Biogeography of Heterotrophic Flagellate Populations Indicates the Presence of Generalist and Specialist Taxa in the Arctic Ocean

MARY THALER and CONNIE LOVEJOY*

Département de Biologie, Québec-Océan and Institut de Biologie Intégrative et des Systèmes (IBIS), Unité Mixte Internationale Takuvik CNRS France –Canada, Université Laval, Québec, Québec G1V 0A6, Canada

*Corresponding Author. Address: Département de Biologie, Québec-Océan and Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Québec G1V 0A6, Canada. Telephone: 1-418-656-2007 Fax: 1-418-656-7176. Email: connie.lovejoy@bio.ulaval.ca
Abstract

Heterotrophic marine flagellates (HF) are ubiquitous in the world’s oceans and represented in nearly all branches of the Domain Eukaryota. However, the factors determining distributions of major taxonomic groups are poorly known. The Arctic Ocean is a good model environment for examining the distribution of functionally similar but phylogenetically diverse HF because the physical oceanography and annual ice cycles result in distinct environments that could select for microbial communities or favour specific taxa. We re-analyzed new and previously published high throughput sequencing data from multiple studies in the Arctic Ocean, to identify broad patterns in the distribution of individual taxa. HF accounted for fewer than 2% to over half of the reads from the water column, and up to 60% of reads from ice, which was dominated by *Cryothecomonas*. In the water column, many HF phylotypes belonging to Telonemia, Picozoa, uncultured marine stramenpiles (MAST), and choanoflagellates were geographically widely distributed. However, for two groups in particular, Telonemia and *Cryothecomonas*, some species level taxa showed more restricted distributions. For example, several phylotypes of Telonemia favoured open waters with lower nutrients such as the Canada Basin and offshore of the Mackenzie Shelf. In summary, we found that while some Arctic HF were successful over a range of conditions others could be specialists that occur under particular conditions. We conclude that tracking species level diversity in HF is not only feasible, but provides a potential tool for understanding the response of marine microbial ecosystems to rapidly changing ice regimes.
Introduction

Heterotrophic flagellates (HF) have a central role in marine food webs, particularly in the Arctic where they can control phytoplankton biomass (1). Modelling studies indicate they likely consume the majority of bacterial biomass in this region (2). In other marine environments they have been found to be more important than viral lysis (3) or ciliate grazers (4) for controlling bacterial concentrations. HF range from pico- to nano-sized cells (0.8–20 µm), and are phylogenetically diverse, with representatives in most branches of the Eukaryota tree. Because taxon-specific differences in feeding behaviour, population dynamics and life histories may be important over broad ecosystem scales (5) the identification and quantification of the diverse groups can add substantially to our understanding of marine ecosystem function. However, HF taxa often have few diagnostic morphological features that enable easy identification (6) and microscopy based studies rarely differentiate between HF taxonomic groups. Molecular based approaches, such as cloning and amplicon sequencing have provided additional information on the deep phylogenetic diversity of HF, especially in the Arctic (7) More recently, high throughput multiplex tag sequencing of hypervariable regions of the 18S rRNA gene is becoming commonplace (8,10) This is considered a semi-quantitative method for retrieving thousands of taxonomically informative sequences from environmental samples. However, the level of taxonomic placement, for example from phylum level to ecotype, is completely dependent on the depth and accuracy of the reference databases. For this reason, alignment and placement within phylogenetic trees remains the best method for accurately portraying diversity within groups of closely related organisms (11) While amplicon sequencing shares some of the inherent biases of
other PCR-based methods (12–13) it is useful for highlighting differences and similarities among samples, provided similar protocols are followed. In addition, diversity can be compared between environments and the high output enables the detection of rarer species. Sufficient coverage will depend on the taxonomic richness of an environment and even using high throughput sequencing it may be difficult to achieve in extreme cases (14).

Current changes in the Arctic Ocean, such as the loss of multiyear ice, may be impacting marine photosynthetic protist communities. For example, Li et al. (15) documented a shift in size structure toward smaller cells. However, the consequences for HF taxa are not known, in part because few studies have characterized the diversity and biogeography of different HF (16-17). Although there have been a number of studies using V4 18S rRNA gene amplicon pyrosequencing in the Arctic (10, 18–20), none were focused on pan-Arctic comparisons. All of these studies have detected an effect of environmental gradients on one or more HF taxa, but the extent to which HF with similar putative ecological roles (e.g. bacterivory) might vary in relative abundance over space and time is not yet understood, and a comprehensive analysis is lacking.

Our aim was to examine smaller HF taxa (nano- and picoplankton, 0.2–20 µm) whose sporadic and low occurrence in individual studies has eluded detailed analysis, and to increase the environmental coverage by additional high throughput sequencing from samples collected mostly from offshore regions of the Eastern and Western Arctic. We chose to focus on HF taxa of smaller cell-size that are usually considered bacterivores
and detritivores. These included the cercozoan order Cryomonadida, the phylogenetically
diverse marine stramenopiles (MAST), Telonemia, Picozoa (formerly known as
picobiliphytes (21) and Choanoflagellida. All of these HF taxa have been detected in 18S
rRNA gene clone libraries from the Canadian Arctic (7, 22–23). Katablepharids were
excluded from the analysis, although they have been detected in enrichment studies in the
Arctic (24) because they were rare in our samples. Although chrysophytes are detected in
the Arctic (25) this group includes both photosynthetic and heterotrophic taxa that
cannot readily be distinguished at this time based solely on the V4 region 18S rRNA gene
sequence that we used to identify taxa, and was therefore also omitted. Our large dataset
of 117 samples, collected from different regions, depths and ice conditions, enabled us to
examine the distribution of the five major groups at multiple taxonomic scales. Our first
step was to identify the taxonomic diversity within the five groups. In general, MAST
phylogeny is well-studied, and named clades form robust clusters based on sequence
similarity (17, 26, 28) and reads can be identified by comparison with a reference
database. However, clades from the other four HF groups are less well defined.
Cryomonadida, Telonemia and Picozoa have very few cultured representatives, and
choanoflagellate phylogeny has recently been extensively revised (29), requiring an
update of tools and reference databases used for taxonomic identification from sequence
data. Because the uncertainty of the phylogenetic placement of short reads makes simple
identification by sequence similarity ineffective, our approach for the four taxa with
problematic phylogenies was to map reads onto phylogenetic trees using an
environmental placement algorithm (EPA). Reads placed at the same node of a
phylogenetic tree were considered to belong to the same phylotype. We then determined
where these phylotypes occurred across Arctic Seas.

Methods

Experimental Design

Existing data was drawn from four different studies that used pyrosequencing of the V4
variable region of the 18S rRNA gene to describe eukaryotic communities. All four were
carried out over the Beaufort Sea, and included the Canada Basin, Amundsen Gulf and
across the Mackenzie Trough (Table 1; Figure 1; Table S1). The studies sampled surface
waters or the subsurface chlorophyll maximum (SCM) as defined in Martin et al. (30)
and sometimes both depths. The Mackenzie Trough study, which was in conjunction
with the French International Polar Year project ‘Malina’, included additional depths
above and below the SCM. One of the projects focused on sea ice sampled in the
Amundsen Gulf region in spring 2008 (18) Sample collection and pyrosequencing are
described in associated publications (10, 18, 20, 25, 31), and were similar for all datasets.
To increase pan-Arctic coverage, additional samples that had been collected during
various missions between 2005 and 2011, were also sequenced and analysed using
similar methods; these samples were from the Chukchi Sea, the Canadian Arctic
Archipelago (Archipelago), Baffin Bay, Hudson Bay and the Laptev Sea as detailed
below. For most samples, 2000-12 000 reads were analyzed per sample (Supplementary
Table S1). Accession numbers for the NCBI Sequence Read Archive (SRA) for all
datasets are given in Table 1. Metadata for stations from all studies, including sampling
depths, is provided in Supplementary Table S1.
Additional samples

Chukchi Sea samples were collected onboard the USCGS Healy mostly from the Chukchi Shelf (Figure 1) from 13 June–22 July 2010 as part of the ICESCAPE project (URL: http://www.espo.nasa.gov/icescape/). Seawater was collected into rinsed carboys from 30-liter Niskin-type bottles mounted onto a SeaBird carousel rosette equipped with a SBE9plus CTD. Genomic DNA was collected by sequentially filtering 5 liters of seawater through a 3-µm pore size, 47-mm diameter polycarbonate (PC) filter (Millipore) and a 0.22-µm pore size Sterivex cartridge (Fisher Scientific) to obtain two size fractions, 0.22–3 µm (small) and > 3 µm (large), to enrich for and concentrate organisms in the targeted size classes (32). Samples were preserved in a lysis buffer (50 mM Tris HCl pH 8.3, 40 mM EDTA pH 8.0, and 0.75 M sucrose), and frozen at -80°C.

The Archipelago, Baffin Bay and Hudson Bay samples were collected in conjunction with ArcticNet missions (http://www.arcticnet.ulaval.ca/) from 2005-2011. Samples were collected aboard the CCGS Amundsen as described above but with the following modifications. Seawater was collected using 12-liter Niskin-type bottles mounted on a rosette system (10, 20), and 6 liters were sequentially filtered as above. In 2010 and 2011, RNAlater (Qiagen, Germantown, MD) was used as a preservative instead of the lysis buffer. Laptev Sea samples were collected onboard the Russian RV Viktor Buynitsky from 17–30 September 2007 as part of the 2007 Nansen and Amundsen Basins Observational System (NABOS) mission (http://nabos.iarc.uaf.edu/). Water was collected directly from 5-liter Niskin-type bottles and was filtered sequentially as above, preserved...
in the lysis buffer and frozen at -80 °C. These ArcticNet and NABOS samples were sequenced as part of the Arctic Ocean Survey (AOS) project (33).

DNA was extracted from filters and Sterivex cartridges using an AllPrep DNA/RNA MiniKit (Qiagen). Primers and conditions for amplification and generating amplicons and reads followed the protocols described in Comeau et al. (10, 18). For the Chukchi Sea, the two size fractions were sequenced separately except for the < 3-µm CHA1B (Kotzebue Sound) sample, which was lost (Table 2). Because of limited funds, the large and small size fractions for the AOS project were pooled after the first amplification step of the amplicon tag sequencing protocol (10) following the ratio of small and large cells as determined by microscopy counts.

Reads from the Chukchi Sea - ICEASCAPE project were deposited in the NCBI SRA with accession number SRP029300, while reads from the multiple sites of the Arctic Ocean Survey project (33) are together under accession number SRP040734 (Table 1).

Sequence processing

The software Qiime (34) was used for all sequence processing and downstream analyses. Reads were denoised using the default FLX settings in Qiime. Reads < 200 bp or > 1000 bp were discarded. Chimeras were removed using the UCHIME algorithm (35) by checking both de novo and against the SILVA database. Only non-chimeric sequences detected by both methods were retained. Reads were clustered into Operational Taxonomic Units (OTUs) at ≥ 98% similarity as in (10) Representatives from each OTU
were aligned using MUSCLE (36) and assigned taxonomy using Mothur (37) with a minimum confidence score of 0.8 against our own curated 18S rRNA gene database, which includes Arctic Ocean 18S rRNA gene sequences from clone libraries (10),

Comparative phylogenetic community diversity was analyzed using unweighted UniFrac distance metrics (38), which clusters samples based on phylogenetic lineages. UniFrac distances were calculated using reads belonging to all five major HF groups, and to MAST reads alone. Dendrograms were constructed using Randomized Accelerated Maximum Likelihood (RAxML v.7.2.7) software (39-40) with a general time reversible (GTR) model of nucleotide substitution using four discrete rate categories to approximate a gamma distribution. Distance matrices were used to build a UPGMA tree. The robustness of UPGMA clustering was determined using jackknife analyses with the default parameters of 10 iterations of 140 reads per sample, which represented 70% of the total number of sequences per sample. A Similarity Percentage (SIMPER) test using OTU abundance, as implemented in PAST v3.01 (41) was used to test which taxa contributed most to the clustering by UniFrac.

Placement of Reads on Phylogenetic Trees
For phylogenetic analysis, 18S rRNA gene reference sequences were selected based on previously published phylogenies: Cryomonadida (16, 42) Telonemia (43), Picozoa (21) and Choanoflagellida (29, 44). Because there are relatively few Picozoa-related full-length 18S rRNA gene sequences in GenBank, we constructed two trees: one with only full-length sequences, and one that omitted the V1 and V3 variable regions but included
more environmental sequences. Many of the choanoflagellate sequences contained large
insertions that were omitted from the analysis, as they are not taxonomically informative.

Alignments of short read OTUs were checked against the reference sequences to look for
potential artefacts, particularly errors in homopolymer length, which is poorly resolved
using pyrosequencing (45) Short read and reference alignments were then assembled into
a single alignment by manually adding gaps in the sequence editing program SeaView
(46). Aligned reference sequences were then extracted and used to reconstruct a reference
tree using RAxML as described above. Short reads were mapped onto the reference tree
using the EPA of RAxML, which sequentially places each short query sequence (read) at
each edge of a reference tree previously constructed with longer sequences, and
calculates the likelihood of the resulting tree (47). The resulting tree was visualized using
NJ Plot (48) Only read placements with likelihood weight > 0.5 were retained. A
combination of EPA and ordination with nonmetric multidimensional scaling (NMDS)
rung in R used the nodes identified in EPA as taxa, with an aim to identify phylotypes that
could be indicators of regional differences or sampling techniques. Phylotypes were
defined as reads which were placed at the same node.

Results

In total, approximately 1 020 000 reads passed the quality control steps. HF groups
comprised 1.6–54% of eukaryote sequences in the water column samples and up to 60%
in the ice samples (Supplementary Figure S1). All of the HF phyla level groups were
found in all of the samples.
Phylogenies and evolutionary placement of reads

The reference trees of Cryomonadida, Telonemia, Picozoa and the Choanoflagellida indicated the high phylogenetic diversity of the four groups, with well-supported branching orders, and usually recovered previously published clusters. Specifically, the *Cryothecomonas* clade (16) was recovered with high bootstrap support in the Cryomonadida phylogeny (Figure 2). Within Telonemia, we recovered the groups TEL1 and TEL2 described by Bråte et al. (43) with very high bootstrap support, along with their freshwater clades 1d and 2e (Figure 3). Two Picozoa phylogenies were constructed, one from an alignment of 16 sequences and 1633 characters and a second one from an alignment of 55 sequences and 1168 characters. There was good agreement between the topologies, so the tree with a shorter residue range and more sequences was used for EPA. Some of the same clades described in Seenivasan et al. (21) were recovered, albeit with different branching order (Figure 4). In the choanoflagellate phylogeny we recovered the Acanthoecida and Crespedida clades found by Nitsche et al. (29) although the latter had low bootstrap support (Figure 5). Within Acanthoecida, we recovered Stephanoecidae as paraphyletic and containing the Acanthoecidae clade. As a group, the choanoflagellates had only moderate bootstrap support, and overall the RAxML tree should be interpreted cautiously, as the 18S rRNA gene is poor at resolving deeper branches among the choanoflagellates (29).

Reads were mapped onto the reference RAxML trees to identify key taxa in different datasets. In total, 3165 Cryomonadida reads, out of an original 23 870, could be placed...
One phylotype (node \(a\)), found in sea-ice and the Laptev Sea, was associated with the putative diatom parasite *Cryothecomonas aestivalis*. The node \(c\) phylotype within the *Cryothecomonas* clade was broadly distributed in all regions, though not found in sea-ice, while the node \(d\) phylotype was found mostly in the Chukchi Sea. The most abundant phylotype (\(e\); 1394 reads) branched at the base of the *Cryothecomonas* clade and was comprised almost exclusively of sea-ice reads. One phylotype comprising 596 reads, mostly from sea-ice, branched outside Cryomonadida (node \(k\)), likely with the genus *Ebria*.

For Telonemia, 9288 out of 14 874 reads could be placed on the reference tree (Figure 3). The largest number of reads (4174) came from the Mackenzie Trough (Malina dataset); the lowest number (79) came from the Laptev Sea. Phylotypes nodes \(b\), \(g\), \(h\) and \(i\) were broadly distributed in most or all of the regions studied. Three Telonemia phylotypes (nodes \(a\), \(c\) and \(e\)) were found mostly in the Canada Basin, with \(a\) restricted to surface water samples. Nodes \(c\) and \(e\) were found in higher abundances in the RNA than the DNA sample from station PP-6.5, although this difference was not observed for other sites where both RNA and DNA were sampled (data not shown). Over 90% of the reads from the Mackenzie Trough were placed with environmental sequence Nor26.Telo.6, in clade 2d (node \(f\)). In contrast, reads from the Canada Basin were phylogenetically more broadly distributed through the marine clades in TEL2.

For Picozoa, 12 902 out of 37 904 reads could be placed on the reference tree (Figure 4). The largest number of reads came from the Amundsen Gulf Time Series (3925) and the
Mackenzie Trough (2685). Very few came from studies which sampled the > 3 µm size fraction exclusively, consistent with the < 2 µm size range reported for Picozoa (21). Reads were mapped to just four nodes, all of which were broadly distributed.

For the choanoflagellates, 8690 out of 17 216 reads could be placed on the reference tree (Figure 5). The largest number of reads (2402) came from the Hudson Bay; the lowest number (122) came from the > 3 µm size fraction from the Chukchi Sea. The majority of reads (7742 total) were placed with the tectiform acanthoecid *Didymoeca costata* (also referred to as *Diplotheca costata* in Doweld (49); nodes *f*, *g*, and *h*). A few nudiform acanthoecid reads (64; node *a*) and Crespedida reads (420; nodes *j*, and *k*) were detected. 160 reads were placed at the base of the choanoflagellate tree (node *l*), and could not be definitely identified.

The combination of EPA and ordination with NMDS enabled identification of phylotypes that appeared to be indicators of regional differences or sampling techniques. While broadly-distributed taxa were found in all groups, more region-specific taxa were found for organisms that are usually > 3 µm such as Cryomonadida and Telonemia, than for smaller organisms such as MAST and Picozoa. For example, while Cryomonadida reads were found in all three > 3 µm datasets, the phylotypes found in each one were different, with very little overlap (Canada Basin: node *f*; Sea-ice: nodes *a*, *b*, *e*, *h*, *i*, *k*; Chukchi Sea: nodes *c*, *d*, *g*, *j*; Figures 2 and 6).

**HF community composition and beta diversity analysis**
Unweighted UniFrac distance measures revealed some clustering by region, albeit with low jackknife support (Supplementary Figure S1). Both sea-ice samples and the > 3-µm size fraction of the Chukchi Sea water column formed separate clusters with a higher proportion of Cryomonadida sequences. The August and October 2009 Amundsen Gulf time series and SCM samples from the Mackenzie Trough (Malina dataset) mostly clustered together. Within the Canada Basin surface water cluster, samples based on either RNA or DNA from the same site tended to cluster together, with the exception of station PP-6.5. Telonemia and choanoflagellates were more dominant in the Canada Basin. A SIMPER test identified the same four OTUs as the largest contributors to dissimilarity between clusters (defined at 85% similarity) for weighted and unweighted UniFrac, accounting for 17% of total dissimilarity. These were an acanthoecid choanoflagellate, a MAST-4, a MAST-7 and a Telonemia. All were either not recovered, or very rare in samples from the > 3 µm fraction in the Chukchi Sea, Canada Basin and sea ice. MASTs were always a greater proportion of reads in datasets from the < 3 µm fraction, and were dominated by clades MAST-1 and 7 (Supplementary Figure S2). As with analysis of all HF, clusters had very low jackknife support.

Discussion

Here we examined high throughput sequence data from nine regions across the Arctic, focusing on heterotrophic flagellates, a functional guild that consumes bacteria and picoeukaryotes and a key component of microbial food webs (50-51). Although all of the major HF groups were found in all of the samples examined in this study, the taxonomic resolution provided by the EPA analysis revealed differences in distribution between
clades for some of the HF groups. All Picozoa and choanoflagellate clades, as well as most MAST clades, were widely distributed. In contrast some Cryomonadida and Telonemia phylotypes had more restricted distributions.

Data Interpretation

Inferring taxon abundance from proportion of reads and interpreting this information in an ecological context is subject to a number of caveats, including limitations inherent to amplicon sequencing, biases introduced by the collection and analysis of samples from different studies, and incomplete taxon sampling in reference trees and databases.

The greater sampling depth achievable by high throughput sequencing enabled the detection of taxa that had not been recorded using FISH or 18S rRNA gene clone libraries. For example, a phylotype related to *Cryothecomonas aestivalis*, a species that parasitizes diatoms (52) was detected by high throughput sequencing of sea-ice and in the Laptev Sea (Figure 2). While cryothecomonad-like cells have been observed associated with diatoms during the spring bloom in northern Baffin Bay (C. Lovejoy, personal communication), Thaler and Lovejoy (16) did not find any Arctic *C. aestivalis*-related sequences in 18S rRNA clone libraries or in public databases. *Cryothecomonas* was originally described as a free-living predator in sea-ice (53) however, to our knowledge no similar FISH studies using a *Cryothecomonas*-specific probe have been carried out in ice environments.
While high throughput sequencing provided a more complete picture of taxonomic diversity compared to earlier studies, some anomalies were evident. For example, the MAST-1A clade had the highest proportion of reads in nearly all samples, in contrast with the cell count data of Thaler and Lovejoy (17) and Lin et al. (27) who found MAST-1A at much lower relative concentrations compared to MAST-1B or -1C in the Arctic and the North Pacific respectively. MAST-1A also tends to be retrieved more frequently in Arctic Ocean clone libraries (7, 23). This could be due to PCR-bias towards certain taxa (12), possibly because of preferential primer binding or RNA structure (13).

Alternatively, FISH studies suggest that MAST-1A cells tend to be larger than 1B (17), and could also have multiple 18S rRNA gene copies. In this context, as with other eukaryotic microbes, variable copy-number of 18S rRNA genes in different taxa make inference of taxon abundance relative, and should not be treated as equivalent to absolute estimates (54).

The opportunistic nature of the collection and analysis of the different samples from studies with slightly different methods could introduce biases that mask true biogeographical signal. Of particular concern is interannual variability resulting from the changes that have taken place in the Arctic over the last 10 years (10, 15). The effect of this type of variability could not be analysed directly since, with the exception of the Amundsen Gulf Time Series, no region was sampled more than once.

Finally, EPA could place only 15–60% of reads. This loss of information resulted from our rather stringent cut-off of likelihood weight > 0.5 in EPA. Likelihood weights are
calculated for the entire tree after adding the inserted query sequence; it is therefore logical that lower likelihood weights were found for taxa such as Cryomonadida and Picozoa, for which large numbers of reads mapped to poorly supported regions of the reference tree (for example, Cryomonadida phylotype e, or Picozoa phylotype c), whereas higher likelihood weights were found for Telonemia, for which most reads were mapped to well-supported nodes. This highlights the need for improved reference trees for better identification.

Because of the caveats discussed above, we make an effort to discuss only those trends that are supported by very marked differences in read abundance, or by a combination of more than one line of evidence (EPA, clustering, NMDS), and are therefore presumably robust.

Regional and water mass signals

Over most of the Arctic, different water masses result in strongly stratified water columns (55), that display very sharp vertical gradients in resources (30). Protist community composition in the Arctic is closely associated with water masses rather than geographic location (20, 56–57). Samples from the Amundsen Gulf Time Series were all from the interface of the Polar Mixed Layer and water originating from the Pacific; the SCM forms at this point because of tradeoffs between sufficient light and nutrients (30). Vertical structuring of the HF community by water mass along the Mackenzie Shelf has already been discussed by Monier et al. (20), who found that samples from the same water mass were more similar to each other than samples from different water masses at...
the same station. Our new analysis, comparing a much larger dataset, found much less
evident separation detected by UniFrac analysis of the HF communities (Supplementary
Figure S2). Factors that could mask the water mass signal by contributing to similarities
and differences between samples include interannual variability as well as the
hydrography of the Arctic Ocean itself, where water masses are modified over distances
(58). There was little difference between surface and SCM samples from Hudson Bay
and the Archipelago, which may reflect a shallower, well-mixed water column in the
Archipelago, and the complex hydrography of the Hudson Bay Region. The samples
from Baffin Bay were collected with an aim to document the maximum diversity and
included both Arctic and Atlantic influenced water masses (33). The complex
interleaving of water masses in Baffin Bay (59) could also mask effects of water mass or
depth, because sampling did not occur at a sufficiently fine vertical scale to capture the
structure of water column. Interestingly, many of the < 3-µm size fraction Chukchi Sea
samples tended to cluster with Amundsen Gulf SCM samples, consistent with the Pacific
Origin water being nearer the surface in the Chukchi Region (58).

Indicator taxa

A combination of EPA and ordination with NMDS enabled identification of phylotypes
that appear to be indicators of regional differences or sampling techniques. More region-
specific taxa were found for larger organisms such as Cryomonadida and Telonemia, than
for smaller organisms such as MAST, Picozoa and choanoflagellates, although broadly-
distributed taxa were found in all groups.
Cryomonadida reads dominated sea-ice samples from the Beaufort Sea and were also abundant in water column samples from the Chukchi Sea (Figure S1). However, there was little overlap of phylotypes between the two datasets (Figure 2, 6), and the most common phylotypes (nodes e and k; Figure 2) found in the Amundsen Gulf Sea Ice were rare in the Chukchi Sea. This suggests that some Cryomonadida may be more adapted to open water rather than ice, such as species associated with nodes e and f (Figure 2).

The Chukchi Sea samples were collected from sites within or near the marginal ice zone, and associated ice-melt may have released sea-ice organisms, including Cryomonadida, into the water column. However, in July 2010 the region was characterized by widening and merging polynyas (U.S. National Ice Centre; URL: http://www.natice.noaa.gov), as is typical for this time of year (60), that would be more consistent with the surface waters having been ice free for much longer. Interestingly, no sea-ice reads were placed with reference sequence MC5-1, a cultured bacterivorous strain isolated from sea-ice in Antarctica (61). Instead, nearly all reads from this dataset were found in the node e phylotype (node k, also with a high abundance of sea-ice reads, is probably not a Cryomonadida but an Ebriida). This phylotype must have a very broad halotolerance, since it includes an environmental sequence (KRL01E39) from a brackish lake, while the salinity of brine channels within sea ice can be many times that of seawater (62-63).

Telonemia clades were most abundant in the offshore Mackenzie Trough, associated with the node f phylotype, and the Canada Basin, with the greatest relative proportions in surface waters (Supplementary Figure S1), associated with phylotypes a, c and e. In contrast, Telonemia in the Canada Basin were both more abundant and phylogenetically...
more diverse (Figure 3). In the Canada Basin, the anticyclonic Beaufort Gyre causes the vertical downward stacking of water masses, with different densities at the center creating a bowl-like structure (58). The down-welling effect depletes the surface waters of nutrients and deepens the nitracline and the SCM to approximately 60 m or deeper (64). Low nutrient concentrations found in surface waters supports low concentrations of phytoplankton, as exhibited by low concentrations of chlorophyll \(a\) (chl \(a\)) offshore of the Mackenzie shelf at the time of sampling (65). While chl \(a\) concentrations were not measured in all samples, average surface concentration in the Canada Basin was half that of the southern Beaufort Sea, and more than two orders of magnitude lower than in the Chukchi Sea (Supplementary Table S1). Such distributions suggest that Telonemia have an advantage in low productivity systems, and are able to maintain high populations of one or a few phylotypes, for example node \(f\) in the Mackenzie Trough (Malina dataset), and nodes \(c\) and \(e\) in the Canada Basin (Figure 3); however, this trend cannot be statistically evaluated with our present data. Higher relative abundances of the Telonemia phylotypes \(c\) and \(e\), along with lower abundance of MAST-1A (UniFrac Analysis, Supplementary Figure S2), were found in the RNA compared to the DNA template samples from site PP-6.5, which was adjacent to the marginal ice zone of Canada Basin (Lovejoy and Scarcella personal communication). No such difference was observed for sites CB-15 and CB-17.5, located under > 90 % ice cover, implying that ribosomal activity detected from 18S rRNA for these taxa may be associated with community changes in a mobile ice environment; unfortunately, no RNA sample was available from the other marginal ice site, PP-2. The node \(l\) phylotype, containing over half the reads
from sea-ice, was associated with an environmental sequence from a sediment sample and may represent a taxon adapted to interstitial environments.

In the Chukchi Sea samples, where size fractions were analysed separately, the smaller size fraction was enriched in Picozoa and MAST (about 20-fold enrichment in the Chukchi Sea data), consistent with cell diameters given for HF taxa in the literature (17, 21, 26, 66). For all samples, Picozoa and MAST were geographically less distinct in the Arctic than Cryomonadida and Telonemia, which are generally larger and in the nanoplankton size range (2–20 µm) (16, 67). On the contrary, our results highlighted the ubiquity and abundance of the clade MAST-7 (Supplementary Figure S2), which comprised 25% of all MAST reads in our data, and was enriched in the < 3-µm datasets. While this clade has been identified in 18S rRNA gene clone libraries (68), to our knowledge there have been no studies on MAST-7 using fluorescent in situ hybridization (FISH), which could give a more reliable picture of its abundance in the Arctic.

Similarly, EPA analysis showed that Picozoa in the Arctic contained only four broadly distributed phylotypes (Figure 4), indicating that this group may be dominated by a few taxa. The single cultured representative of Picozoa is thought to consume colloidal material and perhaps small viruses (21). If so, this group might be less constrained by food availability compared to other HF that may have a preference for specific bacteria or phytoplankton prey, thus enabling the Arctic picozoans to grow across different regions and depths. However, it may also indicate that better taxon coverage is needed in the reference tree or that V4 rRNA gene sequences are highly conserved in this group.
While many eukaryotic environmental gene surveys, including in the Arctic, have focused on the < 3-µm size fraction (10, 69), the above results suggest that the large size fraction of HF can be indicators of regional differences. In NMDS, samples from both size fractions in the Chukchi Sea overlapped (Figure 6), but the > 3-µm size fraction occupied a larger ordination space, suggesting that small organisms were being retrieved in the larger size fraction, but not vice versa. Size fractionation by filtration is imperfect, as flexible cells may be able to pass through pores smaller than their longest dimension, while pores that are clogged will retain cells smaller than the pore diameter. In addition, free DNA from larger cells can be retained on the 0.2-µm filter (70). In practice, size fractionation enriches for a particular size class (32), and needs to be taken into account.

Most choanoflagellate phylotypes were found in all of the studied regions. There is a rich literature on the distribution of choanoflagellate taxa using microscopic methods; however, comparison with molecular-based results is complicated by historic misidentifications and a lack of cultured representatives. Most studies in polar water columns and sea-ice have focused on Acanthoecida (71–74), reflecting both the comparative wealth of diagnostic features in this loricate group, and the dominance of acanthoecids in marine planktonic environments, which is corroborated by our EPA results. Nearly all choanoflagellate reads that could be mapped onto the tree belonged to the order Acanthoecida, which are considered to be mainly planktonic. Ambiguities in the reference tree made it impossible to place reads at a species level; however, in most regions the majority were placed in the genus Didymoeca. Del Campo and Massana (44) identified a Didymoeca clade containing sequences from the Arctic water column, the
Conclusions

We reanalysed new and publicly available V4 18S rRNA gene short read data to detect biogeographic signal for heterotrophic flagellates in the Arctic Ocean. Placement on phylogenetic trees allowed reads to be identified at a high level of taxonomic resolution, although improved reference trees will be needed to fully characterize communities.

While many phylotypes were widespread, others were specific to regions and environments. Phylotypes with restricted distributions were more likely to be found among taxa of larger cell-size, such as Cryomonadida and Telonemia. These taxa may make good targets for future research, functioning as indicators of the dramatic changes in ice regimes that are already taking place across the Arctic.

Acknowledgments

We thank Marcel Babin (the Canadian Excellence Research Chair in remote sensing of Canada's New Arctic Frontier) for access to ICESCAPE samples. The ICESCAPE DNA samples in the Chukchi Sea were collected and extracted by Eva Ortega-Retuerta, Laptev Sea samples were collected by Marie-Éve Garneau. Many thanks to André Comeau and Adam Monier for help with pyrosequencing analysis. We also thank Jill Watkins (Department of Fisheries and Oceans, Canada) for encouragement and support for the Arctic Ocean Survey, which is a contribution to the Circumpolar Biodiversity Monitoring Plan. This study was made possible by fellowships to MT and Discovery grants to CL.
from the Natural Science and Engineering Council of Canada (NSERC). Computing
support from CLUMEQ/Compute Canada was appreciated. The Fond Québécois de
Recherches Nature et Technologies (FQRNT) to Québec Océan and the NCE ArcticNet
are acknowledged for their support.


Figure Legends

Figure 1. Stations sampled by the multiple studies in this paper. a) Beaufort Sea; b) Baffin and Hudson Bays; c) Chukchi Sea; d) Laptev and Siberian Sea. The map was created in Ocean Data View version 4 (Schlitzer, R., Ocean Data View, http://odv.awi.de, 2014).

Figure 2. Phylogenetic mapping of Cryomonadida reads from all studies. Rooted Cryomonadida reference phylogenetic tree was constructed using maximum likelihood from an alignment of 92 sequences and 1689 characters. Some non-Cryomonadida reference sequences have been omitted for clarity. Closed circles indicate nodes with bootstrap support > 50 (out of 100). Reads are mapped onto the nodes marked with blue circles using RAxML evolutionary placement algorithm; only placements with likelihood weight > 0.5 are shown. Pie charts show the proportion of reads from each study at a given node. Left scale bar indicates number of substitutions per position. Protaspis longipes (*) was formerly identified as Cryothecomonas longipes (75) and Rhogostoma minus (**) was formerly identified as Lecythium sp. (76). Outgroups (not shown) are two radiolarian sequences and an acantharean (Spongaster tetrə AB101542, Lithomelissa setosa HQ651801, Acanthometra fusca KC172856).

Figure 3. Phylogenetic mapping of Telonemia reads from all studies. Rooted Telonemia reference phylogenetic tree was constructed using maximum likelihood from an alignment of 52 sequences and 1942 characters. Closed circles indicate nodes with
bootstrap support > 50 (out of 100). Reads are mapped onto the nodes marked with blue
circles using RAxML evolutionary placement algorithm; only placements with likelihood
weight > 0.5 are shown. Pie charts show the proportion of reads from each study at a
given node. Left scale bar indicates number of substitutions per position. Labelled clades
correspond to Bråte et al. (43). Outgroups (not shown) are a haptophyte and a
katablepharid (*Prymnesium parvum* AJ246269 and *Katablepharis japonica* AB231617)

**Figure 4.** Phylogenetic mapping of Picozoa reads from all studies. Rooted Picozoa
reference phylogenetic tree was constructed using maximum likelihood from an
alignment of 55 sequences and 1168 characters. Closed circles indicate nodes with
bootstrap support > 50 (out of 100). Reads are mapped onto the nodes marked with blue
circles using RAxML evolutionary placement algorithm; only placements with likelihood
weight > 0.5 are shown. Pie charts show the proportion of reads from each study at a
given node. Left scale bar indicates number of substitutions per position. Clades are
labelled following Seenivasan et al. (21). Outgroups (not shown) are a haptophyte, a
katablepharid, and a Telonemia (*Prymnesium parvum* AJ246269, *Katablepharis japonica*
AB231617 and *Telonema antarcticum* (AJ564773)

**Figure 5.** Phylogenetic mapping of choanoflagellate reads from all studies. Rooted
choanoflagellate reference phylogenetic tree was constructed using maximum likelihood
from an alignment of 49 sequences and 1963 characters. Closed circles indicate nodes
with bootstrap support > 50 (out of 100). Reads are mapped onto the nodes marked with
blue circles using RAxML evolutionary placement algorithm; only placements with
likelihood weight > 0.5 are shown. Pie charts show the proportion of reads from each
study at a given node. Left scale bar indicates number of substitutions per position.
Outgroups (not shown) are two metazoan sequences, *Mnemiopsis leidyi* (AF293700) and
*Beroe ovata* (AF293694), a sponge (AY348876), two ichthyosporeans (Y16260 and
AF232303), and *Corallochytrium* (L42528)

Figure 6. Non-metric multidimensional scaling (NMDS) of heterotrophic flagellate reads
in the Arctic, excluding sea-ice, using taxa identified by EPA. Letters correspond to
nodes from Figures 2–5, while numbers refer to MAST clades.
Table 1 **Summary:** Studies analyzed for heterotrophic nanoflagellates: geographic area (Region) where samples were collected (see text for details); the names of the datasets (Dataset), with previously published references in parenthesis; collection dates of the datasets, years and months when samples were collected (Dates); number of sites for each data set (Sites); size fraction(s) selected for sequencing (Size) and the Short Read Archive (SRA) refers GenBank deposited raw read data.

<table>
<thead>
<tr>
<th>Region</th>
<th>Dataset</th>
<th>Dates</th>
<th>Sites</th>
<th>Size</th>
<th>SRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaufort Sea</td>
<td>Amundsen</td>
<td>2003-2010</td>
<td>11</td>
<td>&lt; 3 µm</td>
<td>SRA029114</td>
</tr>
<tr>
<td></td>
<td>Gulf (10)</td>
<td>July-Nov</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beaufort Sea</td>
<td>Sea Ice (17)</td>
<td>2008</td>
<td>5</td>
<td>pooled</td>
<td>SRA054160</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mar-May</td>
<td></td>
<td>&lt; 3 µm &amp; &gt; 3 µm</td>
<td></td>
</tr>
<tr>
<td>Mackenzie Trough</td>
<td>Malina (19)</td>
<td>2009 July</td>
<td>6</td>
<td>&lt; 3 µm</td>
<td>SRA063446</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beaufort Sea</td>
<td>Canada</td>
<td>2007</td>
<td>4</td>
<td>&gt; 3 µm</td>
<td>SRA099217</td>
</tr>
<tr>
<td></td>
<td>Basin</td>
<td>Aug</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chukchi Sea</td>
<td>ICESCAPE</td>
<td>2010</td>
<td>12</td>
<td>&lt; 3 µm &amp; &gt; 3 µm</td>
<td>SRP029300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>June-July</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archipelago</td>
<td>Arctic Ocean</td>
<td>2005–2011</td>
<td>20</td>
<td>pooled</td>
<td>SRP040734</td>
</tr>
<tr>
<td>Hudson Bay Survey</td>
<td></td>
<td>2011</td>
<td></td>
<td>&lt; 3 µm &amp; &gt; 3 µm</td>
<td></td>
</tr>
<tr>
<td>Baffin Bay Survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laptev Sea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Pyrosequencing raw data, filtering and OTU statistics for data from ICESCAPE 2010 study. Percentages are given relative to reads available from the preceding step.

<table>
<thead>
<tr>
<th>Processing Step</th>
<th>&gt; 3 µm size fraction</th>
<th>0.22–3 µm size fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(12 samples)</td>
<td>(11 samples)</td>
</tr>
</tbody>
</table>

**Pre-filtering:**
- Total Reads: 138,658 (137,727)
- Mean Length: 343 bp (308 bp)

**Post-filtering:**
- Reads retained after quality control: 64,139 (46%) (49,451 (36%))
- Mean Length: 435 bp (431 bp)
- Metazoans, fungi and land plants removed: 54,319 (85%) (47,594 (96%))
- Chimeras and poorly aligned reads removed: 53,600 (99%) (47,200 (99%))
- Singletons removed: 50,645 (94%) (44,643 (95%))

**OTU analysis (97% level)**
- OTUs all samples (non-redundant): 1,132 (986)
- OTUs all samples (cumulative): 3,111 (3,228)
- Mean OTUs per sample: 283 (293)

*a* Reads from station CN3 are excluded from the OTU analysis because after metazoan reads (mostly a single genus of copepod) were removed, there were insufficient remaining reads for the analysis.
on October 25, 2017 by guest

http://aem.asm.org/