Early detection of invasive species in marine ecosystems using high-throughput sequencing: technical challenges and possible solutions

Wei Xiong^{1,2} · Haitao Li³ · Aibin Zhan^{1,2}

Received: 15 December 2015 / Accepted: 11 May 2016 / Published online: 23 May 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Introduction and spread of invasive non-indigenous species (NIS) has been clearly recognized as a growing global problem in marine and coastal ecosystems. Early detection of NIS has been recognized as one of the first priorities and effective ways in management programs. However, the detection of an array of rare NIS based on traditional morphological methods represents enormous technical challenges in the marine realm, as numerous poorly studied organisms may be small and/or have insufficient morphological distinctive features. The use of DNA-based approaches has largely increased the detection accuracy and efficiency. Owing to many obvious advantages such as the high capacity to detect extremely low-abundance species and low cost per sequence, high-throughput sequencing (HTS)-based methods have been popularly proposed for early detection of marine NIS. Although HTS-based approaches have been approved robust for the census of NIS in aquatic ecosystems, especially for those present at

extremely low abundance, many factors in experimental design and data collection can cause errors including both false negatives and false positives. Here we identify factors responsible for both error types, discuss causes and consequences, and propose possible solutions when utilizing HTS for early detection of NIS in marine and coastal ecosystems.

Introduction

The introduction and spread of invasive non-indigenous species (NIS) has been clearly recognized as a growing global problem in marine and coastal waters, causing significantly adverse consequences in ecosystems, economy, and even human society growth and development (e.g., Lawler et al. 2006; Lodge et al. 2006). Compared to terrestrial ecosystems, marine habitats are more uniform (i.e., have less distinct habitat boundaries). Such hydrogeographical feature leads to high direct connectivity both among invaded habitats and between native and invaded regions, thus making marine ecosystems have higher probabilities to be invaded. In addition, the lack of “physical barriers” in marine ecosystems also results in the difficulties in management of invasive species (see examples in Zhan et al. 2015). Recent studies have illustrated that prevention of NIS can be more cost-effective than post-establishment management to control the establishment and spread of newly introduced NIS (Lodge et al. 2006; Finnoff et al. 2007). As the early detection of incipient invasions by harmful species increases the feasibility of rapid responses to eradicate the species or constrain its spread, early detection has been recommended as one of the first priorities in management plans and programs (e.g., Lodge et al. 2006; Vander Zanden et al. 2010).

Responsible Editor: T. Reusch.

Reviewed by Undisclosed expert.

This article is part of the Topical Collection on Invasive Species.

✉ Aibin Zhan
zhanaibin@hotmail.com; azhan@rcees.ac.cn

¹ Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing 100085, China

² University of Chinese Academy of Sciences, Chinese Academy of Sciences, 19A Yuquan Road, Shijingshan District, Beijing 100049, China

³ South China Sea Environmental Monitoring Center, State Oceanic Administration, 155 Xingang Road West, Guangzhou 510300, Guangdong, China

The biggest technical challenge for early detection is the identification of newly introduced NIS present at extremely low abundance in communities (Darling and Mahon 2011; Zhan et al. 2013; Zhan and MacIsaac 2015). In addition, as only a small proportion of introduced species (e.g., 1/10; Williamson and Fitter 1996) can establish, form self-sustaining populations, and finally become invasive in introduced ranges, it adds the difficulty to monitor a huge number of introduced species, given the circumstances that thousands of species are transplanted daily by human activities such as shipping (e.g., Carlton and Geller 1993). Such detection challenges become more obvious in some particular ecosystems such as marine and coastal ones where organisms may be small, geographically constrained, and hidden beneath the water surface (McDonald 2004; Jerde et al. 2011). Traditional methods, which rely on capture using different sampling instruments and then identification via assessment of morphological and/or anatomical characteristics, have been empirically proven as low-resolution or low-throughput means for the detection of newly introduced NIS (e.g., <20 % based on bulk samples; Harvey et al. 2009). Thus, the comprehensive identification of rare taxa is time-consuming and laborious in marine and coastal ecosystems. In addition, correct identification of marine organisms is often impossible, especially for closely related species, morphologically similar species, and cryptic species complexes, and also for particular life stages such as eggs and immature individuals (e.g., Briski et al. 2011; Mahon et al. 2011; Carugati et al. 2015). Consequently, cost-effective, accurate, high-throughput, and common methods across an array of taxa are desired for early detection of NIS in marine ecosystems (Darling and Mahon 2011; Zhan et al. 2013).

Recent technological advances have promoted rapid development of DNA-based methods designed to facilitate early detection of NIS in aquatic ecosystems (see review by Darling and Mahon 2011; Zhan and MacIsaac 2015). DNA-based methods have radically changed our ability to identify biodiversity in communities (e.g., Mackie and Geller 2010; Briski et al. 2011; Jerde et al. 2011; Zhan et al. 2013). In particular, the use of high-throughput sequencing (HTS)-based methods has highly improved the detection efficiency, accuracy, and sensitivity for rare species in aquatic communities, including native endangered species and newly introduced NIS (e.g., Creer 2010; Creer et al. 2010; Zhan et al. 2013; Carugati et al. 2015). For example, the biomass of spiked rare species in complex plankton communities could be recovered as low as 2.3×10^{-5} % (Zhan et al. 2014a). Compared to other molecular methods such as species-specific PCR, HTS-based methods do not require genetic background information for species to be detected, and more importantly, can detect a large number of species using a single effort (i.e., high throughput).

The obvious advantages, especially the high throughput for detecting an array of unknown species and high sensitivity, have intensified researchers' interest to use HTS-based methods for early detection of invasive NIS in aquatic ecosystems (Lodge et al. 2012; Comtet et al. 2015; Zhan and MacIsaac 2015).

However, as the prerequisite of accurate detection, a number of technical issues should be properly addressed when applied HTS-based methods for early detection of NIS in marine ecosystems (Cowart et al. 2015; Ficetola et al. 2015; Pompanon and Samadi 2015; Zaiko et al. 2015). In particular, technical issues, especially those responsible for errors, need to be well acknowledged, tested, and characterized before HTS-based methods can be broadly applied to early detection programs and decision-making frameworks (Zhan and MacIsaac 2015). In summary, technical issues can lead to false negatives (i.e., failure to detect a species when it is present) and false positives (i.e., identification of a species not present in a community). Here, we identify major technical error sources for both false negatives and false positives, discuss causes and consequences, and propose possible solutions for future studies (Fig. 1). Despite the fact that PCR-free strategies such as DNA capture and direct shotgun sequencing (i.e., metagenomics) have been proposed to avoid biased amplification and/or due to the lack of universal PCR primers (Taberlet et al. 2012), these methods have not been well validated and/or have limited capacities to recover rare taxa (Zhou et al. 2013a; Srivathsan et al. 2015; Zhan and MacIsaac 2015). We therefore focus on the commonly used PCR-method strategy (i.e., metabarcoding strategy; Fig. 1) in the present study.

False negatives

Random sampling of rare taxa

As the population abundance may remain extremely low at early stages of biological invasions, random sampling of rare taxa occurs across the whole procedure of early detection using HTS, including field sampling, sample preparation for DNA extraction, PCR and sequencing (Zhan and MacIsaac 2015). Random sampling processes can result in the absence of rare NIS in samples and/or inconsistent occurrence of rare NIS in replicate samples (Zhou et al. 2011; Zhan et al. 2014a; Zhou et al. 2013b). False negatives caused by random sampling processes have been confirmed by both laboratory work and mathematical modeling (Zhou et al. 2011; Ihrmark et al. 2012; Zhou et al. 2013b; Zhan et al. 2014a). For example, rare species in zooplankton usually cannot be consistently identified when their abundance was low (Table 1), and the error rate trends

Fig. 1 Flowchart for early detection of marine non-indigenous species (NIS) using high-throughput sequencing based on the metabarcoding strategy and a summary of error sources and possible solutions for both false-positive and false-negative errors

Table 1 The detection of rare species using spiked known indicator species (bay scallop *Argopecten irradians* and golden mussel *Limnoperna fortunei*) to existing natural plankton samples based on 454 pyrosequencing

Species	Biomass level ($\times 10^{-6}$ %)		Biomass level ($\times 10^{-5}$ %)		Biomass level ($\times 10^{-4}$ %)	
	Replicate—biomass	Detection	Replicate—biomass	Detection	Replicate—biomass	Detection
Bay scallop <i>Argopecten irradians</i>	I— 1.8×10^{-6} %	No	I— 2.7×10^{-5} %	No	I— 2.6×10^{-4} %	Yes
	II— 2.4×10^{-6} %	No	II— 2.3×10^{-5} %	Yes	II— 1.9×10^{-4} %	No
	III— 2.4×10^{-6} %	No	III— 2.0×10^{-5} %	No	III— 2.3×10^{-5} %	Yes
Golden mussel <i>Limnoperna fortunei</i>	I— 4.0×10^{-6} %	No	I— 4.1×10^{-5} %	No	I— 4.7×10^{-4} %	Yes
	II— 3.8×10^{-6} %	No	II— 4.3×10^{-5} %	Yes	II— 5.9×10^{-4} %	Yes
	III— 3.7×10^{-6} %	No	III— 4.2×10^{-5} %	No	III— 5.9×10^{-4} %	Yes

The sequencing depth is an equivalent of 1/24 picotiter plate for each replicate. The biomass levels ($\times 10^{-6}$ %), ($\times 10^{-5}$ %), and ($\times 10^{-4}$ %) are equivalent of 0.01, 0.1, and 1 larva spiked into approximately 100 mg plankton samples, respectively. The false negatives caused by random sampling of rare taxa can be detected when the spiked biomass of indicator species was above the detection threshold (e.g., $\times 10^{-4}$ % for the bay scallop), and the error rate trends to increase as the species abundance decreases. All original data are derived from Zhan et al. (2014a)

to increase as the species abundance decreases (Zhan et al. 2014a, b). Using an analogous ball example under Poisson sampling with limited sampling efforts, the developed

mathematical framework clearly showed that random sampling was a major error source for false negatives, and such predication was further approved by laboratory work (Zhou

et al. 2013b). False negatives caused by random sampling processes can lead to extremely low reproducibility among technical replicate samples, especially for low-abundance taxa (e.g., as low as <20 %; Zhou et al. 2011; Zhan et al. 2014a).

Owing to the low population abundance, false negatives associated with random sampling of rare taxa unavoidably exist throughout the whole procedure of HTS-based methods (Fig. 1). In addition to traditionally used repeated sampling, several more technical issues should be considered to decrease or possibly correct the error. These technical issues include the use of (1) species occupancy models with an optimized replication level, (2) blocking primers for dominant species, and (3) deep sequencing.

It is well known that the use of both biological and technical replicates can largely reduce the rate of false negatives. Biological replicates refer to different biological samples, such as plankton samples, collected from the same sampling site, while technical replicates refer to repeated measurements from the same biological sample. Under the HTS context, technical replicates usually refer to multiple DNA extractions and PCRs from the same biological sample for downstream analyses. The use of biological replicates can increase the possibility of rare species collection from sampling sites, while the use of technical replicates can largely improve the detection probability of rare taxa in biological samples (see below the replication level recommended by Ficetola et al. 2015). Moreover, in order to avoid biased amplification of rare taxa, usually multiple PCR replicates (such as eight replicates) were used for a single DNA isolate (Zhan et al. 2013, 2014a). Therefore, we may largely decrease false negatives by increasing the biological and technical replicates and the ratio of technical/biological replicates.

Deep investigation showed interesting results that the replication level was strongly dependent on the detection probability of taxa (Ficetola et al. 2015). To limit the issue of false negatives, ecologists developed several models, one of which targeted on rare taxa is species occupancy models (SOM, MacKenzie et al. 2002). The use of SOM can successfully estimate true prevalence, detection probability, and false-positive rates, and the performance of SOM was largely improved with higher numbers of replicates, for example eight technical replicates for the detection of rare taxa (Ficetola et al. 2015). Therefore, the replication level should be determined based on sample characteristics, and the commonly used three replicates could not be enough to rule out false negatives. Consequently, as suggested by Ficetola et al. (2015), we need to run SOM to assess the detection probability and error rate of false negatives and to evaluate whether the replication level is good enough to control false negatives.

Usually, aquatic communities are dominated by several species, while some species remain low abundance and many are extremely rare (e.g., Zhan et al. 2014c). Many studies based on environmental samples have clearly shown that the use of blocking primers for abundant taxa can largely increase the detection probability for rare taxa (e.g., Vestheim and Jarman 2008, Boessenkool et al. 2012). Among several strategies for designing blocking primers, the use of annealing inhibiting primers has proved to be the most efficient based on a diet study in the Antarctic krill (Vestheim and Jarman 2008). However, it should be noted that the use of blocking primers usually is community specific and therefore involves prior knowledge and time to establish protocols. As the sequencing technologies are developing extremely fast, resulting in much deeper sequencing of a sample with less effort and cost, the use of deep sequencing can partially solve both technical problems associated with the development of blocking primers and the false negatives derived from the random sampling of amplicons in the sequencing procedure.

In addition, studies showed that NIS had caused varied influence on different types of habitats. For example, as a result of intense shipping and opening of new transport routes, brackish habitats have been increasingly influenced by NIS (Paavola et al. 2005). Coastal zones, especially those suitable for aquaculture, have been largely occupied by intensive aquaculture activities (Lin et al. 2015). Both the number of used introduced species and the aquaculture area have been increasingly extremely fast in the past two decades, especially in developing countries (see statistics in Lin et al. 2015). Studies showed that the number of NIS detected in these habitats was sampling dependent, that is, more intensive sampling effort leads to a larger number of detected NIS (e.g., Paavola et al. 2005). Consequently, in order to avoid false negatives, habitat-specific sampling design should be performed, and more factors such as environmental variables including salinity, water currents and local community composition should be fully considered (e.g., Paavola et al. 2005; Zhan and MacIsaac 2015).

Limited detected efficiency of genetic markers

For all metabarcoding-based studies, the selection of genetic markers and associated versatile primers with high taxonomic resolution power represents the first and also the most important step (Blaalid et al. 2014; Zhan et al. 2014c; Zhan and MacIsaac 2015). The versatile primers here refer to PCR primers that can systematically amplify all species of interest in a community. In order to get such robust primers, conserved regions of different genes among diverse taxonomic groups are surveyed to locate versatile primers using bioinformatic analyses of available sequences, while hypervariable regions between primer annealing sites are

also identified to ensure the high-resolution power for species delimitation across a wide range of taxa (Zhan et al. 2014c). Indeed, such “perfect” primers do not naturally exist, and selected primers, even after detailed evaluation, can only satisfy several of many expectations, that is, the selected primers have their inherent advantages and disadvantages (Table 2). In summary, the lack of versatile primers with high taxonomic resolution power can lead to three major categories of false-negative errors—(1) low resolution for species delimitation, (2) low amplification efficiency (and/or biomass below detection thresholds), and (3) biased amplification toward certain taxa (Table 2).

As evolutionary histories of selected genes may be different among taxonomic groups, the genes selected for metabarcoding studies can have different resolution power in taxa of interest. For example, although the mitochondrial cytochrome *c* oxidase subunit I gene (COI) has overall greater species-level resolution power (i.e., larger genetic divergence among related species) when compared to the nuclear small subunit ribosomal RNA gene (also known as SSU, 18S in eukaryotes; Tang et al. 2012), low resolution was reported at the COI gene, such as two fish genera—*Trachurus* and *Thunnus* (Keskin and Atar 2013). As many ascidian species are the most notorious invaders in marine ecosystems, we surveyed the resolution power of the commonly used V4 region of SSU based on all available sequences in NCBI GenBank (Table 3). Our results showed that genetic divergence largely varies among genus: >5 % in all species pairs in the genera of *Ciona* and *Didemnum* but <1 % in many genera such as *Ascidella*, *Botrylloides*, and *Botryllus* (Table 3). The low taxonomic resolution power can lead to multiple related species grouped into single operational taxonomic units (OTUs) when the OTU grouping strategy is used at a given divergence threshold such as the commonly used 3 %. However, when the OTU-free strategy is used (i.e., sequences directly subjected for BLAST), the low resolution for species delimitation in these genera can lead to misidentification: The best hits will be suggested based on taxon availability in public databases, leading to false negatives for unknown species, especially for rare species such as newly introduced NIS.

The polymorphism level among genes, even among different regions of the same genes, largely varies across taxonomic groups (e.g., see Fig. 1 in Zhan et al. 2014c). Consequently, the mismatch, as well as its degree, at the primer annealing sites largely affects PCR amplification efficiency. For example, when compared with SSU, the COI gene has less power in biodiversity recovery in plankton communities (Deagle et al. 2014; Zhan et al. 2014c), mainly because the COI gene does not contain suitably conserved regions for designing versatile primers for a wide range of taxa (Tang et al. 2012; Deagle et al. 2014). A test on performance of genetic markers and their corresponding primers

Table 2 A list of commonly used genes for metabarcoding studies, as well as their major known advantages and disadvantages

Gene	Major known advantages	Major known disadvantages	Major detected/possible errors	References
COI	Relatively high taxonomic resolution power; Large public reference databases	Limited biodiversity recovery capacity due to the lack of versatile primers attributed from high polymorphism along the COI gene	Fail to detect species of interest, especially low-abundance species in communities	Tang et al. (2012), Deagle et al. (2014), Zhan et al. (2014c)
16S	Relatively high taxonomic resolution power; Relatively easy to design versatile primers	Limited biodiversity recovery capacity	Fail to detect species of interest, especially low-abundance species in communities	Zhan et al. (2014c)
18S	Easy to design versatile primers; Wide taxonomy coverage; High sensitivity for rare species detection	Relative low-resolution power at the species level	Misidentify species in communities	Tang et al. (2012), Zhan et al. (2014c)
ITS	Easy to design versatile primers; Wide range coverage of taxonomic groups; Relative high taxonomic resolution power	Difficult to assign sequences to taxa by sequence similarity owing to frequent and/or large insertions/deletions	Misidentify species in communities due to the lack of barcoding gaps; Fail to detect species of interest owing to high intraspecific divergence	Schoch et al. (2012), Machida and Knowlton (2012)

Owing to different evolutionary histories of genes (polymorphisms and divergences at the inter- and intraspecific levels) in plants and fungi, we focus on animals here. COI = mitochondrial cytochrome *c* oxidase subunit I; 16S = mitochondrial 16S rRNA gene; 18S = nuclear small subunit ribosomal RNA gene (SSU, 18S in eukaryotes); ITS = nuclear internal transcribed spacer

Table 3 Genetic divergence at the commonly used V4 region of the nuclear small subunit ribosomal RNA gene (also known as SSU, 18S in eukaryotes) in a representative class of Ascidiacea

Order	Family	Genus	No. of species	Genetic divergence range (%)	Average $\pm$ SD (%)
Enterogona	Asciidiidae	<i>Ascidia</i>	3	2.25–2.92	2.62 $\pm$ 0.34
		<i>Ascidiella</i>	2	0	–
		<i>Phallusia</i>	3	1.30–2.60	0.89 $\pm$ 0.66
	Cionidae	<i>Ciona</i>	2	5.86	–
	Corellidae	<i>Chelyosoma</i>	2	1.52	–
		<i>Corella</i>	3	2.46–5.90	4.09 $\pm$ 1.72
	Didemnidae	<i>Didemnum</i>	3	5.86–13.18	10.24 $\pm$ 3.86
		<i>Diplosoma</i>	11	0–4.40	1.53 $\pm$ 1.21
		<i>Lissoclinum</i>	6	0–32.14	11.17 $\pm$ 15.17
	Perophoridae	<i>Ecteinascidia</i>	2	2.00	–
<i>Perophora</i>		2	0.65	–	
Stolidobranchia	Molgulidae	<i>Molgula</i>	14	0–13.77	6.71 $\pm$ 3.78
	Pyuridae	<i>Boltenia</i>	2	0.65	–
		<i>Halocynthia</i>	5	0–1.09	0.57 $\pm$ 0.48
		<i>Herdmania</i>	5	0.22–4.30	2.18 $\pm$ 1.31
		<i>Microcosmus</i>	5	0.65–3.76	2.28 $\pm$ 1.13
		<i>Pyura</i>	15	0–3.78	2.00 $\pm$ 1.10
		Styelidae	<i>Asterocarpa</i>	2	0.50
	<i>Botrylloides</i>		2	0.22	–
	<i>Botryllus</i>		2	0.65	–
	<i>Cnemidocarpa</i>		2	0.25	–
	<i>Dendrodoa</i>		2	0.25	–
	<i>Polyandrocarpa</i>		3	2.02–3.69	3.03 $\pm$ 0.89
	<i>Polycarpa</i>		5	0–3.69	2.32 $\pm$ 1.26
	<i>Styela</i>		5	0.22–2.46	1.22 $\pm$ 0.76
	<i>Symplegma</i>		3	0.23–2.04	1.34 $\pm$ 0.97

All available sequences within each genus were downloaded from NCBI GenBank. The results showed that resolution power for species delimitation largely varied among genus. For example, the genetic divergence was $>5\%$ in all species pairs in the genera of *Ciona* and *Didemnum* but $<1\%$ in many genera such as *Ascidiella*, *Botrylloides*, and *Botryllus*

in a plankton community showed that 38 versus two orders were recovered based on primers for the SSU and COI genes, respectively (Zhan et al. 2014c). The difference in biodiversity recovery was also observed in different regions of SSU, such as V4 versus V5–7 (Zhan et al. 2014c). A large number of taxa, particularly those rare ones, were not recovered from the examined communities when using primers with low amplification efficiency, leading to a high rate of false negative errors. Moreover, although the well-evaluated primers of SSU can detect extremely rare species (down to $2.3 \times 10^{-5}\%$ of sample biomass; Zhan et al. 2013), false negative errors were still detected when the biomass of species became much rarer (e.g., $\times 10^{-6}\%$ of sample biomass; Table 2). Despite the fact that all methods have their own detection thresholds, and the reported detection threshold (i.e., $2.3 \times 10^{-5}\%$ of sample biomass) based on SSU-metabarcoding may represent the most sensitive one, low amplification efficiency (and/or biomass below

detection thresholds) is still one of the most crucial error sources for false negatives, especially when using environmental DNA (i.e., eDNA) where trace of DNA is left in environments by rare species.

In addition to low resolution and low amplification efficiency, another important error source for false negatives is biased PCR amplification (i.e., over-amplification of some taxa but non-amplification of some others). This error source has been widely reported in both laboratory- 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). Many factors can lead to biased amplification in metabarcoding surveys, including mismatch in primer annealing sites (i.e., universality of the used primers), length of amplified non-coding regions among taxa (i.e., insertions/deletions), and taxonomic composition of communities of interest (Bellemain et al. 2010; Engelbrektson et al. 2010). Interestingly,

biased amplification occurs in technical replicates based on the exactly same communities when using the same protocols throughout the whole HTS procedure (Fig. 2), suggesting that random sampling events such as the random picking of DNA fragments in each round of PCR can potentially contribute to biased amplification.

Several technical improvements and/or tests could decrease the error rate of false negatives derived from the limited detection efficiency of selected markers: (1) development of study-specific versatile primers with high taxonomic resolution power. Obviously, versatile primers with high taxonomic resolution power are still lacking and more efforts are needed to develop and then characterize such a type of primers (Zhan and MacIsaac 2015). As “perfect” primers do not exist across taxa and the performance for biodiversity recovery largely varies among primers (e.g., Zhan et al. 2014c), it is necessary to select candidate genes and design study-specific primers based on research aims/goals. (2) The use of two-step strategy, that is, survey first using relatively versatile primers such as those derived from SSU to get an overview of biodiversity and then further investigate by using/designing robust primers for narrower taxa of interest. Given that it is difficult to design highly versatile primers for fast-evolving genes such as the COI gene and primers derived from some relatively conserved genes such as SSU can amplify a wide range of taxa (e.g., Zhan et al. 2014c), we can still use available primers based on these relatively conserved genes to recover biodiversity at higher ranks such as genus or family levels. Subsequently, primers with high taxonomic resolution power can then be more easily designed for specific narrower taxa based on fast-evolving genes for species delimitation. As the speed of pipelines for BLASTing large datasets has been highly improved, the OTU-free strategy (i.e., sequences directly subjected for BLAST) is recommended for species annotation when primers with low taxonomic resolution are used. (3) Detailed tests on amplification efficiency for selected primers such as universality across a range of taxonomic groups, biased amplification, and capacity for the recovery of multiple taxa. After candidate genes and associated primers are selected, all parameters of concern based on research aims and goals should be tested to avoid a high rate of false negatives. (4) The use of multiple primer pairs for rare NIS detection. The amplification and sequencing of a DNA library derived from a single marker unlikely recover all biodiversity in a community (e.g., Creer et al. 2010; Zhan et al. 2014c). The use of multiple primer pairs, especially those derived from different genes, can increase the detection probability of rare NIS. Multiple datasets can be cross-referenced to potentially correct false negatives. However, it should be clarified that cross-referencing among multiple datasets can be complicated, mainly owing to incomplete reference libraries (see below the section for more detail),

Fig. 2 Distribution of sequence reads for each shared operational taxonomic unit (OTU) derived from the two replicates of a plankton community collected from the Nanaimo Harbour, British Columbia, Canada. The two replicates were made based on the exactly same communities and protocols throughout the whole HTS procedure. The data is shown by the natural log (ln)-transformed ratio between sequence percentage of each OTU in Fraction I and Fraction II. The ln-transformed ratio is expected to be 0 (dotted line) if the distribution of sequence reads for each OTU matches perfectly in the two fractions (i.e., no biased amplification). The results here suggest that biased amplification occurs even in replicates prepared using the exactly same communities and protocols. Data are derived from Zhan et al. (2014a) and Sun et al. (2015)

poor linkage and crossing-reference among genes used, and even poorer taxonomic assignments at multiple loci (Zhan and MacIsaac 2015).

Poor reference libraries

Despite the fact that the rapid development of numerous DNA barcoding projects has populated large datasets in public databases such as GenBank and the Barcode of Life Database (BOLD, Ratnasingham and Hebert 2007), the availability of reference sequences in these public genetic databases largely varies among taxonomic groups (Briski et al. 2011). For example, only 3.5 % described Copepoda, one of the dominant groups in marine communities, has been sequenced at the COI gene or small subunit ribosomal 16S rRNA (16S) gene in the Barcode of Life Database (BOLD) or GenBank; however, these two public databases cover 54 % of known Branchiopoda species (Briski et al. 2011). Indeed, the taxon bias and the lack of available sequences in public databases are particularly acute at the community level. Among 353 OTUs derived from the high-throughput sequencing of a plankton community collected from the Hamilton Harbour (Ontario, Canada), only 32 (9.1 %) showed significant hits to available sequences in public databases when using the 99 % similarity (Zhan et al. 2014c).

Comprehensive reference libraries have become valuable resources for a more extensive and rigorous documentation of local biodiversity and identification of NIS.

Fortunately, efforts have been made to construct reference libraries (e.g., Lobo et al. 2015; Kneibelsberger et al. 2014; Zhou et al. 2011). However, broad collaborations are largely required to develop community-level reference libraries as different research groups have their own taxa of interest.

Raw data filtering

Available evidence clearly showed that errors/artifacts derived from sequencing are a major cause of inflated biodiversity estimates (e.g., Quince et al. 2009; Kunin et al. 2010). Particularly, errors are prone to appear in low-abundance sequence reads such as singletons, doubletons, and tripletons (Reeder and Knight 2009; Huse et al. 2010; Tedersoo et al. 2010). Consequently, the filtering of raw sequence reads represents one of the most important steps for removing sequences containing errors, and almost all studies use data filtering as a error control strategy (see Flynn et al. 2015 and references therein). However, as the general trend of positive correlation exists between the sequence abundance and species abundance (Sun et al. 2015), rare species are usually recovered by low-abundance sequences (Brown et al. 2014; Zhan et al. 2014a). Owing to the low number of sequences for each rare taxon, any nucleotide ambiguity and/or low-quality nucleotide in such low-abundance reads can result in complete removal of these sequences during quality filtering. The erroneous removal of real rare species during data filtering has been clearly detected by using both internal and external controls in marine plankton communities (Zhan et al. 2014b). Unfortunately, the erroneous removal occurred in all tested filtering stringencies, even at very low stringencies such as Q (Phred score) <10. In addition, rarer species were erroneously deleted as filtering stringencies increased and the probability of erroneous deletion increased as the biomass decreased (Zhan et al. 2014a).

Data filtering has become a standard procedure in data processing (see Flynn et al. 2015 and references therein). However, as data filtering is performed at the beginning phase of data processing, it is easily overlooked as a source responsible for false negatives, especially for the detection of rare species such as newly introduced NIS. This is no question that data filtering largely improves the accuracy of diversity estimates (see Flynn et al. 2015 and references therein); however, we must realize that data filtering can erroneously remove a large proportion of rare species. Therefore, raw data, rather than “standard protocols,” should be directly subjected for BLAST or other analyses for the detection of newly introduced NIS to avoid false negatives. In addition, there is an urgent need to develop powerful bioinformatic pipelines to allow accurate sorting of real rare taxa out of errors/artifacts. The development

and use of more expansive reference libraries will allow better testing of bioinformatic pipelines, permit better filtering of errors/artifacts, and assure accurate rare species assignment (Zhan and MacIsaac 2015).

False positive

Cross-contamination

The early detection of NIS mainly focuses on environmental and/or impure samples (e.g., net-towed samples). DNA isolated from such types of samples may have contaminants derived from multiple sources, such as fish tissues from food processing plants and restaurants and litters from tourists. Consequently, these contaminants can be identified as newly introduced NIS after sequencing, leading to false positives in detection programs. In addition, as HTS has high sequencing capacities, thus allowing many samples (e.g., up to several hundred) pooled together for parallel sequencing. Cross-contamination can occur during sample preparation for such a large number of samples in a short period of time in the same laboratories. As HTS-based methods are extremely sensitive, a trace of contaminants can be recovered as possible NIS after sequencing. Owing to these reasons, a study demonstrated that possible contamination occurred in 71.8 % metagenomes, with as high as 64 % contaminating sequences (Schmieder and Edwards 2011).

Despite the fact that it is extremely difficult to control, detect, or eliminate cross-contamination when various types of samples such as those from environments are processed in a short period of time in the same laboratories, careful manipulation, good organization, and the use of strictly standard experimental protocols may reduce false positives. In addition, a comprehensive survey and knowledge of contamination sources such as locations and types of food processing plants and long-term surveillance on well-designated areas may help rule out false negatives caused by human activities.

Tag switching (also known as tag jumping)

When multiple samples are pooled together for parallel sequencing, tags, also known as molecular identifiers (MIDs) which usually consist of 6–10 bp, are uniquely added to each sample when preparing samples for sequencing. As the unique tag is uniquely designed and added to a specific sample, sequences can be bioinformatically traced back to the samples from which they originated. However, this efficient method is a largely over-looked source of errors—unique tags may switch among pooled samples, resulting in the identification of a species not present in

Table 4 Genetic divergence (%) at the commonly used V4 region of the nuclear small subunit ribosomal RNA gene (also known as SSU, 18S in eukaryotes) in nine representative species of the genus *Molgula* (Asciacea: Stolidobranchia: Molgulidae)

	<i>M. arenata</i>	<i>M. bleizi</i>	<i>M. citrina</i>	<i>M. complanata</i>	<i>M. manhattensis</i>	<i>M. oculata</i>	<i>M. pacifica</i>	<i>M. provisionalis</i>
<i>M. bleizi</i>	5.29							
<i>M. citrina</i>	3.70	3.17						
<i>M. complanata</i>	2.12	5.82	4.76					
<i>M. manhattensis</i>	1.06	5.29	3.70	2.12				
<i>M. oculata</i>	3.70	2.12	1.59	4.76	3.70			
<i>M. pacifica</i>	3.70	1.59	1.59	4.76	3.70	0.53		
<i>M. provisionalis</i>	0.00	5.29	3.70	2.12	1.06	3.70	3.70	
<i>M. retortiformis</i>	1.06	5.29	3.70	2.12	0.00	3.70	3.70	1.06

The names of invasive species are bolded

a community (i.e., false positives; see Carlsen et al. 2012 and references therein). Tag switching is detected when mismatch occurs between forward primer tags and reverse primer tags in sequences. Consequently, when using only one tagged primer (i.e., only forward or reverse primer is tagged) or sequencing long amplicons (i.e., amplicons are larger than sequencing length capacity), tag switching cannot be detected. Consequently, the report on errors caused by tag switching is rare in metabarcoding studies. Indeed, tag switching is common in multiple sequencing platforms such as Roche/454 (e.g., Blaaid et al. 2014; Botnen et al. 2014) and Illumina sequencing platforms (e.g., Esling et al. 2015; Schnell et al. 2015), and even in non-metabarcoding Illumina-based studies (e.g., Kircher et al. 2011). Studies showed that up to 16 % and more than 2 % sequence reads had non-compatible tag combinations based on Roche/454 and Illumina sequencing platforms (van Orsouw et al. 2007; Schnell et al. 2015). Due to tag switching, Carew et al. (2013) found that all species incorrectly identified in a study site were already in the experiment.

For the major causes of tag switching, some are common among sequencing platforms, while others are platform specific. Among the common ones, the following three are well documented—(1) low concentrations of unused tagged primers may interfere with the amplicons during sequencing procedures to complete tag switching (Carlsen et al. 2012); (2) sporadic cross-contamination of primers carrying different tags occurs during primer synthesis or handling (Kircher et al. 2011); (3) chimera sequences are formed during PCR (Kircher et al. 2011). For the platform-specific, taking the popularly used Illumina sequencing platform for example, the two, including tag switching caused by T4 DNA polymerase activity during the blunt-end step of the library preparation and mixed clusters on the Illumina flowcell, have been suggested as the mechanisms for tag switching (van Orsouw et al. 2007; Kircher et al. 2011; Schnell et al. 2015).

As suggested by Schnell et al. (2015), the use of the following five measures can decrease the error rate during experimental design: (1) using matching tags; (2) doing PCR replicates; (3) minimizing and only carefully handling tagged amplicons; (4) incorporating negative controls at all steps and sequencing a subset, even if negative; and (5) using a tag only once in each library. In addition, robust chimera detection pipelines should be used to remove chimera sequences with tag switching. Amplicons should be tagged at both ends in order to detect and remove sequences with tag switching (Carlsen et al. 2012). 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).

Low-resolution power of genetic markers

In addition to false negatives caused by low-resolution power of genetic markers for species delimitation, it can cause false positives as well. For example, when we surveyed genetic divergence of the commonly used V4 region of SSU in nine representative species of the genus *Molgula* (Asciacea: Stolidobranchia: Molgulidae), two species pairs, *M. provisionalis*–*M. arenata* and *M. retortiformis*–*M. manhattensis*, showed the identical sequences (i.e., no genetic divergence; Table 4). As *M. manhattensis* is an invasive species in many parts globally such as the Pacific coast of North America, Black Sea, and China (Zhan et al. 2015), when for example sequences of *M. retortiformis* are subjected for BLAST, the best hit of *M. manhattensis* may be suggested. The good knowledge of examined taxa and the genetic markers used, as well as the use of genetic markers with high taxonomic resolution power for related species delimitation, may solve the false positives here.

Application cases

Obviously, current biosecurity and management programs are not sufficient to provide effective monitoring methods for the early detection of NIS, particularly in marine and coastal ecosystems (e.g., Pochon et al. 2013; Zhan and MacIsaac 2015). There is no doubt that molecular methods have the potential to dramatically improve our capacity for the early detection and monitoring of NIS (Darling and Mahon 2011; Zhan and MacIsaac 2015). Fortunately, several countries, including Canada, New Zealand, and Australia, have adopted molecular methods to support management decisions (Darling and Mahon 2011 and references therein).

Based on HTS-based methods, recent studies have successfully identified rare NIS from both artificially assembled and environmental samples. After spiking known NIS into natural complex communities, NIS could be successfully recovered when their biomass abundance was $>0.64\%$ in Pochon et al.'s study (Pochon et al. 2013) and $2.3 \times 10^{-5}\%$ in Zhan et al.'s study (Zhan et al. 2014a). The difference on detection sensitivity between these studies is mainly derived from PCR primers, as different primers have varied amplification efficiency (Zhan and MacIsaac 2015). The high detection sensitivity has largely increased the detection probability; for example, when more than one larva of a highly invasive species, golden mussel *Limnoperna fortunei*, occurred in plankton samples, the detection probability was 100 % (Zhan et al. 2014a). In contrast, typical sampling protocols have a low probability (<0.2) of detecting the species unless population density is high (Harvey et al. 2009). Using environmental samples, Zaiko et al. (2015) identified four NIS from plankton samples collected from southeastern Baltic Sea coastal zones based on HTS metabarcoding; we found 24 NIS in a comprehensive survey on 16 major Canadian ports, 11 of which were detected in previously unreported locations (unpublished data). It should be noted that the early detection of marine NIS based on HTS is at the infant stage; however, available application cases above clearly show that HTS-based methods have tremendous potential for the early detection of marine NIS. Therefore, HTS-based early detection is expected to largely adopt and use for science-aided management of marine NIS.

Although various impediments occur in the way of wide application of HTS-based methods for early detection (Zhan and MacIsaac 2015), deep understanding of technical challenges, especially development of strategies to remove and/or avoid errors, can allow us to develop efficient methods for rapid detection of new incursions of marine pests. Meanwhile, such methods can aid us to take appropriate remedial actions to eradicate incursions before

new NIS spread in "open" waters. In addition, HTS-based methods can allow us to deeply understand biological and ecological causes and consequences on how rare species, including both native and non-native ones, can survive at low population density and finally become abundant in marine and coastal ecosystems.

Conclusions

As the introduction and spread of NIS, as well as associated negative economic and ecological effects, is increasing globally in marine and coastal ecosystems, there is an urgent need to establish functional and efficient management actions. Successful early detection in management programs may have a higher chance to eradicate NIS and also to prevent further spread of NIS. Owing to the obvious advantages including high detection efficiency, accuracy, and sensitivity, HTS-based methods are expected to serve as robust and powerful tools for early detection and warning of newly introduced NIS in management programs. However, many technical issues can potentially lead to both false positives and false negatives. As genetic detection tools are being adopted in decision-making agents, the error sources discussed in the present study, as well as causes and consequences should be fully acknowledged and managed to correct and/or reduce errors in management programs.

Acknowledgments This work was supported by National Natural Science Foundation of China (31272665) and 100 Talents Program of the Chinese Academy of Sciences to A.Z.

References

- Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kause-
rud H (2010) ITS as an environmental DNA barcode for fungi:
an in silico approach reveals potential PCR biases. *BMC Micro-
biol* 10:189
- Blaalid R, Davey ML, Kause-
rud H, Carlsen T, Halvorsen R, Høi-
land K, Eidesen PB (2014) Arctic root-associated fungal com-
munity composition reflects environmental filtering. *Mol Ecol*
23:649–659
- 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
- Botnen S, Vik U, Carlsen T et al (2014) Low host specificity of root-
associated fungi at an Arctic site. *Mol Ecol* 23:975–985
- Briski E, Cristescu ME, Bailey SA, MacIsaac HJ (2011) Use of DNA
barcoding to detect invertebrate invasive species from diapausing
eggs. *Biol Invasions* 13:1325–1340
- Brown SP, Veach AM, Rigdon-Huss AR, Grond K, Lickteig SK,
Lothamer K, Oliver AK, Jumpponen A (2014) Scraping the bot-
tom of the barrel: are rare high throughput sequences artifacts?
Fungal Ecol 13:221–225

- Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013) Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Front Zool* 10:45
- Carlsen T, Aas AB, Lindner D, Vra Istad T, Chumacher T, Kausrud 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
- Carlton JT, Gellar JB (1993) Ecological roulette: the global transport of nonindigenous marine organisms. *Science* 261:78–82
- Carugati L, Corinaldesi C, Dell'Anno A, Danovaro R (2015) Metagenetic tools for the census of marine meiofaunal biodiversity. *Mar Genomics* 24:11–20. doi:[10.1016/j.margen.2015.04.010](https://doi.org/10.1016/j.margen.2015.04.010)
- Clarke LJ, Soubrier J, Weyrich LS, Cooper A (2014) Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Mol Ecol Resour* 14:1160–1170
- Comtet T, Sandionigi A, Viard F, Casiraghi M (2015) DNA (meta) barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens. *Biol Invasions* 17:905–922
- Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Miné J, Arnaud-Haond S (2015) Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS ONE* 10(2):e0117562
- Creer S (2010) Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. *Mol Ecol* 19:2829–2831
- Creer S, Fonseca VG, Porazinska DL, Giblin-Davis M, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lamshead PJD, Thomas WK (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls, and promises. *Mol Ecol* 19:4–20
- 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
- Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biol Lett* 10:20140562
- 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
- Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res.* doi:[10.1093/nar/gkv107](https://doi.org/10.1093/nar/gkv107)
- Ficetola GF, Pansu J, Bonin A, Coissac E, Giguët-Covex C, De Barba M, Gielly L, Lopes CM, Boyer F, Pompanon F, Rayé G, Taberlet P (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol Ecol Resour* 15:543–556
- Finnoff D, Shogren JF, Leung B, Lodge D (2007) Take a risk: preferring prevention over control of biological invaders. *Ecol Econ* 62:216–222
- Flynn JM, Brown EA, Chain FJJ, MacIsaac HJ, Cristescu ME (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecol Evol* 5:2252–2266
- Harvey CT, Qureshi SA, MacIsaac HJ (2009) Detection of a colonizing, aquatic, non-indigenous species. *Divers Distrib* 15:429–437
- Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 12:1889–1898
- 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
- Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) “Sight-unseen” detection of rare aquatic species using environmental DNA. *Conserv Lett* 4:150–157
- Keskin E, Atar HH (2013) DNA barcoding commercially important fish species of Turkey. *Mol Ecol Resour* 13:788–797
- Kircher M, Sawyer S, Meyer M (2011) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* doi:[10.1093/nar/gkr771](https://doi.org/10.1093/nar/gkr771)
- Kneibelsberger T, Landi M, Neumann H, Kloppmann M, Sell AF, Campbell PD, Campbell PD, Laakmann S, Raupach MJ, Carvalho GR, Costa FO (2014) A reliable DNA barcode reference library for the identification of the North European shelf fish fauna. *Mol Ecol Resour* 14:1060–1071
- 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
- Lawler JJ, Aukema JE, Grant JB, Halpern BS, Kareiva P, Nelson CR, Ohleth K, Olden JD, Schlaepfer MA, Silliman BR, Zaradic P (2006) Conservation science: a 20-year report card. *Front Ecol Environ* 4:473–480
- Lin Y, Gao Z, Zhan A (2015) Introduction and use of non-native species for aquaculture in China: status, risks and management solutions. *Rev Aquac* 7:28–58
- 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
- Lobo J, Teixeira MAL, Borges LMS, Ferreira MSG, Hollatz C, Gomes PT, Sousa R, Ravara A, Costa MH, Costa FO (2015) Starting a DNA barcode reference library for shallow water polychaetes from the southern European Atlantic coast. *Mol Ecol Resour.* doi:[10.1111/1755-0998.12441](https://doi.org/10.1111/1755-0998.12441)
- Lodge DM, Williams S, MacIsaac HJ, Hayes K, Leung B, Reichard S, Mack RN, Moyle PB, Smith M, Andow DA, Carlton JT, McMichael A (2006) Biological invasions: recommendations for U.S. policy and management. *Ecol Appl* 16:2035–2054
- Lodge DM, Turner CR, Jerde CL, Barnes MA, Chadderton L, Egan SP, Feder JL, Mahon AR, Pfrender ME (2012) Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. *Mol Ecol* 21:2555–2558
- Machida RJ, Knowlton N (2012) PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS One* 7:e46180
- MacKenzie DIM, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83:2248–2255
- Mackie JA, Geller J (2010) Experimental parameters affecting quantitative PCR of *Artemia franciscana*: a model for a marine zooplankton target in natural plankton samples. *Limnol Oceanogr Methods* 8:337–347
- Mahon AR, Barnes MA, Senapati S, Feder JL, Darling JA, Chang H, Lodge DM (2011) Molecular detection of invasive species in heterogeneous mixtures using a microfluidic carbon nanotube platform. *PLoS ONE* 6:e17280. doi:[10.1371/journal.pone.0017280](https://doi.org/10.1371/journal.pone.0017280)
- McDonald LL (2004) Sampling rare populations. In: Thompson WL (ed) Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters. Island Press, New York, pp 11–42
- 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
- Paavola M, Olenin S, Leppakoski E (2005) Are invasive species most successful in habitats of low native species richness across European brackish water seas? *Estuar Coast Shelf Sci* 64:738–750

- 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
- Pompanon F, Samadi S (2015) Next generation sequencing for characterizing biodiversity promises and challenges. *Genetica* 143:133–138
- 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
- Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (www.barcodinglife.org). *Mol Ecol Notes* 7:355–364
- Reeder J, Knight R (2009) The ‘rare biosphere’: a reality check. *Nat Methods* 6:636–637
- Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288
- Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated—reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* 15:1289–1303
- 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 (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Mol Ecol Resour* 15:250–261
- Sun C, Zhao Y, Li H, Dong Y, MacIsaac HJ, Zhan A (2015) Unreliable quantitation of species abundance based on high-throughput sequencing data of zooplankton communities. *Aquat Biol* 24:9–15
- Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. *Mol Ecol* 21:1789–1793
- Tang CQ, Leasi F, Obertegger U, Kienke A, Barraclough TG, Fontaneto D (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proc Natl Acad Sci USA* 109:16208–16212
- Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytol* 188:291–301
- 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, Hvd Poel, Jv Oeveren, Verstege H, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJT (2007) Complexity reduction of polymorphic sequences (CRoPSTM): 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
- Vander Zanden MJ, Hansen GJA, Higgins SN, Kornis MS (2010) A pound of prevention, plus a pound of cure: early detection and eradication of invasive species in the Laurentian Great Lakes. *J Great Lakes Res* 36:199–205
- Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples—a case study on prey DNA in Antarctic krill stomachs. *Front Zool* 5:12
- Williamson M, Fitter A (1996) The varying success of invaders. *Ecology* 77:1661–1666
- Zaiko A, Martinez JL, Schmidt-Petersen J, Ribicic D, Samuiloviene A, Garcia-Vazquez E (2015) Metabarcoding approach for the ballast water surveillance—an advantageous solution or an awkward challenge? *Mar Pollut Bull* 92:25–34
- Zhan A, MacIsaac HJ (2015) Rare biosphere exploration using high-throughput sequencing: research progress and perspectives. *Conserv Genet* 16:513–522
- 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, Xiong W, He S, MacIsaac HJ (2014b) Influence of artifact removal on rare species recovery in natural complex communities using high-throughput sequencing. *PLoS ONE* 9:e96928
- Zhan S, He EA, Brown FJJ, Chain TW, Therriault CL, Abbott DD, Heath ME, MacIsaac Cristescu HJ (2014c) Reproducibility of pyrosequencing data for biodiversity assessment in complex communities. *Methods Ecol Evol* 5:881–890
- Zhan A, Briski E, Bock DG, Ghabooli S, MacIsaac HJ (2015) Ascidiaceans as models for studying invasion success. *Mar Biol* 162:2449–2470
- 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 J, Jiang Y, Deng Y, Shi Z, Zhou B, Xue K, Wu L, He Z, Yang Y (2013a) Random sampling process leads to overestimation of biodiversity of microbial communities. *mBio* 4:324. doi:10.1128/mBio.00324-13
- Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013b) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* 2:4