Diversity and Distributions, (Diversity Distrib.) (2016) 22, 493–504

BIODIVERSITY RESEARCH

# Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports

Frédéric J. J. Chain[1]*, Emily A. Brown[1,2], Hugh J. MacIsaac[2] and Melania E. Cristescu[1]

[1]Department of Biology, McGill University, 1205 Docteur Penfield, Montreal, Quebec, H3A 1B1, Canada, [2]Great Lakes Institute for Environmental Research, University of Windsor, Windsor, Ontario, N9B 3P4, Canada

## ABSTRACT

**Aim** The urgent need for large-scale spatio-temporal assessments of biodiversity in the face of rapid environmental change prompts technological advancements in species identification and biomonitoring such as metabarcoding. The high-throughput DNA sequencing of bulk samples offers many advantages over traditional morphological identification for describing community composition. Our objective was to evaluate the applicability of metabarcoding to identify species in taxonomically complex samples, evaluate biodiversity trends across broad geographical and temporal scales and facilitate cross-study comparisons.

**Location** Marine and freshwater ports along Canadian coastlines (Pacific, Arctic and Atlantic) and the Great Lakes.

**Methods** We used metabarcoding of bulk zooplankton samples to identify species and profile biodiversity across habitats and seasons in busy commercial ports. A taxonomic assignment approach circumventing sequence clustering was implemented to provide increased resolution and accuracy compared to pre-clustering.

**Results** Taxonomic classification of over seven million sequences identified organisms spanning around 400 metazoan families and complements previous surveys based on morphological identification. Metabarcoding revealed over 30 orders that were previously not reported, while certain taxonomic groups were underrepresented because of depauperate reference databases. Despite the limitations of assigning metabarcoding data to the species level, zooplankton communities were distinct among coastlines and significantly divergent among marine, freshwater and estuarine habitats even at the family level. Furthermore, biodiversity varied substantially across two seasons reaching a beta diversity of 0.9 in a sub-Arctic port exposed to high vessel traffic.

**Main Conclusions** Metabarcoding offers a powerful and sensitive approach to conduct large-scale biodiversity surveys and allows comparability across studies when rooted in taxonomy. We highlight ways of overcoming current limitations of metabarcoding for identifying species and assessing biodiversity, which has important implications for detecting organisms at low abundance such as endangered species and early invaders. Our study conveys pertinent and timely considerations for future large-scale monitoring surveys in relationship to environmental change.

**Keywords**
**18S, biodiversity, bioinformatics, biomonitoring, DNA barcoding, OTU clustering.**

*Correspondence: Frédéric J. J. Chain, McGill University, Department of Biology, 1205 Docteur Penfield, Montreal, QC, H3A 1B1, Canada.
E-mail: frederic.chain@mcgill.ca



DOI: 10.1111/ddi.12427
http://wileyonlinelibrary.com/journal/ddi



A Journal of Conservation Biogeography
Diversity and Distributions

## INTRODUCTION

Contemporary environmental change is elevating extinction rates and altering biodiversity patterns across the globe (Urban, 2015). Increases in global trade, urbanization and travel are propelling the movement of organisms across previously isolated regions and advancing biotic homogenization and extinctions (Mooney & Cleland, 2001; Baiser *et al.*, 2012). Aquatic systems are particularly vulnerable, with the spread of invasive species threatening biodiversity, especially around marine and freshwater ports (Miller & Ruiz, 2014). Changing climate is causing extensive loss of Arctic sea ice, which is opening up new channels for commercial ships and creating more colonization opportunities for organisms (Vermeij & Roopnarine, 2008; Smith & Stephenson, 2013). Therefore, increasing the scale and frequency of monitoring surveys to assess biodiversity baselines and seasonal fluxes in vulnerable regions is in pressing demand. However, biodiversity assessments of aquatic organisms are challenged by the time and cost of surveys, and by the large number of species that are morphologically cryptic and remain undescribed (Appeltans *et al.*, 2012). Molecular techniques offer powerful and reliable alternatives for studying marine communities and aquatic invasions (Holland, 2000). Methods such as DNA barcoding (Hebert *et al.*, 2003) alleviate many of the limitations of traditional morphological taxonomy by increasing taxonomic resolution and precluding the need for expert taxonomists (Radulovici *et al.*, 2010). When coupled with high-throughput sequencing, barcoding of bulk samples – metabarcoding – enables rapid screening of taxonomically complex communities (Taberlet *et al.*, 2012; Yu *et al.*, 2012) with the sensitivity to detect rare organisms such as endangered species and early invaders. Metabarcoding is emerging as a tool for monitoring biodiversity in numerous kinds of environmental samples for the assessment of aquatic ecosystem health (Baird & Hajibabaei, 2012; Aylagas *et al.*, 2014) and global biodiversity trends (Cristescu, 2014).

Metabarcoding is more sensitive in identifying meiofauna and microbial aquatic organisms than morphological taxonomic identification (Cowart *et al.*, 2015; Zimmermann *et al.*, 2015). This is partly attributed to the ability to genetically discriminate cryptic species, larval stages (Lindeque *et al.*, 2013) and species with very low abundance such as early invaders that might be missed by traditional identifications (Darling & Mahon, 2011). Despite the great promises this technology offers for large-scale surveys and understanding global biodiversity patterns, most metabarcoding studies have examined regional communities from a single time point. Only very recently has metabarcoding been used to describe the relationship between the environment and aquatic communities, for example across a range of estuarine ecological conditions (Chariton *et al.*, 2015) and in coastal samples before and after an oil spill (Bik *et al.*, 2012). A major problem that such studies face is their inability to correctly identify taxa, especially at the species level, and thereby to differentiate between native and non-indigenous species and to compare across studies that use different identification methods. Identification is hampered by the complexity of aquatic communities relative to incomplete reference sequence databases. Another limitation stems from bioinformatics pipelines that implement sequence clustering prior to taxonomic assignments, reducing taxonomic resolution and leading to incorrect assignments particularly at the species level (Schloss & Westcott, 2011; Brown *et al.*, 2015). Careful taxonomic characterization and scaling up of biodiversity assessments across large geographical regions, multiple seasons, and broad taxonomic scope can provide essential baseline references for growing conservation management programs. In this study, we used metabarcoding data from bulk zooplankton samples to identify species and assess spatio-temporal patterns of biodiversity from Canadian coastlines (three oceans) and the Great Lakes. Linking taxonomy with sequence data enabled comparisons with previous morphological surveys and showed general congruency. Our approach circumvents sequence clustering, which has implications for the consistent and reproducible characterization of native and non-indigenous biodiversity in ports that may sustain future biological invasions.

## METHODS

### Biological sampling

Zooplankton samples were collected using geo-referenced plankton net hauls of 80 and 250 μm mesh size and were immediately preserved in 95% ethanol. Sampling was performed in 16 major Canadian ports (see Table S1 in Supporting Information) that covered four geographical regions: the Pacific, including the Strait of Juan de Fuca and Strait of Georgia (Nanaimo, Robert's Bank, Victoria, Vancouver); the Arctic, including Hudson Bay and the Hudson Strait (Churchill, Deception Bay, Iqaluit, Steensby Inlet); the Atlantic, including the Bay of Fundy and Gulf of Saint Lawrence (Bayside, Baie de Sept-Îles, Halifax, Hawkesbury); and the Laurentian Great Lakes basin, including Lake Ontario, Lake Erie, Lake Superior, and the Saint Lawrence River (Hamilton, Nanticoke, Thunder Bay, and Montreal, respectively). For the majority of ports, samples were collected from six sites (transects) on the same day using oblique tows or multiple vertical tows (for under-ice sampling) over two seasons between May 2011 and December 2012 (see Table S1). Due to weather restrictions, Steensby Inlet was only sampled during the summer season.

### DNA preparation and sequencing

The 80- and 250-μm net samples from each transect from the same season were pooled prior to DNA extraction. Most ports were represented by six independent pooled samples per season, with the exception of Steensby (only one season),

Sept-Îles (seasons combined), and Vancouver and Hamilton (all sites and seasons combined). We processed 147 pooled samples (hereafter referred to as 'samples'; see Table S1). Around 100 mg from each sample was centrifuged at 12,000 rpm for 3 min and placed in a fume hood for 10–15 min to remove ethanol. DNA extractions were carried out for each sample, and four replicate PCRs were run per DNA extraction. DNA extraction, PCR conditions and clean-up followed the protocol of Brown *et al.* (2015). Amplification of an approximately 400–600 bp stretch of the 18S rRNA gene (in the V4 region) was conducted using primers that had a higher amplification success rate than other primers tested in zooplankton groups (Zhan *et al.*, 2014). To ensure sample recognition in downstream analyses, each sample was amplified with a unique 10-bp MID barcode approved by Roche (Technical bulletin 005-2009, Roche Diagnostics Corp., Basel, Switzerland). The combined samples from each port were pyrosequenced at ½ PicoTiter plate scale using 454 FLX Adaptor A on a GS-FLX Titanium platform (454 Life Sciences, Branford CT, USA) by Engencore at the University of South Carolina or by Genome Quebec at McGill University.

## Data processing

In total, 10,277,272 reads were pyrosequenced (Table 1). Each sequenced read was assigned to a sample based on its barcode, and forward primers were removed using USEARCH v8.0.1616 (Edgar, 2010). Reverse primers were removed, reads were trimmed based on quality (minimum Phred score of 20), and only reads at least 200-bp long were kept using FASTX-toolkit (hannonlab.cshl.edu/fastx_toolkit). For each sample, reads were collapsed to unique sequences (dereplicated) using derep_fulllength in USEARCH to hasten downstream bioinformatical steps. UCHIME was used to remove chimeras (Edgar *et al.*, 2011). In total, 7,733,541 reads (75%) were left for taxonomic assignment.

## Taxonomic identification

Taxonomic identification used the lowest common ancestor approach based on nucleotide BLAST searches against a local reference database constructed from 18S sequences from the NCBI nucleotide database (in August 2014) and the SILVA database version NR99_119 (Pruesse *et al.*, 2007), excluding uncultured organisms. All dereplicated reads were directly used as queries in parallel BLAST searches, taking less than a day to process all reads. The reads with top BLAST hits greater than 370 bp were retrieved using custom Perl scripts (the length of the reads after removing primers ranged from 345 bp to 664 bp, with a mean of 432 bp and median of 428 bp). Only hits to metazoans with a minimum of 75% sequence identity were retained for further analysis. Operational taxonomic units (OTUs) were defined as unique top BLAST hits that were separated according to sequence similarity thresholds (97–100%, 94–96%, 90–93% and 75–89%) as a way to estimate diverged taxonomic lineages (see Table S2). Note that this is not an exact species assignment as OTUs could represent several species, and species could be distributed across OTUs with different thresholds. Furthermore, reads equally matching more than one species (by BLAST identity) due to low interspecific divergence formed a shared OTU. The assignments for higher taxonomic levels

Table 1 Summary of reads and taxonomic assignment. Reads in OTUs are all filtered reads that matched zooplankton in BLAST searches (> 75% sequence identity and 370 bp), and the numbers of OTUs with singletons (OTUs all) and without singletons (OTUs filtered) are derived using the lowest common ancestor approach. Higher level taxa (family, order, class and phylum) are reported after removing singletons.

| Port | Raw Reads | Reads in OTUs | OTUs all | OTUs filtered | Family | Order | Class | Phylum |
|---|---|---|---|---|---|---|---|---|
| Total | 10,277,272 | 7,162,599 | 2799 | 1867 | 368 | 101 | 38 | 19 |
| Vancouver | 1,008,358 | 223,361 | 377 | 298 | 84 | 26 | 14 | 7 |
| Victoria | 456,391 | 194,237 | 670 | 545 | 144 | 44 | 22 | 14 |
| Nanaimo | 789,405 | 349,559 | 649 | 527 | 130 | 43 | 23 | 10 |
| Robert's Bank | 715,442 | 659,515 | 321 | 294 | 105 | 43 | 19 | 12 |
| Churchill | 605,049 | 560,045 | 429 | 376 | 117 | 46 | 24 | 13 |
| Deception Bay | 787,293 | 729,091 | 229 | 211 | 71 | 32 | 15 | 10 |
| Steensby | 799,089 | 750,180 | 232 | 211 | 69 | 36 | 20 | 11 |
| Iqaluit | 767,297 | 713,050 | 321 | 281 | 92 | 41 | 22 | 14 |
| Bayside | 656,488 | 409,208 | 268 | 236 | 83 | 37 | 21 | 13 |
| Sept-Îles | 502,688 | 252,931 | 465 | 391 | 88 | 35 | 18 | 12 |
| Hawkesbury | 444,315 | 240,579 | 494 | 413 | 123 | 41 | 21 | 12 |
| Halifax | 770,511 | 715,428 | 314 | 274 | 113 | 49 | 28 | 13 |
| Thunder Bay | 556,984 | 331,814 | 511 | 411 | 76 | 31 | 19 | 10 |
| Nanticoke | 480,962 | 252,639 | 430 | 370 | 73 | 30 | 18 | 11 |
| Montreal | 634,126 | 579,619 | 147 | 131 | 33 | 17 | 12 | 8 |
| Hamilton | 1,099,458 | 201,343 | 234 | 190 | 30 | 17 | 12 | 10 |

were inferred based on the lowest common ancestor of the top BLAST hits acquired from the NCBI taxonomy database. Undetermined taxonomic levels remained because of lack of information or top BLAST hits equally matching species from different groups (genera, family, etc.). A subset of taxonomic names was changed to the naming scheme of the WoRMS database (WoRMS Editorial Board, 2015) when discrepancies were found. To investigate the potential phylogenetic placement of taxa with top BLAST matches below 90% sequence identity, multiple sequence alignments were performed with default settings in MAFFT v7.150b (Katoh & Standley, 2013). Phylogenetic reconstruction was performed using FASTTREE version 2.1.7 (Price *et al.*, 2010), and genetic similarity was estimated using dnadist from the PHYLIP package version 3.695 (Felsenstein, 1989).

### Clustering of 18S sequences prior to taxonomic assignment

Clustering reads together based on sequence similarity permits a taxonomy-independent approach for the estimation of biodiversity. However, evolutionary properties of 18S rRNA such as length variation and different levels of genetic diversity among taxonomic groups challenge the accuracy of clustering for complex zooplankton communities, as we have previously shown using mock communities (Brown *et al.*, 2015; Flynn *et al.*, 2015). To assess whether pre-clustering affected taxonomic assignment, we clustered dereplicated reads from all ports together using a 97% sequence similarity threshold with UPARSE implemented in USEARCH v8.0.1616 (Edgar, 2013). The representative sequence of each cluster was then used as a BLAST query to identify each cluster of reads following the same criteria as described for the assignment of individual reads above. Overall, diversity in terms of both taxonomic richness and abundance was highly similar between taxonomic assignments with and without pre-clustering, but fewer taxa were identified after clustering (118 fewer families). This can be explained by the decreased sensitivity of clustering, which grouped sequences from closely related organisms and failed to detect low abundance taxa because of low interspecific divergence. For example, over 1000 reads matched the calanoids *Scolecithricella* and *Euchaeta* (> 97% identity) when clustering was not performed. After clustering, these reads were grouped with closely related calanoids with high sequence similarity (see Figs S1 and S2), and therefore were misclassified. In contrast, high intraspecific diversity contributed to numerous clusters matching the same species (for example 76 clusters matched *Paracalanus parvus*). We therefore present taxonomic assignment without pre-clustering because of its greater sensitivity in distinguishing closely related species and identifying reads at low abundance, which are important for conservation monitoring. The main difference between this approach and most metabarcoding bioinformatics workflows is that sequence clustering is skipped altogether and each read is directly identified.

### Taxonomic richness and abundance

Reads predominantly matched zooplankton groups, although several nonmetazoans were also sparsely captured in our data including plants, algae, and fungi (combined abundance < 1000 reads), as well as ciliates and other Chromalveolata groups (42,431 reads; see Table S2). Only metazoans were kept for analyses. For comparisons across samples, taxa represented by a single read (singletons) were excluded to reduce the impact of potentially spuriously amplified taxa. Taxonomic richness extrapolation curves were performed using INEXT (Hsieh *et al.*, 2013).

Rarefaction was performed with the vegan package (Oksanen *et al.*, 2015) in R version 3.1.3 (R Core Team, 2015) to accommodate differences in sequencing depth across ports. Proportional abundance of taxa following rarefaction was used to compare relative abundances among samples and ports rather than absolute numbers. Taxonomic abundance mosaic plots were generated with ggplot2 (Wickham, 2009) and hierarchical charts were created using Krona (Ondov *et al.*, 2011).

### Spatial and temporal variation

Diversity indices (Richness R, Chao1 and ACE estimators, Shannon diversity H′, Simpson diversity D, Pielou's evenness J and taxonomic distinctness Δ*) were calculated with the vegan package in R after rarefaction and removing singletons. Similar results were found between R and Chao1 and ACE estimators, and between H' and D, so only R and H' are presented. Pairwise beta-diversity coefficients for both incidence (Jaccard dissimilarity index) and abundance (Bray–Curtis dissimilarity index) were calculated with the VEGAN package in R. Community composition differences among samples were evaluated using PERMANOVA (Anderson, 2001), and similarities were assessed using the Pearson product-moment correlation coefficient. Principal component analysis and creation of heatmap plots were performed in R after log-transformation of rarified read counts.

## RESULTS

### Taxonomic identification

Taxonomic assignment of bulk zooplankton samples from sixteen of Canada's busiest regional ports (Fig. 1) revealed 7,162,599 (92%) metazoan reads. The majority of these reads (94%) matched a reference sequence above or equal to a 90% identity threshold, with 81% having at least a 97% identity threshold. We detected 2799 OTUs (operational taxonomic units) as defined by sequence identity thresholds (see Methods). However, the combination of incomplete reference databases and low interspecific genetic variation among certain taxa resulted in the identification of only 379 species from 61% of all reads; those uniquely matching one species from online reference databases beyond 97%

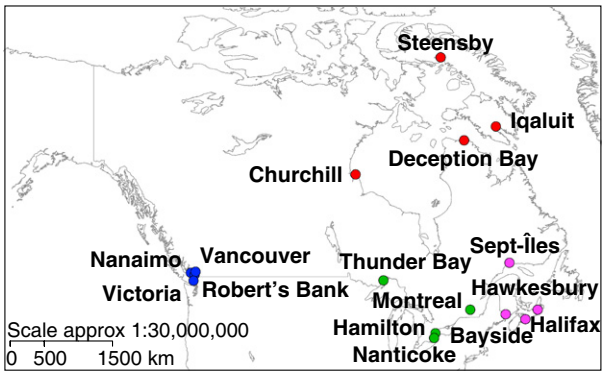


Figure 1 Location of the 16 sampled ports. Samples are from four geographical regions (Pacific Ocean, Arctic Ocean, Atlantic Ocean and Great Lakes).

sequence identity (40% of reads matched 245 species with at least 99% identity; Fig. 2). The rest of the OTUs could not be assigned a unique species name: 373 OTUs matched more than one species that could not be distinguished, 72 OTUs matched a database entry not taxonomically classified to the species level, and 1975 OTUs matched a database entry below 97% identity (see Table S2). Less than 5% of all reads (37% of the OTUs) could also not be assigned to the genus level, mainly because they matched species from different genera equally. Taxonomy was therefore assessed using the lowest common ancestor approach, and analyses were performed at the family level to maximize data inclusion. For example, 5% of all reads matched calanoid copepods of the family Acartiidae below 90% identity (constituting more than 85% of the reads with best BLAST hits below 90%). These could not be assigned a species name, but they form geographical clusters and are monophyletic with other *Acartia* species (see Fig. S1). *Acartia* species have unusually high genetic diversity among calanoid copepods; nine species of *Acartia* have a mean genetic similarity of 73% in the V4 region, compared to 99% among 37 species from 12 genera in the calanoid family Diaptomidae. Overall, 33% of all OTUs were represented by a single read (singletons), and 66% had fewer than 10 reads. Although singleton OTUs (932) were excluded from further analyses to reduce the possible effects of spurious signals, they could represent organisms at very low abundance and are included in Table S2.

## Taxonomic richness and abundance

Taxonomic richness rarefaction and extrapolation curves level off as sequencing depth increases, suggesting that most ports were reasonably well sampled and the data approached saturation (see Fig. S3). Moreover, after removing singleton OTUs, there was a significant negative linear correlation between the number of reads and the number of OTUs ($P = 0.04$, adjusted $R^2 = 0.22$) and no significant correlation with the number of family-level taxa ($P = 0.75$, adjusted $R^2 = -0.06$). Of the 1867 nonsingleton metazoan OTUs, 19 phyla were represented across 38 classes, 101 orders and 368 families (Table 1). Taxonomic richness varied substantially across the 16 ports in which 39% of all families and 24% of all orders belonged to the phylum Arthropoda. The most diverse taxonomic groups were Copepoda (601 OTUs) and Decapoda (129 OTUs). The copepods spanned 64 families from 7 orders including 19 Calanoida, 14 Harpacticoida and 13 Cyclopoida (Fig. 3). Most OTUs occurred at low abundance; 99% of all reads belonged to 5–16 OTUs and 2–14 families within a port. Arthropoda was the most prevalent phylum with 93% of all reads, the majority being copepod crustaceans (see Fig. S4). The most abundant were the calanoid families Temoridae (22%; *Eurytemora* spp. and *Temora* spp. predominantly in the Atlantic and Great Lakes samples), and Clausocalanidae (18%; *Pseudocalanus* spp. predominantly in the Arctic and Pacific; Fig. 4). Most of the dominant taxa could be identified to genus and many to species (Fig. 4). While some species were commonly found across marine ports (*Pseudocalanus elongatus* and *Oithona similis*), others were more confined to a particular coast (*Paracalanus parvus* in the Pacific and *Temora longicornis* in the Atlantic).

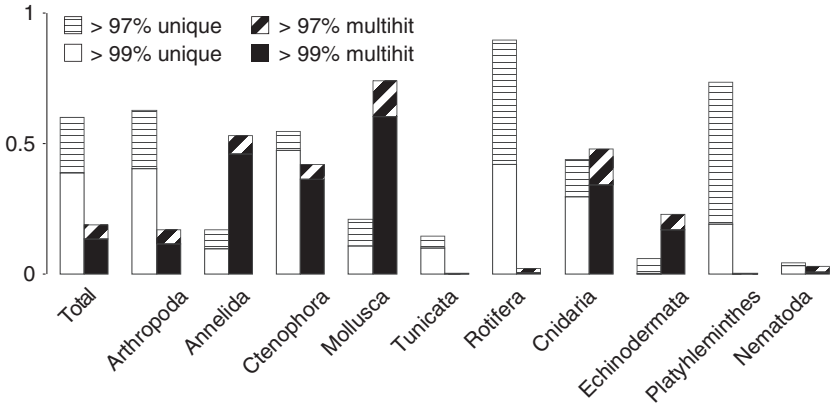


Figure 2 Taxonomic assignment and richness. Proportion of reads with a single best (unique) hit to 18S database reference sequences with more than 97% and 99% sequence identity (over a minimum length of 370 bp); and reads equally matching more than one species (multihit) or matching reference sequences without species information with more than 97% and 99% sequence identity. Categories include all reads (Total) and proportion of reads classified to major higher level taxa.

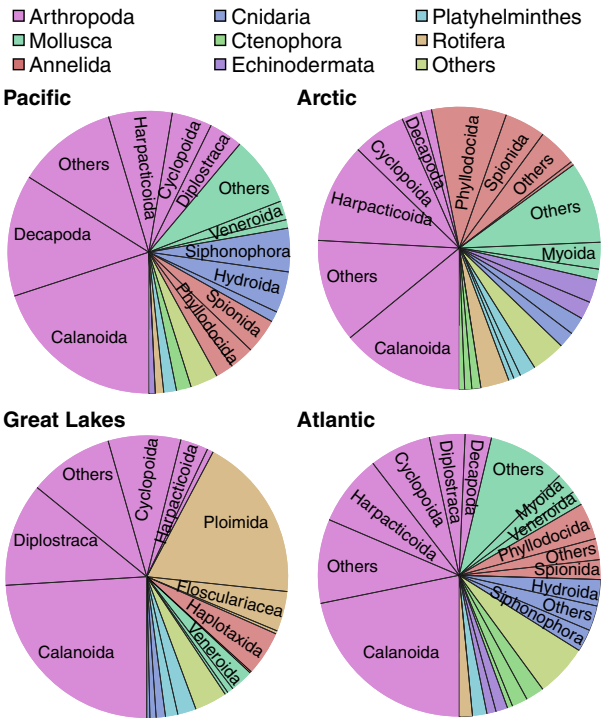


Figure 3 OTU richness among orders across sampled regions. The proportion of taxa detected from different orders and phyla are shown for each region.

### Comparison with previous morphological surveys

Although many organisms could not be identified to species, most of the zooplankton groups that have been previously reported along Canadian marine coasts were detected. Of 14 classes and 39 orders reported from a meta-analysis performed in 2010 spanning the Pacific, Arctic and Atlantic oceans (Archambault *et al.*, 2010), we found 12 and 29, respectively (see Table S3). As for freshwater zooplankton, the majority of the common crustacean families were detected. We recovered 16 of 25 families and 32 of 74 genera reported by Balcer *et al.* (1984) and Hudson & Lesko (2003). All but one undetected genus had partial or no 18S reference sequences available, precluding identification in our study using metabarcoding (see Table S3). It is therefore likely that other organisms in our samples are not represented in 18S databases but generated a BLAST hit to a related species. In contrast, we detected 29 freshwater crustacean families and 35 marine orders that were not reported in these surveys, possibly including some non-indigenous species (Brown *et al.*, unpublished), highlighting the sensitivity of metabarcoding (see Table S3).

### Spatial and temporal variation

Zooplankton community composition was generally consistent across independent transects from the same port and season (Fig. 4), with a mean Pearson's $r$ almost twice as high as among ports from the same geographical region (see Table S4); there were no significant differences in community composition among transects for either OTUs (Jaccard: $F = 0.7$, $P = 0.994$; Bray–Curtis: $F = 0.6$, $P = 0.999$; PERMANOVA) or families (Jaccard: $F = 0.6$, $P = 0.996$; Bray–Curtis: $F = 0.5$, $P = 0.996$; PERMANOVA). In contrast, there were significant differences in community composition among ports and among coasts at both OTU and family levels, and in both incidence and abundance (all $P < 0.001$; PERMANOVA; see Fig. S5). Freshwater ports had lower taxonomic richness R ($P = 0.025$; Mann–Whitney U-test) than marine ports at the family level (see Fig. S6). Arctic ports had higher taxonomic evenness than ports in other regions ($P = 0.020$ for OTUs, $P = 0.030$ for families; Mann–Whitney U-test). Taxonomic distinctness (Warwick & Clarke, 1995), a measure of taxonomic relatedness capturing phylogenetic diversity, was particularly high in two Arctic ports (Steensby and Churchill; see Fig. S6). Among marine ports, incidence-based analyses showed prominent differences between the Arctic and the Pacific and Atlantic coastal regions, whereas abundance-based analyses revealed higher similarity between Arctic and Pacific samples (Fig. 5). The Atlantic Bayside samples formed a distinct cluster between marine and freshwater samples in both incidence and abundance analyses, with high average pairwise beta diversity among marine samples (see Table S4). Bayside is, in fact, an estuarine port with a characteristic community dominated by two calanoid copepod genera, *Eurytemora* (59% of reads) and *Acartia* (33% of reads; see Fig. S7a), leading to low alpha diversity (see Fig. S6). Several taxonomic groups occurred predominantly or uniquely in one port or geographical region, the main differences being between marine and freshwater environments (see Figs S7b and S8).

For 12 of the 16 ports (three from each geographical region) biodiversity could be compared across two sampling seasons. Even though sampling at each port was on six independently sampled transects that introduced spatial variation within seasons, at each port, community composition at both the OTU and family levels was overall significantly different between seasons (Jaccard: $F = 13.2$, $P < 0.001$ and Bray–Curtis: $F = 6.9$, $P < 0.001$ for OTUs; Jaccard: $F = 17.1$, $P < 0.001$ and Bray–Curtis: $F = 6.4$, $P = 0.002$ for families; PERMANOVA). Seven of the 12 ports had at least 1 species that differed in proportional abundance between seasons by more than 35% (see Fig. S9). Pronounced seasonal differences included the proportional abundance of *Daphnia* among Montreal samples from the Saint Lawrence River (42% in July, < 1% in September), *Temoridae* species among Halifax samples from the Atlantic (62% in August, 13% in December) and *Paracalanus parvus*, *Metridia* spp. and *Pseudocalanus* spp. from Pacific samples (Nanaimo: 62% *P. parvus* and 1% *Metridia* spp. in July vs. 25% and 40% in December; Robert's Bank: 41% *P. parvus* and 12% *Pseudocalanus* spp. in July vs. 5% and 82% in December). The most striking community turnover occurred between spring (May) and summer (August) among the subarctic Churchill samples (Jaccard dissimilarity of 0.77 for OTUs and 0.67 for

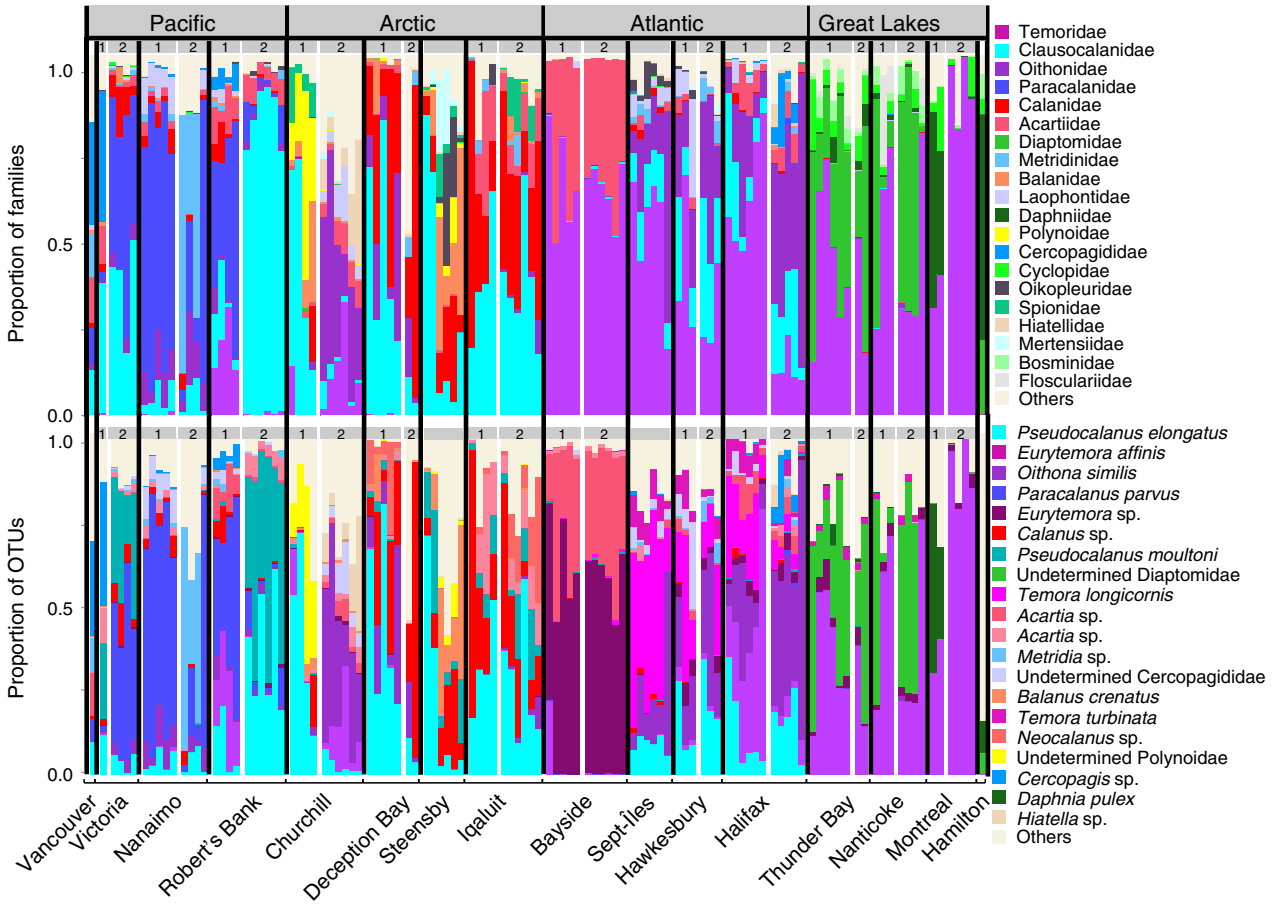


Figure 4 Proportional taxonomic abundance of each family and OTU for each sample. Ports for which we have samples from different seasons are divided into seasons 1 and 2.

families; Bray–Curtis dissimilarity of 0.91 for OTUs and 0.90 for families; see Table S4; Fig. 5). In this port, spring samples were dominated by *Pseudocalanus elongatus* (45%) and polychaete annelids (36%), whereas the summer samples had relatively few *Pseudocalanus* (3%) and annelids (3%) but were dominated by *Oithona similis* (38%) and other copepods (30%), as well as molluscs (15%; Fig. 6; see Fig. S9).

## DISCUSSION

Our biodiversity assessment based on metabarcoding data from ports that experience heavy shipping traffic supplements previous surveys using morphological taxonomy. We recovered the main taxonomic groups in addition to identifying organisms from over 30 orders that are probably difficult to detect using traditional morphological methods and/or present in low abundance. The taxonomy-dependent method implemented for characterizing metabarcoding data allowed us to improve the accuracy and sensitivity of taxonomic assessment compared to methods involving pre-clustering of sequences. Community profiles strongly reflected regional and seasonal differences, which are essential components to capture in monitoring surveys given globally changing environments.

Taxonomic assignment can be efficient for the fraction of species that are represented in sequence databases, and particularly useful for higher taxonomic levels. Our approach allowed us to link 61% of all metabarcoding reads with a species name (> 97% identity threshold; Fig. 2). A total of 94% of reads matched a reference sequence with at least 90% identity, and if these can be reliably assigned to higher levels of taxonomy, our results indicate that species from certain phyla like nematodes and tunicates are clearly underrepresented in current 18S databases (Fig. 2). Previous work using mock zooplankton communities has shown that conventional OTU clustering protocols for 18S metabarcoding data are sensitive to similarity thresholds and can lead to both under- and overestimations of biodiversity; these are due to inherent challenges in properly aligning and clustering 18S sequences of variable length, and the presence of different amounts of variation within and among species (Brown *et al.*, 2015; Flynn *et al.*, 2015). With pre-clustering, we found that reads seemingly belonging to different species clustered together because of low interspecific sequence divergence (see Fig. S2) and that reads matching the same species formed separate clusters because of high intraspecific sequence divergence. Direct taxonomic assignment of each read without pre-clustering was more effective in

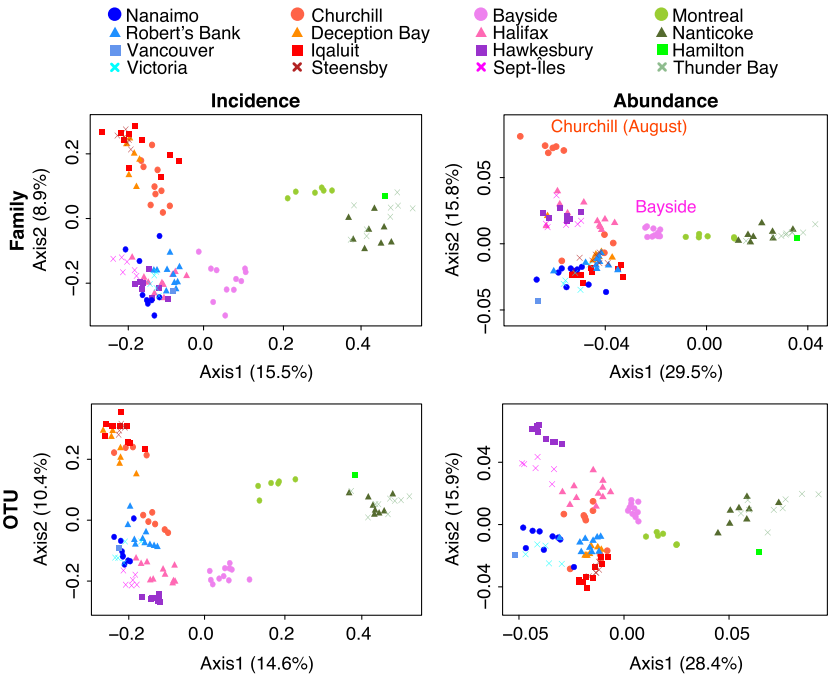


Figure 5 Biodiversity differences among ports. Principal component analysis of taxonomic composition (families and OTUs) calculated using the Jaccard index (incidence based) and the Bray–Curtis index (abundance based), with each data point representing a particular sample (site and season) from each port. Each port is represented from the Pacific (blue), Arctic (red), Atlantic (purple) or Great Lakes (green).

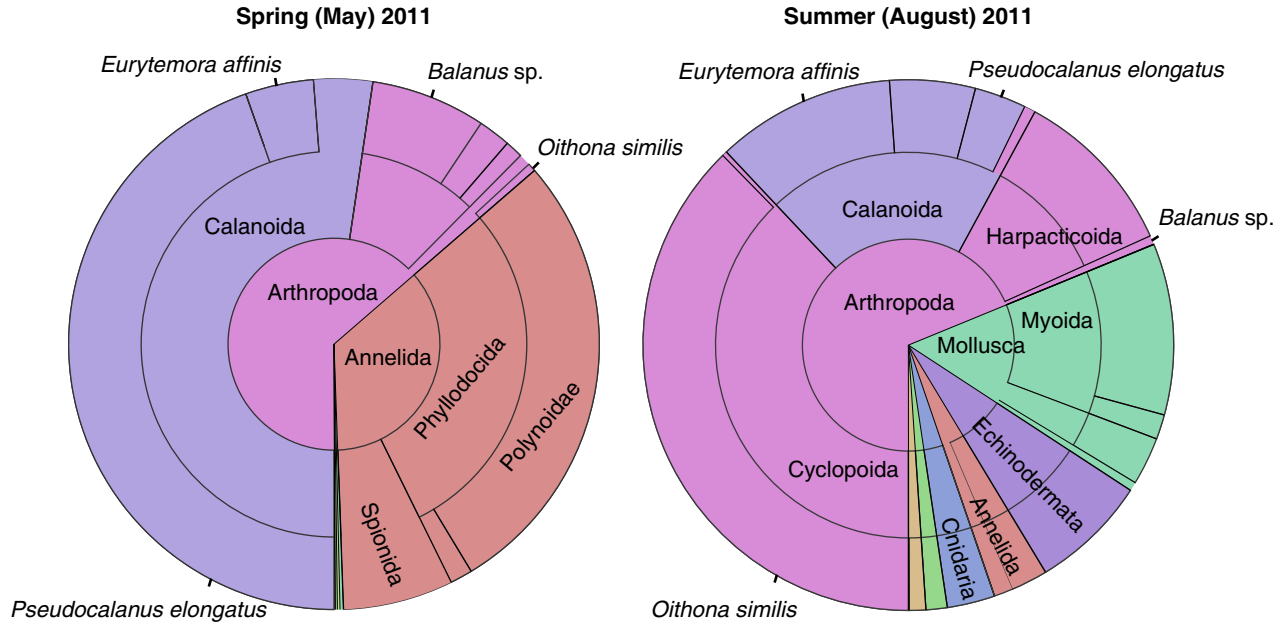


Figure 6 Zooplankton community differences between seasons in Churchill. The different layers represent phyla (central), orders and families (peripheral), with prominent arthropod species labelled.

distinguishing between closely related species and more sensitive in detecting taxa from reads at low abundance, both crucial elements for detecting early stages of species invasions. We therefore advocate that for the purpose of assigning taxonomy and identifying species, particularly those occurring at low abundance, taxonomic classification of metabarcoding data should be performed without sequence pre-clustering. Overall biodiversity profiles, however, which are driven by commonly occurring taxa, are largely unaffected by clustering.

Proportional abundance was not significantly different among transects from the same port (Fig. 4) despite being sampled at different times of the day and tidal cycle, suggesting that large proportional abundance differences among ports and between seasons reveal gross community composition differences. For example, both incidence and abundance biodiversity profiles revealed strong differences between freshwater and marine communities, as well as estuarine communities. An important caveat in interpreting species abundance from metabarcoding data is that estimations are

influenced by technical and biological factors including PCR and primer biases, filtering thresholds, intraspecific copy number differences and the variation in cell density contributed to the initial pool of DNA (Pompanon *et al.*, 2012). Although the relative abundance based on read counts can reflect natural biodiversity (Porazinska *et al.*, 2010; Leray & Knowlton, 2015), the strong relationship between read abundance and biomass will lead to the underestimation of the number of smaller organisms (Lindeque *et al.*, 2013).

Our metabarcoding study found similar diversity patterns to those reported previously for both marine and freshwater environments based on morphological taxonomy. For example, calanoid copepods were the most diverse group of zooplankton across samples from the three Canadian coastlines, the same as reported in a meta-analysis of biodiversity in Canadian oceans (Archambault *et al.*, 2010). For some taxa, we found comparatively more or less diversity than previous zooplankton surveys based on morphological identification. Differences among studies could result from sampling parameters (e.g. collection method, time, location), technological biases (e.g. primer and amplification biases) and reference database extensiveness. On the one hand, we found higher diversity than morphological zooplankton surveys among Mollusca, Annelida, Platyhelminthes and Crustacea (Diplostraca branchiopods and Siphonostomatoida copepods), as well as Cnidaria, Bryozoa, harpacticoid and cyclopoid copepods in the Pacific and polychaetes in the Arctic (Archambault *et al.*, 2010; Carr, 2011). Some of these differences are probably due to the inherent difficulties in distinguishing closely related species morphologically and in capturing larval stages of certain taxonomic groups (Lindeque *et al.*, 2013), showcasing the power of metabarcoding. On the other hand, we found lower diversity of taxa among the phyla Chaetognatha, Echiura and Sipuncula, as well as the classes Ostracoda and Malacostraca (in particular the order Amphipoda). The under-representation of species in metabarcoding data is partly due to the lack of interspecific genetic diversity of the chosen marker (see Fig. S2). This illustrates the limitation of using a single genetic locus for taxonomic identification across a broad taxonomic range. Overall, however, our study recovered similar phylogenetic diversity to that recovered using degenerate COI primers for marine metazoans (Geller *et al.*, 2013; Lobo *et al.*, 2013). Metabarcoding with multiple loci may be crucial for differentiating certain species that are otherwise difficult to characterize using single markers (Cowart *et al.*, 2015). An equally important limitation further preventing species classification is the taxonomic diversity that is unaccounted for in reference databases. For example, NCBI currently holds 304 18S sequences and 1631 COI sequences for organisms in the class Ostracoda, but only 8 species are represented by both markers. Furthermore, these sequences represent only 108 (18S) and 113 (COI) species because of multiple entries for the same species and others identified only above the species level. The only way to resolve this issue is to continue developing molecular databases for broadly used barcoding markers based on vouchered specimens, and in particular those of invasive species for targeted conservation monitoring programmes.

Our data describe significant and substantial seasonal community turnover within ports. Most ports had at least one abundant OTU that changed in proportional abundance more than 35% between seasons. Our results suggest that the port of Churchill in Hudson Bay exhibited a major shift in community composition from spring to summer, during which river inflows and nutrient availability change considerably (Link *et al.*, 2011). The summer season had a distinct community profile for all samples indicating relatively low variation among transects (Fig. 5). These observed seasonal differences could have been influenced for example by melting sea ice or the extensive amount of freshwater runoff received by the Hudson Bay (VanderZwaag *et al.*, 2012), possibly reflected in the rotifer community that was twice as rich and more abundant in summer samples (see Table S5). Differences between seasons could, however, also be influenced by stochastic environmental fluctuations unrelated to season-specific changes. Among Canadian Arctic ports, Churchill has the highest relative risk of species invasion because of the high vessel traffic (Chan *et al.*, 2012). Continued thawing of the Arctic icepack is projected to increase ship traffic (Smith & Stephenson, 2013), raising concerns about potential range expansions and non-indigenous species invasions between oceans (Chan *et al.*, 2015). We provide zooplankton biodiversity baselines that can be useful in further monitoring efforts to help in tracking local communities and differentiating between native and non-indigenous species (Goldsmit *et al.*, 2014). Our findings highlight the importance of routine sampling across seasons for monitoring programs to capture the broad range of organisms the relative abundances of which can fluctuate drastically through time.

Evaluating community composition and species distributions is important for understanding long-term ecological repercussions of environmental perturbations. Large-scale biodiversity surveys, especially those that can uncover low abundance taxa and cryptic diversity, are necessary to assess ongoing spatio-temporal changes and for continued monitoring and protection of ecosystems. Metabarcoding offers a powerful tool to achieve these goals, although taxonomic approaches depend on a vouchered reference database to assign names, highlighting the need for the identification of reference specimens by expert taxonomists coupled with continued development of high-quality molecular databases (Cristescu, 2014; Cowart *et al.*, 2015). We nonetheless detected significant community composition differences among ports and habitats at higher taxonomic levels, suggesting that metabarcoding can be useful for describing coarse level biodiversity trends. The intimate integration of taxonomy into the metabarcoding framework enhances our ability to make relevant comparisons among studies that use different identification approaches and markers, as well as to resolve conservation issues such as the status of endangered

species and the introduction of non-indigenous organisms. Continued improvements should enhance the efficiency of genetic tools for monitoring biodiversity in globally changing environments.

## ACKNOWLEDGEMENTS

We thank our NSERC Canadian Aquatic Invasive Species Network (CAISN) colleagues, in particular the sample collection teams including Kimberly Howland, Rob Young, Jesica Goldsmit, Aibin Zhan and Samir Qureshi. We also thank Cathryn Abbott, Sally Adamowicz, Teri Crease, Jesica Goldsmit and Rob Young for insightful discussions. This project has been supported by the NSERC CAISN and by NSERC Discovery Grants and Canada Research Chairs to HJM and MEC. The authors declare no conflict of interests.

## DATA ACCESSIBILITY

Sequencing data are archived in the European Nucleotide Archive (ENA) under project PRJEB11768 (http://www.ebi.ac.uk/ena/data/view/PRJEB11768).

## REFERENCES

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Appeltans, W., Ahyong, S.T., Anderson, G. *et al.* (2012) The magnitude of global marine species diversity. *Current Biology*, **22**, 2189–2202.

Archambault, P., Snelgrove, P.V.R., Fisher, J.A.D., Gagnon, J.-M., Garbary, D.J., Harvey, M., Kenchington, E.L., Lesage, V., Levesque, M., Lovejoy, C., Mackas, D.L., McKindsey, C.W., Nelson, J.R., Pepin, P., Piché, L. & Poulin, M. (2010) From sea to sea: Canada's three oceans of biodiversity. *PLoS One*, **5**, e12182.

Aylagas, E., Borja, A. & Rodríguez-Ezpeleta, N. (2014) Environmental status assessment using DNA metabarcoding: towards a genetics based Marine Biotic Index (gAMBI). *PLoS One*, **9**, e90529.

Baird, D.J. & Hajibabaei, M. (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Baiser, B., Olden, J.D., Record, S., Lockwood, J.L. & McKinney, M.L. (2012) Pattern and process of biotic homogenization in the New Pangaea. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 4772–4777.

Balcer, M.D., Korda, N.L. & Dodson, S.I. (1984) *Zooplankton of the Great Lakes: a guide to the identification and ecology of the common crustacean species*. The University of Wisconsin Press, Madison, Wisconsin.

Bik, H.M., Halanych, K.M., Sharma, J. & Thomas, W.K. (2012) Dramatic shifts in benthic microbial eukaryote communities following the Deepwater Horizon oil spill. *PLoS One*, **7**, e38550.

Brown, E.A., Chain, F.J.J., Crease, T.J., MacIsaac, H.J. & Cristescu, M.E. (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecology and Evolution*, **5**, 2234–2251.

Carr, C.M. (2011) Polychaete diversity and distribution patterns in Canadian marine waters. *Marine Biodiversity*, **42**, 93–107.

Chan, F.T., Bronnenhuber, J.E., Bradie, J.N., Howland, K.L., Simard, N. & Bailey, S.A. (2012) Risk assessment for ship-mediated introductions of aquatic nonindigenous species to the Canadian Arctic. Department Fisheries and Oceans, Canadian Science Advisory Secretariat Research Document 2011/105. Available at: http://www.dfo-mpo.gc.ca/csas-sccs/Publications/ResDocs-DocRech/2011/2011_105-eng.html (accessed 19 March 2015).

Chan, F.T., MacIsaac, H.J. & Bailey, S.A. (2015) Relative importance of vessel hull fouling and ballast water as transport vectors of nonindigenous species to the Canadian Arctic. *Canadian Journal of Fisheries and Aquatic Sciences*, **72**, 1230–1242.

Chariton, A.A., Stephenson, S., Morgan, M.J., Steven, A.D.L., Colloff, M.J., Court, L.N. & Hardy, C.M. (2015) Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental pollution*, **203**, 165–174.

Cowart, D.A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J. & Arnaud-Haond, S. (2015) Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS One*, **10**, e0117562.

Cristescu, M.E. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology and Evolution*, **29**, 566–571.

Darling, J.A. & Mahon, A.R. (2011) From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research*, **111**, 978–988.

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

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**, 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Felsenstein, J. (1989) PHYLIP – phylogeny inference package. *Cladistics*, **5**, 164–166.

Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J. & Cristescu, M.E. (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

Geller, J., Meyer, C., Parker, M. & Hawk, H. (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase

subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, **13**, 851–861.

Goldsmit, J., Howland, K.L. & Archambault, P. (2014) Establishing a baseline for early detection of non-indigenous species in ports of the Canadian Arctic. *Aquatic Invasions*, **9**, 327–342.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.

Holland, B.S. (2000) Genetics of marine bioinvasions. *Hydrobiologia*, **420**, 63–71.

Hsieh, T.C., Ma, K.H. & Chao, A. (2013) iNEXT: An R package for interpolation and extrapolation in measuring species diversity. Available at: http://chao.stat.nthu.edu.tw/blog/software-download/inext-r-package (accessed 3 July 2014).

Hudson, P.L. & Lesko, L.T. (2003) Free-living and Parasitic Copepods of the Laurentian Great Lakes: Keys and Details on Individual Species. Available at: http://www.glsc.usgs.gov/greatlakescopepods/default.php (accessed 29 March 2015).

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Leray, M. & Knowlton, N. (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the USA*, **112**, 2076–2081.

Lindeque, P.K., Parry, H.E., Harmer, R.A., Somerfield, P.J. & Atkinson, A. (2013) Next generation sequencing reveals the hidden diversity of zooplankton assemblages. *PLoS One*, **8**, e81327.

Link, H., Archambault, P., Tamelander, T., Renaud, P.E. & Piepenburg, D. (2011) Spring-to-summer changes and regional variability of benthic processes in the western Canadian Arctic. *Polar Biology*, **34**, 2025–2038.

Lobo, J., Costa, P.M., Teixeira, M.A.L., Ferreira, M.S.G., Costa, M.H. & Costa, F.O. (2013) Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. *BMC Ecology*, **13**, 34.

Miller, A.W. & Ruiz, G.M. (2014) Arctic shipping and marine invaders. *Nature Climate Change*, **4**, 413–416.

Mooney, H.A. & Cleland, E.E. (2001) The evolutionary impact of invasive species. *Proceedings of the National Academy of Sciences of the USA*, **98**, 5446–5451.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H. & Wagner, H. (2015) vegan: Community Ecology Package. Available at: http://CRAN.R-project.org/package=vegan (accessed 22 March 2015).

Ondov, B.D., Bergman, N.H. & Phillippy, A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N. & Taberlet, P. (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931–1950.

Porazinska, D.L., Sung, W., Giblin-Davis, R.M. & Thomas, W.K. (2010) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources*, **10**, 666–676.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. & Glöckner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, **35**, 7188–7196.

R Development Core Team (2015) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radulovici, A.E., Archambault, P. & Dufresne, F. (2010) DNA barcodes for marine biodiversity: moving fast forward? *Diversity*, **2**, 450–472.

Schloss, P.D. & Westcott, S.L. (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology*, **77**, 3219–3226.

Smith, L.C. & Stephenson, S.R. (2013) New Trans-Arctic shipping routes navigable by midcentury. *Proceedings of the National Academy of Sciences of the USA*, **110**, E1191–E1195.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Urban, M.C. (2015) Climate change. Accelerating extinction risk from climate change. *Science*, **348**, 571–573.

VanderZwaag, D.L., Hutchings, J.A., Jennings, S. & Peterman, R.M. (2012) Climate change, fisheries, and aquaculture: trends and consequences for Canadian marine biodiversity. *Environmental Reviews*, **20**, 312–352.

Vermeij, G.J. & Roopnarine, P.D. (2008) The coming Arctic invasion. *Science*, **321**, 780–781.

Warwick, R.M. & Clarke, K.R. (1995) New "biodiversity" measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, **129**, 301–305.

Wickham, H. (2009) *ggplot2*. Springer New York. Available at: http://ggplot2.org/ (accessed 16 June 2014).

WoRMS Editorial Board (2015) World Register of Marine Species. Available at: http://www.marinespecies.org at VLIZ (accessed 20 May 2015).

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhan, A., Bailey, S.A., Heath, D.D. & MacIsaac, H.J. (2014) Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in

complex communities. *Molecular Ecology Resources*, **14**, 1049–1059.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. & Gemeinholzer, B. (2015) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, **15**, 526–542.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Figure S1** Phylogenetic diversity of copepods.

**Figure S2** Genetic similarity among copepods.

**Figure S3** Rarefaction and extrapolation curves.

**Figure S4** Proportional abundance.

**Figure S5** Taxonomic similarity among ports.

**Figure S6** Diversity indices among ports.

**Figure S7** Proportional abundance differences among ports.

**Figure S8** Differences in families among ports.

**Figure S9** Proportional abundance differences between seasons.

**Table S1** Data collection and sequencing information.

**Table S2** Full taxonomic list.

**Table S3** Comparison between our study and morphological surveys.

**Table S4** Pearson correlations and beta diversity among ports.

**Table S5** Rotifer differences between Churchill seasons.

## BIOSKETCH

**Frédéric Chain** is an evolutionary biologist studying evolutionary genomics and bioinformatics. All authors are part of the Canadian Aquatic Invasive Species Network (CAISN) www.caisn.ca.

Author contributions: HJM and MEC conceived and designed the experiment; EAB collected and prepared the DNA; FJJC and EAB developed and designed the methodology; FJJC implemented the analyses and wrote the manuscript; and FJJC, EAB, HJM and MEC structured and edited the manuscript.

Editor: Robert Cowie