FOOD COMPOSITION AND ADDITIVES

A Single-Laboratory Validated Method for the Generation of DNA Barcodes for the Identification of Fish for Regulatory Compliance

SARA M. HANDY and JONATHAN R. DEEDS
U.S. Food and Drug Administration, Office of Regulatory Science, Center for Food Safety and Applied Nutrition, College Park, MD 20740
NATALIA V. IVANOVA and PAUL D.N. HEBERT
Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, University of Guelph, 50 Stone Rd East, Guelph, ON, Canada, N1G 2W1
ROBERT H. HANNER
University of Guelph, Department of Integrative Biology, 50 Stone Rd East, Guelph, ON, Canada, N1G 2W1
ANDREA ORMOS and LEE A. WEIGT
Smithsonian Institution, National Museum of Natural History, Laboratories of Analytical Biology, MRC 534, Washington, DC 20013-7012
MICHELLE M. MOORE
U.S. Food and Drug Administration, Office of Regulatory Affairs, Pacific Regional Laboratory Northwest, Applied Technology Center, 22201 23rd Dr SE, Bothell, WA 98021-4421
HAILE F. YANCY
U.S. Food and Drug Administration, Office of Research, Center for Veterinary Medicine, 8401 Muirkirk Rd, Laurel, MD 20708

The U.S. Food and Drug Administration is responsible for ensuring that the nation's food supply is safe and accurately labeled. This task is particularly challenging in the case of seafood where a large variety of species are marketed, most of this commodity is imported, and processed product is difficult to identify using traditional morphological methods. Reliable species identification is critical for both foodborne illness investigations and for prevention of deceptive practices, such as those where species are intentionally mislabeled to circumvent import restrictions or for resale as species of higher value. New methods that allow accurate and rapid species identifications are needed, but any new methods to be used for regulatory compliance must be both standardized and adequately validated. "DNA barcoding" is a process by which species discriminations are achieved through the use of short, standardized gene fragments. For animals, a fragment (655 base pairs starting near the 5′ end) of the cytochrome *c* oxidase subunit 1 mitochondrial gene has been shown to provide reliable species level discrimination in most cases. We provide here a protocol with single-laboratory validation for the generation of DNA barcodes suitable for the identification of seafood products, specifically fish, in a manner that is suitable for FDA regulatory use.

Received June 22, 2010. Accepted by SG July 23, 2010.
Corresponding author's e-mail: Sara.Handy@fda.hhs.gov

The overarching mission of the U.S. Food and Drug Administration (FDA) is to promote and protect the public's health. According to the Federal Food, Drug, and Cosmetic Act, the Fair Packaging and Labeling Act, and the Public Health Service Act, aquatic animals that are harvested, processed, distributed, and sold in the United States as food must be safe, wholesome, and properly labeled. Recently, seafood has gained increased attention due to the potential health-related risks associated with misbranding of species (1). Another area in which seafood has garnered significant attention is seafood fraud (2). This practice can occur in several forms, such as over-breading, processing, glazing (where excessive ice is added to a product that is reweighted to include the ice in the label weight), short-weighting, trans-shipping (where an imported product is shipped through an additional port to avoid tariffs), and species substitution.

Species substitution, or misbranding, is a practice where low value species or a species with a potential food safety hazard is intentionally mislabeled and substituted in whole or in part for a more expensive species, or for one with no potential food safety hazard. Often, substitution is impossible to detect by simple inspection of an aquatic

Table 1. Comparison of quality value and read length for different sequencing polymers

Sample	Polymer	Quality of F[a]	Quality of R[b]	Length F/R
Snapper-1	Pop-7	99.4	99.4	650/637
Snapper-1	Pop-6	99.8	98.9	655/651
Snapper-2	Pop-7	99.7	99.1	655/638
Snapper-2	Pop-6	99.7	99.1	655/652
Snapper-3	Pop-7	98.9	99.5	645/634
Snapper-3	Pop-6	99.1	99.6	655/652
Snapper-4	Pop-7	98.6	98.9	650/630
Snapper-4	Pop-6	99.8	98.6	655/651

[a] F = Forward sequence.
[b] R = Reverse sequence.

product. Processing often removes or damages diagnostic characteristics crucial for the identification of species by conventional taxonomic means (3). In these cases, traditional morphological methods are often insufficient to provide species resolution. An extreme example would be the identification of cooked meal remnants involved in food poisoning outbreaks (1). Proper species identification is critical in these investigations to ensure that the suspect product is correctly identified so that additional product can be removed from public commerce.

In a 2009 report by the U.S. Government Accountability Office to the U.S. Senate, the FDA was identified as one of three key U.S. agencies responsible for the identification and prevention of seafood fraud in the marketplace (4). Within the FDA, the Center for Food Safety and Applied Nutrition and the Office of Regulatory Affairs have the responsibility to develop, apply, and provide the necessary technology and methods for the effective identification of fish species for the purposes of protecting the public from improperly and illegally imported products that may constitute a health threat. The FDA is also responsible for having the necessary tools to detect and enforce regulations against substitution of one species for another in cases of economic adulteration (i.e., seafood fraud).

Several years ago, the FDA developed a web-based resource known as the Regulatory Fish Encyclopedia (RFE) to aid in the identification of commercially important species of fish (5). The method for species identification of fish fillets recommended in the RFE is protein isoelectric focusing (IEF). Even though this analysis requires subjective interpretations of gel results and is not effective in the case of processed, cooked, or degraded samples, it remains the only method that has been validated for regulatory compliance (6). In addition, this method requires the inclusion of perishable frozen standards in each run that must be replaced periodically. The online RFE was designed so that it could be expanded to include additional data and accommodate the use of newer analytical tools as they became available. Recently, DNA sequence data were generated for 172 individual authenticated fish representing 72 species from 27 families contained in the RFE (7). The methodology used to generate this sequence data was based on the international Barcode of Life (iBOL) initiative (http://www.ibolproject.org/; accessed April 4, 2010), specifically the FISH-BOL campaign (http://www.fishbol.org/; accessed June 11, 2010).

The iBOL initiative represents an ambitious effort to develop an identification system for eukaryotic life based upon the analysis of sequence diversity in short, standardized gene regions (a.k.a. barcodes). For animals, a target gene region has been selected: a fragment consisting of 648-655 base pairs starting near the 5′ end of the cytochrome *c* oxidase subunit 1 (COI) mitochondrial gene (8, 9). This region has been shown to reliably discriminate most commercial species of fish (10–12). Even relatively short nucleotide sequences (100–200 bp) from this barcode region can provide accurate identifications in some cases, allowing barcodes to identify some specimens whose DNA is degraded due to heavy processing (e.g., canning; 13). After Yancy et al. (7) showed that DNA barcoding could be used to reliably discriminate the fish species contained in the RFE, we set out to validate a method for DNA barcode generation in fish to make this method fully acceptable for regulatory compliance. Based primarily on equipment and methodologies used at the large, high-throughput, "barcoding factories" at the Biodiversity Institute of Ontario at the University of Guelph and the Smithsonian Institution National Museum of Natural History's Laboratories of Analytical Biology, we first attempted to develop a standardized method for DNA barcode generation for fish and published this method in the FDA Laboratory Information Bulletin (LIB No. 4420: A Protocol for Validation of DNA-Barcoding for the Species Identification of Fish for FDA Regulatory Compliance), available at: http://www.fda.gov/Food/ScienceResearch/LaboratoryMethods/ucm169034.htm. After an initial three-laboratory pilot trial, we further refined this method and modified it based on equipment available in FDA regulatory laboratories. We provide here a detailed protocol with single-laboratory validation (SLV) for the generation of DNA barcodes suitable for the identification of seafood products, specifically fish, in a form that is suitable for FDA regulatory use.

Methods for Pilot Study and Ruggedness Testing

Pilot Study Design

As a first step toward validation of a standardized method for the generation of DNA barcodes in fish, a three-laboratory, pilot trial was conducted using FDA LIB method No. 4420. Blind test plates containing quadruplicate tissue samples from 24 individual fish were prepared at the FDA and sent to the laboratories of Paul Hebert (Lab 1) at the Canadian Center for DNA Barcoding (CCDB); Robert Hanner (Lab 2) at the Department of Integrative Biology, both located at the University of Guelph, Ontario, Canada; and Lee Weigt

Table 2. Authenticated fish species standards used for both the interlaboratory trial of LIB No. 4420 and the SLV

Sample No.	Species	Common name[a]	FDA acceptable market name[b]	NMNH ID No.
FDA-14	*Rhomboplites aurorubens*	Vermillion snapper	Snapper	394149
FDA-15	*Morone saxatilis*	Striped bass	Bass	394145
FDA-17	*Paralichthys lethostigma*	Southern flounder	Flounder or Fluke	394061
FDA-18	*Morone americana*	White perch	Perch, white	394152
FDA-19	*Centropristis striata*	Black sea bass	Bass, sea	394056
FDA-20	*Oreochromis niloticus*	Nile tilapia	Tilapia	394147
FDA-21	*Urophycis chuss*	Red hake	Hake	394058
FDA-22.1	*Merluccius bilnearis*	Silver hake	Whiting	394060
FDA-22.2	*Merluccius bilnearis*	Silver hake	Whiting	394060
FDA-23	*Micropogonias undulatus*	Atlantic croaker	Croaker	394164
FDA-24	*Salmo salar*	Atlantic salmon	Salmon	394151
FDA-25	*Leiostomus xanthurus*	Spot	Spot	394055
FDA-26	*Stenotomus chrysops*	Scup	Porgy or Scup	394153
FDA-27	*Pomatomus saltatrix*	Bluefish	Bluefish	394142
FDA-28	*Ictiobus cyprinellus*	Bigmouth buffalo	Buffalofish	394165
FDA-29	*Ocyurus chrysurus*	Yellowtail snapper	Snapper	394057
FDA-32	*Scomberomorus regalis*	Cero	Mackerel, Spanish	394146
FDA-33	*Mugil cephalus*	Striped mullet	Mullet	394059
FDA-34	*Peprilus paru*	American harvestfish	Butterfish	394062
FDA-35	*Haemulon plumieri*	White grunt	Grunt	394144
FDA-36	*Lagodon rhomboides*	Pinfish	Porgy	394054
FDA-37	*Archosargus probatocephalus*	Sheepshead	Sheepshead	394053
FDA-38	*Sciaenops ocellatus*	Red drum	Drum or Redfish	394143
FDA-39	*Ictalurus punctatus*	Channel catfish	Catfish	394150

[a] According to the American Fisheries Society (19).
[b] According to the Seafood List (20).

(Lab 3), at the Smithsonian Institution's L.A.B. (Laboratories of Analytical Biology), located in Suitland, MD. Additional components of the method were further subjected to ruggedness testing (*see* below) based on data obtained during the pilot interlaboratory trial. Full results from these pilot studies are not provided here, but relevant data that were used to develop an optimized protocol for DNA barcode generation in fish suitable for formal validation are included below. This optimized protocol was further subjected to a formal SLV at the FDA (full results below).

Primer and PCR Cocktail Selection

Using all 94 extracted tissues from one of the pilot interlaboratory test plates (Lab 1), the primer sets from Ivanova et al. (14) with PCR cocktails from Ivanova and Grainger (15) and modified from Baldwin et al. (16; *see* note under *Procedure*) were compared for their ability to produce a PCR product of approximately 700 bases with a high-quality sequence, defined here as sequence data that were either bidirectional with <2% ambiguous bases in the final contig (contiguous consensus sequence based on the individual bidirectional reads), or single reads of ≥98% quality, based on KB basecaller software scores and >500 bases in length. [*Note*: This quality value was generated using the program Sequencher 4.9 (Gene Codes Corp., Ann Arbor, MI) and is based on the percentage of bases in a sequence that is above the low confidence range threshold. In this case, the low confidence range was a KB basecaller score of 20 or less, which means a 1/100 chance of an incorrect base call (from Sequencher 4.8 user manual for Macintosh).] This scoring is similar to another commonly used base quality value called Phred (17), and the scores can be used interchangeably (18). In addition, the PCR cocktails (mixture of all PCR reagents besides the DNA template) from each study were compared using each primer set. The greatest difference between the two cocktails, besides the primer design and concentration (lower concentration in CCDB cocktail), was the presence (CCDB) or absence (SI) of 10% trehalose and differences in the concentration of deoxynucleoside triphosphates (dNTPs). The PCR thermocycling conditions were: 2 min at 94°C; 35 cycles of 94°C for 30 s; 52°C for 40 s; 72°C for 1 min; and then a final extension at 72°C for 10 min.

Sequencing Polymer Selection

To assess the ruggedness of the sequencing polymers used in this study, extracted DNA from four different species of snapper (Family Lutjanidae) were sequenced on two different instruments, an ABI 3730xl using Pop-7™ Polymer and an ABI 3130xl using Pop-6™ Polymer. These four samples were extracted, amplified, and sequenced using the optimized protocol below. According to the manufacturer's literature, Pop-6 is appropriate for standard and rapid sequencing, Pop-7 for DNA sequencing and fragment analysis.

Results and Discussion for Pilot Studies and Ruggedness Testing

Pilot Study

Due to insufficient detail in the initial protocol, two laboratories involved in the pilot interlaboratory trial used slightly different methods to identify the test plate of fish tissues (mainly variation in PCR primers and PCR cocktails). Upon analysis of the data, this appeared to result in different efficiencies between laboratories (both in success of PCR and sequencing). This information was used to design specific ruggedness tests to confirm these initial observations and to clarify specific steps that needed to be followed to generate a unique barcode for fish specimens, in an efficient manner, that met a set of minimum criteria (defined below).

PCR Cocktail Selection

One of the ruggedness tests that was undertaken was to determine whether the primer sets of Ivanova et al. (14; here called CCDB) or modified from Baldwin et al. (16; here called SI; *see* note under *Procedure*), each tested with both of their respective PCR cocktails, would provide a higher number of successful PCRs and sequencing reactions (reads of at least 500 bases, with a quality value of <98% for single reads or contigs, with <2% ambiguous bases). The goal was to accomplish this in a single run with no hand-editing of sequences. The following was determined: SI primers with SI PCR cocktail yielded 92 of 94 single amplicons of approximately 650 bases. CCDB primers did not amplify as well with the SI PCR cocktail (only 82 of 94 had single amplicons of approximately 650 bases). Both the SI and CCDB primers had equal PCR success with the CCDB PCR cocktail (94 out of 94 had single amplicons of approximately 650 bases). Sequencing success (reads of at least 500 bases, with quality values of <98% for single reads or contigs with <2% ambiguous bases) for the CCDB PCR cocktail was 91 of 94 for SI primers and 88 of 94 for CCDB primers. As a result, the fish barcoding primers modified from Baldwin et al. (16; SI; *see* note under *Procedure*) and the PCR cocktail from Ivanova and Grainger (15; CCDB), which contained 10% trehalose, were deemed most efficient. This primer set/PCR cocktail was then used in the SLV of DNA barcode generation in fish.

Buffers

A second ruggedness test compared using Pop-7 Polymer on an ABI 3730xl or Pop-6 Polymer on an ABI 3130xl; both generated sequences that were >98% quality and around 650 bases in length for the four snapper samples compared (Table 1). It was determined that either polymer would be acceptable for barcode generation. The SLV for DNA barcode generation in fish used Pop-7 polymer on an ABI 3730 instrument, but it is known that some FDA field laboratories will be using Pop-6 polymer on an ABI 3130xl; therefore, this ruggedness test was also needed.

SLV Experimental

Apparatus

(**a**) *Eppendorf Mastercycler® ep gradient S thermocycler.*—Cat. No. 950010045 (Eppendorf, Hamburg, Germany).

(**b**) *Gel Doc 2000 gel documentation system.*—Bio-Rad (Hercules, CA).

(**c**) *E-Base® integrated power supply.*—Cat. No. EB-M03 (Invitrogen, Carlsbad, CA).

(**d**) *AirClean® systems ductless PCR workstation.*—Cat. No. 36 099 3859 (Fisher Scientific, Waltham, MA).

(**e**) *Nanodrop ND 1000 spectrophotometer.*—Thermo Scientific (Wilmington, DE).

(**f**) *Sorvall Evolution RC.*—Thermo Scientific.

Reagents and Consumables

(**a**) *Reagent alcohol, histological (EtOH 96%).*—Cat. No. A962-4 (Fisher Scientific).

(**b**) *DNeasy blood and tissue kit.*—Cat. No. 69504 (Qiagen, Valencia, CA).

(**c**) *Molecular grade water (dd H_2O).*—Cat. No. 10977023 (Invitrogen).

(**d**) *D – (+) – Trehalose dihydrate.*—Cat. No. 90210-50g (Sigma-Aldrich, St. Louis, MO).

(**e**) *10X PCR buffer, minus Mg.*—Cat. No. 10966-034 (Invitrogen).

(**f**) *Deoxynucleotide solution mixture.*—Cat. No. N0447L (New England Biolabs, Ipswitch, MA).

(**g**) *Oligonucleotide primers.*—Integrated DNA Technologies (Coralville, IA); *see PCR* section below for primer sequences.

(**h**) *Platinum Taq DNA polymerase.*—Cat. No. 10966-034 (Invitrogen).

(**i**) *2% E-Gel® 96 precast agarose gels.*—Cat. No. G7008-02 (Invitrogen).

(**j**) *5X Sequencing buffer (400 nm Tris-HCl, pH 9.0 + 10 mM $MgCl_2$).*—(Applied Biosystems, Carlsbad, CA).

(**k**) *BigDye® Terminator v3.1 cycle sequencing kit.*—Cat. No. 4337457 (Applied Biosystems).

(**l**) *Pop-7 Polymer for 3730 DNA analyzers.*—Cat. No. 4363929 (Applied Biosystems).

(**m**) *3730 DNA analyzer.*—Cat. No. 3730S (Applied Biosystems).

(**n**) *3730 DNA analyzer capillary array, 50 cm.*—Cat. No. 4331250 (Applied Biosystems).

(**o**) *Edge Bio PERFORMA DTR V3 96-well short plate kit.*—Cat. No. 899 39 (Edge Bio, Gaithersburg, MD).

(**p**) *Hi-Di™ formamide*.—Cat. No. 4311320 (Applied Biosystems).

(**q**) *96-Well semiskirted PCR plate*.—Cat. No. 14-230-244 (Fisher Scientific).

(**r**) *ExoSAP-IT*.—Cat. No. 78201 (USB Corp., Cleveland, OH).

Tissue Collection

Twenty-four whole fish specimens were purchased by the FDA from local markets in the Washington, DC, and Baltimore, MD, areas. Samples of white muscle from the right filet were taken from each fish, and the remainder of the specimen was submitted to the Fish Division of the National Museum of Natural History (NMNH) for authentication, preservation, vouchering, and curation. Tissue samples were stored at –20°C until subsampled for this study. Sample details, including NMNH catalog numbers, are listed in Table 2.

Procedure

(**a**) *Cell lysis and DNA extraction.—Goal.*—Extract DNA to be used in PCR.

Criteria for success.—Obtain ≥5 ng/μL of DNA for all samples (*Note*: This was the lowest quantity used for the SLV; lower quantities may work) measured on a Nanodrop ND 1000 spectrophotometer with a 260/280 nm ratio of approximately 1.8. In addition, a negative control with no added DNA should give a reading of 0 ng/μL.

On 4 separate days, a small piece of tissue (approximately 10 mg) from each of the 24 filets was removed using flame-sterilized (with EtOH) tweezers and added to a sterile 1.5 mL microcentrifuge tube. DNA was extracted from tissue by use of a DNeasy Blood & Tissue kit. A negative control was included with each set of 24 samples with no added tissue. Reagent volumes were reduced to a quarter of the volume listed in the manual (50 μL buffer ATL with 5.56 μL proteinase K, followed by 55.56 μL buffer AL and 55.6 μL EtOH). For the wash steps, 140 μL AW1 and AW2 were used, followed by elution with 50 μL buffer AE.) Besides these changes, the manufacture's protocol was followed, with the additional step of incubating the washed filters and elution buffer at 37°C for 30 min to increase successful elution of DNA. After extraction, all samples were quantified using a Nanodrop ND 1000 spectrophotometer.

(**b**) *PCR.—Goal*.—To amplify approximately 700 bases starting near the 5′ end of the CO1 mitochondrial gene suitable for sequencing.

Criteria for success.—A sample that produced a single band of approximately 700 bases in size as visualized on a precast 1% agarose gel. Each set of reactions also required two negative controls, one of which consisted of the PCR cocktail with no additional DNA template added and another with an addition of the extraction negative control. These samples had to be negative on the agarose gel (no band produced).

PCR primers FISHCO1LBC: 5′-TCAACYAATCAY AAAGATATYGGCAC, and FISHCO1HBC: 5′-ACTTCY GGGTGRCCRAARAATCA called Fish-BCL and Fish-BCH in Baldwin et al. (16), were selected for use in this validation study. [*Note*: FISHCO1HBC has degeneracies at positions 6, 12, 15, and 18 compared to Fish-BCH published in Baldwin et al. (16)]. The primer sets were tailed to stream-line sequencing and to allow for a longer read in the CO1 gene using M13F-29: 5′-CACGACGTTGTAAAACGAC-3′ and M13R: 5′GGATAACAATTTCACACAGG-3′.

Four separate PCRs (one for each of the individual extractions) were run that included each of the 24 samples, plus a negative control. The PCR cocktail from Ivanova and Grainger (15), which consisted of 6.25 μL 10% trehalose solution, 2 μL dd H_2O, 1.25 μL 10x PCR buffer, 0.625 μL 50 mM $MgCl_2$, 0.125 μL 10 μM of both primers, 0.062 μL 10 mM dNTPs, 0.060 μL Platinum Taq (5 U/μL), and 1 μL undiluted DNA template/reaction (11.5 μL total). An Eppendorf Mastercycler® ep gradient S thermocycler was used for the PCRs with the following conditions: 94°C for 2 min; 35 cycles of 94°C for 30 s; 55°C for 40 s; and 72°C for 1 min, with a final extension at 72°C for 10 min.

All products were verified using pre-cast 1% E-gel 48 agarose gels (two sets of PCR products/gel) according to manufactures' protocols with the E-Base® Integrated power supply. Gels were run for 5 min and then visualized using a Gel Doc 2000 gel documentation system. The gel was also photographed with this instrument, and the picture was retained for records.

(**c**) *PCR cleanup.—Goal*.— To produce an amplicon free of extra dNTPs and excess primers that might interfere with the sequencing reaction.

Criteria for success.—At least 97–303 ng/μL DNA measured on a Nanodrop ND 1000 spectrophotometer with a 260/280 nm ratio of approximately 1.8. (*Note*: This quantity of DNA worked well for this study; however, it is possible that lower or higher quantities of DNA would also produce acceptable results.)

Successfully amplified products were purified by adding 2 μL Exosap-IT to 5 μL PCR product, and incubating at 37°C for 15 min, followed by 15 min at 80°C. PCR products were then quantified using a Nanodrop ND 1000.

(**d**) *Cycle sequencing reaction.—Goal*.—To attach fluorescently labeled dideoxy nucleoside triphosphates (ddNTPs) using PCR-amplified template DNA.

Criteria for success.—None. Success of dye incorporation cannot be assessed until Step (**g**) *Post-sequencing analysis.*

Two sequencing reactions (one each in the forward and reverse direction) were run on each sample (24 × 4). Each set of 24 (48 samples each) was processed in a different thermocycler run. Each reaction contained: 0.25 μL BigDye® Terminator v3.1; 1.875 μL 5X sequencing buffer; 5 μL 10% trehalose; 1 μL primer (either M13F-29 or M13R); and 0.875 μL molecular grade water, for a total of 9 μL to which

Table 3. List of samples with ambiguities in the contigs

Sample	Ambiguities in contig	Ambiguous, %
FDA1-21	8	1.22
FDA4-21	7	1.07
FDA2-14	2	0.31
FDA1-18	1	0.15
FDA2-18	1	0.15
FDA3-18	1	0.15
FDA4-18	1	0.15
FDA1-19	1	0.15
FDA2-22.2	1	0.15
FDA1-23	1	0.15
FDA2-23	1	0.15
FDA4-23	1	0.15
FDA2-24	1	0.15
FDA3-23	1	0.15
FDA1-26	1	0.15
FDA2-27	1	0.15
FDA1-32	1	0.15
FDA1-36	1	0.15
FDA1-39	1	0.15
FDA3-39	1	0.15
FDA4-39	1	0.15
FDA1-17	1	0.15

1 μL cleaned up PCR product was added. Sequencing reactions were also conducted on an Eppendorf Mastercycler® ep gradient S Thermocycler with the following conditions: 96°C for 2 min; 30 cycles of 96°C for 30 s, 55°C for 15 s; and 60°C for 4 min, followed by a 4°C hold.

(**e**) *Sequencing reaction cleanup.—Goal*.—To remove unincorporated dye terminators and salts from sequencing reactions so that they will not interfere with the base pair determination of the fragment.

Criteria for success.—A 96-well sequencing plate containing between 10 and 15 μL cleaned sequencing reaction product after centrifugation. Success of unincorporated dye removal cannot be assessed until Step (**g**) *Post-sequencing analysis*.

After the sequencing reaction had finished, products were transferred directly into an Edge Bio PERFORMA DTR V3 96-well short plate kit that had been prepared according to the manufacturer's instructions. After products were added to the plate, the samples were spun at 850 × *g* for 2 min [instead of 5 min as the manufacturer suggests, because of specific instructions from Applied Biosystems (21)], using a Sorvall Evolution RC centrifuged into a Fisherbrand 96-well semi-skirted PCR plate. After the spin, plates were visually inspected to ensure that between 10 and 15 μL liquid was contained in the wells. [*Note*: Some residual liquid (3–5 μL) will come through even in the empty wells, according to the manufacturer, which will cause the final volume to be greater than the 10 μL added.] Finally, 10 μL Hi-Di™ formamide was added to each sample and pipetted to mix.

(**f**) *Sequencing.—Goal*.—To determine the accurate base pair composition of the CO1 amplicon.

Criteria for success.—Generation of an AB1 file, though successful determination of the accurate base pair composition cannot be assessed until Step (**g**), *Post-sequencing analysis*.

At this point, the samples could be put directly on the ABI 3730 sequencer, which was maintained according to manufacturer's specifications, using instrument protocol default_50cm along with analysis protocol 3730BDTv3-KB-DeNovo_v5.2.

(**g**) *Post-sequencing analysis.—Goal.*–To process the sequence from the AB1 file into a usable unit for comparison with other sequences in a database.

Criteria for success.–Bidirectional sequences, of at least 500 bp in length, with <2% ambiguous bases in the contig, or a single read with 98% quality value.

All ABI files were imported into the Sequencher 4.9. To trim off primers, all samples were aligned to a reference degenerate bony fish barcoding sequence developed and provided by Lee Weigt at the Smithsonian L.A.B. (sequence available upon request). The samples were then trimmed to the reference sequence. Samples were then trimmed further for quality based on the following: for the 5′-end, trimming no more than 25%, trim until the first 25 bases contain less than three ambiguities, and trimming no more than 25%, trim until the first 25 bases contain less than three with confidences below KB basecaller score of 20. For the 3′-end, starting 100 bases after the 5′ trim, trim the first 25 bases containing more than three ambiguities; trim from the 3′-end until the last 25 bases contain less than three ambiguities; and trim from the 3′-end until the last 25 bases contain less than three bases with confidence below a KB basecaller score of 20. The "postfix" was set to: maximum desired length after trimming is 655 bases, trim more from 3′-end if necessary, and remove leading and trailing ambiguous base. Any sequences whose lengths were less than 500 bp were considered failures and removed from the analysis. Next, the bidirectional sequences were assembled into contigs (with default settings: using dirty data algorithm, realigner and prefer 3′ gap placement, as well as a 20-base minimum overlap and an 85% minimum match percentage). At this point, if any contig contained >2% ambiguities, those samples were also removed. Any remaining single read sequences were used if their quality value was better than 98%. No hand-editing was performed on the sequences.

To test the hypothesis that 24 visually distinct whole fish samples would consistently produce 24 unique COI barcode sequences four consecutive times, contigs >500 bases of bidirectional coverage with less than 2% ambiguous bases or single reads >500 bases, with a quality value of >98%, were exported as fasta-concatenated files into the bioinformatic software, Geneious Pro [Biomatters Ltd, Auckland, New Zealand (22)], and a sequence list was created. An alignment

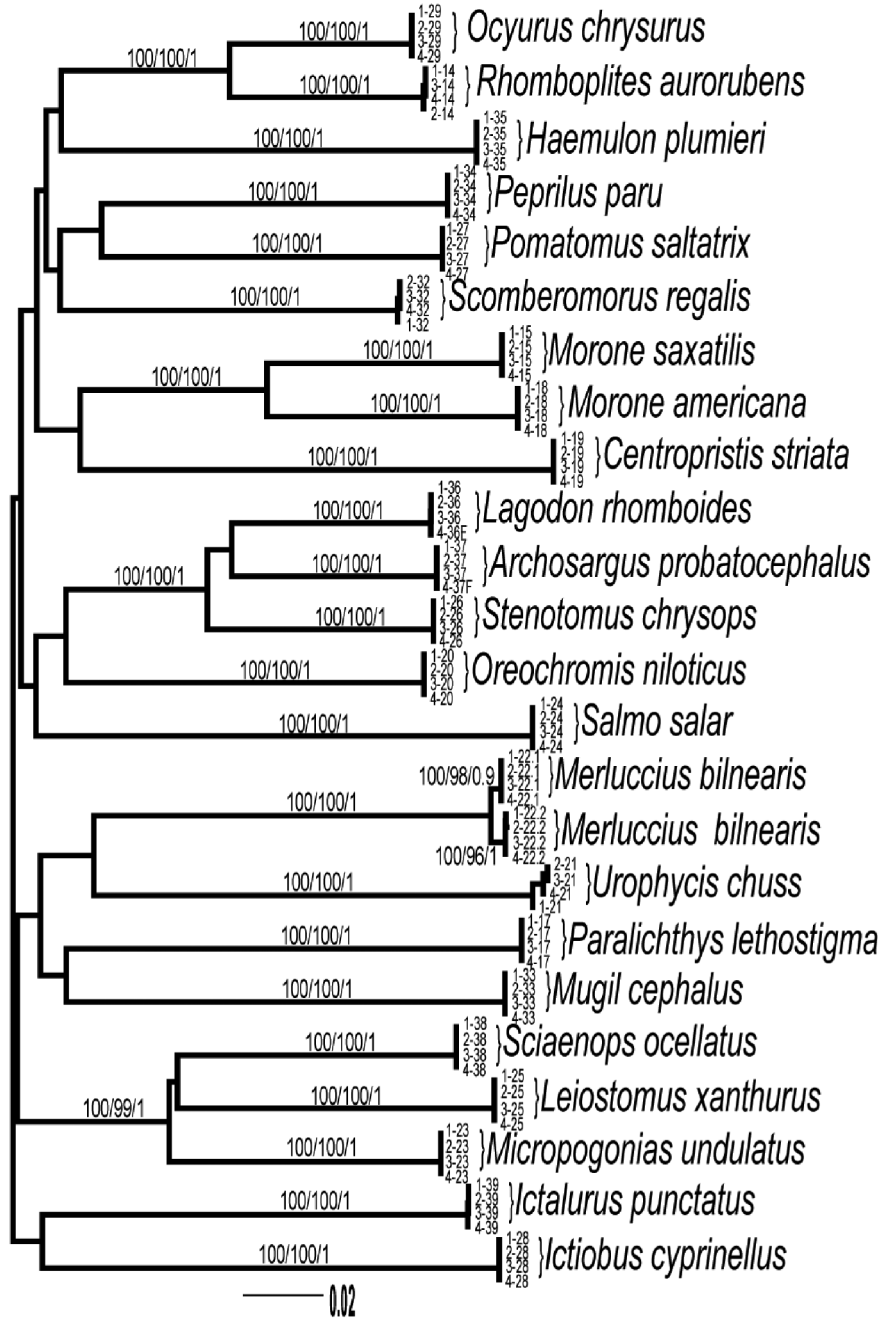

Figure 1. A neighbor-joining consensus tree of four replicates of 24 fish specimens examined in the SLV. Values located on the branches are neighbor-joining bootstraps/maximum likelihood bootstraps/Bayesian posterior probabilities, all generated in the program Geneious Pro. The 0.02 refers to the substitutions per site.

was constructed in Geneious from this using the "Muscle Alignment" tab, with default settings. Next, a neighbor-joining consensus tree with a Jukes-Cantor genetic distance model (23) was constructed, and bootstrap support values were generated. Finally, a maximum likelihood analysis using an HKY85 substitution model (24) and Bayesian analysis using an HKY85 nucleotide substitution model and gamma rate variation with a total chain length of 1 100 000, a subsample frequency of 200, a burn-in length of 110 000, with four heated chains, and a temperature of 0.2 (25) were run on both the individual data sets and the combined 24 × 4 set using the software Geneious Pro.

(**1**) *Precision*.—Each of the four sets of generated sequences was compared base-by-base visually in the alignment to determine the precision of acquiring the same barcode for each species.

(**2**) *Specificity*.—Each barcode was examined to determine that it was unique (variable by more than 2%) based

on a neighbor-joining tree using a Jukes-Cantor genetic distance model.

(**3**) *Uncertainty*.—Two of the selected fish (Nos. 22-1 and 22-2) are considered the same species. This allowed us to examine what is potentially real intraspecies variation in a barcode versus variation between barcodes due to ambiguous bases in the contig. In addition, one fish (sample No. 21) was somewhat difficult to sequence (it contained as many as eight ambiguous bases, six more than any other sample in the set of 96), making it an ideal sample for comparison.

SLV Results and Discussion

Molecular Analysis

DNA was successfully extracted from all 24 fish samples (sample list in Table 1) four separate times with resulting DNA concentrations ranging from 5.2 to 89.3 ng/μL, a mean of 33.7 ± 21.9 ng/μL, and a median of 25.75 ng/μL. All of these extractions (96 total) resulted in PCR products clearly visible as single bands on agarose gels. All negative controls, which were run alongside each separate PCR, were negative. Exo-SAP-it purified products yielded concentrations ranging from 97.5 to 303 ng/μL with a mean of 193.1 ± 35.8 ng/μL and a median of 191.7 ng/μL.

Of all 192 sequencing reactions, there were only two failures (samples FDA4-36R and FDA4-37R), but each of these samples worked well in the forward direction (quality better than 99.4%). This was a success rate of 99%. Four samples, FDA1-21, FDA3-21, FDA2-37, and FDA4-37, had contigs of 647, 644, 643, and 653 bases, respectively (shortened on the 3′-end). The rest of the forward and reverse sequences formed contigs of 655 bp in length. Of these, 22 out of the 96 had at least one ambiguity. Table 3 shows a list of ambiguities by sample; however, even the highest number of ambiguities (sample FDA1-21) still only had 1.2% ambiguities. Average quality values for the forward sequences were 99.1 ± 1.85% and for the reverse sequences was 98.98 ± 1.92%. Therefore, we feel that it is achievable to generate a good quality sequence from a range of different fish species in one attempt, with no hand-editing, by following this protocol. DNA sequences and raw trace files were deposited in a public folder on BOLD (26) under Campaign: Barcoding Fish (FISH-BOL), Project Title: FDA Single Lab Validation for Fish, Project Code: SLV. DNA sequences were also deposited in Genbank (accession numbers HQ024929-HQ025024).

Generation of Unique Barcodes

To determine if the same unique barcode could be generated from the 24 individual fish four consecutive times, a neighbor-joining tree was constructed using a Jukes-Cantor genetic distance model (Figure 1). Because no evolutionary history should be assumed based on short barcodes, Bayesian and maximum likelihood analyses were also run to provide statistical support for sequence groupings. Bootstraps from neighbor-joining and maximum likelihood analysis, as well as posterior probabilities, supported a unique branch (group) for each individual four consecutive times. Much of the early work on the use of DNA barcodes to identify species used techniques such as neighbor-joining analysis and distance thresholds, with a generalized cutoff of 2% sequence divergence, to delimit species (9). Newer studies have used more elaborate statistical methods to define species boundaries (27). For example, a character-based approach was recently used to successfully distinguish many of the large commercial species of tuna, some of which are endangered, and are impossible to distinguish in their marketed form (28). In this work, we describe a method for the generation of DNA barcodes in a reliable and repeatable manner. How that barcode is best used to match an unknown sample to a library of sequences from authenticated standards will ultimately depend on the target group in question (29).

Assessment of Uncertainty

Of the 23 species used in this study, only sample FDA-21 had a relevant amount of sequence variation between the four trials (1.7% bases when all four were compared). Other samples had one or at the most two ambiguities. All variation in FDA-21 was due to ambiguities in sequence, suggesting that some property of this particular sample made it more difficult to sequence. Even so, each of the resulting barcodes were 122 bases (18.6%) different from the next closest sample (FDA-22.1), so without hand-editing some degree of ambiguity was acceptable. The number of sequence ambiguities that will be acceptable while attempting to discriminate species will depend on the taxonomic group being tested, which will require further testing of closely related species and will be the subject of subsequent publications. As a general rule, we consider ≤2% sequence ambiguities acceptable in most cases.

In contrast, samples FDA-22.1 and FDA-22.2 represented the same species of fish. Here, slight individual variation can be seen as well (Figure 1), but within the four replicates of each individual, there is no sequence variation, while between individuals there are four bases consistently different (a 0.61% difference). Even with this subtle variation, the differences between these two individuals and the next closest individual were clearly evident (122 bases, 18% difference compared to sample FDA-21), making them easily distinguishable. It must be pointed out that these observations are based on a small sample size containing, apparently, distantly related species. Which commercial fish species DNA barcodes can and cannot differentiate will have to be determined on a group-by-group basis. For regulatory purposes, the level of species or group resolution required can vary greatly depending on the nature of the investigation. These studies are ongoing at FDA and will be reported separately.

Summary and Conclusions

Based on the findings of this SLV study, we have determined that the described method for DNA barcode

generation in fish can be used to produce consistent barcode sequences for different species in a reliable and repeatable manner. This method has several advantages over the current accepted regulatory method of protein IEF. The barcoding method is based on DNA, which is more stable than proteins, and works on a larger variety of fish products, such as dried, salt-cured, smoked, stewed, etc. (1, 12, 28, 30, 31, FDA unpublished data). Even products subjected to extremes in processing, such as canning, which utilizes both high temperatures and pressures, can often still yield short (100–200 bp) barcodes that can be used to help identify the product in some cases (13, FDA unpublished data). In further contrast to IEF, barcoding removes the requirement for perishable tissue standards, which must not only be run with each sample, but must be regularly replaced as the proteins degrade. Once a barcode is generated, it is an electronic file that never goes "bad" like the frozen tissue standards used for IEF. Perhaps most importantly, since the "authenticated standards" used to ultimately identify fish products in the current IEF method (6) are now electronic files, regulatory standard collections can be shared and combined, both nationally and internationally, as long as they are generated under a standardized set of conditions (32). In collaboration with several U.S. state and federal agencies, as well as academic institutions both domestically and abroad, the FDA is developing a library of regulatory barcode reference sequences for commercial and recreational fish based on vouchered specimens. This collection will be curated permanently at the Smithsonian Institution National Museum of Natural History. The criteria used for the development of these "authenticated standard" barcodes as well as potential applications for this library will be the subject of subsequent publications.

Acknowledgments

We would like to thank Karen Blickenstaff (FDA Center for Veterinary Medicine) for instruction and support on the ABI 3730 sequencer; Jeff Williams and Jerry Finan (Fish Division of NMNH) for fish authentication; Errol Strain (FDA Center for Food Safety and Applied Nutrition) for his help with data analysis; and Yolanda Jones (FDA Center for Veterinary Medicine) for sample preparation. The pilot studies carried out at CCDB were supported by grants to P.D.N.H. from Genome Canada through the Ontario Genomics Institute, and Natural Sciences and Engineering Research Council of Canada. We also acknowledge the laboratory assistance of Heather Braid during pilot studies conducted in the Department of Integrative Biology at the University of Guelph, with grant support to R.H.H. from Advanced Foods and Materials Network.

References

(1) Cohen, N.J., Deeds, J.R., Wong, E.S., Hanner, R.H., Yancy, H.F., White, K.D., Thompson, T.M., Wahl, M., Pham, T., Guichard, F.M., Huh, I., Austin, C., Dizikes, G., & Gerber, S. (2009) *J. Food Prot.* **72**, 810–817

(2) Jacquet, J.L., & Pauly, D. (2008) *Mar. Policy* **32**, 309–318

(3) Ogden, R. (2008) *Fish Fish.* **9**, 462–472

(4) *Government Accountability Office Report GAO-09-258* (2009) http://www.gao.gov/ new.items/d09258.pdf

(5) Tenge, B.J, Dang, N.L., Fry, F.S., Savary, W.E., Rogers, P.L., Barnett, J.D., Hill, W.E., Wiskerchen, J.E., & Wekell, M.M. (1997) in *Fish Inspection, Quality Control, and HACCP - A Global Focus*, R.E. Martin, R.L. Collette, & J.W. Slavin (Eds), Technomic Publishing, Lancaster, PA, pp 214–226

(6) *Official Methods of Analysis* (1980) 35.1.41 AOAC INTERNATIONAL, Gaithersburg, MD, Method **980.16**

(7) Yancy, H.F., Zemlak, T.S., Mason, J.A., Washington, J.D., Tenge, B.J., Nguyen, N.T., Barnett, J.D., Savary, W.E., Hill, W.E., Moore, M.M., Fry, F.S., Randolph, S.C., Rogers, P.L., & Hebert, P.D.N. (2008) *J. Food Prot.* **71**, 210–217

(8) Hebert, P.D.N., Cywinska, A., Ball, S.L., & deWaard, J.R. (2003) *Proc. Biol. Sci*. **270**, 313–321

(9) Hebert, P.D.N., Ratnasingham, S., & deWaard, J.R. (2003) *Proc. Biol. Sci.* **270**, Suppl. 1, S96–S99

(10) Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R., & Hebert, P.D.N. (2005) *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **360**, 1847–1857

(11) Hubert, N., Hanner, R., Holm, E., Mandrak, N.E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J., & Bernatchez, L. (2008) *PLoS One* **3**, (6) e2490

(12) Wong, E.H.K., & Hanner, R. (2008) *Food Res. Int.* **41**, 828–837

(13) Hajibabaei, M., deWaard, J.R., Ivanova, N.V., Ratnashingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M., & Hebert, P.D.N. (2005) *Philos. Trans. R. Soc. B. Biol. Sci.* **360**, 1959–1967

(14) Ivanova, N.V., Zemlak, T.S., Hanner, R.H., & Hebert, P.D.N. (2007) *Mol. Ecol. Notes* **7**, 544–548

(15) Ivanova, N.V., & Grainger, C.M. (2007) *CCDB protocols*, http://www.dnabarcoding.ca/CCDB_DOCS/CCDB_Amplification.pdf

(16) Baldwin, C.C., Mounts, J.H., Smith, D.G., & Weigt, L.A. (2009) *Zootaxa* 1–22

(17) Ewing, B., Hillier, L., Wendl, M.C., & Green, P. (1998) *Genome Res*. **8**, 175–185

(18) *ABI User Bulletin Part No. 4362968 Rev. B* (2007) Carlsbad, CA

(19) Robins, R.C., Bailey, R.M., Bond, C.E. Brooker, J.R., Lachner, E.A., Lea, R.N., & Scott, W.B. (1993) *Common and Scientific Names of Fishes from the United States and Canada*, 5th Ed., American Fisheries Society Special Publication 20, AFS, Bethesda, MD

(20) *The Seafood List: FDA Guide to Acceptable Market Names for Food Fish Sold in Interstate Commerce* (2008) Office of Food Safety, Division of Seafood Safety, FDA Center for Food Safety and Applied Nutrition, U.S. Department of Health and Human Services, http://www.fda.gov/Food/GuidanceComplianceRegulatoryInformation/GuidanceDocuments/Seafood/ucm113260.htm

(21) *ABI User Bulletin Part No. 4337035 Rev. A* (2002) Carlsbad, CA

(22) Drummond, A.J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T., & Wilson, A. (2009) *Geneious v4.7*, http://www.geneious.com/

(23) Saitou, N., & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425

(24) Guindon, S., & Gascuel, O. (2003) *Syst. Biol*. **52**, 696–704

(25) Ronquist, F., & Huelsenbeck, J.P. (2001) *Bioinformatics* **17**, 754–755

(26) Ratnasingham, S., & Hebert, P.N.D. (2007) *Mol. Ecol. Notes* **7**, 355–364

(27) Kerr, K.C.R., Birks, S.M., Kalyakin, M.V., Red'kin, Y.A., Koblik, E.A., & Hebert, P.D.N. (2009) *Front. Zool*. **6**, 1–13

(28) Lowenstein, J.H., Amato, G., & Kolokotronis, S.-O. (2009) *PLoS One* **4**, (11) e7866

(29) Casiraghi, M., Labra, M., Ferri, E., Galimberti, A., & De Mattia, F. (2010) *Briefings Bioinf*., doi:101093/bib/bbq004

(30) Holmes, B.H., Steinke, D., & Ward, R.D. (2009) *Fish. Res*. **95**, 280–288

(31) Smith, P.J., McVeagh, S.M., & Steinke, D. (2008) *J. Fish Biol*. **72**, 464–471

(32) Ward, R.D., Hanner, R., & Hebert, P.D.N. (2009) *J. Fish Biol*. **74**, 329–356