Conserv Genet (2015) 16:513–522
DOI 10.1007/s10592-014-0678-9

PERSPECTIVE

Rare biosphere exploration using high-throughput sequencing: research progress and perspectives

Aibin Zhan · Hugh J. MacIsaac

Received: 26 July 2014 / Accepted: 6 November 2014 / Published online: 16 November 2014

Abstract Identification of rare species and mapping their distributions is crucial for understanding natural species distributions and causes and consequences of accelerating species declines. However, detection of rare species in both terrestrial and especially aquatic communities typically dominated by numerous microscopic species (i.e. rare biosphere) represents a formidable technical challenge. Rapid advances in high-throughput sequencing (HTS) technologies have revolutionized biodiversity studies in the rare biosphere, and also stimulated associated debates. Here we summarize research progress, discuss debates and problems, and propose possible solutions and future studies to address these issues. In addition, we provide take-home messages for experimental design and data interpretation when utilizing HTS techniques for rare biosphere exploration in ecology and conservation biology.

Keywords Biodiversity · Metabarcoding · Next-generation sequencing · Rare species · Type I error · Type II error

Introduction

Species in natural environments are increasingly placed at risk by interacting anthropogenic stressors that include over-exploitation, chemical pollution, climate change, and introductions of alien invasive species. These human-mediated disturbances contribute to changes in the global distribution of organisms, and to the sixth major extinction event in the history of life (Chapin III et al. 2000). The global decline in biodiversity is estimated 100–1,000 times faster than pre-human rates (Pimm et al. 1995; Barnosky et al. 2011). Understanding the direct causes and consequences of these dramatic changes, as well as the indirect, knock-on effects on ecosystems, is a critical issue in many disciplines of ecology and conservation biology. In addition, better understanding of these phenomena may also help develop effective conservation plans to stanch future biodiversity loss. Identification of rare species and mapping their distributions represent crucial prerequisites to understanding biodiversity distribution patterns and trends.

Communities are typically dominated by a few species, with an additional but varying number present at (very) low population abundance (e.g. Pedrós-Alió 2012). Such a large number of diverse rare species is collectively referred as to the "rare biosphere". The rare biosphere mainly includes native rare species and recently introduced non-indigenous species (NIS). Some of these native rare species may be vulnerable to extirpation due to demographic stochasticity (Wilson et al. 2011), while new NIS usually remain rare for a long period of time, even in cases where they eventually become dominant to cause large-scale negative effects (Crooks and Soulé 1999). Consequently, the rare biosphere must be well characterized to identify native endangered rare species for conservation, and to identify recently introduced NIS for early detection and rapid response.

The detection of an array of rare species represents a major technical challenge in all habitats, but particularly so in communities where species are not typically visible, such as microbes (Pedrós-Alió 2012) and plankton (Zhan et al. 2013). Recent technological advances have driven

A. Zhan (✉)
Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing 100085, China
e-mail: zhanaibin@hotmail.com; azhan@rcees.ac.cn

H. J. MacIsaac
Great Lakes Institute for Environmental Research, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada

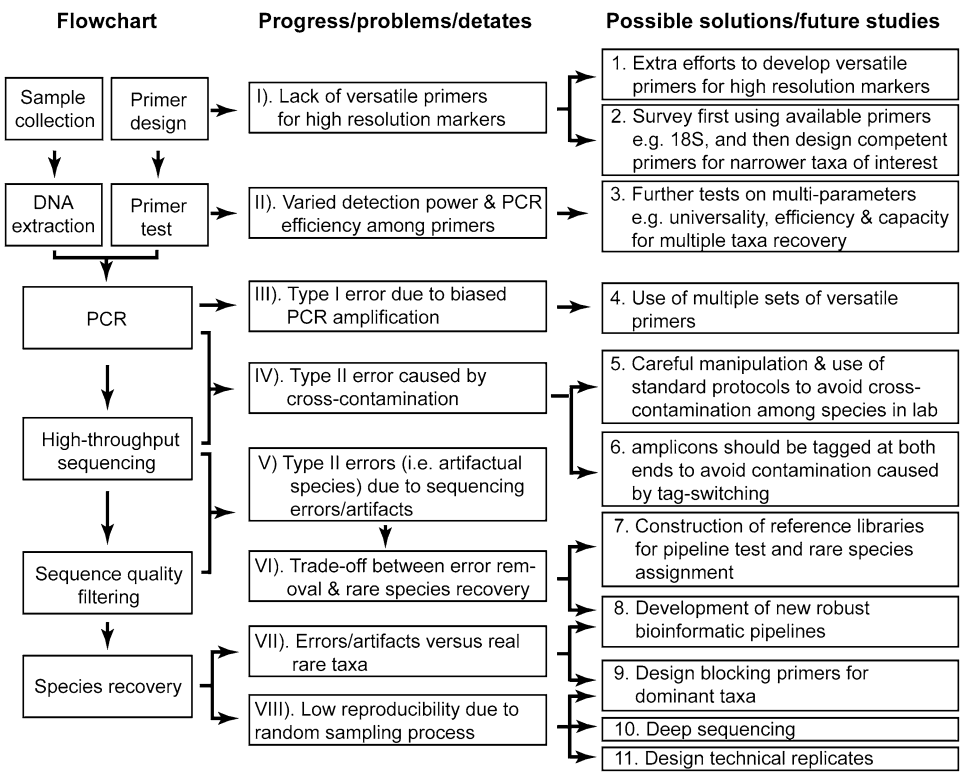

Fig. 1 Flowchart for rare biosphere exploration using high-throughput sequencing, and a summary of research progress, observed problems and/or debates raised, as well as possible solutions for future studies

rapid development of many DNA-based methods designed to facilitate biodiversity assessment in complex communities (see reviews by Darling and Mahon 2011; Zimmerman et al. 2014). In particular, the advent of high-throughput sequencing (HTS) technologies is revolutionizing biodiversity studies, especially for the rare biosphere in complex communities (e.g. Creer 2010; Hajibabaei 2012; Zhan et al. 2013; Bohmann et al. 2014).

Even though HTS technologies provide powerful tools to access otherwise difficult to detect communities, the use of these technologies is clouded by a number of issues including data interpretation, availability of reference collections, and a variety of technical issues. Here we summarize research progress, discuss debates and problems, and propose future studies to address these issues (Fig. 1). In addition, we provide take-home messages for experimental design and data interpretation when utilizing HTS techniques for rare species detection in conservation biology and ecology (Fig. 1).

Selection of versatile primers based on high resolution markers

The commonly used protocol for detection of rare species using HTS-based tools relies on PCR amplification (Fig. 1), which uses versatile (or "universal") primers to amplify DNA extracted from bulk samples and then PCR amplicons are subjected to sequencing for species identification (i.e. DNA metabarcoding). Selection of candidate gene regions, as well as associated versatile primers, represents the first most crucial step (e.g. Blaalid et al. 2013; Zhan et al. 2014a). Versatile primers in the paper refer to PCR primers that can be used to successfully amplify all species members in communities of interest.

In order to overcome the limitation of PCR-based methods caused by lack of versatile primers for high resolution markers (see next section below), PCR-free methods such as DNA capture and direct shotgun sequencing (i.e. metagenomics) have been proposed as alternatives (Taberlet et al. 2012). However, these methods remain largely untested on environmental samples and/or low input–output efficiency (i.e. available data versus sequencing cost/depth, e.g. Zhou et al. 2013a). Further, in a recent study on diet analysis the authors were unable to recover rare species using metagenomic methods (Srivathsan et al. 2014). Consequently, we focus on only commonly used PCR-based methods here.

Lack of versatile primers for high resolution markers: more efforts needed

An ideal gene region for PCR primer design is expected to have conserved regions among diverse taxonomic groups and hypervariable regions between primer annealing sites for species delimitation, while ideal primers are expected to be highly versatile in that they effectively amplify across

a wide range of taxa. Closely related taxa can be reliably differentiated using gene regions with fast evolutionary rates (i.e. hyperpolymorphism in DNA sequences among related taxa), however, ease of designing versatile primers is inversely related to evolutionary rate (Machida and Knowlton 2012). Despite inherent technical difficulties and the fact that primers often display specificity for a certain group of species, development and detailed testing of versatile primers represents the first priority for metabarcoding surveys (Machida and Knowlton 2012; Leray et al. 2013; Zhan et al. 2014a).

Available evidence clearly demonstrates that performance—including biodiversity detection power, PCR efficiency, and degree of universality—varies widely among primer pairs that were designed based on different gene regions (Machida and Knowlton 2012; Zhan et al. 2014a). For example, taxa recovery for a complex plankton community varied widely: 38 orders based on primers for nuclear small subunit ribosomal DNA (SSU rDNA) versus only 10 orders for mitochondrial 16S (mt16S), and many of the unrecovered taxa using mt16S were probably rare taxa (Zhan et al. 2014a). Owing to relatively high taxonomic resolution capability and a large public reference database for species annotation, mitochondrial cytochrome *c* oxidase subunit I (COI) is expected as a competent candidate gene for metabarcoding surveys for animals. However, owing to high polymorphism of COI sequences among taxa, only a few studies have successfully developed competent primers for COI (e.g. Leray et al. 2013; Gibson et al. 2014). More successful examples have been reported for the rDNA region, including the internal transcribed spacer (ITS), SSU rDNA, and large subunit ribosomal DNA (LSU rDNA), for diverse communities such as microbes (Nossa et al. 2010), fungi (Ihrmark et al. 2012; Blaalid et al. 2013) and plankton (Zhan et al. 2014a).

For eukaryotes, currently metabarcoding surveys for environmental samples rely mainly on rDNA regions, however there are technical problems (e.g. Tang et al. 2012; Lindahl et al. 2013). Despite the fact that ITS provides high taxonomic resolution power (e.g. Schoch et al. 2012), the frequent occurrence of insertions/deletions may render it difficult to assign sequences to taxa by sequence similarity, especially to higher taxonomic ranks (Machida and Knowlton 2012). Owing to a relatively slow evolutionary rate, SSU rDNA likely underestimates true species richness. Consequently, metabarcoding data based on SSU rDNA should be interpreted with caution for species-level patterns (Tang et al. 2012), even though this marker can recover a wide range of taxa in complex communities (e.g. Zhan et al. 2014a).

It is obvious that versatile primers for high resolution markers are required for metabarcoding surveys, especially those targeting the rare biosphere. However, given the inherent technical difficulties involved in designing highly versatile primers for fast-evolving genes, the rare biosphere can still be explored using available primers such as SSU rDNA at higher taxonomic ranks (e.g. family or higher); competent primers can then be more easily designed for specific narrower taxa based on fast-evolving genes such as COI, allowing for species delimitation. This two-step process may thus accommodate species exploration in the largely unexplored rare biosphere.

Varied detection sensitivity: another consideration for primer selection

HTS-based methods have been identified as sensitive tools for biodiversity assessment of the rare biosphere, allowing for simultaneous detection of an array of rare species using a single effort (Hajibabaei et al. 2011; Pochon et al. 2013; Zhan et al. 2013). Hajibabaei et al. (2011) determined that 454 pyrosequencing could accurately identify macro-invertebrate species present at more than 1 % abundance in a pooled mixture. A slightly higher level of sensitivity, 0.64 % abundance, was detected when using pooled DNA/PCR samples of marine species (Pochon et al. 2013). The survey used newly developed versatile primers to demonstrate that indicator species spiked into complex plankton communities could be recovered at much lower levels, as low as 2.3×10^{-5} % biomass (Zhan et al. 2013). It should be noted that there are difficulties in directly comparing these sensitivity levels, mainly because these studies effectively measured different organisms of interest. Available evidence suggested that differences in detection sensitivity across studies were not likely due to differing community complexity, since the plankton communities used in Zhan et al. (2013) are more complex than artificially assembled ones that used <30 species in the former two studies. In addition, sequencing depth cannot explain the sensitivity difference, as we had higher detection capability at 1/24 PicoTiter plate (detection limit: <0.021 % biomass across all tested species; Zhan et al. 2013) than Hajibabaei et al. (2011) had at 1/4 PicoTiter plate (detection limit: 1 % abundance). Collectively, the observed difference may be determined by selected primers for different taxa/communities analyzed. Such primer-associated efficiency differences were detected in other communities such as fungi (e.g. Ihrmark et al. 2012).

The consequences of primer-associated PCR efficiency among taxa remain unevaluated for metabarcoding surveys at a community level. A logical expectation in the rare biosphere is that the efficiency should be positively correlated with the number of species recovered. However, primer-associated PCR efficiency can be taxa-specific, leading to both biased detection of certain taxa (see next section for more detail) and highly varied detection thresholds among taxa in communities (Panel I). Indeed,

different detection sensitivity was observed across species tested, even though the versatile primer pair used exhibited high detection sensitivity across a wide range of taxa ranging from Mollusca, Echinodermata to Crustacea (see Table 1 in Zhan et al. 2013).

Ideal versatile primers are characterized by high PCR efficiency across an entire community. However, technical challenges exist for designing/choosing such versatile primers. Consequently, a test for PCR efficiency and detection sensitivity of primers in taxa of interest is expected. A practical way of testing efficiency is the use of representative indicator species to cover as wide a taxonomic range as possible. Technically, the carefully selected indicator species can be added into analyzed communities using a gradient of biomass proportions before DNA extraction, and then are identified during data analysis by bioinformatic tools (e.g. Ihrmark et al. 2012; Zhan et al. 2013). Although a test on a limited number of taxa cannot reflect the performance of primers across an entire community, it can provide both a preliminary evaluation of the utility of available primers, and possible suggestions for extra effort for modifying primers for biased amplified taxa.

Type I and Type II errors

Two critical issues of concern in data analysis and interpretation are Type I (failure to detect a species when it is present, i.e. incorrectly reject null hypothesis that a rare species is present in communities) and Type II errors (identification of a species not present in a community, i.e. presence of artifactual species). We address these issues in turn below.

Type I error caused by biased PCR amplification

A common and severe problem affecting metabarcoding studies is biased PCR amplification of different taxa in complex communities, leading to biased identification of certain taxa, i.e. type I error (Panel I). Type I error has been widely reported from both laboratory-based and silico-based surveys for biodiversity assessment using HTS (e.g. Clarke et al. 2014; Liu et al. 2013; Ovaskainen et al. 2013; van Velzen et al. 2012; Toju et al. 2012). This type of error is especially acute for those studies focusing on rare taxa (Bellemain et al. 2010; Engelbrektson et al. 2010). For example, estimates of species richness were highly influenced by primers used: short PCR amplicons (<400 bp) produce higher number of operational taxonomic units (OTUs) than do long ones (Huber et al. 2009; Engelbrektson et al. 2010). Biased PCR amplification can be caused by many factors, including universality of the primers, length of amplified noncoding regions (i.e. insertions/deletions) among taxa, and taxonomic composition of communities of interest (Bellemain et al. 2010; Engelbrektson et al. 2010).

The use of multiple sets of versatile primers can minimize biased detection of taxa in the rare biosphere. The use of multiple sets of primers based on different conserved regions can decrease the probability of biased PCR amplifications towards the same taxa. Results from multiple sets of primers can be cross-referred to obtain missed taxa caused by biased amplifications. However, it should be noted that cross-referencing based on multiple genes can be complicated, mainly owing to incomplete reference databases and poor taxonomic assignments among different genes used. The development of novel pipelines and reference databases are highly desired to integrate results based on multiple genes for a comprehensive survey on biodiversity.

Random sampling of rare taxa: a major cause for low reproducibility (Type I error)

Recent tests showed that the reproducibility of metabarcoding data was low among parallel replicates, for example, <30 % of OTUs for soil microbes when using two or three tagged replicates (Zhou et al. 2013b). When reproducibility was examined based on the abundance of OTUs, the reproducibility of low-abundance OTUs was much lower than high-abundance ones (Panel II), for example, as low as <25 % for singletons (OTUs represented by one sequence) versus 100 % for OTUs with the number of sequences >100 (Zhan et al. 2014c). Both laboratory work and a newly developed mathematical framework support the view that low reproducibility was a result of random sampling process during both biological sample preparation and data generation (Zhou et al. 2011; Ihrmark et al. 2012; Zhou et al. 2013b; Zhan et al. 2014c). For biological sample collection and preparation, low population density may lead to inconsistent presence of rare species in subsamples. For data generation, many steps involve random sampling processes, including DNA extraction, PCR, and random selection of amplicons during sequencing procedures (Zhan et al. 2014c).

Low reproducibility has profound influences on HTS-based studies. For example, biodiversity differences among communities (i.e. β-diversity) can be highly over-estimated due to a random sampling process. Even worse, wrong conclusions may be made if variation of intra-samples (i.e. replicates within a sample) was higher than that of inter-samples. It may be necessary to revisit some of these studies to re-assess β-diversity among communities. The use of deeper sequencing, blocking primers against dominant taxa (i.e. specific primers annealed to dominant taxa to

block PCR amplification), and pooling of repeated PCRs may help resolve this problem. In addition, for future experimental design, technical replicates (i.e. preparation of multiple DNA extractions from the same bulk sample) are needed to statistically correct intra-sample variation, while field-based replicate samples are desirable to substantiate results.

Real rare species versus artifacts: more efforts needed to reduce Type II error

Available evidence strongly supports the conclusion that errors/artifacts are a major cause of inflated biodiversity estimates (i.e. Type II error, presence of artifactual species), with error estimates as high as an order of magnitude (e.g. Quince et al. 2009; Kunin et al. 2010). The most problematic issue is the presence of OTUs represented by low-abundance reads such as singletons, doubletons and tripletons (i.e. OTUs represented by one, two and three sequences, respectively). Even after noises and error-prone sequences are removed during data processing, low-abundance OTUs may still account for a very large fraction of all detected OTUs. For example, 66.3 % of all recovered OTUs were singletons, doubletons and tripletons in a zooplankton community (Zhan et al. 2014a). Studies suggest that a majority of these reads may be errors/artifacts (e.g. Reeder and Knight 2009), although real rare species may exist among these low-abundance OTUs (Zhan et al. 2013; Brown et al. 2014; Zhan et al. 2014a).

One important point to keep in mind is that rare taxa may be represented by low frequency reads (e.g. Zhan et al. 2013; Brown et al. 2014). A central problem, however, is that the data filtering process will remove not only errors/artifacts, but also real rare species recovered as low-abundance OTUs (see the sequence quality filtering section for more detail). Rare sequences can be valuable and informative in reflecting unique lineages in communities. These issues highlight the urgent need to develop more robust bioinformatic algorithms to allow accurate sorting of informative reads from errors/artifacts. In addition, construction of more expansive reference libraries will allow better testing of bioinformatic algorithms, permit better filtering of errors/artifacts, and assure accurate rare species assignment.

Type II error caused by cross-contamination

Many HTS-based studies use environmental and/or impure samples. Nucleic acid isolated from such types of samples may contain contaminants. In addition, since sequencing capacity has been highly improved in the past several years, many samples (e.g. several hundred) can be pooled and sequenced in parallel. Cross-contamination can occur when preparing such a large number of samples in a short period of time in a laboratory. Cross-contamination represents a serious concern on downstream data analysis and interpretation, and even leading to erroneous conclusions (Schmieder and Edwards 2011). A survey showed that possible contamination occurred in 145 out of 202 metagenomes, with as high as 64 % contaminating sequences (Schmieder and Edwards 2011). Cross-contamination can definitely lead to the presence of artifactual species in HTS-based studies (i.e. Type II error). To reduce cross-contamination, careful manipulation, good organization, and the use of strictly standard experimental protocols are required when preparing a large number of samples. Moreover, the use of contamination removal pipelines, such as DeconSeq (Schmieder and Edwards 2011), can facilitate the detection and elimination of possible contaminant sequences. However, it should be noted that there are challenges to control, detect and then eliminate cross-contamination when dealing with various types of samples such as those from environments.

In addition to cross-contamination during laboratory work, there is an highly overlooked technical source of errors. In HTS-based studies, numerous samples are usually tagged, pooled and sequenced in parallel. Tags, also known as molecular identifiers (MIDs) which usually consist of 6–10 bp, are uniquely assigned to each sample either by linking to the amplified fragments directly during PCR (i.e. using fusion primers) or by ligation after PCR. The former method is widely used; however, recent studies clearly showed that such a time- and cost-saving strategy is a source of errors, such as Type II error caused by tag switching (i.e. contamination from pooled samples; see Carlsen et al. 2012 and references therein) and amplification bias (Berry et al. 2011). After pooling samples together, low concentrations of unused tagged primers may interfere with the amplicons during sequencing procedures to complete tag switching (Carlsen et al. 2012). Tag switching can result in erroneous assignment of sequence reads to wrong samples, causing cross contamination among pooled samples (i.e. Type II error). Type II error caused by tag switching may be a common but largely overlooked phenomenon in HTS-based studies (Westra et al. 2011; Carlsen et al. 2012). For example, 0.1–16 % sequence reads had non-compatible tag combinations in a 454 sequencing setup with mixed samples (van Orsouw et al. 2007). In many HTS-based studies, amplicons are usually tagged at one end only, leading to no power to detect and control for the presence of non-compatible tag combinations after sequencing. To correct Type II error caused by tag switching, amplicons should be tagged at both ends (Carlsen et al. 2012). In addition, thorough rinsing of PCR products, cold storage of pooled amplicon libraries immediately after mixing, and reduced sample

storage time between the final steps in the laboratory preparations may be used to avoid tag switching (Carlsen et al. 2012). Some pipelines such as CLOTU have implemented options for filtering out sequences with non-compatible tag combinations (Kumar et al. 2011). The application of such filtering options may highly decrease Type II error caused by tag switching. In addition to tag switching, empirical studies clearly demonstrated that tagged primers introduced amplification biases during PCR, leading to less reproducible data sets (Berry et al. 2011). Several strategies are suggested to avoid this problem, such as a nested PCR approach (Davey et al. 2014) and a 2-step PCR procedure, i.e. use conventional PCR primers to amplify the template during the first step of PCR amplification, and a dilution of yielded amplicons from the first step to serve as a template in a successive low-cycle-number PCR amplification using tagged primers (Berry et al. 2011).

Sequence quality filtering: a trade-off between Type I and Type II errors

In order to eliminate errors/artifacts, sequence data should be subjected to stringent sequence quality filtering (Quince et al. 2009; Kunin et al. 2010). However, rare species, which will be probably represented by low-abundance reads, may be sensitive to sequence quality filtering process (Panel III). Owing to the low number of sequences in final datasets, nucleotide ambiguity and/or low-quality nucleotides in real low-abundance reads can lead to complete removal of these sequences during quality filtering (Panel III). Consequently, there exists a trade-off between reducing Type I and Type II errors (i.e. failing to detect real rare species versus eliminating artifactual species, Zhan et al. 2014b). Using both internal (i.e. reliable OTUs selected from natural communities) and external (i.e. known spiked indicator species) references, three patterns were detected when testing this trade-off in plankton communities: (1) rare species were eliminated at all tested filtering stringencies; (2) more rare taxa were deleted as filtering stringency increased; and (3) elimination of rare species intensified as the relative biomass of a species decreased (Zhan et al. 2014b). Consequently, caution should be taken to avoid loss of rare taxa during data processing, especially for the use of sequence filtering strategies.

When handling low-abundance reads, the objective of a study must be considered. If an investigator is seeking to determine presence of specific species or those of management significance such as invasive alien species at early stages of biological invasions, reducing Type I error should be the primary concern. If, on the other hand, a researcher is seeking a conservative biodiversity estimate for a community, then considerations to limit Type II error should prevail.

As the awareness of sequencing errors increases, many new pipelines have been developed to either correct or remove sequencing errors, such as Blue (Greenfield et al. 2014), BLESS (Heo et al. 2014), UPARSE (Edgar 2013), Coral (Salmela and Schröder 2011), ECHO (Kao et al. 2011), HiTEC (Ilie et al. 2011), HSHREC (Salmela 2010), Reptile (Yang et al. 2010) and others (see review by Yang et al. 2013). These pipelines have their own advantages to handle certain types of data. For example, Reptile, HiTEC and ECHO usually provide more accurate results when compared with other methods for Illumina data; only pipelines HSHREC and Coral can handle insertion/deletion errors, and Coral can provide better accuracy (Yang et al. 2013). The use of these newly developed pipelines can improve the accuracy of biodiversity estimates. However, further investigation is needed to test whether, and the degree to which, these pipelines may influence rare species detection in natural complex communities. In addition, as sequencing capacity is improving at an extraordinarily high pace, the use of deeper sequencing may partially solve the artifact problem. Alternatively, blocking primers may be designed and applied against dominant taxa as a practical solution. Such a strategy has been effectively applied to environmental samples (e.g. Deagle et al. 2009; Boessenkool et al. 2012).

Take-home messages

(1) More effort is required regarding development of versatile primers for high resolution markers. Selection of competent versatile primers represents a crucial first step for experimental design. Several important characteristics of chosen primers should be tested by laboratory- and/or in silico-based work, including resolution power, universality/biased amplification among taxa of interest, PCR efficiency/detection limit, and multiple taxa recovery capability. In addition, the use of multiple sets of versatile primers should also be considered to avoid Type I error.

(2) There is no doubt that errors/artifacts can largely inflate biodiversity estimates (Type II error), and that data filtering improves the accuracy of diversity estimates. However, there exists a trade-off between error removal (reduced Type II error) and rare taxa recovery (reduced Type I error). This trade-off holds important consequences for biodiversity assessments, and investigators must make conscious decisions whether their objective is to ensure that they do not miss rare species of interest in their surveys, or

they wish to conservatively estimate community biodiversity. The use of blocking primers for dominant taxa and/or deep sequencing may help solve this dilemma. Many pipelines have recently been developed to remove/correct sequencing errors/artifacts; however, it remains largely untested whether, and the degree to which, these pipelines may potentially influence rare species recovery from natural complex communities. Obviously, more robust bioinformatic algorithms and enhanced reference taxonomic libraries are desired to sort out and identify real rare taxa. In addition, careful manipulation and the use of standard protocols, as well as proper design and use of MIDs (e.g. sample-specific tags), are needed to reduce Type II error caused by cross-contamination.

(3) Random sampling of a large number of rare taxa in complex communities may lead to low reproducibility. This problem has many profound influences on HTS-based studies, such as over-estimation of β-diversity among samples. Careful experimental design including the use of technical replicates for intra-sample variation correction, and technical improvements including the use of deep sequencing and blocking primers for dominant taxa, may help substantiate observed patterns.

(4) As biodiversity is rapidly declining globally, there is an urgent need for data-driven prioritization of conservation actions. Many conservation actions rely largely on fast and effective monitoring of rare species. Owing to the fast, sensitive, and effective nature of HTS-based methods, they have been successfully applied to monitoring endangered species for conservation, such as in freshwater ecosystems (Thomsen et al. 2012). Although HTS has its inherent disadvantages, errors can be corrected/reduced as long as these issues are fully acknowledged and managed. When using HTS-based methods for conservation plans, good experimental design and proper selection of competent protocols are needed using recommendations mentioned above. In addition, big efforts are required to develop robust bioinformatic algorithms and to expand public database for species annotation. As genetic detection tools are being adopted in decision-making agents (e.g. Darling and Mahon 2011; Ojaveer et al. 2013), we expect that HTS-based tools will play a key role in conservation management of rare species based on environmental DNA, especially in detection and mapping distributions of rare species at large geographical scales. The wide use of HTS-based methods can serve as a bridge to close the gap between research and management (Darling 2014).

Acknowledgments This work was supported by the National Natural Science Foundation of China (31272665), the One-Three-Five Program (YSW2013B02) of the Research Center for Eco-Environmental Sciences and 100-Talent Program of the Chinese Academy of Sciences to A.Z., by Discovery grants from Natural Sciences and Engineering Research Council of Canada (NSERC), the NSERC Canadian Aquatic Invasive Species Network (CAISN), and Canada Research Chair to H.J.M.

Appendix

Panel I

Selection of candidate gene regions, as well as associated versatile primers, represents the first most crucial step for high-throughput sequencing (HST)-based rare biosphere exploration. Despite that the universality of primer pairs selected has been tested using either Sanger sequencing or small-scale HST, the biodiversity detection efficiency may still be highly varied, especially when characterizing complex communities. Such varied detection efficiency occurred among primer pairs based on both different types of genetic markers (Fig. 2) and even different regions of the same genetic markers (Fig. 3). In addition, biased PCR amplification of different taxa in complex communities may effect both relatively abundant and rare taxa (Fig. 2), leading to Type I error (i.e. failure to detect a species when it is present). For example, Cyclopidae, a relatively abundant taxon recovered by18S (33.3% of all total observed Operational Taxonomic Units, OTUs) was not detected when using mt16S (Fig. 2). The Type I error becomes more severe when targeting relatively rare taxa. Four taxa detected by 18S with relative abundance lower than 2% were not recovered by mt16S (Fig. 2). The number of taxa was 19 versus 15 for V4 and V5-7 regions of 18S, respectively (Fig. 3). All these results suggest that Type I error often occurs when using HTS for rare biosphere exploration. Type I error may become a serious problem

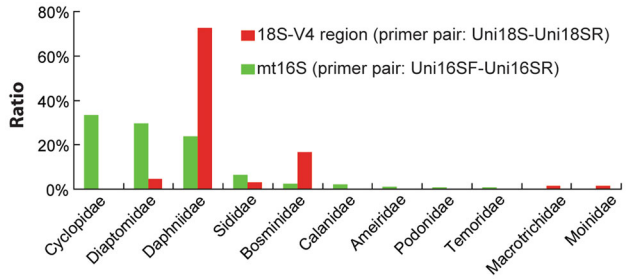

Fig. 2 Comparison of family-level taxa of Crustacea recovered from the complex plankton community collected from Hamilton Harbour, Ontario, Canada using 454 pyrosequencing based on two types of genetic markers, 18S (nuclear) and mt16S (mtDNA). The sequencing depth was 1/2 PicoTiter plate for each marker. Data is derived from Zhan et al. (2014a)

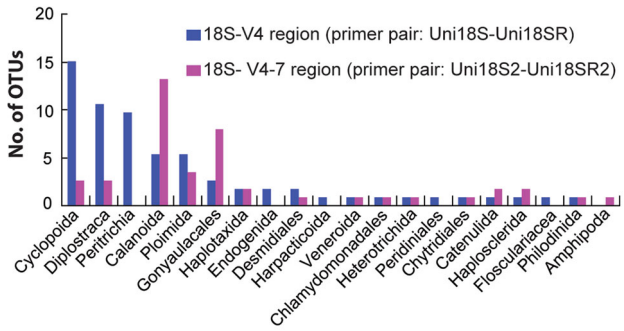

Fig. 3 Comparison of order-level taxa recovered from the complex plankton community collected from Hamilton Harbour, Ontario, Canada using a small-scale run of 454 pyrosequencing (i.e. an equivalent of 1/48 PicoTiter plate) based on two regions (V4 and V5-7) of 18S. Data is derived from Zhan et al. (2014a)

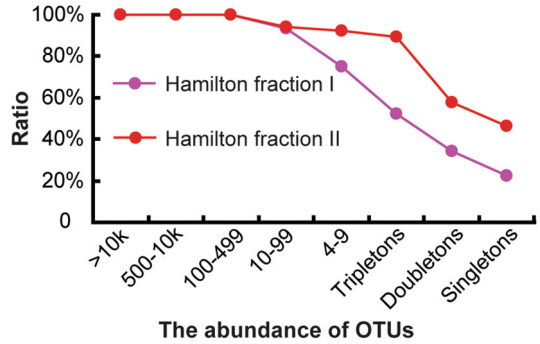

Fig. 4 Reproducibility analysis of Operational Taxonomic Units (OTUs) for the two parallel fractions (1/2 PicoTiter plate) of a complex plankton community collected from Hamilton Harbor, Ontario, Canada. Reproducibility refers to the capacity of an entire pyrosequencing dataset to be thoroughly replicated when using exactly the same protocol throughout the whole experiment. OTUs were grouped based on their abundance. Singletons, doubletons and tripletons denote OTUs represented by one, two and three sequences, respectively. Data is derived from Zhan et al. (2014c)

when using HTS-based methods for protection of endangered species and early detection and rapid response of recently introduced alien invasive species.

Panel 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 out of a community. Random sampling occurs during the whole experimental procedure including field biological sampling, PCR and sequencing. Low population density leads to inconsistent presence of rare species among parallel replicates during sample preparation and inconsistent presence of amplicons/sequences during PCR and sequencing. Low reproducibility is particularly serious for rare taxa (Fig. 4). By analyzing two parallel replicates, the results showed that the reproducibility was 100% for high-abundance Operational Taxonomic Units (OTUs) (>100 sequences). However, the reproducibility was lower for low-abundance OTUs, and sometimes <25% for singletons (Fig. 4). These low-abundance OTUs, including those irreproducible, were often assigned to different taxa of interest (Fig. 5). Errors and artefacts may inflate the number of low-abundance reads; however, these factors cannot create new taxonomic groups. Collectively, low-abundance OTUs, at least some of them, reflect unique rare lineages in communities (Fig. 5). Low reproducibility due to random sampling results in multiple problems including influence on both α- and β-diversity estimates, as well as Type I error (Zhan et al. 2014c).

Panel III

There is no doubt that sequencing errors can largely inflate biodiversity estimates. In order to eliminate overestimation, high-throughput sequencing data is usually subjected

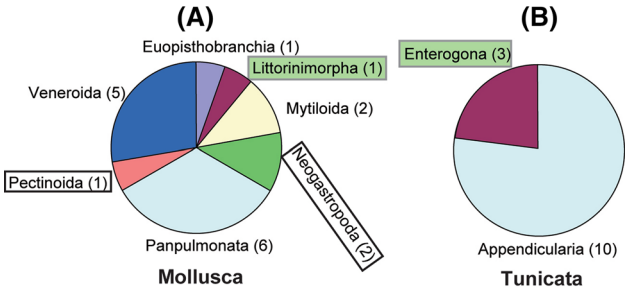

Fig. 5 Reproducibility analysis of two groups (mollusca, A and tunicata, B) of interest based on two parallel fractions (1/2 PicoTiter plate) of a complex plankton community collected from Nanaimo Harbor, British Columbia, Canada. Reproducibility refers to the capacity of an entire pyrosequencing dataset to be thoroughly replicated when using exactly the same protocol throughout the whole experiment. Taxa in boxes indicate recovery by singletons only, while highlighted ones indicate recovery by irreproducible singletons between two parallel fractions. Numbers in brackets show the number of OTUs detected in this taxon. Data is derived from Zhan et al. (2014c)

to sequence quality filtering. Based on surveys using both internal (Fig. 6) and external references (Fig. 7), rare species represented by low-abundance sequences in datasets have been approved to be more sensitive to artifact removal process when compared to abundant species. Loss of rare species occurred at even low filtering stringencies, such as $Q = 10$ for cases based on both internal and external references. Generally, elimination of rare species was intensified as the relative biomass of a species was decreased (Fig. 7). All these patterns clearly support a trade-off between reducing Type I (failing to detect real rare species) and Type II (eliminating artifactual species) errors.

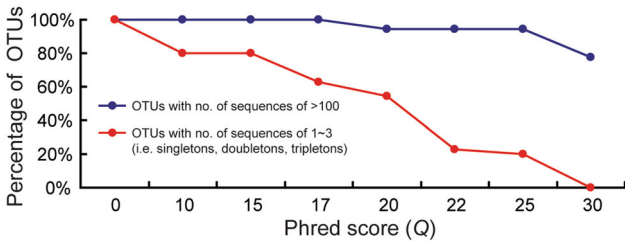

Fig. 6 Influence of sequence quality filtering on detection of rare taxa based on internal references, i.e. Operational Taxonomic Units (OTUs) with high similarities (similarity >99%, query coverage >99%) to available records in GenBank were chosen from the community itself as internal references. OTUs were grouped based on their abundance, and ratios of internal references retained in two groups are shown at filtering stringencies of Q (Phred score) = 0 (no filtering) to 30. Data is derived from a plankton community collected from Nanticoke Harbour, Ontario, Canada by Zhan et al. (2014b)

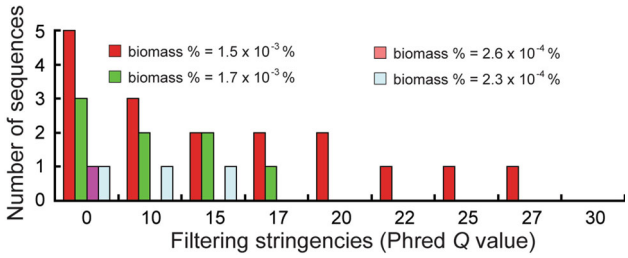

Fig. 7 Influence of sequence quality filtering on detection of rare taxa based on external references, i.e. known indicator species. These known indicator species were spiked into complex plankton communities using a series of biomass gradients. After 454 pyrosequencing, data was subjected to a series of filtering stringencies at Q (Phred score) = 0 (no filtering) to 30, and then the known spiked rare species were identified using local BLAST from each dataset. Data is derived from Zhan et al. (2013) and Zhan et al. (2014b)

References

Barnosky AD, Matzke N, Tomiya S, Wogan GO, Swartz B, Quental TB, Marshall C, McGuire JL, Lindsey EL, Maguire KC, Mersey B, Ferrer EA (2011) Has the Earth's sixth mass extinction already arrived? Nature 471:51–57

Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. BMC Microbiol 10:189

Berry D, Mahfoudh KB, Wagner M, Loy A (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl Environ Microbiol 77:7846–7849

Blaalid R, Kumar S, Nilsson RH, Abarenkov K, Kirk PM, Kauserud H (2013) ITS1 versus ITS2 as DNA metabarcodes for fungi. Mol Ecol Resour 13:218–224

Boessenkool S, Epp LS, Haile J, Bellemain E (2012) Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. Mol Ecol 21:1806–1815

Bohmann K, Evans A, Gilbert TP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends Ecol Evol 29:358–367

Brown SP, Veach AM, Rigdon-Huss AR, Grond K, Lickteig SK, Lothamer K, Oliver AK, Jumpponen A (2014) Scraping the bottom of the barrel: are rare high throughput sequences artifacts? Fungal Ecol. doi:10.1016/j.funeco.2014.08.006

Carlsen T, Aas AB, Lindner D, Vrålstad T, Chumacher T, Kauserud H (2012) Don't make a mista(g)ke: Is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecol 5:747–749

Chapin FS III, Zavaleta ES, Eviner VT, Naylor RL, Vitousek PM, Reynolds HL, Hooper DU, Lavorel S, Sala OE, Hobbie SE, Mack MC, Díaz S (2000) Consequences of changing biodiversity. Nature 405:234–242

Clarke LJ, Soubrier J, Weyrich LS, Cooper A (2014) Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. Mol Ecol Resour. doi:10.1111/1755-0998.12265

Creer S (2010) Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. Mol Ecol 19:2829–2831

Crooks JA, Soulé ME (1999) Lag times in population explosions of invasive species: Cases and implications. In: Sandlund OT, Schei PJ, Viken A (eds) Invasive species and biodiversity management. Kluwer Academic Publishers, Dordrecht

Darling JA (2014) Genetic studies of aquatic biological invasions: closing the gap between research and management. Biol Invasions. doi:10.1007/s10530-014-0726-x

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

Davey ML, Kauserud H, Ohlson M (2014) Forestry impacts on the hidden fungal biodiversity associated with bryophytes. FEMS Microbiol Ecol. doi:10.1111/1574-6941.12386

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

Edgar RC (2013) UPARSE: highly accurate OUT sequences from microbial amplicon reads. Nat Methods 10:996–998

Engelbrektson A, Kunin V, Wrighton K, Zvenigorodsky N, Chen F, Ochman H, Hugenholtz P (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J 4:642–647

Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. P Natl Acad Sci USA. doi:10.1073/pnas.1406468111

Greenfield P, Duesing K, Papanicolaou A, Bauer DC (2014) Blue: correcting sequencing errors using consensus and context. Bioinformatics. doi:10.1093/bioinformatics/btu368

Hajibabaei M (2012) The golden age of DNA metasystematics. Trends Genet 28:535–537

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One 6:e17497

Heo Y, Wu X, Chen D, Ma J, Hwu WM (2014) Bless: bloom filter-based error correction solution for high-thoughput sequencing reads. Bioinformatics 30:1354–1362

Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environ Microbiol 11:1292–1302

Ihrmark K, Bödeker ITM, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandström-Durling M, Clemmensen KE, Lindahl BD (2012) New primers to amplify the fungal ITS2 region—evaluation by 454-sequencing of artificial and natural communities. FEMS Microbiol Ecol 82:666–677

Ilie L, Fazayeli F, Ilie S (2011) HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics 27:295–302

Kao WC, Chan AH, Song YS (2011) ECHO: a reference-free short-read error correction algorithm. Genome Res 21:1181–1192

Kumar S, Carlsen T, Mevik B-H, 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. Environ Microbiol 12:118–123

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front Zool 10:34

Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjøller R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J, Kauserud H (2013) Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. New Phytol 199:288–299

Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y, Yu DW, Zhou X (2013) SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. Methods Ecol Evol 4:1142–1150

Machida RJ, Knowlton N (2012) PCR Primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. PLoS One 7:e46180

Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, De Santis TZ, Brodie EL, Malamud D, Poles MA, Pei Z (2010) Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroentero 16:4135–4144

Ojaveer H, Galil BS, Minchin D, Olenin S, Amorim A, Canning-Clode J, Chainho P, Copp GH, Collasch S, Jelmert A, Lehtiniemi M, McKenzie C, Mikušm J, Miossecn L, Occhipinti-Ambrogio A, Pećarevićm M, Pedersonp J, Quilez-Badiaq G, Wijsmanr JWM, Zenetoss A (2013) Ten recommendations for advancing the assessment and management of non-indigenous species in marine ecosystems. Mar Policy 44:1–6

Ovaskainen O, Schigel D, Ali-Kovero H, Auvinen P, Paulin L, Nordén B, Nordén J (2013) Combining high-throughput sequencing with fruit body surveys reveals contrasting life-history strategies in fungi. ISME J 7:1696–1709

Pedrós-Alió C (2012) The rare bacterial biosphere. Annu Rev Mar Sci 4:449–466

Pimm SL, Russell GJ, Gittleman JL, Brooks TM (1995) The future of biodiversity. Science 269:347–350

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

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6:639–641

Reeder J, Knight R (2009) The 'rare biosphere': a reality check. Nat Methods 6:636–637

Salmela L (2010) Correction of sequencing errors in a mixed set of reads. Bioinformatics 26:1284–1290

Salmela L, Schröder J (2011) Correcting errors in short reads by multiple alignments. Bioinformatics 27:1455–1461

Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6:e17288

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. P Natl Acad Sci USA 109:6241–6246

Srivathsan A, Sha JCM, Vogler AP, Meier R (2014) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). Mol Ecol Resour. doi:10.1111/1755-0998.12302

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. Mol Ecol 21:1789–1793

Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. P Natl Acad Sci USA 109:16208–16212

Thomsen PF, Kielgast J, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP, Orlando L, Willerslev E (2012) Monitoring endangered freshwater biodiversity using environmental DNA. Mol Ecol 21:2565–2573

Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-coverage ITS primers for the DNA-based identification of Ascomycetes and Basidiomycetes in environmental samples. PLoS One 7:e40863

van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, Poel Hvd, Oeveren Jv, Verstege H, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJT (2007) Complexity reduction of polymorphic sequences (CRoPS^{TM}): A novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 2:e1172

van Velzen R, Weitschek E, Felici G, Bakker FT (2012) DNA barcoding of recently diverged species: relative performance of matching methods. PLoS One 7:e30490

Westra H-J, Jansen RC, Fehrmann RSN, te Meerman GJ, van Heel D, Wijmenga C, Franke L (2011) MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics 27:2104–2111

Wilson HB, Joseph LN, Moore AL, Possingham HP (2011) When should we save the most endangered species? Ecol Lett 14:886–890

Yang X, Dorman KS, Aluru S (2010) Reptile: representative tiling for short read error correction. Bioinformatics 26:2526–2533

Yang X, Chockalingam SP, Aluru S (2013) A survey of error-correction methods for next-generation sequencing. Brief Bioinform 14:56–66

Zhan A, Hulák M, Sylvester F, Huang X, Adebayo AA, Abbott CL, Adamowicz SJ, Heath DD, Cristescu ME, MacIsaac HJ (2013) High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Methods Ecol Evol 4:558–565

Zhan A, Bailey SA, Heath DD, MacIsaac HJ (2014a) Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. Mol Ecol Resour 14:1049–1059

Zhan A, He S, Brown EA, Chain FJJ, Therriault TW, Abbott CL, Heath DD, Cristescu ME, MacIsaac HJ (2014b) Reproducibility of pyrosequencing data for biodiversity assessment in complex communities. Methods Ecol Evol 5:881–890

Zhan A, Xiong W, He S, MacIsaac HJ (2014c) Influence of artifact removal on rare species recovery in natural complex communities using high-throughput sequencing. PLoS One 9:e96928

Zhou J, Wu L, Deng Y, Zhi X, Jiang Y, Tu Q, Xie J, Van Nostrand JD, He Z, Yang Y (2011) Reproducibility and quantitation of amplicon sequencing-based detection. ISME J 5:1303–1313

Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013a) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. GigaScience 2:4

Zhou J, Jiang Y, Deng Y, Shi Z, Zhou B, Xue K, Wu L, He Z, Yang Y (2013b) Random sampling process leads to overestimation of β-diversity of microbial communities. mBio 4:e00324

Zimmerman N, Izard J, Klatt C, Zhou J, Aronson E (2014) The unseen world: environmental microbial sequencing and identification methods for ecologists. Front Ecol Environ 12:224–231