MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12254

Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities

AIBIN ZHAN,*† SARAH A. BAILEY,‡ DANIEL D. HEATH† and HUGH J. MACISAAC†

*Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing 100085 China, †Great Lakes Institute for Environmental Research, University of Windsor, 401 Sunset Avenue, Windsor, Ontario, Canada, N9B 3P4, ‡Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans Canada, 867 Lakeshore Road, Burlington, Ontario, Canada L7R 4A6

Abstract

Metabarcode surveys of DNA extracted from environmental samples are increasingly popular for biodiversity assessment in natural communities. Such surveys rely heavily on robust genetic markers. Therefore, analysis of PCR efficiency and subsequent biodiversity estimation for different types of genetic markers and their corresponding primers is important. Here, we test the PCR efficiency and biodiversity recovery potential of three commonly used genetic markers – nuclear small subunit ribosomal DNA (18S), mitochondrial cytochrome *c* oxidase subunit I (COI) and 16S ribosomal RNA (mt16S) – using 454 pyrosequencing of a zooplankton community collected from Hamilton Harbour, Ontario. We found that biodiversity detection power and PCR efficiency varied widely among these markers. All tested primers for COI failed to provide high-quality PCR products for pyrosequencing, but newly designed primers for 18S and 16S passed all tests. Furthermore, multiple analyses based on large-scale pyrosequencing (i.e. 1/2 PicoTiter plate for each marker) showed that primers for 18S recover more (38 orders) groups than 16S (10 orders) across all taxa, and four vs. two orders and nine vs. six families for Crustacea. Our results showed that 18S, using newly designed primers, is an efficient and powerful tool for profiling biodiversity in largely unexplored communities, especially when amplification difficulties exist for mitochondrial markers such as COI. Universal primers for higher resolution markers such as COI are still needed to address the possible low resolution of 18S for species-level identification.

Keywords: 454 pyrosequencing, aquatic community, cytochrome *c* oxidase subunit I (COI), efficiency test, mitochondrial 16S ribosomal RNA (mt16S), small subunit ribosomal DNA (18S)

Received 18 December 2013; revision received 27 February 2014; accepted 10 March 2014

Introduction

Aquatic ecosystems are among the most threatened habitats globally (Dudgeon *et al.* 2006; Hambler *et al.* 2011; Holland *et al.* 2012). Owing to severe threats from interacting stressors including overexploitation, chemical pollution and introductions of invasive species, extinction risk of aquatic species appears higher than that for terrestrial species (Dudgeon *et al.* 2006; Pereira *et al.* 2012). For example, extinction rates of freshwater animals were estimated as more than five times higher than those for terrestrial species in North America (Ricciardi & Rasmussen 1999). By 2012, more than 4600 freshwater animal species were identified as threatened or recently extinct, accounting for more than 25% of all identified freshwater animals (Thomsen *et al.* 2012). There exists a pressing need, therefore, to identify aquatic ecosystems with high endemicity and to develop effective conservation plans to halt biodiversity loss in these ecosystems.

Effective conservation plans largely rely on robust information regarding species composition. In terms of abundance and biomass, aquatic ecosystems are dominated by an array of drifting microscopic organisms collectively referred to as plankton (Machida *et al.* 2009). Plankton play an important role in maintaining ecosystem function (e.g. key role in food webs, Telesh 2004). Empirical studies have shown that change/loss of biodiversity in plankton communities may alter ecosystem function and services (e.g. McMahon *et al.* 2012).

Correspondence: Aibin Zhan, Fax: +86-10-6284-9882; E-mail: zhanaibin@hotmail.com or azhan@rcees.ac.cn

However, it is often difficult, and sometimes impossible, to identify planktonic organisms by traditional methods such as light microscopy, mainly due to poor species integrity following preservation, cryptic immature stages and the vast number of taxa present (Briski *et al.* 2011; Darling & Mahon 2011; Uusitalo *et al.* 2013). In addition, many plankton communities are comprised of a few very abundant species and numerous very rare species, making it difficult to detect and identify all taxa (e.g. Galand *et al.* 2009; Cheung *et al.* 2010). Recently, the advent of robust and sensitive metabarcode methods based on high-throughput sequencing has made environmental DNA-based biodiversity assessments of complex communities possible, especially those dominated by microscopic organisms (e.g. Fonseca *et al.* 2010; Lodge *et al.* 2012; Zhan *et al.* 2013).

A prerequisite for metabarcode surveys using high-throughput sequencing is the selection of robust genetic markers and corresponding universal PCR primers. Good candidate primer pairs are expected to effectively amplify and differentiate a wide range of species in complex communities. One of the best tested markers/genes based on traditional Sanger sequencing is mitochondrial cytochrome *c* oxidase subunit I (COI), mainly attributable to coordinated global barcoding initiatives (Hebert *et al.* 2003). Therefore, we hypothesized that COI could be an ideal candidate marker for high-throughput sequencing-based biodiversity assessment. However, some taxa in aquatic communities, including dominant taxonomic groups such as Copepoda and Cladocera, can be difficult to PCR amplify when using COI for barcoding analysis. For example, it was almost impossible to amplify a globally distributed cladoceran taxon, *Holopedium*, collected from Churchill, Manitoba, Canada at COI (Jeffery *et al.* 2011). Such recent evidence challenged our a priori hypothesis. Consequently, a comparison between different types of genetic markers and their corresponding primer sets is crucial before employing high-throughput sequencing technologies for biodiversity assessment in aquatic communities.

In this study, we examined the efficiency of both nuclear (i.e. small subunit ribosomal DNA, also known as 18S rDNA for eukaryotes, referred to as 18S hereafter) and mitochondrial markers (i.e. cytochrome c oxidase subunit I and 16S ribosomal RNA, hereafter COI and mt16S, respectively) and their corresponding primers for biodiversity assessment of a freshwater plankton community collected from Hamilton Harbour in Lake Ontario, Canada. Based on the results obtained from our evaluation, we sought to select robust primers for biodiversity assessment based on high-throughput sequencing technologies for complex zooplankton communities.

Materials and methods

Field sampling

Zooplankton samples were collected from Hamilton Harbour in September 2011. We used a conical plankton net (80-μm mesh) to collect six geo-referenced zooplankton samples using oblique tows from the bottom to water surface. The collected zooplankton samples were immediately preserved in 100% ethanol and stored at −20 °C until further analyses.

Molecular markers and corresponding primers

We included one nuclear marker, small subunit ribosomal DNA (i.e. 18S), and two mitochondrial DNA (mtDNA) markers, COI and mt16S, for our analysis of the efficiency of plankton biodiversity estimation using 454 pyrosequencing. Because zooplankton communities consist of an array of taxonomic groups, we focused mainly on the dominant group, Crustacea (Adamowicz & Purvis 2005), for the design of new primers and selection of published primers. Moreover, to make newly designed primers applicable to other communities such as living in the benthos in both freshwater and marine habitats, we also included more taxonomic groups from Arthropoda, as well as Mollusca, Tunicata, Echinodermata, Annelida, Nematoda and Platyhelminthes for primer design.

For primer design for each of the three markers, we recovered sequences of representative species in the selected taxonomic groups from GenBank (http://www.ncbi.nlm.nih.gov/nuccore). We initially screened DNA polymorphism among sequences within each taxonomic group to select representative sequences for seeking conserved regions among taxonomic groups. Subsequently, the selected sequences were aligned to find conserved regions to locate universal primers across all taxonomic groups (Fig. 1; and sequence alignment for COI, 18S and mt16S in Appendices S1, S2 and S3, Supporting information, respectively). All primer pairs were designed to amplify approximately 400–600 bp based on the read length (~500 bp) of the 454 GS-FLX Titanium platform and the availability of conserved regions of each gene for primer design. For the COI gene, we failed to identify any conserved region, even within Crustacea (see sequence alignment in Appendix S1, Supporting information). Thus, we chose a number of published 'universal' primers to assess their efficiency for biodiversity assessment in plankton communities.

To characterize pooled PCR product sequences after pyrosequencing, all forward primers selected for pyrosequencing were tagged specifically for each sample using 8nt nucleotide codes (Parameswaran *et al.* 2007). In addition, the 454 adaptors (A and B) were added to the

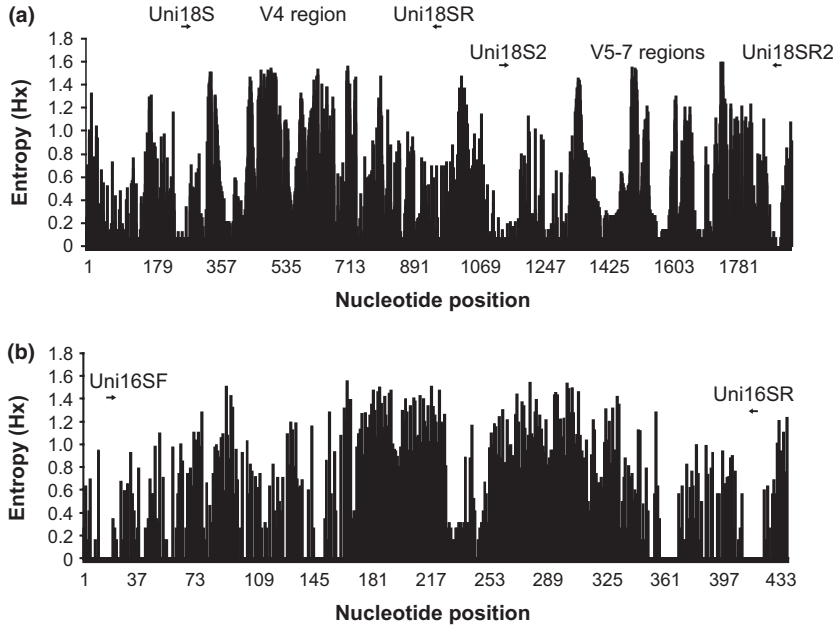

Fig. 1 Information content at each nucleotide position (alignment gaps included) for 'universal' primer design for the two commonly used genetic markers, (a) small subunit ribosomal DNA (18S) and (b) mitochondrial 16S ribosomal RNA (mt16S). Plots are based on aligned sequences of selected sequences deposited in GenBank (see Appendices S2 and S3, Supporting information, for 18S and mt16S, respectively). Entropy (Hx), which is low in conserved sites and high in variable sites, is shown on the *y*-axis. Universal PCR primers were designed in regions of low entropy (i.e. conserved regions). Locations for primer design are shown for each marker.

5′-end of the forward and reverse primers, respectively, to make them compatible with the pyrosequencing process (Table 1).

Primer test

Selected primers were subjected to a series of tests (Fig. 2). First, we subjected the primers to amplification of several crustacean species – the waterfleas *Daphina pulex* and *Cercopagis pengoi* and the European green crab *Carcinus maenas* – to test specificity (i.e. to avoid random amplifications). Second, to explore whether these primers could effectively amplify multiple species in a mixed-species assemblage of zooplankton, we amplified bulk DNA isolated from a zooplankton sample collected in Hamilton Harbour, Ontario. PCR products obtained from successful primer pairs were cloned into a vector using the TA cloning kit (Invitrogen Inc., ON, Canada). Twenty-four clones were randomly selected for each primer pair to perform traditional Sanger sequencing for the multiple species recovery test. Third, we used a small-scale run of 454 pyrosequencing (i.e. an equivalent of 1/48 PicoTiter plate) to assess taxonomic coverage using the same bulk DNA as was used for Sanger sequencing. Finally, we used large-scale pyrosequencing (i.e. 1/2 PicoTiter plate for each primer pair) to test the performance of the selected primer pairs for biodiversity assessment.

DNA extraction, PCR and pyrosequencing

Total genomic DNA was extracted from 100 mg zooplankton sample using the DNeasy Blood and Tissue Kit (Qiagen Canada Inc., ON, Canada). The quality and quantity of extracted DNA were measured by NanoDrop spectrophotometer (NanoDrop Technologies, DE, USA). PCR mixtures (25 μL) were prepared in eight replicates for each sample to avoid possible biased amplification. Each duplicate consisted of 100 ng of genomic DNA, 1 × PCR buffer, 2 mM of Mg^{2+}, 0.2 mM of dNTPs, 0.4 μM of each primer and 2 U of *Taq* DNA polymerase (Genscript). PCR cycling parameters consisted of an initial denaturation step at 95 °C for 5 min, followed by 25 amplification cycles of 95 °C for 30 s, locus-specific annealing temperature (Table 1) for 30 s, 72 °C for 90 s and a final elongation at 72 °C for 5 min. PCR products of duplicates were pooled and purified using the solid-phase reversible immobilization (SPRI) paramagnetic bead-based method (Agencourt, Beverly, MA, USA). Pyrosequencing was performed using 454 Adaptor A (Table 1) on a GS-FLX Titanium platform (454 Life Sciences, Branford, CT, USA) by Engencore at the University of South Carolina.

Pyrosequencing data analysis

After pyrosequencing, raw reads were denoised in MOTHUR version 1.31.2 (Schloss *et al.* 2009) implemented in the pipeline SEED version 1.1.35 (Větrovský & Baldrian 2013). Subsequently, each sequence was sorted based on its unique tag code on the 5′-end of the forward primer. Raw sequence reads were trimmed and filtered prior to downstream analyses using RDP pyrosequencing pipeline (http://rdp.cme.msu.edu/). Generally, we removed sequences that (i) did not perfectly match the tag codes and forward primer sequences, (ii) contained ambiguous nucleotide (N's) and (iii) were short (<250 bp). In

Table 1 Sequences of primers used for performance evaluation of biodiversity assessment based on high-throughput sequencing. Fusion primers were used: 454 adaptor A and unique tag were added to 5′-end of each forward primer, and 454 adaptor B was added to 5′-end of each reverse primer. 454 adaptors (A and B) were used to make PCR products compatible with the GS-FLX Titanium sequencing process, while unique tags (i.e. tags consisting of unique eight nucleotides, Parameswaran *et al.* 2007) were used to differentiate PCR products pooled together for pyrosequencing. T_a = annealing temperature for PCR. The IUPAC codes were used for degenerate nucleotides: K = G + T, Y = C + T, R = A + G, D = G + A + T, W = A + T, H = A + C + T

Genetic marker	Primer name	Sequence	T_a (°C)	Reference
Nuclear SSU	Uni18S	5′-AGGGCAAKYCTGGTGCCAGC-3′	50	Zhan *et al.* 2013;
	Uni18SR	5′-GRCGGTATCTRATCGYCTT-3′		
	Uni18S2	5′-CTTAATTTGACTCAACACGG-3′	50	This study
	Uni18SR2	5′-TAGCGACGGGCGGTGTGTAC-3′		
Mitochondrial 16S	Uni16SF	5′-TRACYGTGCDAAGGTAGC-3′	50	This study
	Uni16SR	5′-YTRRTYCAACATCGAGGTC-3′		
Mitochondrial COI	LCO1490	5′-GGTCAACAAATCATAAAGATATTGG-3′	40	Folmer *et al.* 1994;
	HCO2198	5′-TAAACTTCAGGGTGACCAAAAAATCA-3′		
	UniMinibarF1	5′-TCCACTAATCACAARGATATTGGTAC-3′	46 for first 5 cycles, then 53 for 35 cycles	Meusnier *et al.* 2008;
	UniMinibarR1	5′-TGAAAATCATAATGAAGGCATGAGC-3′		
	CrustF1	5′-TTTTCTACAAATCATAAAGACATTGG-3′	42	Costa *et al.* 2007;
	CrustF2	5′-GGTTCTTCTCCACCAACCACAARGAYATHGG-3′		
	CrustDF1	5′-GGTCWACAAAYCATAAAGAYATTGG-3′	45	Radulovici *et al.* 2009
	CrustDR1	5′-TAAACYTCAGGRTGACCRAARAAYCA-3′		
-	454 adaptor A	5′-GCCTCCCTCGCGCCATCAG-3′	-	454 Life Sciences
	454 adaptor B	5′-GCCTTGCCAGCCCGCTCAG-3′		

addition, we detected and then deleted PCR-mediated recombinants (i.e. chimeras) in amplification products from each data set using UCHIME (Edgar *et al.* 2011).

Sequence reads from each sample were clustered into similarity-based operational taxonomic units (OTUs) using the CD-HIT method (Li & Godzik 2006) implemented in the pipeline CLOTU (Kumar *et al.* 2011). Subsequently, OTUs were grouped taxonomically (e.g. order and family) by searching against the nucleotide data base of GenBank. We used BLASTn implemented in the pipeline Seed with the parameters of E value $<10^{-80}$ and minimum query coverage >80%. In addition, because a run of 1/2 plate for each marker yielded different numbers of sequence reads, we used rarefaction analysis to compare the taxon recovery efficiency of the final selected primer pairs with our large-scale pyrosequencing data at a common sequencing depth. Rarefaction analysis was performed at two levels, OTU and order level, using 5000 random iterations in ECOSIM version 7.72 (Gotelli & Entsminger 2006).

Results

Design and selection of primers

After an initial screen among seven taxonomic groups (i.e. Arthropoda, Mollusca, Tunicata, Echinodermata, Annelida, Nematoda and Platyhelminthes), we chose 77 representative sequences for 18S (see sequence alignment in Appendix S2, Supporting information) and 46 sequences for mt16S (see sequence alignment in Appendix S3, Supporting information) for primer design. The polymorphism survey revealed several conserved regions of 18S (Fig. 1a) and mt16S (Fig. 1b) for universal primer design. Based on the sequencing capacity (approximately 500 bp) and identified sequence polymorphisms in the target regions, we designed one primer pair (Uni18S2-Uni18SR2) spanning V5-7 regions for 18S, and one primer pair (Uni16SF-Uni16SR) for 16S (Table 1; Fig. 1). In addition, one universal primer pair (Uni18S-Uni18SR) spanning the V4 region of 18S (Zhan *et al.* 2013) was also evaluated in this study. However, we did not detect any acceptable conserved regions for universal primer design in the alignment of COI sequences (Appendix S1, Supporting information). Hence, we chose five published COI primer pairs (Table 1), including the commonly used universal primers LCO1490-HCO2198 (Folmer *et al.* 1994), a primer pair (UniMinibarF1-UniMinibarR1) for amplification of a short 'minibarcode' fragment (Meusnier *et al.* 2008), and three primer pairs (CrustF1-HCO2198, CrustF2-HCO2198 and CrustDF1-CrustDR1) for Crustacea (Costa *et al.* 2007; Radulovici *et al.* 2009).

Efficiency evaluation of selected primers

All primers were subjected to step-by-step evaluation (Fig. 2). The evaluation for specificity using several

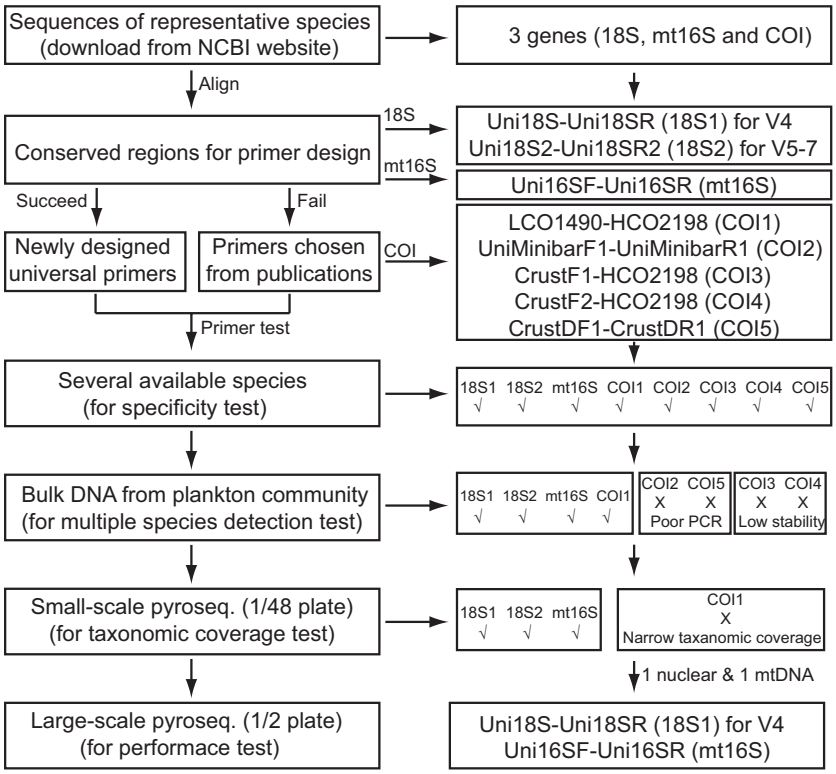

Fig. 2 Methodological flow chart used to perform step-by-step performance evaluation for biodiversity assessment in a complex plankton community collected from Hamilton Harbour, Ontario, using 454 pyrosequencing based on three genetic markers and their corresponding primers.

common species showed that the newly designed primer pairs for 18S and mt16S amplified all species effectively, exhibiting bright and sharp bands on agarose gels. The primers selected from publications for COI were functional, showing weak but detectable bands on agarose gels. Hence, we included all eight primer pairs for the three genetic markers to further assess PCR amplification efficiency using bulk DNA extracted from the zooplankton community. After PCR, two primer pairs, CrustF1-HCO2198 and CrustF2-HCO2198, showed weak or no bands with high background noise on agarose gels. Such poor PCR amplification was not improved after PCR optimization for Mg^{2+} (1.5–3.5 mM) and annealing temperature (40–60 °C). Another two primer pairs, UniMinibarF1-UniMinibarR1 and CrustDF1-CrustDR1, showed inconsistent amplification when repeated using the same DNA template and PCR conditions. Consequently, we cloned the PCR products from four primer pairs: two for 18S (Uni18S-Uni18SR and Uni18S2-Uni18SR2); one for 16S (Uni16SF-Uni16SR); and one for CO1 (LCO1490-HCO2198). After Sanger sequencing, multiple species were recovered using these four primer pairs. Therefore, we included these four primer pairs for small-scale pyrosequencing.

A small-scale pyrosequencing run (i.e. an equivalent of 1/48 PicoTiter plate) yielded approximately 13 000–17 000 sequence reads for each of the four primer pairs. To assess taxonomic coverage, all sequences were grouped taxonomically by order (higher ranks were used when order was not available for a given taxon) using BLASTn implemented in the pipeline Seed (Fig. 3). The number of observed order-level taxa varied greatly among the types of genetic markers, with 19 and 15 based on the two primer pairs for 18S vs. two for COI and seven for mt16S (Fig. 3). We observed a similar pattern for Crustacea: we detected five orders based on 18S, but only one and two orders using COI and mt16S, respectively. In addition, the two primer pairs for 18S recovered a wide range of other taxonomic groups, including animals, plants (algae), fungi, blue-green algae and protists (Fig. 3), despite the primers not being designed to capture that level of biodiversity. However, only a limited number of those groups were recovered using either COI or mt16S. When comparing the two mitochondrial markers, mt16S recovered more groups than COI, including platyhelminthes, rotifers and cyanobacteria. Considering both the number of taxonomic groups recovered (Fig. 3) and the DNA sequence variation that the primers span (Fig. 1), we chose one representative primer pair for each of two types of genetic markers for further performance tests, that is Uni18S-Uni18SR for 18S (nuclear) and Uni16SF-Uni16SR for mt16S (mtDNA).

A total of 686064 and 299045 sequences were obtained for 18S and mt16S, respectively, in runs of 1/2 PicoTiter plate for each marker. After preprocessing to remove

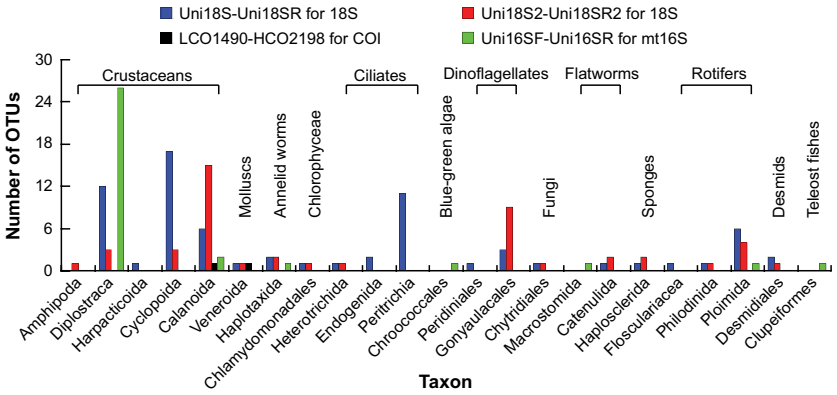

Fig. 3 Order-level taxa recovered from the complex plankton community collected from Hamilton Harbour, Ontario, Canada, using a small-scale run of 454 pyrosequencing (i.e. an equivalent of 1/48 PicoTiter plate) based on three markers and corresponding primer: Uni18S-Uni18SR and Uni18S2-Uni18SR2 for 18S; LCO1490-HCO2198 for COI; and Uni16SF-Uni16SR for mt16S. Operational taxonomic units (OTUs) were grouped at 3% genetic divergence using CD-HIT method.

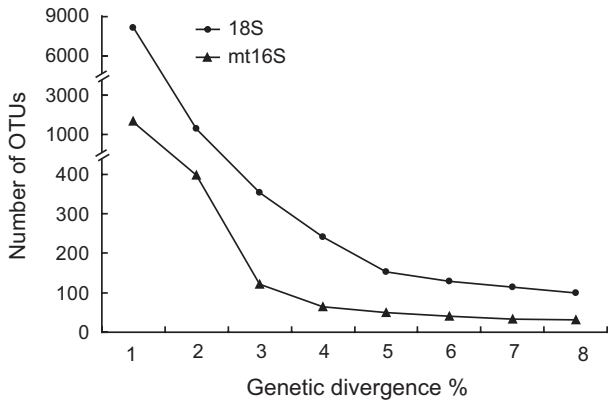

Fig. 4 Number of operational taxonomic units (OTUs) at a range of genetic divergences from 1% to 8% based on 1/2 PicoTiter plate pyrosequencing of the plankton sample collected from Hamilton Harbour, Ontario, Canada, using 18S (primer pair: Uni18S-Uni18SR) and mt16S (primer pair: Uni16SF-Uni16SR).

low-quality sequences, we obtained 138246 and 38718 sequences for 18S and mt16S, respectively, which were subjected to similarity-based OTU clustering using genetic divergences ranging from 1% to 8% (Fig. 4). As expected, the number of OTUs decreased as genetic divergence increased (Fig. 4). The number of observed OTUs based on 18S was larger than that based on mt16S at all genetic divergence levels examined (Fig. 4). After OTUs were assigned to orders (see Appendix S4, Supporting information), we detected a large difference in the number of order-level taxa recovered for these two genetic markers (Table 2). Overall, 38 orders were recovered when using 18S primers, while only 10 orders were detected based on mt16S. Similar to the results obtained from small-scale pyrosequencing, a wide range of taxonomic groups, including numerous animal groups, fungi, algae and protists, was recovered based on 18S, though recovery with mt16S was limited to several animal groups including molluscs, bryozoans and crustaceans (Table 2). Almost all taxonomic groups recovered by mt16S were also recovered by 18S, with two exceptions (Phylactolaemata and Clupeiformes; Table 2). Among all order-level taxa detected by mt16S, more than 78% belong to Diplostraca, though we did not detect such dominance by any one group using 18S. The most abundant taxonomic groups detected by 18S were two crustacean orders Cyclopoida (18.7%) and Diplostraca (18.7%), followed by Calanoida (18.1%) and Peritrichia (13.3%). Relative contributions of sequences by remaining taxa were lower than 6%, ranging from 0.28% to 5.9% (Table 2).

When crustaceans were examined at the family-level, we detected a similar pattern to that observed at the order level. Overall, we detected more families based on 18S than mt16S, nine vs. six families (Table 3). For family detection based on 18S, the most abundant was Cyclopidae (18.7%); however, this family was not detected in the mt16S sequence data (Table 2). The most abundant family identified in the mt16S data was Daphniidae (39.3%), followed by Bosminidae (35.2%).

When both 18S and mt16S data were subjected to rarefaction analysis, the curves did not plateau for either OTU- or order-level analyses, even after 120000 high-quality sequences had been added for 18S (Fig. 5). The curves for 18S showed more OTUs/orders at a common sequencing depth when compared to mt16S. This pattern became more apparent at the order level (Fig. 5b). For example, when sequencing depth was set to 20000 sequences, the number of orders for 18S was double that of mt16S (Fig. 5b).

Discussion

High-throughput sequencing of environmental samples has revolutionized the exploration and quantification of biodiversity, especially in organisms such as microscopic species for which traditional morphological identification is problematic and sometimes impossible (e.g. Fonseca *et al.* 2010; Lodge *et al.* 2012). Despite the allure of such technology for analysing the diversity and

Table 2 Composition of order-level taxa recovered from the complex plankton community collected from Hamilton Harbour, Ontario, using 454 pyrosequencing based on two types of genetic markers, 18S (nuclear) and 16S (mtDNA). The sequencing depth was 1/2 PicoTiter plate for each marker. Operational taxonomic units (OTUs) were grouped at 3% genetic divergence using CD-HIT method. '-' indicates no detection

Group	Order	Percentage of OTUs-18S (%)	Percentage of OTUs-16S (%)
Acari	Astigmata	0.28	—
Annelida (annelid worms)	Haplotaxida	1.42	0.82
Bacillariophyta (diatoms)	Fragilariales	0.28	—
Bryozoa (bryozoans)	Phylactolaemata	—	0.82
	Ctenostomatida	0.28	—
Chlorophyta (green algae)	Mychonastes	0.28	—
	Sphaeropleales	0.28	—
	Chlamydomonadales	1.13	—
Ciliophora (ciliates)	Cyclotrichida	0.28	—
	Cyrtolophosidida	0.28	—
	Heterotrichida	0.28	—
	Hypotrichia	0.28	—
	Oligotrichia	0.28	—
	Stichotrichia	0.28	—
	Endogenida	5.95	—
	Peritrichia	13.31	—
Cnidaria (cnidarians)	Hydroida	0.57	—
Crustacea (crustaceans)	Harpacticoida	0.28	—
	Calanoida	18.13	2.46
	Cyclopoida	18.70	—
	Diplostraca	18.70	78.69
Cryptophyta (cryptomonads)	Cryptomonadales	0.28	—
Cyanobacteria (blue-green algae)	Chroococcales	0.28	3.28
Desmidiales (desmids)	Desmidiales	0.57	—
Dinophyceae (dinoflagellates)	Suessiales	0.28	—
	Gonyaulacales	0.57	—
Fungi (fungi)	Monoblepharidales	0.28	—
	Spizellomycetales	0.85	—
Ichthyosporea	Ichthyophonida	0.28	—
Mollusca (molluscs)	Veneroida	0.28	4.10
Nematoda (roundworms)	Chromadorida	0.28	—
Oomycetes	Saprolegniales	0.28	—
	Myzocytiopsidales	0.28	—
Perkinsea	Perkinsea	0.28	—
Platyhelminthes (flatworms)	Macrostomida	2.27	1.64
Porifera (sponges)	Haplosclerida	0.28	—
Rotifera (rotifers)	Ploimida	5.10	4.92
	Flosculariacea	5.67	—
Zygnematales	Zygnematales	0.57	—
Teleostei (teleost fishes)	Clupeiformes	—	3.28

distribution of species at the community level, successful application to some communities, such as zooplankton, relies largely on the power and efficiency of genetic markers and corresponding primers (Tang *et al.* 2012). In this study, we compared the performance of three commonly used markers, 18S, mt16S and COI, and their corresponding primers for biodiversity assessment based on 454 pyrosequencing of a complex zooplankton community. Our results showed that the power and efficiency varied widely among these markers and their corresponding primers, with the best overall performance obtained with 18S.

The COI gene is the most commonly used DNA barcode marker for identifying and differentiating animal species (Hebert *et al.* 2003). Studies based on traditional Sanger sequencing have confirmed that COI is a better indicator of true diversity than both traditional morphology and other genetic markers, including 18S

Table 3 Comparison of order- and family-level of crustacean taxa recovered from the complex plankton community collected from Hamilton Harbour, Ontario, using 454 pyrosequencing based on two types of genetic markers, 18S (nuclear) and 16S (mtDNA). The sequencing depth was 1/2 PicoTiter plate for each marker. Operational taxonomic units (OTUs) were grouped at 3% genetic divergence using CD-HIT method. Morphological taxonomy data were based on surveys on Hamilton Harbour conducted from 1984–2007 (see Appendix S5, Supporting information, for species list). '-' indicates no detection

Order	Family	18S (%)	16S (%)	Detected by morphological taxonomy?
Diplostraca	Bosminidae	1.42	35.25	Yes
	Daphniidae	13.31	39.34	Yes
	Sididae	3.68	1.64	Yes
	Podonidae	0.28	—	No
	Macrotrichidae	—	0.82	No
	Moinidae	—	0.82	No
Cyclopoida	Cyclopidae	18.69	—	Yes
Calanoida	Diaptomidae	16.72	2.46	Yes
	Calanidae	1.13	—	Yes
	Temoridae	0.28	—	No
Harpacticoida	Ameiridae	0.57	—	No

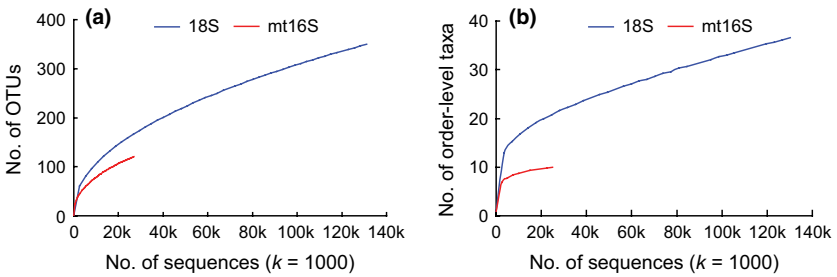

Fig. 5 Rarefaction curves for 18S (primer pair: Uni18S-Uni18SR) and mt16S (primer pair: Uni16SF-Uni16SR) at the operational taxonomic unit (OTU) level (a) and order level (b) based on 1/2 PicoTiter plate pyrosequencing of the plankton sample collected from Hamilton Harbour, Ontario, Canada.

(Tang *et al.* 2012). In addition, global barcoding initiatives have accumulated a large number of reference sequences for annotating metabarcoding data derived from environmental samples. Collectively, the COI gene was expected to be an ideal candidate marker for high-throughput sequencing-based biodiversity assessment. However, performance comparisons in this study revealed that the tested COI primers were affected by three different problems: (i) weak/no bands with high background noise (CrustF1-HCO2198 and CrustF2-HCO2198); (ii) unstable amplification (UniMinibarF1-UniMinibarR1 and CrustDF1-CrustDR1); and (iii) narrow taxonomic coverage (LCO1490-HCO2198) after small-scale pyrosequencing. Similarly, poor PCR amplification success was observed in other microscopic organisms, such as crustaceans (Elías-Gutiérrez *et al.* 2008; Jeffery *et al.* 2011) and nematodes (Bhadury *et al.* 2006). To address poor PCR amplification and narrow taxonomic coverage, new primer pairs were designed based on the taxonomic groups of interest (e.g. Costa *et al.* 2007; Meusnier *et al.* 2008; Radulovici *et al.* 2009). However, our tests indicate that the more specific primer pairs are not useful for broad application using bulk DNA from a complex aquatic community. Several aspects of the COI molecule, such as poorly conserved priming sites associated with high substitution rates, and biased substitution patterns with high AT content, could be responsible for our observed poor PCR amplification and narrow taxonomic coverage (Sanna *et al.* 2009; Creer *et al.* 2010). Poor PCR amplification or lack of primer universality are the major reasons why the COI gene has had limited use for biodiversity surveys in environmental samples using high-throughput sequencing (Sanna *et al.* 2009; Creer *et al.* 2010).

Currently, biodiversity assessment for eukaryote environmental samples relies mainly on the nuclear 18S rDNA gene (Creer *et al.* 2010; Fonseca *et al.* 2010; Bik *et al.* 2012). Our study which includes pyrosequencing results and rarefaction analyses clearly demonstrated that 18S recovered more taxonomic groups than COI and mt16S for both the dominant taxonomic group (i.e. Crustacea) in zooplankton and others (Fig. 3; Tables 2 and 3). However, several studies have suggested that 18S likely underestimates true species

richness (e.g. Tang *et al.* 2012). The use of 18S might be adequate at higher taxonomic scales, but species-level patterns should be interpreted with caution (Tang *et al.* 2012). As demonstrated in this study, 18S and newly designed primers are powerful tools for profiling biodiversity in complex communities, especially in largely unexplored communities such as zooplankton where a large number of taxa remain unknown and some dominant taxa are difficult to amplify using mitochondrial markers such as COI. A possible solution for the low resolution of 18S at the species level is the use of other high-resolution markers, such as the mt16S primers developed here or COI primers designed specifically for taxonomic groups of interest after broad-scale community composition has been comprehensively explored using 18S. Very recently, new COI primer pairs including ZplankF1_t1-ZplankR1_t1 (Prosser *et al.* 2013) and jgHCO2198-jgLCO1490 (Geller *et al.* 2013) were tested using traditional Sanger sequencing. Those primer pairs were not included in this study due to unavailability at the time when this project was performed. After the performance of those primers is fully tested using high-throughput sequencing of bulk DNA isolated from complex communities, it may be possible to determine whether they can be used for whole community-level biodiversity assessments.

Mt16S is also a commonly used genetic marker for the identification of animals (Deagle *et al.* 2009; Mitani *et al.* 2009). Compared to COI, conserved mt16S regions were readily available for universal primer design to cover a relatively wide range of taxa (Fig. 1b). A small-scale run of pyrosequencing also confirmed that 16S recovered more taxa than COI (Fig. 3). However, relative to 18S, mt16S seriously underperformed with respect to the number of taxa detected (Fig. 3; Table 2). This problem became more apparent when a large-scale run of pyrosequencing (i.e. 1/2 PicoTiter plate for each marker) was employed, with 38 vs. 10 orders for all taxonomic groups (Table 2), and four vs. two orders and nine vs. six families for Crustacea (Table 3). In addition, more than 78% of detected orders belonged to a single group (Diplostraca), suggesting possible biased amplification with this primer pair. Biased amplification and/or different degree of universality of primers may therefore be responsible for the difference in biodiversity recovery between 18S and mt16S.

Indeed, biased PCR amplification was observed even when using different primers for the same markers/genes (e.g. Bellemain *et al.* 2010; Pinto & Raskin 2012). Biased PCR amplification can be caused by several factors, such as specificity/universality of the primers, length variation of amplified regions among taxa and taxonomic composition of communities of interest (Huber *et al.* 2009; Bellemain *et al.* 2010; Engelbrektson *et al.* 2010). The first two factors are derived from the nature of markers/genes selected, such as the availability of conserved regions for universal primer design and the existence of large insertions/deletions in amplified regions (e.g. Fig. 1; Appendices S1–S3, Supporting information). For environmental DNA-based or community-based studies, redesign of 'universal' primers based on a competent marker (e.g. 18S in this study) and/or use of multiple sets of primers targeting different regions of selected marker genes (e.g. Uni18S-Uni18SR for V4 region and Uni18S2-Uni18SR2 for V5–V7 regions) may minimize biased amplification (Bellemain *et al.* 2010; Nossa *et al.* 2010; Pinto & Raskin 2012).

Interestingly, all crustacean families with occurrence lower than 1% by pyrosequencing were not detected in surveys conducted between 1984 and 2007 using traditional morphological taxonomy (see Appendix S5, Supporting information, for species list). This disparity might be due to spatial and temporal variation in sample collection, or higher sensitivity of pyrosequencing for detection of rare taxa compared to traditional field collection surveys and identification by microscopy. Our results, as well as others (e.g. Pochon *et al.* 2013), indicate that high-throughput sequencing is a sensitive and effective method for rare taxon detection in mixed-species communities. Indeed, the high sensitivity of 454 pyrosequencing was documented by Zhan *et al.* (2013) who spiked known indicator species into complex plankton communities, and determined a detection limit as low as 2.3×10^{-5}% of sample biomass. Reliable detection of such rare species in nature holds important implications for the conservation of species at risk and for rapid-response programmes targeting the eradication of invading nonindigenous species.

Acknowledgements

This work was supported by the One-Three-Five Program (YSW2013B02) of the Research Center for Eco-Environmental Sciences and 100 Talents Program of the Chinese Academy of Sciences to AZ, by Discovery grants from Natural Sciences and Engineering Research Council of Canada (NSERC) to DDH, SAB and HJM, by the NSERC Canadian Aquatic Invasive Species Network (CAISN), and by an NSERC Discovery Accelerator Supplement to HJM. Great thanks to Dr. Melania Cristescu for providing laboratory space and instrument.

References

Adamowicz SJ, Purvis A (2005) How many branchiopod crustacean species are there? Quantifying the components of underestimation. *Global Ecology and Biogeography*, **14**, 455–468.

Bellemain E, Carlsen T, Brochmann C *et al.* (2010) ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *BMC Microbiology*, **10**, 189.

Bhadury P, Austen MC, Bilton DT *et al.* (2006) Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Marine Ecology Progress Series*, **320**, 1–9.

Bik HM, Porazinska DL, Creer S *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution*, **27**, 233–243.

Briski E, Cristescu ME, Bailey SA *et al.* (2011) Use of DNA barcoding to detect invertebrate invasive species from diapausing eggs. *Biological Invasions*, **13**, 1325–1340.

Cheung MK, Au CH, Chu KH *et al.* (2010) Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. *The ISME Journal*, **4**, 1053–1059.

Costa FO, de Waard JR, Boutillie J *et al.* (2007) Biological identifications through DNA barcodes: the case of the *Crustacea*. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 272–295.

Creer S, Fonseca VG, Porazinska DL *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

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

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Dudgeon D, Arthington AH, Gessner MO *et al.* (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews*, **81**, 163–182.

Edgar RC, Haas BJ, Clemente JC *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Elías-Gutiérrez M, Martínez-Jerónimo F, Ivanova NV *et al.* (2008) DNA barcodes for *Cladocera* and *Copepoda* from Mexico and Guatemala, highlights and new discoveries. *Zootaxa*, **1849**, 1–42.

Engelbrektson A, Kunin V, Wrighton K *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal*, **4**, 642–647.

Folmer O, Black M, Hoeh W *et al.* (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Fonseca VG, Carvalho GR, Sung W *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Galand PE, Casamayor EO, Kirchman DL *et al.* (2009) Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. *The ISME Journal*, **3**, 860–869.

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.

Gotelli NJ, Entsminger GL (2006) EcoSim: null models software for ecology. Version 7. Acquired Intelligence Inc. and Kesey-Bear. Jericho, VT05465. http://garyents-minger.com/ecosim.ht.

Hambler C, Henderson PA, Speight MR (2011) Extinction rates, extinction-prone habitats, and indicator groups in Britain and at larger scales. *Biological Conservation*, **144**, 713–721.

Hebert PDN, Cywinska A, Ball SL *et al.* (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B Biological Sciences*, **270**, 313–322.

Holland RA, Darwall WRT, Smith KG (2012) Conservation priorities for freshwater biodiversity: the key biodiversity area approach refined and tested for continental Africa. *Biological Conservation*, **148**, 167–179.

Huber JA, Morrison HG, Huse SM *et al.* (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology*, **11**, 1292–1302.

Jeffery NW, Eías-Gutiérrez M, Adamowicz SJ (2011) Species diversity and phylogeographical affinities of the *Branchiopoda* (*Crustacea*) of Churchill, Manitoba, Canada. *PLoS ONE*, **6**, e18364.

Kumar S, Carlsen T, Mevik B *et al.* (2011) CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics*, **12**, 182.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lodge DM, Turner CR, Jerde CL *et al.* (2012) Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, **21**, 2555–2558.

Machida RJ, Hashiguchi Y, Nishida M *et al.* (2009) Zooplankton diversity analysis through single-gene sequencing of a community sample. *BMC Genomics*, **10**, 438.

McMahon TA, Halstead NT, Johnson S *et al.* (2012) Fungicide-induced declines of freshwater biodiversity modify ecosystem functions and services. *Ecology Letters*, **15**, 714–722.

Meusnier I, Singer GAC, Landry J *et al.* (2008) A universal DNA minibarcode for biodiversity analysis. *BMC Genomics*, **9**, 214.

Mitani T, Akane A, Tokiyasu T *et al.* (2009) Identification of animal species using the partial sequences in the mitochondrial 16S rRNA gene. *Legal Medicine*, **11**, S449–S450.

Nossa CW, Oberdorf WE, Yang L *et al.* (2010) Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World Journal of Gastroenterology*, **16**, 4135–4144.

Parameswaran P, Jalili R, Tao L *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, **35**, e130.

Pereira HM, Navarro LM, Martins IS (2012) Global biodiversity change: the bad, the good, and the unknown. *Annual Review of Environment and Resources*, **37**, 25–50.

Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE*, **7**, e43093.

Pochon X, Bott NJ, Smith KF, Wood SA (2013) Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests. *PLoS ONE*, **8**, e73935.

Prosser S, Martínez-Arce A, Elías-Gutiérrez M (2013) A new set of primers for COI amplification from freshwater microcrustaceans. *Molecular Ecology Resources*, **13**, 1151–1155.

Radulovici AE, Sainte-Marie B, Dufresn F (2009) DNA barcoding of marine crustaceans from the Estuary and Gulf of St Lawrence: a regional-scale approach. *Molecular Ecology Resources*, **9**, 181–187.

Ricciardi A, Rasmussen JB (1999) Extinction rates of North American freshwater fauna. *Conservation Biology*, **13**, 1220–1222.

Sanna D, Lai T, Francalacci P *et al.* (2009) Population structure of the *Monocelis lineata* (*Proseriata*, *Monocelididae*) species complex assessed by phylogenetic analysis of the mitochondrial Cytochrome *c* Oxidase subunit I (COI) gene. *Genetics and Molecular Biology*, **32**, 864–867.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Tang CQ, Leasi F, Obertegger U *et al.* (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 16208–16212.

Telesh IV (2004) Plankton of the Baltic estuarine ecosystems with emphasis on Neva Estuary: a review of present knowledge and research perspectives. *Marine Pollution Bulletin*, **49**, 206–219.

Thomsen PF, Kielgast J, Iversen LL *et al.* (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Uusitalo L, Fleming-Lehtinen V, Hällfors H *et al.* (2013) A novel approach for estimating phytoplankton biodiversity. *ICES Journal of Marine Science*, **70**, 408–417.

Vĕtrovský T, Baldrian P (2013) Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the

introduction of the pipeline SEED. *Biology and Fertility of Soils*, **49**, 1027–1037.

Zhan A, Hulák M, Sylvester F *et al.* (2013) High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods in Ecology and Evolution*, **4**, 558–565.

A.Z., D.D.H. and H.J.M. conceived and designed the project. S.A.B. collected traditional morphological taxonomy data from Hamilton Harbour. A.Z. did laboratory work, analysed the data and led manuscript writing. All authors contributed to revision of the manuscript.

Data Accessibility

454 pyrosequencing data for the three markers (i.e. 18S, mt16S, COI): NCBI SRA: SRR1171114 for 18S, SRR1171115 for mt16S and SRR1171155 for COI. Representative sequences for primer design for COI, 18S and mt16S uploaded as online supplemental material (Appendices S1–S3, Supporting information). Operational taxonomy units (OTUs) derived from 18S and mt16S uploaded as online supplemental material (Appendix S4, Supporting information).

Supporting Information

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

Appendix S1 Alignment (in FASTA format) of selected representative sequences (519 sequences) of Crustacea for seeking conserved regions to design primers for mitochondrial cytochrome *c* oxidase subunit I (COI).

Appendix S2 Alignment (in FASTA format) of finally selected representative sequences for primer design across all taxonomic groups (Arthropoda, Mollusca, Tunicata, Echinodermata, Annelida, Nematoda, Platyhelminthes) for small subunit ribosomal DNA (18S).

Appendix S3 Alignment (in FASTA format) of finally selected representative sequences for primer design across all taxonomic groups (Arthropoda, Mollusca, Tunicata, Echinodermata, Annelida) for mitochondrial 16S ribosomal RNA (mt16S).

Appendix S4 BLAST results for pyrosequencing data based on small subunit ribosomal DNA (18S) and mitochondrial 16S ribosomal RNA (mt16S) using 1/2 PicoTiter plate for each marker.

Appendix S5 Crustacean species identified by traditional morphological taxonomy based on surveys on Hamilton Harbour conducted from 1984 to 2007.