Reproducibility of pyrosequencing data for biodiversity assessment in complex communities

Aibin Zhan¹,²*, Song He¹†, Emily A. Brown³, Frédéric J.J. Chain³, Thomas W. Therriault⁴, Cathryn L. Abbott⁴, Daniel D. Heath², Melania E. Cristescu²,³ and Hugh J. MacIsaac²

¹Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing 100085, China; ²Great Lakes Institute for Environmental Research, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada; ³Department of Biology, McGill University, 1205 Docteur Penfield, Montreal, QC H3A 1B1, Canada; and ⁴Pacific 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
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 (Ghiring, 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/generate 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

---

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 13 523 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 x PCR buffer, 2 mM of MgCl$_2$, 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 (Vétrovský & 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 errors/artifact sources for pyrosequencing, we identified and then deleted chimeras from each data set using the algorithmUCHIME (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. Chao 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$_d$) and Bray-Curtis’s abundance (BC$_a$)-based methods were calculated using EstimateS version 9.1.0 (http://viceroy.eeb.uconn.edu/estimates/). Their complements, S$_d$ ($S_d = 1 − S_d$) and BC$_a$ (BC$_a = 1 − BC_a$), 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 chimaera 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.

<table>
<thead>
<tr>
<th></th>
<th>Hamilton Fraction I</th>
<th>Hamilton Fraction II</th>
<th>Hamilton All</th>
<th>Nanaimo Fraction I</th>
<th>Nanaimo Fraction II</th>
<th>Nanaimo All</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of OTUs detected</td>
<td>353</td>
<td>244</td>
<td>427</td>
<td>566</td>
<td>592</td>
<td>849</td>
</tr>
<tr>
<td>No. of OTUs recovered by singletons</td>
<td>187 (53.0%)</td>
<td>116 (47.5%)</td>
<td>243 (56.9%)</td>
<td>313 (55.3%)</td>
<td>313 (52.9%)</td>
<td>514 (60.5%)</td>
</tr>
<tr>
<td>No. of taxa detected</td>
<td>38</td>
<td>29</td>
<td>42</td>
<td>61</td>
<td>47</td>
<td>68</td>
</tr>
<tr>
<td>No. of taxa recovered by singletons</td>
<td>29 (76.3%)</td>
<td>29 (100%)</td>
<td>38 (90.5%)</td>
<td>47 (77.0%)</td>
<td>38 (80.9%)</td>
<td>56 (82.4%)</td>
</tr>
<tr>
<td>No. of taxa recovered only by singletons</td>
<td>18 (47.4%)</td>
<td>12 (41.4%)</td>
<td>23 (54.8%)</td>
<td>21 (34.4%)</td>
<td>17 (36.2%)</td>
<td>25 (36.8%)</td>
</tr>
<tr>
<td>( S_d ) between two replicates (All data/singleton removed/single-, double-, triple-tots removed)</td>
<td>0.431/0.252/0.103</td>
<td>–</td>
<td>–</td>
<td>0.467/0.260/0.119</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>( BC_d ) between two replicates (All data/singleton removed/single-, double-, triple-tots removed)</td>
<td>0.459/0.458/0.457</td>
<td>–</td>
<td>–</td>
<td>0.269/0.267/0.262</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Representative sequence ID</th>
<th>Recovered by fraction (I or II)</th>
<th>BLAST result</th>
<th>Coverage (%)</th>
<th>Expect (E) value</th>
<th>Max identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamilton</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fragilariida</td>
<td>HAVE4QVO1DIJU0Y</td>
<td>I</td>
<td>Fragilariida crotonensis</td>
<td>96</td>
<td>0</td>
<td>97</td>
</tr>
<tr>
<td>Choanofoflagellida</td>
<td>HKQYOUX01D57S</td>
<td>II</td>
<td>Sphaeroeca volvox</td>
<td>90</td>
<td>1.55E-141</td>
<td>90</td>
</tr>
<tr>
<td>Cyrtolophosida</td>
<td>HAVE4QVO1BF3BV</td>
<td>I</td>
<td>Cyrtolophosida minor</td>
<td>92</td>
<td>2.14E-171</td>
<td>96</td>
</tr>
<tr>
<td>Cryptomonadales</td>
<td>HAVE4QVO1BSKT</td>
<td>I</td>
<td>Cryptomonas curvata</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Chroococcales</td>
<td>HKQYOUX01BIGK</td>
<td>II</td>
<td>Microcystis aeruginosa</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Hypotrichia</td>
<td>HAVE4QVO1DFLCJ</td>
<td>I</td>
<td>Parabistrichella variabilis</td>
<td>93</td>
<td>1.27E-161</td>
<td>93</td>
</tr>
<tr>
<td>Oligotrichia</td>
<td>HAVE4QVO1AI0P2F</td>
<td>I</td>
<td>Strombidium sp.</td>
<td>95</td>
<td>5.75E-166</td>
<td>95</td>
</tr>
<tr>
<td>Phylactolaemata</td>
<td>HKQYOUX01EUTU</td>
<td>II</td>
<td>Plumatella sp.</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Sphaeropleakes</td>
<td>HAVE4QVO1DUVE3</td>
<td>I</td>
<td>Coelastrum microporum</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Parachela</td>
<td>HKQYOUX01A9864</td>
<td>II</td>
<td>Murrayon pullari</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Zygmenatales</td>
<td>HKQYOUX01DUCSA</td>
<td>II</td>
<td>Monogea sp.</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>HKQYOUX01CT928</td>
<td>I</td>
<td>Spirogyra sp.</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Nanaimo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acari</td>
<td>HMWF7SE02JR4IE</td>
<td>I</td>
<td>Cacculidiae sp.</td>
<td>93</td>
<td>2.98E-163</td>
<td>94</td>
</tr>
<tr>
<td>Gymnosomata</td>
<td>HMWF7SE02G2YFB</td>
<td>I</td>
<td>Pneuromedera atlantica</td>
<td>98</td>
<td>0</td>
<td>98</td>
</tr>
<tr>
<td>Actiniaria</td>
<td>HOFZNNK02JL8NZ</td>
<td>II</td>
<td>Edwardsiella lineata</td>
<td>95</td>
<td>1.35E-167</td>
<td>95</td>
</tr>
<tr>
<td>Gymnodiales</td>
<td>HMWF7SE02J16D</td>
<td>I</td>
<td>Gymnodium fusiforme</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Kentrogonida</td>
<td>HMWF7SE02HYLN6</td>
<td>I</td>
<td>Loxothylacus panopaei</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Leptotheceae</td>
<td>HMWF7SE02B5GX</td>
<td>I</td>
<td>Phialella quadrata</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Littorinimorpha</td>
<td>HMWF7SE02EVEZS</td>
<td>I</td>
<td>Lucerna pallidula</td>
<td>99</td>
<td>1.70E-139</td>
<td>99</td>
</tr>
<tr>
<td>Monostilifera</td>
<td>HMWF7SE02J2MPI</td>
<td>I</td>
<td>Zygomenetis viriscens</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Nassellaria</td>
<td>HMWF7SE02J20FW</td>
<td>I</td>
<td>Chlorastrum tricoliopum</td>
<td>96</td>
<td>1.54E-179</td>
<td>96</td>
</tr>
<tr>
<td>Hymenoptera</td>
<td>HMWF7SE02FS78Z</td>
<td>I</td>
<td>Anagrus epos</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Prorodontida</td>
<td>HMWF7SE02FH6S8</td>
<td>I</td>
<td>Urotricha sp.</td>
<td>95</td>
<td>6.35E-159</td>
<td>95</td>
</tr>
<tr>
<td>Unclassed Dinophyceae</td>
<td>HOFZNNK02IS2FEN</td>
<td>II</td>
<td>Stockeriasp.</td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
</tbody>
</table>

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).
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). Sorensen and Bray-Curtis indices suggested a relatively high level of dissimilarity, with values of 0.431 and 0.467 for $S_0$ 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...
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 artificial 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 2.3 × 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 \( \alpha \) - and \( \beta \)-diversity in complex communities. With respect to \( \beta \)-diversity, Zhou et al. (2011) determined that random sampling processes could pose a problem (e.g. over-estimation of \( \beta \)-diversity) if variation between replicates was higher than that between samples. In this study, we observed high values for the two \( \beta \)-diversity indices for both harbours (Table 1), suggesting that caution be applied in interpreting \( \beta \)-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 \( \beta \)-diversity when using phylogenetic diversity-based methods (Lozupone et al. 2011). For \( \alpha \)-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 \( \alpha \)-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 \( \alpha \) - and \( \beta \)-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.
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.