Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 881–890 doi: 10.1111/2041-210X.12230

Reproducibility of pyrosequencing data for biodiversity assessment in complex communities

Aibin Zhan[1,2]*, Song He[1]†, Emily A. Brown[3], Frédéric J.J. Chain[3], Thomas W. Therriault[4], Cathryn L. Abbott[4], Daniel D. Heath[2], Melania E. Cristescu[2,3] and Hugh J. MacIsaac[2]

[1]*Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing 100085, China;* [2]*Great Lakes Institute for Environmental Research, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada;* [3]*Department of Biology, McGill University, 1205 Docteur Penfield, Montreal, QC H3A 1B1, Canada; and* [4]*Pacific Biological Station, Fisheries and Oceans Canada, 3190 Hammond Bay Road, Nanaimo, BC V9T 6N7, Canada*

Summary

1. High-throughput sequencing is rapidly becoming a popular method to profile complex communities and has generated deep insights into community biodiversity. However, the reproducibility of this method for biodiversity assessment remains largely unexplored.

2. Here we evaluated reproducibility by analysing 454 pyrosequenced biological replicates of two complex plankton communities collected from one freshwater port and one marine port. We also tested whether reproducibility potentially influences biodiversity estimates, notably α- and β-diversity.

3. Our evaluation of reproducibility revealed a complex scenario, having both technical and biological significance. At the Operational Taxonomic Unit (OTU) level, reproducibility was 100% for high-abundance OTUs (> 100 sequences), although it was lower for low-abundance OTUs, and sometimes < 25% for singletons. BLAST searches showed that > 88% of irreproducible OTUs had high sequence similarity to existing records, suggesting that some singletons may reflect rare lineages/genotypes in communities. However, spurious amplification of distantly related taxonomic groups generated mainly low-abundance OTUs that were characterized by low reproducibility. At a broad taxonomic level (i.e. order level), reproducibility decreased as the abundance of OTUs decreased and was particularly low for distantly related taxonomic groups such as algae and protists that were not the targets of our zooplankton biodiversity survey. At a lower taxonomical level (i.e. family-level), overall reproducibility was high (> 80%) for crustaceans, the dominant group in zooplankton samples. Therefore, we suggest that random variation during both sample collection and sequencing processes can be responsible for low reproducibility. Our analyses also suggest that random sampling processes may influence both α- and β-diversity estimates.

4. Our results add to growing evidence that caution needs to be applied when designing and interpreting experiments utilizing high-throughput sequencing data for biodiversity assessments. Technical replicates are needed to statistically correct intra-sample variation, while field-based replicate samples are desirable to substantiate results. An overestimation of species diversity can occur when OTUs are uniquely characterized by spuriously amplified sequences and errors/artifacts. Therefore, careful management of low-abundance OTUs is required to reveal unique/rare lineages. Our results suggest that further studies are needed to determine the ecological significance of low-abundance OTUs in complex communities.

Key-words: α diversity, β diversity, community ecology, 454 pyrosequencing, nuclear small subunit (nSSU) rDNA (18S), plankton

Introduction

The advent of high-throughput sequencing (HTS) such as 454 pyrosequencing has thoroughly revolutionized scientific strategies and approaches in medical and biological sciences, resulting in an enormous growth of studies in a variety of disciplines (e.g. Schuster 2008; Creer 2010). Indeed, many studies largely rely on HTS to genotype DNA polymorphisms (mostly single nucleotide polymorphisms, i.e. SNPs; Gruber, Colligan & Wolford 2002), assess biodiversity in communities (Fonseca *et al.* 2010; Yu *et al.* 2012; Ji *et al.* 2013), and evaluate gene expression (i.e. RNA-Seq) at the whole genome level (Wang, Gerstein & Snyder 2009). Among these applications, biodiversity assessment of ecological communities using HTS represents one of its most popular uses. For example, studies of soil

*Correspondence author. Email: zhanaibin@hotmail.com or azhan@rcees.ac.cn

†Present address: Red Sea Research Center, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

microbial communities suggest extant biodiversity orders of magnitude higher than previously recognized (Rousk *et al.* 2010), while similar results were obtained for communities in extreme environments such as deep within the Earth and in polar regions (Chivian *et al.* 2008; Schütte *et al.* 2010). More recently, the approach was used to detect rare species in plankton communities (Hajibabaei *et al.* 2011; Zhan *et al.* 2013). HTS-based studies may also provide high resolution of temporal and spatial biodiversity dynamics of communities (Creer 2010; Pommier *et al.* 2010).

While HTS is becoming a promising approach for biodiversity assessments, several technical problems can impact biodiversity estimates, including nucleotide base calling errors, poor alignment for large datasets and biased amplification of taxa when targeting broad taxonomic groups (Gihring, Green & Schadt 2012). These problems contribute to the potential for overestimating biodiversity (Quince, Curtis & Sloan 2008; Gomez-Alvarez, Teal & Schmidt 2009; Kunin *et al.* 2010). When environmental DNA derived from complex communities is subjected to HTS, each unique sequence read is interpreted as an identifier of a community member. Consequently, intrinsic sequencing errors/artifacts can inflate biodiversity estimates, especially when using deep sequencing of individual genes such as nuclear small subunit ribosomal DNA (nSSU rDNA) (Quince, Curtis & Sloan 2008; Kunin *et al.* 2010). Despite the awareness of such overestimation, reproducibility of HTS data for biodiversity assessment remains largely unknown, especially for complex communities. Reproducibility represents a critical technical aspect for HTS-based biodiversity assessment not only because it is an important indicator for data stability and reliability, but also because it can affect biodiversity comparisons among different communities in space and/or time (Prosser 2010; Zhou *et al.* 2011).

To make technical issues easily understandable, we clarify terms used in the study as follows: (i) reproducibility refers to the capacity of an entire pyrosequencing dataset to be thoroughly replicated when using exactly the same protocol throughout the whole experiment; (ii) random sampling means that an individual is collected from a defined group through unpredictable and random means, that is, all individuals have an equal chance of being chosen of a community during biological sample collection, and sequences have an equal chance of being selected/generated during pyrosequencing processes including library preparation, PCR, sequencing and other procedures; (iii) biological replicates are biological bulk samples (i.e. plankton samples in this study) collected from the same sampling site but treated separately in the experiment; (iv) technical replicates refer to multiple DNA extractions from the same bulk sample for downstream analyses.

In this study, we examined the reproducibility of 454 pyrosequencing data by analysing two complex plankton communities from two harbours, one freshwater (Hamilton, on Lake Ontario) and one marine (Nanaimo, on the Pacific coast of Canada). We set up two parallel 454 pyrosequencing fractions for each community represented by 1/2 PicoTiter plate for each fraction. By examining the two fractions for each community, we sought to determine the reproducibility of 454 pyrosequencing data at an Operational Taxonomic Unit (OTU)-level and at a higher traditional taxonomic level (i.e. order level). We further assessed reproducibility at a moderate taxonomic level (i.e. family-level) for crustaceans, one of our major groups of interest. In addition, we assessed whether reproducibility influences α- and β-diversity estimates.

Materials and methods

FIELD SAMPLING

Plankton samples were collected from Hamilton Harbour, Ontario and Nanaimo Harbour, British Columbia in September and July 2011, respectively. We used geo-referenced oblique plankton tows with both a small mesh (80 μm) and larger mesh (250 μm) plankton net to sample from the bottom to the water surface to collect plankton samples (Fig. 1). The collected plankton samples were preserved in 100% ethanol and stored in −20°C freezer prior to further analyses.

DNA EXTRACTION, PCR AND PYROSEQUENCING

To test reproducibility, two parallel fractions representing biological replicates were set up for each community (Fig. 1). Briefly, preserved

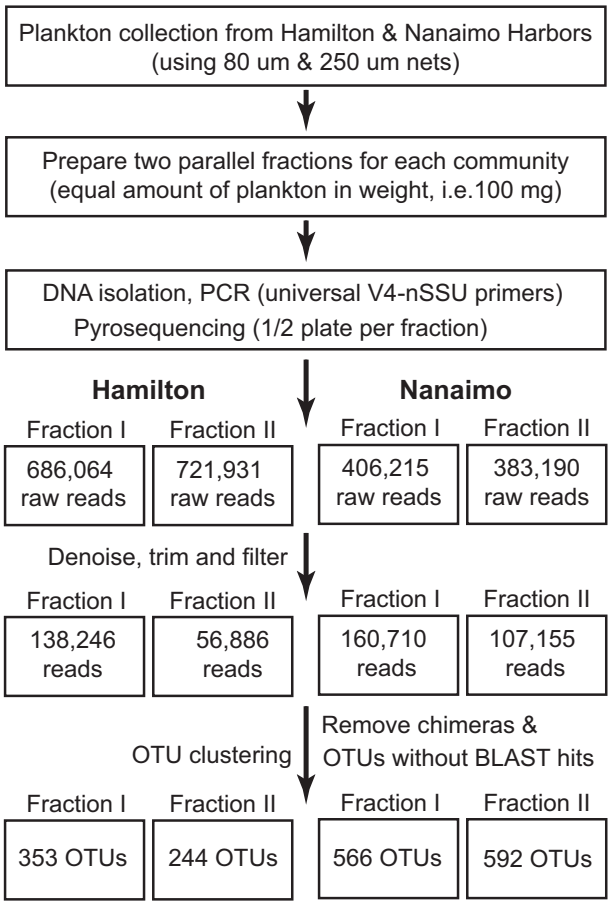

Fig. 1. Flow chart for setting up parallel fractions for the two plankton communities derived from the two harbours, Hamilton (freshwater) and Nanaimo (marine). Sequence reads were grouped into OTUs at a commonly used similarity cut-off value of 97%.

plankton samples were vigorously shaken and transferred into eppendorf tubes to prepare two parallel fractions for each harbour. Tubes containing preserved plankton were centrifuged at 13523 g for 3 min to remove ethanol, and then opened in a fume hood for 10–15 min to evaporate residual ethanol. For each fraction, the total genomic DNA was separately extracted from an equal amount of plankton sample (100 mg, weighed by a balance) using the DNeasy Blood and Tissue Kit (Qiagen Inc., Toronto, ON, Canada). The quality and quantity of these four DNA samples were measured by a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). PCRs were performed using the primer pair Uni18S–Uni18SR spanning the hypervariable V4 region of nSSU rDNA (Zhan *et al.* 2013, 2014a). This primer pair, which amplifies a wide range of eukaryotes with an amplicon size between 400–600 bp depending on taxa, was designed for pyrosequencing zooplankton species (Zhan *et al.* 2013, 2014a). Each 25 μL PCR cocktail contained 100 ng of genomic DNA, 1× PCR buffer, 2 mM of Mg^{2+}, 0·2 mM of dNTPs, 0·4 μM of each fusion primer (Zhan *et al.* 2014a) and 2U of *Taq* DNA polymerase (Genscript). PCR cycling parameters consisted of an initial denaturation step at 95°C for 5 min, followed by 25 cycles of 95°C for 30 s, 50°C for 30 s, 72°C for 90 s, and a final elongation step at 72°C for 10 min. To increase the PCR amplification efficiency, eight replicate PCRs were conducted for each of the two separate fractions from both port samples. PCR products from these eight replicates within each fraction were pooled and then purified using the solid-phase reversible immobilization (SPRI) paramagnetic bead-based method (Agencourt Bioscience Corporation, Beverly, MA, USA). Pyrosequencing was performed 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. We performed 1/2 PicoTiter plate for each fraction for the two communities.

DATA ANALYSIS

After pyrosequencing, raw reads were denoised by Mothur version 1.31.2 (Schloss *et al.* 2009) using default settings implemented in the pipeline Seed version 1.1.35 (Větrovský & Baldrian 2013). Subsequently, we used the RDP pyrosequencing pipeline (http://rdp.cme.msu.edu/) to remove low-quality sequences that: (i) contained any mismatch for the forward primer; (ii) contained any undetermined nucleotide (N's); (iii) were too short (i.e. <250 bp); (iv) contained homopolymers greater than eight; or (v) had Phred scores (Q) lower than 20. We used only the first 400 bp after the forward PCR primer of each sequence for further analyses, mainly because the quality of sequences degrades sharply beyond this point (quality figures not shown). The end trimming process was performed using the pipeline Seed.

The processed sequences from the two fractions of each community were clustered into similarity-based OTUs at a commonly used similarity cut-off value (97%; Kunin *et al.* 2010) using the CD-HIT method (Li & Godzik 2006) implemented in the pipeline CLOTU (Kumar *et al.* 2011). The CD-HIT method is based on a heuristic search strategy, whereby the longest sequence is chosen as a representative sequence for each OTU after a heuristic search (Li & Godzik 2006). Given that PCR-mediated recombination in PCR amplification products (i.e. chimeras) is one of the major error/artifact sources for pyrosequencing, we identified and then deleted chimeras from each data set using the algorithm UCHIME (Edgar *et al.* 2011).

To assess reproducibility at a high (i.e. order level) and moderate (i.e. family-level) taxonomic level, OTUs were grouped taxonomically. The reproducibility at the family-level was conducted for crustaceans, one of the major targeted groups represented 50–80% of total OTUs. Reproducibility at lower taxonomic levels was not assessed to avoid any possible misinterpretation due to possible low resolution of nSSU rDNA (Tang *et al.* 2012). The taxonomic assignment was conducted by searching against the nucleotide database of GenBank using BLASTn implemented in the pipeline Seed with the parameters of E value < 10^{-80} and minimum query coverage >80%. A few low-abundance OTUs without significant BLAST hits (60 and 14 for Hamilton and Nanaimo, respectively) were excluded from subsequent analyses (Fig. 1).

To investigate the relationship between the recovered OTUs represented by both high- and low-abundance sequences, multiple sequence alignments were performed using MAFFT version 7.147b (Katoh & Standley 2013), and approximately-maximum-likelihood phylogenies were reconstructed using FastTree version 2.1.7 (Price, Dehal & Arkin 2009).

Because the two fractions from each of the two harbours in this study yielded different numbers of sequence reads (Fig. 1), and nonparametric estimates (e.g. Chao1 and abundance-based coverage estimator, ACE) and parametric estimates (e.g. Shannon and Simpson's indices) are sensitive to sample size and/or the number of rare OTUs (Gihring, Green & Schadt 2012), we used rarefaction analysis to estimate diversity richness at a common sequencing depth. Rarefaction analysis is a straightforward comparison of diversity richness and unbiased by sample size (Gihring, Green & Schadt 2012). Individual-based species rarefaction analyses were performed for each fraction of both communities using 5000 random iterations in Ecosim version 7.72 (Gotelli & Entsminger 2006). In addition, two popular metrics, Sørensen's incidence (S_s)-based and Bray-Curtis's abundance (BC_s)-based methods were calculated using EstimateS version 9.1.0 (http://viceroy.eeb.uconn.edu/estimates/). Their complements, S_d ($S_d = 1 - S_s$) and BC_d ($BC_d = 1 - BC_s$), are widely used to assess dissimilarity between communities (i.e. β diversity; Sørensen 1948; Bray & Curtis 1957). The range of these two indices is from 0 (when all OTUs are shared between two communities) to 1 (when no OTUs are shared).

Results

PYROSEQUENCING AND OPERATIONAL TAXONOMIC UNITS (OTUS) GROUPING

A total of 686 064 (NCBI SRA accession: SRR1171114) and 721 931 sequences (NCBI SRA accession: SRR1177666) were obtained for the two fractions for Hamilton, while 406,215 (NCBI SRA accession: SRR1177768) and 383 190 sequences (NCBI SRA accession: SRR1177769) were obtained for Nanaimo (Fig. 1). After pre-processing to remove low-quality sequences, the percentage of remaining sequences for downstream analyses varied widely, ranging from 7·9% for fraction II in Hamilton to 39·6% for fraction I in Nanaimo (Fig. 1). A large number of OTUs were recovered after clustering and chimera removal (Fig. 1). Differences in the number of OTUs detected between parallel fractions were substantial: 353 and 244 OTUs for Hamilton, and 566 and 592 OTUs for Nanaimo. These differences cannot be attributed to variation in the number of pre-processed sequences because the Hamilton fraction with more sequences resulted in a larger number of OTUs, but the opposite occurred for the Nanaimo fractions (Table 1; Fig. 1). Among OTUs, singletons (i.e. OTUs represented by a

Table 1. Reproducibility of Operational Taxonomic Units (OTUs) and order-level taxa across the two parallel fractions of the plankton communities derived from the two harbours, Hamilton (freshwater) and Nanaimo (marine), with percentages in brackets. S_d and BC_d = Sørensen's incidence- and Bray-Curtis's abundance-based dissimilarity indices

	Hamilton			Nanaimo		
	Fraction I	Fraction II	All	Fraction I	Fraction II	All
No. of OTUs detected	353	244	427	566	592	849
No. of OTUs recovered by singletons	187 (53·0%)	116 (47·5%)	243 (56·9%)	313 (55·3%)	313 (52·9%)	514 (60·5%)
No. of taxa detected	38	29	42	61	47	68
No. of taxa recovered by singletons	29 (76·3%)	29 (100%)	38 (90·5%)	47 (77·0%)	38 (80·9%)	56 (82·4%)
No. of taxa recovered only by singletons	18 (47·4%)	12 (41·4%)	23 (54·8%)	21 (34·4%)	17 (36·2%)	25 (36·8%)
S_d between two replicates (All data/singleton removed/single-, double-, triple-tons removed)	0·431/0·252/0·103		–	0·467/0·260/0·119		–
BC_d between two replicates (All data/singleton removed/single-, double-, triple-tons removed)	0·459/0·458/0·457		–	0·269/0·267/0·262		–

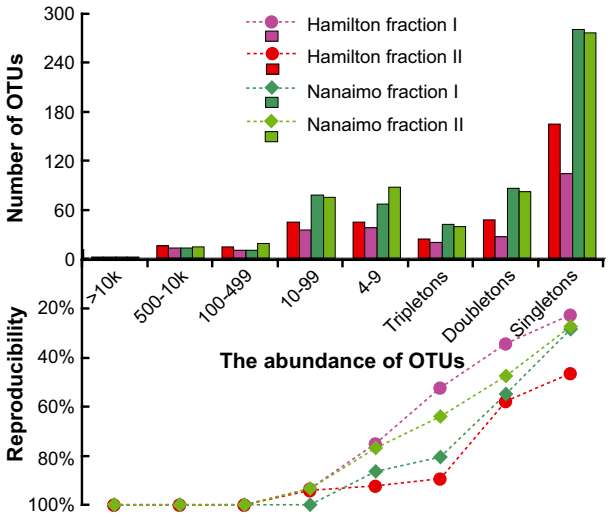

Fig. 2. Reproducibility of Operational Taxonomic Units (OTUs) for the two parallel fractions when they served as references for each other (line chart), and the number of OTUs grouped by the abundance of OTUs (bar chart) for the two plankton communities derived from the two harbours, Hamilton (freshwater) and Nanaimo (marine).

single sequence) were the most abundant, accounting for approximately 50% of OTUs in all four samples (Table 1; Fig. 2).

Significant BLAST hits to existing records in GenBank were obtained for the majority of OTUs (88% in Hamilton and 98% in Nanaimo), including for singletons (81% in Hamilton and 98% in Nanaimo; Table 2). OTUs without significant BLAST hits were mainly present in low abundance (84% were singletons), and they were mostly irreproducible: only 7% and 0% of singletons without BLAST hits were reproducible in Hamilton and Nanaimo, respectively. These OTUs, which were removed for subsequent analyses, exhibited very high sequence divergence from the major taxa identified, as inferred by long internal and external phylogenetic branches (Figs S1 & S2). Phylogenetic exploration of the data also revealed that most clades contained multiple OTUs with consistent BLAST results to one genus or species, and that many low-abundance OTUs including singletons formed closely related clades with high-abundance OTUs (Figs S1 & S2). This suggests that many of the low-abundance OTUs reported herein may represent intraspecific biodiversity in the form of rare genotypes rather than distinct lineages/species. These OTUs may remain informative in this context.

REPRODUCIBILITY AT OTU-LEVEL

For Hamilton, a total of 427 OTUs were recovered when data from the two fractions were considered together, of which only 170 (39·8%) were shared between the two fractions. For Nanaimo, a larger number of OTUs (849) were detected, but a smaller proportion (36·4%, 309 OTUs) were shared between the two fractions (Table 1; Figs 2 & 3). When singletons were removed from the analysis, the percentages of shared OTUs increased to 59·8% and 58·8% for Hamilton and Nanaimo, respectively (Fig. 3). OTU sharing increased to approximately 80% when singletons, doubletons and tripletons were all removed (Fig. 3).

Overall OTU-level reproducibility was low: 48·2% and 69·7% between the two fractions for Hamilton, and 54·6% and 52·2% for Nanaimo (Fig. 2). The fraction with smaller numbers of sequences in Hamilton had a higher reproducibility, but the opposite was detected in Nanaimo (Figs 1 & 2). In general, reproducibility decreased as the abundance of OTUs decreased (Fig. 2). For both harbours, OTUs represented by more than 100 sequences were 100% reproducible (Fig. 2). For OTUs represented by more than three sequences, reproducibility was relatively high, ranging from >75% to >92% across different samples (Fig. 2). For tripletons, reproducibility decreased, ranging from >52% to >89%. Reproducibility further declined for doubletons and singletons and was lowest for singletons (22·6%) in fraction I in Hamilton (Fig. 2).

REPRODUCIBILITY AT ORDER LEVEL

After OTUs were assigned to order-level taxa, we detected a wide range of taxa in both harbours, including many animal

Table 2. Representative taxa recovered only by singletons in one of the two fractions (i.e. irreproducible) in the two plankton communities derived from Hamilton (freshwater) and Nanaimo (marine). Representative sequence ID for Operational Taxonomic Units (OTUs) and detailed information from BLASTn searches are shown. Max identity shows per cent similarity between the query and subject sequences over the length of the coverage area

			BLAST result			
Taxonomic group	Representative sequence ID	Recovered by fraction (I or II)	Best hit	Coverage (%)	Expect (*E*) value	Max identity (%)
Hamilton						
Fragilariales	HAVE4QV01D1U0Y	I	*Fragilaria crotonensis*	96	0	97
Choanoflagellida	HKQYOUX01DI57S	II	*Sphaeroeca volvox*	90	1·55E-141	90
Cyrtolophosidida	HAVE4QV01BT3BV	I	*Cyrtolophosis minor*	96	2·14E-171	96
Cryptomonadales	HAVE4QV01BSTKR	I	*Cryptomonas curvata*	99	0	99
Chroococcales	HKQYOUX01BIGKY	II	*Microcystis aeruginosa*	99	0	99
Hypotrichia	HAVE4QV01DFLCJ	I	*Parabistichella variabilis*	93	1·27E-161	93
Oligotrichia	HAVE4QV01A0P2F	I	*Strombidium* sp.	95	5·75E-166	95
Phylactolaemata	HKQYOUX01EGTU6	II	*Plumatella* sp.	100	0	100
Sphaeropleales	HAVE4QV01DUVE3	I	*Coelastrum microporum*	99	0	99
Parachela	HKQYOUX01A9864	I	*Murrayon pullari*	100	0	100
Zygnematales	HKQYOUX01DUCSA	II	*Mougeotia* sp.	99	0	99
	HKQYOUX01CT928	II	*Spirogyra* sp.	99	0	99
Nanaimo						
Acari	HMWF7SE02JR4IE	I	*Caeculidae* sp.	93	2·98E-163	94
Gymnosomata	HMWF7SE02G2YFB	I	*Pneumoderma atlantica*	98	0	98
Actiniaria	HOFZNNK02JL8NZ	II	*Edwardsiella lineata*	95	1·35E-167	95
Gymnodiniales	HMWF7SE02J16DV	I	*Gyrodinium fusiforme*	99	0	99
	HMWF7SE02G8W0O	I	*Karenia papilionacea*	97	4·96E-120	97
Kentrogonida	HMWF7SE02HYNL6	I	*Loxothylacus panopaei*	100	0	100
Leptothecatae	HMWF7SE02I8GIX	I	*Phialella quadrata*	100	0	100
Littorinimorpha	HMWF7SE02IEZ5J	I	*Lacuna pallidula*	99	1·70E-139	99
Monostilifera	HMWF7SE02J2MPI	I	*Zygonemertes virescens*	100	0	100
Nassellaria	HMWF7SE02I0JFW	I	*Cladoscenium tricolpium*	96	1·54E-179	96
Hymenoptera	HMWF7SE02FS7SZ	I	*Anagrus epos*	99	0	99
Prorodontida	HMWF7SE02HF8VS	I	*Urotricha* sp.	95	6·35E-159	95
Unclassed Dinophyceae	HOFZNNK02IS2FN	II	*Stoeckeria* sp.	99	0	99

groups, some plants (algae), and a few fungi, protists and one prokaryote (Cyanobacteria) (Tables 1 & S1). We detected 38 and 29 order-level taxa for the two fractions in Hamilton (a combined total of 42 taxa), and 61 and 47 taxa for the two fractions in Nanaimo (a combined total of 68 taxa; Tables 1 & S1). For Hamilton, the most abundant taxon was Crustacea, followed by Ciliophora and Rotifera (Table S1). Crustacea also dominated the plankton community in Nanaimo, followed by Cnidaria and Algae (Table S1).

Similar to the OTU-level results, both harbours exhibited a comparable percentage of shared taxa between the two fractions at the order level: 59·5% and 58·8% for Hamilton and Nanaimo, respectively (Fig. 3). When singletons were removed, the percentage of shared taxa increased more in Hamilton than in Nanaimo (89·5% vs. 65·0%). When OTUs represented by very low-abundance sequences (singletons, doubletons and tripletons) were removed, the percentages of shared taxa between parallel fractions remained similar (87·5%) in Hamilton but were much increased (95·8%) in Nanaimo (Fig. 3).

Using the singletons that could be assigned to order-level taxa, we found that more than 75% of taxa from our full dataset were recovered, and between 30–50% of taxa from our full dataset were solely represented by singletons (Tables 1 & S1). When these taxa were subjected to reproducibility analysis, only six and seven taxa were reproducible between the two fractions for Hamilton and Nanaimo, respectively (Table S1). A large number of irreproducible taxa were derived from singletons (17 in Hamilton and 18 in Nanaimo), accounting for 40·5% and 26·5% of all taxa obtained in these two harbours (Tables 1 & S1). Moreover, many irreproducible taxa derived from singletons represent taxonomically divergent groups that were not directly targeted in our biodiversity survey (Table 2).

REPRODUCIBILITY OF CRUSTACEANS AT FAMILY-LEVEL

Crustacea was the dominant taxon in our samples and was well represented by both high- and low-abundance OTUs (Table S1). We found that reproducibility of OTUs in this subset of the data was very similar to that of the full dataset, but family-level taxa reproducibility was comparatively higher (87·5% for Hamilton and 82% for Nanaimo; Tables S2 & S3). The irreproducible family-level taxa were characterized by low-abundance OTUs: 80% were singletons, and the rest had at most four sequences. Compared with order-level analyses, singletons contributed less to the number of family-level taxa detected. Less than 15% of families were solely recovered by singletons (Tables S2 & S3).

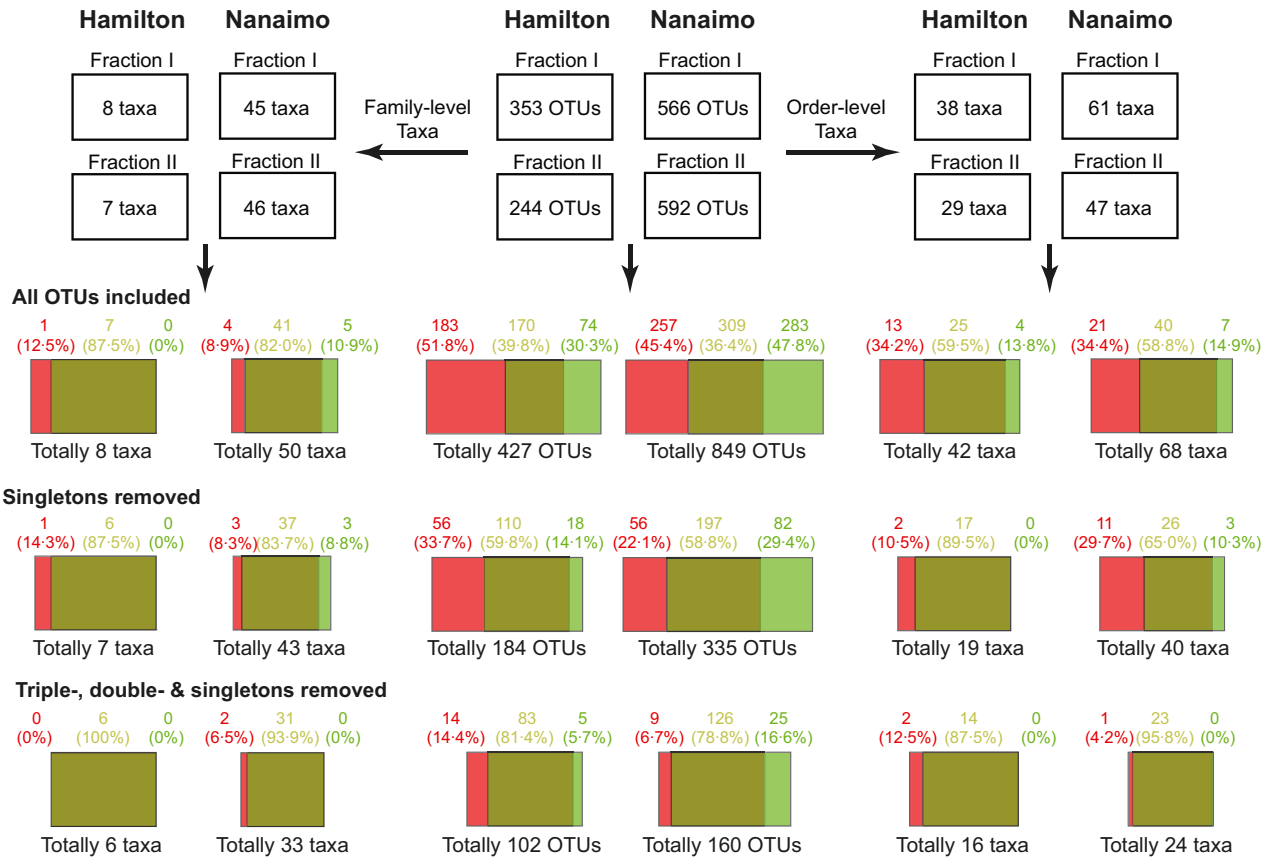

Fig. 3. Reproducibility of Operational Taxonomic Units (OTUs), order-level taxa and family-level taxa (only crustaceans of interest were considered) for the two plankton communities derived from the two harbours, Hamilton (freshwater) and Nanaimo (marine). Reproducibility is shown for all data combined, data with singletons excluded, and data with singletons, doubletons and tripletons excluded. The percentages for unique OTUs/taxa for each fraction were calculated using the number of unique OTUs/taxa divided by the total number of OTUs/taxa in a given fraction, while the percentages for shared OTUs/taxa were calculated using the number of shared OTUs/taxa divided by the combined total number of OTUs/taxa from two fractions. Red, dark green and light green indicate OTUs/taxa recovered only in fraction I, shared between two fractions, and only in fraction II, respectively.

INFLUENCE ON BIODIVERSITY ESTIMATES

When all data were subjected to rarefaction analysis, we found that the curves did not plateau for either fraction in either harbour, even after 120 000 high-quality sequences had been added (Fig. 4). The curve for fraction I from Nanaimo showed approximately 30% more OTUs at a common sequencing depth when compared with fraction II. A similar difference between two fractions was observed when low-abundance OTUs were removed, even though the curves did reach saturation (Fig. 4). The difference between the two fractions in Hamilton was not as great as that in Nanaimo, although a slight difference was still observed for all three datasets (i.e. all data included, singletons removed, and singletons, doubletons and tripletons excluded; Fig. 4). Sørensen and Bray-Curtis indices suggested a relatively high level of dissimilarity, with values of 0·431 and 0·467 for S_d, and 0·459 and 0·269 for BC_d for Hamilton and Nanaimo, respectively. In general, removal of low-abundance OTUs resulted in decreased values for both methods (i.e. assemblages became more similar; Table 1), but it had less effect on Bray-Curtis index than on Sørensen index (Table 1).

Discussion

Metagenomic technologies such as large-scale HTS have created tools that have been used to explore complex communities at an unprecedented depth, identifying orders of magnitude more biodiversity than was previously recognized (Creer 2010; Fonseca *et al.* 2010). Consequently, HTS has become a popular method for assessing community composition and structure. However, to date, reproducibility of HTS-based biodiversity assessment has not been well established. In this study, we assessed this important technical question using large-scale 454 pyrosequencing data from two parallel biological replicates derived from two communities. We further discussed the potential implications with respect to interpreting findings, especially biodiversity measures.

Overall, the reproducibility of OTUs was surprisingly low: 39·8% and 36·4% for Hamilton and Nanaimo, respectively (Table 1; Fig. 3). Low reproducibility has also been observed in other complex communities including soil microbes (Zhou *et al.* 2011), where tagged primers were used to generate replicates. However, the reproducibility percentages obtained in this study are higher than those (13·1% ± 1·5%) reported by

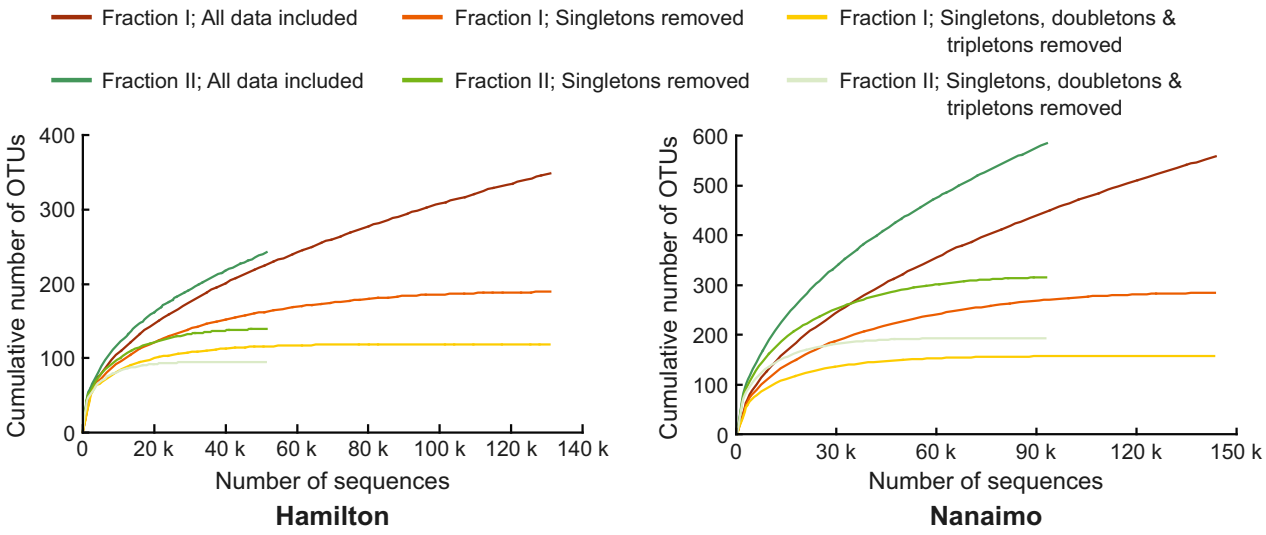

Fig. 4. Rarefaction curves for the two parallel fractions of the plankton communities from Hamilton (freshwater) and Nanaimo (marine).

Zhou *et al.* (2011) and are particularly good when focusing on the most representative taxonomic group (Crustacea) using a family-level resolution (>80%). Our higher reproducibility may be due to several factors including primers with high efficiency and deeper sequencing, as we used 1/2 PicoTiter plate for each replicate versus one plate for 24 samples in Zhou *et al.* (2011). Moreover, our communities are likely less complex than microbial soil communities (1121 ± 390 OTUs for soil communities vs. 427 OTUs for Hamilton and 849 OTUs for Nanaimo when using the same similarity-based clustering threshold, i.e. 3%, and an overlap of gene used, i.e. V4-nSSU rDNA).

When we assigned OTUs taxonomically at the order level, low reproducibility was recovered for taxa inferred from low-abundance OTUs such as singletons (Tables 1 & S1). While several studies have suggested that low-abundance OTUs (such as singletons, doubletons and tripletons) can be artifactual and should be removed from further analyses (Kunin *et al.* 2010; Tedersoo *et al.* 2010), others have clearly shown that some singletons can reflect rare species in communities (Kauserud *et al.* 2012; Zhan *et al.* 2013). For example, our earlier study showed that indicator species spiked at exceptionally low levels (as low as $2{\cdot}3 \times 10^{-5}$% biomass) into plankton communities were recovered as singletons (Zhan *et al.* 2013). Although low-abundance taxa in this study may be represented by singletons, sampling and sequencing artifacts may also contribute noise; the subset of OTUs without BLAST hits were highly diverged from our target sequences (Figs S1 & S2), and many irreproducible order-level taxa represented by singletons were from taxonomically divergent groups which were not direct targets in our biodiversity survey. However, the family-level analysis of crustaceans showed that low-abundance OTUs had a far reduced impact on taxa identification in which less than 15% of families were solely recovered by singleton OTUs (Table S2). This suggests that focusing on groups of interest that are consistently amplified can mitigate the spurious effects of low-abundance OTUs and consequently improve reproducibility. Thus, proper management of low-abundance OTUs remains crucial for extracting accurate and precise biodiversity estimates for complex communities, especially for studies focusing on rare and/or unique species such as conservation of species at risk and early detection of invading non-indigenous species (Zhan *et al.* 2014b). In addition, much more investigation is needed to understand the ecological significance of rare/unique taxa. For example, previous studies have suggested that it is possible that large numbers of species that exist at very low frequency could collectively impact community dynamics, as well as ecosystem structure and function (Li *et al.* 1983; Lyons & Schwartz 2001).

The observed low reproducibility between parallel replicates is most likely a result of random sampling processes during both biological sample collection and pyrosequencing processes (Zhou *et al.* 2008, 2011). For biological sample collections, low population density may lead to inconsistent presence/absence of rare species in collected plankton samples. If primers are designed to amplify certain taxonomic groups, as is the case here, species may also appear as rare in the dataset simply because they are rare within PCR products, rather than in the original biological samples. Our results indicate that low-abundance sequences can also be the result of spurious PCR amplification (inconsistent amplification) of distant taxonomic groups. This may affect reproducibility as well as inflate the proportion of truly low-abundance lineages. In addition, many steps in the pyrosequencing and biodiversity assessment procedures involve random sampling, including sample preparation, DNA extraction, emulsion and immobilization of beads, and bead deposition into wells on PicoTiter plates. Reproducibility results obtained in this study are consistent with the consequences of such random sampling processes. Random sampling processes are unlikely to lead to the absence of high-abundance OTUs, but could produce inconsistent presence/absence of low-abundance OTUs, leading to low reproducibility in parallel replicates (Figs 2 & 3).

A major concern is whether high variation derived from random sampling processes can affect diversity estimates, such as α- and β-diversity in complex communities. With respect to β-diversity, Zhou *et al.* (2011) determined that random sampling processes could pose a problem (e.g. over-estimation of β-diversity) if variation between replicates was higher than that between samples. In this study, we observed high values for the two β-diversity indices for both harbours (Table 1), suggesting that caution be applied in interpreting β-diversity patterns. The popularly used phylogenetic diversity-based methods, such as UniFrac (Lozupone & Knight 2005), can increase the accuracy and efficiency for comparing community structure (McDonald *et al.* 2013). However, sampling variation, especially for rare lineages, can lead to inflated phylogenetic distance estimates (Lozupone *et al.* 2011; Chen *et al.* 2012). Rarefaction is one method often used to overcome this, however, when sampling depth is highly variable among samples (as is the case here), rarefaction tends to eliminate a large number of sequence reads to get a consistent sampling depth across samples analysed. Such elimination contributes to a high level of variation and potentially reduces diversity estimates (Chen *et al.* 2012). Further investigation into the degree to which intra-sample variation contributes to community similarity comparisons is needed when using phylogenetic diversity-based methods. The use of technical replicates, coupled with multivariate statistical methods such as the phylogenetic distance matrix-based PERMANOVA, may serve as an experimental design and analytical method to weaken, or overcome, the influence of random sampling processes on estimation of β-diversity when using phylogenetic diversity-based methods (Lozupone *et al.* 2011). For α-diversity, Zhou *et al.* (2011) suggested that high variation among replicates could be less problematic for detecting new taxa. However, caution should be applied when using the number of OTUs or other methods such as rarefaction analysis to compare α-diversity among samples, because high variation derived from random sampling processes may lead to large differences even when using the same sequencing depth (e.g. in Nanaimo, Fig. 4).

Several methods have been proposed to avoid problems associated with random sampling processes when estimating biodiversity, including increasing the number of biological replicates, combining multiple methods (e.g. pyrosequencing and microarray), and removing low-abundance OTUs (Zhou *et al.* 2011). The former two methods were effective in the study of microbial communities (Zhou *et al.* 2008, 2011). However, the latter one should be used with caution, because removing low-abundance OTUs may reduce the ability to detect rare taxa in communities. We suggest that technical replicates be performed to assess the degree of variation among replicates, which could be used for statistical corrections and adjustments that would facilitate comparisons among studies (i.e. identify which low-abundance OTUs may be biologically meaningful). Indeed, it is relatively easy to conduct multiple replicates using tagged primers for HTS. As the expense of HTS decreases, large-scale sequencing for multiple replicates becomes increasingly attainable. Moreover, biological replicates are desirable to draw confident biological and ecological conclusions. Collectively, confidence in HTS results depends on sound experimental design and data interpretation.

Conclusions

Scientific reproducibility is critical in all studies, yet its determination in HTS-based biodiversity assessments remains poorly investigated. Our study reveals a complex but interesting scenario for HTS-based biodiversity assessment studies, having both technical significance and biological implications. Low reproducibility for low-abundance OTUs and taxa between parallel replicates likely stems from random sampling processes that occur during sample collection, sample preparation and sequencing. These random sampling processes may profoundly affect assessments of both α- and β-diversity. Our study indicates that replicates are required to assess the degree of variation for statistical corrections and adjustments to accurately measure biodiversity. In addition, proper management of low-abundance OTUs, rather than simple removal from datasets, is required to avoid underestimating biodiversity and losing unique and/or rare lineages/genotypes in complex communities. For example, the phylogenetic relationships between low-abundance and high-abundance OTUs, as well as a focused analysis on a targeted taxonomic group may permit more informed interpretations of the sampled biodiversity. Although our results were based on 454 pyrosequencing, they likely apply to other HTS technologies, including Illumina and SOLiD sequencing platforms.

Besides the low reproducibility for low-abundance OTUs obtained here, additional technical issues remain in data processing and interpretation for biodiversity assessments using HTS, such as filtering stringency for error/artifact removal and identification of thresholds for clustering species-level OTUs. The uncertainty on filtering stringency and clustering threshold may largely inflate the number of OTUs. Solving these stringency/threshold problems relies on both lab work such as establishment of testing artificial communities and advances in bioinformatics approaches such as development of robust bioinformatic algorithms. Our results add to a growing body of literature urging that caution be applied when designing metagenomics studies using HTS.

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 MEC, DDH and HJM, and by the NSERC Canadian Aquatic Invasive Species Network (CAISN), an NSERC Discovery Accelerator Supplement, and Canada Research Chair to HJM.

Data accessibility

454 pyrosequencing data for two fractions of both harbours surveyed in this study has been deposited into NCBI Sequence Read Archive (SRA) database with the accessions as follows: NCBI SRA accession SRR1171114 and SRR1177666 for the two fractions for Hamilton Harbour, and NCBI SRA accession SRR1177768 and SRR1177769 for the two fractions for Nanaimo Harbour.

References

Bray, J. & Curtis, J. (1957) An ordination of the upland forest communities in southern Wisconsin. *Ecology Monographs*, **27**, 325–349.

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D. & Li, H. (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, **28**, 2106–2113.

Chivian, D., Broidie, E.L., Alm, E.J., Culley, D.E., Dehal, P.S., DeSantis, T.Z. *et al.* (2008) Environmental genomics reveals a single-species ecosystem deep within Earth. *Science*, **322**, 275–278.

Creer, S. (2010) Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. *Molecular Ecology*, **19**, 2829–2831.

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.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P. *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Gihring, T.M., Green, S.J. & Schadt, C.W. (2012) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library. *Environmental Microbiology*, **14**, 285–290.

Gomez-Alvarez, V., Teal, T.K. & Schmidt, T.M. (2009) Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal*, **3**, 1314–1317.

Gotelli, N.J. & Entsminger, G.L. (2006) *EcoSim: Null Models Software for Ecology. Version 7*. Acquired Intelligence Inc. and Kesey-Bear, Jericho, VT 05465. http://garyents-minger.com/ecosim.ht.

Gruber, J.D., Colligan, P.B. & Wolford, J.K. (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using pyrosequencing. *Human Genetics*, **110**, 395–401.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A. *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

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.

Kauserud, H., Kumar, S., Brysting, A.K., Norden, J. & Carlsen, T. (2012) High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal plant root samples. *Mycorrhiza*, **22**, 309–315.

Kumar, S., Carlsen, T., Mevik, B., Enger, P., Blaalid, R., Shalchian-Tabrizi, K. & Kauserud, H. (2011) CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics*, **12**, 182.

Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.

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.

Li, W.K.W., Subba Rao, D.V., Harrison, W.G., Smith, J.C., Cullen, J.J., Irwin, B. & Platt, T. (1983) Autotrophic picoplankton in the tropical ocean. *Science*, **219**, 292–295.

Lozupone, C. & Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J. & Knight, R. (2011) UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal*, **5**, 169–172.

Lyons, K.G. & Schwartz, M.W. (2001) Rare species loss alters ecosystem function - invasion resistance. *Ecology Letters*, **4**, 358–365.

McDonald, D., Vázquez-Baeza, Y., Walters, W.A., Caporaso, J.G. & Knight, R. (2013) From molecules to dynamic biological communities. *Biology and Philosophy*, **28**, 241–259.

Pommier, T., Neal, P.R., Gasol, J.M., Coll, M., Acinas, S.G. & Pedrós-Alió, C. (2010) Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquatic Microbial Ecology*, **61**, 221–233.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, **26**, 1641–1650.

Prosser, J.I. (2010) Replicate or lie. *Environmental Microbiology*, **12**, 1806–1810.

Quince, C., Curtis, T.P. & Sloan, W.T. (2008) The rational exploration of microbial diversity. *The ISME Journal*, **2**, 997–1006.

Rousk, J., Bååth, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., Knight, R. & Fierer, N. (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. *The ISME Journal*, **4**, 1340–1351.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16–18.

Schütte, U.M.E., Abdo, Z., Foster, J., Ravel, J., Bunge, J., Solheim, B. & Forney, L.J. (2010) Bacterial diversity in a glacier foreland of the high Arctic. *Molecular Ecology*, **19**, 54–66.

Sørensen, T.A. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter*, **5**, 1–34.

Tang, C.Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T.G. & Fontaneto, D. (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.

Tedersoo, L., Nilsson, R.H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I. *et al.* (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, **188**, 291–301.

Větrovský, 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.

Wang, Z., Gerstein, M. & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

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

Zhan, A., Bailey, S.A., Heath, D.D. & MacIsaac, H.J. (2014a) Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources*, doi:10.1111/1755-0998.12254.

Zhan, A., Xiong, W., He, S. & MacIsaac, H.J. (2014b) Influence of artifact removal on rare species recovery in natural complex communities using high-throughput sequencing. *PLoS ONE*, **9**, e96928.

Zhou, J., Kang, S., Schadt, C.W. & Garten, C.T. (2008) Spatial scaling of functional gene diversity across various microbial taxa. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 7768–7773.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y., Tu, Q. *et al.* (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME Journal*, **5**, 1303–1313.

Received 15 September 2013; accepted 8 July 2014
Handling Editor: Daniel Faith

Supporting Information

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

Fig. S1. The approximately-maximum-likelihood phylogeny reconstructed using FastTree (numbers shown above branches in red represent local support values) for all Operational Taxonomic Units (OTUs) of Hamilton harbour. The OTUs were labelled as Tcluster#_#Fraction1Reads_#Fraction2Reads_Taxa_%identityBlastHit.

Fig. S2. The approximately-maximum-likelihood phylogeny reconstructed using FastTree (numbers shown above branches in red represent local support values) for all Operational Taxonomic Units (OTUs)

of Nanaimo harbour. The OTUs were labelled as Tcluster#_#Fraction1Reads_#Fraction2Reads_Taxa_%identityBlastHit.

Table S1. Taxon composition for the two plankton communities derived from Hamilton (freshwater) and Nanaimo (marine).

Table S2. Reproducibility of OTUs and family-level taxa when only considering OTUs matching crustacean families, across both parallel fractions of plankton communities from Hamilton (freshwater) and Nanaimo (marine), with percentages in brackets.

Table S3. Family-level reproducibility for crustaceans across two parallel fractions for (a) Hamilton and (b) Nanaimo.