Evaluating bioinformatics pipelines for population-level inference using environmental DNA

Bastien Macé, Régis Hocdé, Virginie Marques, Pierre-Edouard Guerin, Alice Valentini, V. Arnal, Loïc Pellissier, S. Manel

Evaluating bioinformatics pipelines for population-level inference using environmental DNA

Author(s):
Macé, Bastien; Hocdé, Régis; Marques, Virginie; Guerin, Pierre-Edouard; Valentini, Alice; Arnal, Véronique; Pellissier, Loïc; Manel, Stéphanie

Publication date:
2022-05

Permanent link:
https://doi.org/10.3929/ethz-b-000550930

Originally published in:
Environmental DNA 4(3), https://doi.org/10.1002/edn3.269

Received: 25 March 2021 | Revised: 5 November 2021 | Accepted: 16 November 2021

DOI: 10.1002/edn3.269

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

WILEY

ORIGINAL ARTICLE

Evaluating bioinformatics pipelines for population-level inference using environmental DNA

Bastien Macé[1] | Régis Hocdé[2] | Virginie Marques[1,2] | Pierre-Edouard Guerin[1] | Alice Valentini[3] | Véronique Arnal[1] | Loïc Pellissier[4,5] | Stéphanie Manel[1]

[1]CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France

[2]MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[3]SPYGEN, Le Bourget -du-Lac, France

[4]Landscape Ecology, Department of Environmental Systems Science, Institute of Terrestrial Ecosystems, ETH Zürich, Zürich, Switzerland

[5]Unit of Land Change Science, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

Correspondence
Stéphanie Manel, UMR 5175, CEFE, EPHE-PSL, Campus CNRS 1919 route de Mende, 34293 Montpellier, Cedex 5, France.
Email: stephanie.manel@ephe.psl.eu

Funding information
Université de Montpellier, Grant/Award Number: KIM Sea & Coast 2020; LabEx CeMEB, Grant/Award Number: ANR-10-LABX-04-01

Abstract

Environmental DNA is mainly not only used at the interspecific level, to quantify species diversity in ecosystems, but can also be used to quantify intraspecific genetic variability, thus avoiding the need to sample individual tissue. However, errors in the amplification and sequencing of eDNA samples can blur this intraspecific signal and strongly over-estimate genetic diversity. Existing bioinformatics pipelines therefore need to be tested to evaluate whether reliable levels of intraspecific genetic variability can be derived from eDNA samples. Here, we compare the ability of twelve metabarcoding pipelines to detect intraspecific genetic variability combining five programs. All pipelines have common pre-processing steps, a processing data step using programs among *obiclean*; DADA2; SWARM; and LULU. An additional chimera removal step is also investigated based on two programs (VSEARCH or DADA2). The case study was the natural intraspecific variation within *Mullus surmuletus* in experimental settings. We developed specific primers for this species, located on the mitochondrial D-loop fragment (barcode MS-DL06). Thirty-nine individuals were collected from the Mediterranean Sea, placed into four aquariums, and their DNA was sequenced on this marker to build an intraspecific reference database. After filtering the aquarium water, DNA was extracted, amplified, and sequenced using the primer pair developed. We then quantified the number of true haplotypes returned by each pipeline and its capacity to eliminate most of the erroneous sequences. We show that the program DADA2 with a two-parent chimeric sequence removal step is the best tool to estimate intraspecific diversity from eDNA. Furthermore, our approach was also able to detect true *M. surmuletus* haplotypes in two eDNA samples collected in the Mediterranean Sea. We conclude that the combination of an appropriate intrapopulation barcode and a denoising pipeline like DADA2 with a chimeric sequence removal step is promising to make population-level inference using environmental DNA possible.

KEYWORDS
bioinformatics, environmental DNA, fish, genetic diversity, marine ecology

1 | INTRODUCTION

Environmental DNA (eDNA) metabarcoding is a promising tool for improving biodiversity assessment down to species level (Cordier et al., 2021; Ficetola et al., 2008). Organisms leave genetic material including feces, gametes, and epidermal cells in the environment, which can be detected in filtered water samples using high-throughput sequencing and bioinformatics (Taberlet et al., 2018). Environmental DNA metabarcoding studies have been suggested as a noninvasive and effective environmental monitoring tool for biodiversity assessment, with higher detection capabilities and better cost-effectiveness than traditional methods (Thomsen & Willerslev, 2015) even in marine ecosystems (Polanco Fernández et al., 2021). Recent aquatic applications have demonstrated that eDNA can be used to quantify the richness and compositional variation of seasonal species (Milhau et al., 2021; Sigsgaard et al., 2017; Stoeckle et al., 2017). By developing specific primers to define barcodes, eDNA can also be used to detect the presence of species, such as the threatened Maugean skate (*Zaeraja maugeana*) in Tasmania (Weltz et al., 2017) or invasive species, like the Atlantic rangia (*Rangia cuneata*) in the Baltic Sea (Ardura et al., 2015).

One emerging application for eDNA methods is the study of intraspecific genetic variation. Using eDNA for population-level inference opens up alternatives to conventional tissue-based methods, with the advantage of eDNA sampling being noninvasive and generally more cost-efficient (Adams et al., 2019; Sigsgaard et al., 2020). One promising perspective for the application of eDNA at the population level is the spatial context of landscape genetics (Manel & Holderegger, 2013), where a large number of sites and populations need to be sampled to establish patterns of connectivity and adaptive genetic variation. Currently, only preliminary studies have been published in this field, focusing on just a few sites, but they have shown that eDNA can be used to estimate genetic diversity within a population or to detect genetic differentiation among populations (e.g., Baker et al., 2018; Elbrecht et al., 2018; Parsons et al., 2018; Sigsgaard et al., 2016; Stat et al., 2017; Székely et al., 2021; Turon et al., 2020; Uchii et al., 2016, 2017). Most recent studies have used PCR amplification and sequencing of mitochondrial DNA sequences, but the studies of Minamoto et al. (2017) and more recently of Andres et al. (2021) are based on nuclear markers.

One of the main obstacles to using eDNA for intraspecific diversity assessments is the relatively high error rate during PCR and sequencing (Furlan et al., 2020). PCRs can generate substitutions and chimeric sequences (or PCR-mediated recombinants), representing a merged DNA fragment from closely related templates (Holcomb et al., 2014; Potapov & Ong, 2017; Smyth et al., 2010). A chimera removal step can be added to bioinformatics pipelines using dedicated programs (Furlan et al., 2020). Erroneous sequences can also be generated by sequencing errors. All these errors can be reduced using abundance-based filters or cleaning tools implemented in bioinformatics pipelines (Sigsgaard et al., 2020). However, no comparisons of the performance of such programs have been made for eDNA intrapopulation analysis and very few studies have been conducted using datasets from *in situ* samples (e.g., Elbrecht et al., 2018; Parsons et al., 2018). Although different programs have been applied in previous diversity studies to eliminate as many erroneous sequences as possible, no consensus set of bioinformatics tools have emerged for studying intraspecific diversity using eDNA.

The general objective of this study was to evaluate and compare the ability of existing bioinformatics programs to reveal haplotypes for intraspecific analyses from the case study of one species, the striped red mullet (*Mullus surmuletus*). We first designed an experimental setting to capture the intraspecific variation within the striped red mullet. Primers located on the mitochondrial D-loop fragment (barcode MS-DL06) were developed for this species. Thirty-nine individuals were collected in the Mediterranean Sea and placed in four aquariums, and their DNA was sequenced on this marker to build an intraspecific reference database. After filtering the water from the aquariums, DNA was extracted, amplified, and sequenced using the primer pair developed. Secondly, we analyzed *in situ* marine eDNA samples with the best-performing pipeline. These comprised two field samples collected in the Mediterranean Sea, known to contain *M. surmuletus* from a previous independent study (Boulanger et al., 2021). Specifically, we aimed to (i) develop a barcode adapted to the intraspecific analysis of *Mullus surmuletus*, (ii) compare twelve pipelines based on five bioinformatics programs to analyze intraspecific barcode data, and (iii) make recommendations regarding best practices for analyzing intraspecific eDNA.

2 | MATERIAL AND METHODS

2.1 | Species biology and experimental conditions

Mullus surmuletus is a teleost fish from the *Mullidae* family. This benthic carnivorous fish is widely distributed on the continental shelf along the coast of Western Europe, up to the North Sea, and Western Africa, up to Dakar, including the Mediterranean Sea and the Black Sea (Whitehead et al., 1986). This species has been fished since Antiquity (Mahe et al., 2005) and currently has a high commercial value (Alomar et al., 2017). Genetic barriers have been detected between the Atlantic Ocean and the Mediterranean Sea, and within the Mediterranean Sea (Matić-Skoko et al., 2018). Developing eDNA's potential to quantify intraspecific genetic *M. surmuletus* variability can be useful in the future.

Thirty-nine juvenile *M. surmuletus* were caught while closed-circuit rebreather diving, at a depth of between 19 and 40 meters, in the Mediterranean Sea (coordinates: 43.6850262 and 7.30168527) at Villefranche-sur-Mer (France) on July 14, 2019 (Figure 1). The use of the closed-circuit rebreather allows us to have the time to work at 40 m and then 19 m to capture the mullets in good conditions. Moreover, the silence due to the lack of bubbles of the divers allows to better approach and less stress the mullets which are very sensitive animals, and to capture them more easily. We used a 6 × 1.20 m net with fine-meshed to prevent fish from becoming entangled in it, floating on the high line and ballasted on the low line and deployed

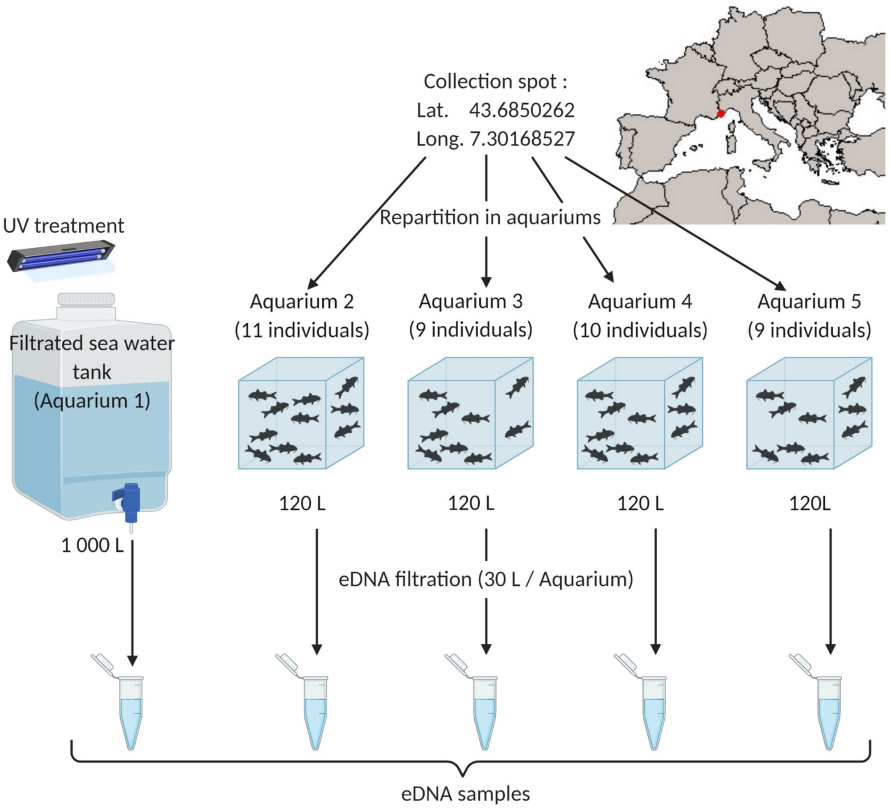

FIGURE 1 Localization of the collection point in the Mediterranean Sea on July 14, 2019 and summary of the sampling steps. The map was produced using the mapproj R package. The figure was created using Biorender.com (2020)

on the sea bed to act as a barrier. The fish were then caught with a small hand net. They were placed in a live fish bait bucket that was brought to the surface carefully to prevent swim bladder trauma and to avoid thermal shock due to the variations in the seawater temperature. The seawater temperature measured using a conductivity–temperature–depth (CTD) profiler was 21.4°C at 19 m depth and 15.4°C at 40 m depth (Deter et al., 2020). The fish caught were transferred under controlled low-temperature conditions, and all necessary precautions were taken to reduce animal stress. Authorization to catch fish in the Mediterranean Sea for the purposes of this study was given by the French Interregional Direction of the Mediterranean Sea (Order No. 387 of June 24, 2019).

Four 250-liter aquariums were bleached clean one day prior to be used (filled with seawater; fish transfer) in Montpellier (France). Seawater collected by the French Research Institute for Exploitation of the Sea at Palavas-les-Flots (France) was first stored in a 1000 L tank for two weeks, under UV treatment to avoid any contamination. The aquariums were then filled with 120 L of this water. Each aquarium had a closed-circuit water circulation and was equipped with an air bubbles exhauster in a tube that brought up the water on a neutral synthetic foam filter. The aquariums were thus oxygenated, and the coarsest suspended matter was filtered out. The remaining seawater in the tank was used as a negative control (Aquarium 1). Nine to eleven fish were added to each of the four aquariums (Figure 1). The aquarium water was sampled six hours after introducing the fish into the aquariums using an Athena peristaltic pump with a nominal flow of 1.0 L/min to filter 30 L and VigiDNA 0.2 µm crossflow filtration capsules (SPYGEN) with disposable sterile tubing. After filtration, 80 ml of CL1 conservation buffer (SPYGEN) was added before storing the samples at ambient temperature.

We reanalyzed here two eDNA samples of 30 L replicate each, collected in the Mediterranean Sea, at Banyuls (France, coordinates: 42.41568 and 3.17110) and Calvi (France, coordinates: 42.62964 and 8.89161) published in a previous metabarcoding analysis and known to contain *M. surmuletus* sequences (detected with the metabarcode teleo 12S) (Boulanger et al., 2021). These two Mediterranean eDNA samples were amplified and sequenced using the primers developed for this study and then analyzed using the best-performing pipeline as determined by our evaluation. These two samples were used as proof of concept of the possibility to estimate within-site variability in real conditions.

2.2 | Designing an intraspecific barcode

The first step in developing a barcode for the intraspecific analysis of *M. surmuletus* is to define a primer pair for the PCR amplification that can detect the species (*M. surmulletus*) and target a sequence with intraspecific variability. The D-loop region of the mtDNA is the most variable region of the mtDNA with high nucleotide variation and is a good candidate region to look for an intraspecific barcode (Najjar Lashgari et al., 2017; Xie et al., 2006). A search for *Mullidae* D-loop sequences in the nucleotide NCBI (National Center for Biotechnology Information) database (NCBI Resource Coordinators, 2016) detected partial sequences, but no complete referenced D-loop for *M. surmuletus*. We sequenced the

complete D-loop for 21 individuals collected by fishers from the Mediterranean Sea (12 individuals), Atlantic Ocean (4 individuals), and North Sea (5 individuals) using Sanger sequencing. DNA extraction and amplification were performed in separate, dedicated rooms. Tissue DNA from the fins was extracted using the DNeasy Blood & Tissue kit (QIAGEN, Venlo, Netherlands). The complete D-loop was amplified with the forward primer MS-CYTB-1 (5′-AAGGCCCTGCAATGAACA-3′) and the reverse primer MS-12S-1 (5′-GGTGGCTGGCACGAGTTT-3′). MS-CYTB-1 was built from an alignment of the complete cyt-b sequences available in the nucleotide NCBI database for *M. surmuletus*, and it is positioned at the end of the cyt-b gene. MS-12S-1 is the reverse V05F-898 primer (Thomsen et al., 2016). The amplified fragment was 1396 bp long, including the sequences tRNA-Thr, tRNA-Pro, the complete D-loop, tRNA-Phe, and a short section of 12S rRNA. PCRs were carried out in 10 µl volume containing 1X of the REDExtract-N-Amp PCR reaction mix (Sigma-Aldrich Co. LLC, Saint Louis, MO, USA), 0.25 pM of each primer, and 2 µl of DNA. The PCR mixture was denatured at 94°C for 30 s, followed by 45 cycles of 30 s at 94°C, 30 s at 59°C, and 1 min at 72°C and a final elongation step at 72°C for 5 min. The purification and sequencing of PCR products were carried out by Eurofins Genomics Germany GmbH (Ebersberg, Germany). The chromatograms were checked using CodonCode Aligner version 4.2.7 (CodonCode Corporation, Dedham, MA, USA).

We used the ECOPRIMERS program (Riaz et al., 2011) to detect all the pairs of primers that could potentially amplify a marker sized from 100 to 300 bp of the partial D-loop sequences downloaded from NCBI and the additional 21 "home" complete sequences. We then tested these primer pairs with *in silico* PCR using the ECOPCR program (Ficetola et al., 2010), allowing three mismatches in the primer sequence. The amplifications all returned sequences with a maximum length of 1000 bp amplified from the ENA (European Nucleotide Archive) database (Amid et al., 2020, release 140). Intraspecific variation is measured by the number of polymorphic sites.

2.3 | Reference database

Tissue fragments were collected from the fins of the 39 fish individuals. Tissue DNA was extracted following the same protocol as for the primer design. For PCR amplification, we defined the forward primer MS-DL06-F (5′-TGATATAGGACACGATAT-3′) and the reverse primer MS-DL06-R (5′-TGTCCCTCACCTTCAATA-3′). PCRs were carried out in 10 µl volume containing 1X of the REDExtract-N-Amp PCR reaction mix (Sigma-Aldrich Co. LLC), 0.2 pM of each primer, and 1 µl of DNA. The PCR mixture was denatured at 95°C for 30 s, followed by 35 cycles of 30 s at 95°C, 1 min at 47°C, and 30 s at 72°C and a final elongation step at 72°C for 10 min. The purification and sequencing of PCR products were carried out by Eurofins Genomics Germany GmbH. The chromatograms were checked using CodonCode Aligner version 4.2.7 (CodonCode Corporation).

The sequences of these 39 individuals constitute the reference database for the aquarium study. Ambiguities (bases other than ACTG) were observed at the ends of each of the 39 sequences. Since sequences with a similar length were needed for the study, all the sequences were trimmed after a common pattern around 225 pb (5′-CCTACCATC-3′). Trimming shortened the sequence length from 235–259 bp to 233–235 pb. After trimming, 37 out of the 39 sequences were distinct. The individuals with the same haplotype after trimming were not in the same aquarium. This reference database is presented in Table S1.

For the re-analysis of the two eDNA samples collected from the Mediterranean Sea and known to contain *M. surmuletus*, we added the 21 individuals sampled in the Mediterranean Sea, the Atlantic, and the Northern Sea, and sequenced for the D-loop in this study to the reference database (Table S2).

2.4 | Environmental DNA and genetic data

DNA extraction and amplification from eDNA samples were performed by the company SPYGEN (Le Bourget-du-Lac, France) in separate, dedicated rooms following the protocol described by Polanco Fernández et al. (2021). The amplification was performed in a final volume of 25 µl including 1 U of AmpliTaq Gold DNA Polymerase (Applied Biosystems, Foster City, CA, USA), 10 mM of Tris-HCl, 50 mM of KCl, 2.5 mM of MgCl2, 0.2 mM of each dNTP, 0.2 µM of each primer, 0.2 µg/µl of bovine serum albumin (Roche Diagnostics, Basel, Switzerland), and 3 µl of DNA template. The PCR mixture was denatured at 95°C for 10 min, followed by 50 cycles of 30 s at 95°C, 30 s at 47°C, 1 min at 72°C, and a final elongation step at 72°C for 7 min. The primers were 5'-labelled with an eight-nucleotide tag unique to each DNA sample, allowing each sequence to be assigned to the corresponding sample during the sequence analysis. Twelve replicate PCRs were run per sample. Two libraries were prepared using the MetaFast protocol (Fasteris 2020, https://www.fasteris.com/dna/), and the sequencing was performed by Fasteris (Geneva, Switzerland) on two separate runs on an Illumina MiSeq (2x250 bp) (Illumina, San Diego, CA, USA) and the Miseq Kit v3 (Illumina) following the manufacturer's instructions. Two negative extraction controls and one negative PCR control (12 replicates of ultrapure water) were amplified and sequenced to monitor for possible contaminants (Polanco Fernández et al., 2021).

2.5 | Bioinformatics pipelines

We tested twelve pipelines combining five common bioinformatics programs: the *obiclean* program available in the OBITOOLS toolkit (Boyer et al., 2016); the denoising program DADA2 (Callahan et al., 2016), which corrects and gathers sequences into amplicon sequence variants (ASVs); the clustering program SWARM (Mahé et al., 2014) alone; and followed by the post-clustering LULU program (Frøslev et al., 2017), and an additionnal chimeric removal step

based on DADA2 and VSEARCH. These five bioinformatics programs constitute the key processing steps of the four groups of pipelines, respectively A, B, C, and D. After this key processing step, the only other differences between the pipelines in each group are the post-processing steps (Figure 2). Either there is no post-processing (pipelines A1, B1, C1, and D1), or there is a chimeric sequences removal step, using the *removeBimeraDenovo* function of the program DADA2 (pipelines A2, B2, C2, and D2), or the *uchime3_denovo* command (pipelines A3, B3, C3, and D3) from the VSEARCH toolkit (Rognes et al., 2016) (Figure 2). A pipeline combines the successive steps resolved by various pre-existing programs to produce the output.

The first steps to pre-process data are identical for all pipelines: *illuminapairedend* (OBITOOLS) is used to align and merge paired-end reads, *ngsfilter* (OBITOOLS) to demultiplex using the tags and remove primer sequences, allowing a maximum of two mismatches for the primer sequences; *filterAndTrim* (DADA2) to discard any sequences containing ambiguities (nucleotides other than ACTG), with the maximum number of expected errors tolerated in a read set at 1 (calculated from quality score), and to trim sequences at 235 bp, which corresponded to the smallest size of reference sequences before their initial trimming; and *derepFastq* (DADA2) to dereplicate sequences (Figure 2). Then, the four groups of pipelines differ according to the processing step (Figure 2). For all pipelines, an abundance filter is applied after the key processing step, to remove sequences with less than 10 reads using *obigrep* (OBITOOLS). This filter of at least 10 reads is commonly applied in similar studies (*e.g.*, Duarte et al., 2021; Marques et al., 2020). Finally, the post-processing steps will produce 12 different pipelines (Figure 2).

For group A pipelines (pipelines A1, A2, and A3), the core program tested was the *obiclean* algorithm in which sequence variants identified as potential PCR errors were flagged and removed. The *obiclean* program (Boyer et al., 2016) functions using both sequence dissimilarity (number of mismatches) and the ratio of abundance between a pair of sequences, with both parameters controlled by the user. We used thresholds of 1 mismatch and a ratio of abundance of 0.05, so any sequence which has 1 mismatch with another sequence and less than 5% of its abundance within the same sample is considered as a variant of the most abundant sequence and is discarded.

For group B pipelines (pipelines B1, B2, and B3), we used the denoising program DADA2 (Callahan et al., 2016). The purpose of DADA2 was to denoise sequences by correcting errors, which are mostly generated during sequencing (Callahan et al., 2016). It uses the sequencing quality scores to filter sequences considered to be erroneous, returning only corrected sequences known as amplicon sequence variants (ASVs). The program first partitions the sequences and defines the most abundant sequences as the core of the cluster. All other sequences are compared with this core based on an abundance p-value defined by default in the pipeline. If the *p*-value of the sequence with the lowest *p*-value is below the threshold, the sequence is removed from the partition, to become the core of a new partition. The comparison is repeated in each partition until there are no sequences with an abundance p-value below the threshold. Each partition is then represented by its core, which is considered to be the original sequence from which all other sequences in the partition are created. Each of those sequences are then considered as individual ASVs. The error estimation model is produced using the *learnErrors* function, and the correction of the sequences into ASVs using the *dada* function.

For group C pipelines (pipelines C1, C2, C3, and C4), we used the clustering program SWARM (Mahé et al., 2015), which clusters similar sequences into operational taxonomic units (OTUs). The algorithm makes a pairwise alignment between sequences, counts the mismatches, and makes a network based on the sequence's relative abundance. The network is then broken at the most appropriate section to form OTUs, based on a user-chosen threshold for the minimum distance between a pair of OTUs (here, 1). The most

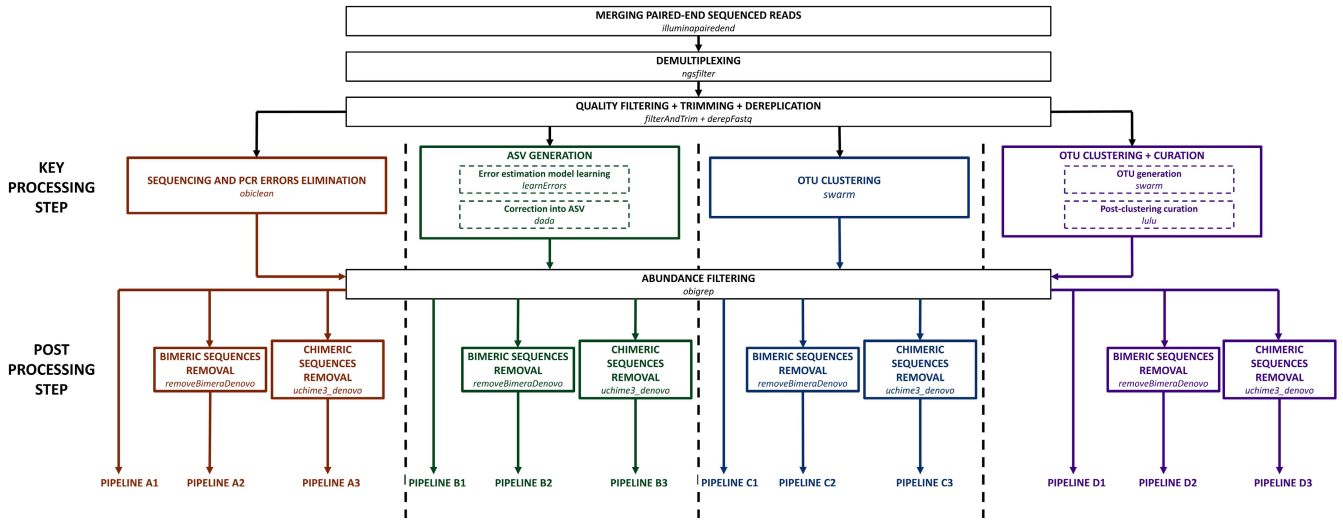

FIGURE 2 Summary of the twelve pipelines compared, with the key processing steps and the post-processing steps colored for each pipeline group

abundant sequence within each OTU is assigned as the representative sequence. Only the representative sequences were analyzed. The other sequences forming the OTUs were considered as errors.

For group D pipelines (pipelines D1, D2, D3, and D4), we combined SWARM algorithm with an additional post-clustering step implemented in the program LULU (Frøslev et al., 2017). LULU eliminates OTUs by flagging the alleged erroneous OTUs of more abundant OTUs and merges them. The algorithm requires an OTU match list to provide the pairwise similarity scores of the OTUs, with a minimum threshold of sequence similarity set at 84% as recommended by the authors. Only OTU pairs with a sequence similarity above 84% can then be interpreted as "parent" for the most abundant one and "daughter" for the other. Both OTU will possibly be merged provided that the co-occurrence pattern of the OTU pair among samples is higher than 95% and the abundance ratio between the "potential parent" and "potential daughter" is higher than a minimum ratio set by default as the minimum observed ratio.

A post-processing step based on chimeric sequences removal is added to all pipelines except pipelines A1, B1, C1, and D1. For pipelines A2, B2, C2, and D2, sequences considered as bimeras, or two-parent chimeras, are removed using *removeBimeraDenovo* (DADA2 function). This function mostly points out bimeras by aligning each sequence with all more abundant sequences and detecting a combination of an exact "right parent" and an exact "left parent" of this sequence (Callahan et al., 2016). For pipelines A3, B3, C3, and D3, chimeras are removed using *uchime3_denovo* (VSEARCH). This command is based on the UCHIME2 algorithm (Edgar, 2016). Each sequence is divided into four segments, and the command mostly searches for similarity for each segment to all other sequences using a heuristic method. The best potential parent sequences are then selected, and the query sequence is considered as chimera if a set of default parameters is not exceeded (Rognes et al., 2016).

The comparison between the twelve pipelines was made for each aquarium. For each pipeline output, the sequences were trimmed following the same pattern as for the reference database, to find which correspond to true haplotypes.

2.6 | Statistical analyses

Three metrics were calculated for each aquarium to compare the twelve pipelines: sensitivity, haplotype precision, and read precision. Sensitivity represents the ratio between the number of validated true haplotypes returned by each pipeline and the number of known true haplotypes, which is the number of individuals in the aquarium considered. A sensitivity value of one means that all true haplotypes were detected, and a sensitivity value below one means that some true haplotypes were not detected. Haplotype precision represents the ratio between the number of true haplotypes returned by each pipeline and the total number of haplotypes returned. A haplotype precision value below one means that some haplotypes recovered represent errors and not real haplotype diversity. Read precision represents the ratio between the number of reads corresponding to true haplotypes and the total number of reads returned, values closer to one indicate that most reads correspond to true haplotypes. All statistical analyses were performed using R version 4.0.2 (R Core Team, 2021). Kruskal–Wallis tests were carried out to test mean differences in metrics between pipelines. A Wilcoxon–Mann–Whitney test was used to test the mean difference between the number of true haplotype reads and false-positive haplotype reads returned by pipeline B2. Unless otherwise specified, all figures were produced using ggplot2 R package.

The two eDNA Mediterranean samples were analyzed using the best pipeline, as determined by this study, and using the completed reference database. We investigated both haplotypes from the reference database, and potential new haplotypes, and compared the haplotype composition between the two sites.

3 | RESULTS

3.1 | Primer design

The program ECOPRIMERS defined six primer pairs on the D-loop (Figures S1, S2, Table S3). The marker length obtained for all the potential candidate markers ranged from 119 pb (marker MS-DL02), with 17 polymorphic sites identified, to 259 bp, with 39 polymorphic sites (Table S3). The *in silico* PCR analysis of the six pairs of primers conducted using ECOPCR revealed that the primer pair MS-DL02 amplified three other marine fish species with no mismatches in the forward or reverse primer sequences. This pair therefore could not be used for an intraspecific study and was excluded. Allowing up to three mismatches in both primer sequences, sequences from species other than *Mullus surmuletus* were amplified for all the primer pairs (Figures S1, S2). Among the remaining five potential candidate primer pairs, MS-DL06 was found to be the best candidate. It amplified *Mullus surmuletus* sequences with no mismatches in the forward and reverse primers for the 21 D-loop sequences and only one sequence of marine fish species when three mismatches on both primer sequences were allowed (Figure S2). This primer pair amplifies a sequence of 259 bp that has the highest number of polymorphic sites (39).

3.2 | Ratio of truly returned haplotypes to true known haplotypes (sensitivity)

None of the twelve pipelines detected any haplotypes in the negative control (Aquarium 1), and this aquarium was not considered in further analysis. The mean sensitivity across aquariums ranged from 0.741 (pipelines D1, D2, and D3) to 1.000 (pipelines A1, A2, A3, B1, B2, and B3) indicating that in average, at least 74.1% of true haplotypes were returned by the pipelines D1, D2, and D3, while all true haplotypes were returned by pipelines A1, A2, A3, B1, B2 and B3 (Figure 3a). The difference in the mean sensitivity between pipelines is significant (Kruskal–Wallis test, $p < 0.05$). The sensitivity in each

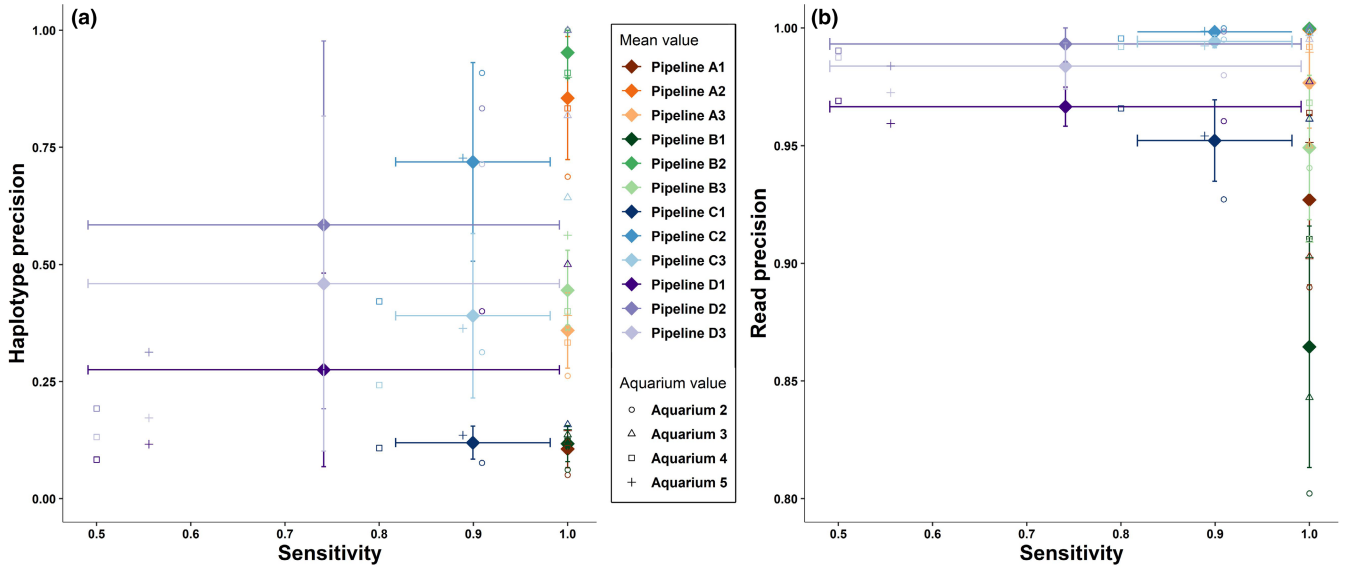

FIGURE 3 Pipeline sensitivity and precision. The scatter plot represents the values of precision in relation to sensitivity in each aquarium for each pipeline (Aquarium values) and the mean values across aquarium (Mean value) for haplotypes (a) or reads (b). Error bars correspond to the standard deviation

aquarium ranged from 0.500 in aquarium 4 for pipelines D1, D2, and D3 (5 true haplotypes lost in Aquarium 4) to 1 for pipelines of groups A and B (Figure 3a; Table S4). Similar pipelines ranking was obtained for reads sensitivity (Figure 3b).

3.3 | Ratio of true haplotypes/reads to all haplotypes/reads (precision)

Mean haplotype precision across aquariums ranged from 0.106 (pipeline A1) to 0.952 (pipeline B2) and average read precision from 0.864 (pipeline B1) to 0.999 (pipelines A2 and B2) (Figure 3b; Table S4). Pipeline B2 has the highest precision of all the pipelines, returning 95.2% of true haplotypes and 99.9% of reads fitting to true haplotypes, while the other pipelines on average only returned 40.3% of true haplotypes and 96.4% of reads fitting to the true haplotypes. The differences in mean read and haplotype precisions between pipelines are significant (Kruskal–Wallis test, $p < 0.05$).

3.4 | ASVs pipelines and the effect of removing bimeras

Pipelines B1 and B2, which are similar except for the additional bimeras removal step in pipeline B2, returned the same number of true haplotypes, that is, 9 (Aquarium 5) to 11 (Aquarium 2) (Figure 4a,c). No true haplotypes were lost during the bimera removal step while the number of haplotypes returned was divided by 9.44 on average (from 96.75 to 10.25), increasing mean haplotype precision from 0.117 to 0.952 (Table S4). These results show that pipeline B1 identified a large number of sequences identified as bimeras (87.3% on average). Removing the bimeras increased read precision from 0.864 to 0.999 (Table S4) on average. The reads generally contained fewer bimera sequences than true haplotypes (Figure 4b,d), and bimeras were generally less abundant than true haplotypes (Figure 4e). However, in some cases, the bimera read counts were more abundant than true haplotypes (Aquarium 5, Figure 4e). After bimera removal, only two false-positive haplotypes were returned by pipeline B2 (Figure 4f), and all of those had a lower read count compared with the true haplotypes, and 13 (Aquarium 4) and 27 (Aquarium 5) reads returned, when the weakest read count for a true haplotype was 226 (Aquarium 5). The order of magnitude of the mean read count for these false-positive haplotypes was significantly weaker than for the true haplotypes (Wilcoxon–Mann–Whitney test, $p < 0.05$). The maximum ratio of the read count for a false-positive haplotype to the total read count for the aquarium considered was 7.86×10^{-4}, while this ratio was at least 6.57×10^{-3} for the true haplotypes.

3.5 | Haplotype detection in eDNA sea samples

With no additional bioinformatics filters, pipeline B2 returned a total of 187 haplotypes, composed of 138 totally different haplotypes in the Banyuls eDNA sample and 49 haplotypes in the Calvi sample, two of which (one from Banyuls and one from Calvi) were in the full reference database (Figure 5). These two haplotypes were different and corresponded to referenced haplotypes for the Mediterranean individuals. Using a threshold of 6.57×10^{-3} per site for the minimal relative read count in both samples, four unreferenced haplotypes were identified as possible true haplotypes in the Banyuls eDNA sample, and 14 in the Calvi sample (Figure 5).

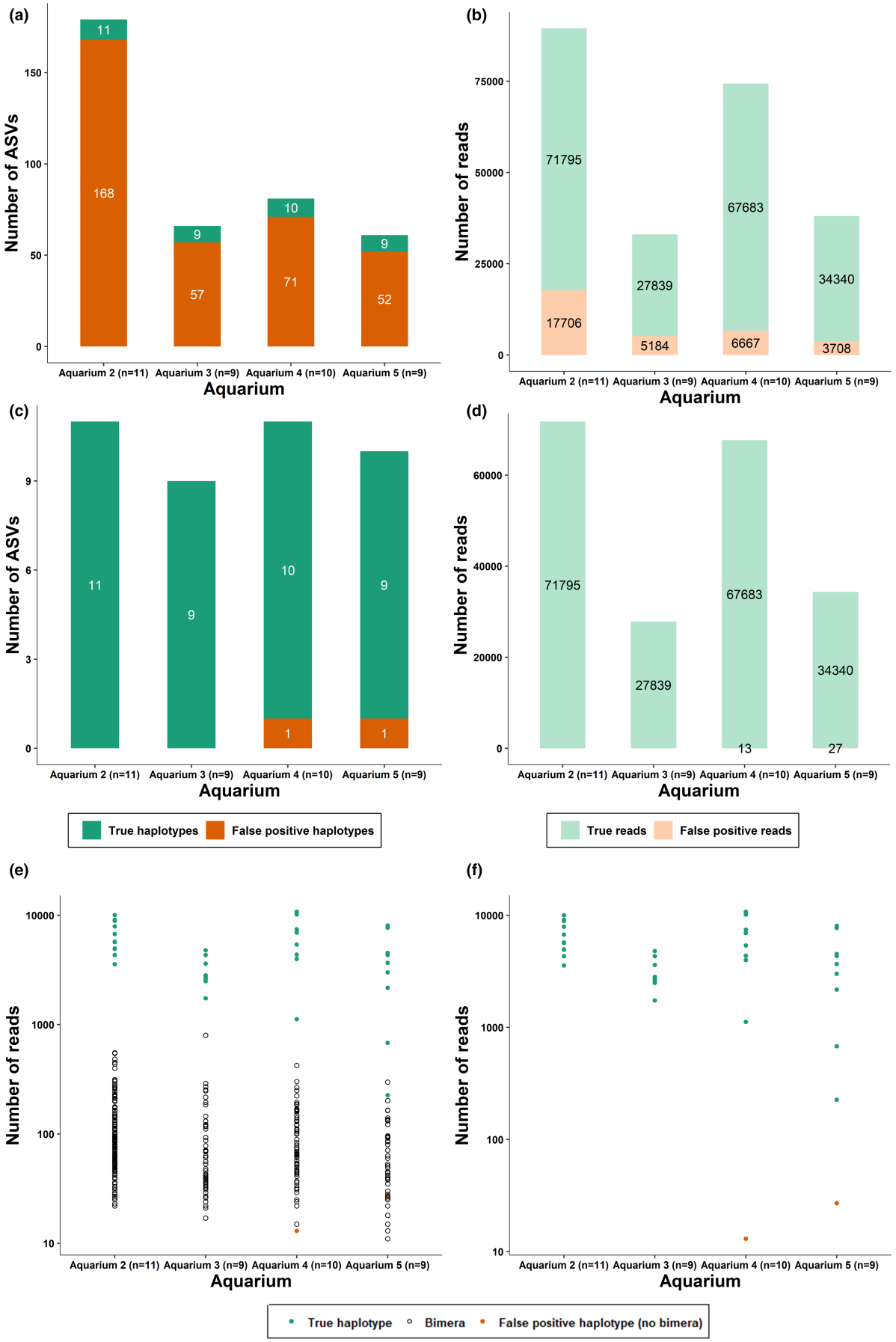

FIGURE 4 Legend on next page

FIGURE 4 Comparison of pipelines B1 with B2. Bar charts representing the number of ASVs corresponding to true- and false-positive haplotypes returned by pipeline B1 (a) and pipeline B2 (c); and the number of reads corresponding to true- and false-positive haplotypes returned by pipeline B1 (b) and pipeline B2 (d). Scatterplot overlapping the number of reads for each true haplotype and false-positive haplotype (bimeras and others) returned by pipeline B1 (e) and pipeline B2 (f). A decimal logarithmic scale is used for the vertical axis

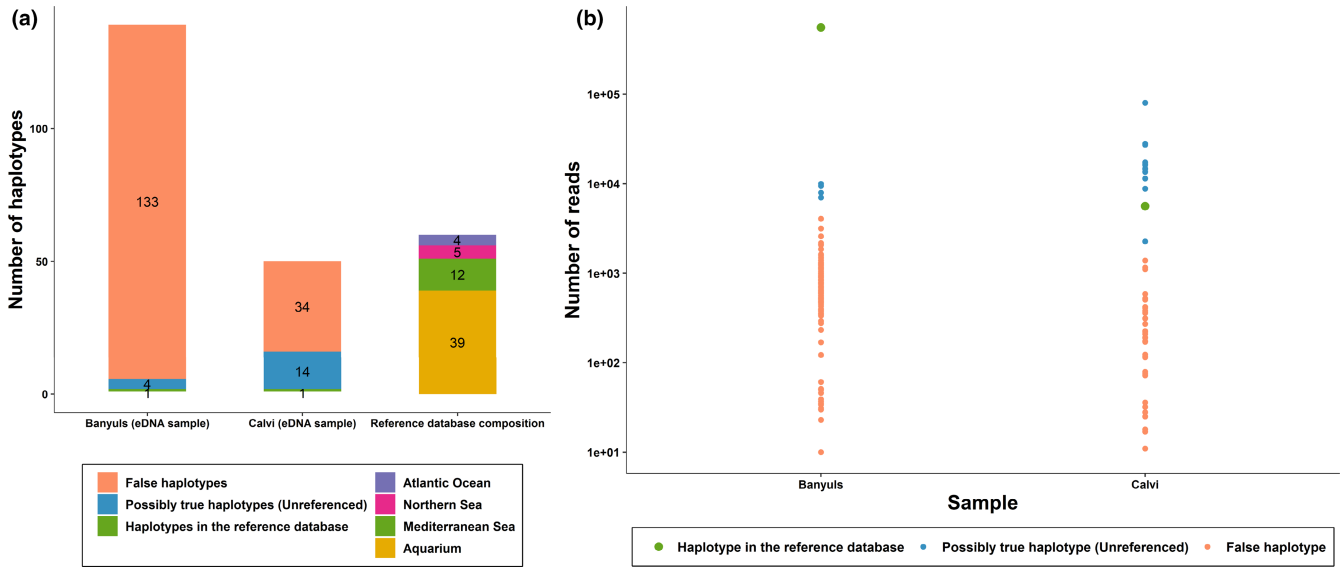

FIGURE 5 Results for ocean eDNA samples analyzed with pipeline B2 with a threshold of 6.57×10^{-3} per site for the minimum relative read count in both samples. (a) Bar charts representing the number of haplotypes in the reference database, false-positive haplotypes, and possibly true haplotypes returned for eDNA samples from Banyuls and Calvi, in comparison with the composition of the reference dataset built to design the barcode. (b) Scatterplot overlapping the number of reads for haplotypes in the reference database, false-positive haplotypes, and "possibly true haplotypes." A decimal logarithmic scale is used for the vertical axis

4 | DISCUSSION

Using a species-specific polymorphic mitochondrial barcode, we found that correcting sequences with a denoising algorithm (DADA2) combined with the removal of two-parent chimeric sequences (DADA2) significantly improve the accuracy of intraspecific genetic diversity estimation using eDNA in a controlled environment. The same algorithm was also able to detect two haplotypes from our reference database in the eDNA sea samples, and 18 other possibly true haplotypes not present in our tissue collection. The haplotypes were different between the two sites, suggesting genetic variation. Overall, the denoising method with an additional bimeric sequence removal step combined with a specific polymorphic MT barcode is promising for intraspecific studies using eDNA possible.

An important result of our study is that for all groups of pipelines, the pipelines with the bimeric sequences removal step with DADA2 always provide the best precision (Figure 3). Then, the OBITOOLS pipeline combined with DADA2 (A2) ranks as the second best pipeline (Figure 3). A noticeable result is the fact that except for aquarium 3, true haplotypes are systematically lost with SWARM OTU clustering in C and D pipelines. This can be explained by the fact that haplotypes are not different enough, particularly in aquariums 4 and 5, and they are clustered in the same OTU. It also explains why for these two aquariums, sensitivity is lower with LULU, because this algorithm merges the closest OTUs. However, those results have been obtained in ideal condition with low DNA degradation (aquarium) for one species, *Mullus surmuletus*, with one marker in the D-loop, and applied to only two sites *in situ*. More studies are needed to confirm their application to eDNA intrapopulation studies *in situ*.

The large number of sequences identified as bimeric decreases the precision of true haplotype detections (Figure 4). Our results concur with previous studies which have shown that a large number of chimeric sequences can impact the results of downstream analyses if these sequences are not detected (Schloss et al., 2011; Smyth et al., 2010). We showed that both bimeras and other false-positive haplotypes generally have a low read count compared with true haplotypes (Figure 4) suggesting that a read relative abundance filter could help to remove false positives and chimeras. However, bimera removal should be done before applying a read relative abundance filter since some bimeras can have a higher or similar number of reads than true haplotypes. Applying a read relative abundance filter without having first removed the bimeras would also remove some low-abundance true haplotypes. In our study, removing all ASVs with a read count ratio of less than 6.57×10^{-3} of the total sample read count after the bimeras removal step eliminated all the remaining false-positive haplotypes without losing any true haplotypes. Following our experimental results in the aquarium, we selected this threshold as the minimal ratio for the read count of a true-positive haplotype to the total read count for the aquarium considered, that is, 6.57×10^{-3}.

However, this type of threshold-based filtering should be tested and adapted to each study upstream since it can lead to the loss of true haplotypes (Prodan et al., 2020), especially if some haplotypes are less abundant than others. As eDNA concentrations are lower *in situ* in a marine environment (Doi et al., 2017; Tsuji et al., 2020a,b), we expect a lower threshold value would be required for studies in natural conditions. However, in our two Mediterranean samples, this threshold (6.57×10^{-3}) allowed us to define a total of 20 haplotypes, two of which were assigned to the reference database (Figure 5). Using a lower threshold of 1/1000, we found 72 possibly true haplotypes in Banyuls and 27 in Calvi, but potentially increased the risk of obtaining false positives (Figure S3). Overall, adequate bioinformatics filters (chimera removal and read count) substantially improve haplotype detection.

Reducing the number of PCR cycles has been recommended as a way of limiting chimera formation during PCR (Lahr & Katz, 2009). Smyth et al. (2010) demonstrated that chimera formation could be considerably decreased by not exceeding 29 cycles and Holcomb et al. (2014) recommended no more than 28 cycles. In conditions with low DNA concentrations, Lahr and Katz (2009) demonstrated that no chimeras were formed when the number of PCR cycles was reduced to 30. In our study, we applied 50 PCR cycles, which might contribute to the large number of chimeras. Reducing the number of PCR cycles for future analyses could therefore help prevent the formation of chimeras. This includes chimeras generated by having more than two initial templates when the *removeBimeraDenovo* function fails to remove them. Nevertheless, a large number of PCR cycles (>40) are necessary in eDNA studies due to the very low concentration of DNA in environmental samples (Klymus et al., 2020). Further studies might need to find a compromise between reducing the number of PCR cycles to limit chimera formation and maintaining a sufficient number of cycles to detect the eDNA of target species at low concentrations.

In order to detect false-positive haplotypes, Tsuji et al. (2020a,b) used a filter not solely based on read count, but also on haplotype presence rate for different PCR replicates from one sample. We were not able to test this filter since the 12 PCR replicates in our experiment were pooled. Having unique tags for each PCR replicate instead of pooling all replicates under the same tag could help to distinguish true haplotypes. Erroneous sequences are expected to have a lower incidence among PCR replicates as they are generated randomly during amplification and sequencing. It would however increase the cost of data production. Sigsgaard et al. (2016) also suggested discarding sequences present in only one PCR replicate, using a total of six PCR replicates per sample. However, other authors observed that the same false haplotypes can be present in multiple PCR replicates (Elbrecht et al., 2018). Keeping only the haplotypes present in all PCR replicates could be a suitable filtering method, but it is unlikely to work *in situ* in a marine environment where DNA concentration is low and true sequences are rarely present in all PCR replicates.

Fish primers as Mifish are usually used to detect species or above taxonomic level in metabarcoding studies. They do not contain enough variability to detect intraspecific variability (Miya et al., 2015). This is why we have developed new primer pair for *M. surmuletus* containing intraspecific variations, but the primer design is associated with some limitations. Out of all the primer pairs generated, we selected MS-DL06 due to its ability to amplify the target species' DNA and avoid non-specific amplifications. Only one other marine species (*Gouania willdenowi*) can be amplified with this primer pair (Figure S2) with three nucleotide mismatches on each primer based on the current genetic NCBI database (accession: LR131985). Since the sequence amplified for *G. willdenowi* is a nuclear sequence, presenting with lower numbers of copies compared with mitochondrial sequences, the risk of cross-detection in marine samples should remain low (Birky-Jr et al., 1989; Olson et al., 2012). Moreover, since the *G. willdenowi* sequence amplified *in silico* was 198 bp long compared to at least 235 bp for *M. surmuletus*, it could easily be detected after sequencing. Other species are also amplified by the primer pair when more than three mismatches are allowed in forward and reverse primers, but it is recognized that cross-amplification decreases dramatically when the number of mismatches per primer is higher than three (Housley et al., 2006). Nevertheless, we also checked that MS-DL06 primers do not amplify sequences of *Mullus barbatus*, a closely related and sympatric species of our target species *M. surmuletus* (Lombarte et al., 2000). Therefore, we expect the careful definition of a specific polymorphic barcode to improve population genetic inference from eDNA.

Our study investigates the methodological considerations around using eDNA methods to obtain reliable estimations of intraspecific genetic variation in the seas. It was based on the use of a barcode to amplify DNA with PCR and sequencing steps. Recently, the use of hybridization capture method has opened up promising perspectives in the field for using low concentration nuclear genes (Jensen et al., 2021). All recent studies in this area have demonstrated that although eDNA population-level inference is in its infancy, it will soon be possible to move from proof of concept to use, in order to address evolutionary questions.

ACKNOWLEDGEMENTS

We would like to thank GOMBESSA 5 cruise and Zembra vessel for supporting our fieldwork and the PLATAX experimental aquaculture platform of ISEM for their support with the experimental set-up. We thank Rémi Dugué and Christophe Cochet for assisting us with the animal experiments in the aquariums. We thank Daniel Pinelli and Apolline Gorry for their involvement preliminary analysis. We thank Thomas Broquet, François Leven, An Distro Ruadhan Gillespie-Mule, and Peter J Wright to sample individuals of *Mullus surmuletus*. We are grateful to Émilie Boulanger for her help in choosing the appropriate Mediterranean eDNA samples. This project was funded by the MUSE KIM SEA & COAST AAP International 2019 and the SEAMONT ANR. The pipelines deposit "edna_intra_pipeline_comparison" benefited from the Montpellier Bioinformatics Biodiversity platform supported by the LabEx CeMEB, an ANR "Investissements d'avenir" program (ANR-10-LABX-04-01).

CONFLICT OF INTEREST

The authors declared no conflicts of interest.

AUTHOR CONTRIBUTION

SM conceived the study. SM and RH designed and performed the fieldwork. SM, VA, and AV designed and performed the molecular laboratory work. PEG designed the bioinformatics pipelines. BM analyzed the data. BM, SM, LP, VM, RH, and AV were involved in the writing of the manuscript. All authors approved the final version.

DATA AVAILABLITY STATEMENT

The raw dataset containing the NGS eDNA sequencing is available at https://zenodo.org/record/4570303. The pipelines are available at https://gitlab.mbb.univ-montp2.fr/edna/exploitation/edna_intra_pipeline_comparison

ORCID

Bastien Macé https://orcid.org/0000-0001-7721-3993
Régis Hocdé https://orcid.org/0000-0002-5794-2598
Pierre-Edouard Guerin https://orcid.org/0000-0001-7909-3729
Alice Valentini https://orcid.org/0000-0001-5829-5479
Véronique Arnal https://orcid.org/0000-0002-4037-490X
Loïc Pellissier https://orcid.org/0000-0002-2289-8259
Stéphanie Manel https://orcid.org/0000-0001-8902-6052

REFERENCES

Adams, C. I. M., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool? *Genes*, *10*(3), 192. https://doi.org/10.3390/genes10030192

Alomar, C., Sureda, A., Capó, X., Guijarro, B., Tejada, S., & Deudero, S. (2017). Microplastic ingestion by Mullus surmuletus Linnaeus, 1758 fish and its potential for causing oxidative stress. *Environmental Research*, *159*, 135–142. https://doi.org/10.1016/j.envres.2017.07.043

Amid, C., Alako, B. T. F., Balavenkataraman Kadhirvelu, V., Burdett, T., Burgin, J., Fan, J., Harrison, P. W., Holt, S., Hussein, A., Ivanov, E., Jayathilaka, S., Kay, S., Keane, T., Leinonen, R., Liu, X., Martinez-Villacorta, J., Milano, A., Pakseresht, A., Rahman, N., … Cochrane, G. (2020). The European nucleotide archive in 2019. *Nucleic Acids Research*, *48*(D1), D70–D76. https://doi.org/10.1093/nar/gkz1063

Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Molecular Ecology*, *30*(3), 685–697. https://doi.org/10.1111/mec.15765

Ardura, A., Zaiko, A., Martinez, J. L., Samulioviene, A., Semenova, A., & Garcia-Vazquez, E. (2015). EDNA and specific primers for early detection of invasive species—A case study on the bivalve *Rangia cuneata*, currently spreading in Europe. *Marine Environmental Research*, *112*, 48–55. https://doi.org/10.1016/j.marenvres.2015.09.013

Baker, C. S., Steel, D., Nieukirk, S., & Klinck, H. (2018). Environmental DNA (eDNA) from the wake of the whales: Droplet digital PCR for detection and species identification. *Frontiers in Marine Science*, *5*. https://doi.org/10.3389/fmars.2018.00133

Birky-Jr, C. W., Fuerst, P., & Maruyama, T. (1989). Organelle gene diversity under migration, mutation, and drift: Equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics*, *121*(3), 613–627. https://doi.org/10.1093/genetics/121.3.613

Boulanger, E., Loiseau, N., Valentini, A., Arnal, V., Boissery, P., Dejean, T., Deter, J., Guellati, N., Holon, F., Juhel, J.-B., Lenfant, P., Manel, S., & Mouillot, D. (2021). Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1949). https://doi.org/10.1098/rspb.2021.0112

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*(13), 2937–2958. https://doi.org/10.1111/mec.15472

Deter, J., Juhel, J.-B., Boulanger, E., Guellati, N., Mauron, S., Holon, F., Mouillot, D., & Hocdé, R. (2020). Gombessa 5 cruise: CTD profiles in western Mediterranean, July 2019. *SEANOE*. https://doi.org/10.17882/71814

Doi, H., Inui, R., Akamatsu, Y., Kanno, K., Yamanaka, H., Takahara, T., & Minamoto, T. (2017). Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology*, *62*, 30–39. https://doi.org/10.1111/fwb.12846

Duarte, S., Vieira, P. E., Lavrador, A. S., & Costa, F. O. (2021). Status and prospects of marine NIS detection and monitoring through (e) DNA metabarcoding. *Science of the Total Environment*, *751*, 141729. https://doi.org/10.1016/j.scitotenv.2020.141729

Edgar, R. C. (2016). UCHIME2: Improved chimera prediction for amplicon sequencing. *BioRxiv*, 74252. https://doi.org/10.1101/074252

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, *6*, e4644. https://doi.org/10.7717/peerj.4644

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*(1), 434. https://doi.org/10.1186/1471-2164-11-434

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*(4), 423–425. https://doi.org/10.1098/rsbl.2008.0118

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1188. https://doi.org/10.1038/s41467-017-01312-x

Furlan, E. M., Davis, J., & Duncan, R. P. (2020). Identifying error and accurately interpreting environmental DNA metabarcoding results: A case study to detect vertebrates at arid zone waterholes. *Molecular Ecology Resources*, *20*(5), 1259–1276. https://doi.org/10.1111/1755-0998.13170

Holcomb, C. L., Rastrou, M., Williams, T. C., Goodridge, D., Lazaro, A. M., Tilanus, M., & Erlich, H. A. (2014). Next-generation sequencing can reveal in vitro-generated PCR crossover products: Some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. *Tissue Antigens*, *83*(1), 32–40. https://doi.org/10.1111/tan.12269

Housley, D. J., Zalewski, Z. A., Beckett, S. E., & Venta, P. J. (2006). Design factors that influence PCR amplification success of cross-species

primers among 1147 mammalian primer pairs. *BMC Genomics*, *7*(1), 253. https://doi.org/10.1186/1471-2164-7-253

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., Møller, P. R., & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources*, *21*(3), 690–702. https://doi.org/10.1111/1755-0998.13293

Klymus, K. E., Ramos, D. V. R., Thompson, N. L., & Richter, C. A. (2020). Development and testing of species-specific quantitative PCR assays for environmental DNA applications. *Journal of Visualized Experiments*, *165*, e61825. https://doi.org/10.3791/61825

Lahr, D. J. G., & Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, *47*(4), 857–866. https://doi.org/10.2144/000113219

Lombarte, A., Recasens, L., González, M., & de Sola, L. G. (2000). Spatial segregation of two species of Mullidae (*Mullus surmuletus* and *M. barbatus*) in relation to habitat. *Marine Ecology Progress Series*, *206*, 239–249. https://doi.org/10.3354/meps206239

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. https://doi.org/10.7717/peerj.593

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420. https://doi.org/10.7717/peerj.1420

Mahe, K., Destombe, A., Coppin, F., Koubbi, P., Vaz, S., Le Roy, D., & Carpentier, A. (2005). Le rouget barbet de roche Mullus surmuletus (L. 1758) en Manche orientale et mer du Nord. RAPPORT DE CONTRAT IFREMER/CRPMEM NORD-PAS-DE-CALAIS. https://archimer.ifremer.fr/doc/00000/2351/

Manel, S., & Holderegger, R. (2013). Ten years of landscape genetics. *Trends in Ecology & Evolution*, *28*(10), 614–621. https://doi.org/10.1016/j.tree.2013.05.012

Marques, V., Guérin, P.-É., Rocle, M., Valentini, A., Manel, S., Mouillot, D., & Dejean, T. (2020). Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. *Ecography*, *43*(12), 1779–1790. https://doi.org/10.1111/ecog.05049

Matić-Skoko, S., Šegvić-Bubić, T., Mandić, I., Izquierdo-Gomez, D., Arneri, E., Carbonara, P., Grati, F., Ikica, Z., Kolitari, J., Milone, N., Sartor, P., Scarcella, G., Tokaç, A., & Tzanatos, E. (2018). Evidence of subtle genetic structure in the sympatric species *Mullus barbatus* and *Mullus surmuletus* (Linnaeus, 1758) in the Mediterranean Sea. *Scientific Reports*, *8*(1), 676. https://doi.org/10.1038/s41598-017-18503-7

Milhau, T., Valentini, A., Poulet, N., Roset, N., Jean, P., Gaboriaud, C., & Dejean, T. (2021). Seasonal dynamics of riverine fish communities using eDNA. *Journal of Fish Biology*, *21*(98), 387–398. https://doi.org/10.1111/jfb.14190

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., & Doi, H. (2017). Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in common carp Cyprinus carpio. *Molecular Ecology Resources*, *17*(2), 324–333. https://doi.org/10.1111/1755-0998.12586

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & 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. https://doi.org/10.1098/rsos.150088

Najjar Lashgari, S., Rezvni Gilkolaei, S., Kamali, A., & Soltani, M. (2017). Study of genetic diversity of wild Caspian trout Salmo trutta caspius in the Sardabrud and Astara Rivers, using D-Loop region sequencing. *Iranian Journal of Fisheries Sciences*, *16*, 352–365.

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–19. https://doi.org/10.1093/nar/gkv1290

Olson, Z. H., Briggler, J. T., & Williams, R. N. (2012). An eDNA approach to detect eastern hellbenders (*Cryptobranchus a. Alleganiensis*) using samples of water. *Wildlife Research*, *39*(7), 629–636. https://doi.org/10.1071/WR12114

Parsons, K. M., Everett, M., Dahlheim, M., & Park, L. (2018). Water, water everywhere: Environmental DNA can unlock population structure in elusive marine species. *Royal Society Open Science*, *5*(8), 180537. https://doi.org/10.1098/rsos.180537

Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J.-B., Borrero-Pérez, G. H., Cheutin, M.-C., Dejean, T., Corredor, J. D. G., Acosta-Chaparro, A., Hocdé, R., Eme, D., Maire, E., Spescha, M., Valentini, A., Manel, S., Mouillot, D., Albouy, C., & Pellissier, L. (2021). Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environmental DNA*, *3*(1), 142–156. https://doi.org/10.1002/edn3.140

Potapov, V., & Ong, J. L. (2017). Examining sources of error in PCR by single-molecule sequencing. *PLoS One*, *12*(1), e0169774. https://doi.org/10.1371/journal.pone.0169774

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, *15*, e0227434. https://doi.org/10.1371/journal.pone.0227434

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, *39*(21), e145. https://doi.org/10.1093/nar/gkr732

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, *6*(12), e27310. https://doi.org/10.1371/journal.pone.0027310

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, *13*(2), 245–262. https://doi.org/10.1111/eva.12882

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., Pedersen, M. W., Jaidah, M. A., Orlando, L., Willerslev, E., Møller, P. R., & Thomsen, P. F. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution*, *1*, 1–5. https://doi.org/10.1038/s41559-016-0004

Sigsgaard, E. E., Nielsen, I. B., Carl, H., Krag, M. A., Knudsen, S. W., Xing, Y., Holm-Hansen, T. H., Møller, P. R., & Thomsen, P. F. (2017). Seawater environmental DNA reflects seasonality of a coastal fish community. *Marine Biology*, *164*(6), 128. https://doi.org/10.1007/s00227-017-3147-4

Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., Davenport, M. P., & Mak, J. (2010). Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, *469*(1), 45–51. https://doi.org/10.1016/j.gene.2010.08.009

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, *7*, 12240. https://doi.org/10.1038/s41598-017-12501-5

Stoeckle, M. Y., Soboleva, L., & Charlop-Powers, Z. (2017). Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. *PLoS One*, *12*(4), e0175186. https://doi.org/10.1371/journal.pone.0175186

Székely, D., Corfixen, N. L., Mørch, L. L., Knudsen, S. W., McCarthy, M. L., Teilmann, J., Heide-Jørgensen, M. P., & Olsen, M. T. (2021). Environmental DNA captures the genetic diversity of bowhead whales (Balaena mysticetus) in West Greenland. *Environmental DNA*, *3*(1), 248–260. https://doi.org/10.1002/edn3.176

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.

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

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

Tsuji, S., Maruyama, A., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020a). Environmental DNA analysis shows high potential as a tool for estimating intraspecific genetic diversity in a wild fish population. *Molecular Ecology Resources*, *20*(5), 1248–1258. https://doi.org/10.1111/1755-0998.13165

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020b). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*, *2*(1), 42–52. https://doi.org/10.1002/edn3.44

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: Separating the wheat from the chaff. *Ecological Applications*, *30*(2), e02036. https://doi.org/10.1002/eap.2036

Uchii, K., Doi, H., & Minamoto, T. (2016). A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Molecular Ecology Resources*, *16*(2), 415–422. https://doi.org/10.1111/1755-0998.12460

Uchii, K., Doi, H., Yamanaka, H., & Minamoto, T. (2017). Distinct seasonal migration patterns of Japanese native and non-native genotypes of common carp estimated by environmental DNA. *Ecology and Evolution*, *7*(20), 8515–8522. https://doi.org/10.1002/ece3.3346

Weltz, K., Lyle, J. M., Ovenden, J., Morgan, J. A. T., Moreno, D. A., & Semmens, J. M. (2017). Application of environmental DNA to detect an endangered marine skate species in the wild. *PLoS One*, *12*, e0178124. https://doi.org/10.1371/journal.pone.0178124

Whitehead, P. J. P., Bauchot, M.-L., Hureau, J. C., Nielsen, J., & Tortonese, E. (1986). Mullidae. In *Fishes of the North-eastern Atlantic and the Mediterranean* (Vol. *2*., pp. 877–882). https://bibliotheques.mnhn.fr/medias/doc/EXPLOITATION/HORIZON/65206/fishes-of-the-north-eastern-atlantic-and-the-mediterranean-poissons-de-l-atlantique-du-nord-est-et-d

Xie, Z.-Y., Du, J.-Z., Chen, X.-Q., Wang, Y.-X., & Murray, B. W. (2006). The significance of mitochondria control region (D-Loop) in intraspecific genetic differentiation of fish. *Hereditas*, *28*(3), 362–368.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Macé, B., Hocdé, R., Marques, V., Guerin, P.-E., Valentini, A., Arnal, V., Pellissier, L., & Manel, S. (2022). Evaluating bioinformatics pipelines for population-level inference using environmental DNA. *Environmental DNA*, 4, 674–686. https://doi.org/10.1002/edn3.269