Metabarcoding primers for fish biodiversity assessment: a multi-marker comparative study

Bakker J.[1], Wangensteen O. S.[1], Collins R. A.[2], Soto A. Z.[1], Genner M. J.[2], Sims D.?, EA staff?, Henderson P.?, Mariani S.[1*]

[1] School of Environment and Life Sciences, University of Salford
[2] School of Biological Sciences, University of Bristol

4.1 Abstract

Rapid and accurate large-scale species identification and diversity assessment are essential in monitoring and conservation programs for marine species. Environmental DNA (eDNA) metabarcoding is increasingly being applied for the assessment of a diverse range of vertebrate species, including teleosts. However, currently used teleost primers present several shortcomings, which still hinders species-level diversity assessment in taxonomically diverse groups. Here we evaluate the performance of four metabarcoding primers for the assessment of teleost diversity from aqueous environmental DNA samples from UK transitional coastal sites, by specifically focusing on the following variables: (1) taxonomic coverage (2) taxonomic resolution (3) primer specificity, and (4) overlap between DNA-based identification and morphological surveys. By comparing the data generated by an established 12S primer set with that of three alternative COI primer sets, with different lengths, we test if the enhanced reference databases and taxonomic resolution of COI provide better results. This study shows that eDNA metabarcoding is a promising method for teleost biodiversity assessment and that in future applications, a multi-marker approach will most likely be the most appropriate one. However, significant improvements in both reference databases and taxonomic coverage of primers, are essential.

4.2 Introduction

The development of effective management and conservation strategies for marine fishes depends on accurate population status, biodiversity, and species distribution data. However, detecting species occurrences is often even more challenging in the aquatic environment than on land (Webb & Mindel 2015), and obtaining this data by traditional capture and observation-based sampling methods, is often time-consuming, expensive, and invasive in nature. Moreover, it is subject to intrinsic biases related to catchability and requires taxonomic expertise (e.g. Dejean *et al*. 2012; Taberlet *et al*. 2012; Takahara *et al*. 2013; Wheeler 2004).

Currently, environmental DNA (eDNA) metabarcoding of complex samples for aquatic biodiversity assessments is becoming an increasingly popular method for the detection of fish communities in a more resource efficient, comprehensive and non-invasive manner (Bakker et al. 2017; Evans & Lamberti 2017; Yamamoto *et al.* 2017). The detection of multiple species from eDNA is based on the retrieval of genetic material (e.g. skin cells, metabolic waste, blood), naturally released by organisms in their environment (Ficetola *et al*. 2008; Taberlet *et al*. 2012), and the subsequent amplification, sequencing, and taxonomic assignment of this material, through metabarcoding (Ji *et al*. 2013; Taberlet *et al*. 2012b; Thomsen & Willerslev 2015).

It has previously been demonstrated that eDNA metabarcoding has the ability to outperform traditional survey methods for diverse taxa, including fish, both in freshwater (Civade *et al*. 2016; Deiner *et al*. 2016; Hänfling *et al*. 2016; Valentini *et al*. 2016), and in marine ecosystems (Boussarie et al. 2018; Port *et al*. 2016; Thomsen *et al*. 2016; Yamamoto *et al*. 2017). Not only does high throughput sequencing of eDNA provide a non-invasive and easy to standardize means to rapidly identify multiple taxa without the need for taxonomic identification, it also enables the identification of species that might not be detected using conventional survey methods, such as cryptic species and the juvenile stages of many species, of which the distributions during these phases can only be assumed.

Environmental DNA metabarcoding holds considerable promise to revolutionise our understanding of the spatial and temporal patterns of fish diversity in aquatic environments, particularly in improving estimates of species richness, by revealing the composition of entire fish communities in locations of interest (Handley 2015; Evans & Lamberti 2017; Yamamoto *et al*. 2017).

As opposed to the detection of a single species, for biodiversity assessment using eDNA metabarcoding, primers are designed to amplify ‘universal’ barcoding regions that are

evolutionarily conserved across the full spectrum of target species (but still contain enough sequence variability to allow taxonomic resolution of the different species within the targeted community), while having minimal affinity to non-target taxa, such that most of the sequencing depth is dedicated to detecting the species of interest. Hence, the choice of the 'right' primers for the job, is of critical importance.

Several markers, of differing lengths, have been proposed for both the detection of individual fish species, and for the characterization of fish communities from aqueous eDNA (Ficetola *et al*. 2010; Kelly *et al*. 2014; Leray *et al*. 2013; Miya *et al*. 2015; Riaz *et al*. 2011; Stoeckle *et al*. 2017; Thomsen *et al*. 2012; Valentini *et al*. 2016), however, consensus on an optimal generic fish marker has yet to be reached (Shaw *et al*. 2016), and studies comparing different markers on the same samples, particularly pertaining to marine species, are scarce.

Currently, the mitochondrial cytochrome oxidase I (COI) and the 12S marker gene regions are the most widely used in fish DNA barcoding studies (Hardy et al., 2011; Shaw et al. 2016). The (COI) barcode region (Hebert *et al*. 2003), is one of the most commonly sequenced regions for analysis of species diversity among marine animals (Bucklin *et al*. 2011, 2016), due to its high taxonomic resolution resulting from high mutation rate. COI variability is large enough to allow the discrimination of closely related species in most groups, and can even inform on intraspecific variation associated with geographic structure (Bucklin 2011). Additionally, the availability of the large Barcode of Life Data (BOLD) system greatly facilitates taxonomic assignment (Clarke 2017; Hebert *et al*. 2003; Ratnasingham & Hebert 2007). However, since the COI region shows high codon degeneracy ('third codon wobble') throughout its sequence, it lacks highly conserved primer-binding sites (potentially causing taxonomic bias through primer-template mismatches when targeting genetically diverse taxonomic groups), making the design of universal primers, such as is essential for biodiversity metabarcoding studies, very difficult (Clarke et al. 2014; Deagle et al. 2014; Piñol et al. 2015; Sharma & Kobayashi 2014). Instead, it has been argued that the use of mitochondrial non-coding, ribosomal markers, with more conserved primer binding regions, such as the 12S region, may be more appropriate (Clarke et al. 2014; Deagle et al. 2014; Miya et al. 2015; Yang et al. 2014). Nonetheless, due to the higher level of sequence conservation, ribosomal markers often have limited taxonomic resolution (potentially underestimating species diversity within a community), and additionally, have less exhaustive reference databases, compared to COI.

Regardless of marker choice, whether a target species in a complex mixed eDNA sample can be detected, will depend on a number of factors (Fig. 4.1) (1) completeness of the reference database; if a reference sequence for this species is available, (2) taxonomic

coverage of the chosen primer set; if the species-specific barcode sequence can be amplified, (3) taxonomic resolution of the barcoding marker; if the marker sequence is able to characterize the species (and not just the genus or family it belongs to), and (4) the taxonomic specificity of the primer set; if the produced reads of all target species are of sufficient abundance, compared to non-target taxa that may be present in the same sample.

In this study, we evaluate and compare the theoretical and practical performances of four primer sets, targeting both the mitochondrial COI and 12S regions (Table 4.1), for the diversity assessment of teleost fish communities from coastal and transitional waters in the United Kingdom. We used *in silico* PCR to compare taxonomic coverage and resolution of novel COI fragments of two lengths (SeaDNA-Short 55 bp and SeaDNA-Mid 130 bp), specifically designed for the amplification of teleosts. Another, more universal, COI primer set (Leray-XT, 313 bp) targeting eukaryotic diversity (Wangensteen et al. 2017), and one established 12S, teleost specific, primer set (MiFish, 170 bp) (Miya et al. 2015). Subsequently we compared their efficiency in teleost eDNA amplification, *in vitro*, in order to determine primer efficiency in the amplification and characterization of teleost diversity in natural eDNA samples, from four locations in North Sea and English Channel coastal sites in the Britain (Fig. 4.2).. Additionally, an important aspect in testing the potential of eDNA methods to be applied as a tool in monitoring marine species, is to survey the natural environment using both traditional and eDNA methods in concert. Hence, we also compare the fish communities identified from the metabarcoding surveys with those identified from concurrent morphological surveys.

Figure 4.1 Factors determining whether the DNA of a species, present in a complex, mixed eDNA sample, will be detected by eDNA metabarcoding

4.3 Material and methods

4.3.1 Water sampling

A total of 15 sites were sampled from four coastal British locations. These included: the Tees Estuary two sites within the Esk Estuary, The Test Estuary and Whitsand Bay, sampled between October and November of 2016. The former three are estuarine sites, while the latter is a coastal shelf area. A map of sampling locations is shown in Fig. 4.2. Three replicate samples per site, consisting of 2 litres of water, were collected by hand (while wearing disposable gloves) in sterile collection bottles. In order to minimize the amount of plankton (non-target taxa) in the samples and to remove algae and sediment (reducing the severity of filter clogging) prior to passing the samples through the filters, collection bottles were covered with a 250 µm pore size nylon mesh pre-filter, attached to the neck of the bottle by an elastic band. The mesh was discarded after the water sample was collected.

Figure 4.2 Map of the United Kingdom, showing the four sampling locations. Map made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.

4.3.2 Fishing

Fish sampling in the Esk estuary was done by two replicate fyke nets (Esk-fyke) and two replicate beach-seine nets (Esk-seine), during October 2016. At the Tees sampling site, fish surveys were conducted by both two beach-seine nets and two shallow beam trawls, also during October of 2016. During November of the same year, impingement monitoring of fish species was conducted in the Test Estuary and four otter trawls as described in Sims et al. 2004, were conducted in Whitsand Bay. The variety of fishing techniques used in the different sampling locations are part of the currently ongoing monitoring programmes implemented by our local collaborating organization (mention stakehlders).

4.3.3 Sample processing and DNA extraction

After collection, the water samples were individually covered, put into individual sterile plastic bags, and stored in a cooler box filled with ice, while being transported back to the dedicated controlled eDNA laboratory facilities at the University of Salford. Within five hours after collection, each 2-litre sample was filtered through a sterile 0.22 µm Sterivex™-GP filter (polyethersulfone (PES), Merck Millipore, Germany) using a 100 mL polypropylene syringe (Merck Millipore, Germany). After filtration, the filters were stored at -20 ºC, prior to extraction. DNA was extracted from the filters with the Mo-Bio PowerWater DNA Isolation Kit (www.mobio.com), following the manufacturers' protocol, with the addition of an initial 2-hour agitation step, during which the membrane filters were placed in tubes with lysis buffer (C1) and garnet beads from the PowerWater Isolation kit. The tubes were subsequently placed in an orbital shaker at low speed, at 65 ºC, in order to promote the release of DNA from the membrane. Filtration blank controls were processed in parallel. Purified extracts

were assessed for DNA concentration in a Qubit fluorometer (Thermo Fisher Scientific). Disposable nitrile gloves were worn during all stages of sample and filter processing and laboratory equipment and surfaces were cleaned using a 50% bleach solution.

4.3.4 Primer Design

We designed two new metabarcoding primers, SeaDNA-short and SeaDNA-mid, specifically targeting teleosts. These primers were designed using a dataset of complete COI genes for fishes. A total of 2,317 complete fish mitochondrial genomes were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genome/browse/) in November 2016. The COI genes were extracted using a hidden Markov model in the program HMMER v3.1 http://hmmer.org/ (Eddy, 1998), taxonomic annotations were added using rfishbase v2.1.2 (Boettiger et al., 2012), and the sequences were aligned using MAFFT v7.123 (Katoh et al., 2013). Primers were then designed manually in Geneious 8.8.1 (Kearse et al., 2012), 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 new primer sets are both internal to the Folmer fragment (Folmer et al. 1994), which is commonly used for DNA barcoding.

4.3.5 Library preparation and sequencing

Four different metabarcoding markers were amplified, for which the details are presented in table 4.1. To all four primers, 8-base sample specific oligo-tags, differing in at least 3 bases were attached (Guardiola et al., 2015). In order to increase variability of the amplicon sequences, a variable number (2, 3 or 4) of fully degenerate positions (Ns) was added at the beginning of each primer (Wangensteen and Turon, 2017). The PCR conditions for the 12S MiFish primers followed Miya et al. (2015), and for the Leray-XT primers, Wangensteen *et al*. (2017). For PCR amplification with the newly designed SeaDNA-Short and SeaDNA-Mid primers, a two-step protocol was used, attaching the 8-base tagged primers, after an initial amplification, in the second PCR. The mix recipe for the first PCR included AmpliTaq Gold DNA polymerase, with 1 µl of each 5 µM forward and reverse primers, 0.16 µl of bovine serum albumin and 10 ng of purified DNA in a total volume of 20 µl per sample. The recipe for the second PCR was identical, except for the primers now being the forward and reverse 8-base tagged primers. For the first stage PCR, the profile included an initial denaturing step of 95 ºC for 10 minutes, 40 cycles of 94 ºC 30 sec, 47 ºC 45 sec and 72 ºC 30 sec and a final extension step of 72 ºC for 5 minutes. The profile for the second stage PCR was identical, except for the annealing temperature being 50 ºC instead of 47 ºC.

All PCR amplifications were done in duplicate reactions to minimize PCR bias. The quality of all amplifications was assessed by electrophoresis, running the products through a

1.5% agarose gel stained with Gel Red (Cambridge Bioscience) and visualized on a UV light platform. Between the first and second PCR step, amplicons were purified using MinElute PCR purification columns (www.qiagen.com) and diluted ten times prior to being used as a template for the second PCR. After the second PCR, all tagged amplicons were pooled by marker, purified using MinElute columns and each pool was eluted in a total volume of 45 µl, in order to concentrate the amplicons approximately 15 times, for NGS library preparation.

Libraries (one for each marker) were built using the ligation-based NetFlex PCR-free library preparation kit (BIOO Scientific). The libraries were quantified using the NEBNext qPCR quantification kit (New England Biolabs) and pooled in equimolar concentrations along with 1% PhiX (v3, Illumina) serving as a positive sequencing quality control. For each primer, the 15 samples were run using two PCR duplicates, along with one filtration and one PCR blank. The libraries were sequenced on an Illumina MiSeq platform, using V3 chemistry (2x75bp paired-end run) for the SeaDNA-Short library, which was run along two other libraries (from an unrelated project). For the MiFish and SeaDNA-Mid libraries V2 chemistry (2x150bp paired-end run) was used, and these were sequenced in the same run. The Leray-XT library was run using V2 (2x250bp paired-end run) chemistry along with one different library (from an unrelated project). Sequencing depth for all libraries was approximately similar.

A dedicated controlled eDNA lab, with separate rooms designated for the physical separation of eDNA extraction, pre-PCR preparations and post-PCR procedures, was used for all laboratory work. Moreover, to identify potential contamination, every library included one filtration blank (DNA extraction from a Sterivex filter after passing 2L of commercial drinking water), and one PCR blank.

Name	Locus	Primer sequence (5'-3')	Amplicon length (bp)	Full length; amplicon + primers + sample tags + leading N's (bp)	Reference
SeaDNA-Short	COI	GGAGGCTTTGGMAAYTGRYT GGGGGAAGAARYCARAARCT	55	236	This study
SeaDNA-Mid	COI	GGAGGCTTTGGMAAYTGRYT TAGAGGRGGGTARACWGTYCA	130	312	This study
Leray-XT	COI	GGWACWRGWTGRACWITITAYCCYCC TAIACYTCIGGRTGICCRAARAAYCA	313	510	Wangensteen *et al.* (2017)
MiFish	12S	GTCGGTAAAACTCGTGCCAGC CATAGTGGGGTATCTAATCCCAGTTTG	165	345	Miya *et al.* (2015)

Table 4.1 Details of the PCR primers used in this study. Amplicon lengths are given both excluding and including primer sequences, individual sample tags, and leading N's.

4.3.6 In silico evaluation of the primer sets

In order to compare taxonomic coverage and species-level resolution of all four primer sets, they were evaluated *in silico* against 160 teleost species that are found in UK transitional coastal waters. A species list, provided by the Water Framework Directive, United Kingdom Technical Advisory Group, is available in Supplementary Material 4.1. Available full COI and 12S sequences for these species were downloaded from Genbank and the ability of each primer set to amplify the different species (taxonomic coverage) was assessed using ecoPCR and the ecotaxstat function (Ficetola et al. 2010), allowing 3 mismatches per primer, and no mismatches in the two base pairs at the 3' end. The program ecoPCR uses a pattern-matching algorithm to identify sequences within a database that can be amplified with a given primer pair by constraining the relative orientation of and maximum distance between primer-binding sites, as well as the number of mismatches between primer and target sequences (Ficetola et al. 2010). This approach could not be used directly for partial COI sequences (corresponding to the standard Folmer barcoding region) to evaluate the Leray-XT primer set, since most barcode sequences in the Genbank and BOLD databases usually lack the reverse primer binding sequence (as it lays just outside the Folmer fragment) (Wangensteen et al. 2017). Thus, for those teleosts for which no complete COI sequences were available, ecopcr and ecotaxstat were ran against sets of COI barcode sequences with an artificial jgHCO2198-matching sequence attached to the 3' end. This made it possible to test the coverages of the internal forward primers and to compare the performance of the Leray-XT primer set with the other three primer sets, although the taxonomic coverage produced by this primer set could be overestimated using this approach, as it does not account for potential mismatches in the reverse primer region of the partial COI sequences. To check the taxonomic resolution of the markers, the ecotaxspecificity function was used (Ficetola et al. 2010). This function calculates the ratio of the amplified species that can be unambiguously identified by the amplified fragment.

4.3.7 Bioinformatic and statistical analysis

The metabarcoding pipelines were based on the OBITools software suite (Boyer et al., 2016). Quality of the raw reads was assessed using FastQC and the length of raw reads was trimmed to a median Phred quality score >30, after which paired-end reads were assembled using illuminapairedend. The reads with alignment quality scores >40 were kept and the resulting dataset demultiplexed using ngsfilter. A length filter (obigrep) was applied to the aligned reads (45-65 bp for SeaDNA-Short, 120-140 bp for SeaDNA-Mid, 303-323 bp for Leray-XT and 155-185 for MiFish) in order the select only the fragments with the correct target size.

Reads containing ambiguous bases were also removed. The reads were then dereplicated using obiuniq, and chimeric sequences were detected and removed using the uchime-denovo algorithm (Edgar et al., 2011) implemented in vsearch (Rognes et al., 2016). The MOTUs were subsequently delimited using the step by step aggregation clustering algorithm implemented in SWARM 2.0 (Mahe et al. 2015) with a d-value of 13 for Leray-XT (Wangensteen & Turon 2017), d=3 for Miya, d=2 for SeaDNA-short, d=6 for SeaDNA-Mid. The values for this parameter were chosen taking into consideration the natural variability and the length of the different fragments, and the homogeneity of taxonomic assignment of the clustered sequences was checked for selected MOTUs, in order to validate the used d values. The SWARM 2.0 algorithm results in variable thresholds for delimiting MOTUs across different branches of the taxonomic tree, following the natural organization of the clusters in multidimensional sequence space, and includes one final step that breaks chained MOTUs that could lead to artificial over clustering and hence to an underestimation of MOTU richness values. Taxonomic assignment of the representative sequences for each Molecular Operational Taxonomic Unit (MOTU) was performed using the ecotag algorithm (Boyer et al., 2016), which uses a bespoke reference local database and a phylogenetically-based approach for assigning unmatched sequences to the last common ancestor of the most closely related sequences in the reference database. A COI database (Wangensteen et al. 2017) containing 191.295 Eukarya sequences, retrieved from the BOLD database (Ratnasingham & Hebert, 2007) and the EMBL repository (Kulikova et al., 2004), was used for the Leray-XT primers. Bespoke teleost databases for the MiFish 12S (6868 sequences), SeaDNA-Mid COI (67070 sequences), and SeaDNA-Short COI (259696 sequences) primer sets were built with sequences retrieved from Genbank. Reference databases are publicly available from: http://github.com/metabarpark/Reference-databases. After taxonomic assignment, the final refining of the datasets included taxonomic clustering of MOTUs assigned to the same species and minimal abundance filtering; unassigned MOTUs with less than 5 reads for SeaDNA-Mid, less than 3 reads for SeaDNA-Short, 2 reads for Leray-XT and 60 reads for MiFish, were discarded. The minimum read abundance cut-off rates per marker were optimized to reduce the number of false positive MOTUs, considering the number of teleost reads generated by each of the markers. After taxonomic assignment, the resulting datasets from the PCR duplicates from each sample, were combined in order to reduce the effect of PCR amplification bias and sequencing errors (Alberdi 2017; Burgar et al., 2014; Leray & Knowlton, 2015) in order to maximize diversity detection. The pipelines used for data analysis, for both metabarcoding markers, are summarized in Supplementary

material 4.2. All statistical analyses were performed in R v 3.3.0 (https://www.R-project.org/). Package vegan (function speccacum) (Oksanen et al. 2016) was used for obtaining MOTU accumulation curves. Custom R scripts are publicly available from http://github.com/metabarpark.

4.4 Results

4.4.1 *In silico* evaluation of the primer sets

We have tested the theoretical performance of the COI and 12S primer sets, by computationally evaluating them against a list of 160 teleost species found in estuarine and transitional waters in the UK. Based on the four factors determining species detectability as shown in Figure. 4.1, the theoretical performances of each of the four primer sets are displayed in Table 4.2. All three COI primer sets, have a reference sequence available for 94.4% of the 160 species, while for the 12S primer set, a reference sequence is available for only 51.9% of the 160 different teleost species. Additionally, the COI primer sets have relatively high taxonomic coverage and taxonomic resolution, compared to the 12S primers set. However, the performance in taxonomic specificity of the 12S primer set, is superior compared to all three COI primer sets, indicating that 77% of the amplified reads will belong to teleost species, as opposed to non-target taxa. Thus, based on *in silico* analysis, the 12S marker has an incomplete reference database, moderate taxonomic coverage, moderate taxonomic resolution, but a very high specificity for the amplification of teleost DNA. Hence, it is expected to produce a high number of teleost reads, but it will possibly yield a relatively low number of detected species, with moderately accurate species identification compared to the COI markers, which have a much more exhaustive reference database, good taxonomic coverage, very good taxonomic resolution, but very low taxonomic specificity. Accordingly, the COI primer sets are expected to produce a very low number of teleost reads (compared to reads from non-target taxa), but from a more diverse assemblage of species. Moreover, they are expected to show highly accurate species identification ability compared to the 12S primer set.

Primers	**Reference database completeness**	**Taxonomic coverage**	**Taxonomic resolution**	**Specificity**
12 S Miya MiFish 170 bp	83 spp.: **51.9%**	69 spp.: **83.1%**	62 spp.: **89.8%**	High: **77%** teleost reads
COI Leray-XT 313 bp	151 spp.: **94.4%**	144 spp.: **95.5%**	142 spp.: **98.6%**	Very low: **0.05%** teleost reads
COI SeaDNA-	151 spp.: **94.4%**	128 spp.:	108 spp.: **84.4%**	Low: **1.2%**

Mid 130 bp		**84.8%**		teleost reads
COI SeaDNA-Short 55 bp	151 spp.: **94.4%**	147 spp.: **97.3%**	121 spp.: **82.3 %**	Low: **1.2%** teleost reads

Table 4.2 Result from the *in silico* analysis, comparing the theoretical performance of the four primer sets in the detection of 160 teleost species from UK transitional waters.

4.4.2 Teleost detection

A total of 15 samples, from 4 locations (5 sites), in 4 different amplicon libraries, were sequenced using three separate Illumina MiSeq runs. The read statistics for all four primer sets after sample assignment, quality and sequence-length filtering, and combining the results from the PCR duplicates from each sample, are displayed in Table 4.3, No teleost reads were detected in the negative controls. The final data set for Leray-XT comprised a total number of 3.062.194 reads. While only 0.051% of the reads were assigned to teleosts, the number of teleost MOTUs detected was almost equal to the number of teleost MOTUs detected by the MiFish primers, for which the percentage of teleost reads was 77.7%. And while the percentage of teleost reads for both SeaDNA primers was slightly higher compared to Leray-XT, the number of species detected was lower (17 and 18 teleost MOTUs). Overall, 60 teleost MOTUs were detected by eDNA metabarcoding, of which 36 species were exclusively detected by eDNA. In total, 41 different species were detected by the morphological

surveys, of which 18 species were not detected by eDNA. A total of 27 species was detected by both methods (Figure 4.3).

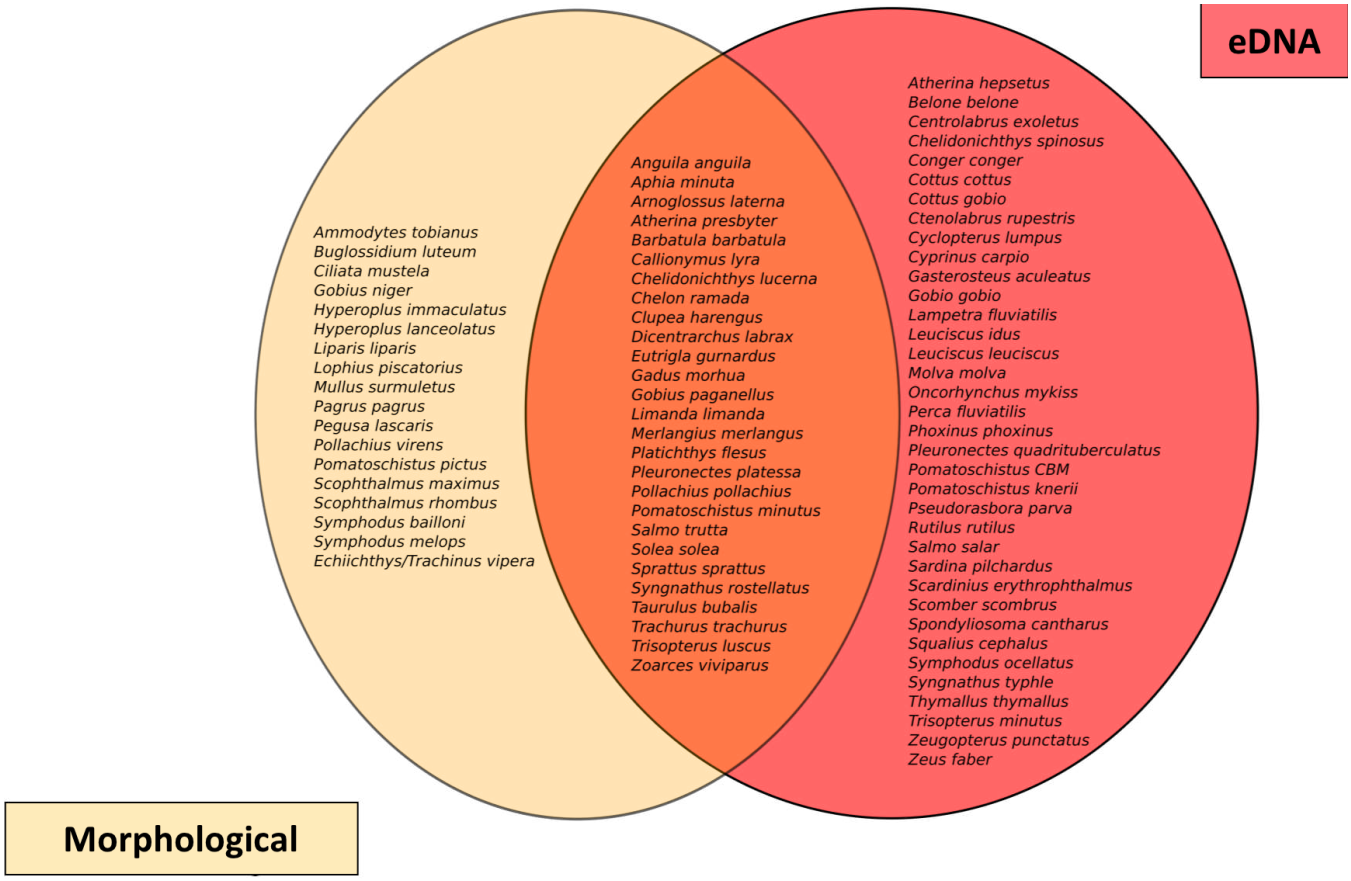

Figure 4.3 Comparison of the different species detected by morphological surveys and by eDNA metabarcoding. The overlap represents the 27 species detected by both methods.

Primer set	Total reads	Teleost reads	Teleost reads (%)	Teleost MOTUs	Metazoa	Metazoa (%)	Eukarya	Unidentified Eukarya	Eukarya (%)	Bacteria	Bacteria (%)
COI Leray-XT	3062194	1554	0.051	32	15968	0.521	1242435	407511	40.573	1396280	45.597
COI-SeaDNA-Mid	8201135	97589	1.19	17	276619	3.373	5313981	1308357	64.796	1302178	15.878
COI-SeaDNA-Short	4447441	54392	1.223	18	103442	2.326	1275366	553237	28.676	2515396	56.558
12S Miya MiFish	2312505	1795728	77.653	33	2295310	99.256	17195	0	0.744	0	0

Table 4.3. Sequencing read statistics for the four primer sets

As shown in Figure 4.4, there is only a relatively small overlap in species detection between the four different primer sets. Eight species are detected by all four primers, six marine species, *Aphia minuta* (transparent goby), *Gasterosteus aculeatus* (three-spined stickleback), *Oncorhynchus mykiss* (rainbow trout), *Salmo salar* (Atlantic salmon), *Salmo trutta* (brown trout) and *Trisopterus luscus* (whiting-pout), and two fresh water species *Barbatula barbatula* (stone loach) and *Gobio gobio* (gudgeon). While a total number of 60 species was detected, half (30 species) was detected by only one of the four primer sets. Leray-XT uniquely amplified 10 species, while 13 species were only amplified by MiFish, 4 by SeaDNA-Mid and 3 by SeaDNA-Short. If one would have to choose a combination of two primer sets, Leray-XT and MiFish combined, detect a total of 51 different teleost MOTUs. However, SeaDNA-Short is the only primer set that has detected *Pseudorasbora parva* (indicated by a blue star in Fig. 4.4), the topmouth gudgeon (or stone moroko), which is native to Asia, but has been introduced and is now considered an invasive species in Europe (Britton et al. 2010; Pinder et al. 2005). Moreover, 5 out of the 33 MOTU's that were detected by the MiFish primer set, were taxonomically assigned, with 100% accuracy, to species that do not occur in UK waters. Both *Chelidonichthys spinosus* (spiny red gurnard) and *Pleuronectes quadrituberculatus* (Alaska plaice) are Pacific species (red stars in Fig. 4.4) while their Atlantic counterparts *Chelidonichthys cuculus* (red gurnard) and *Pleuronectes platessa* (European plaice), indeed do occur in the North Sea. Moreover, three MOTU's delimited by the MiFish primer set, have been taxonomically assigned to Mediterranean species, *Atherina hepsetus* (Mediterranean sand smelt), *Pomatoschistus knerii* (Kner's goby) and *Symphodus ocellatus* (ocellated wrasse) (green stars in Fig. 4.4). Most likely, these sequences are derived instead from similar species occurring in the North Sea; *Atherina presbyter* (sand smelt), *Pomatoschistus minutus* (sand goby) and *Symphodus melops* (corkwing wrasse). The incorrect taxonomic assignment to sister species is most likely caused by the fact that many sequences deposited in the 12S teleost database are derived from Pacific counterparts while the sequences of the European species are absent from the reference database. Details for the number of reads per species, per location, for each primer set, are available in Supplementary material 4.3.

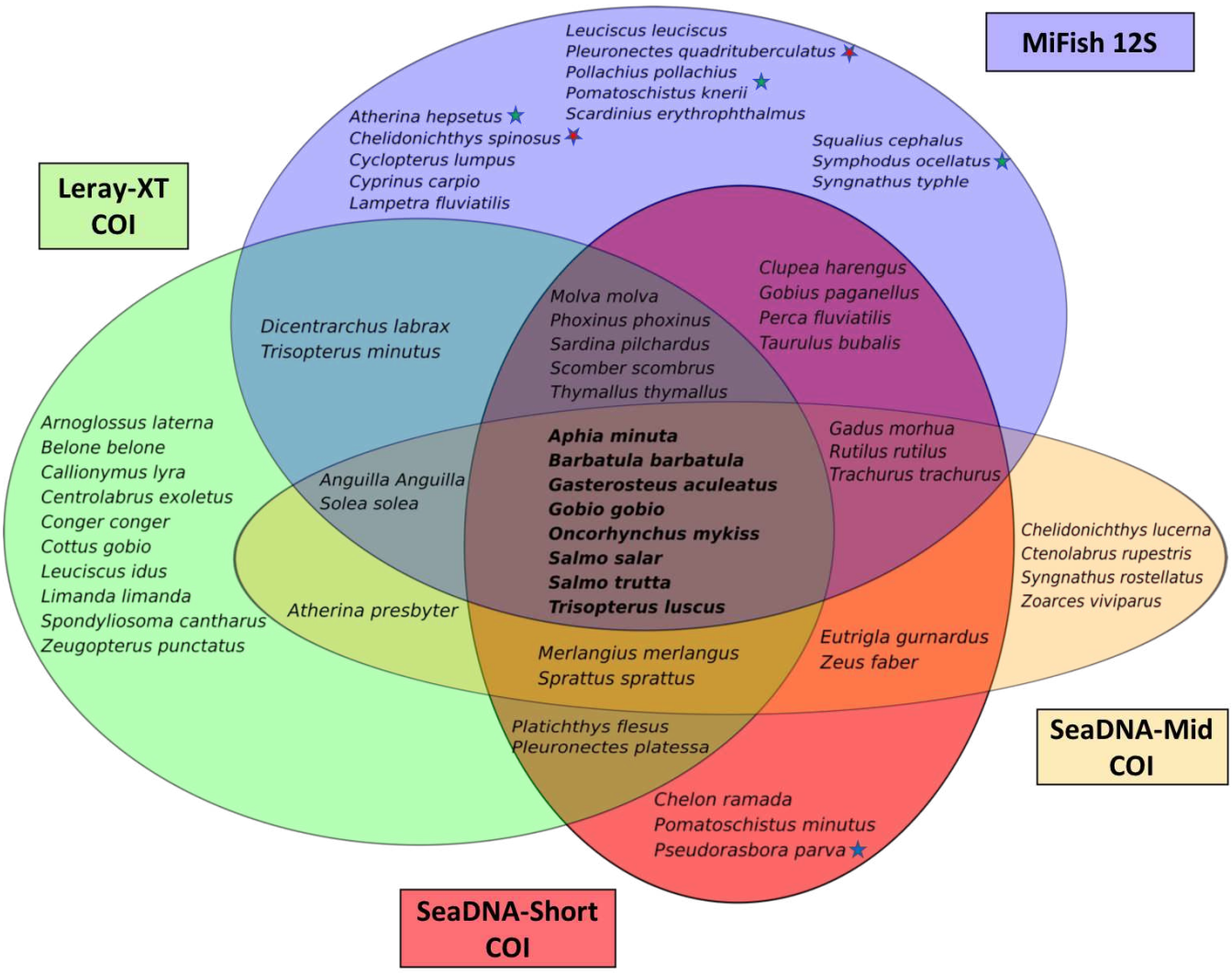

Figure 4.4 Venn diagram showing the distinctive and overlapping species detection between all four primer sets. Red stars indicating MOTUs taxonomically assigned to Pacific species and green stars indicating MOTUs assigned to Mediterranean species. The blue star indicates the invasive topmouth gudgeon (*Pseudorasbora parva*).

Figure 4.5 shows all the species detected by both traditional morphological sampling, and the eDNA surveys, for the four different primers, indicated per sampling site. The site with the highest density of species diversity, detected by both morphological and eDNA surveys, is the Test estuary. Here 22 species were identified by morphological surveys, and 44 species by eDNA metabarcoding. At every site, more teleost species were detected by the combined four primer sets than with the morphological surveys (Table 4.4.) However, none of the four primer sets individually outperforms the morphological sampling at all sites (only at 1,2 or 3 out of the 5 sites). Furthermore, the detection patterns overall are very irregular, and the

matches between the morphological surveys and metabarcoding results are not very robust, indicating that a larger number of replicates would be needed (see also Figure 4.3). Ten freshwater species were detected by the eDNA but not the morphological surveys; *Barbatula barbatula* (stone loach), *Cottus gobio* (European bullhead), *Cyprinus carpio* (European carp), *Gobio gobio* (gudgeon), *Leuciscus idus* (ide), *Perca fluviatilis* (European perch), *Phoxinus phoxinus* (Eurasian minnow), *Pseudorasbora parva* (topmouth gudgeon), *Scardinius erythrophthalmus* (common rudd), and *Squalius cephalus* (Chub). It is most likely that the DNA from these species was transported down to the sampling estuaries from the upstream rivers (Deiner et al. 2016).

Figure 4.5 Presence/absence diagram showing all the species (in alphabetical order) detected in the five sampling sites. For every site, species detection is shown for traditional sampling ('Trad'; total number of individuals detected), and for the four different primers, for which the total number of sequences is shown. Indicated by the red arrows are the ten fresh water species

Site	Number of species					
	Morphological sampling	eDNA sampling primers combined	Leray -XT	SeaDNA -Mid	SeaDNA -Short	Miya-MiFish
Esk –Fyke	**8**	**23**	**9**	**4**	**12**	**8**
Esk –Seine	**4**	**22**	**7**	**5**	**11**	**6**
Tees	**7**	**14**	**5**	**4**	**5**	**4**
Test	**22**	**44**	**23**	**9**	**16**	**25**
Whitsand	**21**	**27**	**10**	**7**	**9**	**9**

Table 4.4 Numbers of teleost species detected at each of the sampling sites, by both morphological and eDNA surveys. Number of species for eDNA surveys are given for the four primers combined and for each primer set individually. Coloured numbers indicate eDNA species detection performance compared to morphological surveys. Green indicates same or higher number of species detected and red indicates less species detected compared to morphological surveys.

4.4.3 Teleost diversity and read abundance patterns

Violin plots of MOTU richness for each primer set (Fig. 4.6), show how the different sample values are distributed, by comparing the variable sample size distribution across the five different sites. The distribution of density (number of MOTUs per sample) is represented by the width of the plots. Even though there are pronounced differences in species detection between the primers (Fig. 4.4), and additionally, in the numbers of different MOTUs detected (Table 4.3), the patterns of MOTU richness between the four primers is strikingly similar. The first thing that becomes apparent is that when using any of the four primer sets, the Test estuary is the location with the highest MOTU (species) diversity. With a maximum of >20 MOTUS in one sample with both MiFish and SeaDNA-Short primers, and a minimum of 9 MOTUs per sample (for SeaDNA-Mid). While the Tees estuary shows the lowest MOTU richness per sample (between 2 and 7 MOTUs), with any of the four primer sets. MOTU richness is most diverse among the different samples of the Esk-Fyke site, as indicated by the size and shape of its plots. The sites can be ordered from highest to lowest MOTU richness; (1) Test, (2) Esk-Fyke & Whitsand, (3) Esk-Seine, and (4) Tees. No samples contain less than 3 teleost MOTUs.

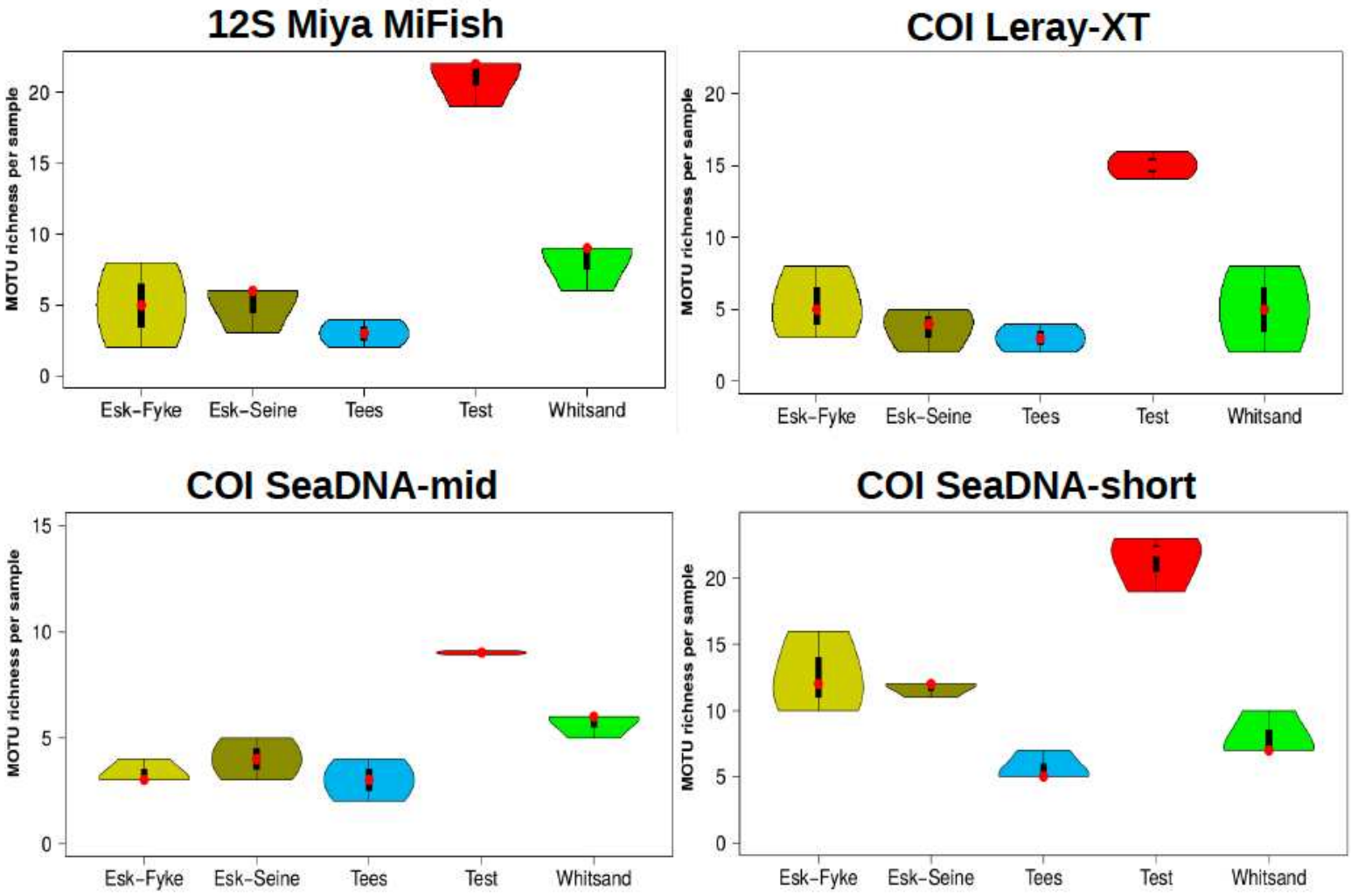

Figure 4.6 Violin plots showing detected teleost diversity (MOTU richness), per sample in the different sites by the four primers sets. The shapes indicate the density distribution of the samples, extending from the minimum to the maximum observed values. The median values are indicated by the red dots. The thick black bars are the interquartile ranges. The thin black extending lines represent the 95% confidence intervals such that the values in the wider parts of the plots are more probable than those in the narrower parts.

4.4.4 Patterns of β-diversity

Patterns of community structure between the five sampling sites are presented for all four primer sets in Figure 4.7, visualizing the dissimilarities among the different sites. Even though the species detection patterns between the four primers, within each site, are irregular (Fig. 4.5), it is apparent that samples cluster according to sampling site, with all four primers. However, the pattern of dissimilarity is amplified (greater spread of the five sampling sites) by both the SeaDNA-Mid and SeaDNA-Short primers. Which could be explained by the fact

that a smaller amount of species has been detected by these primers, magnifying the differences in species composition between the sites, resulting in an enhanced spread in the MDS. The samples from Whitsand are most dissimilar from the other sites (with all primers), which is in accordance with the fact that this is a continental shelf sampling location, while the others are estuarine.

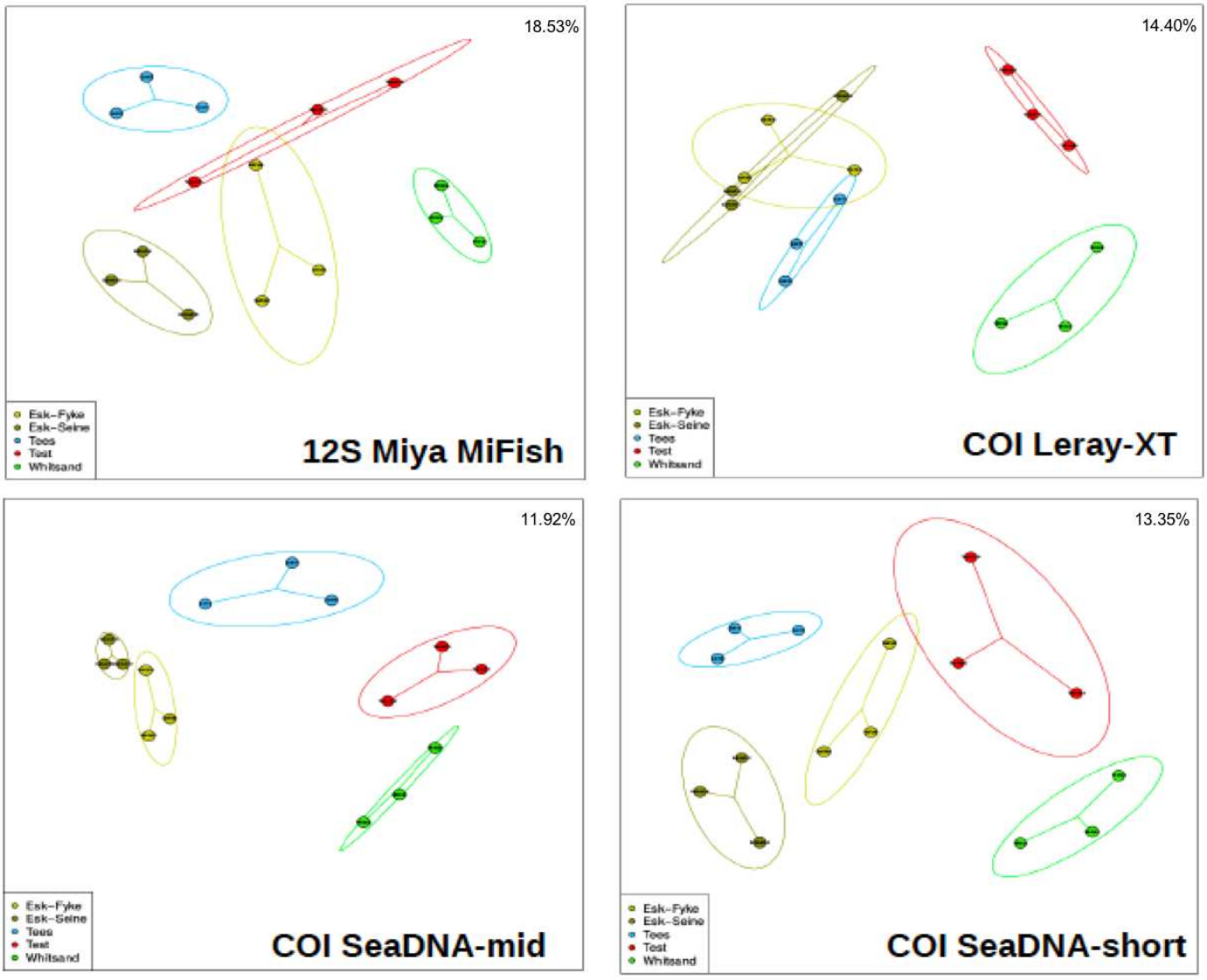

Figure 4.7 nMDS ordinations of the three replicate samples per location, as produced by each of the four primer sets. Results are based on the Jaccard index, based on the presence-absence data of MOTUs. Numbers in the upper right corners indicate stress of the final configurations

Species accumulation curves are plotted for each location, in a separate graph for each of the primer sets (Fig. 4.8). The curves show teleost diversity (MOTU richness) as a function of the

number of samples taken in each of the five sites. Error bars indicate standard errors after 100 permutations. It shows that none of the sampled sites (three replicates for each site), with any of the primers, tend to reach plateau in MOTU richness. The overall non-saturation of the species accumulation curves, indicates that increased sampling effort is essential for capturing total teleost diversity in each of the sites, regardless of the primer set used.

Figure 4.8 Species accumulation curves showing teleost diversity (MOTU richness) as a function of the number of samples taken at each location, separately for the four different primer sets. Error bars indicate standard errors after 100 permutations.

4.5 Discussion

The lack of suitable universal primers currently hinders the application of eDNA metabarcoding techniques for the characterization of teleost communities in estuarine and marine waters. Therefore, the goal of this study was to evaluate and compare the theoretical and practical performances of four primer sets, in amplifying a wide array of teleost diversity present in UK estuarine and marine ecosystems. Both *in silico*, and *in vitro* analysis allowed us to evaluate and compare the efficiency of two established (COI and 12S), and two novel (both COI) eDNA metabarcoding primer sets, in describing teleost diversity in five UK sampling sites, and additionally, to compare these results with concurrent morphological sampling.

As was estimated by *in silico* analysis (Table 4.2), the 12S MiFish primers are superior in primer specificity (77.7% of all reads was assigned to teleosts), compared to the three COI primers (Table 4.3). But strikingly, while with the Leray-XT primers (which are not designed specifically for the amplification of teleosts, but rather for eukaryots in general), teleost read abundance only accounted for 0.05% of all reads (very low primer specificity), the number of species detected (MOTU richness) was nearly identical to that of the MiFish primers (32 and 33 respectively). However, there is an overlap of only 17 species between these two primers (Fig. 4.4), indicating that both are underperforming in species detection. Thus, while both primers show potential for applied eDNA metabarcoding in teleost diversity assessment in UK waters, significant improvements are required. Albeit, it is a different area for each of the primers, where the improvements will have to be implemented. For the 12S MiFish primers, that already have very high teleost specificity, it is mostly the reference database that is currently insufficient, particularly for the identification of Atlantic teleost species, as was indicated by the *in silico* analysis (only 51.9% of the 160 species was present in the reference database, Table 4.2), and highlighted by the taxonomic assignment of eDNA reads from UK sampling sites, to species that only occur in the Pacific Ocean or in the Mediterranean Sea (Fig. 4.4). While the COI Leray-XT markers allow for a high level of accurate species identification (taxonomic resolution of 98.6%, Table 4.2), their selectivity for teleost DNA is very low. Improving primer selectivity, while at the same time keeping broad taxonomic coverage for teleosts is very challenging, and it is this same challenge that lays at the base of the design of metabarcoding primers for any taxonomic group. Alternatively, or simultaneously, a higher sequencing depth may improve teleost DNA amplification by the Leray-XT primers.

Even though the two other COI primers, SeaDNA-Mid and SeaDNA short, were designed specifically for this study, with the aim of improving selectivity for teleost sequences, they clearly performed worse than the MiFish and Leray-XT primer sets. Both of these primers were expected to perform at least equally well as the Leray-XT primers, based on the premises of producing a significantly shorter amplicon (130 and 55 bp respectively, compared to 313 bp for the Leray-XT primers), increasing the chances of amplifying the full length of the marker, from the degraded small fragments of DNA present in the water and enhancing the selectivity for teleost sequences. However, in order to improve taxonomic coverage of teleost sequences by keeping low levels of primer bias, both SeaDNA primers contain considerable levels of ambiguity (Table 4.1) in order to improve their universality (Wangensteen et al. 2017), which in combination with the shorter amplicon length, was expected to result in better teleost detection performance. It has been suggested that 50-55bp may be the optimal amplicon length to acquire maximum resolution while concurrently optimizing (degraded) eDNA amplification by using a minimum amplicon length (Hajibabaei et al. 2007; Wangensteen et al. 2017). However, targeting short fragments may increase the probability of amplifying more of the eDNA present in a sample, it will also inevitably decrease taxonomic resolution of the marker, due to increased synonymies between species within these short sequences, consequently rendering these primers less discriminant at the species level. Consequently, a large part of the teleost reads from both SeaDNA primers, could only be taxonomically assigned to genus or family level. Additionally, while for both SeaDNA primers, the forward primer is identical, the reverse primer of SeaDNA-Mid is prone to an increased level of primer bias, as is shown by a smaller taxonomic coverage (table 4.2), leading to a decreased level of taxonomic resolution compared to SeaDNA-Short. Accordingly, 17 and 18 teleost MOTUs were identified by SeaDNA-Mid and SeaDNA-Short respectively, but in each sampling site, SeaDNA-Short has detected more species compared to SeaDNA-Mid (Table 4.4). Thus, our efforts to achieve an increase in the number of teleost reads while keeping high coverage levels and low primer bias, using COI-based markers, have been mostly unsuccessful.

As fish monitoring in the sampled range of transitional water bodies cannot be consistently applied, the different sampling approaches at each site, influence the catch data. However, the purpose of this study was firstly, to compare the performance of the different primers sets among each other and secondly, to investigate how eDNA metabarcoding compares to the standard traditional fish sampling, implemented in the different locations, irrespective of sampling method. In spite of the fact that with a combination of the four

primer sets, eDNA species detection was superior at every sampling site, and overall, 19 more species were detected with eDNA metabarcoding compared to the morphological surveys, the detection patterns between the two methods are quite dissimilar and 18 species present in the morphological samples have not been detected by eDNA (Figure 4.3). We expect that this issue will largely be resolved with increased eDNA sample sizes.

Our study reiterates that eDNA metabarcoding can be used to assess teleost diversity from natural water samples, however, even when a single taxonomic group is targeted, the use of different primer sets may still produce biased results in species detection in the same sample, indicating that metabarcoding approaches for the detection of species-level diversity, using a single marker, still face significant challenges. Thus, in order for eDNA metabarcoding to develop into a tool that can be used to assist fisheries professionals in the assessment and monitoring of fish communities, it is expected that in future applications for the assessment of community diversity for such taxonomically diverse groups as teleost, a multi-marker combination of primer sets may be most suitable to reduce taxonomic biases and increase taxonomic coverage and species detection probability, and hence provide a more accurate picture of species diversity.

4.6 Acknowledgements

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