

UNCORRECTED PROOF

Research Article

Intact vs. homogenized subsampling: Testing impacts of pre-extraction processing of multi-species samples on invasive species detectionKatrina M. Pagenkopp Lohan^{1,*}, Tracy L. Campbell², Jinchen Guo², Melinda Wheelock², Ruth A. DiMaria¹ and Jonathan B. Geller²¹Smithsonian Environmental Research Center, Edgewater, MD 21037, USA²Moss Landing Marine Laboratories, Moss Landing, California 95039, USAAuthor e-mails: lohank@si.edu (KMPL), campbell.tracyl@gmail.com (TLC), jguo@mlml.calstate.edu (JG), mwheelock@mlml.calstate.edu (MW), dimariar@si.edu (RAD), geller@mlml.calstate.edu (JBG)

*Corresponding author

Citation: Pagenkopp Lohan KM, Campbell TL, Guo J, Wheelock M, DiMaria RA, Geller JB (2019) Intact vs. homogenized subsampling: Testing impacts of pre-extraction processing of multi-species samples on invasive species detection. *Management of Biological Invasions* 10 (in press)

Received: 9 August 2018**Accepted:** 21 October 2018**Published:** xx xxxxx 2019**Handling editor:** Joana Dias**Thematic editor:** Katherine Dafforn**Copyright:** © Pagenkopp Lohan et al.This an open access article published under terms of the Creative Commons Attribution License ([Attribution 4.0 International - CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

OPEN ACCESS

Abstract

With the increasing use of metagenetics for invasive species monitoring in aquatic habitats, fully assessing sample pre-processing is essential for appropriate data interpretation. This includes the impact of subsampling before DNA extraction, a common practice for processing zooplankton and other mixed samples where the volume of the sample exceeds the maximum allowed in a commercial DNA extraction kit. Our goal for this study was to assess the impact of these pre-processing methods on detecting invasive species through 1) examining operational taxonomic unit (OTU) richness assessments and 2) detecting low abundance OTUs in mixed samples. We used tagged amplicon high throughput sequencing (HTS) to amplify the mitochondrial cytochrome oxidase I (COI) gene from zooplankton samples collected from multiple bays. We then used a step-wise approach to assess the impacts of replication and homogenizing DNA extracts on the resulting OTU richness. We compared the results from 1) a single, intact subsample, 2) multiple, intact subsample replicates, 3) a single, homogenized subsample, or 4) multiple, homogenized subsample replicates. We expected OTU richness to be highest for multiple subsamples and for those homogenized, with subsequent impacts on rare OTU detection. Our results showed that homogenizing a sample prior to DNA extraction increased the number of OTUs recovered, particularly low abundance OTUs. Additionally, processing a greater number of extraction replicates increased the OTU richness of each sample, regardless of the extraction pre-processing conducted. However, with the increasing OTUs, more sequences per replicate were required to detect all OTUs. Thus, when detection of low abundance OTUs is a goal, particularly for early detection of invasive species, homogenizing multi-species samples is recommended due to the clear increase in rare OTUs detected.

Key words: homogenization, metagenetic, non-native, non-indigenous, replication, zooplankton**Introduction**

High throughput sequencing (HTS), particularly metagenetics or metabarcoding, is a popular tool for assessing aquatic biodiversity (Valentini et al. 2016; Borrell et al. 2017; Ransome et al. 2017). Metagenetics, a technique that allows for community level assessments of multispecies

samples through the amplification of a single locus (Taberlet et al. 2012), is increasingly being used for invasive species detection (Xiong et al. 2016; Borrell et al. 2017). Given its high sensitivity for detecting low abundance or rare taxa (Zhan et al. 2013), this method has great appeal for early detection of aquatic invasive species (Xiong et al. 2016), an essential step to prevent establishment of nuisance taxa.

Previous studies have assessed the usefulness of metagenetics for aquatic invasive species monitoring in multi-species samples. In order to detect non-native species entering coastal waters via maritime trade, researchers have used metagenetics to determine the bacterial (Lymperopoulou and Dobbs 2017), protistan (Pagenkopp Lohan et al. 2016, 2017), and metazoan (Zaiko et al. 2015) communities in ballast water as well as the metazoan and protistan communities in the bilge water of recreational boats (Fletcher et al. 2017; Pochon et al. 2017). Multiple studies have shown this technique can also detect non-native species in recipient waters (Abad et al. 2016; Chain et al. 2016; Borrell et al. 2017). For example, Abad et al. (2016) compared the results obtained from metagenetics with those obtained from traditional microscopic analysis of zooplankton samples. They found detection of low abundance organisms was more likely with metagenetics, but was limited by the incompleteness of the sequence databases required for identification. These studies demonstrate the high sensitivity of metagenetics, particularly for detecting low abundance species (Hajibabaei et al. 2011; Zhan et al. 2013; Pochon et al. 2013; Abad et al. 2016), a requirement for identifying early introductions of non-native species.

Searches for non-native species are often conducted within unsorted environmental samples, which generally include a wide-ranging mixture of organisms (e.g., zooplankton, phytoplankton, bacteria) and sometimes contain chemical PCR inhibitors. Thus, researchers have recognized the importance of determining the impacts of different techniques used in sample processing on the subsequent data, to uncover the best ways to reduce bias and inhibition while maximizing the likelihood of detecting low abundance organisms. Zhan et al. (2013) used biological replicates from zooplankton samples to show that there was high reproducibility among common OTUs, but lower reproducibility for low abundance OTUs. These results indicated that sequence depth of samples should be carefully considered and that replicates are needed to assess total diversity. Ransome et al. (2017) assessed the impact of multiple sample preservation methods and field-processing techniques on biodiversity assessments associated with metabarcoding samples from autonomous reef monitoring structures (ARMS). Their results demonstrated that these decisions are critical and can impact the diversity results generated. Some best practices for generating metagenetic libraries currently include careful selection of one or multiple primer sets depending on the research question and taxa

targeted (Freeland 2017), generating PCR replicates to decrease the influence of primer bias and binding stochasticity (Polz and Cavanaugh 1998; Alberdi et al. 2018), and maximizing the fragment length for taxonomic assignment (Wommack et al. 2008; Morales et al. 2009). Mock communities have proven very useful for microbial studies to ground truth bioinformatic analyses (Schloss et al. 2011; Lee et al. 2012; Bjørnsgaard Aas et al. 2017). However, this technique is not as useful for metazoans, which are ecosystems themselves and contain complex bacterial and protistan communities that inflate the number of OTUs well beyond the number expected (Geisen et al. 2015; Leray and Knowlton 2017). Thus, due to the complicated nature of the samples, the sensitivity and limitations of the metagenetics protocol, and the important need to rapidly and quickly identify newly arriving non-native species, it is imperative to examine all aspects of the laboratory workflow used to generate the sequencing libraries in order to maximize the likelihood of detecting low abundance organisms.

One aspect of sample processing that has not been as thoroughly assessed is the impact of pre-processing prior to subsampling for DNA extraction. This is a problem when the volume of biomass collected in a given sample is larger than the total volume that can be processed as a single sample using a commercial DNA extraction kit. There are multiple ways that researchers could approach subsampling, each with pros and cons. While researchers could collect smaller volumes, doing so could exclude a large amount of the present biodiversity (if a lower total volume is collected) or increase sample collection and processing time while simultaneously increasing the cost to process the additional replicates. Thus, if researchers want to collect large volume samples to screen a greater volume for invasive species in a particular area, they need to extract DNA from either 1) an intact or 2) a homogenized sample. In both cases, this can be done with a single or replicate subsamples, depending on the research question and available funds.

In this study, we assessed the impact of pre-extraction processing on 1) capturing OTU richness and 2) detecting low abundance OTUs to determine the impact of this step on non-native species detection. To do this, we designed a step-wise progression to optimize a standard method for processing zooplankton samples for metagenetic analyses with the following progression: 1) a single, intact subsample, 2) multiple, intact subsample replicates, 3) a single, homogenized subsample, or 4) multiple, homogenized subsample replicates. We expected that OTU richness per sample would be lowest for a single, intact subsample and highest after multiple, homogenized subsample replicates, with subsequent impacts on detection of rare OTUs. While we conducted this test with zooplankton samples, there are many other samples that can be utilized for detection of non-native species and these results are applicable to any multi-species sample where homogenization is possible (e.g., sediment, ARMS).

Table 1. For all samples included in the intact sample dataset (top) or homogenized sample dataset (bottom), the following calculations were determined using the replicates for each sample: OTU variance, the average number of OTUs, the standard deviation of OTUs across replicates, the coefficient of variation, the average difference in number of OTUs across replicates, and the percentage of OTUs included in a single replicate. Note that 586A and B were collected from a single sampling event, but were split upon collection due to the volume of the sample. Thus, they were treated as independent samples for the purpose of this study.

| Intact Dataset | OTU Variance | Average # of OTUs | Standard deviation | Coefficient of Variation | Average Difference | % OTUs in 1 replicate |
|---------------------|--------------|-------------------|--------------------|--------------------------|--------------------|-----------------------|
| 580 | 393.2 | 229.8 | 19.8 | 8.6 | 177.2 | 43.5 |
| 582 | 4,946.3 | 149.6 | 70.3 | 47.0 | 144.4 | 49.1 |
| 586A | 240.7 | 212.8 | 15.5 | 7.3 | 194.2 | 47.7 |
| 586B | 2,172.8 | 234.6 | 46.6 | 19.9 | 245.4 | 51.1 |
| 656 | 1,115.0 | 523.5 | 33.4 | 6.4 | 279.5 | 34.8 |
| 755 | 2,826.3 | 510.6 | 53.2 | 10.4 | 226.4 | 30.7 |
| Homogenized Dataset | | | | | | |
| 1359 | 880.3 | 594.3 | 29.7 | 5.0 | 35.7 | 5.7 |
| 1379 | 6,181.3 | 668.3 | 78.6 | 11.8 | 98.7 | 12.9 |
| 654 | 0.5 | 322.5 | 0.7 | 0.2 | 55.5 | 14.7 |
| 759 | 1,164.0 | 317.0 | 34.1 | 10.8 | 150.0 | 32.1 |
| 805 | 2,354.3 | 414.3 | 48.5 | 11.7 | 115.7 | 21.8 |
| 770 | 5,100.5 | 257.5 | 71.4 | 27.7 | 98.5 | 27.7 |

Materials and methods

Sample collection

Zooplankton samples were collected in San Francisco, Morro, Mission, and San Diego bays in 2013 and Bodega Harbor in 2014 (all in California, USA) using a combination of pump and tow methods (Supplementary material Table S1). These are a subset of a larger dataset of zooplankton samples, so we purposefully included in this study samples from multiple bays that appeared to vary in species richness. Samples were collected within marinas using a modified trash pump (North Star S106120 model; Honda GX160 gas motor) coupled to a plankton net assembly (0.75 m diameter net; 80 μ m mesh size) to collect and filter zooplankton at 1 m depth over 10 minutes, totaling 5 m³ water volume filtered per sample. Additionally, a tow sample was collected in open water adjacent to each marina using a weighted plankton net (0.50 m diameter; 80 μ m mesh size) deployed to 5 m depth and pulled vertically up through the water column. All samples were preserved in *ca.* 200 mL 95% ethanol.

DNA extraction

Prior to extraction, each plankton sample was sieved through a clean 75 μ m mesh and rinsed well with 1X TE (Tris-EDTA) buffer. All samples were initially subsampled intact as follows: total weight was recorded and a single 0.25 g (wet weight) intact subsample of each intact plankton sample was added to a PowerBead tube of a MoBio PowerSoil extraction kit. This was repeated for a total of five, independent 0.25 g (wet weight) subsample replicates of each intact plankton sample. When total sample weight did not exceed 5 \times 0.25 g, a replicate was omitted (sample 656 only). This procedure resulted in a total of 29 intact (unhomogenized) DNA replicates in the intact dataset (Table 1). Additionally, we generated a single intact replicate

(as described above) for each of the samples in the homogenized dataset ($n = 6$; described below). Remaining material was stored in the original bottle with the 95% ethanol retained after sieving the sample.

Samples that were bulk homogenized were processed by taking ethanol preserved plankton (from which a 0.25 g intact subsample had been previously removed) and filtering the remaining plankton again through 75 μm mesh and rinsing well with 1X TE (Tris-EDTA) buffer (Table 1). Samples were individually re-weighed, transferred to a clean, autoclaved mortar and homogenized by pulverizing with a pestle. An additional 2–3 subsample replicates of each homogenized sample (~ 0.25 g each) were transferred to a PowerBead tube of a MoBio PowerSoil extraction kit. This resulted in a total of 16 homogenized replicates in the homogenized dataset (Table 1). For both intact and homogenized subsamples, the extraction continued by following the manufacturer's protocol, except that samples were eluted into 80 μL of elution buffer. A 20 μL aliquot of each DNA sample was transferred to a 96 well plate for downstream applications. Different samples were used to generate the two datasets, as there was not enough material to test both methods on the same set of samples (Table 1).

Library preparation

Genomic DNA was quantified using the Invitrogen Quant-iT PicoGreen dsDNA Assay, according to the manufacturer's protocol. Genomic DNA was standardized to 0.5 and 2.5 ng/ μL for samples used in the intact and homogenized datasets (Table 1), respectively. The COI gene was amplified, in triplicate, using primers mlCOIintF forward (5'-GGWACWGGWTGAACWGTWTAYCCYCC-3'; Leray et al. 2013) and jgHCO2198 reverse (5'-ACYTCIGGRTGICCRARAAYCA-3'; Geller et al. 2013) with partial Nextera barcode indices added to the 5' ends (Illumina support 2013). For the PCR reaction, (0.5 intact or 2.5 ng homogenized) genomic DNA was amplified with final concentration of 1X Kapa Robust Hot Start Ready Mix, 0.2 mg/mL BSA, 2 mM MgCl_2 , and 0.4 μM of each primer in a 25 μL reaction. Thermocycling conditions consisted of an initial 3 min at 95 $^\circ\text{C}$, followed by 27 cycles of 1 min at 95 $^\circ\text{C}$, 45 s at 47 $^\circ\text{C}$, and 1 min at 72 $^\circ\text{C}$ with a final 72 $^\circ\text{C}$ hold for 5 min. Amplicons were viewed on a 2% agarose gel stained with ethidium bromide under UV light. Triplicates were pooled and purified with 1.4X the sample volume of Agencourt Ampure beads, according to the manufacturer's protocol, PicoGreen quantified, and concentrations were standardized.

To attach the remaining portion of the Nextera barcodes, 1 μL of pooled, purified amplicons were PCR amplified using a final concentration of 1X Kapa Robust Hot Start Ready Mix, 0.2 mg/mL BSA, 0.2 μM each forward and reverse barcode, and 2 mM MgCl_2 in a final volume of 25 μL . Thermocycling conditions consisted of an initial 3 min at 95 $^\circ\text{C}$, followed

by 8 cycles of 30 s at 95 °C, 30 s at 55 °C, and 30 s at 72 °C with a final 72 °C hold for 5 min. PCR products were viewed on a 2% agarose gel stained with ethidium bromide under UV light. Amplicons were purified with 1.4X the sample volume of Agencourt Ampure beads, according to the manufacturer's protocol. Purified samples were quantified using the PicoGreen Assay, according to the manufacturer's protocol. Barcoded amplicons were pooled evenly according to their concentration in ng/μL. The purified barcode PCR product was pooled into a sterile 1.5 mL microcentrifuge tube and purified with 1.4X the pooled library volume of Agencourt Ampure beads, according to the manufacturer's protocol, and eluted in TE buffer. One microliter of the pooled library was loaded on a High Sensitivity DNA chip (Agilent Technologies) and quantified with the Agilent 2100 Bioanalyzer System to determine the final concentration before Illumina sequencing. The library was denatured with 0.2 N sodium hydroxide, combined with 20% PhiX (also denatured with 0.2 N sodium hydroxide), diluted to a concentration of 8 pM, then run on an Illumina MiSeq using a 600 cycle v3 cartridge, according to the manufacturer's recommended protocols. The sequence data is available from the NCBI Short Read Archive (SRA) BioProject PRJNA501910.

Bioinformatics

Using USEARCH v9.2 (Edgar 2010), paired end reads were merged (-fastq_mergepairs) using a minimum final length of 356 bp, a maximum final length of 374 bp, and allowing ≤ 12 differences in the merged alignment. These parameters were chosen to allow for potential variation in the size of the COI fragment, in the case of the maximum and minimum final lengths, and to follow the software recommendations for read pairs with long, overlapping merge areas (> 100 bp). To filter low quality reads from the dataset, we used a strict maximum expected error rate of 0.5 (-fastq_filter). Primers were trimmed from the ends of all reads (-fastx_truncate). Sequences were then dereplicated (-fastq_uniques) and sorted by abundance (-sortbysize). Singletons were removed at this step to increase the speed of clustering. Clustering was conducted in two steps. First, sequences were clustered at the default 97% similarity using UPARSE (Edgar 2013; -cluster_otus), which also includes chimera detection and removal. Sequences were then sorted by length (-sortbylength) and clustered a second time at 95% similarity (-cluster_smallmem) to more accurately approximate metazoan species richness with genetic similarity across the COI gene region.

The representative set of sequences generated from the 95% clustering was imported into Geneious v10.2.3 (<http://www.geneious.com>, Kearse et al. 2012). Alignments were generated using ClustalW (Thompson et al. 1997) with the IUB cost matrix, a gap open cost of 50, and a gap extend cost of 6.66. These alignments were reviewed by eye to determine which

sequences were unlikely to be from COI, based on the size and number of indels present. Sequences that did not appear to be COI were then queried using BLAST within Geneious against the nr database at NCBI. Sequences not identified as COI were removed from the representative set. The remaining sequences were used to generate OTU tables in USEARCH (-usearch_global) with reads mapping to the filtered representative set at 95% similarity, including singletons. Two separate OTU tables were generated, one containing all samples used for comparisons of the intact-plankton subsampling (intact dataset) and one containing all samples that were homogenized (homogenized dataset; Table 1).

Using Quantitative Insights Into Microbial Ecology (QIIME; Caporaso et al. 2010), rarefaction curves were generated for both datasets (-alpha_rarefaction) to determine an appropriate rarefaction cut-off to optimize both the number of samples and sequences per sample included in the analyses. The intact-sample dataset was rarefied to 75,000 sequences and the homogenized sample dataset was rarefied to 85,000 sequences. With these rarefied datasets, the mean, standard deviations, and variance of OTUs across replicates were calculated in Microsoft Excel. OTU accumulation curves were generated using the vegan package (Oksanen et al. 2016) in the statistical program R (R Core Team 2015). We also assessed the distribution of high abundance (those with > 10,000 reads) and low abundance (those with < 100 reads) OTUs across the datasets. Venn diagrams to compare across replicates were generated using mothur (Schloss et al. 2009). We examined the number of shared OTUs across 50, 75, and 100% of samples (-compute_core_microbiome.py) in both datasets as well as those OTUs in high and low abundance.

Results

Intact sample dataset

For the intact sample dataset, we generated 78,193–208,647 reads per subsample replicate per sample. After rarefaction, there were 89–581 OTUs per replicate per sample. The variance and standard deviation of OTUs generated across replicates for the same sample were high (Table 1), with the average differences in OTU number between replicates per sample being 144.4–279.5 OTUs (Table 1). Thus, only 30.7–51.1% of all recovered OTUs were unique and only captured in a single subsample replicate using intact subsampling.

The OTU accumulation curves showed that adding subsamples adds substantial additional OTU richness and that none of these curves appeared to asymptote (Figure 1), indicating that additional richness is likely undetected in these samples. When examining the percent OTU richness captured per subsample (assuming that the total number of subsamples generated was sufficient for the total richness present), we found

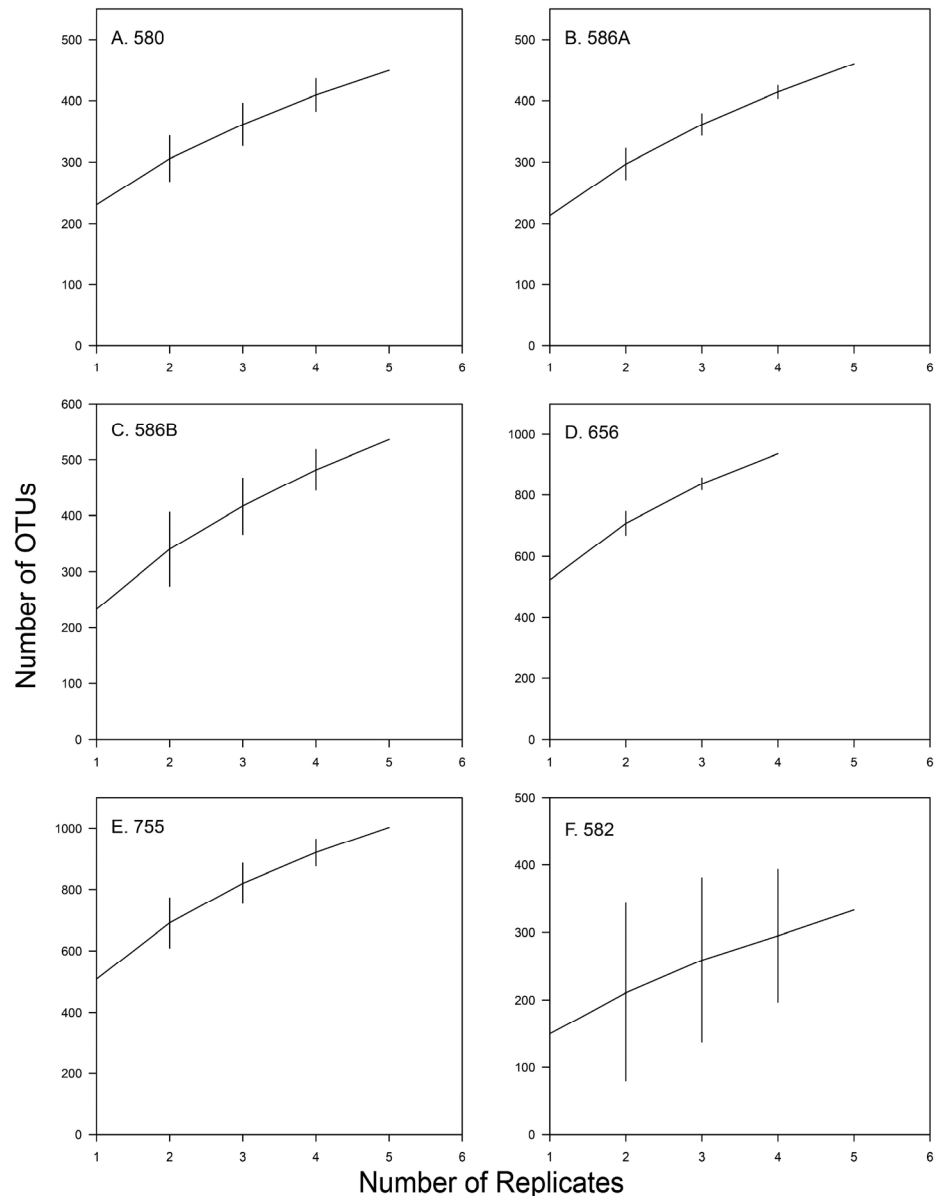


Figure 1. OTU accumulation curves generated in the vegan package (Oksanen et al. 2016) in the statistical program R (R Core Team 2015) for all samples in the intact dataset.

that a single subsample was only capable of capturing 43.5–56% of the richness (Figure 2). Adding a second replicate added ~ 20% additional OTU richness to the dataset, while additional 3–4 replicates each added ~ 10–15% more OTUs (Figure 2).

There were 63 OTUs with > 10,000 reads, which we considered high abundance, accounting for 1,892,361 reads. That is, 2.3% of the OTUs accounted for 62.5% of the sequences. The accumulation curves using only these high abundance OTUs did not show such sharp inclines and some appeared to asymptote (Figure S1), indicating that for many samples the richness of the high abundance OTUs was covered with 2–3 subsamples.

Repeating these analyses for low abundance sequences, there were 1,991 OTUs with < 100 reads, which account for 37,278 reads. That is, 73.9% of the OTUs were found within 1.2% of the sequences. Not surprisingly, the

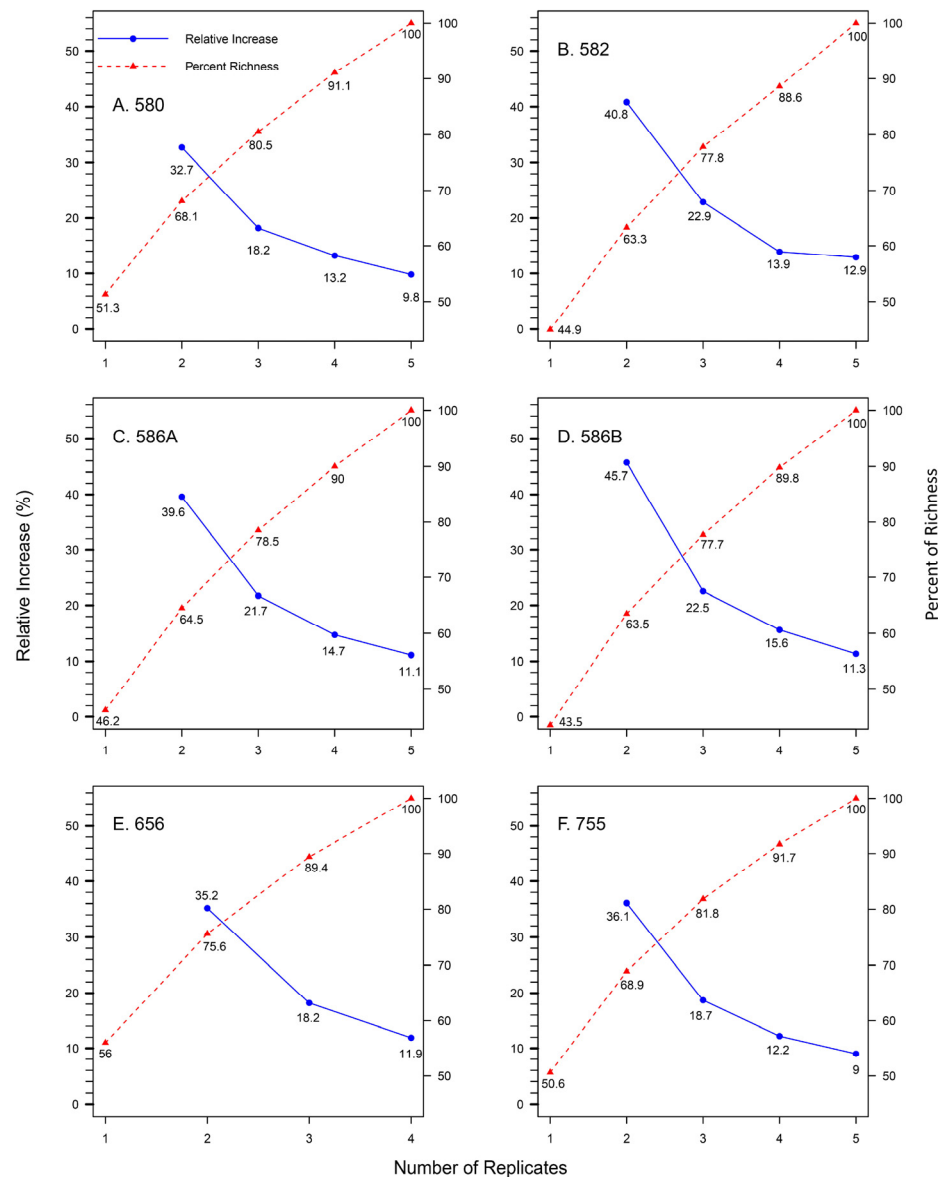


Figure 2. The percent OTU richness that is added with each additional replicate for all samples in the intact dataset is shown with a red dashed line. The relative increase in OTU richness with each additional replicate for all samples in the intact dataset is shown with a blue solid line.

OTU accumulation curves for this subset show sharp inclines, large standard error bars, and no asymptotes (Figure S2), indicating that much of the OTU richness for these low abundance organisms remains undetected even with 4–5 replicates. When examining the percent OTU richness per subsample (assuming that the total number of subsamples generated was sufficient for the total richness present), we found that a single subsample was only capable of capturing 23.8–46.7% of the richness of low abundance taxa (Figure S3).

Homogenized sample dataset

For the homogenized sample dataset, we generated 85,275–515,785 reads per subsample replicate per sample. After rarefaction, there were 207–727 OTUs per replicate per sample, higher than for the intact sample dataset.



Figure 3. The number of OTUs that were shared across four samples in the homogenized sample dataset, where a single intact subsample (left circle) was compared with a single homogenized subsample (right circle). This comparison was conducted across all OTUs (top), high abundance OTUs (middle), and low abundance OTUs (bottom).

For replicates of the same sample, the variance and standard deviation of OTUs were also high for this dataset (Table 1), with the average difference in OTU number between replicates per sample being 35.7–150 OTUs (Table 1). Thus, only 5.7–32.1% of all OTUs were unique and only captured in a single replicate using homogenized subsamples.

From four samples within this dataset, we compared one unhomogenized subsample to one homogenized subsample from the same sample to see if homogenization changed overall OTU richness or detection of rare OTUs. For these four samples, the two different replicate treatments shared between 41–50% of the total OTUs generated (Figure 3); however, for every sample, > 100 more OTUs were recovered from only the homogenized replicate. Upon further examination, we found that most high abundance OTUs were detected regardless of the preprocessing used. Most of the additional OTUs present in the homogenized subsamples were low abundance OTUs (Figure 3).

The OTU accumulation curves show that adding replicates substantially increases OTU richness as none of these curves appear to asymptote (Figure 4), indicating that additional richness is likely undetected in these samples. We then examined the percent OTU richness captured per subsample (treating total recovered OTU richness as an estimate of total actual richness) and found that a single homogenized subsample was capable of detecting 69.7–83.9% of the richness (Figure 5). Adding a second replicate added 24–43% additional OTUs to the dataset, while adding a third replicate added 13–30% more OTUs (Figure 5).

Finally, we compared the distribution of core OTUs (those recovered in 50%, 75%, and 100% of replicates within samples) in the intact *vs.* homogenized subsample datasets. These results confirm the trends observed in other analyses, indicating that more OTUs were observed with

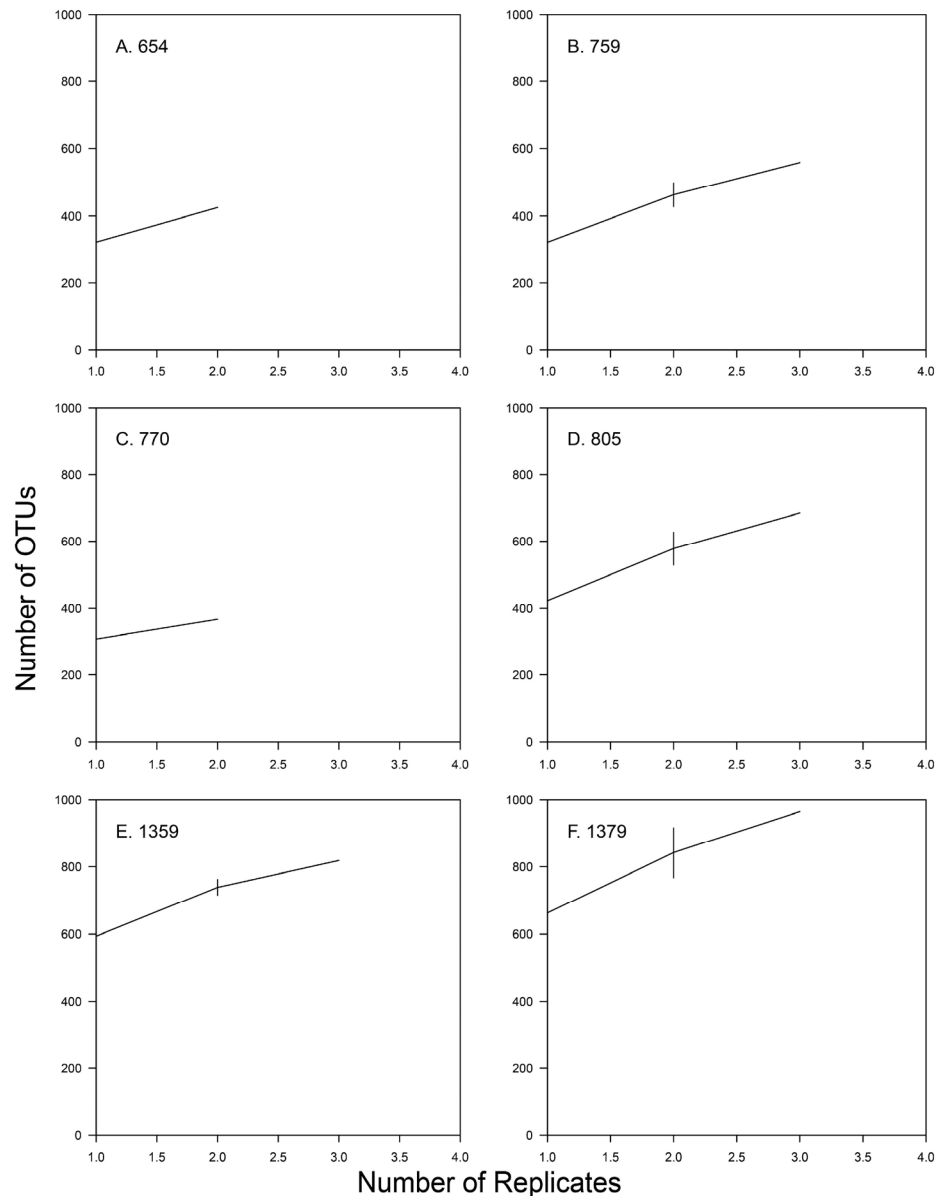


Figure 4. OTU accumulation curves generated in the vegan package (Oksanen et al. 2016) in the statistical program R (R Core Team 2015) for all samples in the homogenized sample dataset.

homogenization than without (Figure 6A, B) and that high abundance OTUs were well covered with the current sequencing effort (Figure 6C, D). This analysis also shows that when looking at all data (Figure 6A, B) or the low abundance OTUs (Figure 6E, F), there are more core OTUs present across homogenate than intact subsamples, indicating that more OTUs are shared across homogenized replicates, which could increase the likelihood of detection for any single OTU, particularly if a single replicate is processed. For the high abundance OTUs, these were readily shared across replicates in both the intact subsample and homogenized replicates, indicating that these OTUs are more likely be detected in any single replicate using either technique. This analysis also shows that homogenized subsamples had a greater number of low abundance OTUs (Figure 6E, F), though this is likely impacted by the varying richness in the

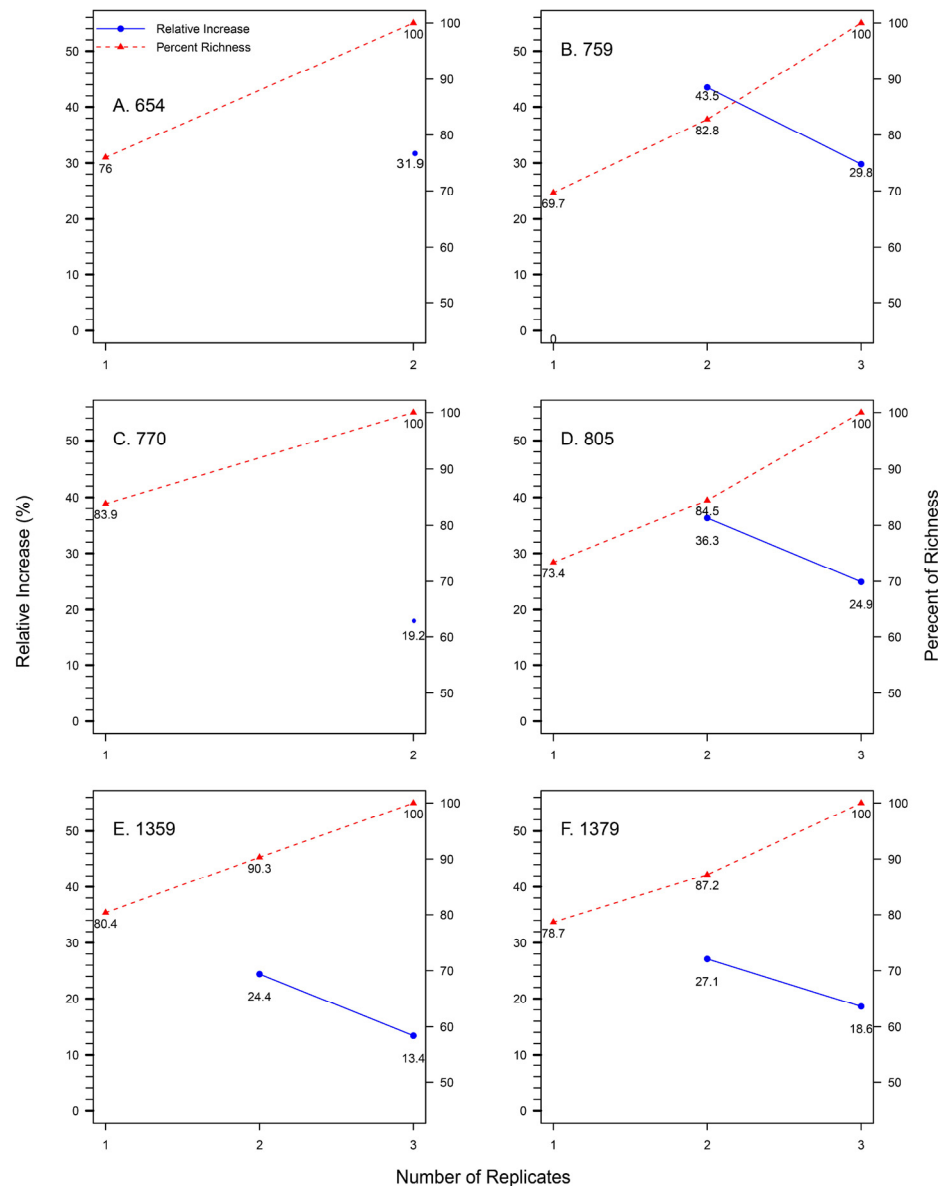


Figure 5. The percent OTU richness that is added with each additional replicate for all samples in the homogenized sample dataset is shown with a red dashed line. The relative increase in OTU richness with each additional replicate for all samples in the homogenized sample dataset is shown with a blue solid line.

initial sample (i.e., samples 656 and 755 appear to have higher richness than the other samples chosen for intact subsampling).

Discussion

Our results demonstrate that homogenizing a zooplankton sample prior to DNA extraction increased the number of OTUs recovered, particularly rare or low abundance OTUs, increased the richness captured in a single subsample, and resulted in more OTUs shared across all subsamples, thus increasing the likelihood of detecting a new invader in a single or multiple subsamples. However, the increase in OTUs comes at a cost, as more sequences per replicate are required to assess the rare OTUs within a subsample. Additionally, adding extraction replicates per sample always

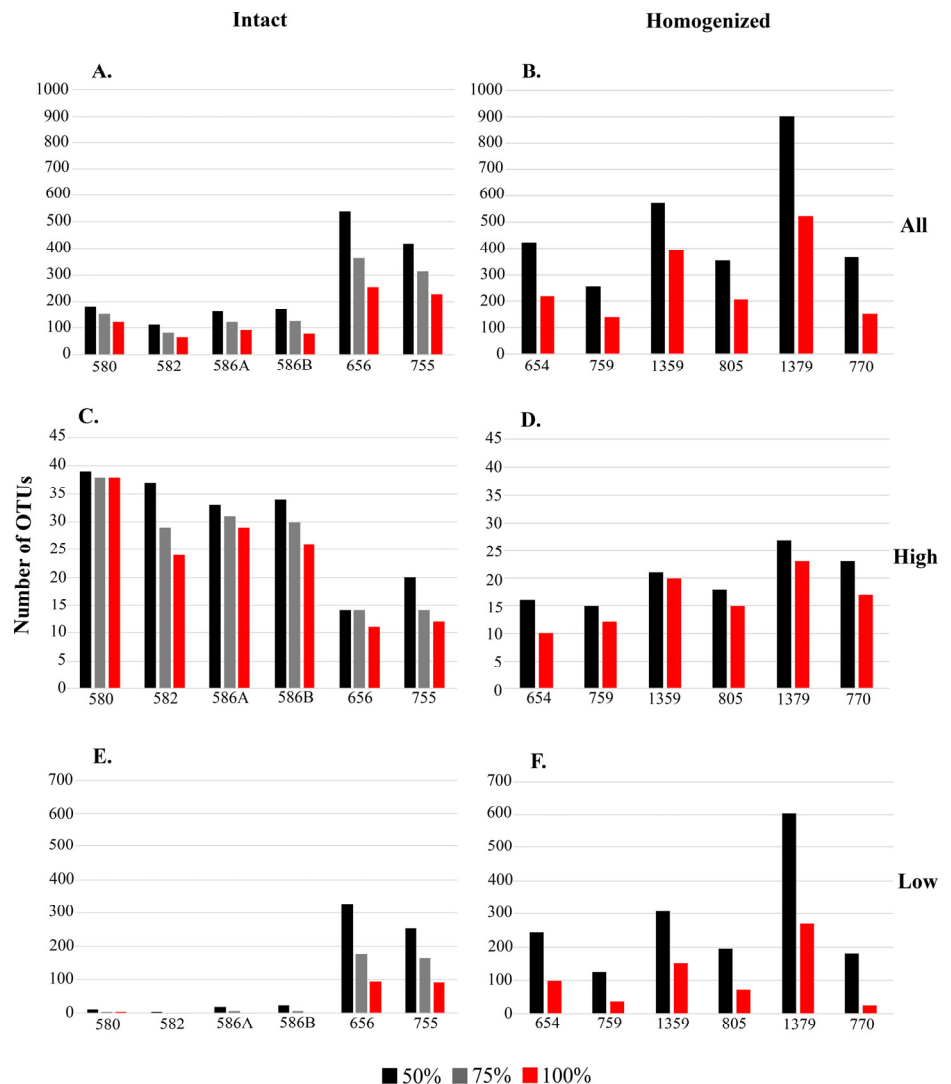


Figure 6. The distribution of core OTUs (those recovered in 50%, 75%, and 100% of samples) in the intact vs. homogenized datasets across all OTUs (A, B), the high abundance OTUs (C, D), and the low abundance OTUs (E, F).

increased OTU richness regardless of the dataset, with 2–3 subsample replicates per sample generally being sufficient to encompass the majority of diversity detected within a sample, though this might be lower for the homogenized compared to intact samples. Given these results, when detection of low abundance non-native species is the primary goal, homogenizing bulk community samples should be performed due to the clear increase in low abundance OTUs detected. However, we caution that to do this, researchers need to factor in the additional processing time and costs. Further, sample homogenization is a step that is vulnerable to cross-sample contamination, so extreme caution should be taken when performing this step.

The use of metagenetics combined with high throughput sequencing (HTS) technology has proven to be a promising method for non-indigenous species (NIS) detection. This method can readily detect low abundance organisms, including newly introduced non-native species and rare natives (Zhan and MacIsaac 2015). It also allows for high-throughput

processing of samples, which have and will continue to be cheaper as sequencing costs decline, negating the need for costly and time intensive morphological identification of microscopic marine organisms. Finally, this method also allows for the possibility of lower-level identification of small, morphologically non-descript marine organisms, including those that can be difficult or impossible to identify otherwise, including many species of marine larvae. While there are many benefits to metagenetics, the downsides to this methodology include: 1) lack of reproducibility for recovering low abundance organisms (Zhan et al. 2013), 2) difficulty with obtaining accurate or low-level taxonomic assignments due to sparsity of sequences in reference databases (Ransome et al. 2017), 3) trade-offs associated with marker choice (e.g., 18S, 16S, COI; Clarke et al. 2017), and 4) a lack of consensus on parameters for bioinformatic pipelines. Finally, while species introductions, which are more readily detected with metagenetics than by morphological methods, are likely high in number, many introduced species will not establish and only a few that do establish will become pests (Williamson and Fitter 1996). This means researchers who use metagenetics will have a lot of sequence data to parse through in order to find a relatively small number of potential invaders and difficulties determining which species are more likely to establish to prioritize their early eradication or prevent future introductions. Alternatively, zooplankton is relatively easy to collect and compactly store, such that extra, unprocessed samples could be saved and retrospectively analyzed by genetic methods to study early stages of successful invasions. Even with these issues, metagenetics provides better resolution and is more cost-effective than traditional, microscopy surveys for NIS taxa.

As a lack of material forced us to use different samples for the different datasets, we were prevented from conducting any direct comparisons between intact and homogenized subsamples (with the exception of the data in Figure 3), so instead focused on and compared the overall conclusions in both datasets. Our results clearly show that homogenizing zooplankton samples led to an increased detection of low abundance OTUs overall (Figure 3) and across replicates (Figure 6), indicating an increased detection of rare species. Our comparisons also demonstrated that increasing the number of subsample replicates that are extracted from a single sample increases both overall richness (Figures 2, 4) and the likelihood of detecting low abundance OTUs (Figure S3), regardless of whether or not the samples are homogenized, though homogenized replicates contained more richness per replicate (Figures 2, 5). While some of these may be spurious OTUs that are not biologically relevant, we took extra precautions to find and remove those OTUs including strict filtering based on sequence quality, chimera detection, and making alignments with the representative sequences. What this data indicates is that when a

sample is pulverized (i.e., homogenized), this process allows for smaller fragments from a greater variety of organisms to be present in a single subsample. Clearly, a major factor in detection for this process (with or without homogenization) is the starting richness of the original sample. Thus, the richness in the homogenized subsamples after pulverization is higher because those samples include more of the organisms from the total pool of the original sample, though likely in smaller increments. This is where there is certainly a “Catch-22” in the process: in allowing for smaller components of more individuals, the overall richness of the subsample increases as does the likelihood that rare organisms would be detected (due to their inclusion in the homogenized subsample), but being able to capture the rare OTUs in the HTS pipeline requires greater sequence depth per sample. Finally, the general conclusions across the intact and homogenized datasets appear to hold regardless of the sampling method (pump vs. tow) used and previous comparisons on a larger dataset have demonstrated that there is no difference in richness across the two methods (Geller et al., *unpubl. data*). Thus, it appears that homogenizing complex, multi-species samples prior to extraction may increase the likelihood of detecting early introductions of NIS and native species occurring at low abundance.

Replicates

Multiple studies have assessed the impact of PCR replicates on determining the biodiversity in a multi-species sample. For example, Alberdi et al. (2018) showed that 95% more OTUs and 30% more species were detected when PCR replicates were increased from one to three. Additionally, they showed that if taxonomic richness is high and abundance is variable, then multiple replicates were more likely to capture the total diversity present. In contrast, fewer studies have generated extraction replicates, as doing so can substantially increase the processing time and costs, though it has been suggested as a way to statistically correct for intra-sample variation (Zhan and MacIsaac 2015). Xiong et al. (2016) showed that increasing technical replicates (i.e., repeated DNA extractions and PCRs from the same sample) will likely decrease false negatives. Our results support this finding, demonstrating that additional taxa (24–43%), were recovered with additional DNA extraction replicates, with this increase being even more drastic for the rare taxa (46–84%). These results indicate that multiple technical replicates could be an important step in thoroughly assessing rare taxa. However, given that individual homogenized replicates contained more overall diversity (69–84%, Figure 5), compared to intact replicates (43.5–56%, Figure 2), it is possible that researchers could use less homogenized than intact replicates to assess the same richness. Also, researchers must carefully consider the tradeoff between fully assessing the

biodiversity, including rare taxa, within a single sample through multiple DNA extractions and assessing the biodiversity in a given area through additional sample replicates (Xiong et al. 2016).

Sequencing Depth

Another way to increase the diversity detected in a single sample is to generate more sequences per sample, though there are clear limitations to this approach. As shown in Alberdi et al. (2018), the detected biodiversity in a sample tends to level off at high sequencing depth (25,000 reads per faecal sample for their study), which is expected due to the inevitability of eventually sequencing all the organisms present in an aliquot. Alberdi et al. (2018) also showed that increasing sequencing depth did not fully negate the impacts of PCR stochasticity and primer biases, as even at higher sequence depth they detected substantial differences in diversity across PCR replicates. In this study, we showed that homogenized subsamples had greater OTU richness compared to intact subsamples (Figure 3); however, as described above, a component of assessing this richness means increasing the sequencing depth per subsample, which increases the cost of sample processing. Thus, there appears to be a trade-off between increasing sequencing depth for a single replicate and increasing the number of either PCR or extraction replicates to more fully characterize diversity.

Acknowledgements

We thank Captain Steve Neil (UC Davis Bodega Marina Lab), Rob Brewster, Tom Moylan, and Jason Felton (CalPoly Center for Coastal Marine Sciences) for access to boats and assistance with sample collection. We also thank Arianne Balsom and Darrick Sparks with their assistance in organizing field collections and collecting samples. Finally, we thank the anonymous reviewers for their thorough assessment and helpful comments on an earlier version of this manuscript. All samples were collected under permit number SC-10124. K. Lohan is a Robert and Arlene Kogod Secretarial Scholar. This project was funded by an award from the California Department of Fish and Wildlife to J. Geller.

References

- Abad D, Albaina A, Aguirre M, Laza-Martínez A, Uriarte I, Iriarte A, Villate F, Estonba A (2016) Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. *Marine Biology* 163: 149, <https://doi.org/10.1007/s00227-016-2920-0>
- Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K, Mahon A (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution* 9: 134–147, <https://doi.org/10.1111/2041-210X.12849>
- Bjørnsgaard Aas A, Davey ML, H Kausrud (2017) ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources* 17: 730–741, <https://doi.org/10.1111/1755-0998.12622>
- Borrell YJ, Miralles L, Do Huu H, Mohammed-Geba K, Garcia-Vazquez E (2017) DNA in a bottle-Rapid metabarcoding survey for early alerts of invasive species in ports. *PLoS ONE* 12: e0183347, <https://doi.org/10.1371/journal.pone.0183347>
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7: 335–336, <https://doi.org/10.1038/nmeth.f.303>

- Chain FJJ, Brown EA, MacIsaac HJ, Cristescu ME, Cowie R (2016) Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. *Diversity and Distributions* 22: 493–504, <https://doi.org/10.1111/ddi.12427>
- Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology & Evolution* 7: 873–883, <https://doi.org/10.1002/ece3.2667>
- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461, <https://doi.org/10.1093/bioinformatics/btq461>
- Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10: 996, <https://doi.org/10.1038/nmeth.2604>
- Fletcher LM, Zaiko A, Atalah J, Richter I, Dufour CM, Pochon X, Wood SA, Hopkins GA (2017) Bilge water as a vector for the spread of marine pests: a morphological, metabarcoding and experimental assessment. *Biological Invasions* 19: 2851–2867, <https://doi.org/10.1007/s10530-017-1489-y>
- Freeland JR (2017) The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA. *Genome* 60: 358–374, <https://doi.org/10.1139/gen-2016-0100>
- Geisen S, Laros I, Vizcaino A, Bonkowski M, De Groot GA (2015) Not all are free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa. *Molecular Ecology* 24: 4556–4569, <https://doi.org/10.1111/mec.13238>
- Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome *c* oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources* 13: 851–861, <https://doi.org/10.1111/1755-0998.12138>
- Hajibabaei M, Shokralla S, Zhou X, Singer GA, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6: e17497, <https://doi.org/10.1371/journal.pone.0017497>
- Illumina Support (2013) 16S Metagenomic Sequencing Library Preparation. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649, <https://doi.org/10.1093/bioinformatics/bts199>
- Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, McDonald IR, Cary SC (2012) Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS ONE* 7: e44224, <https://doi.org/10.1371/journal.pone.0044224>
- Leray M, Knowlton N (2017) Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ* 5: e3006, <https://doi.org/10.7717/peerj.3006>
- 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. *Frontiers in Zoology* 10: 34, <https://doi.org/10.1186/1742-9994-10-34>
- Lymperopoulou DS, Dobbs FC (2017) Bacterial Diversity in Ships' Ballast Water, Ballast-Water Exchange, and Implications for Ship-Mediated Dispersal of Microorganisms. *Environmental Science & Technology* 51: 1962–1972, <https://doi.org/10.1021/acs.est.6b03108>
- Morales SE, Cosart TF, Johnson JV, Holben WE (2009) Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. *Applied and Environmental Microbiology* 75: 668–675, <https://doi.org/10.1128/AEM.01757-08>
- Oksanen J, Guillaume Blanchet F, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2016) vegan: Community Ecology Package, October 28, 2018. <https://cran.r-project.org/web/packages/vegan/vegan.pdf>
- Pagenkopp Lohan KM, Fleischer RC, Carney KJ, Holzer KK, Ruiz GM (2016) Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases. *Microbial Ecology* 71: 530–542, <https://doi.org/10.1007/s00248-015-0684-6>
- Pagenkopp Lohan KM, Fleischer RC, Carney KJ, Holzer KK, Ruiz GM (2017) Molecular characterisation of protistan species and communities in ships' ballast water across three U.S. coasts. *Diversity and Distributions* 23: 680–691, <https://doi.org/10.1111/ddi.12550>
- 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, <https://doi.org/10.1371/journal.pone.0073935>
- Pochon X, Zaiko A, Fletcher LM, Laroche O, Wood SA (2017) Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. *PLoS ONE* 12: e0187636, <https://doi.org/10.1371/journal.pone.0187636>

- Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* 64: 3724–3730
- R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org>
- Ransome E, Geller JB, Timmers M, Leray M, Mahardini A, Sembiring A, Collins AG, Meyer CP (2017) The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. *PLoS ONE* 12: e0175066, <https://doi.org/10.1371/journal.pone.0175066>
- Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6: e27310, <https://doi.org/10.1371/journal.pone.0027310>
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* 75: 7537–7541, <https://doi.org/10.1128/AEM.01541-09>
- Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21: 2045–2050, <https://doi.org/10.1111/j.1365-294X.2012.05470.x>
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 25: 4876–4882, <https://doi.org/10.1093/nar/25.24.4876>
- Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E, Besnard A, Coissac E, Boyer F, Gaboriaud C, Jean P, Poulet N, Roset N, Copp GH, Geniez P, Pont D, Argillier C, Baudoin JM, Peroux T, Crivelli AJ, Olivier A, Acqueberge M, Le Brun M, Møller PR, Willerslev E, Dejean T (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology* 25: 929–942, <https://doi.org/10.1111/mec.13428>
- Williamson M, Fitter A (1996) The varying success of invaders. *Ecology* 77: 1661–1666, <https://doi.org/10.2307/2265769>
- Wommack K, Bhavsar E, Ravel J (2008) Metagenomics: read length matters. *Applied and Environmental Microbiology* 74: 1453–1463, <https://doi.org/10.1128/AEM.02181-07>
- Xiong W, Li HT, Zhan AB (2016) Early detection of invasive species in marine ecosystems using high-throughput sequencing: technical challenges and possible solutions. *Marine Biology* 163: 139, <https://doi.org/10.1007/s00227-016-2911-1>
- Zaiko A, Martinez JL, Ardura A, Clusa L, Borrell YJ, Samuiloviene A, Roca A, Garcia-Vazquez E (2015) Detecting nuisance species using NGST: Methodology shortcomings and possible application in ballast water monitoring. *Marine Environmental Research* 112(Pt B): 64–72, <https://doi.org/10.1016/j.marenvres.2015.07.002>
- Zhan A, MacIsaac HJ (2015) Rare biosphere exploration using high-throughput sequencing: research progress and perspectives. *Conservation Genetics* 16: 513–522, <https://doi.org/10.1007/s10592-014-0678-9>
- Zhan A, Hulák M, Sylvester F, Huang X, Abebayo 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 in Ecology and Evolution* 4: 558–565, <https://doi.org/10.1111/2041-210X.12037>

Supplementary material

The following supplementary material is available for this article:

Figure S1. OTU accumulation curves generated in the vegan package (Oksanen et al. 2016) in the statistical program R (R Core Team 2015) for the high abundance OTUs from all samples in the subsample dataset.

Figure S2. OTU accumulation curves generated in the vegan package (Oksanen et al. 2016) in the statistical program R (R Core Team 2015) for the low abundance OTUs from all samples in the subsample dataset.

Figure S3. The percent OTU richness that is added with each additional replicate for the low abundance OTUs from all samples in the subsample dataset is shown with a red dashed line.

Table S1. Collection information for all the samples used in this study.

This material is available as part of online article from:
<http://www.reabic.net/journals/mbi/2019/Supplements/>