LJMU Research Online

Collins, RA, Bakker, J, Wangensteen, OS, Soto, AZ, Corrigan, L, Sims, DW, Genner, MJ and Mariani, S

Non-specific amplification compromises environmental DNA metabarcoding with COI

http://researchonline.ljmu.ac.uk/id/eprint/11575/

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Collins, RA, Bakker, J, Wangensteen, OS, Soto, AZ, Corrigan, L, Sims, DW, Genner, MJ and Mariani, S (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution. ISSN 2041-210X

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

http://researchonline.ljmu.ac.uk/

LJMU Research Online

Collins, RA, Bakker, J, Wangensteen, OS, Soto, AZ, Corrigan, L, Sims, DW, Genner, MJ and Mariani, S

Non‐specific amplification compromises environmental DNA metabarcoding with COI

http://researchonline.ljmu.ac.uk/id/eprint/11575/

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Collins, RA, Bakker, J, Wangensteen, OS, Soto, AZ, Corrigan, L, Sims, DW, Genner, MJ and Mariani, S Non‐specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution. ISSN 2041-210X

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

http://researchonline.ljmu.ac.uk/

Non-specific amplification compromises environmental DNA metabarcoding with COI

Rupert A. Collins*[1], Judith Bakker*[2,3], Owen S. Wangensteen[3,4], Ana Z. Soto[3], Laura Corrigan[5], David W. Sims[6,7], Martin J. Genner[1], and Stefano Mariani[3,8]

[1]**School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK**
[2]**Department of Biological Sciences, Florida International University, 11200 S.W., 8th Street, Miami, Florida, 33199, USA**
[3]**Ecosystems & Environment Research Centre, School of Environment & Life Sciences, University of Salford, Salford M5 4WT, UK**
[4]**Norwegian College of Fishery Science, UiT The Arctic University of Norway, N-9037, Tromsø, Norway**
[5]**Environment Agency, Tyneside House, Skinnerburn Road, Newcastle upon Tyne NE4 7AR, UK**
[6]**Marine Biological Association of the United Kingdom, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK**
[7]**Ocean and Earth Science, University of Southampton, National Oceanography Centre Southampton, European Way, Southampton SO14 3ZH, UK**
[8]**School of Natural Sciences & Psychology, Liverpool John Moores University, Liverpool, L3 3AF, UK**

Corresponding author:
Stefano Mariani

Email address: s.mariani@salford.ac.uk

*Both authors contributed equally to this work.

ABSTRACT

1. Metabarcoding extra-organismal DNA from environmental samples is now a key technique in aquatic biomonitoring and ecosystem health assessment. However, choice of genetic marker and primer set is a critical consideration when designing experiments, especially so when developing community standards and legislative frameworks. Mitochondrial cytochrome *c* oxidase subunit I (COI), the standard DNA barcode marker for animals, with its extensive reference library, taxonomic discriminatory power, and predictable sequence variation, is the natural choice for many metabarcoding applications such as the bulk sequencing of invertebrates. However, the overall utility of COI for environmental sequencing of targeted taxonomic groups has yet to be fully scrutinised.

2. Here, by using a case study of marine and freshwater fishes from the British Isles, we quantify the *in silico* performance of twelve mitochondrial primer pairs from COI, cytochrome *b*, 12S and 16S, in terms of reference library coverage, taxonomic discriminatory power, and primer universality. We subsequently test *in vitro* three COI primer pairs and one 12S pair for their specificity, reproducibility, and congruence with independent datasets derived from traditional survey methods at five estuarine and coastal sites in the English Channel and North Sea coast.

3. Our results show that for aqueous extra-organismal DNA at low template concentrations, both metazoan and fish-targeted COI primers perform poorly in comparison to 12S, exhibiting low levels of reproducibility due to non-specific amplification of prokaryotic and non-target eukaryotic DNAs.

4. An ideal metabarcode would have an extensive reference library for which custom primer sets can be designed for either broad assessments of biodiversity or taxon specific surveys, but unfortunately, low primer specificity hinders the use of COI, while the paucity of reference sequences is problematic for 12S. The latter, however, can be mitigated by expanding the concept of DNA barcodes to include whole mitochondrial genomes generated by genome-skimming existing tissue collections.

[Keywords: 12S, COI, eDNA, Environmental DNA, metabarcoding, primer design.]

INTRODUCTION

DNA barcoding and metabarcoding techniques are now established and indispensable tools for the assessment and monitoring of past and present ecosystems (Valentini et al., 2016; Leray and Knowlton, 2015; Thomsen and Willerslev, 2015; Pedersen et al., 2015), and are being increasingly incorporated into policy and management decisions (Kelly et al., 2014b; Mariani et al., 2015; Rees et al., 2014; Hering et al., 2018). A remarkably wide range of biological substrates can now be sequenced to identify presence of a particular species or reconstruct communities, and can include restaurant sushi meals (Vandamme et al., 2016), deep sea sediments (Guardiola et al., 2015), permafrost ice cores (Willerslev et al., 2003), terrestrial insect collections (Ji et al., 2013), animal faeces (Kartzinel et al., 2015) and seawater samples (Thomsen et al., 2012a).

The term "DNA metabarcoding" encompasses two distinct methodologies: (i) bulk sample metabarcoding, which is the direct amplification of a concentrated mixture of organisms, from for example, plankton (Clarke et al., 2017), mass arthropod collections (Yu et al., 2012) or gut material (Leray et al., 2013); or (ii) "environmental DNA (eDNA) metabarcoding", which is indirect amplification via extra-organismal DNA in water, sediments, or soils (Taberlet et al., 2012). This latter methodology involves first isolating and concentrating DNA using filters, rather than homogenising entire organisms or parts of organisms (Macher et al., 2018; Yu et al., 2012; Spens et al., 2017). The detection of macrobial fauna such as vertebrates and insects using aquatic eDNA has been recognised as a highly sensitive survey technique and a key use-case of metabarcoding (Valentini et al., 2016; Rees et al., 2014). However, DNA from environmental samples such as seawater is likely to be degraded (Collins et al., 2018), and also have a significant quantity of co-extracted microbial DNA that may co-amplify with the targeted metazoan DNA molecules (Andújar et al., 2018; Stat et al., 2017).

Early eDNA metabarcoding studies targeting fishes used the cytochrome *b* gene (Thomsen et al., 2012b,a; Minamoto et al., 2012), but more recent studies have used the 12S ribosomal rRNA locus (Kelly et al., 2014a; Port et al., 2016; Hänfling et al., 2016; Stoeckle et al., 2017; Ushio et al., 2018; Yamamoto et al., 2017), and also 16S rRNA (Berry et al., 2017; Bylemans et al., 2018; Shaw et al., 2016; Stat et al., 2018; Jeunen et al., 2018). Various regions of 12S have been proposed as metabarcoding markers, including a ca. 63 bp fragment (Valentini et al., 2016), a ca. 106 bp fragment (Riaz et al., 2011; Kelly et al., 2014a), and a ca. 171 bp fragment (Miya et al., 2015). Modified versions of some of these primers have also been published by Taberlet et al. (2018). Ribosomal genes such as 12S and 16S offer the advantage of conserved priming sites (Deagle et al., 2014; Valentini et al., 2016), and amplification across a broad range of fish taxa (Bylemans et al., 2018; Miya et al., 2015). However, taxonomic resolution can be low (Hänfling et al., 2016; Andruszkiewicz et al., 2017; Miya et al., 2015), with relatively short length ribosomal markers being unable to distinguish commercially important species of the cod family Gadidae (Thomsen et al., 2016), for example. A problem for studies using ribosomal markers are the reference libraries, which are usually poorly populated, and often have to be developed for each project on an ad hoc basis (Thomsen et al., 2016; Stoeckle et al., 2017; Miya et al., 2015). Assembling reference libraries for ribosomal genes is further complicated by frequently-used primer sets amplifying different regions, so any two given 12S references from GenBank, for example, may not be homologous.

For animals, the primary DNA barcode is the 5′ "Folmer" region of COI, the cytochrome *c* oxidase subunit I gene (Folmer et al., 1994; Hebert et al., 2003). In comparison to ribosomal markers, the advantages of

For Review Only

COI are high interspecific variability (Ward, 2009), an extensive reference database (BOLD; Barcode of Life Database; Ratnasingham and Hebert, 2007), and due to the protein-coding constraints of the gene, more straightforward bioinformatic procedures such as alignment and denoising (Andújar et al., 2018). Inside of the 5′ Folmer fragment, multiple primer sets have been developed, targeting shorter regions in the 100–400 bp range, which are more suitable than a full length barcode (ca. 658 bp) for analyses of degraded DNA, or for sequencing on short read platforms such as Illumina (Elbrecht and Leese, 2017; Leray et al., 2013; Shokralla et al., 2015). However, due its nucleotide variation, finding conserved priming regions within the Folmer fragment is difficult, and concerns have been raised about the suitability of some COI primers in terms of species-specific primer-template mismatches, which can result in inefficient, biased amplifications that may hinder quantitative analyses (Deagle et al., 2014). Addressing this issue with bias requires incorporating a high degree of degeneracy into COI primers (Leray et al., 2013; Marquina et al., 2019), particularly by the use of multiple inosine sites (Elbrecht and Leese, 2017; Shokralla et al., 2015; Wangensteen et al., 2018). Despite this problem, Andújar et al. (2018) argue that COI should be the standard marker for metabarcoding, and COI markers are increasingly being used for eDNA metabarcoding (Stat et al., 2017; Kelly et al., 2017; Bakker et al., 2017; Macher et al., 2018; Jeunen et al., 2018; Singer et al., 2019). However, studies comparing efficacy markers have done so in a bulk-sample metabarcoding context (Clarke et al., 2017; Elbrecht and Leese, 2017), or have compared only ribosomal markers for vertebrate eDNA applications (Bylemans et al., 2018). Therefore, there lacks a clear assessment of how degenerate COI primers compare to 12S and 16S rRNA when used on low-template-concentration environmental samples, where non-target DNA molecules are found in abundance.

Given the importance of marker choice in metabarcoding studies (Alberdi et al., 2018), and the need to thoroughly scrutinise the utility of COI in comparison with the widely used ribosomal markers (Deagle et al., 2014; Andújar et al., 2018), we use a case study of fishes from the British Isles—a well studied and important group in terms of ecosystem health and human food security—to ask the following questions: (i) can COI primer sets be used as eDNA metabarcoding markers appropriate for aquatic vertebrate biodiversity assessment; and (ii) how do they compare to alternative markers including 12S, 16S and cytochrome *b*? We survey a range of published primer sets both *in silico* and *in vitro*, and include a degenerate metazoan COI primer pair as well as novel fish-targeted COI sets with reduced degeneracy. Using *in silico* methods we assess a number of factors: (i) the reference database coverage for the individual fragments, i.e. how many species and individuals of each species are represented in public databases; (ii) the taxonomic discrimination of each fragment, i.e. is each unique DNA sequence unambiguously associated with a single species name; and (iii) the universality of the primer set, i.e. are all species of the target taxonomic group predicted to amplify equally well. Then, we test using a series of water samples taken from locations with corresponding data from traditional fish survey methods, three COI primer sets against a best performing alternative set, as based upon the results of the *in silico* analyses. By PCR amplifying and sequencing these water samples we compare: (i) the specificity of the primer set, i.e. the proportion of the reads that came from the target taxonomic group; (ii) the power of the primer set, i.e. the total species richness estimated; (iii) the reproducibility of the primer set, i.e. are the same species consistently represented in replicate water samples and PCRs; and (iv) the congruence of the primer set, i.e. are the same species detected in the traditional surveys as the eDNA surveys.

For Review Only

METHODS

In silico analyses

Reference library construction

A list of fish species recorded from the marine and freshwater environments of the British Isles was compiled from three sources: (i) the Global Biodiversity Information Facility (https://www.gbif.org; *rgbif v1.1.0*; Chamberlain and Boettiger, 2017); (ii) FishBase (https://www.fishbase.org); and (iii) the European Water Framework Directive United Kingdom Technical Advisory Group list of transitional fish species (https://www.wfduk.org/resources/transitional-waters-fish; Annex 1). These species were then cross-referenced for all synonyms using *rfishbase v3.0.0* (Boettiger et al., 2012). The subsequent list of valid species names and all their synonyms was then searched using *rentrez v1.2.1* (Winter, 2017) against NCBI GenBank release 230 (nucleotide database; https://www.ncbi.nlm.nih.gov/nucleotide/) for any of the following terms: "COI, 12S, 16S, rRNA, ribosomal, cytb, CO1, cox1, cytochrome, subunit, COB, CYB, mitochondrial, mitochondrion". The Barcode of Life Database BOLD (http://www.boldsystems.org/) was also searched for the same species using *bold v0.8.6* (Chamberlain, 2018).

Hidden Markov models of the alignments of each primer set were then constructed using *HMMER v3.1b2* (http://hmmer.org/; Eddy, 1998) and the fish mitochondrial genome database (http://mitofish.aori.u-tokyo.ac.jp/; Iwasaki et al., 2013). These profiles were used to extract homologous regions of nucleotides from the total mitochondrial data obtained from the GenBank and BOLD searches. The resulting sequences were then annotated with metadata using *traits v0.3.0.9310* (Chamberlain et al., 2018). A phylogenetic quality control step was then carried out by aligning the sequences in *MAFFT v7.271* (Katoh and Standley, 2013) and constructing a maximum likelihood tree using *RAxML v8.2.12* (Stamatakis et al., 2008). Sequences with putatively spurious annotations—i.e. those indicative of misidentifications—were filtered out if the following criteria were met: (i) individual(s) of species *x* being identical to or nested within a cluster of sequences of species *y*, but with other individuals of species *x* forming an independent cluster; and (ii) the putatively spurious sequences coming from a single study, while the putatively correct sequences of species *x* and *y* coming from multiple studies. Records flagged by NCBI as "unverified" were also omitted. The full reference library and code to reproduce it can be found at https://doi.org/10.6084/m9.figshare.7464521.

For Review Only

Primer design

We designed two new COI metabarcoding primers targeting fishes (Table 1): "SeaDNA-short" and "SeaDNA-mid", which share a forward primer, and are internal to the Folmer fragment. The new primer pairs were designed manually in *Geneious v8.8.1* (Kearse et al., 2012) using the same fish mitochondrial genome dataset as described above, with the assistance of *Primer3* (Untergasser et al., 2012) and the sliding window functions in *spider v1.3.0* (Boyer et al., 2012; Brown et al., 2012). The primers were tested on a range of fish tissue extractions from elasmobranchs and actinopterygians, and produced strong clean PCR amplicons of the expected size.

In silico *PCR and taxonomic discrimination*

Primers were evaluated using a subset of 955 unique sequences from 184 species obtained in the British Isles fish reference library construction step, for which full mitochondrial genomes were available. Twelve primer pairs were chosen for the *in silico* PCRs, representing COI, cytochrome *b*, ribosomal 12S and ribosomal 16S

Table 1. Primer sets assessed in this study. The approximate fragment length is based upon the length of that region in the *Anguilla anguilla* mitochondrial genome (AP007233.1). The asterisks represent the sequences of the Leray-XT primer set that were simplified by changing inosines to double-base ambiguities to allow an *in silico* assessment with *MFEprimer*. The standard DNA barcode marker for fishes (Ward et al., 2005) is presented for reference.

Primer set	Locus	Primer names	Oligonucleotide 5′–3′	Fragment length (bp)	Reference
Leray-XT	COI	mlCOIintF-XT	GGWACWRGWTGRACWITITAYCCYCC	313	Wangensteen et al. (2018)
		mlCOIintF-XT*	GGWACWRGWTGRACWGTYTAYCCYCC		
		jgHCO2198	TAIACYTCIGGRTGICCRAARAAYCA		
		jgHCO2198*	TAKACYTCWGGRTGRCCRAARAAYCA		
SeaDNA-short		coi.175f	GGAGGCTTTGGMAAYTGRYT	55	This study
		coi.226r	GGGGGAAGAARYCARAARCT		
SeaDNA-mid		coi.175f	GGAGGCTTTGGMAAYTGRYT	130	This study
		coi.345r	TAGAGGRGGGTARACWGTYCA		
Ward-barcode		FishF1	TCAACCAACCACAAAGACATTGGCAC	655	Ward et al. (2005)
		FishR1	TAGACTTCTGGGTGGCCAAAGAATCA		
Minamoto-fish	Cytb	L14912-CYB	TTCCTAGCCATACAYTAYAC	235	Minamoto et al. (2012)
		H15149-CYB	GGTGGCKCCTCAGAAGGACATTTGKCCYCA		
MiFish-U	12S	MiFish-U-F	GTCGGTAAAACTCGTGCCAGC	171	Miya et al. (2015)
		MiFish-U-R	CATAGTGGGGTATCTAATCCCAGTTTG		
MiFish-E		MiFish-E-F	GTTGGTAAATCTCGTGCCAGC	171	Miya et al. (2015)
		MiFish-E-R	CATAGTGGGGTATCTAATCCTAGTTTG		
Taberlet-tele02		Tele02-f	AAACTCGTGCCAGCCACC	167	Taberlet et al. (2018)
		Tele02-r	GGGTATCTAATCCCAGTTTG		
Taberlet-elas02		Elas02-f	GTTGGTHAATCTCGTGCCAGC	171	Taberlet et al. (2018)
		Elas02-r	CATAGTAGGGTATCTAATCCTAGTTTG		
Valentini-tele01		L1848	ACACCGCCCGTCACTCT	63	Valentini et al. (2016)
		H1913	CTTCCGGTACACTTACCATG		
Riaz-V5		12S-V5f	ACTGGGATTAGATACCCC	106	Riaz et al. (2011)
		12S-V5r	TAGAACAGGCTCCTCTAG		
Berry-fish	16S	Fish16sF/D	GACCCTATGGAGCTTTAGAC	219	Berry et al. (2017)
		16s2R	CGCTGTTATCCCTADRGTAACT		

(Table 1). *MFEprimer v2.0* (Qu et al., 2012) was used to perform the *in silico* PCR on the untagged primers. Amplification universality was estimated using the Primer Pair Coverage (PPC) statistic from *MFEprimer*, where $PPC = \frac{Fm}{Fl} \times \frac{Rm}{Rl} \times (1 - CVfr)$, with Fl and Rl the length of the forward and reverse primers, and $CVfr$ the coefficient of variability of matched lengths Fm and Rm to the template. Therefore, a PPC value of 100% indicates complete binding of both primers to a template. The highest PPC value was then selected for each species, and averaged over all species to provide the PPC for each primer set. Predicted non-amplifications with a default 5 bp 3′ binding stability of $> 0\Delta G$ were set to a PPC of 0%. In order for sufficient RAM to be available to complete the analysis of the highly degenerate Leray-XT primer set, the inosine sites were simplified to double-base ambiguities. This was achieved by choosing the most frequent base combination in the mitogenome alignment. None of the altered inosine sites were within 8 bp of the 3′ end of the primer (Table 1).

Taxonomic discrimination (= resolution) was assessed first using all available species from the British Isles fish reference library for each primer set individually, and then secondly on a subset of species for which sequences were present for all of the primer sets. Discrimination as a proportion of the total number of species was calculated following Ficetola et al. (2010): "A taxon unambiguously identified by a primer pair owns a barcode sequence associated to this pair that is not shared by any other taxa".

Primer evaluation *in vitro*

Field sites and traditional fish survey

Five locations in the United Kingdom were surveyed for fishes using eDNA and traditional methods between October and November of 2016. These included: the River Tees, County Durham (54.631327,-1.164447); two sites within the River Esk estuary, North Yorkshire (54.491633,-0.611833; 54.48975,-0.612617); the River Test, Hampshire (50.901563,-1.440836); and Whitsand Bay, Devon (50.329616,-4.243751), The former four are estuarine sites, while the latter is an inshore coastal area, approximately 1 km from shore. Fish sampling in the River Esk estuary was done by duplicate fyke nets (Esk-fyke) and duplicate beach-seine nets (Esk-seine), in different locations. At the River Tees sampling site, duplicate beach-seine netting and two shallow beam trawls were carried out. The River Test site comprised a 24 h fish impingement survey conducted at Marchwood Power Station. Whitsand Bay was surveyed by four otter trawls, as described in McHugh et al. (2011). The variety of fishing techniques used in the different sampling locations are part of the currently ongoing fish monitoring programmes implemented by local collaborating organisations: the Environment Agency, PISCES Conservation Ltd. and the Marine Biological Association. Further details are presented in Supplementary Information.

Water processing and DNA extraction

Three 2 L water sample replicates per site were collected immediately prior to the traditional fish survey commencing, using Nalgene HDPE collection bottles pre-sterilised with a 10% bleach solution. Water was pre-strained with a 250 μm nylon mesh filter to remove debris, if required. After collection, the water samples were put into individual sterile plastic bags, and stored in an ice box while being transported back to the laboratory. Within five hours, each 2 L sample was filtered through an 0.22 μm Sterivex-GP PES filter (Merck Millipore) using a 100 mL polypropylene syringe or a peristaltic pump, and cleared of water. When the full 2 L could not be passed due to filter clogging, the volume of water was recorded. After filtration, the filters were stored at -20°C. DNA was extracted from the filters using the DNeasy PowerSoil DNA Isolation Kit (MoBio/Qiagen), following the manufacturers' protocol, with the addition of an initial 2 h agitation step to promote the release of DNA from the filter, during which the filter membranes were placed in tubes with lysis buffer C1 and garnet beads from the PowerWater Isolation kit and shaken at 65°C. Filtration blank controls were processed in parallel. All processing was carried out in dedicated eDNA extraction laboratories, and equipment and surfaces were regularly cleaned using a 10% bleach solution. The eDNA extraction, pre-PCR preparations and post-PCR procedures were carried out in separate rooms.

PCR and library preparation

Four primer sets were selected to go forward for *in vitro* testing: three COI primer sets (Leray-XT, SeaDNA-short, SeaDNA-mid), and one best-performing primer set from the *in silico* analysis (12S MiFish-U). All PCR amplifications were done in duplicate reactions each with a unique 7/8-mer oligo-tag barcode, differing by at least three bases (Guardiola et al., 2015). In order to increase variability of the amplicon sequences, a variable number (two, three or four) of fully degenerate positions (Ns) were added at the 5′ end of the oligo tags (Wangensteen et al., 2018). For PCR amplification with the newly designed SeaDNA-short and SeaDNA-mid primers, a two-step protocol was used, first using untagged primers, then tagged primers in a second PCR round. The reaction for the first PCR step included 10 μL AmpliTaq Gold 360 Master Mix

For Review Only

(Thermofisher), with 1 μL of each 5 μM forward and reverse primer, 0.16 μL of bovine serum albumin and 10 ng of purified DNA in a total volume of 20 μL per sample. Thermocycling profile for the first step included an initial denaturation at 95°C for 10 minutes, then 40 cycles of 94°C for 30 sec, 47°C for 45 sec and 72°C for 30 sec, and then a final extension of 72°C for 5 minutes. The profile for the second PCR step was identical, except for the annealing temperature being 50°C instead of 47°C. Amplifications were assessed by electrophoresis on a 1.5% agarose gel, and the field and laboratory controls were checked for the presence of amplicons. Between the first and second PCR step, amplicons were purified using MinElute PCR purification columns (QIAGEN) and diluted by a factor of ten prior to being used as a template for the second PCR. After the second PCR, all tagged amplicons were pooled by marker, purified again using MinElute columns and eluted into a total volume of 45 μL, in order to concentrate the amplicons approximately 15 times. For 12S MiFish and Leray-XT we used a one-step procedure with tagged PCR primers, with PCR cycling conditions following Miya et al. (2015) and Wangensteen et al. (2018), respectively. Reagents and volumes were the same as for the two-step protocol.

Libraries (one for each primer set) were built using the PCR-free NEXTflex library preparation kit (BIOO Scientific). The libraries were quantified using the NEBNext qPCR quantification kit (New England Biolabs) and spiked with with 1% PhiX (Illumina). The libraries were sequenced on an Illumina MiSeq platform, using V3 chemistry (2×75 bp paired-end) for the SeaDNA-short library, which was run along with two other libraries from unrelated projects. For the MiFish-U and SeaDNA-mid libraries, V2 chemistry (2×150 bp paired-end) was used, and these were sequenced in the same run. The Leray-XT library was run using V2 chemistry (2×250 bp paired-end) along with another library from an unrelated project.

For Review Only

Bioinformatic processing

Raw sequencing data were converted to fastq format using *bcl2fastq v2.20* (https://support.illumina.com/sequencing/sequenc conversion-software.html). The remaining bioinformatic steps were carried out using *cutadapt v2.3* (Martin, 2011) and *dada2 v1.10.1* (Callahan et al., 2016). Because a PCR-free library preparation kit was used, adapters could have been ligated to either the 5′ or the 3′ end of the amplicon, and in order to take advantage of the Illumina error profiling in the *dada2* denoising step, the sense- and antisense-orientated sequences were first isolated and processed independently. This was achieved using *cutadapt* by filtering the R1 fastq files for reads with the forward PCR primer, and then for those with the reverse PCR primer. The reads were then demultiplexed by tag, followed by primer and adapter trimming. Quality trimming was carried out in *dada2* using default settings, but with read truncation length "truncLen" determined to give an approximate 30 bp overlap between forward and reverse reads. The reads were then denoised, dereplicated, merged, cleaned of chimaeras and reorientated, using the *dada2* workflow. Our reference library sequences for each primer set were used as priors to avoid low abundance but valid sequences being discarded during denoising. A homology filter was then implemented by aligning the ASVs against a hidden Markov model of the expected fragment using *HMMER hmmsearch*, and the non-homologous reads discarded.

Taxonomy assignment of the amplicon sequence variants (ASVs) produced by *dada2* was carried out using a multi-step procedure, incorporating distance-based and phylogenetic methods. First, a preformatted "nt" blast database was downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/db/v5; 21 March 2019). Each ASV sequence was then locally blasted against this database using *blastn v2.9.0* ('-task blastn -evalue 1000 -word_size 11 -max_target_seqs 500'), and the results filtered to obtain a rough taxonomic classification based on the best-scoring blast hit. Next, a more stringent procedure was carried out, with the putative fish sequences extracted from this initial blast result subjected to a second *blastn* search, this time using our curated reference library of British Isles fishes as the blast database (same settings as the "nt" search but with '-word_size 7'). The same reads were then run through the Evolutionary Placement Algorithm (*EPA-ng v0.3.5*, *gappa v0.2.0;* Barbera et al., 2018; Czech and Stamatakis, 2018). Species name(s) were assigned based on either of the following rules: (i) species-level EPA placement same as the best scoring blast hit, with an aligned match length of $\geq 90\%$ of the modal length of the fragment, and an identity of $\geq 97\%$; or (ii) highest likelihood EPA placement same as the best scoring blast hit, with an EPA probability $\geq 90\%$ and blast identity $\geq 90\%$. Rule (i) finds assignments that are congruent between both the *EPA-ng* and *blastn* methods, but rejects assignments with low similarity and short match lengths. Rule (ii) allows for dissimilar hits, but only ones that have a high phylogenetic probability, and which are usually indicative of low abundance variants with errors. Our prior knowledge of the expected fish fauna of the sites was used to set these cut-off values, with the aim of conservatively minimising false positive assignments. The fish reads were also summarised by OTU clustering using *Swarm v2.2.2* (Mahé et al., 2015), with $d = 1$ and the "fastidious" option enabled. This step permitted an evaluation of possible misassigned and unassigned species.

RESULTS

In silico analyses

A total of 531 species were identified as part of the United Kingdom marine and freshwater fish fauna. Of these, 176 names were flagged as "common" species, having been identified as relatively widespread marine or freshwater taxa that are likely to be encountered during survey work of coastal and inland habitats (Henderson, 2014; Kottelat and Freyhof, 2007). The remainder were mostly highly localised species, deep water offshore species, or rare pelagic migrants. The combined reference library for all primer sets, after cleaning, duplicate removal and quality control, comprised 43,366 sequences from 491 total species, and 25,799 sequences from 172 common species.

In terms of reference database coverage for individual primer sets (Table 2), COI primers had the greatest number of reference sequences at 23,911–24,058, covering 91% of species. The "Minamoto-fish" cytochrome *b* set had 15,405 sequences and a species coverage of 65%. Of the ribosomal primer sets, the "Berry-fish" 16S set had the greatest number of sequences at 4,089, with species coverage at 77%. Among the 12S sets, the "Riaz-V5" primers had the greatest number of sequences (2,416; species coverage 69%), while the "Valentini-tele01" set had the fewest sequences (1,699; species coverage 51%). The "MiFish" primers and their variants (MiFish-U/E, Taberlet-tele02, Taberlet-elas02) had 1,904 sequences, and a coverage of 61%. Per species, the average number of reference sequences was greatest for the COI primer sets (mean 49–50; median 24), followed by cytochrome *b* (mean 45; median 7), 16S (mean 9.9; median 4), and then 12S (mean 5.9–6.6; median 2–3). When only the subset of common species was considered, the species coverage increased for all primer sets, as did the average number of sequences per species (Table 2).

In terms of taxonomic discrimination of the fragments obtained from each primer set (Table 2), the proportion of British Isles fish species where all individuals could be unambiguously identified was greatest for the Leray-XT COI fragment at 95%, while the shorter SeaDNA-mid and SeaDNA-short COI fragments resolved 91% and 87% respectively. The cytochrome *b* fragment discriminated 91%. The MiFish fragment had

Table 2. Statistics for reference library coverage, taxonomic discriminatory power, and primer universality as estimated by *in silico* PCR, for twelve primer sets from COI, cytochrome *b*, 16S and 12S. Library coverage is calculated as the number of species for which at least one sequence was available out of the total ($n = 531$) or common species subset ($n = 176$) of British Isles marine and freshwater fishes (proportion in parentheses). Library sequences per species is the mean (median in parentheses) number of sequences available for each species. Taxonomic discrimination is the proportion of species for which all individuals can be unambiguously identified by a unique DNA sequence, with values in parentheses showing the proportion for the subset of species that are shared over all primer sets ($n = 221$ for all; $n = 88$ for common). Primer universality represents the mean Primer Pair Coverage (PPC) percent statistic from *MFEprimer*, and was calculated using the 184 British Isles fish species for which data were available for all species. The standard DNA barcode marker for fishes (Ward et al., 2005) is presented for reference. The highly degenerate Leray-XT primers were simplified to overcome analytical RAM limitations (see Table 1).

Locus	Primer pair	Species subset	Total number sequences	Library species coverage	Library sequences per species	Fragment taxonomic discrimination	Primer % universality (Actinopterygii)	Primer % universality (Elasmobranchii)
COI	Leray-XT	All	24,058	481 (0.91)	50 (24)	0.95 (0.96)	27.8	39
	SeaDNA-mid		24,045	481 (0.91)	50 (24)	0.91 (0.94)	23	22.9
	SeaDNA-short		23,911	481(0.91)	49.7 (24)	0.87 (0.9)	34.5	21.5
	Ward-barcode		23,975	481 (0.91)	49.8 (24)	0.95 (0.97)	6.3	1.2
CYTB	Minamoto-fish		15,405	344 (0.65)	44.8 (6.5)	0.91 (0.91)	13.5	14.4
12S	MiFish-U		1,904	322 (0.61)	5.9 (3)	0.93 (0.91)	71.3	2.4
	Taberlet-tele02		1,904	322 (0.61)	5.9 (3)	0.93 (0.91)	85.3	7.7
	MiFish-E		1,904	322 (0.61)	5.9 (3)	0.93 (0.91)	0.4	39.3
	Taberlet-elas02		1,904	322 (0.61)	5.9 (3)	0.93 (0.91)	0.4	68.8
	Valentini-tele01		1,699	273 (0.51)	6.2 (2)	0.86 (0.85)	68.2	60.4
	Riaz-V5		2,416	364 (0.69)	6.6 (2)	0.79 (0.78)	92.2	11.2
16S	Berry-fish		4,089	411 (0.77)	9.9 (4)	0.89 (0.86)	47.5	0
COI	Leray-XT	Common	12,698	170 (0.97)	74.7 (38.5)	0.97 (1)	23.3	49.3
	SeaDNA-mid		12,639	170 (0.97)	74.3 (37.5)	0.93 (1)	17	29
	SeaDNA-short		12,553	170 (0.97)	73.8 (37.5)	0.93 (1)	32.8	28.9
	Ward-barcode		12,579	170 (0.97)	74 (37.5)	0.97 (1)	6.3	0
CYTB	Minamoto-fish		10,936	143 (0.81)	76.5 (16)	0.94 (1)	13.6	9.1
12S	MiFish-U		941	109 (0.62)	8.6 (3)	0.94 (0.94)	75.6	0
	Taberlet-tele02		941	109 (0.62)	8.6 (3)	0.94 (0.94)	89.3	0
	MiFish-E		941	109 (0.62)	8.6 (3)	0.94 (0.94)	0	52.4
	Taberlet-elas02		941	109 (0.62)	8.6 (3)	0.94 (0.94)	0	82.3
	Valentini-tele01		852	99 (0.56)	8.6 (2)	0.93 (0.94)	67.6	60.4
	Riaz-V5		1,398	143 (0.81)	9.8 (3)	0.85 (0.83)	96.4	0
16S	Berry-fish		2,296	167 (0.95)	13.7 (6)	0.87 (0.91)	50.3	0

the greatest discrimination among the ribosomal primer sets at 93%, with the Berry-fish 16S, Valentini-tele01, and Riaz-V5 pairs having lower rates (89%, 86%, and 79% respectively). When a standardised dataset of species common to all primer sets ($n = 88$) was used, the overall pattern remained similar (Table 2).

In terms of primer universality as estimated by *in silico* PCR for British Isles fish species with comparable data available for all markers ($n = 184$; Table 2), the 12S primer sets targeting actinopterygians had a higher mean PPC than all other markers, at between 68.2% (Valentini-tele01) and 92.2% (Riaz-V5), compared to between 13.5% (cytochrome *b*) and 47.5% (16S). The best performing COI marker for actinopterygians (SeaDNA-short) had a PPC value of 34.5%. For elasmobranchs, three 12S primer pairs had the highest mean PPC values, with Taberlet-elas02 at 68.8%, Valentini-tele01 at 60.4%, and MiFish-E at 39.3%. The 12S Riaz-V5 primers, the cytochrome *b* primers, and the 16S primers, had the lowest PPC values (11.2%, 14.4% and 0% respectively), while the COI primers had PPC values between 21.5% (SeaDNA-short) and 39% (simplified Leray-XT). These patterns remained when only common species were compared (Table 2).

In vitro analyses

Total reads from Illumina sequencing (Table 3) varied between 3.4 million (12S MiFish-U) and 14.3 million (COI SeaDNA-mid). After bioinformatic processing, the proportions of reads retained were 46% (COI SeaDNA-short), 54% (COI Leray-XT), 61% (COI SeaDNA-mid) and 63% (12S MiFish-U). Mean cleaned reads recovered per sampling event (triplicate water samples, duplicate PCR tags; $n = 6$) were: 107,458 (SD = 46,924) for Leray-XT; 290,104 (SD = 118,592) for SeaDNA-mid; 135,804 (SD = 44,993) for SeaDNA-short; and 71,912 (SD = 13,682) for 12S MiFish-U. Supplementary Figure 1 shows distributions of read depths per sample for each site and primer set. The 12S MiFish-U primers provided the greatest proportion of chordate and fish reads (100% and 76% of cleaned reads, respectively), resulting in more than 1.6 million putative fish reads and 156 fish ASVs. From these fish reads, 96% were assigned to 41 species and 67 *Swarm* OTU clusters. A total of 73,377 fish reads comprising 18 *Swarm* OTUs could not be assigned, and in addition to PCR and sequencing artefacts, these likely represent at least nine species not present in the reference library (Supplementary Table 1). For the COI primer sets, chordate reads comprised between 0.2% (Leray-XT) and 6% (SeaDNA-short) of the total cleaned reads, with between 0.1% and 5% putative fish reads comprising between 22 (Leray-XT) and 29 (SeaDNA-short) assigned species. Between 42% (Leray-XT) and 85% (SeaDNA-short) of the putative fish reads were unassigned to species. The non-chordate reads were inferred from the preliminary blast search to consist of DNA from other metazoans (4–10%) and eukaryotes (41–83%), or bacteria (17–59%).

Table 3. Number of reads remaining after seven bioinformatic steps, as well as the number of estimated reads for taxonomic groups (assignments were carried out on the reads remaining after the homology search step 7). Fish reads (putative) are reads assigned to fishes based on the best scoring *blastn* hit using the NCBI "nt" blast database. Fish reads (assigned) are reads assigned to fish species by the stringent taxonomic identification step using *blastn* and *EPA-ng* on our curated reference library. Fish reads (unassigned) are putative fish reads that could not be assigned to species by the stringent taxonomic identification step.

Filtering step	COI Leray-XT	COI SeaDNA-mid	COI SeaDNA-short	12S MiFish-U
Total passing filter	5,967,313	14,291,168	8,881,088	3,436,278
(1) Detect primers	4,828,799	11,535,904	6,428,030	2,776,073
(2) Demultiplex	4,648,811	10,879,223	5,994,815	2,473,594
(3) Trim primers	4,618,236	10,300,907	5,852,555	2,462,936
(4) Quality filter	4,519,097	10,344,024	5,856,045	2,455,532
(5) Merge	3,395,057	9,658,709	4,804,502	2,383,162
(6) Remove chimaeras	3,225,240	9,404,746	4,416,647	2,271,541
(7) Homology search	3,223,743	8,703,109	4,074,123	2,157,365
Bacteria	1,476,994	1,388,681	2,242,220	4
Eukaryota	1,745,295	7,294,762	1,815,928	2,157,361
Metazoa	321,590	1,161,769	412,871	2,157,361
Chordata	6,351	337,901	250,650	2,157,361
Fish (putative)	2,371	234,219	193,593	1,637,728
Fish (assigned)	1,368	109,486	30,026	1,564,351
Fish (unassigned)	1,003	124,733	163,567	73,377

Per sampling location the 12S MiFish-U primer set detected a consistently greater number of total species across sites than the COI markers, at between 2.2 (River Test) and 2.6 (Whitsand Bay) fold higher (Figure 1). The SeaDNA-short primers detected a greater number of species than both the SeaDNA-mid and Leray-XT primers, except at the River Tees site where SeaDNA-mid detected one more.

In terms of reproducibility (Figure 2), the 12S MiFish-U primer set showed a greater proportion of shared

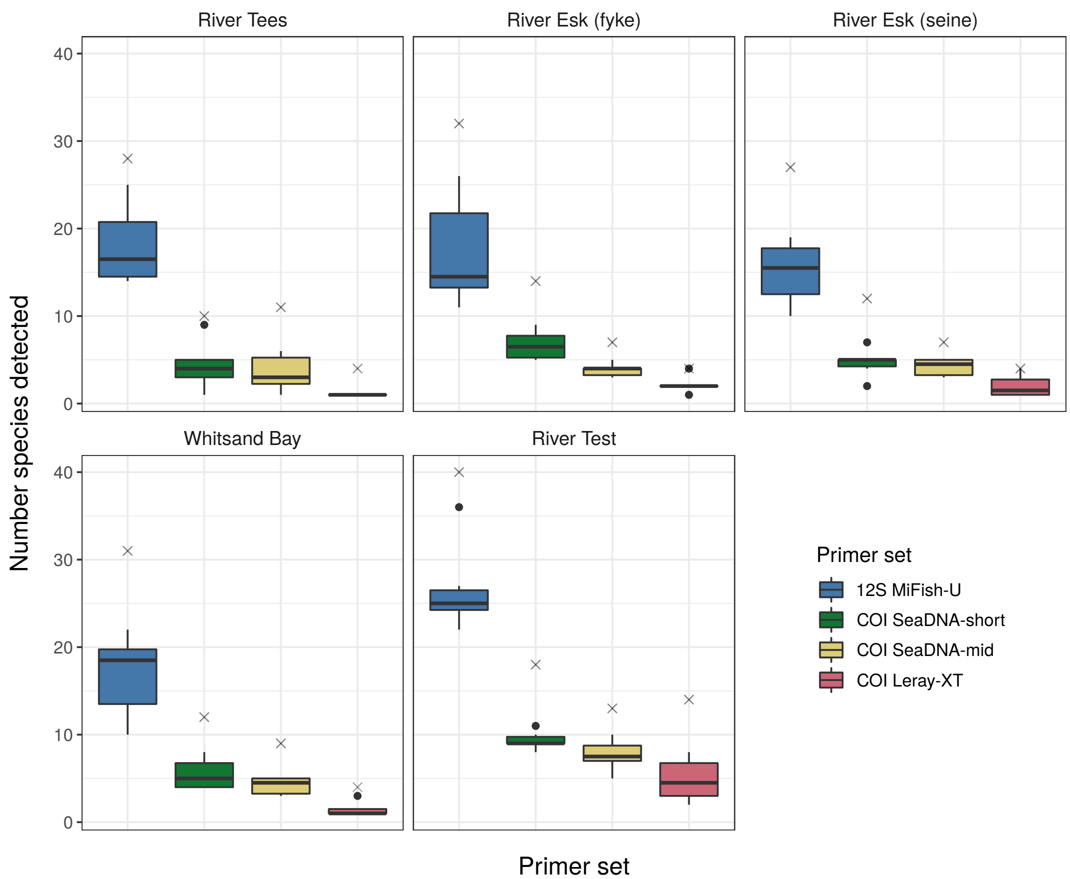


Figure 1. Fish species richness as estimated by four primer pairs at five sampling locations. Per primer-location combination there are three water sample replicates and two uniquely tagged PCR replicates ($n = 6$). The horizontal represents the median value, the boxes represent the 25–75th percentiles, the whiskers represent the values less than 1.5 times the interquartile range, dots represent the outlying data points, and crosses represent the cumulative number of species.

species—the top ten species by read abundance at each location—amplified across water sample and PCR replicates, with a 71% mean reproducibility over all sampling locations. The COI primer sets had mean reproducibility values of 36% (SeaDNA-short), 29% (SeaDNA-mid) and 12% (Leray-XT).

When compared to traditional survey methods—with the freshwater species omitted from the eDNA results as they were not expected to be found on the traditional fish surveys of the estuarine and coastal habitats—the 12S MiFish-U primer set showed the greatest congruence (Figure 3), at between 15% (Whitsand Bay) and 54% (River Test). The COI primers were between 9% (Leray-XT) and 13% (SeaDNA-short) congruent overall. The MiFish-U primer set also amplified a greater number of marine/estuarine species to the traditional survey methods at all locations except for Whitsand Bay (26 versus 23 species). The COI primer sets amplified fewer marine/estuarine species than the traditional surveys in all cases, except for the SeaDNA-short primer set at the River Tees and River Esk sites. For each site survey, reads per species (eDNA survey) and individuals per species (traditional survey) are presented in Supplementary Tables 2–6.

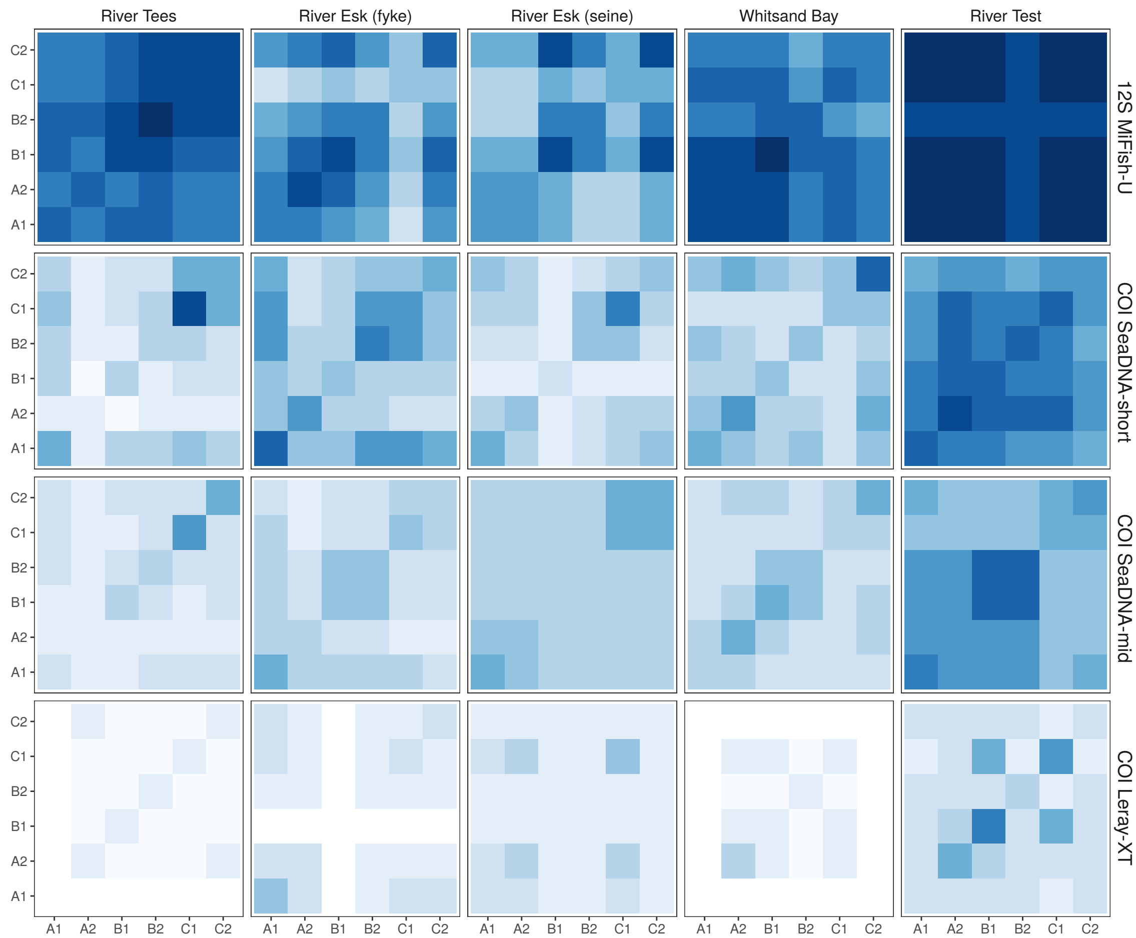


Figure 2. Reproducibility heatmaps of four primer pairs at five sampling locations for the top ten fish species found at each location by read abundance. Letters A, B, and C represent the three water samples taken, while numbers 1 and 2 represent the independent PCR reactions with uniquely tagged primers. There are ten shades showing 10% increments. The darkest shade shows a reproducibility of 100%, i.e. reads from all of the ten species were common to both PCRs. The lightest shade shows 0% reproducibility, i.e. none of the species were present in both of the PCRs. Diagonals show the proportion of the top ten species amplified in that single PCR.

DISCUSSION

A single metabarcoding marker for fishes?

Of arguably the greatest importance in the ability of metabarcoding to answer a particular question, is that of the choice of marker and primer (Alberdi et al., 2018; Elbrecht and Leese, 2017; Clarke et al., 2017; Deagle et al., 2014; Valentini et al., 2016). The ideal genetic marker for eDNA metabarcoding marker should be flexible, allowing different primer sets to target different taxonomic groups, but requiring only a single reference library. Each individual primer set must also be designed with the following qualities: (i) it must be universal, i.e. amplifying a large proportion of the target taxonomic group; (ii) it must be specific, i.e. it must not amplify other taxa at the expense of the target group; (iii) it must be unbiased, i.e. not preferentially amplifying a subset of the target group; (iv) it must be discriminatory, i.e. the DNA fragment recovered should differentiate at the appropriate taxonomic level for the question; and (v) it must be replete, i.e. associated

Figure 3. Overlap between fish species found by eDNA metabarcoding (red) and traditional fish surveying (blue). Sizes of circles are proportional only within each primer-location comparison, and not between. Numbers represent number of species in each set. Only marine and estuarine species are shown; freshwater species recorded by the eDNA surveys were removed to allow an equivalent comparison.

with a reference library enabling identifications within the target taxonomic group. Here, we assess these characteristics for COI, cytochrome *b*, 12S, and 16S primer sets using the example of marine and freshwater fishes from the British Isles.

Which primers have the best reference library?

In terms of reference libraries, the COI primers were substantially better endowed than all other marker genes, with between 1.6 times (cytochrome *b*) and 14 times (Valentini-tele01) more public sequence data available for all species. This was also reflected in the common species coverage, at up to 97% for COI. The 16S (95%), cytochrome *b* (81%), and 12S Riaz-V5 libraries (81%) were also well developed for common species, but coverage for other 12S primer sets was lower, at 56–62%. A reference library with broad taxonomic depth will allow inferences beyond a comparison of anonymous MOTUs, thereby leveraging the wealth of scientific information that a taxonomic name brings with it (Ward et al., 2009). Deep coverage in the COI reference library—i.e. the number and geographic distribution of sequences per species—also has advantages in terms of potential for population level assignments, and for flagging spuriously identified sequences (due to the lesser weight of evidence from the low numbers of sequences, misidentifications were harder to confirm for 12S during the quality control step). Furthermore, in terms of voucher specimen and location data etc, much of the ribosomal data on GenBank are not validated to the same standard as COI data on BOLD are (Ward et al., 2009). However, it is important to remember that despite the success of 15 years of the DNA barcode initiative producing COI coverage spanning the majority of northern European fish species, the BOLD database still remains seriously underdeveloped for many other taxonomic groups such as marine invertebrates (Bucklin et al., 2011; Leray and Knowlton, 2016).

Which primers best discriminate species?

In terms of the discriminatory power for our dataset of British Isles fish species, all primer sets gave a resolution above 90% except for SeaDNA-short (COI), Valentini-tele01 (12S), Riaz-V5 (12S) and Berry-fish (16S). Predictably, the longer COI fragments resolved more species than the shorter ones, at 95% for the 313 bp Leray-XT and 87% for the 55 bp SeaDNA-short fragment. The 12S primers did not show this pattern as clearly, with the shorter Valentini-tele01 fragment having a better taxonomic resolution (86%) than the longer Riaz-V5 fragment (79%); the longest, MiFish-U/E and Taberlet-tele02/elas02 primers, had the greatest species resolution at 93%. While discriminatory power may depend on the range of species in that particular library, the observed patterns held up when a dataset of sequences that were shared for all primer sets was used. Discriminatory power also tended to remain the same or increase when only the common species were considered, most likely because rare congeners were excluded.

For Review Only

Which primers are most universal?

Primer universality as estimated by *in silico* PCR varied greatly. Our results show that the metabarcode primers targeting protein-coding genes—COI and cytochrome *b*—are likely to exhibit a greater degree of species-level primer bias (i.e. lower universality) than ribosomal 12S and 16S, as indicated by the lower mean PPC values; a mean PPC of 96% was estimated for common actinopterygian species amplified with the Riaz-V5 primers. Previous studies have also reported or predicted less primer bias with rRNA targets than protein coding ones (Clarke et al., 2014; Elbrecht et al., 2016; Deagle et al., 2014; Marquina et al., 2019). It is also important to note again that due to the high level of degeneracy the Leray-XT primers were simplified to overcome RAM limitations of the analysis, and therefore the value presented is likely to be an underestimate of their true potential, as highly degenerate COI primers have been shown to reduce bias substantially (Marquina et al., 2019).

Regarding higher level taxonomic bias, for the 12S and 16S primers tested here, no set except Valentini-tele01 appeared suited to amplify actinopterygians and elasmobranchs equally. The COI primers were, however, unbiased in regard to higher taxonomic group. The MiFish primers and the Taberlet et al. (2018) variants of the same sets were both published with actinopterygian (MiFish-U) and elasmobranch (MiFish-E) versions, due to a number of mismatches in the conserved regions (Miya et al., 2015). Unsurprisingly, both of these performed substantially better for their respective taxa. The Taberlet et al. (2018) primers were also predicted here to exhibit substantially less species-level primer bias than the original MiFish versions, for both elasmobranchs and actinopterygians.

Many studies computationally predict primer amplification by the number of mismatches between primer and template (e.g. Riaz et al., 2011), or by the number of mismatches and their type and position (e.g. Elbrecht et al., 2017), but often do not fully consider the thermodynamics of a primer-template reaction. We used the thermodynamics-based PCR simulation implemented in MFEprimer (Qu et al., 2012), but regardless of whether this method is more realistic or accurate than alternative methods, it is important to remember that these are predicted amplifications, and were used here to compare relative performances between primer sets. Therefore, the lower values estimated do not represent amplification failure *per se*, but rather are indicative of increased bias associated with that primer set (Deagle et al., 2014). For example, the standard COI DNA barcode primers for fishes (Ward-barcode) had a very low PPC, but these are tried-and-tested primers for amplifying a wide range of fish taxa in standard PCR for Sanger sequencing (Ward et al., 2005). The use of mock communities is an important step in quality controlling an assay if primer bias is suspected (Piñol et al., 2015; Elbrecht and Leese, 2017; Bista et al., 2018), but *in silico* PCR has been demonstrated to be an effective proxy in its absence (Clarke et al., 2014).

We used the results of our *in silico* analyses to inform our choices for the *in vitro* experiments. All COI primer sets were selected for testing *in vitro* because of the advantages in terms of reference library and taxonomic discrimination. We chose only one 12S set for comparison, and here we chose the MiFish-U primer pair because this pair had better predicted universality for actinopterygians and more reference sequences available than the Valentini-tele01 primers, and greater taxonomic discrimination than the Riaz-V5 primers. Due to the better predicted universality of the Taberlet-tele02 primer set compared to MiFish-U, these would have been chosen had they been publicly available at the time the experiment was implemented. Despite the well developed reference libraries and good taxonomic discrimination, we did not select cytochrome *b* or 16S because of the lower predicted universality of these primers in comparison to 12S.

Which primers are the most specific?

Despite having the fewest total raw reads, the MiFish-U primer set produced the greatest number and proportion of usable fish reads (76% of processed reads, 48% of raw reads), the greatest overall species richness (41 species), and the greatest proportion of fish reads that were assigned to species (96%). The COI primers amplified a very low proportion of chordate and fish reads compared to the overall sequencing depth (maximum 5% of cleaned reads were fishes). The majority of the SeaDNA-short and SeaDNA-mid reads were estimated by preliminary blast search to have come from bacteria or non-metazoan eukaryotes (86–90%).

That the highly degenerate Leray-XT primers produced a low proportion of fish reads is unsurprising given that previous studies on environmental samples using degenerate COI primers have demonstrated that they can amplify widely beyond their target taxa, and can produce large proportions of unassigned reads (Macher et al., 2018; Stat et al., 2017; Lim et al., 2016; Singer et al., 2019). The proportion of bacterial reads are generally lower when metabarcoding bulk organismal samples, however, with most reads belonging to metazoans (Wangensteen et al., 2018; Leray and Knowlton, 2015; Macher et al., 2018). More surprising was the poor specificity of the SeaDNA-short and SeaDNA-mid primers, which were designed to target fishes, and with minimal degeneracy. These data are, however, consistent with those of an analysis of shark diversity by Bakker et al. (2017), who used COI mini-barcode primers designed on sharks, and reported a similar level of non-specific amplification.

The cause of this non-specific amplification is likely to be the extensive homoplasy (nucleotide convergence) apparent in the mutationally saturated COI gene and its homologs. Siddall et al. (2009) demonstrated that metazoan-targeted COI primers are likely to co-amplify many marine prokaryote groups—

gammaproteobacteria being a particularly diverse and abundant lineage (Sunagawa et al., 2015)—thereby compromising the specificity of these primer sets. Optimisation of PCR protocols or library preparation methods may increase specificity of the assay (Siddall et al., 2009), but it is probably unlikely that it can increase to a level that makes the proportion of usable reads viable for eDNA metabarcoding of targeted taxonomic groups. While this phenomenon was first observed in marine prokaryotes, studies on freshwater and soil faunas have shown a similar pattern, also with large numbers of unassigned reads (Lim et al., 2016; Yang et al., 2014).

Which primers give the most reproducible results?

The low number of usable fish reads for the COI primers is reflected in the reproducibility of the assays across water sample and PCR replicates. For the most frequently amplified species at each site, the COI primers were less consistent than 12S MiFish-U overall. Low quantities of template DNA and stochasticity in early PCR cycles is a known factor in causing poor reproducibility (Leray and Knowlton, 2017; Alberdi et al., 2018; Collins et al., 2018), and can be ameliorated by performing multiple PCR technical replicates (Ficetola et al., 2015). We show that this effect is exacerbated when primer specificity is low and non-target organisms are abundant, as is the case in highly diverse environmental samples such as seawater. For many applications repeatability between assays or sampling sites is a requirement, such as the detection of an endangered or invasive species (Grey et al., 2018). Our results, even considering only the top ten common species, show that detectability can vary between sites with the same genetic marker, and that many more than two PCRs will be required if the rare species are to be detected across multiple PCR and water sample replicates (Dopheide et al., 2018).

Species richness estimates at all sampling sites were greatest with 12S MiFish-U, and this was despite the deficiencies in the reference library, at only 61% species coverage. For example, species including the European plaice (*Pleuronectes platessa*) and European flounder (*Platichthys flesus*)—both common fishes present at all sampling locations—were missing from the reference library and therefore not represented when comparing with the traditional fish surveys. Most of the large number of reads that were assigned to American plaice, *Hippoglossoides platessoides* ($n = 198,445$), were likely misassigned and actually belong to European plaice and flounder (Supplementary Table 1). The *Swarm* OTU analysis showed a greater number of clusters (67) than assigned species (41), also suggesting that some species missing from the reference library are likely to have been misassigned. While a small number of the 73,377 unassigned 12S fish reads were low abundance sequences derived from artefacts, almost all could be could be inferred by phylogenetic analysis or by similarity to geographically disjunct congeners, to belong to at least eight species that were known to be missing from the reference library (Supplementary Table 1). Despite this major handicap, the 12S MiFish primers remained superior to COI in terms of congruence with the traditional fish surveys, by recovering a greater overlap of species in all cases. The 12S MiFish primers amplified more species than the traditional surveys at all sites, except Whitsand Bay. This was mainly due to the underrepresentation of the fauna of that site in the 12S reference library, with over half of the surveyed species absent from the library, and a higher proportion of elasmobranchs (five species) than the other sites, which the MiFish-U primers fail to amplify. Overall, no species that were recorded in the traditional surveys were missing from the COI reference libraries, but eighteen species were missing from the 12S MiFish library (37%). The low numbers of species recorded by the traditional surveys at the Esk and Tees sites in comparison to the Whitsand Bay and River Test sites, is

For Review Only

partly due to the inherently less diverse fauna of these northerly estuaries, as well as a reflection of the survey techniques, with fyke and seine netting likely to detect fewer species than otter trawling (Whitsand Bay) or a 24 h power station impingement (River Test). It should also be noted that there is no *a priori* assumption that the eDNA and traditional survey data will be completely congruent, as most fish survey methods are imperfect, sampling a moving target of diversity and abundance over difficult-to-define spatio-temporal points. For example, eDNA can be transported in or out by tides, while some species are difficult to sample using particular fishing gears, due to effects of size, behaviour or abundance. Therefore, overlap between eDNA and traditional survey data is best interpreted as a relative measure between the primer sets.

CONCLUSIONS

While PCR-free methods are being actively investigated, it is clear that despite the limitations in quantification, the majority of environmental metabarcoding will be based around amplicon sequencing, at least for the medium term (Wilcox et al., 2018; Stat et al., 2017; Bista et al., 2018; Creer et al., 2016). Particularly important for regulatory applications, or where researchers wish to compare results over time or between studies, some degree of standardisation is desirable (Hering et al., 2018). Our results—and those of previous studies using similar primer sets (Macher et al., 2018; Stat et al., 2017; Lim et al., 2016; Bakker et al., 2017; Yang et al., 2014; Jeunen et al., 2018; Singer et al., 2019)—show that environmental metabarcoding for restricted taxonomic groups using degenerate COI primers results in excessive volumes of "wasted" sequencing effort. This co-amplification of prokaryotic and non-target eukaryotic DNAs and subsequent lack of specificity is due to the nature of mutation patterns in COI (Siddall et al., 2009). Therefore, while we fully support the arguments presented by Andújar et al. (2018) regarding the overall advantages of COI as a bulk-sample metabarcoding marker, we find it difficult to recommend for metabarcoding environmental samples with low target template concentrations and high microbial and plankton diversity, such as natural water bodies.

While the use of multiple primer sets and markers are probably required for a comprehensive view of total biodiversity (Stat et al., 2017; Drummond et al., 2015), for specific taxonomic groups such as fishes a single assay should be a feasible proposition. Unfortunately, no single 12S primer set was shown to be optimal for eDNA fish surveys. The MiFish-U primer set—and *in silico*, the Taberlet et al. (2018) modified versions—performed well in terms of specificity, discriminatory power, and reproducibility. Despite this, MiFish-U is not universal for all fishes, because a separate MiFish-E assay is required to amplify elasmobranchs. The MiFish reference library was also inadequate in this case, missing large numbers of common taxa. The Valentini-tele01 primer set amplifies actinopterygians and elasmobranchs in a single assay, but suffers from an even more poorly populated reference library than MiFish-U, and weaker taxonomic resolution. The Riaz-V5 primers had the most complete reference library of the 12S primer pairs, but also do not amplify elasmobranchs and have the poorest discriminatory power.

Because no single alternative primer set to COI will be optimal for all applications, it is clear that the current DNA barcode reference libraries will need to be augmented with data from multiple mitochondrial regions to enable their wider utility for vertebrate metabarcoding. However, rather than sequencing individual 12S regions on an ad hoc basis, a better solution is to generate whole mitochondrial genomes which can act as an extended or linking barcode if sequenced from the same collection material (Coissac et al., 2016; Collins

For Review Only

and Cruickshank, 2014). Low coverage genome skimming techniques now produce high quality mitogenomes, and are compatible with existing—frequently ethanol-based—tissue collections, and therefore will not require the recollection of specimens (Linard et al., 2016; Gillett et al., 2014). Environmental DNA techniques could potentially be the default survey methodology for aquatic ecosystems, but the existing gap between recovered genotypes and their corresponding phenotypic and historical data can only be filled with substantially more comprehensive reference libraries.

ACKNOWLEDGEMENTS

For assistance with logistics, eDNA sampling and providing the traditional survey data, we thank: Peter Henderson and Robin Somes from Pisces Conservation Ltd.; Tony Gray, Hannah Westoby and the staff at the Environment Agency in Newcastle Upon Tyne; and Aisling Smith and Sophie Rainbird from the Marine Biological Association in Plymouth. This work was co-funded by the Natural Environment Research Council grants NE/N005759/1 and NE/N005937/1 (project SeaDNA), and the University of Salford R&E strategy funding.

DECLARATION OF INTEREST

The authors declare that they have no competing interests.

DATA ACCESSIBILITY

The full reference library and code to reproduce it can be found at https://doi.org/10.6084/m9.figshare.7464521. Code to reproduce all other analyses in this study can be found at https://doi.org/10.6084/m9.figshare.8291660.

REFERENCES

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., and Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, 9:134–147, DOI: 10.1111/2041-210X.12849.

Andruszkiewicz, E. A., Starks, H. A., Chavez, F. P., Sassoubre, L. M., Block, B. A., and Boehm, A. B. (2017). Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. *PLoS ONE*, 12:e0176343, DOI: 10.1371/journal.pone.0176343.

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., and Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the Metazoa. *Molecular Ecology*, 27:3968–3975, DOI: 10.1111/mec.14844.

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., Hertler, H., Mouillot, D., Vigliola, L., and Mariani, S. (2017). Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. *Scientific Reports*, 7:16886, DOI: 10.1038/s41598-017-17150-2.

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2018). EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, DOI: 10.1101/291658.

Berry, T. E., Osterrieder, S. K., Murray, D. C., Coghlan, M. L., Richardson, A. J., Grealy, A. K., Stat, M., Bejder, L., and Bunce, M. (2017). Metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (Neophoca cinerea). *Ecology and Evolution*, pages 1–19, DOI: 10.1002/ece3.3123.

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., and Creer, S. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, 18:1020–1034, DOI: 10.1111/1755-0998.12888.

Boettiger, C., Lang, D. T., and Wainwright, P. C. (2012). rfishbase: exploring , manipulating and visualizing FishBase data from R. *Journal of Fish Biology*, 81:2030–2039, DOI: 10.1111/j.1095-8649.2012.03464.x.

Boyer, S., Brown, S. D. J., Collins, R. A., Cruickshank, R. H., Lefort, M.-C., Malumbres-Olarte, J., and Wratten, S. D. (2012). Sliding window analyses for optimal selection of mini-barcodes, and application to 454-pyrosequencing for specimen identification from degraded DNA. *PLoS ONE*, 7:e38215, DOI: 10.1371/journal.pone.0038215.

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C. J., and Cruickshank, R. H. (2012). Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, 12:562–565, DOI: 10.1111/j.1755-0998.2011.03108.x.

Bucklin, A., Steinke, D., and Blanco-Bercial, L. (2011). DNA Barcoding of Marine Metazoa. *Annual Review of Marine Science*, 3:471–508, DOI: 10.1146/annurev-marine-120308-080950.

Bylemans, J., Gleeson, D. M., Hardy, C. M., and Furlan, E. (2018). Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). *Ecology and Evolution*, 8:8697–8712, DOI: 10.1002/ece3.4387.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13:581–583, DOI: 10.1038/nmeth.3869.

Chamberlain, S. (2018). bold: Interface to Bold Systems API. https://cran.r-project.org/package=bold.

Chamberlain, S. and Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data. *PeerJ PrePrints*, DOI: 10.7287/peerj.preprints.3304v1.

Chamberlain, S., Foster, Z., Bartomeus, I., LeBauer, D., Black, C., and Harris, D. (2018). traits: species trait data from around the web. https://github.com/ropensci/traits.

Clarke, L. J., Beard, J. M., Swadling, K. M., and Deagle, B. E. (2017). Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*, 7:873–883, DOI: 10.1002/ece3.2667.

Clarke, L. J., Soubrier, J., Weyrich, L. S., and Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, 14:1160–1170, DOI: 10.1111/1755-0998.12265.

Coissac, E., Hollingsworth, P. M., Lavergne, S., and Taberlet, P. (2016). From barcodes to genomes: Extending

For Review Only

the concept of DNA barcoding. *Molecular Ecology*, 25:1423–1428, DOI: 10.1111/mec.13549.

Collins, R. A. and Cruickshank, R. H. (2014). Known knowns, known unknowns, unknown unknowns and unknown knowns in DNA barcoding: A comment on Dowton et al. *Systematic Biology*, 63:1005–1009, DOI: 10.1093/sysbio/syu060.

Collins, R. A., Wangensteen, O. S., Sims, D. W., Genner, M. J., and Mariani, S. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, 1:185, DOI: 10.1038/s42003-018-0192-6.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, K., Potter, C., and Bik, H. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, 56:68–74, DOI: 10.1111/2041-210X.12574.

Czech, L. and Stamatakis, A. (2018). Scalable methods for post-processing , visualizing , and analyzing phylogenetic placements. *bioRxiv*, pages 1–36, DOI: 10.1101/346353.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., and Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, 10:20140562, DOI: 10.1098/rsbl.2014.0562.

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., and Newcomb, R. D. (2018). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution*, DOI: 10.1111/2041-210X.13086.

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C. M., Heled, J., Ross, H. A., Tooman, L., Grosser, S., Park, D., Demetras, N. J., Stevens, M. I., Russell, J. C., Anderson, S. H., Carter, A., and Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, 4:46, DOI: 10.1186/s13742-015-0086-1.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14:755–763, DOI: 10.1093/bioinformatics/14.9.755.

Elbrecht, V. and Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, 5:1–11, DOI: 10.3389/fenvs.2017.00011.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., Coissac, E., Boyer, F., and Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, 4:e1966, DOI: 10.7717/peerj.1966.

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., and Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 8:1265–1275, DOI: 10.1111/2041-210X.12789.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., and Shehzad, W. (2010). An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, 11:434, DOI: 10.1186/1471-2164-11-434.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., and Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15:543–556, DOI: 10.1111/1755-0998.12338.

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3:294–299.

Gillett, C. P. D. T., Crampton-Platt, A., Timmermans, M. J. T. N., Jordal, B., Emerson, B. C., and Vogler, A. P. (2014). Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, 31:2223–2237, DOI: 10.1093/molbev/msu154.

Grey, E. K., Bernatchez, L., Cassey, P., Deiner, K., Deveney, M., Howland, K. L., Lacoursière-Roussel, A., Leong, S. C. Y., Li, Y., Olds, B., Pfrender, M. E., Prowse, T. A., Renshaw, M. A., and Lodge, D. M. (2018). Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys. *Scientific Reports*, 8:2–11, DOI: 10.1038/s41598-018-27048-2.

Guardiola, M., Uriz, M. J., Taberlet, P., Coissac, E., Wangensteen, O. S., and Turon, X. (2015). Deep-sea, deep-sequencing: Metabarcoding extracellular DNA from sediments of marine canyons. *PLoS ONE*, 10:e0139633, DOI: 10.1371/journal.pone.0139633.

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., and Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25:3101–3119, DOI: 10.1111/mec.13660.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270:313–321, DOI: 10.1098/rspb.2002.2218.

Henderson, P. A. (2014). *Identification Guide to the Inshore Fish of the British Isles*. Pisces Conservation Ltd., Pennington, UK.

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., and Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, 138:192–205, DOI: 10.1016/j.watres.2018.03.003.

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., and Nishida, M. (2013). Mitofish and mitoannotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, 30:2531–2540, DOI: 10.1093/molbev/mst141.

Jeunen, G.-J., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., Bunce, M., and Gemmell, N. J. (2018). Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. *Molecular Ecology Resources*, DOI: 10.1111/1755-0998.12982.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B. C., and Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16:1245–1257, DOI: 10.1111/ele.12162.

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., Rubenstein, D. I., Wang, W., and Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences*, 112:8019–8024, DOI:

10.1073/pnas.1503283112.
Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780, DOI: 10.1093/molbev/mst010.
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., and Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28:1647–1649, DOI: 10.1093/bioinformatics/bts199.
Kelly, R. P., Closek, C. J., O'Donnell, J. L., Kralj, J. E., Shelton, A. O., and Samhouri, J. F. (2017). Genetic and manual survey methods yield different and complementary views of an ecosystem. *Frontiers in Marine Science*, 3:1–11, DOI: 10.3389/fmars.2016.00283.
Kelly, R. P., Port, J. A., Yamahara, K. M., and Crowder, L. B. (2014a). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE*, 9:e86175, DOI: 10.1371/journal.pone.0086175.
Kelly, R. P., Port, J. A., Yamahara, K. M., Martone, R. G., Lowell, N., Thomsen, P. F., Mach, M. E., Bennett, M., Prahler, E., Caldwell, M. R., and Crowder, L. B. (2014b). Harnessing DNA to improve environmental management. *Science*, 344:1455–1456, DOI: 10.1126/science.1251156.
Kottelat, M. and Freyhof, J. (2007). *Handbook of European Freshwater Fishes*. Publications Kottelat, Cornol, Switzerland.
Leray, M. and Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, 112:2076–2081, DOI: 10.1073/pnas.1424997112.
Leray, M. and Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371:20150331, DOI: 10.1098/rstb.2015.0331.
Leray, M. and Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, 5:e3006, DOI: 10.7717/peerj.3006.
Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., and Machida, R. J. (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, DOI: 10.1186/1742-9994-10-34.
Lim, N. K. M., Tay, Y. C., Tan, J. W. T., Kwik, J. T. B., Baloğlu, B., Meier, R., and Yeo, D. C. J. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and communities. *Royal Society Open Science*, 3:160635, DOI: 10.1098/rsos.160635.
Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A., and Vogler, A. P. (2016). Lessons from genome skimming of arthropod-preserving ethanol. *Molecular Ecology Resources*, 16:1365–1377, DOI: 10.1111/1755-0998.12539.
Macher, J. N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, C. D., and Leese, F. (2018). Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome c oxidase I primers. *Molecular Ecology Resources*, 18:1456–1468, DOI: 10.1111/1755-0998.12940.
Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and

high-resolution amplicon clustering. *PeerJ*, 3:e1420, DOI: 10.7717/peerj.1420.

Mariani, S., Griffiths, A. M., Velasco, A., Kappel, K., Jerome, M., Perez-Martin, R. I., Schroder, U., Verrez-Bagnis, V., Silva, H., Vandamme, S. G., Boufana, B., Mendes, R., Shorten, M., Smith, C., Hankard, E., Hook, S. A., Weymer, A. S., Gunning, D., and Sotelo, C. G. (2015). Low mislabeling rates indicate marked improvements in European seafood market operations. *Frontiers in Ecology and the Environment*, 13:536–540, DOI: 10.1890/150119.

Marquina, D., Andersson, A. F., and Ronquist, F. (2019). New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Molecular Ecology Resources*, 19:90–104, DOI: 10.1111/1755-0998.12942.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17:10–12, DOI: 10.14806/ej.17.1.200.

McHugh, M., Sims, D. W., Partridge, J. C., and Genner, M. J. (2011). A century later: Long-term change of an inshore temperate marine fish assemblage. *Journal of Sea Research*, 65:187–194, DOI: 10.1016/j.seares.2010.09.006.

Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M. N., and Kawabata, Z. (2012). Surveillance of fish species composition using environmental DNA. *Limnology*, 13:193–197, DOI: 10.1007/s10201-011-0362-4.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., and Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*, 2:150088, DOI: 10.1098/rsos.150088.

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Der Sarkissian, C., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, F., Schmidt, A. M. Z., Gilbert, M. T. P., Hansen, A. J., Orlando, L., and Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370:20130383, DOI: 10.1098/rstb.2013.0383.

Piñol, J., Mir, G., Gomez-Polo, P., and Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, 15:819–830, DOI: 10.1111/1755-0998.12355.

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., and Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, 25:527–541, DOI: 10.1111/mec.13481.

Qu, W., Zhou, Y., Zhang, Y., Lu, Y., Wang, X., Zhao, D., Yang, Y., and Zhang, C. (2012). MFEprimer-2.0: A fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Research*, 40:205–208, DOI: 10.1093/nar/gks552.

Ratnasingham, S. and Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7:355–364, DOI: 10.1111/j.1471-8286.2006.01678.x.

Rees, H. C., Maddison, B. C., Middleditch, D. J., Patmore, J. R. M., and Gough, K. C. (2014). The detection of aquatic animal species using environmental DNA - a review of eDNA as a survey tool in ecology.

For Review Only

Journal of Applied Ecology, 51:1450–1459, DOI: 10.1111/1365-2664.12306.

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., and Coissac, E. (2011). EcoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, 39:e145, DOI: 10.1093/nar/gkr732.

Shaw, J. L. A., Clarke, L. J., Wedderburn, S. D., Barnes, T. C., Weyrich, L. S., and Cooper, A. (2016). Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, 197:131–138, DOI: 10.1016/j.biocon.2016.03.010.

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., and Hajibabaei, M. (2015). A DNA mini-barcoding system for authentication of processed fish products. *Scientific Reports*, 5:15894, DOI: 10.1038/srep15894.

Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., and Erséus, C. (2009). Barcoding bamboozled by bacteria: convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, 58:445–451, DOI: 10.1093/sysbio/syp033.

Singer, G. A. C., Fahner, N. A., Barnes, J. G., Mccarthy, A., and Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Scientific Reports*, 9:5991, DOI: 10.1038/s41598-019-42455-9.

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S. T., Sigsgaard, E. E., and Hellström, M. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution*, 8:635–645, DOI: 10.1111/2041-210X.12683.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology*, 57:758–771, DOI: 10.1080/10635150802429642.

Stat, M., Huggett, M. J., Bernasconi, R., Dibattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., and Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, 7:0–22, DOI: 10.1038/s41598-017-12501-5.

Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., and Harvey, E. S. (2018). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. *Conservation Biology*, 33:196–205, DOI: 10.1111/cobi.13183.

Stoeckle, M. Y., Soboleva, L., and Charlop-Powers, Z. (2017). Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. *PLoS ONE*, 12:e0175186, DOI: 10.1371/journal.pone.0175186.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sieracki, M., Velayoudon, D., Bowler, C., De Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, 348:1–10, DOI: 10.1126/science.1261359.

For Review Only

Taberlet, P., Bonin, A., Zinger, L., and Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford University Press, Oxford, DOI: 10.1093/oso/9780198767220.001.0001.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21:2045–2050, DOI: 10.1111/j.1365-294X.2012.05470.x.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., and Willerslev, E. (2012a). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS ONE*, 7:e41732, DOI: 10.1371/journal.pone.0041732.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., and Willerslev, E. (2012b). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, 21:2565–2573, DOI: 10.1111/j.1365-294X.2011.05418.x.

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., and Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS ONE*, 11:e0165252, DOI: 10.1371/journal.pone.0165252.

Thomsen, P. F. and Willerslev, E. (2015). Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, 183:4–18, DOI: 10.1016/j.biocon.2014.11.019.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40:e115, DOI: 10.1093/nar/gks596.

Ushio, M., Murakami, H., Masuda, R., Sado, T., Miya, M., Sakurai, S., Yamanaka, H., Minamoto, T., and Kondoh, M. (2018). Quantitative monitoring of multispecies fish environmental DNA using high-throughput sequencing. *Metabarcoding and Metagenomics*, 2:1–15, DOI: 10.3897/mbmg.2.23297.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., Peroux, T., Crivelli, A. J., Olivier, A., Acqueberge, M., Le Brun, M., Møller, P. R., Willerslev, E., and Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25:929–942, DOI: 10.1111/mec.13428.

Vandamme, S. G., Griffiths, A. M., Taylor, S.-A., Di Muri, C., Hankard, E. A., Towne, J. A., Watson, M., and Mariani, S. (2016). Sushi barcoding in the UK: another kettle of fish. *PeerJ*, 4:e1891, DOI: 10.7717/peerj.1891.

Wangensteen, O. S., Palacín, C., Guardiola, M., and Turon, X. (2018). DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. *PeerJ*, 6:e4705, DOI: 10.7717/peerj.4705.

Ward, R. D. (2009). DNA barcode divergence among species and genera of birds and fishes. *Molecular Ecology Resources*, 9:1077–1085, DOI: 10.1111/j.1755-0998.2009.02541.x.

Ward, R. D., Hanner, R., and Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, 74:329–56, DOI: 10.1111/j.1095-8649.2008.02080.x.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., and Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1847–1857, DOI:

For Review Only

10.1098/rstb.2005.1716.

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., and Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources*, 18:1392–1401, DOI: 10.1111/1755-0998.12928.

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., Bunce, M., Wiuf, C., Gilichinsky, D. A., and Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, 300:791–795, DOI: 10.1126/science.1084114.

Winter, D. J. (2017). rentrez: an R package for the NCBI eUtils API. *The R Journal*, 9:520–526.

Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., and Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, 7:40368, DOI: 10.1038/srep40368.

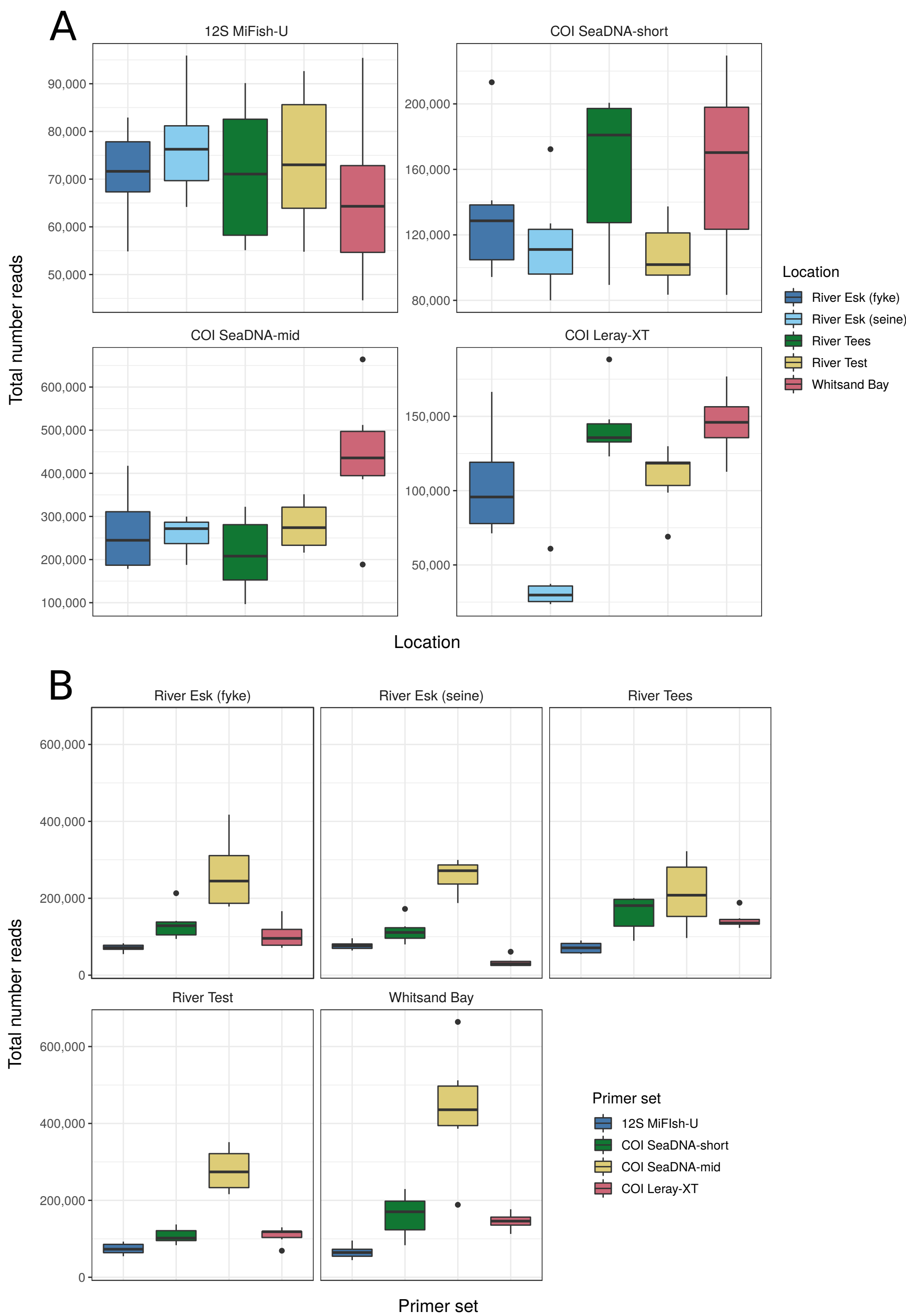
Yang, C., Wang, X., Miller, J. A., De Blécourt, M., Ji, Y., Yang, C., Harrison, R. D., and Yu, D. W. (2014). Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, 46:379–389, DOI: 10.1016/j.ecolind.2014.06.028.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., and Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3:613–623, DOI: 10.1111/j.2041-210X.2012.00198.x.

Supplementary information for:
Non-specific amplification compromises environmental DNA metabarcoding with COI

Rupert A. Collins, Judith Bakker, Owen S. Wangensteen, Ana Z. Soto, Laura Corrigan, David W. Sims, Martin J. Genner, Stefano Mariani

June 17, 2019



Supplementary Figure 1: (A) Read depth (after bioinformatic processing) per location by primer set. (B) Read depth per primer set by location. Per primer-location combination there are three water sample replicates and for each of these, two uniquely tagged PCR replicates ($n = 6$). The horizontal represents the median value, the boxes represent the 25–75th percentiles, the whiskers represent the values less than 1.5 times the interquartile range, and dots represent the outlying data points.

Supplementary Table 1: Unassigned 12S fish reads (n = 73,377) obtained from the *in vitro* analyses of water samples taken from five sites. These were reads that were not assigned to species using our curated reference database of UK fishes, under the criteria of: (i) species-level EPA placement same as the best scoring blast hit, with an aligned match length of $\geq 90\%$ of the modal length of the fragment, and an identity of $\geq 97\%$; or (ii) highest likelihood EPA placement same as the best scoring blast hit, with an EPA probability $\geq 90\%$ and blast identity $\geq 90\%$. The assumed identification is reported after conducting additional phylogenetic analyses, additional BLAST searches, and considering the most common species in the area and the species missing from the reference library. There was also a total of 198,445 reads assigned to *Hippoglossoides platessoides*, which were most probably mis-assigned, and actually belong to the more common pleuronectiform species, such as *Pleuronectes platessa* and *Platichthys flesus*, that were absent from the reference library. Likewise 18,746 reads were assigned to *Syngnathus typhle*, but are more likely to belong to *Syngnathus acus* or *Syngnathus rostellatus*. OTUs (operational taxonomic units) were clustered using *Swarm* from the ASVs (amplicon sequence variants) produced by *dada2*.

OTU	Number ASVs	Total reads	GenBank BLAST match	EPA identification	Assumed species	Comment	Locations
otu11	2	33,143	*Chelidonichthys spinosus*	Actinopterygii	*Chelidonichthys lucerna*	Reference not in library	Test, Tees, Esk-Seine, Whitsand Bay, Esk-Fyke
otu23	3	14,512	*Gaidropsarus argentatus*	Lotidae	*Gaidropsarus vulgaris*	Reference not in library	Esk-Seine, Tees, Whitsand Bay
otu26	1	10,086	*Labrus merula*	*Labrus mixtus*	*Labrus bergylta*	Reference not in library	Tees, Whitsand Bay, Esk-Fyke, Test
otu27	2	9,179	*Ammodytes personatus*	*Ammodytes americanus*	*Ammodytes tobianus*	Reference not in library	Tees, Whitsand Bay, Esk-Fyke, Test
otu35	1	2,677	*Symphodus ocellatus*	*Symphodus melops*	*Symphodus melops*	Misidentified by BLAST due to short reference	Test
otu46	2	1,851	*Parablennius yatabei*	Vertebrata, Gobiidae	*Parablennius gattorugine*	Reference not in library	Tees, Test, Esk-Seine, Esk-Fyke
otu45	1	1,119	*Eleutherochir mccaddeni*	Lophiiformes	*Callionymus lyra*	Reference not in library	Whitsand Bay, Esk-Seine, Test
otu50	2	746	*Psettina iijimae*	Stomiiformes	Pleuronectiformes sp.	Possibly *Arnoglossus*	Test, Whitsand Bay
otu56	1	12	*Clupea harengus*	*Clupea harengus*	*Clupea harengus*	Sequencing/PCR error	Tees
otu57	1	9	*Sprattus sprattus*	*Clupea harengus*	*Clupea* or *Sprattus*	Sequencing/PCR error	Tees
otu58	1	7	*Clupea harengus*	*Clupea harengus*	*Clupea* or *Sprattus*	Sequencing/PCR error	Tees
otu59	1	7	*Lepidopsetta bilineata*	Actinopterygii	Pleuronectiformes sp.	Possible sequencing/PCR error	Esk-Fyke
otu60	1	7	*Chelidonichthys kumu*	*Alepisaurus ferox*	Clupeidae	Sequencing/PCR error	Whitsand Bay
otu61	1	7	*Conger erebennus*	*Nessorhamphus ingolfianus*	*Conger conger*	Reference not in library	Whitsand Bay
otu62	1	7	*Salmo trutta*	*Salmo trutta*	*Salmo trutta*	Sequencing/PCR error	Esk-Fyke, Esk-Seine
otu64	1	3	*Chelidonichthys spinosus*	Perciformes	*Chelidonichthys lucerna*	Reference not in library	Esk-Seine
otu66	1	3	*Enchelyopus cimbrius*	Lotidae	*Gaidropsarus vulgaris*	Reference not in library	Tees
otu67	1	2	*Clupea pallasii*	*Clupea harengus*	*Clupea harengus*	Sequencing/PCR error	Tees

Supplementary Table 2: Metabarcoding and traditional fish survey results for the River Tees site survey. Values correspond to the number of reads identified to species for the molecular markers, and the number of individuals caught on the traditional surveys. Species separated by semicolon are those for which matches were ambiguous. Predominantly freshwater species that are generally not caught on the traditional surveys, are highlighted with an asterisk.

Species	Traditional	12S MiFish-U	COI Leray-XT	COI SeaDNA-mid	COI SeaDNA-short
Ammodytes tobianus	1				
Anguilla anguilla		85			
Aphia minuta		5		2	211
Atherina boyeri		34	3		
*Barbatula barbatula**		42		2	
Chelon labrosus; Liza ramada		43			
Clupea harengus	29	98,907			10
Clupea harengus; Sprattus sprattus		137,123			
*Cottus gobio**		25			
Cyclopterus lumpus		8			
Dicentrarchus labrax		265			
Gadus morhua		41,495			
*Gasterosteus aculeatus**		30		4	
*Gobio gobio**		22		2	
Gobius paganellus		33			3
Hippoglossoides platessoides		13,968			
Limanda limanda	1				
Melanogrammus aeglefinus; Merlangius merlangus		71			
Merlangius merlangus				14	38
Molva molva					31
*Oncorhynchus mykiss**		139		83	159
*Perca fluviatilis**		19			
*Phoxinus phoxinus**		38			
Platichthys flesus	1				
Pleuronectes platessa	12				
Pomatoschistus microps					6
Pomatoschistus minutus	3	24,247			
*Salmo salar**				7	
*Salmo trutta**		13,086		713	158
Sardina pilchardus		307			
Scomber scombrus		101	3		
Sprattus sprattus	233		3	4	
*Squalius cephalus**		8			
Syngnathus acus					47
Syngnathus rostellatus				4	
Syngnathus typhle		16			
Taurulus bubalis		29,189			
Trachurus trachurus		198		10	
Trisopterus luscus					3
Trisopterus minutus		28			
Zeugopterus punctatus			7		

868

Supplementary Table 3: Metabarcoding and traditional fish survey results for the River Esk (fyke) site survey. Values correspond to the number of reads identified to species for the molecular markers, and the number of individuals caught on the traditional surveys. Species separated by semicolon are those for which matches were ambiguous. Predominantly freshwater species that are generally not caught on the traditional surveys, are highlighted with an asterisk.

Species	Traditional	12S MiFish-U	COI Leray-XT	COI SeaDNA-mid	COI SeaDNA-short
Anguilla anguilla	3	364			
Aphia minuta		25			1,105
Atherina boyeri		50			
*Barbatula barbatula**		183		4	
Chelidonichthys lucerna				179	
Chelon labrosus; Liza ramada		18			
Ciliata mustela	11				
Clupea harengus		9,258			65
Clupea harengus; Sprattus sprattus		367			
*Cottus gobio**		26			
Cyclopterus lumpus		8			
Dicentrarchus labrax		165			
Eutrigla gurnardus				27	5
Gadus morhua	16	23,958		53	690
*Gasterosteus aculeatus**		51			
*Gobio gobio**		85			10
Gobius paganellus		97			7
Hippoglossoides platessoides		45,006			
*Lampetra fluviatilis; Lampetra planeri**		1,562			81
Melanogrammus aeglefinus; Merlangius merlangus		13			
Merlangius merlangus					159
Molva molva		16,319	12		4,443
*Oncorhynchus mykiss**		271		32	92
*Perca fluviatilis**		14			
*Phoxinus phoxinus**		171			
Platichthys flesus	12				
Pleuronectes platessa	2				
Pollachius pollachius	2	1,704			
Pollachius virens	11				
Pomatoschistus minutus		10,794			42
*Salmo salar**		13	22	415	
*Salmo trutta**		73,142	172	15,963	1,871
Sardina pilchardus		81			
Scomber scombrus		15,844			
*Squalius cephalus**		10			
Syngnathus acus					340
Taurulus bubalis	19	17			
Trachurus trachurus		38			14
Trisopterus luscus		5			
Trisopterus minutus		12			
Zeugopterus punctatus			16		
Zoarces viviparus	2				

Supplementary Table 4: Metabarcoding and traditional fish survey results for the River Esk (seine) site survey. Values correspond to the number of reads identified to species for the molecular markers, and the number of individuals caught on the traditional surveys. Species separated by semicolon are those for which matches were ambiguous. Predominantly freshwater species that are generally not caught on the traditional surveys, are highlighted with an asterisk.

Species	Traditional	12S MiFish-U	COI Leray-XT	COI SeaDNA-mid	COI SeaDNA-short
Anguilla anguilla		20,004			31
Aphia minuta		4			191
Atherina boyeri		17			
*Barbatula barbatula**		11,688		558	8
Chelon labrosus; Liza ramada		83			
Clupea harengus		225			309
Clupea harengus; Sprattus sprattus		278			
*Cottus gobio**		7			
Dicentrarchus labrax		184			
Gadus morhua		224			
*Gasterosteus aculeatus**		6,633			
*Gobio gobio**		4,349		331	697
Gobius paganellus		21			
Hippoglossoides platessoides		45,936			
*Lampetra fluviatilis; Lampetra planeri**					10
Melanogrammus aeglefinus; Merlangius merlangus		14			
Merlangius merlangus					165
Molva molva		9			
*Oncorhynchus mykiss**		114		32	94
*Phoxinus phoxinus**		43,149	31		
Platichthys flesus	2				
Pleuronectes platessa	1				
Pomatoschistus microps					89
Pomatoschistus minutus		79			
*Salmo salar**		3,424	71	3,770	290
*Salmo trutta**	2	177,271	703	74,004	6,108
Sardina pilchardus		260			
Scomber scombrus		220			
Spondyliosoma cantharus			4		
Sprattus sprattus	1				
Syngnathus acus					50
Syngnathus typhle		97			
Taurulus bubalis		33			
Trachurus trachurus		99		4	
Trisopterus minutus		47			
Zoarces viviparus				53	

Supplementary Table 5: Metabarcoding and traditional fish survey results for the Whitsand Bay site survey. Values correspond to the number of reads identified to species for the molecular markers, and the number of individuals caught on the traditional surveys. Species separated by semicolon are those for which matches were ambiguous. Predominantly freshwater species that are generally not caught on the traditional surveys, are highlighted with an asterisk.

Species	Traditional	12S MiFish-U	COI Leray-XT	COI SeaDNA-mid	COI SeaDNA-short
Ammodytes tobianus	52				
Anguilla anguilla		90			
Aphia minuta		11			1,322
Arnoglossus laterna	13		2		
Atherina boyeri		26			
*Barbatula barbatula**		31			
Buglossidium luteum	16				
Callionymus lyra	21		3		
Centrolabrus exoletus			2		
Chelidonichthys lucerna	1				
Chelon labrosus; Liza ramada		11,013			
Clupea harengus		7,016			
Clupea harengus; Sprattus sprattus		177			
Conger conger				2	
*Cottus gobio**		19			
Ctenolabrus rupestris				44	
Cyclopterus lumpus		4			
Dicentrarchus labrax		30,746			
Echiichthys vipera	7				
Eutrigla gurnardus	26				
Gadus morhua		167			39
*Gasterosteus aculeatus**		8			
*Gobio gobio**		24			
Gobius paganellus		67			
Hippoglossoides platessoides		90,664			
Hyperoplus immaculatus	24				
Hyperoplus lanceolatus	1				
Limanda limanda	8				
Lophius piscatorius	3				
Melanogrammus aeglefinus; Merlangius merlangus		13,398			
Merlangius merlangus	6			16	87
Molva molva		10			
Mullus surmuletus	7				
*Oncorhynchus mykiss**		191		6	32
Pagrus pagrus	10				
Pegusa lascaris	4				
*Perca fluviatilis**		5			
*Phoxinus phoxinus**		49			
Pleuronectes platessa	71				
Pomatoschistus microps					5
Pomatoschistus minutus	192	55			
Raja brachyura	1				
Raja clavata	3				
Raja microocellata	2				
Raja montagui	6				
*Salmo trutta**		427		237	149
Sardina pilchardus		89,488			150
Scomber scombrus		15,546			
Scophthalmus maximus	8				
Scophthalmus rhombus	3				3
Scyliorhinus canicula	1				
Solea solea	3	4		4	
*Squalius cephalus**		6			
Syngnathus acus					287
Syngnathus rostellatus				122	
Syngnathus typhle		18,597			
Taurulus bubalis		19			4
Trachurus trachurus	4	49,801		274	209
Trisopterus luscus		7			
Trisopterus minutus		12,953	7		
Zeus faber				4	5

Supplementary Table 6: Metabarcoding and traditional fish survey results for the River Test site survey. Values correspond to the number of reads identified to species for the molecular markers, and the number of individuals caught on the traditional surveys. Species separated by semicolon are those for which matches were ambiguous. Predominantly freshwater species that are generally not caught on the traditional surveys, are highlighted with an asterisk.

Species	Traditional	12S MiFish-U	COI Leray-XT	COI SeaDNA-mid	COI SeaDNA-short
*Abramis brama**					8
Anguilla anguilla	2	1,704			
Aphia minuta	111	6,493	7	242	220
Atherina boyeri	240	6,154	5	159	
*Barbatula barbatula**		2,470	7	11	
Belone belone			6		
Chelidonichthys lucerna				6	
Chelon labrosus; Liza ramada		17,658			
Ciliata mustela	3		3		
Clupea harengus	24	30,097			288
Clupea harengus; Sprattus sprattus		21,893			
*Cottus gobio**		11,718			
Cyclopterus lumpus		1,609			
*Cyprinus carpio**		597			
Dicentrarchus labrax	4	39,417	8		
Gadus morhua		2,841			14
*Gasterosteus aculeatus**		13,671		306	14
*Gobio gobio**		390	2		
Gobius niger	18				
Gobius paganellus	170	21,225			727
Hippoglossoides platessoides		2,871			
*Lampetra fluviatilis; Lampetra planeri**		215			30
*Leuciscus leuciscus**		2,151	2		
Limanda limanda		1,399			
Liparis liparis	1				
Liza aurata		875			
Liza ramada	1				17
Melanogrammus aeglefinus; Merlangius merlangus		5,364			
Merlangius merlangus	11			56	500
Molva molva		640			46
*Oncorhynchus mykiss**		94,231	139	6,263	7,275
*Perca fluviatilis**		2,196			
*Phoxinus phoxinus**		22,812			
Platichthys flesus	1				
Pleuronectes platessa	1				
Pollachius pollachius		92			
Pomatoschistus microps				14	83
Pomatoschistus minutus	114	12,195			228
Pomatoschistus pictus	3				
*Pseudorasbora parva**					35
Raja clavata	1				
*Rutilus rutilus**		888		16	13
*Salmo salar**			10	46	
*Salmo trutta**		12,049	87	5,362	545
Sardina pilchardus		293			
*Scardinius erythrophthalmus**		1,361			
Scomber scombrus		505	4		
Scyliorhinus canicula	1				
Solea solea	3	784			
Sprattus sprattus	241		7	24	
*Squalius cephalus**		2,646			
Symphodus bailloni	1				
Symphodus melops	2				
Syngnathus rostellatus	1				
Syngnathus typhle		36			
Taurulus bubalis	2	3,171			
*Thymallus thymallus**		1,626			7
Trachurus trachurus		216			
Trisopterus luscus	27	3,046	20	2	52
Trisopterus minutus		461			

Supporting Information: Traditional fish survey protocols

Marchwood Power Station, River Test, Hampshire, Pisces Conservation Ltd.

Outline. Fish entering the station can have four possible fates. They may be returned to sea via the fish return system, they may be washed into the trash basket, captured on the coarse trash screens, or if they are small, they may pass through the station and back to the sea. To estimate the total impingement/entrainment of the station, all possible fates must be quantified. The condition of fish returned to sea is also assessed.

Fish return system monitoring. The fish, invertebrates and weed passing through the fish return system are collected by diverting the flow into a net mounted in the tank built within the system. The water is diverted for a period of 18 hours, usually from 15:15 until 09:15 the following day. A further 6 one-hour samples are then undertaken to complete the full 24-hour monitoring period. The nets used to collect the samples are 1 cm mesh.

From each sample, the debris is sorted and the fish and invertebrates present identified to species. For each fish species present, up to 5 individuals are selected from each size or age class, and their lengths and weights recorded. For fish with no distinct size-classes, individual lengths and weights are recorded for the first 50 individuals. Individual lengths and a combined weight are then recorded for the next 100 individuals of each species. Any further individuals of each species are counted and a combined weight recorded.

For Review Only

Trash basket monitoring. The trash basket is lined with a net, and a 24-hour sample collected and sorted. Fish and invertebrates are measured as described above for the fish return system. The net used to collect the sample is 1 cm mesh.

Trash rake monitoring. A net is placed into the trash skip which receives the rakings from the coarse trash screen. The screens are raked just before the sample is started, and the 24-hr catch is recorded. Mostly the rakings consist of weed and woody debris. The occasional large fish is caught. These data are added to the data on the number of organisms not entering the return system.

Rivers Esk (North Yorkshire) and Tees (County Durham), Environment Agency

Outline. The Water Framework Directive (WFD) monitoring programme consists of two survey approaches: (i) a suite of methods that include fyke nets, seine nets and small (1.5 metre) beam trawl in the shallower, intertidal parts of each water body. These methods are undertaken twice a year during spring and autumn, in combination per site or per water body, depending upon conditions; (ii) a coastal survey vessel to deploy otter trawls in deeper waters. This method is undertaken once a year during autumn where appropriate.

The combination of results from the above methods provides an assessment of the fish communities present throughout the water body.

Seine netting. Two hauls at least within site area, ideally at low slack (high slack may be needed at shallow upstream sites).

Fyke netting. One deployment per sample station. Use two pairs of nets over a full 12 hour tidal cycle.

1.5 metre beam trawl. One tow of 200 metres.

Data. The transitional fish monitoring programme requires the following mandatory data to be collected at each location for each sample: (i) date, time, trawl duration and tide state; (ii) method used; (iii) equipment used, including net dimensions; (iv) sampler names; (v) fish species present; (vi) abundance of each species; (vii) individual length measurements (freshwater and migratory species record fork length, marine species record total length); (ix) water chemistry data (dissolved oxygen, salinity, temperature; and (x) GPS position.

Whitsand Bay, Devon, Marine Biological Association

For the otter trawl methodology, refer to:

McHugh, M., Sims, D. W., Partridge, J. C., and Genner, M. J. (2011). A century later: Long-term change of an inshore temperate marine fish assemblage. *Journal of Sea Research*, 65:187–194, DOI: 10.1016/j.seares.2010.09.006.