1058.4 | 0.96 | 1.25 |
| Chaetocerotaceae_sp._J | 26358.6 | 1772 | 0.95 | 1.24 |
| Lingulodinium_sp._O | 2453.6 | 20477.4 | 0.85 | 1.11 |
| Pseudokeronopsis_sp._Z | 15535 | 1331 | 0.84 | 1.1 |
| Lingulodinium_sp._C | 1477.6 | 18347 | 0.8 | 1.04 |
| Lingulodinium_sp._G | 911.4 | 17655 | 0.79 | 1.02 |
| Stereomyxa_ramosa_S | 12248.6 | 739.6 | 0.77 | 1 |
| Thalassiosira_sp._G | 22105.4 | 233.6 | 0.74 | 0.96 |
| Chaetoceros_cf._neogracile_J | 20285 | 1026.4 | 0.69 | 0.9 |
| Rhizosolenia_setigera_J | 13132.4 | 355 | 0.69 | 0.89 |
| Thalassiosira_sp._P | 19631.2 | 180.6 | 0.66 | 0.86 |
| Paraphysomonas_vestita_Z | 11191 | 924.4 | 0.6 | 0.78 |
| Thalassiosira_sp._E | 17583.2 | 149.6 | 0.59 | 0.77 |
| Lingulodinium_sp._P | 540.8 | 12772.8 | 0.58 | 0.75 |
| Thalassiosira_sp._C | 16150 | 318.2 | 0.55 | 0.71 |
| Chaetoceros_cf._neogracile_R | 14589 | 223.6 | 0.49 | 0.64 |
| Pseudo-nitzschia_pungens_J | 14144 | 45.4 | 0.48 | 0.63 |
| Detonula_confervacea_G | 13704 | 133.2 | 0.46 | 0.6 |
| Minutocellus_polymorphus_J | 11484.4 | 320.6 | 0.44 | 0.58 |
| Chaetocerotaceae_sp._O | 11978.4 | 866.6 | 0.44 | 0.57 |
| Detonula_confervacea_O | 12945 | 309.2 | 0.43 | 0.56 |
| Detonula_confervacea_R | 12566 | 229 | 0.42 | 0.55 |
| Lingulodinium_sp._RTKL | 564.2 | 9220.6 | 0.41 | 0.53 |
| Blepharisma_japonicum_Z | 7367.4 | 520.8 | 0.4 | 0.51 |

Groups Greenland & Norway
Average dissimilarity = 67,82

| Species | Group Greenland Av.Abund | Group Norway Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| Thalassiosira_sp._J | 71428.4 | 12153.4 | 2.08 | 3.06 |
| Chaetocerotaceae_sp._J | 26358.6 | 52341.6 | 1.67 | 2.47 |
| Corethron_hystrix_O | 684.6 | 26733.8 | 1.03 | 1.51 |
| Thalassiosira_sp._R | 33642 | 1602.4 | 0.98 | 1.45 |
| Thalassiosira_sp._O | 32848 | 3544.2 | 0.94 | 1.39 |
| Detonula_confervacea_J | 32176.8 | 4681.4 | 0.9 | 1.32 |
| Chaetoceros_cf._neogracile_O | 28693.6 | 8359.4 | 0.84 | 1.24 |
| Corethron_hystrix_J | 2404 | 22807 | 0.84 | 1.23 |
| Chaetocerotaceae_sp._O | 11978.4 | 26746.6 | 0.83 | 1.22 |
| Corethron_hystrix_R | 552.6 | 21338 | 0.82 | 1.21 |
| Corethron_hystrix_C | 796.4 | 19569.6 | 0.74 | 1.09 |
| Pseudokeronopsis_sp._Z | 15535 | 1652 | 0.7 | 1.03 |
| Thalassiosira_sp._G | 22105.4 | 1845.2 | 0.66 | 0.97 |
| Chaetoceros_cf._neogracile_J | 20285 | 16766.8 | 0.65 | 0.95 |
| Stereomyxa_ramosa_S | 12248.6 | 342.4 | 0.63 | 0.93 |
| Chaetocerotaceae_sp._C | 6326.4 | 19280.6 | 0.62 | 0.91 |
| Thalassiosira_sp._P | 19631.2 | 938.8 | 0.58 | 0.85 |
| Chaetocerotaceae_sp._G | 4490 | 16771.4 | 0.57 | 0.85 |
| Rhizosolenia_setigera_J | 13132.4 | 1991.6 | 0.55 | 0.81 |
| Chaetocerotaceae_sp._R | 7951.6 | 16946.6 | 0.55 | 0.8 |
| Thalassiosira_sp._E | 17583.2 | 1740.2 | 0.52 | 0.77 |
| Corethron_hystrix_E | 386.2 | 13306 | 0.51 | 0.76 |
| Paraphysomonas_vestita_Z | 11191 | 1052.4 | 0.5 | 0.74 |
| Thalassiosira_sp._C | 16150 | 3390.8 | 0.47 | 0.69 |
| Pseudo-nitzschia_pungens_J | 14144 | 1954.4 | 0.44 | 0.65 |
| Chaetocerotaceae_sp._P | 4526.4 | 13620.8 | 0.43 | 0.64 |
| Chaetoceros_cf._neogracile_R | 14589 | 3273.2 | 0.43 | 0.64 |

Groups Sweden & Norway
Average dissimilarity = 65,09

| Species | Group Sweden Av.Abund | Group Norway Av.Abund | Av.Diss | Contrib% |
|---|---|---|---|---|
| Chaetocerotaceae_sp._J | 1772 | 52341.6 | 2.57 | 3.95 |
| Lingulodinium_sp._J | 34859.2 | 6842.4 | 1.5 | 2.31 |
| Corethron_hystrix_O | 196.8 | 26733.8 | 1.4 | 2.15 |
| Chaetocerotaceae_sp._O | 866.6 | 26746.6 | 1.33 | 2.04 |
| Corethron_hystrix_J | 180.8 | 22807 | 1.19 | 1.82 |
| Corethron_hystrix_R | 164.4 | 21338 | 1.12 | 1.71 |
| Corethron_hystrix_C | 126.8 | 19569.6 | 1.02 | 1.57 |
| Lingulodinium_sp._R | 23643.4 | 4496.6 | 1.02 | 1.57 |
| Chaetocerotaceae_sp._C | 576 | 19280.6 | 0.96 | 1.47 |
| Lingulodinium_sp._G | 17655 | 1905.2 | 0.84 | 1.3 |
| Chaetocerotaceae_sp._R | 553.8 | 16946.6 | 0.83 | 1.28 |
| Lingulodinium_sp._O | 20477.4 | 4853.2 | 0.83 | 1.28 |
| Chaetocerotaceae_sp._G | 523.6 | 16771.4 | 0.82 | 1.26 |
| Chaetoceros_cf._neogracile_J | 1026.4 | 16766.8 | 0.82 | 1.26 |
| Lingulodinium_sp._C | 18347 | 3935 | 0.77 | 1.18 |
| Corethron_hystrix_E | 64.2 | 13306 | 0.7 | 1.07 |
| Chaetocerotaceae_sp._P | 416.4 | 13620.8 | 0.67 | 1.04 |
| Lingulodinium_sp._P | 12772.8 | 1476.6 | 0.6 | 0.93 |
| Thalassiosira_sp._J | 974.4 | 12153.4 | 0.58 | 0.89 |

## 3.11 **Enrichment Analysis**

Enrichment Analysis allowed to determine both the species and the COG category displaying a significantly increased expression levels. Enrichment Analysis was carried out by conducting a hypergeometric distribution followed by a Bonferroni correction of the generated p values. This allowed for a selection of transcripts which represented a positive enrichment. A short except is given in figure showing the expression of enriched COG categories for the genus *Chaetoceros* of the individual sampling sites of the three regions. This genus displayed a strong presence of enriched transcripts in Greenland and to a lesser extent in Norway while barely any enriched transcripts were detected in Sweden. Within the Greenland region *Chaetoceros* showed a clear divide between the first two sites and the remaining three. COG categories with highest scores of enrichment over all regions were "predicted only", "PTM, protein turnover and chaperone proteins", "Translation" as well as "Amino acid metabolism" and "Lipid metabolism". As the resulting file in its entirety was judged too extensive for depiction a condensation of the data by creation of word cloud diagrams was chosen. The complete file of the enrichment analysis can be found in the supplemental data segment.

***Table 8:*** *Excerpt of the edited results of the enrichment analysis. Displayed are the enriched counts of the genus* Chaetoceros. *First column states the genus; the second column gives the COG category. Greenlandic region displayed in blue, Swedish region in red and Norwegian region in green. Numbers of enriched transcripts per region were highlighted in heat map style.*

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chaetoceros | A | 4 | 27 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1 | 0 | 0 |
| Chaetoceros | B | 6 | 28 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 7 | 1 | 1 |
| Chaetoceros | C | 46 | 223 | 2 | 9 | 8 | 0 | 0 | 0 | 0 | 1 | 8 | 103 | 100 | 32 | 5 |
| Chaetoceros | D | 8 | 41 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 9 | 0 | 0 |
| Chaetoceros | E | 49 | 240 | 0 | 7 | 2 | 1 | 0 | 0 | 0 | 2 | 3 | 77 | 60 | 7 | 3 |
| Chaetoceros | F | 14 | 79 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 15 | 13 | 0 | 0 |
| Chaetoceros | G | 59 | 136 | 1 | 9 | 3 | 1 | 0 | 0 | 0 | 1 | 2 | 59 | 91 | 22 | 4 |
| Chaetoceros | H | 26 | 90 | 0 | 7 | 3 | 0 | 0 | 0 | 0 | 2 | 5 | 41 | 56 | 17 | 2 |
| Chaetoceros | I | 27 | 107 | 1 | 4 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 6 | 41 | 3 | 1 |
| Chaetoceros | J | 105 | 346 | 1 | 39 | 14 | 4 | 0 | 0 | 1 | 1 | 21 | 129 | 161 | 26 | 11 |
| Chaetoceros | K | 46 | 240 | 2 | 10 | 4 | 1 | 0 | 0 | 1 | 1 | 1 | 36 | 37 | 8 | 0 |
| Chaetoceros | L | 39 | 213 | 1 | 10 | 3 | 1 | 0 | 0 | 1 | 1 | 1 | 35 | 46 | 9 | 0 |
| Chaetoceros | M | 15 | 48 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 18 | 6 | 0 |
| Chaetoceros | N | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| Chaetoceros | O | 79 | 377 | 3 | 16 | 11 | 0 | 0 | 0 | 1 | 4 | 7 | 96 | 111 | 34 | 11 |
| Chaetoceros | P | 29 | 118 | 1 | 5 | 6 | 1 | 0 | 0 | 1 | 0 | 0 | 27 | 56 | 28 | 4 |
| Chaetoceros | Q | 7 | 55 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 17 | 1 | 0 |
| Chaetoceros | R | 119 | 507 | 2 | 9 | 7 | 2 | 0 | 0 | 1 | 4 | 4 | 73 | 111 | 19 | 1 |
| Chaetoceros | S | 24 | 138 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 25 | 2 | 1 |
| Chaetoceros | T | 28 | 150 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 11 | 22 | 3 | 0 |
| Chaetoceros | U | 16 | 91 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 15 | 22 | 3 | 0 |
| Chaetoceros | V | 5 | 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 0 |
| Chaetoceros | Y | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Chaetoceros | Z | 7 | 24 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 5 | 4 | 0 |

### 3.11.1 Visualization of the results of Enrichment analysis

***Region of Greenland***

Word cloud plots are a method for visualization and weighing of datasets. Entries with larger values appear as larger words in the word cloud plots.

The word cloud of the enrichment analysis of the Greenlandic region is given in figure 33. The dominating genera of the figure were the diatoms *Detonula*, *Thalassiosira* and *Chaetoceros*. Other phyla besides Bacillariophyta were observed a very limited number with the genera *Symbiodinium* and *Alexandrium*.

Dominant COG categories fall along the lines of previously described most abundant COG categories.


***Figure 38:*** *Visualization of enrichment results for Greenland by word cloud plot. Maximum of words was limited to 100, additional figures for word limits of 75, and - 150 listed in supplemental data. The strongest magnified entries listed genera* Detonula, Thalassiosira *and* Chaetoceros. *COG categories linked to magnified entries were "PTM, protein turnover, chaperone (O)", "Function Unknown (S)", "General function prediction only (R)" and "Translation (J)".*

***Region of Sweden***

The word cloud of the enrichment analysis of the Swedish region is given in figure 34. Unlike the previous Greenlandic region the Dinophyta were the dominating phylum in the region. The most prominently featured entries listed the genera *Lingulodinium*, *Alexandrium* and *Prorocentrum*. Multiple other Dinophyta were listed in smaller font size. Besides Dinophyta no other phylum appeared with the cloud limitation of 100 words.

Largest front size and thereby most enriched transcripts are visibly dominated by the genus *Lingulodinium*. In regards to expressed COG categories to top spot was given to the category "predicted protein only". This presents a deviation in comparison to the other two regions as while the named category was listed as enriched it took a less prominent spot.

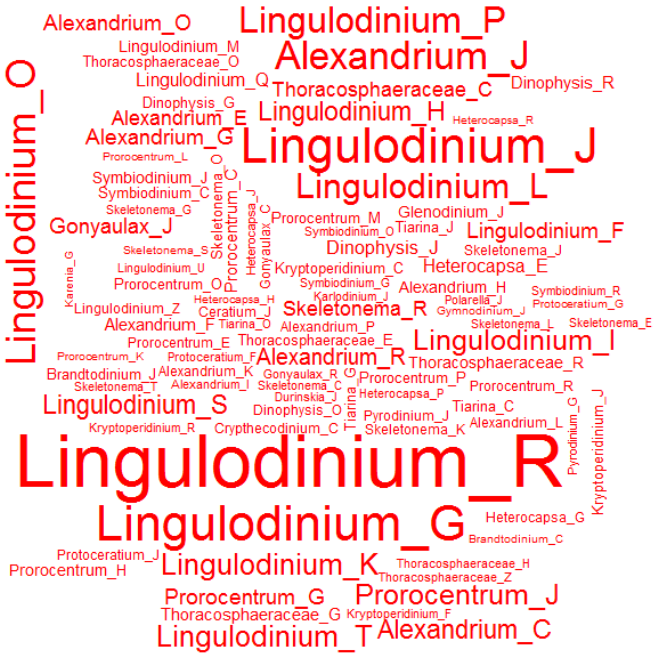



***Figure 39:*** *Visualization of enrichment results for Sweden by word cloud plot. Maximum of words was limited to 100, additional figures for word limits of 75, and - 150 listed in supplemental data. The strongest magnified entries listed genera* Lingulodinium, Alexandrium *and* Prorocentrum. *COG categories linked to magnified entries were "PTM, protein turnover, chaperone (O)", "Carbohydrate metabolism and transport (G)","Lipid metabolism (L)", "General function prediction only (R)" and "Translation (J)".*

### *Region of Norway*

The word cloud of the enrichment analysis of the Norwegian region is given in figure 40. The species composition appeared similar to the region of Greenland given the most prominently featured genera *Corethron*, *Cheatoceros* and other unidentified genera of the family *Cheatocerotaceae* belong to the phylum Bacillariophyta. Besides Diatoms a small number of Dinophyta were listed, namely the genera *Azadinium* and *Gonyaulax* as well as isolated listings of members of the phylum Ciliophoran. Overall the word cloud plots displayed a fairly even distribution of front sizes and in comparison to the other regions a visibly limited number of listed species.

In regards to listed COG categories in the Norwegian region the category "Translation (J)" is featured less noticeably than in the two previous regions. Most prominently featured COG categories linked to magnified entries were "PTM, protein turnover, chaperone (O)",

"Transcription (K)","Energy production and conversion (C)" and "Amino acid metabolism (E)".

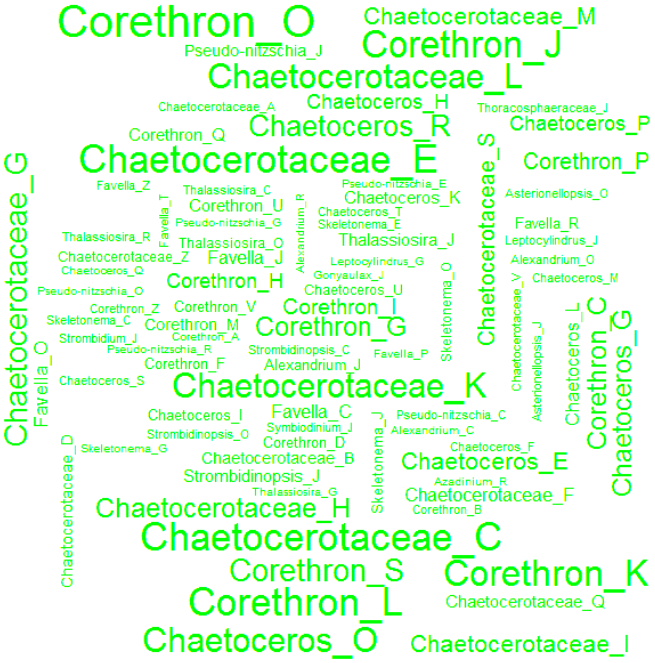



***Figure 40:*** *Visualization of enrichment results for Norway by word cloud plot. Maximum of words was limited to 100, additional figures for word limits of 75, and - 150 listed in supplemental data. The strongest magnified entries were mostly Corethron and unidentified genera of the family Chaetocerotaceae. COG categories linked to magnified entries were "PTM, protein turnover, chaperone (O)", "Transcription (K)","Energy production and conversion (C)", "Lipid metabolism (L) and "Translation (J)".*

# 4 Discussion

The main objective of the project was the characterisation of the three sampled regions in regards to the environmental conditions, their biodiversity and generation of expression profiles by functional metatranscriptomics. In addition for determination of biodiversity and the functional annotation of cDNA, the project used a multi-staged approach. Firstly by sequencing of 28S rRNA LSU region contained in the samples followed by phylogenetic placement of the sequences with in house reference trees for the determination of the biodiversity. For the functional metatranscriptomics the sequenced cDNA was compared by BLAST algorithm against in-house data bank with transcriptomes of eukaryotic marine organisms. Final objectives include an enrichment analysis of the expression profiles to identify whether a species expresses a functional class of protein significantly more than other species express this protein class as well as the determination of correlations between biodiversity of sampled regions, expressed protein classes in the different regions or the expression profiles of detected species.

## 4.1 Environmental characterization of sampled regions

Highest measured temperatures of the sampled regions were the Swedish sample sites which likes was a result of the fjord system being the shallowest water body as well as the most southern. Major influence of terrestrial freshwater was observed in the Greenlandic region around St513 and St514 near the Ilulissat glacier. The localized influx of freshwater resulted in a visible local increase in temperature (figure 6), decrease in salinity (figure 7). By extension the decreased temperature may be the driving factor behind the reduced occurrence of chlorophyll a around the glacier (figure 9). Highest observed amounts of chlorophyll a was observed in the Greenlandic Vaigat strait correlating with the highest amount of observed oxygen (figure 8). Swedish and Norwegian regions displayed little amounts of dissolved nutrients (figures 10 - 14) which might be the explanation of the observed low occurrence of chlorophyll a. For the Greenlandic sample sites nutrient depletion was observed only in areas with high chlorophyll a content while deeper parts of the water columns contained by nitrate, phosphate and ammonium (figures 11, 12 and 14). Silicate was observed in the Swedish region in deeper waters as well as outside of the fjords system in the Kattegat which might be the results of an absence of Bacillariophyta.

## 4.2 Assessment of biodiversity by molecular markers

The two sequential rRNA approaches allowed the comparison of the two methods in regards to the overarching species abundances, plotted on the taxonomic rank phylum, especially the overviews of the three regions (figures 12 and 16) were mostly similar regarding the composition of the observed taxa. Both methods determined for the Greenlandic Disko Bay the presence of both Bacillariophyta and Dinophyta in roughly even abundance, while the

Swedish fjord dominantly displayed Dinophyta with a minor abundance of Bacillariophyta. The sampled sound in the Lofoten predominantly displayed Bacillariophyta. These findings proved that on the taxonomic rank phylum, both methods, the 28S rRNA LSU barcoding and the expanded phylogenetic placement of the results approach provided similar results. So far no publications are available utilising rRNA metabarcoding and subsequently phylogenetic placement with reference trees to refine the results of the metabarcoding. The similarity in composition of the observed taxa between the two methods proves the validity of the chosen approach.

A number of differences in community composition were observed between QIIME – and phylogenetic placement derived taxa abundances and diversity. Foremost the share of sequences allocated to the categories "Unassigned" and "other" were considerably reduced for the metagenomics derived abundances. This was especially visible for the regions Greenland and Norway (figures 22 and 23). The rRNA approach here was not able to assign about a fifth of all sequences to a phylum. The phylogenetic placement approach on the other hand was able to almost all sequences to a taxon. This was the backdrop which promoted the idea of a refining the results of the QIIME search by additional phylogenetic placement with appropriate reference trees. The observed decreased number of unassigned sequences for the metagenomics approach was therefore an indication for the successful refinement of the QIIME data. Other differences observed between the two methods were, that none of the QIIME derived OTU abundances (figures 17 - 20) displayed any Alveolates besides Dinophyta, no other Stramenophiles besides diatoms and no Haptophyta or Viridiplantae at all. This also proved the difficulties the rRNA approach has in differentiating closely related species. After phylogenetic placement all these taxa were observed, proving yet again that the additional step of phylogenetic placement allows for a more differentiated taxonomic OTU determination. In the other direction in the 28S rRNA QIIME derived abundances a small number of sequences were allotted to several Metazoa, a subkingdom that was not included in the reference trees for the phylogenetic placement and thereby could not be detected with the method.

## 4.3 Analysis of metagenomic results

When looking more in depth at the intra-region distribution of taxa and species abundances the findings were to an extent predicted by the habitat clustering tests in figures 16a and 16b. The case of the Greenlandic region both the NMDS (figures 16a) as well as the heat map (figures 16b) displayed the five sample stations forming two separate clusters. The first consisting of stations St511 and St512 and the remaining three stations forming the second cluster. The biodiversity results of both QIIME and phylogenetic placement confirmed that divide and substantiated the observed separate clustering with a deviating composition of the local populations. Both methods determined the Bacillariophyta abundance > 90 % of all OTUs/allotted contigs for St511 and St512. The other three stations contained Bacillariophyta abundance less than 30%. The in the NMDS plot observed divide between the Greenlandic sites can be traced back to the compositions of the local population. Observed species

abundances of Swedish region also matched with the results of the habitat clustering as the determined small clustering of the five Swedish sites mirrored the highly similar species composition of both rRNA methods (figures 20, 24). In case of the Norwegian region the species abundance results were not matching to the results of the habitat clustering. The clustering stressed the individuality of the site St001 while clustering the remaining four sites closely together (figure 16a, 16b). This was not confirmed with depicted species compositions of the Norwegian regions. Instead of observable disparity of St001 to the other stations the subsequent stations displayed a gradual decrease of Bacillariophyta through the Sortlandsound from station St001 at the northern mouth of the sound through station St005 near the Island Hadseløya.

When comparing the determined species diversity with previously published results the best matching results were for the Swedish fjord system (Godhe 2001, McQuoid 2005, Harland 2006). The publications described sediment studies conducted in the Gullmar fjord just north of the fjord system of this study and the observed Dinophyta species are consistent with the ones observed in this study. The high abundance of Dinophyta in the Swedish fjord system can also be linked to anthropological influences. Dinophyta were observed to thrive in hypertrophic waters (Dale 2001, Matsuoka 2003). Of the three sampled regions it stands to reason that the Swedish fjord system would be subjected to the highest anthropological influence. Along this line of thought the highest abundance of Dinophyta are to be found in the Swedish fjord followed by the Norwegian sound and finally the Greenland Disko Bay where the human influence should be the lowest. This is the case for both the diversity derived from the metatranscriptomic data and with some limitations the diversity derived from rRNA. Here the Greenlandic and Norwegian regions displayed approximately equal abundances of Dinophyta. In terms of species composition of the Greenlandic populations the determined species were in accordance of previously observed species of the Greenlandic West coastal waters (Krawczyk 2014).

Two observations have to be made for the taxonomical group of Haptophyta. The fact that Haptophyta were observed in smaller abundances compared to Dinophyta and Bacillariophyta might not necessarily represent the actual abundances of Haptophyta. Due to a lack of available data no phylogenetic tree for phylogenetic placement could be constructed for Haptophyta which might have led to underrepresentation of the phylum. In addition the used primer for amplification of the 28S LSU rRNA region was reported to perform not optimally for Haptophyta (Bittner 2013). This may also have compacted the potential underrepresentation of the Haptophyta.

## 4.4 Determination of biological activity by metatranscriptomics

The analysis of the metatransciptome included the intra – and inter region comparisons of expression profiles and allotting the observed the observed COG categories into the three larger functional groups of “information storage and processing”, “cellular processing and signalling” and “metabolism” to better visualise the metabolically active pathways.

Less than 15 % annotated reads of all regions could not be assigned to a functional COG category. Around a tenth of the reads fell into the COG category “general function only predicted” and around another 2 - 3 % for “function unknown”.

The intra region expression profiles (figures 28 - 30) of the three regions displayed comparable expressions of the different COG categories for all five corresponding sites. The numbers of transcripts evaluated in the individual sampling sites were in a comparable magnitude for the three regions. Greenlandic sites yielded most reads followed by the Norwegian sites while the sites of the Swedish regions yielded the fewest reads. The difference in numbers of transcripts numbers the differences in sequences observed during the metagenomics approach (figures 17 – 24). Those differences may be caused by a general higher number of species in the Greenlandic sample regions. The measured concentrations of chlorophyll a content (figure 9) represent the amount of phototrophic organisms in the waters of the sample sites. As the sample sites of Greenland contained the most chlorophyll it can be equalled to the highest species richness.

For all three regions the highest scored COG categories were “Translation”, “PTM, protein turnover and chaperone proteins” and “Energy production and conversion”. Due to this globally observed pattern of distribution it stands to reason that transcripts allotted into these categories belong to the primary metabolism and have more of a housekeeping role and their expression is to be expected. The observed highest scored categories overlapped with previously published results for all three taxonomical groups Bacillariophyta (Mock 2005), Dinophyta (Jaeckisch 2011) as well as Haptophyta (Claire 2005).

Differences in the three regions were more visible in the regional comparisons of expressions (figures 33 - 34). Due the differences in transcript count between regions the regular pattern of expression appeared to be the Greenlandic region with the most transcripts, the Norwegian sites with around 2/3 of the Greenlandic transcripts followed by the Swedish fjord system with around 1/3 of the Greenlandic transcripts. A possible explanation for the observed differences in numbers may be found in the time of the expedition. The Greenlandic region was sampled during an expedition in July of 2012 while Norwegian and Swedish regions were sampled during August and September of 2014. The observed differences in numbers of transcripts between regions therefore can be the results of a delayed spring bloom in Greenlandic and Norwegian sample sites. As the two regions are located further north the spring bloom occurs later in the year. The sampling of Greenlandic and Norwegian sites may have fallen in the time or shortly after of the spring bloom, resulting in the observed higher number of transcripts. As the expedition moved from north to south, the Swedish region was sampled after the Norwegian region and at the time of sample collection the spring bloom was most likely long past and number of organisms were likely to be lower compared to the other two regions.

Deviations from the tentatively established pattern were distinctive of the three regions. The Greenlandic expression profile displayed decreased activity in the categories “Cell cycle control and mitosis”, “Nucleotide metabolism and transport”, “Nuclear structure”, “RNA processing and modifications” and “Coenzyme metabolism”. The combination of said

categories and the fact that they were all expressed to a lesser extent in Greenland than compared to the other two reasons lend themselves to reason that the Greenlandic sites at the time of the sampling contained mostly mature cells with reduced metabolism.

The Norwegian expression profiles displayed more deviations than the Greenlandic ones. Under expressed categories were “Cytoskeleton” and “Cell motility” while the categories “Energy production”, “Defence mechanisms”, “Coenzyme metabolism” and “Nucleotide metabolism and transport” were expressed more strongly than compared to the other two regions. The expressed proteins of the category “Defence mechanisms” were traced to the two genera *Corethron* and not further characterized members of the *Chaetocerotaceae* family. Within the category “Defence mechanisms” three different COG-Genes: (ABC) transporter (COG1131, COG1132, COG0842), Beta-lactamase (COG1680) and Mate efflux family protein (COG0534) were identified, in part with multiple COG-Ids. So the combination of the differently expressed COG categories led to postulate that at least parts of the Norwegian populations were exposed to stress factors.

Swedish region was less conclusive. Only the two categories “Signal transduction” and “Nucleotide metabolism and transport” were more prominently expressed and it was difficult to explain an eventual context between an overexpression of both those categories.

*Functional metatranscriptomics of the taxonomical group Dinophyta*

In the taxonomical group of the Dinophyta the majority of the expressed COG categories were dominated by the two species *Lingulodinium* and *Alexandrium.* In terms of COG expression a number of other Dinophyta expressed categories to a noteworthy aspect. Especially the tested Swedish fjord system displayed a variety of Dinophyta which made an impact on various COG categories such as the species *Symbiodinium*, *Scrippsiella*, *Prorocentrum* and *Azadinium.* In the Greenlandic region of the Disko Bay the species *Lingulodinium* and *Alexandrium* were noteworthy in most COG categories while the species *Symbiodinium* made a one-time appearance in the category “Chromatin structure and dynamics”. In the Norwegian region only *Lingulodinium* and *Alexandrium* were observed with notable expression levels in most protein categories. In terms of protein expression therefore *Lingulodinium* and *Alexandrium* were observed globally in all regions.

The above described species abundance by protein expression was compared to the two other analyses of species diversity featured in this study by QIIME 28S LSU rRNA analysis (figures 17-21) and the diversity determined by phylogenetic placement (figures 22-25) in. Abundances in species diversity determined by phylogenetic placement were different from the species abundance received by protein expression in the form that only three Dinophyta species were determined. *Prorocentrum* and *Alexandrium* were determined in all regions and in addition *Lingulodinium* occurred in the Swedish and Norwegian region only. QIIME derived abundances differed yet again as only two species were determined. Both *Scrippsiella* and *Alexandrium* were found over all three regions.

*Functional metatranscriptomics of taxonomical group Bacillariophyta*

Expression of COG categories for diatoms differed for the three regions. Swedish stations displayed comparatively little expression of diatoms. In Greenland the species most notable were *Thalassiosira, Cheatoceros* and most of all *Detonula*. Norwegian sites displayed notable COG expression for *Cheatoceros* as well but *Corethron* was observed to a greater effect.

Abundances determined by 28S rRNA phylogenetic placement and metagenomics differed from the species composition derived by protein expression levels. While the expression levels did not show a notable occurrence of diatoms in the Swedish sample sites, both the 28S rRNA phylogenetic placement - and metagenomics results presented notable abundances of three diatoms species. Both methods detected contingents of *Thalassiosira* and by 28S rRNA phylogenetic placement *Cheatoceros* was found in addition. For the Greenlandic region metagenomics only detected *Thalassiosira* in noticeable abundances while rRNA phylogenetic placement proved the presence of large abundances of *Thalassiosira* as well as less pronounced populations of *Cheatoceros* and *Detonula.* In the Norwegian sample sites rRNA phylogenetic placement detected the species *Thalassiosira* and *Cheatoceros* while metagenomics only found *Thalassiosira,* mirroring the results of the two previous regions.

## 4.5 Correlation between sampled regions

Independently from determination of species abundance and biological activity the potential correlations of the sampled region was studied. The Mantel test (table 2) functioned as the quickest and the simplest method. As only the first pairwise comparison of the count-tables before the metagenomics against the reference databank and the 28S rRNA placement enquiry were shown to have a correlation the results were unexpected. Possibly the Mantel test results were negative because of the large number of contigs without a matched species after the BLAST search. As these unmatched contigs were excluded from the further steps the count-tables used in the Mantel test changed considerably after BLAST search thereby mitigating the correlation.

While the Mantel test was inconclusive the subsequently conducted SIMPER Analysis went into far more detail which allowed characterizing the relations of the three regions further. The results in table 3 showed that in terms of biological activity the sites of Sweden and Norway were the most similar. When comparing the species abundances (derived by metagenomics BLAST search) however (table 5) the similarities of pairwise comparisons of Norway and Greenland as well as Norway and Sweden were almost identical. This was compacted when comparing both species abundance and biological activity (table 7) where both pairwise comparisons of Norway with Greenland as well as Norway with Sweden were still fairly comparable.

Therefore the SIMPER analysis showed that the regions of Greenland and Sweden were the most divergent in both species composition as well as biological activity. Norway was comparable in both aspects to both other regions. Both observations were not necessarily relatable to the earlier conducted habitat clustering (figures 16a, 16b).

*Results of the Enrichment Analysis*

An aspect noted during the SIMPER analysis was that in some cases a given species seemed to feature dominantly in a given region while not being present in others, for example the species *Thalassiosira* for Greenlandic region (table 5). This prompted the enrichment analysis to study this aspect in detail.

While the complete results of the enrichment analysis were delegated to the supplemental data section word cloud figures (figures 38 – 40) were used to condense the results. The enrichments analysis verified the elevated occurrence of *Thalassiosira, Cheatoceros* and *Detonula* for the Greenlandic region, *Alexandrium* and especially *Lingulodinium* for Swedish region and *Corethron* and not further characterized members of *Chaetocerotaceae* family for the Norwegian region. So while the domineering presence of these species were hinted at least in part in the RA-plots (figures 25 - 28), the processed expression profiles of individual species in figures 35 - 37 and supplemental data and the results of the SIMPER test (tables 2 - 7), the enrichment analysis showed that these species did indeed occur at elevated levels. As the word cloud figures combined the species with the expressed COG categories these results were compared with the results of protein expression (table 1 and figures 31a-c) and matched those without exception.

The SIMPER - as well as the enrichment analysis included the COG categories *"Function Unknown (S)"* and *"General function prediction only (R)".* Both categories were omitted in the proteomics part of the study as they contained no functional information and were not relevant at that point of the study. They were included in SIMPER and enrichment analysis to verify the need for more completely genome sequences of marine plankton as even enriched parts of the currently studied transcriptome could not be matched to a function.

Apart from the condensed results in the word clouds figures the results of the enrichment analysis in the supplemental data section allow to study the distribution of the enriched species within the sampled regions. In Sweden the dominant Dinophyta *Alexandrium* was distributed fairly evenly throughout all stations. *Lingulodinium* however was found mainly in St029. This could not be explained satisfyingly as no previous parts of the study (environmental characterization, habitat clustering, …) suggested a noticeable difference between any of the Swedish sample sites. For the Greenlandic region the enriched Bacillariophyta species *Detonula* and *Thalassiosira* occurred almost exclusively in the first two sampled sites St511 and St512. Along the same lines in the Norwegian region the diatom *Corethron* was found exclusively in the first site St001.

It was previously determined in the habitat clustering (figure 16b) that in the Norwegian region St001 was clustered apart from the remaining four stations and the Greenlandic stations St511 and St512 formed their own cluster. For the Greenlandic stations in addition to

the findings of the habitat clustering also findings of the biodiversity part of the study (figures 13 – 24) stressed the differences between St511 and St512 and the remaining three sample sites. Therefore the enrichment analysis allowed naming the actual species, which made up the differences in habitat clustering and biodiversity.

### 4.6 Methodical challenges of the chosen approach

The driving reason behind following the step of barcoding with rRNA with the additional phylogenetic placement is determined by the difficulty of rRNA barcoding to distinguish closely related taxa (Whitworth 2007). This limited resolution makes rRNA barcoding in taxonomic studies only reliable to the taxonomic level of phylum and may struggle beyond, especially for taxas with limited verified references – as is often the case for marine eukaryotic microorganisms (Piganeau 2010, Coissac 2012, Kipling 2016).

As metatranscriptomic studies of marine eukaryotic microbes are currently still rare, a problem is the lack of reference datasets for phylogenetic placement. The usage of already available general purpose databanks (NCBI EST collection, NIH GenBank genetic sequence database) as reference for sequences of marine origin has been shown to lead to large amounts of unassigned - or incorrectly identified sequences (Poretsky 2008). Therefore a sub goal to the taxonomical analysis of the assembled cDNA sequences of the metatranscriptomic dataset was the creation of an in-house reference databank. That databank from the available EST dataset of marine microalgae was used instead of established general purpose databanks as the narrower scope would cut down the number of un - or misassigned sequences. This was observed to great effect in the plotted abundances between rRNA LSU (figures 17-20) and the BLAST derived abundances (figures 20-24). The amount of unassigned transcripts was visibly reduced in the refined approach, confirming the hypothesis.

## 5. Conclusion and future outlook

The main objectives associated with the creation of a reference databank were achieved. As a) the results of metagenomic BLAST derived OTU abundances were comparable on a phylum level to the 28S rRNA LSU QIIME OTU abundances and b) the amount of unassigned sequences was reduced considerably. As a considerable amount of contigs in the assembly could not be matched it stands to reason that the created databank might be improved for further studies with more sequenced and annotated genomes as they become available. As the QIIME pipeline had assigned contigs to the phylum Metazoa which currently is not represented in the reference databank this may represent a valuable starting point for a future extension.

Methodological difficulties during the creation of the reference databank were firstly the limited availability of relevant sequenced and annotated marine genomes. The MMETSP project, which has migrated to iPlant since, was a valuable starting point, but the used sequences contained duplicates as well as either out-dated or inaccurate taxonomical

information for the associated species. This represents the second challenge during the construction of the databank, as it was proven to be a time intensive process which will continue with the further maintenance and expansion of the databank. As the groundwork is done and proven to work reliably, in the future the expansion of the databank should be possible and work seamlessly.

Regarding phylogenetic placement by the phylogenetic trees, such as the ones for Dinophyta and Bacillariophyta that were used in this study, it was observed that phylogenetic reference trees differ based on the used base data (Beszteri 2007). Beszteri et al. reported that trees based on rRNA, mitochondrial DNA and plastidal DNA differed from another. This study used Dinophyta and Bacillariophyta trees based on rRNA sequences containing both SSU and LSU motives. It is currently unknown if phylogenetic trees based on rRNA are the most accurate or if DNA from cell organelles is more useful for phylogenetic placements.

The conducted enrichment analysis proved to be valuable in consolidating the observations of NMDS and determination of species abundance. The analysis allowed tracing the difference in clustering of the Greenlandic station down to the individual species of *Thalassiosira* and *Detonula* and in the case of the Norwegian St001 the species *Corethron*. In the currents form featured in this study the enrichment analysis only covers overexpression of transcripts. A future instalment could include under expression which might valuable insights of the metatranscriptome of the samples regions. Along the same lines the generated RA plots can be enriched by adding labels to the individual contigs to allow in depth study of regional distribution of contigs after various functional annotations as displayed by Marchetti et al. for the influence of iron enrichment of transcriptomes (Marchetti 2012). In this regard many datasets generated in this project which could not be fully analysed in depth due to time restrains can serve as a scaffold for future publications.

Key feature of the project was the comparison between three different regions. The data featured in this project were gathered in two expeditions two years apart. Even within an expedition the travel time of the expedition vessel between two regions were several weeks. Within the project during the environmental characterisation as well as during the metagenomics the time-wise differences in collection of the samples may have had an impact. This represents a weakness in the experimental setup which cannot be avoided in future projects as simultaneous sampling of the three regions is neither logistically nor financially viable.

# 6. References


Alexander (2015). "Metatranscriptome analyses indicate resource partitioning between diatoms in the field." PNAS.

Alexander, H. (2015). "Metatranscriptome analyses indicate resource partitioning between diatoms in the field." PNAS.

Bachvaroff (2009). "Expressed sequence tags from Amoebophrya sp. infecting Karlodinium vneficum." J Eukaryot. Microbiol.

Beja (2000). "Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea." Science **289**: 1902-1906.

Beja (2004). "To BAC or not to BAC: marine ecogenomics." Current Opinion in Biotechnology **15**.

Bender (2014). "Transcriptional responses of three model diatoms to nitrate limitation of growth." Fronters in marine science **1**.

Beszteri, B. (2007). "An assessment of cryptic genetic diversity within the Cyclotella meneghiniana species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms." European Journal of Phycology **42**.

Beszteri, B. (2011). "Transcriptomic response of the toxic prymnesiophyte Prymnesium parvum to phosphorus and nitrogen starvation." Harmful Algae.

Bhattacharya, D. (2013). "Genome of the red alga Porphyridium purpureum." Nat. Com.

Bittner, L. (2013). "Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay." Molecular Ecology **22**.

Burki, F. (2014). "Rhizaria." Current Biology **24**(3).

Campo, J. d. (2014). "The others: our biased perspective of

eukaryotic genomes." Trends in Ecology & Evolution **29**(5).

Caporaso (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature Methods.

Chambouvet (2008). "Control of Toxic Marine Dinoflagellate Blooms by Serial Parasitic Killers." Science **322**.

Chan, C. X. (2012). "Analysis of dinoflagellate genes reveals the remarkably complex evolutionary history of a microbial eukaryote." J. Phycol.

Claire, J. W. L. (2005). "Analysis of Expressed Sequence Tags from the Harmful Alga, Prymnesium parvum " Marine Biotechnology **8**.

Clarke (1993). "Non-parametric multivariate analyses of changes in community structure." Australian Journal of Ecology.

Coissac, E. (2012). "Bioinformatic challenges for DNA metabarcoding of plants and animals." Molecular Ecology **21**.

Curtis, B. A. (2012). "Cryptophyte and chlorarachniophyte nuclear genomes reveal evolutionarymosaicism and the fate of nucleomorphs." Nature.

Cuvelier, M. (2010). "Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton." Proc Natl Acad Sci U S A.

Dale, B. (2001). "Marine dinoflagellate cysts as indicators of eutrophication and industrial pollution: a discussion." The Science of the Total Environment **264**.

DeLong, E. F. (2014). "Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior." Science **311**: 496-503.

DMI, D. M. i. (2007). "<dmi_sci.report_2._del Diskobay.pdf>."

Drange, H. (2005). "Nordic Seas: An Integrated Perspective." American Geophysical Union.

Dray (2016). "Package "ade4"."

Edgar (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics.

Edgcomb, V. (2011). "Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness." ISME.

Fellows (2013). "Package "wordcloud"."

Gilbert, J. A. (2011). "Microbial Metagenomics: Beyond the Genome." Annu. Rev. Marine. Sci **3**.

Godhe, A. (2001). "Relationship between planktonic dinoflagellate abundance, cysts recovered in sediment traps and environmental factors in the Gullmar Fjord, Sweden." JOURNAL OF PLANKTON RESEARCH **23**(3).

Grabherr, M. G. (2013). "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data." Nat Biotechnol **29**(7).

Handelsman, J. (1998). "Molecular biological access to the chemistry of unknown 0095 SP soil microbes: a new frontier for natural products." Chemistry & Biology **8**.

Harland, R. (2006). "Dinof lagellate cysts and hydrographical change in Gullmar Fjord, west coast of Sweden." Science of the Total Environment **355**.

Hillis, D. D., Michael (1991). "Ribosomal DNA; Molecular evolution and phylogenic interference.pdf." Quarterly Review of Biology **66**: 411-453.

Jaeckisch, N. (2011). "Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate Alexandrium ostenfeldii." PLoS ONE.

Jaeckisch, N. (2011). "Comparative Genomic and Transcriptomic Characterization of the Toxigenic Marine Dinoflagellate Alexandrium ostenfeldii." PLoS ONE **6**(12).

Jardillier, L. (2010). "Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean." ISME.

John (2001). "A comparative approach to study inhibition of grazing and lipid composition of a toxic and non-toxic clone of Chrysochromulina polylepis." Harmful Algae.

John, T., Hülskötter, Wohlrab (2015). "Intraspecific facilitation by allelochemical mediated grazing protection within a toxigenic dinoflagellate population." Proceedings Royal Society.

John, U., et al. (2003). "The application of a molecular clock based on molecular sequences and the fossil record to explain biogeographic distributions within the Alexandrium tamarense "species complex" (Dinophyceae)." Mol Biol Evol **20**(7): 1015-1027.

Kanehisa, M. (2013). "Data, information, knowledge and principle: back to metabolism in KEGG." Nucleic Acids Research **42**.

Keeling, P. J. (2008). "Horizontal gene transfer in eukaryotic evolution." Nat. Rev. Genet.

Keeling, P. J. (2014). "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing." PLoS **12**(6).

Kipling, W. (2016). "The Perils of DNA Barcoding and the Need for Integrative Taxonomy." Syst. Biol **54**.

Krawczyk (2014). "Description of diatoms from the Southwest to West Greenland coastal and open marine waters." Polar Biol **37**.

Krell (2008). "A new class of ice-binding proteins discovered in a salt-stress-induced cDNA library of the psychrophilic diatom Fragilariopsis cylindrus." European Journal of Phycology **43**.

Lima-Mendez, G. (2015). "Determinants of community structure in the global plankton interactome." Science of the Total Environment.

Love (2016). "Differential analysis of count data - the DESeq2 package."

Lu, Y. (2014). "Genomic Insights into Processes Driving the Infection of Alexandrium tamarense by the Parasitoid Amoebophrya sp." Eukaryotic Cell.

Maheswari (2010). "Digital expression profiling of novel diatom transcripts provides insight into their biological functions." Genome Biology **11**.

Marchetti (2012). "Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability." PNAS **109**(6).

Marchetti (2016). "MANTA: Microbial Assemblage Normalized Transcript Analysis."

Matsuoka, K. (2003). "Modern dinoflagellate cysts in hypertrophic coastal waters of Tokyo Bay, Japan." JOURNAL OF PLANKTON RESEARCH **25**(12).

McQuoid, M. R. (2005). "Influence of salinity on seasonal germination of resting stages and composition of microplankton on the Swedish west coast." MARINE ECOLOGY PROGRESS SERIES **289**.

Medlin (1998). "Phylogenic analysis of marine phytoplankton."

Mock (2005). "Analysis Of Expressed Sequence Tags (ESTS) From The Polar Diatoms Fragilariopsis Cylindrus." J. Phycol **42**.

Moustafa, A. (2009). "Science " Genomic footprints of a cryptic plastid endosymbiosis in diatoms.

Oksanen (2013). "Package "vegan"."

Parker, M. S. (2005). "Synergistic Eeefects Of Light, Temperature, And Nitrogen Source On Transcription Of Genes For Carbon And Nitrogen Metabolism In The Centric Diatom Thalassiosira Pseudonana." J. Phycol **41**.

Pawlowski, J. (2012). "CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms." PLoS ONE.

Pearson (2015). "Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula." ISME.

Piganeau, G. (2010). "How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes." PLoS ONE **6**(2).

Poretsky, R. S. (2008). "Comparative day/night metatranscriptomic analysis

of microbial communities in the North Pacific

subtropical gyre." Environmental Microbiology **11**(6): 1358–1375.

Ribergaard, M. H. (2013). "<Oceanographic Investigations off West Greenland 2012.pdf>." Danish Meterological Institute Center for Ocean and Ice.

Roy, R. S. (2014). "Single cell genome analysis of an uncultured heterotrophic stramenopile." Scientific Reports.

Schlitzer, R. (2016). "Exploring Marine Data with Ocean Data View (ODV) – Advanced Course." MARUM / GLOMAR Basic Skills and Methods Course.

SMHI (2013). "SMHI annual report 2013."

Sonnenberg, R., et al. (2007). "An evaluation of LSU rDNA D1-D2 sequences for their use in species identification." Front Zool **4**: 6.

Spector (1984). "Dinoflagellates." Elsevier(1).

Stecher, A. (2015). "rRNA and rDNA based assessment of sea ice protist biodiversity from the central Arctic Ocean." European Journal of Phycology **51**.

Tatusov (2000). "The COG database: a tool for genome scale analysis of protein functions and evolution." Nucleic Acids Research **28**(1).

Tillmann (2002). "Toxic effects of Alexandrium spp. on heterotropic dinoflagellates: an allelochemical defece mechanism independent of PSP-toxin content." MARINE ECOLOGY PROGRESS SERIES **230**.

Tillmann (2011). "A new non-toxic species in the dinoflagellate genus Azadinium: A. poporum sp. nov." European Journal of Phycology.

Tillmann (2015). "First record of Amphidoma languida and Azadinium dexteroporum from the Irminger Sea." Marine Biodiversity Records.

Toebe, K., et al. (2013). "Molecular discrimination of toxic and non-toxic Alexandrium species (Dinophyta) in natural phytoplankton assemblages from the Scottish coast of the North Sea." European Journal of Phycology **48**(1): 12-26.

Toebe, K., et al. (2012). "Molecular discrimination of taxa within the dinoflagellate genus Azadinium, the source of azaspiracid toxins." JOURNAL OF PLANKTON RESEARCH **35**(1): 225-230.

Toseland (2013). "The impact of temperature on marine phytoplankton resource allocation and metabolism." NATURE CLIMATE CHANGE **3**.

Uhlig, C. (2015). "In situ expression of eukaryotic ice-binding proteins in microbial communities of Arctic and Antarctic sea ice." ISME.

Vargas, C. d. (2015). "Eukaryotic plankton diversity in the sunlit ocean." Science of the Total Environment.

Venter, J. C. (2004). "Environmental Genome Shotgun Sequencing of the Sargasso Sea." Science **304**.

Verity, P. G. (1996). "Organism life cycles, predation, and the structure of marine pelagic ecosystems." MARINE ECOLOGY PROGRESS SERIES **130**.

Waal, V. d. (2015). "Characterization of multiple isolates from an Alexandrium ostenfeldii bloom in The Netherlands " Harmful Algae **49**.

Westphal (2013). "Biodiversity pattern discrimination from Greenlandic and Icelandic coastal water by amplicon deep sequencing."

Whitworth, T. L. (2007). "DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora." Proc. R. Soc. B **274**.

Wohlrab (2010). "A Molecular and Co-Evolutionary Context for Grazer Induced Toxin Production in Alexandrium tamarense." PLoS ONE.

Worden, A. Z. (2015). "Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes." Science **347**(6223).

Yang, I. (2011). "Growth- and nutrient-dependent gene expression in the toxigenic marine dinoflagellate Alexandrium minutum." Harmful Algae.

# 7. List of figures:

# 8. List of tables:



# 9. Supplemental data:

For supplemental data please see the attached CD-ROM