Food Control 79 (2017) 297–308

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Novel nuclear barcode regions for the identification of flatfish species

Valentina Paracchini [a], Mauro Petrillo [a], Antoon Lievens [a], Antonio Puertas Gallardo [a], Jann Thorsten Martinsohn [a], Johann Hofherr [a], Alain Maquet [b], Ana Paula Barbosa Silva [b], Dafni Maria Kagkli [a], Maddalena Querci [a], Alex Patak [a], Alexandre Angers-Loustau [a, *]

[a] European Commission, Joint Research Centre (JRC), via E. Fermi 2749, 21027 Ispra, Italy
[b] European Commission, Joint Research Centre (JRC), Retieseweg 111, 2440 Geel, Belgium

ARTICLE INFO

Article history:
Received 7 November 2016
Received in revised form
5 April 2017
Accepted 6 April 2017
Available online 7 April 2017

Keywords:
Bioinformatics
DNA barcoding
Next-generation sequencing
Seafood identification

ABSTRACT

The development of an efficient seafood traceability framework is crucial for the management of sustainable fisheries and the monitoring of potential substitution fraud across the food chain. Recent studies have shown the potential of DNA barcoding methods in this framework, with most of the efforts focusing on using mitochondrial targets such as the *cytochrome oxidase 1* and *cytochrome b* genes. In this article, we show the identification of novel targets in the nuclear genome, and their associated primers, to be used for the efficient identification of flatfishes of the *Pleuronectidae* family. In addition, different *in silico* methods are described to generate a dataset of barcode reference sequences from the ever-growing wealth of publicly available sequence information, replacing, where possible, labour-intensive laboratory work. The short amplicon lengths render the analysis of these new barcode target regions ideally suited to next-generation sequencing techniques, allowing characterisation of multiple fish species in mixed and processed samples. Their location in the nucleus also improves currently used methods by allowing the identification of hybrid individuals.

1. Introduction

The ongoing global population growth, tied to a general increase of income and urbanization, has led to a steady increase of fish consumption, resulting in a strong expansion of fish production and the development of complex distribution patterns. This is paralleled by an extensive, and often unsustainable, exploitation of fish populations and an ever increasing aquaculture production (FAO, 2014).

In this context, the development of an efficient seafood traceability framework is crucial for the correct management of fisheries and the protection of fish biodiversity. For example, a recent study classified 7.5% of European marine fish species as having an elevated risk of extinction in European waters, while the insufficient availability of data prevented the full assessment for more than 20% of the studied species (Fricke, 2015). Species identification is also needed to ensure a correct labelling through the whole food distribution chain, in order to protect the consumers' right to choose what they eat and to protect them against fraud such as the substitution of some species for cheaper, hard to distinguish alternatives.

The importance of seafood traceability is reflected in the current European Union legal framework. In addition to the established general rules on the provision of food information to consumers (FIC) (EU, 2011) that brought together the existing EU rules on general food and nutrition labelling into one piece of legislation, recent amendments were made to the Common Organisation of the Markets (CMO) in fishery and aquaculture products (EU, 2013). For unprocessed and some processed fish products, these requirements mandate that both the commercial and scientific names are displayed, matching those on the official lists drawn up and published by each EU country (D'Amico, Armani, Gianfaldoni, & Guidi, 2016).

Fish species identification based on external morphological features is, at best, only applicable to whole (and ideally adult) specimens and sometimes is hampered by the presence of sibling species (Knowlton, 1993). Due to its ubiquity and relatively high resilience to processing, DNA has established itself as the target of choice for the analysis of unknown fish samples (Rasmussen & Morrissey, 2009; Teletchea, 2009). One of the associated techniques is DNA barcoding, which involves the amplification of specific DNA regions by the polymerase chain reaction (PCR), followed

* Corresponding author.
E-mail address: Alexandre.Angers@ec.europa.eu (A. Angers-Loustau).

http://dx.doi.org/10.1016/j.foodcont.2017.04.009

by sequencing of the produced amplicon (Hebert, Ratnasingham, & Waard, 2003). Since the species is identified at the sequencing step, this technique has the advantage (over, for example, real-time PCR taxon-specific methods) that it does not require *a priori* knowledge of the potential species in the test sample. On the other hand, this approach depends on two main conditions: the selection of a suitable barcode target and the identification of associated primers, as well as knowledge about the reference barcode sequence for each of the species of interest.

For animals, the mitochondrial *cytochrome oxidase 1* (*COI*) gene has become the *de facto* target of choice, with the development of a whole community dedicated to generate and compile the reference sequences for known species (Stockle & Hebert, 2008). For fish, the FISH-BOL campaign (Ward, Hanner, & Hebert, 2009) has already compiled the sequence of a 648 bp region of the *COI* gene for more than 100 000 specimens representing more than 10 000 species. With these resources, COI DNA barcoding has been used to trace fish species in Egyptian aquafeed formulations (Galal-Khallaf, Osman, Carleos, Garcia-Vazquez, & Borrell, 2016), differentiate between members of the *Sparidae* species (Armani, Guardone et al., 2015; Armani, Tinacci et al., 2015), and verify the labels on seafood products commercialized in Southern Brazil marketplaces (Carvalho, Palhares, Drummond, & Frigo, 2015), to name a few recent examples.

Recent studies aiming to evaluate the extent of fish product mislabelling, due to either fraud or negligence, have shown that the issue remains a serious concern, both in Europe (Armani et al., 2013; Di Pinto et al., 2013; Helyar et al., 2014; Vandamme et al., 2016) and the rest of the world (Cunha et al., 2015; Khaksar et al., 2015; Yan et al., 2016). In 2015, a control plan coordinated at the EU level was launched to assess the prevalence of mislabelling of white fish products at all stages of the food chain. The results report a high number of non-compliances in products declared as Atlantic cod (*Gadus morhua*), in part due to the fact that it was one of the most sampled species. For this species, the observed mislabelling rate was 4%, which is in line with the values obtained in another recent survey (Mariani et al., 2015), although higher rates have also been reported (Filonzi, Chiesa, Vaghi, & Nonnis Marzano, 2010; Herrero, Madriñán, Vieites, & Espiñeira, 2010). When adjusting for the number of samples tested, the flatfishes were among the highest in terms of mislabelling frequencies, including the common sole (*Solea solea*, 24%), the yellowfin sole (*Limanda aspera*, 15%) and the halibut (*Hippoglossus hippoglossus*, 8%). Previous studies in Germany and Spain suggested substitution rates for sole as high as 30%–50% (Herrero, Lago, Vieites, & Espiñeira, 2012; Kappel & Schröder, 2015).

Although the use of DNA barcoding targeting mitochondrial regions such as the *COI* or *cytochrome b* genes for species identification is well established and documented, some limitations have been identified that justify the development of additional targets. Closely related species are sometimes difficult to differentiate with mitochondrial sequences; in these cases, the efficiency of barcoding can be improved by targeting additional genomic positions. For this purpose, the nuclear rhodopsin gene has been used (Sevilla et al., 2007). Still, the length of the sequenced fragments required (usually >600 bp) is above that of the average DNA fragment lengths resulting from certain processing techniques (Armani et al., 2013; Chapela et al., 2007), which renders the amplification problematic. Moreover, those fragment lengths tend to be incompatible with the current Next Generation Sequencing technologies which rely on shorter fragments. This difficulty in assessing highly processed food products consisting of a mixture of species in unknown quantities motivated the development of strategies producing shorter amplicons (Armani, Guardone et al., 2015; Armani, Tinacci et al., 2015; Muñoz-Colmenero, Martínez, Roca, & Garcia-Vazquez, 2017; Shokralla, Hellberg, Handy, King, & Hajibabaei, 2015).

In order to address this need for powerful and efficient DNA-based fish species identification methods with a wider application scope, we describe in this article the identification of novel nuclear barcoding targets, tailored to the identification of flatfish group members. We started by using an automated screening approach to identify short candidate barcode regions within the nuclear genome. We then tested samples of various flatfish species with a selected set of the generated primer pairs, demonstrating the advantages of these novel targets over the mitochondrial ones, such as the capability to identify hybrid individuals, as well as multiple fish species in complex mixtures.

2. Materials and methods

2.1. Primers design

A set of scripts was used to automatically screen a set of candidate ("seed") barcode regions, and to then produce primer pairs on regions matching the requirements of a DNA barcode. This was achieved through the following steps:

2.1.1. Seeds selection

A first strategy to identify seed sequences involved comparing two fish genome sequences using BLAST. The two genomes chosen were *G. morhua* and *T. rubripes* (both from Ensembl, release 78). These genomes were chosen due to the relative evolutionary distance between the two species. The *T. rubripes* genome sequences, in stretches of 100.000 bases, were BLASTed against the *G. morhua* genome. Positive hits with at least 90% identity over a region of at least 200 base pairs were extracted. The final set of seeds were then numbered (1–867) and saved in a FASTA file.

The second source of seed sequences were UCEs identified in humans and available at the UCNE, the Ultra-Conserved Non-coding Element base web portal (ccg.vital-it.ch/UCNEbase). The latest file for humans UCEs (hg19_UCNEs.fasta) was downloaded and potential duplicates eliminated. The resulting sequences were saved in a FASTA file, numbered from 868 to 5052.

The last set of seeds was obtained from the publication by Faircloth et al. (2013). The fish contigs FASTA files from the different species were pooled into a single file and duplicates eliminated. The resulting sequences were saved in a FASTA file, numbered from 5053 to 11836.

To identify potential barcode primers, these seeds were then expanded and their sequences obtained in a small subset of genomes as previously described for plants (Angers-Loustau et al., 2016). The extension size of 2 kb from these analyses was used for the current work. The two genomes selected for the extension were *Danio rerio* and *Oryzias latipes*, and the screen was done using all the 11 fish genomes from Ensembl (release 78), i.e. *Astyanax mexicanus* (Mexican tetra), *Danio rerio* (Zebrafish), *Gadus morhua* (Atlantic cod), *Gasterosteus aculeatus* (Stickleback), *Lepisosteus oculatus* (Spotted gar), *Oreochromis niloticus* (Nile tilapia), *Oryzias latipes* (Medaka), *Poecilia formosa* (Amazon molly), *Takifugu rubripes* (Japanese pufferfish), *Tetraodon nigroviridis* (Green puffer fish) and *Xiphophorus maculatus* (Southern platyfish).

2.1.2. Primer generation

For the seeds giving at least one "valid" hit in each of the eleven genomes, the corresponding sequences were extracted and aligned using MAFFT (Katoh, Misawa, Kuma, & Miyata, 2002). From the aligned sequences, custom scripts were used to identify potential primer pairs (primer length: 20 bp), matching the following criteria: 1) each position is conserved in at least 8 of the 11 species

and 2) the primers produce a predicted amplicon that is between 100 and 400 bp long. These primers were then used in simulated PCR (using the e-PCR tool from NCBI (Schuler, 1997)) against the original 11 genomes in order to determine the expected number of amplicons produced in each species. Only primers producing at least one amplicons in more than 8 of the 11 fish species, while never producing more than 5 amplicons in each species, were kept. Of these, 10 were selected for further studies (see Table 1) and purchased (Eurogentec S.A., Belgium).

2.1.3. Final barcode selection

The capability of the different barcodes to differentiate the flatfish species of interest was predicted by calculating *in silico* the number of differences between species and (when possible) within the same species. The primers were pooled in groups that were predicted to differentiate the two Solea species (2581-aaa, 153-aaa, 284-aac, 355-aab, 357-aaa), the two Scophthalmus (284-aac, 355-aab, 357-aaa), and the other species (357-aaa, 153-aaa, 10410-aad). The other primers were not predicted to be adequate for the identification of the species in the flatfish group with this analysis. The complete picture can be found in Supplementary Fig. 1. These predictions were then tested experimentally, using sets of reads generated by next-generation sequencing of different fish specimens.

For these analyses, we used the glsearch36 tool from the FASTA program package (Pearson & Lipman, 1988). The tool allows matching of a DNA sequence (a barcoding read) to a DNA sequence database (our reference sequence database generated as explained below). In the analysis of a set of NGS reads, each individual read was tested against the database to find sequences with less than 1% differences across the whole length of the read. If more than one record in the database matched this criterion, final classification was obtained by identifying the most recent common ancestor (MRCA) of the matched records. Reads for which the best hit was above the 1% difference threshold were labelled “unassigned”. This was done to minimise wrong species assignment, in particular due to the fact that the reference sequence database still contains a relatively small number of species. At the same time, allowing mismatches gave the system some resilience regarding sequencing errors, which are not uncommon in NGS experiments, and to eventual Single Nucleotide Polymorphisms (SNPs) between members of the same species.

2.2. Generation of a reference sequences database

Reference barcode sequences, linked to specific taxon ids, were then generated for the selected primers using the different strategies described below and summarised in Fig. 1. All references obtained were exported into a FASTA format, including the NCBI taxon ID of the origin species in the name of each sequence.

2.2.1. Whole genome sequences

All genome sequences available from the NCBI's FTP page (ftp.ncbi.nlm.nih.gov/genomes/genbank/) were downloaded and uncompressed locally (501 genomes at the time of this study, only one version per species was used). In addition, the genome sequence from *S. maximus* was downloaded from ENA_Study:PRJEB11743, linked to a recently published article (Figueras et al., 2016). Each genome was then scanned using the *e-PCR* tool to identify potential amplicons with the following parameters: n2 (maximum of 2 mismatches per primer), -g2 (maximum of 2 gaps per primer), -f3 (use 3 discontiguous words) and -t3 (tabular output format, to

Table 1
Candidate barcode targets and primers selected for the study. The table shows the number of predicted amplicons in the eleven Ensembl fish genomes. The gene targeted by each barcode was determined by the annotation of the region in the *Danio rerio* genome.

Primer	Primers sequence	Size (bp)	Number of predicted amplicons per genome											Annotation
			A. mexicanus	*D. rerio*	*G. morhua*	*G. aculeatus*	*L. oculatus*	*O. niloticus*	*O. latipes*	*P. formosa*	*T. rubripes*	*T. nigroviridis*	*X. maculatus*	
19-aaa	Forward: AGCGGTCAATGATGTTCTGC Reverse: GCCCACCGCATCCTGGTGTC	157	1	1	2	1	0	1	2	1	2	2	1	midline 2
153-aaa	Forward: GACATGAACGCCTACCTGCT Reverse: GGGGCCATGTAGAAGTCCGT	305	1	1	1	1	1	1	1	1	1	1	1	PDLIM1 interacting kinase 1 like
284-aac	Forward: GGAAGGAGATGAACGAAATT Reverse: GCCCACATTGAAACAGATCC	191	1	1	2	2	1	2	1	2	2	2	2	WD repeat domain 20b /20a
292-aaa	Forward: CCCATCTGCAGACACTTGAG Reverse: CCACAGGCCTGCACTATGGC	172	1	1	1	1	1	1	1	1	1	1	1	nuclear receptor subfamily 6, group A, member 1a
355-aab	Forward: GATCTGGCACCACACCTTCT Reverse: GAAGGTCTCAAACATGATCT	130	2	2	4	3	3	3	3	2	4	4	2	actin, alpha 1a/1b, skeletal muscle
357-aaa	Forward: CAGGTGGGTGAACTCCATGC Reverse: TTCTCGCAGGCCATCACCAC	279	1	1	1	1	1	1	1	1	1	1	1	ring finger protein 41
1184-aaa	Forward: AAGTTAATAGCCTTGTCTCT Reverse: AAACTAGATTCCTAAGCAGT	217	1	1	1	1	1	1	1	1	1	2	1	zinc finger protein, FOG family member 2a
2034-aad	Forward: CCTTTGTTGTAGGTGCAGTT Reverse: GAGCCTTACAAAAATTCCAG	146	1	1	1	1	0	1	1	1	1	1	1	fibronectin leucine rich transmembrane 3
2581-aaa	Forward: TTTGTGGATTTTGTGGACAT Reverse: TCCACTTACCCGTCTAGTCT	266	2	1	2	2	1	2	2	2	2	2	1	spen family transcriptional repressor
10410-aad	Forward: GGCGGACACCTTGATGTACG Reverse: GCTGCAAGTTCACGTTGTGC	320	0	1	0	1	1	1	1	1	1	1	1	homeobox C13a / C13b

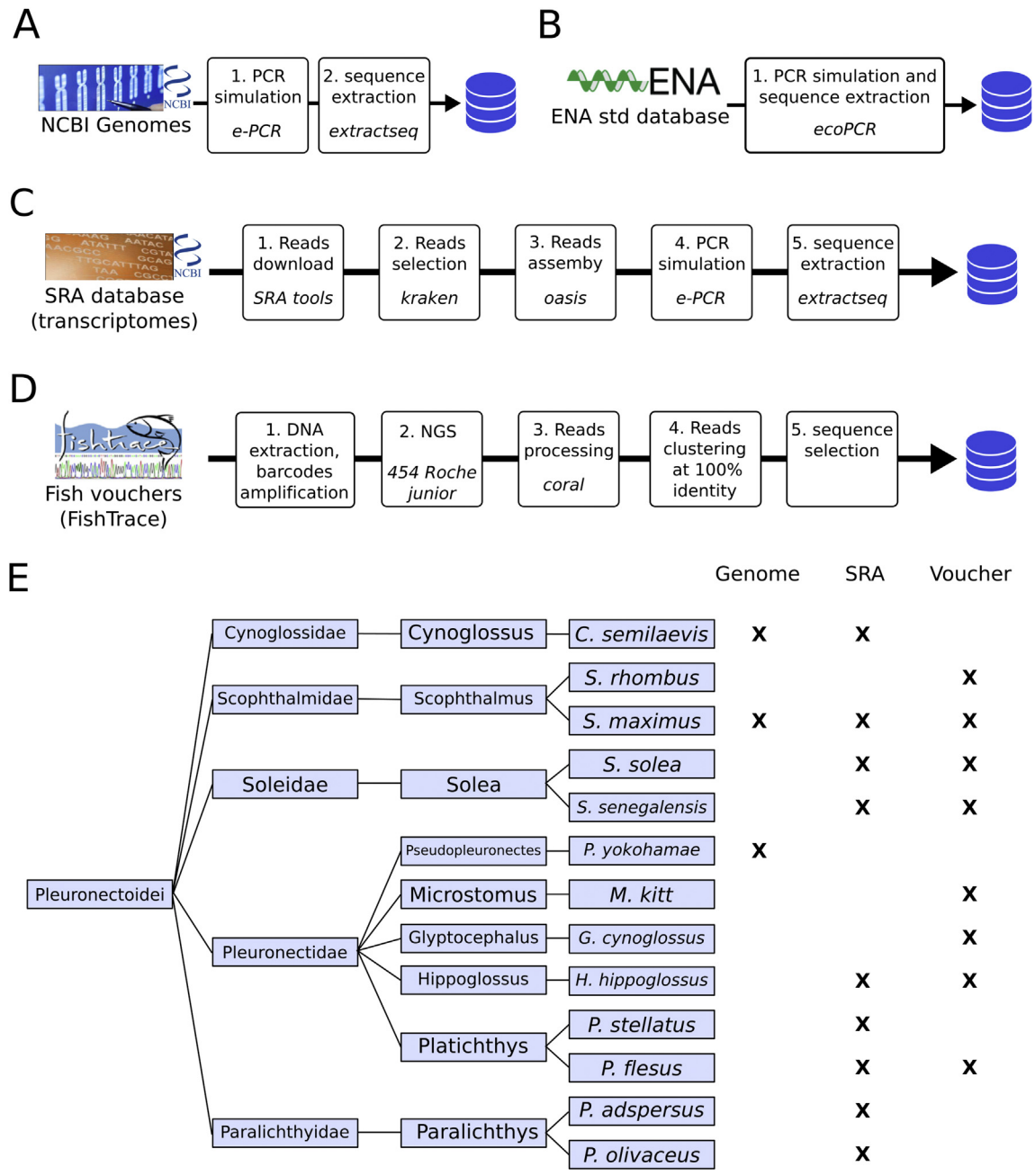

Fig. 1. Generation of reference barcode sequences. For each of the primer pairs, reference barcode regions for different species were generated using a mix of *in silico* and experimental procedures, that included: (A) Scanning the NCBI genome sequences by electronic PCR, (B) Scanning the std section of the ENA database with the ecoPCR tool, (C) Assembling NGS reads from NCBI's Sequence Read Archive and (D) Sequencing the amplicons produced in fish voucher samples. (E) List of species in the Pleuronectoidei family for which at least one reference sequence was obtained, showing the strategies available for each.

simplify parsing in the next step). Based on the produced output, hits with total gaps and mismatches lower than 5 were extracted using the *extractseq* tool of the EMBOSS Open Software Suite version 6.5.7 (Rice, Longden, & Bleasby, 2000) and reverse-complemented when needed.

2.2.2. The European Nucleotide Archive

The whole *standard* section of the ENA database available on the ENA's FTP page (ftp.ebi.ac.uk/pub/databases/ena/sequence/release/std/, release 126) was downloaded locally and uncompressed into a single file. This file was formatted for ecoPCR analyses with the supplied *ecoPCRFormat.py* script (Ficetola et al., 2010). The resulting database was then scanned with the *ecoPCR* script, with the options -e2 (2 max errors allowed by oligonucleotide) and length limits (-l, -L) plus or minus 50bp from the average amplicon length extracted from the genomes.

2.2.3. Sequence Read Archive

Archived reads from fish transcriptomics experiments were identified by the following search in the Sequence Read Archive from NCBI: “transcriptomic AND “bony fishes”[orgn:__txid7898]”. The full table was downloaded from the Run Selector interface, and the file was processed to remove the species for which a genome sequence was available (see 2.2.1), experiments with the platforms

"ABI_SOLID" or "PACBIO_SMRT", and experiments performed on hybrid specimen. When more than one run was available from the same species, all were kept, up to a maximum of 100 Mbases per species. For each of the selected runs, the read sequences were downloaded in the FASTA format using the provided SRA Toolkit. The fraction of "interesting" reads was enriched using the kraken algorithm (Wood & Salzberg, 2014). Briefly, the sequences extracted from the Ensembl genomes in the "genome screen" section above were formatted into a kraken database using the supplied scripts. The downloaded reads were then screened with the kraken script, using the "–only-classified-output" flag to identify the IDs of reads sharing 31-mers with the barcode regions. Those reads were extracted, and assembled using the Velvet algorithm package (Zerbino, 2010). The *velveth* script was launched with the following options and parameters: 31 (hash length), -fasta and either -short (for a SINGLE library layout) or -shortPaired —separate (for a PAIRED library layout). This was followed by the *velvetg* script, with the options -read_trkg yes -exp_cov auto -cov_cutoff 2, adding -ins_length for PAIRED layouts with the value found in the Run Selector table. Finally, the produced contigs file was screened using *e-PCR*, and the sequences of potential hits extracted, as described for the genomes above.

2.2.4. Voucher samples

Biological samples for flatfish species available in the biological reference collection of the FishTrace project (https://fishtrace.jrc.ec.europa.eu/) were obtained from the Swedish Museum of Natural History. These included: *Solea solea* (NRM 52889), *Scophthalmus maximus* (NRM 52878), *Microstomus kitt* (NRM 49458), *Glyptocephalus cynoglossus* (NRM 49432), *Hippoglossus hippoglossus* (NRM 53139) and *Platychthys flesus* (NRM 49641). The original *Scophthalmus rhombus* voucher (NRM 52885) turned out to be a hybrid (see Section 3.5 for details) and was replaced with NRM 52888. A sample of *Solea senegalensis* available *in house* was also included. The exact species of each sample was confirmed by COI barcoding as described in Section 2.5. The voucher samples were sequenced as described below. The original reads files were analysed by *coral*, an error correction algorithm (Salmela & Schröder, 2011). Reads were then clustered at 100% identity and the sequence(s) of the cluster(s) with the highest number of reads was/were entered in the barcode reference sequence database.

2.3. Barcode within- and between-species variability

All calculations were done using R version 3.3.1. Per primer pair, the amplicon sequences were inspected for their origin (species): (I) if multiple sequences were present for a single species, the sequences were aligned (using the multiple sequence alignment available through the package 'msa'), the Levenshtein distance across all pairwise alignments was calculated (using the 'StringDist' function from the package 'Biostrings'), and the consensus sequence was generated (using the 'consensus' function of the package 'seqinr' with 'method = majority'). The latter was used in all downstream operations (i.e. only one amplicon sequence per species). In a second step, (II) a multiple alignment across all species was generated and the Levenshtein distance across all pairwise alignments was calculated, yielding a distance matrix. Subsequently, the diagonal of this distance matrix was populated by inserting the maximal within species Levenshtein distance (as found during the first step).

2.4. Other samples

Seven whole fish specimens were purchased from local markets and supermarkets. The tested specimens were sold as common sole (*Solea solea*), Senegalese sole (*Solea senegalensis*), lemon sole (*Microstomus kitt*), European flounder (*Platichthys flesus*), halibut (*Hippoglossus hippoglossus*), turbot (*Scophthalmus maximus*) and brill (*Scophthalmus rhombus*). Another fish sample was purchased from local market as fillet and sold as salmon (*Salmo salar*). The zebrafish sample was obtained from Dr Kabashi from the Institut du Cerveau et de la Moëlle Epinière, Paris, France. The *S. solea*/*S. senegalensis* mixtures were made cooking individually the two specimens. Samples were then prepared by adding increasing amounts of *S. senegalensis* flesh to *S. solea*, the resulting mixes then mashed with a pylon to produce a uniform paste. The two species were mixed at a flesh weight: weight ratio of 100:0, 99:1, 95:5, 90:10, 50:50 and 0:100. A sample was then taken for each mixture for DNA extraction, barcode amplification and NGS. Other 40 samples each of *Solea solea* and *Scophthalmus maximus* were obtained from a local cafeteria, and flesh fragments of approximately equal weight were cut, combined and homogenised using a pylon, as for the previous mixtures.

2.5. Mitochondrial targets barcoding

The species of voucher and purchased specimen were confirmed *in house* using the traditional COI barcoding method, using published primers and protocols (Ivanova, Zemlak, Hanner, & Hebert, 2007; Ward, Zemlak, Innes, Last, & Hebert, 2005). The COI region was amplified using the following primers in a multiplex PCR: VF2_t1 TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC, FishF2_t1 TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC, FishR2_t1 CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA and FR1d_t1 CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA. The amplicon was sequenced (forward and reverse) with the following primers: M13F (-21) TGTAAAACGACGGCCAGT and M13R (-27) CAGGAAACAGCTATGAC. The identification of species was performed via the Boldsystem portal (http://www.boldsystems.org/).

In addition, all voucher samples obtained from FishTrace had previously been characterised using cytochrome B and nuclear rhodopsin sequences, as part of the procedures establishing the reference collections (Sevilla et al., 2007)

2.6. Metabarcoding sequencing and analysis

For both pure specimens and homogenised mixtures, total DNA was extracted from 2 to 200 mg tissue sample (DNeasy® Blood & Tissue Kit, Qiagen). DNA samples were diluted to a concentration of 10 ng/μl unless lower amounts were obtained. Individual PCR amplifications were performed in 50 μl using 2.5 U/reaction of AmpliTaq Gold DNA Polymerase, 1x Buffer II, 2.5 mM of MgCl2 (Applied Biosystems), 200 μM dNTPs, 200 nM of each primer. DNA samples were amplified in a GeneAmp PCR System 9700 (ABI, USA), with the following cycling parameters, according to the protocol of the AmpliTaq Gold PCR system: initial denaturation at 95 °C for 10 min, followed by 45 cycles of denaturation at 95 °C for 30 s, annealing at 55 °C or 60 °C for 30 s, extension at 72 °C for 30 s, and final extension at 72 °C for 7 min. PCR products were analysed on an agarose gel electrophoresis to verify and confirm the expected size.

Sequencing was performed on a GS Junior System (GS Junior System, 454 Life Sciences, Roche Applied Sciences, Basel, Switzerland). Amplicon libraries were prepared using fusion primers for bidirectional sequencing as described in the Amplicon library preparation manual (Roche Applied Sciences, Basel, Switzerland). Multiplex Identifiers (MIDs) were added in order to allow inclusion of more than one sample in the same experiment. All ten primers were ordered including MID1 (ACGAGTGCGT). For

selected primers (357-aaa and 10410-aad), additional versions with MID2 (ACGCTCGACA) were synthesised in order to allow the simultaneous analysis of four samples (MID1-1, MID1-2, MID2-1 and MID2-2) in the same experiment.

After amplification, amplicons were purified using AMPure XP beads (Roche Applied Sciences, Basel, Switzerland), quantified fluorometrically (Quant-iT PicoGreen dsDNA Assay kit, Life Technologies, Molecular Probes, Eugene, Oregon, USA), diluted, and pooled to a final concentration of 0.5 x 10^6 molecules/μl. Libraries were checked for their quality by performing a Quality Control PCR; they were subsequently visualized using Agilent DNA 1000 Chips (Agilent Bioanalyzer, Agilent Technologies, San Diego, USA). Emulsion PCR containing between 0.6 and 0.75 copies per bead (cpd) was recovered using vacuum and the successive enrichment led to an enrichment rate between 10 and 20%; only 5% of the enriched beads were subsequently loaded on the chip and sequenced in a GS Junior System (Roche Diagnostics). All steps were in accordance with the manufacturers' instructions.

The output (FASTA format) was split using fastx_barcode_splitter from the fastx-toolkit (Gordon & Hannon, 2010) to isolate the reads from the different samples tested in the same run using the MID sequences. The primer sequences were then trimmed using *cutadapt* (Martin, 2011). Scripts were used to analyse each read against the reference files using *glsearch* from the FASTA package to identify matches between the entire length of each read and local regions of the reference barcode sequences. The number of mismatches allowed was set at 1%, i.e. 0 for barcodes less than 100bp, 1 for those between 100 and 200 bp, and 2 for those between 200 and 300 bp. In case of more than one hit, the most recent common ancestor for the identified species was determined using the API of the Open Tree of Life (Watanabe, 2013), reached at https://api.opentreeoflife.org/v3/. If no hit matched the minimum criteria the read was assigned to an "Unassigned" conclusion. A main reason for this would be errors in the sequences from the 454 pyrosequencing process. The occurrences of these remained limited (on average less than 5% of total reads). The results of all the reads for each NGS experiment were compiled and presented as doughnut charts using Excel.

3. Results and discussion

In the current study, we first identified a set of novel barcode regions in nuclear genomes, using a set of criteria compatible with analysis with NGS (Section 3.1) and generated, for these primers, sets of reference sequence data using as much currently available DNA sequence information as possible (Section 3.2). We then selected a subset of these primers based on their predicted performance in our species of interest (Section 3.3), before using them to test a set of samples (Sections 3.4–3.7).

3.1. Identification of new barcode regions in the nuclear genome

A strategy to identify novel DNA barcode regions should take into account three important barcode criteria: (1) to be present in all species of interest, and to include regions of sufficient cross-species identity to allow the design of broad specificity primers for the barcode amplification; (2) to be short enough to allow efficient amplification and sequencing - this is particularly important analysing mixtures since, unlike the Sanger method, current next-generation sequencing technologies allow only the sequencing of relatively short DNA fragments; (3) the sequence of the amplified region itself should contain sufficient differences between species to allow assignment of sequenced amplicons to specific species of origin taking into account the natural variability between members of the same species.

In order to identify regions in the fish nuclear genome that could be used to design DNA barcoding primers, we followed a stepwise approach involving the automated processing and screening of input sequences ("seeds") by a bioinformatics pipeline for the main characteristic of a barcode region: a region of high similarity across the widest range of species. In this optic, ultraconserved elements (UCEs), defined as highly conserved genome regions shared among evolutionary distant taxa (first described in (Bejerano et al., 2004)), were selected as interesting candidates. A final set of 2267 primer pairs were produced, targeting 745 of the original seed regions. Details of the seed sources, pipeline steps and selection criteria can be found in Section 2.1.

From this list of candidate primers, a subset of ten pairs were selected to be evaluated experimentally (Table 1), based on the fact that they produce a single amplicon in each of the genomes. Two primer pairs with multiple targets (284-aac and 355-aab) were also selected, to see whether these multiple copies affect the usefulness of the method.

For each primer pair, the nature of the targeted genomic region was determined using the annotations in Ensembl (in particular, the *Danio rerio* genome), showing that all of them amplify regions within genes.

3.2. Generation of a reference sequence database

To be useable in practice, a DNA barcoding species identification strategy needs the prior establishment of a database of barcode sequences linked to their species or origin. This, in itself, is a daunting task and may represent one of the main obstacles in the use of targets other than those that have been the subject of international efforts in this direction (Stockle & Hebert, 2008; Ward et al., 2009). For our new barcode candidates, we have designed different strategies in order to mine publicly available information with the aim of generating a starting reference set (Fig. 1), detailed in Section 2.2. These public resources used included the genome sequences, the Sequence Read Archive (SRA) - both from Genbank (Benson et al., 2013; Leinonen, Sugawara, & Shumway, 2010)- as well as the European Nucleotide Archive (ENA) (Leinonen, Akhtar, et al., 2010).

3.2.1. Genbank genome sequences

Genome sequences available from Genbank (Benson et al., 2013), both fish and others, were analysed *in silico* by PCR simulation in order to identify and extract predicted amplicons (Fig. 1A). The number of produced barcode sequences is reported in Table 2, showing a relative constant number for all primers except 355-aab. 355-aab targets the actin gene, which is known to be present from fungi to mammals. In addition, being a structural protein, evolutionary pressure might have prevented variations in the gene sequence, allowing better conservation of the primer binding sites.

3.2.2. The European Nucleotide Archive

We used the ecoPCR tool, designed to scan a database of sequences with a pair of primers and extract potential barcode amplicons (Ficetola et al., 2010) against the whole standard section of the ENA database (Fig. 1B). As shown in Table 2, relatively few sequences were found for most of our barcode candidates, possibly due to the fact that they target genes that have not been extensively studied in the scientific literature; once again, the striking exception is actin (355-aab).

3.2.3. The Sequence Read Archive

The SRA stores raw sequencing data generated by high throughput sequencing platforms, with the associated metadata (Kodama, Shumway, & Leinonen, 2012; Leinonen, Sugawara, et al.,

Table 2
Analysis of the generated reference sequences for each barcode primer pair. The table shows the breakdown of the number of unique sequences found by each strategy (see Fig. 1), and by groups of species.

	Number of reference sequences per source				Number of species with unique reference sequences					
Primer	Genomes	ENA	SRA	in house sequencing	Plants	Invertebrates	Fish	Tetrapods	Fungi	Bacteria
19-aaa	183	25	215	10	0	0	118	55	0	0
153-aaa	59	11	124	11	0	0	114	15	0	0
284-aac	142	30	155	14	0	0	107	24	0	0
292-aaa	71	8	71	8	0	0	76	13	0	0
355-aab	2960	10440	189	17	368	565	195	241	191	0
357-aaa	167	15	91	18	0	0	82	93	0	0
1184-aaa	251	59	0	9	0	0	40	61	0	0
2034-aad	81	13	126	17	0	0	79	18	0	0
2581-aaa	100	8	86	14	0	0	88	1	0	0
10410-aad	66	82	45	14	0	0	78	10	0	0

2010). The fact that our new barcodes target genes, thus potentially expressed as RNA, allowed us to focus on experiments that were labelled as "transcriptomics" and with a biosample corresponding to a fish species. The raw reads were scanned to enrich for those similar to our regions of interest, which were then assembled into contigs. These contigs were then subjected to PCR simulations with our primer pairs (Fig. 1C). As shown in Table 2, this strategy produced additional references for all primers except 1184-aaa. Analysis of the expected exon-intron boundaries in Ensembl showed that this primer is the only one found in a predicted intron sequence, which are spliced out and absent from the transcriptomes.

3.2.4. *In house sequencing of voucher samples*

Due to our interest in developing new barcoding solutions specifically for flatfish, such as soles, turbot, and their common substitutions, we complemented our set of reference sequences obtained *in silico* with experimental data generated by sequencing *in house* a set of fish samples. We selected the following species based on the analysis on the frequency of recently documented fraud cases, to ensure an experimental determination or confirmation of their reference sequences: common sole (*Solea solea*), turbot (*Scophthalmus maximus*), brill (*Scophthalmus rhombus*), lemon sole (*Microstomus kitt*), Witch flounder (*Glyptocephalus cynoglossus*), Atlantic halibut (*Hippoglossus hippoglossus*) and European flounder (*Platichthys flesus*). Voucher samples were obtained from the biological reference collection of the FishTrace project (https://fishtrace.jrc.ec.europa.eu/). In addition, a sample of Senegalese sole (*Solea senegalensis*) previously characterised in our laboratory was included. For all these samples, the correct species was confirmed by mitochondrial barcoding (see Supplementary Table). The new barcode regions were then amplified by PCR and sequenced using a 454 Roche Junior (Fig. 1D). At least one sequence was obtained by each of the primers for each of these samples.

3.2.5. *Analyses of the reference sequences*

The main issue faced with these bioinformatics approaches to generate reference sequences is the completeness of the coverage. Whole genome sequences (3.2.1) are published at different stages of completeness, and sequences may be missed for some species if the region in which they are found has not yet been fully reconstructed. Assembling reads from NGS experiments (3.2.3) is a complex procedure that does not guarantee that every region will be found in the generated contigs. Finally, the presence of a sequence in annotated records is dependent on how studied this gene has been in the scientific literature (3.2.2).

Table 2 summarises the species coverage of all the combined reference sequences, showing the number of species for which at least one unique reference sequence was found, grouped as Bacteria, Fungi, Plants, Invertebrates, and Vertebrates, in turn separated in Fish and others (Tetrapods). Most of the barcodes seemed to be specific to vertebrates, with various ratios between fish and other species, the extreme being 2581-aaa, with a single non-fish sequence (from the green anole *Anolis carolinensis*). As discussed earlier, 355-aab, targeting an exon in actin isoforms, is found in the highest number of species, including fungi, animals, and plants. Fig. 1E summarises the information for the flatfish species (Pleuronectoidei) for which reference sequences were found and the strategies used to retrieve them.

To minimise the computational resources required, the assembly of transcriptomics reads was not performed for species for which a genome sequence is available. One exception was

S.maximus, whose genome was published shortly after the transcriptome analyses were performed (Figueras et al., 2016). This situation allowed a direct comparison of the sequences produced. The reference sequence for the 357-aaa primer, which is in a single copy in *S. maximus*, was generated independently from its genomic sequence (ENA_Study:PRJEB11743), a set of transcriptome experiments (SRR1695148-SRR1695159, with the same sequence obtained from each study), and from *in house* sequencing of a voucher sample (FishTrace NRM52878). The three resulting sequences were aligned and found to be identical, suggesting that mining the publicly available information as shown in Sections 3.2.1–3.2.3 is an efficient way to generate reference sequence datasets for new barcoding targets.

3.3. Final barcode analyses

For each of the fish specimen, the barcode regions were amplified and sequenced, and the resulting reads were analysed in order to evaluate what percentage of all produced reads could be assigned to the correct species. Fig. 2 shows the results of testing each voucher sample with each of the barcode primers. The results demonstrate a general good agreement with the *in silico* predictions: three of the five primer pairs predicted to correctly identify the *Solea* species did so indeed; for the *Scophthalmus* species, only one of the three (357-aaa) succeeded, while one additional good primer (19-aaa) was identified. For the other species, the best primer was one of those predicted, i.e. 10410-aad.

As shown in Table 1, the primer pair 357-aaa amplifies a region in a protein called “ring finger protein 41”, also known as Nrdp1, an E3-ubiquitin ligase that was shown to be involved in cell differentiation during zebrafish development (Maddirevula, Anuppalle, Huh, Kim, & Rhee, 2011). The 10410-aad primers target a member of the homeobox C13 proteins, that were shown to be important for zebrafish caudal fin regeneration (Thummel, Ju, Sarras, & Godwin, 2007).

3.4. Samples analyses - local market and supermarket fish samples

The same strategy was then used to analyse whole fish specimens that were bought from local markets and supermarkets. The correctness of the specimen labels was first confirmed using the traditional COI barcoding method (see Supplementary Table), before performing a NGS experiment as explained in the previous section. The results are shown in Fig. 3A–G. As a general rule and based on Fig. 2, the 357-aaa primer was used for the soles, the turbot and the brill, while the others were tested with 10410-aad. In all cases, the conclusion of these analyses was consistent with the species declared for the tested sample and with the results of COI barcoding. The brill specimen was assigned ambiguously to the *Scophthalmus* genus using the 357-aaa primers (Fig. 3G); for this sample, the second primer pair 19-aaa was needed to confirm the correct species.

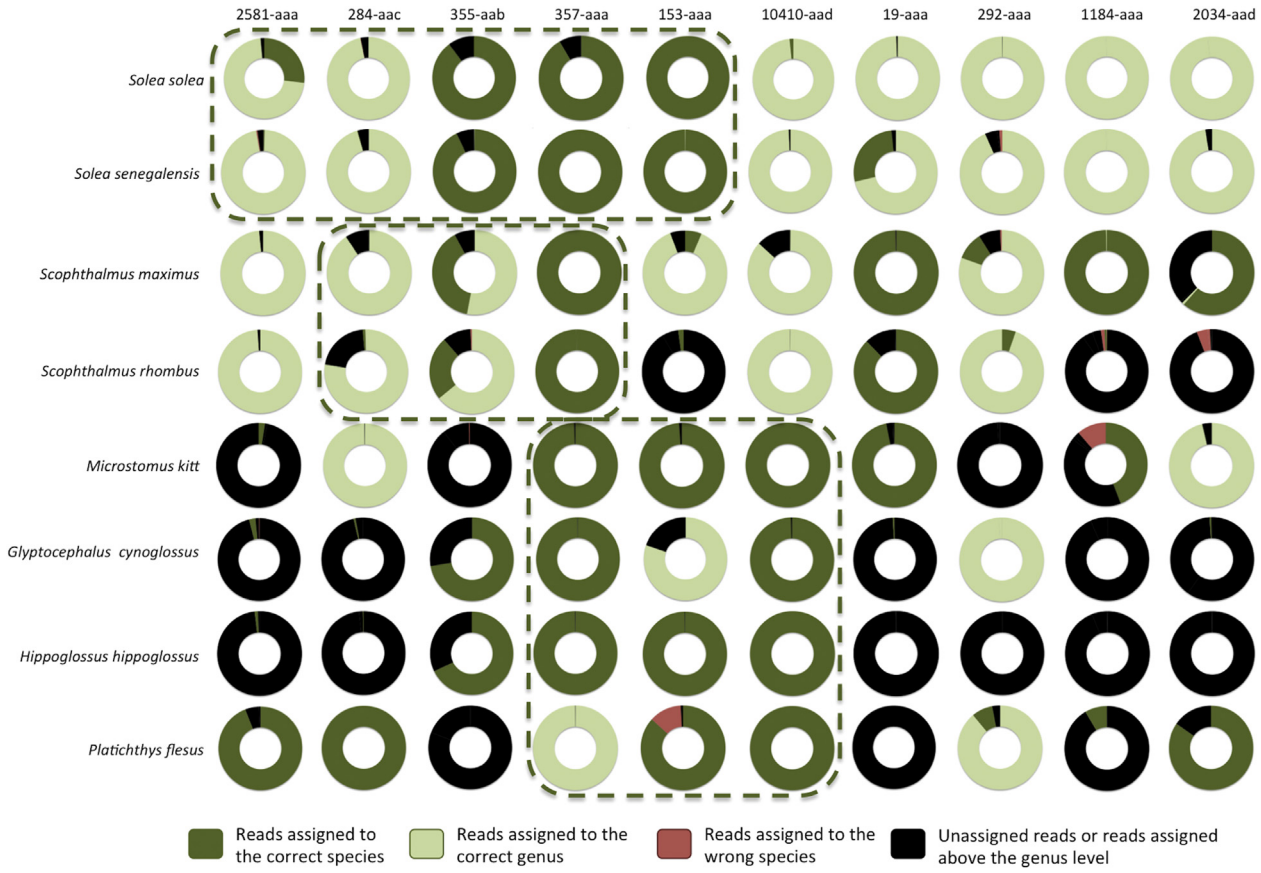

Fig. 2. Identification of the flatfish specimens with the different barcode primers. Individual reads produced by NGS after barcode amplification by each of the primer pairs were assigned to a species according to the strategy described in the text. The figure shows, for each species and primer pair, the ratio of the reads assigned to either the correct species (dark green) or the correct genus (light green). Black shows the proportion of reads for which no sufficiently similar reference sequence were found, or that were assigned above the genus level (for example, at the family level). Red shows the proportion of reads assigned to the wrong species. The dashed rectangles highlight the primers predicted *in silico* to be able to differentiate the soles, turbot/brill, or other species. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

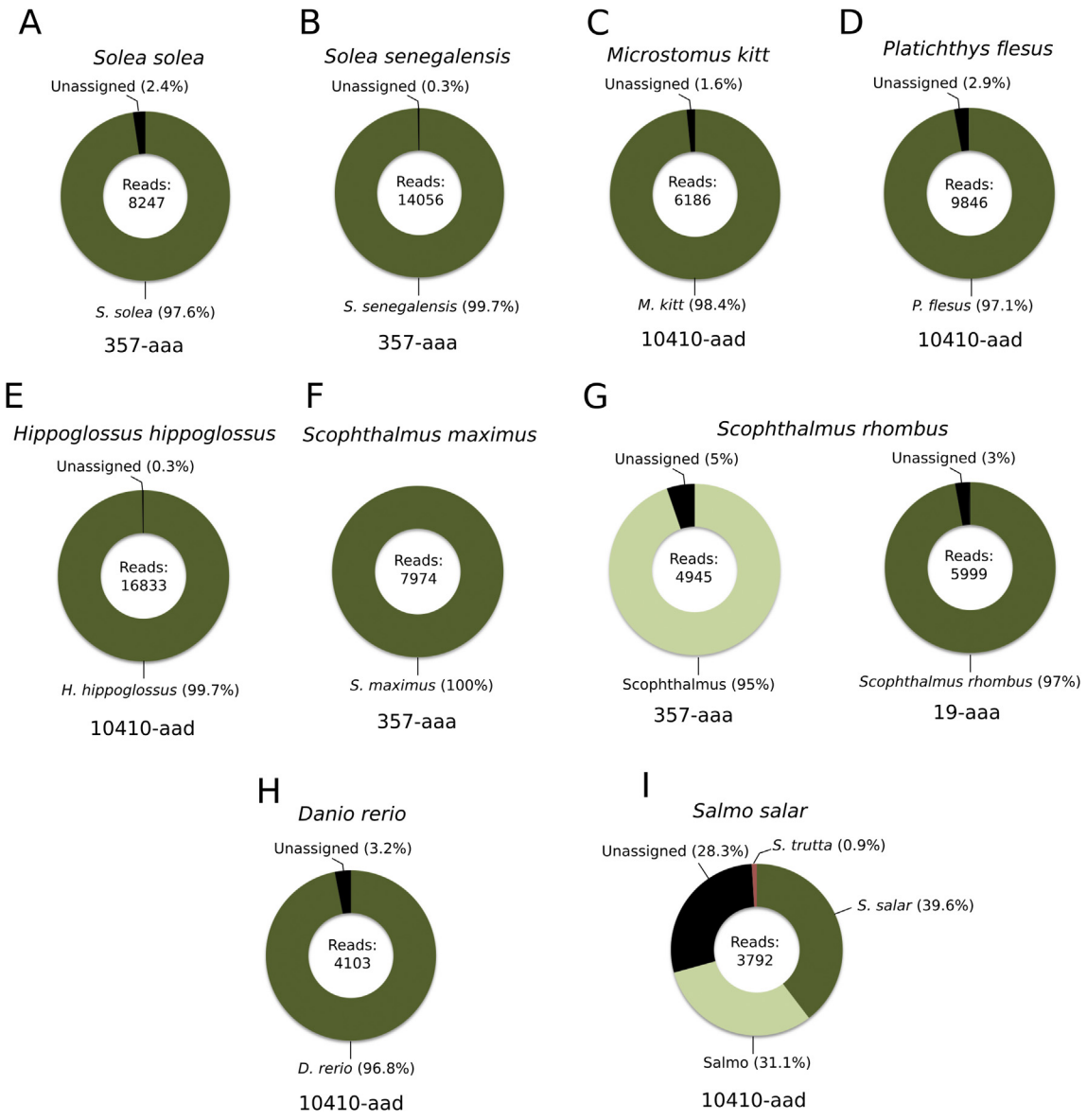

Fig. 3. Testing of fish specimens from the market with the new barcode targets. Only the primers that produced the best results from analysing the voucher samples were used, in particular 357-aaa and 14010-aad. For each sample, the number of produced reads is written in the centre of the graph that then shows the percentage of these reads assigned to different conclusions. The correct species, as determined by the market label and confirmed *in house* by COI barcoding, is shown above each chart.

3.5. Samples analyses - S. maximus x S. rhombus hybrid

During the generation of the reference barcode sequences for the *S. rhombus* voucher (Fig. 1D), the original voucher specimen used showed a systematic duplication in the number of sequences for the different primers, with half of them being, generally, identical to the reference sequences previously produced in the *S. maximus* voucher. In order to verify whether this could be due to the fact that this voucher specimen was actually a *S. maximus* x *S. rhombus* hybrid, a second voucher sample, from a different specimen, was sequenced. In this case, the sequences identical to *S. maximus* were no longer observed, so these sequences were used as the *S. rhombus* reference sequences. Fig. 4 shows the results of testing the *S. maximus* and *S. rhombus* vouchers as well as the voucher believed to be a hybrid, using both primer pairs that were shown to generate good results for both *S. maximus* and *S. rhombus* (i.e. 19-aaa and 357-aaa). The results show a consistent pattern, compatible with the hypothesis that the original voucher specimen was, in fact, a hybrid individual.

Being a voucher sample, the original specimen tested in this study was available, and closer observation did show some weak characteristics compatible with a hybrid individual (Dr Michael Noren, personal communication). Having these barcodes available would greatly simplify studies on the biology and prevalence of fish hybrids, an aspect important also for fisheries management and marine biodiversity monitoring.

3.6. Samples analyses - other species (Salmo salar and Danio rerio)

Fig. 3A–G shows the successful identification, using our novel barcoding target, of fish species for which the reference sequences were obtained by previously sequencing voucher specimens in our laboratory. Fig. 1 shows that additional strategies were used to identify and extract reference sequences from

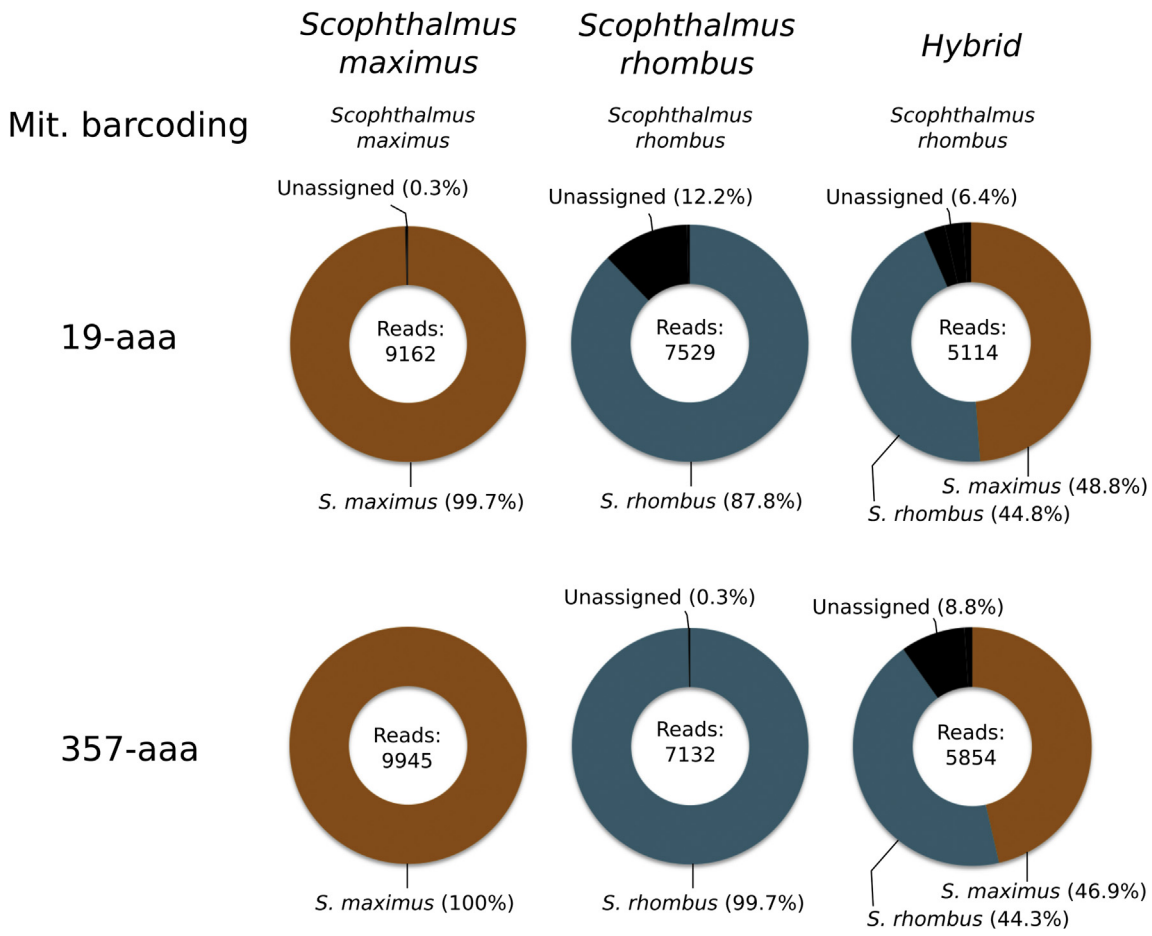

Fig. 4. Identification of a *S. maximus x S. rhombus* hybrid with the nuclear barcodes. The two primers that proved able to differentiate *S. maximus* and *S. rhombus*, i.e. 357-aaa and 19-aaa, were used to demonstrate the hybrid nature of the NRM52885 voucher (right), when compared to the corresponding *S. maximus* and *S. rhombus* vouchers (left). The results of the mitochondrial barcoding for each specimen is also shown.

publicly available database and datasets, and these form the majority of the reference sequences in our database. We then decided to test two species for which the reference was obtained from the published NCBI genome sequences: an Atlantic salmon (*Salmo salar*) fish bought from the market, and zebrafish (*Danio rerio*), a model organism obtained from a laboratory. The same experimental procedure and subsequent bioinformatics analyses was applied to these samples using the 10410-aad primers, and the results are shown in Fig. 3H–I. For both samples, the correct species could be deduced.

3.7. Samples analyses - mixtures

The analyses performed until now have been on pure samples, containing DNA extracted from single specimens. The capacity of this novel barcoding method to detect different species in complex samples was then tested. For this, mixtures of sole specimens (*S. solea* and *S. senegalensis*) at different ratio were prepared (see Section 2.4). A sample was then taken for each mixture for DNA extraction, barcode amplification and NGS. The ratio of reads being assigned to each species for each mixture, using the 357-aaa primers, is shown in Fig. 5A.

These results show that the system could correctly identify the two individual species present in the mixtures; in addition, *S. senegalensis* was detected even at the lowest amount tested, i.e. when it only represented 1% (by weight) in the flesh mixture. The amount of reads assigned to each species is proportional to their relative amount in the original mixtures. Reads not assigned to either of these two species were, at worst, less than 18% of the total, and were mostly assigned to the "*Solea*" genus. This result opens the possibility of using DNA barcoding to test processed mixed samples, and of designing efficient sampling strategies to test large amount of fish specimens at once, by pooling samples prior to DNA extraction.

This possibility was tested in experiments were 40 specimens of the same species (40 samples of *S. solea* and 40 samples of *S. maximus*), obtained from a local cafeteria, were tested in a single experiment. For this, fragments of flesh of similar weights were removed and combined in a single sample, that was then homogenised as for the previous experiment. In this mixed sample, each original specimen would represent more than 2% by weight of the final product, so above the 1% detection level shown in Fig. 5A. This strategy was applied with both 40 samples labelled as common sole (*S. solea*) and turbot (*S. maximus*), and the results shown in Fig. 5B confirm in a single experiment that the 40 specimens are of the correct, expected, species. This strategy falls in the "screening" category, as the identification of reads from a different species would necessitate further investigations to identify the affected specimen(s), if this information is needed.

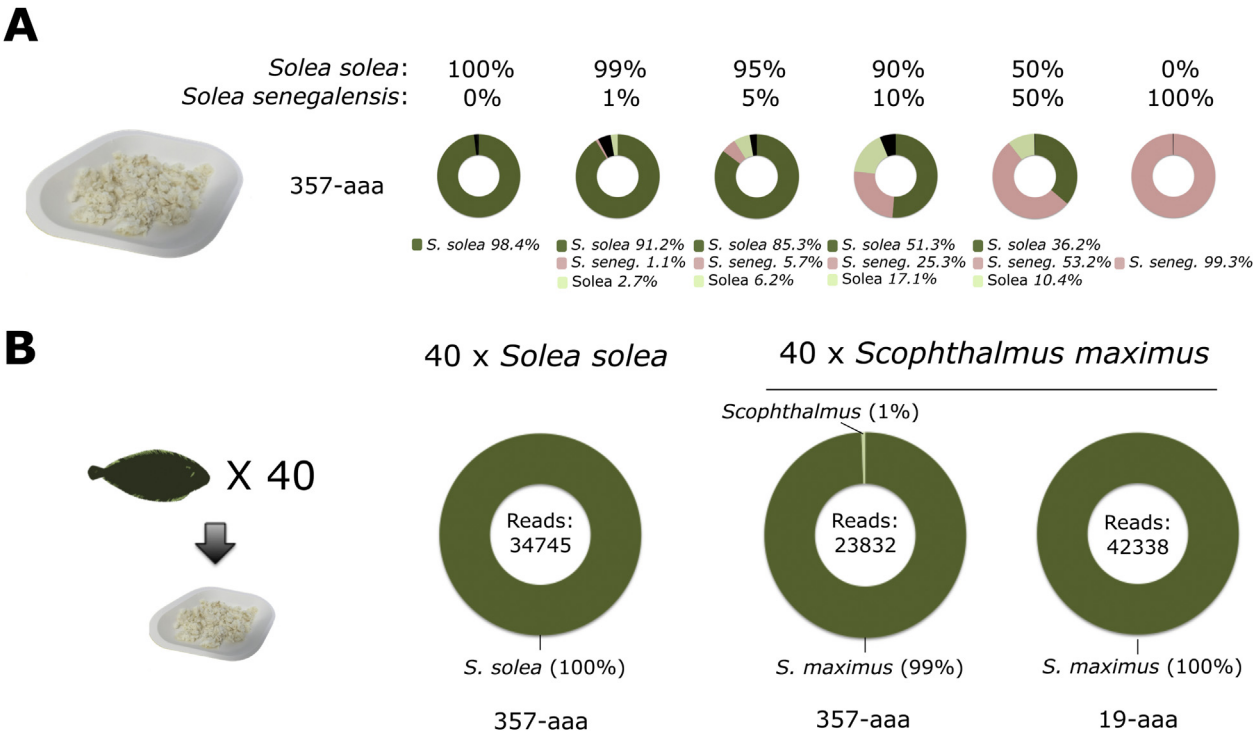

Fig. 5. Analysis of complex, processed fish samples. (A) Mixtures of *S. solea* and *S. senegalensis* were prepared by combining the flesh of the two fishes after cooking and mixed with a pylon to form a uniform paste (pictured). Different ratios were included to vary the amount of *S. senegalensis* from 0 to 100%. The samples were tested with the 357-aaa primer, and the results shown in the graphs, with the black areas representing unassigned reads. (B) Confirmation of the correct species of 40 fish specimens of the same species in a single experiment. Fragments of the flesh of 40 common sole (left) and turbot (right) specimens were pooled and homogenised, and the resulting mix analysed in a single NGS experiments per species. The samples were tested with the 357-aaa and (for *S. maximus*) 19-aaa primers. The results show that all reads were assigned to the expected species.

4. Conclusions

The work presented here describes the identification and use of novel DNA barcode regions for the identification of fish species in different samples. These barcodes could be used to positively identify a hybrid specimen, an identification that can be missed by the mitochondrial targets often used (Fig. 4). Previously described nuclear barcode markers (e.g. 28S) may be too long for sequencing with NGS (Chu, Li, & Qi, 2006), while other recently published mini-barcode solutions were adaptations of existing mitochondrial barcode targets, COI (Shokralla et al., 2015) and the 16S ribosomal RNA (Armani, Guardone et al., 2015; Armani, Tinacci et al., 2015; Muñoz-Colmenero et al., 2017).

The relatively short amplicon lengths allowed sequencing to be done by Next-Generation Sequencing technologies, and the resulting identification of different species mixed in the same sample. This can either be due to the original sample itself being a mixture, or to the combination of multiple specimens in the same test sample as a strategy to facilitate large-scale screening of fish substitutions. Although the results presented in this article were obtained using a Roche 454 junior, whose sequencing technology tends to produce relatively longer reads (Liu et al., 2012), the length of the amplicons produced by the identified primers (<350 bp) makes the methods compatible with other currently available technologies.

On the other hand, as observed in other efforts to generate short barcode solutions (Shokralla et al., 2015), successful resolution at the species level becomes a challenge. A complete solution will probably require a combination of targets, each with their different and complementary capabilities to differentiate species in various families of interest. Follow-up work is now under way to identify primers applicable to other families of commercial interest, including cods, salmons, tunas and sturgeons. Together, these methods will complement existing fish identification strategies contributing to the establishment of an efficient and efficacious framework to detect and prevent substitution frauds along the food chain, as well as to fisheries management and conservation strategies.

Acknowledgements

We are grateful to Dr Michael Noren and Sven Kullander from the Swedish Museum of Natural History and the FP5 FishTrace project consortium for providing the voucher samples, Dr Edor Kabashi of the Institut du Cerveau et de la Moëlle Epinière for the zebrafish DNA, and Antonio Artese from the JRC Ispra cafeteria for the 40 sole and 40 turbot samples.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.foodcont.2017.04.009.

References

Angers-Loustau, A., Petrillo, M., Paracchini, V., Kagkli, D. M., Rischitor, P. E., Gallardo, A. P., et al. (2016). Towards plant species identification in complex samples: A bioinformatics pipeline for the identification of novel nuclear barcode candidates. *PLoS One, 11*(1), e0147692. http://dx.doi.org/10.1371/journal.pone.0147692.

Armani, A., Guardone, L., Castigliego, L., D'Amico, P., Messina, A., Malandra, R., et al. (2015). DNA and Mini-DNA barcoding for the identification of Porgies species (family Sparidae) of commercial interest on the international market. *Food Control, 50*, 589–596. http://dx.doi.org/10.1016/j.foodcont.2014.09.025.

Armani, A., Tinacci, L., Giusti, A., Castigliego, L., Gianfaldoni, D., & Guidi, A. (2013). What is inside the jar? Forensically informative nucleotide sequencing (FINS) of a short mitochondrial COI gene fragment reveals a high percentage of mislabeling in jellyfish food products. *Food Research International, 54*(2), 1383–1393. http://dx.doi.org/10.1016/j.foodres.2013.10.003.

Armani, A., Tinacci, L., Xiong, X., Castigliego, L., Gianfaldoni, D., & Guidi, A. (2015). Fish species identification in canned pet food by BLAST and Forensically Informative Nucleotide Sequencing (FINS) analysis of short fragments of the mitochondrial 16s ribosomal RNA gene (16S rRNA). *Food Control, 50*, 821–830. http://dx.doi.org/10.1016/j.foodcont.2014.10.018.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., et al. (2004). Ultraconserved elements in the human genome. *Science, 304*(5675), 1321–1325. http://dx.doi.org/10.1126/science.1098119.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic Acids Research, 41*(D1), D36–D42. http://dx.doi.org/10.1093/nar/gks1195.

Carvalho, D. C., Palhares, R. M., Drummond, M. G., & Frigo, T. B. (2015). DNA barcoding identification of commercialized seafood in South Brazil: A governmental regulatory forensic program. *Food Control, 50*, 784–788. http://dx.doi.org/10.1016/j.foodcont.2014.10.025.

Chapela, M. J., Sotelo, C. G., Pérez-Martín, R. I., Pardo, M.Á., Pérez-Villareal, B., Gilardi, P., et al. (2007). Comparison of DNA extraction methods from muscle of canned tuna for species identification. *Food Control, 18*(10), 1211–1215. http://dx.doi.org/10.1016/j.foodcont.2006.07.016.

Chu, K. H., Li, C. P., & Qi, J. (2006). Ribosomal RNA as molecular barcodes: A simple correlation analysis without sequence alignment. *Bioinformatics, 22*(14), 1690–1701. http://dx.doi.org/10.1093/bioinformatics/btl146.

Cunha, H. A., Silva, V. M. F. da, Santos, T. E. C., Moreira, S. M., Carmo, N. A. S. do, & Solé-Cava, A. M. (2015). When you get what you haven't paid for: Molecular identification of “Douradinha” fish fillets can help end the illegal use of river dolphins as bait in Brazil. *Journal of Heredity, 106*(S1), 565–572. http://dx.doi.org/10.1093/jhered/esv040.

Di Pinto, A., Di Pinto, P., Terio, V., Bozzo, G., Bonerba, E., Ceci, E., et al. (2013). DNA barcoding for detecting market substitution in salted cod fillets and battered cod chunks. *Food Chemistry, 141*(3), 1757–1762. http://dx.doi.org/10.1016/j.foodchem.2013.05.093.

D'Amico, P., Armani, A., Gianfaldoni, D., & Guidi, A. (2016). New provisions for the labelling of fishery and aquaculture products: Difficulties in the implementation of Regulation (EU) n. 1379/2013. *Marine Policy, 71*, 147–156. http://dx.doi.org/10.1016/j.marpol.2016.05.026.

EU. (2011). *Regulation (EU) No 1169/2011*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32011R1169.

EU. (2013). *Regulation (EU) No 1379/2013*.

FAO. (2014). *The state of world fisheries and aquaculture 2014*.

Faircloth, B. C., Sorenson, L., Santini, F., & Alfaro, M. E. (2013). A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). *PLOS ONE, 8*(6), e65923. http://dx.doi.org/10.1371/journal.pone.0065923FAO (2014). The State of World Fisheries and Aquaculture 2014.

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

Figueras, A., Robledo, D., Corvelo, A., Hermida, M., Pereiro, P., Rubiolo, J. A., et al. (2016). Whole genome sequencing of turbot (Scophthalmus maximus; Pleuronectiformes): A fish adapted to demersal life. *DNA Research*. http://dx.doi.org/10.1093/dnares/dsw007. dsw007.

Filonzi, L., Chiesa, S., Vaghi, M., & Nonnis Marzano, F. (2010). Molecular barcoding reveals mislabelling of commercial fish products in Italy. *Food Research International, 43*(5), 1383–1388. http://dx.doi.org/10.1016/j.foodres.2010.04.016.

Fricke, R. (2015). European red list of marine fishes. *Marine Biology Research, 11*(9), 1004–1007. http://dx.doi.org/10.1080/17451000.2015.1064535.

Galal-Khallaf, A., Osman, A. G. M., Carleos, C. E., Garcia-Vazquez, E., & Borrell, Y. J. (2016). A case study for assessing fish traceability in Egyptian aquafeed formulations using pyrosequencing and metabarcoding. *Fisheries Research, 174*, 143–150. http://dx.doi.org/10.1016/j.fishres.2015.09.009.

Gordon, A., & Hannon, G. J. (2010). *Fastx-toolkit*. FASTQ/A Short-Reads Preprocessing Tools (unpublished) http://hannonlab.Cshl.Edu/fastx_toolkit.

Hebert, P. D. N., Ratnasingham, S., & Waard, J. R. de (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences, 270*(Suppl 1), S96–S99. http://dx.doi.org/10.1098/rsbl.2003.0025.

Helyar, S. J., Lloyd, H. ap D., Bruyn, M. de, Leake, J., Bennett, N., & Carvalho, G. R. (2014). Fish product mislabelling: Failings of traceability in the production chain and Implications for illegal, unreported and unregulated (IUU) fishing. *PLoS One, 9*(6), e98691. http://dx.doi.org/10.1371/journal.pone.0098691.

Herrero, B., Lago, F. C., Vieites, J. M., & Espiñeira, M. (2012). Real-time PCR method applied to seafood products for authentication of European sole (Solea solea) and differentiation of common substitute species. *Food Additives & Contaminants: Part, 29*(1), 12–18. http://dx.doi.org/10.1080/19440049.2011.623682.

Herrero, B., Madriñán, M., Vieites, J. M., & Espiñeira, M. (2010). Authentication of Atlantic Cod (Gadus morhua) using real time PCR. *Journal of Agricultural and Food Chemistry, 58*(8), 4794–4799. http://dx.doi.org/10.1021/jf904018h.

Ivanova, N. V., Zemlak, T. S., Hanner, R. H., & Hebert, P. D. N. (2007). Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes, 7*(4), 544–548. http://dx.doi.org/10.1111/j.1471-8286.2007.01748.x.

Kappel, K., & Schröder, U. (2015). Species identification of fishery products in Germany. *Journal Für Verbraucherschutz Und Lebensmittelsicherheit, 10*(1), 31–34. http://dx.doi.org/10.1007/s00003-015-0988-y.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*(14), 3059–3066. http://dx.doi.org/10.1093/nar/gkf436.

Khaksar, R., Carlson, T., Schaffner, D. W., Ghorashi, M., Best, D., Jandhyala, S., et al. (2015). Unmasking seafood mislabeling in U.S. markets: DNA barcoding as a unique technology for food authentication and quality control. *Food Control, 56*, 71–76. http://dx.doi.org/10.1016/j.foodcont.2015.03.007.

Knowlton, N. (1993). Sibling species in the sea. *Annual Review of Ecology and Systematics, 24*, 189–216.

Kodama, Y., Shumway, M., & Leinonen, R. (2012). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research, 40*(D1), D54–D56. http://dx.doi.org/10.1093/nar/gkr854.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., et al. (2010). The European nucleotide archive. *Nucleic Acids Research*. http://dx.doi.org/10.1093/nar/gkq967. gkq967.

Leinonen, R., Sugawara, H., & Shumway, M. (2010). The sequence read archive. *Nucleic Acids Research*. http://dx.doi.org/10.1093/nar/gkq1019. gkq1019.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *BioMed Research International, 2012*, e251364. http://dx.doi.org/10.1155/2012/251364.

Maddirevula, S., Anuppalle, M., Huh, T.-L., Kim, S. H., & Rhee, M. (2011). Nrdp1 governs differentiation of the melanocyte lineage via Erbb3b signaling in the zebrafish embryogenesis. *Biochemical and Biophysical Research Communications, 409*(3), 454–458. http://dx.doi.org/10.1016/j.bbrc.2011.05.025.

Mariani, S., Griffiths, A. M., Velasco, A., Kappel, K., Jérôme, M., Perez-Martin, R. I., et al. (2015). Low mislabeling rates indicate marked improvements in European seafood market operations. *Frontiers in Ecology and the Environment, 13*(10), 536–540. http://dx.doi.org/10.1890/150119.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal, 17*(1), 10–12. http://dx.doi.org/10.14806/ej.17.1.200.

Muñoz-Colmenero, M., Martínez, J. L., Roca, A., & Garcia-Vazquez, E. (2017). NGS tools for traceability in candies as high processed food products: Ion Torrent PGM versus conventional PCR-cloning. *Food Chemistry, 214*, 631–636. http://dx.doi.org/10.1016/j.foodchem.2016.07.121.

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences, 85*(8), 2444–2448.

Rasmussen, R. S., & Morrissey, M. T. (2009). Application of DNA-based methods to identify fish and seafood substitution on the commercial market. *Comprehensive Reviews in Food Science and Food Safety, 8*(2), 118–154. http://dx.doi.org/10.1111/j.1541-4337.2009.00073.x.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics: TIG, 16*(6), 276–277.

Salmela, L., & Schröder, J. (2011). Correcting errors in short reads by multiple alignments. *Bioinformatics, 27*(11), 1455–1461. http://dx.doi.org/10.1093/bioinformatics/btr170.

Schuler, G. D. (1997). Sequence mapping by electronic PCR. *Genome Research, 7*(5), 541–550.

Sevilla, R. G., Diez, A., Norén, M., Mouchel, O., Jérôme, M., Verrez-Bagnis, V., et al. (2007). Primers and polymerase chain reaction conditions for DNA barcoding teleost fish based on the mitochondrial cytochrome b and nuclear rhodopsin genes. *Molecular Ecology Notes, 7*(5), 730–734. http://dx.doi.org/10.1111/j.1471-8286.2007.01863.x.

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

Stockle, M. Y., & Hebert, P. D. N. (2008). Barcode of life. *Scientific American, 299*(4), 82–88. http://dx.doi.org/10.1038/scientificamerican1008-82.

Teletchea, F. (2009). Molecular identification methods of fish species: Reassessment and possible applications. *Reviews in Fish Biology and Fisheries, 19*(3), 265–293. http://dx.doi.org/10.1007/s11160-009-9107-4.

Thummel, R., Ju, M., Sarras, M. P., Jr., & Godwin, A. R. (2007). Both Hoxc13 orthologs are functionally important for zebrafish tail fin regeneration. *Development Genes and Evolution, 217*(6), 413–420. http://dx.doi.org/10.1007/s00427-007-0154-3.

Vandamme, S. G., Griffiths, A. M., Taylor, S.-A., Di Muri, C., Hankard, E. A., Towne, J. A., et al. (2016). Sushi barcoding in the UK: Another kettle of fish. *PeerJ, 4*, e1891. http://dx.doi.org/10.7717/peerj.1891.

Ward, R. D., Hanner, R., & Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology, 74*(2), 329–356. http://dx.doi.org/10.1111/j.1095-8649.2008.02080.x.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1462), 1847–1857. http://dx.doi.org/10.1098/rstb.2005.1716.

Watanabe, M. E. (2013). Assembling an online tree of life of two Million species. *BioScience, 63*(1). http://dx.doi.org/10.1525/bio.2013.63.1.16, 64–64.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology, 15*, R46. http://dx.doi.org/10.1186/gb-2014-15-3-r46.

Yan, S., Lai, G., Li, L., Xiao, H., Zhao, M., & Wang, M. (2016). DNA barcoding reveals mislabeling of imported fish products in Nansha new port of Guangzhou, Guangdong province, China. *Food Chemistry, 202*, 116–119. http://dx.doi.org/10.1016/j.foodchem.2016.01.133.

Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics/Editoral Board*. http://dx.doi.org/10.1002/0471250953.bi1105s31. Andreas D. Baxevanis… [et Al.], CHAPTER, Unit–11.5.