

## TECHNICAL NOTE

# High-throughput species identification: from DNA isolation to bioinformatics

DAVID E. RICHARDSON, JEFFREY D. VANWYE, AMY M. EXUM, ROBERT K. COWEN and DOUGLAS L. CRAWFORD

*Marine Biology and Fisheries Division, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149-1098, USA*

## Abstract

Ichthyoplankton collections provide a valuable means to study fish life histories. However, these collections are greatly underutilized, as larval fishes are frequently not identified to species due to their small size and limited morphological development. Currently, there is an effort underway to make species identification more readily available across a broad range of taxa through the sequencing of a standard gene. This effort requires the development of new methodologies to both rapidly produce and analyse large numbers of sequences. The methodology presented in this paper addresses these issues with a focus on the larvae of large pelagic fish species. All steps of the methodology are targeted towards high-throughput identification using small amounts of tissue. To accomplish this, DNA isolation was automated on a liquid-handling robot using magnetic bead technologies. Polymerase chain reaction and a unidirectional sequencing reaction followed standard protocols with all template cleanup and transferring also automated. Manual pipetting was thus reduced to a minimum. A character-based bioinformatics program was developed to handle the large sequence output. This program incorporates base-call quality scores in two types of sample to voucher sequence comparisons and provides suggested identifications and sequence information in an easily interpreted spreadsheet format. This technique when applied to tuna and billfish larvae collected in the Straits of Florida had an 89% success rate. A single species (*Thunnus atlanticus*) was found to dominate the catch of tuna larvae, while billfish larvae were more evenly divided between two species (*Makaira nigricans* and *Istiophorus platypterus*).

*Keywords:* DNA barcoding, Istiophoridae, larval fish identification, magnetic bead DNA isolation, Scombridae, *Thunnus*

*Received 4 July 2006; revision accepted 15 October 2006*

## Introduction

Nearly all marine teleost fishes possess a pelagic larval stage that is morphologically distinct from the adults and many orders of magnitude smaller (Kendall *et al.* 1984). Understanding the processes affecting both the survival and transport of this stage is one of the principal challenges in marine fish ecology, as these processes will influence the spatial distribution, population dynamics, migration strategies and evolution of a species (Bakun 1996; Cowen

*et al.* 2006). While many indirect methods have been developed over the years to evaluate larval ecology and transport (e.g. modelling, recruitment/environmental time series, otolith-based studies on juvenile fishes), a comprehensive understanding of these issues still requires the sampling of eggs and larvae in their natural environment. In addition, ichthyoplankton collections are a powerful tool for addressing many other important questions in fish ecology and fisheries management including the identification of spawning locations and seasons of migrating species (Schmidt 1922; McCleave 1993; Govoni *et al.* 2003; Serafy *et al.* 2003), the quantification of population levels or biomass of fished species (Hunter & Lo

Correspondence: David E. Richardson, Fax: (305) 421-4600; E-mail: drichardson@rsmas.miami.edu

## 2 TECHNICAL NOTE

1993; Scott *et al.* 1993; Ralston *et al.* 2003), and the determination of the distribution and diversity of rare and cryptic species (Smith 2002; Richardson & Cowen 2004; Wouthuyzen *et al.* 2005).

Despite considerable effort (e.g. Moser 1996; Leis & Carson-Ewart 2000; Richards 2006a), larval identification remains one of the major factors limiting the questions that can be addressed by field-based ichthyoplankton studies. These limitations are especially acute within certain regions, for certain taxonomic groups and during the earliest stages of larval development (Kendall *et al.* 1984). For example, Kendall & Matarese (1994) calculated that a description of any stage of larval development existed for only 10% of Indo-Pacific fish species. Similarly, problems remain in identifying the larvae of many well-studied and economically valuable species such as grouper (Richards *et al.* 2006), tuna (Richards 2006b) and billfish (Richards & Luthy 2006). While increased effort in developing morphological identification systems will likely be successful for many taxonomic groups of fishes, it appears that the underlying limits have been reached for many others. Notably, similar limitations of morphological identification systems for early life stages are widespread in the animal kingdom, occurring with a variety of marine and terrestrial invertebrates (Medeiros-Bergen *et al.* 1995; Kiesling *et al.* 2002) and amphibians (Vences *et al.* 2005).

Over the past years, molecular means of identification have increased in use in response to the limitations of morphological identification. For fishes, these techniques have most frequently involved the use of polymerase chain reaction (PCR) followed by a restriction enzyme digestion (McDowell & Graves 2002; Chow *et al.* 2003; Luthy *et al.* 2005) or the more efficient utilization of species-specific primers in a single PCR step (Rocha-Olivares 1998; Shivji *et al.* 2002; Chapman *et al.* 2003; Hyde *et al.* 2005). Each of these methodologies has been successfully used in ichthyoplankton studies, with the latter technique even being implemented with basic laboratory equipment on board a research vessel (Hyde *et al.* 2005). While designed for ease of use in modestly equipped laboratory settings, these techniques face a critical drawback of only being suitable for discriminating a limited number (generally five to 15) of species determined at the time of development. Because of this, these two methodologies are generally appropriate for studies targeted at one or two low-diversity taxonomic groups, but are not applicable for the identification of a wide diversity of species.

It has been proposed that the sequence of a single gene may be suitable for discriminating most animal species (Hebert *et al.* 2003a, b). This concept, termed DNA barcoding, has been tested with North American bird species (Hebert *et al.* 2004), Australian fish species (Ward *et al.* 2005), and a variety of invertebrate taxa (Barrett & Hebert 2005; Janzen *et al.* 2005; Smith *et al.* 2005). The conclusions

of these studies, using the cytochrome oxidase subunit 1 (*cox1*) gene, are that sufficient interspecific differences exist for species identification. While most of the focus to date has been on developing the foundations of DNA barcoding, the technique has proven useful in a number of studies. For example, ant diversity in Madagascar was evaluated through DNA barcoding, an otherwise tedious and difficult task using traditional morphological identification methods (Smith *et al.* 2005). Sequencing has also been used in ichthyoplankton work to verify the identity of morphologically distinct larval types (Hare *et al.* 1994). The primary appeal of using a sequence-based identification approach, in these and a variety of other ecological studies, is that a single methodology using a limited amount of tissue can be used to discriminate a broad range of species. Historically, the drawback to the large-scale implementation of sequence-based identification has been the time and cost requirements per sample.

Here we present a methodology that capitalizes on recent advances in molecular techniques and equipment to make sequence-based identification suitable for a large-scale larval fish study. Specifically, this methodology is (i) high throughput (> 800 individuals/week), (ii) highly automated, (iii) suitable for work with small amounts of tissue, (iv) easily modified for use across a broad taxonomic range of species, and (v) provides results readily interpreted with a simple bioinformatics approach. We show the utility of these techniques for a few taxa of great economic importance throughout the tropics, tunas (family: Scombridae) and billfish (family: Istiophoridae), and present preliminary data on the species composition of these taxa in the Straits of Florida. This study supports the scaling up of barcode databases, and demonstrates that sequence-based identification is sufficiently advanced to be applied in ecological studies requiring the identification of large numbers of individuals.

## Materials and methods

### *Collections of larvae and voucher tissue*

Larval fishes used in this study were collected in the Straits of Florida as a part of a project focused on the larval ecology and spawning strategies of billfishes and other pelagic fish species. Larvae were collected with a coupled asymmetrical MOCNESS (multiple opening and closing net and environmental sensing system) equipped to simultaneously trigger a 4-m<sup>2</sup> 800 µm mesh net and a 1-m<sup>2</sup> 150 µm mesh net (Guigand *et al.* 2005) and a combined neuston net (1 × 2 m 1000 µm mesh net attached to a 0.5 × 1 m 150 µm mesh net). Samples were immediately fixed in 95% ethanol. They were placed in 70% ethanol for long-term storage 2–10 days following collection. Samples were genetically identified 1–3 years after collection.

All larval fish were separated from the remaining plankton and identified morphologically to lowest possible taxonomic category. These morphological identifications generally ranged from family to genus level. Larval fishes were measured using either the IMAGE PRO ANALYSIS software (Media Cybernetics) or an ocular micrometer, and ranged in size from 2.5 to 13 mm SL. A majority of these were at the lower end of the size range, reflective of the typical pattern of abundance at size for fishes collected in ichthyoplankton studies. The eyeball and/or tail were removed from each larval fish for genetic analysis. Tissue removal sought to avoid damage to the otoliths (ear bones located in the skull) and guts of the fish, so that each individual could also be used in growth and diet studies.

Voucher sequences for each species were obtained from adult tissue from a variety of sources. Istiophorid voucher specimens (four species: *Istiophorus platypterus*, *Makaira nigricans*, *Tetrapturus albicans*, *T. pfluegeri*) were collected from fishes tagged by the NOAA Southeast Fisheries Science Center billfish tagging program in the western North Atlantic. *Thunnus* adult voucher specimens (five species: *T. atlanticus*, *T. albacares*, *T. thynnus*, *T. obsesus*, *T. alalunga*) were collected by NOAA fishery observers stationed onboard long-line vessels in the Gulf of Mexico. Additional sequences (Finnerty & Block 1995; Chow *et al.* 2003; Hyde *et al.* 2005) of all these species were obtained from GenBank for the cytochrome *b* (*cyt b*) gene used in this study. These additional sequences resulted in interocean differences being included in the voucher sequences for those species occupying both oceans. For each species, two to five sequences of newly collected specimens from the western North Atlantic study region and one to five additional GenBank sequences were used to create the voucher sequences. Sequences (Ward *et al.* 2005) of the *cox1* gene were also obtained from GenBank for comparative purposes. For each species, a single sequence containing polymorphic sites was generated from all the voucher sequences of that species.

#### DNA isolation

All reactions were performed in 96-well plates. Tissue was not added to three of the wells to serve as negative controls for the DNA isolation. These negative controls also functioned as secondary identifiers of the orientation and numbering of the plates. An additional two wells were not included in the isolation procedure, and instead were used as a positive and negative control for the PCR.

DNA isolation was automated on the Evolution P<sup>3</sup> liquid-handling robot (Perkins Elmer) and used the Genfind kit (Agencourt). This kit functions by the absorption of DNA to magnetic beads in the presence of the provided buffers, followed by the removal of contaminants, and then the release of DNA from the beads in the

presence of aqueous buffers. Prior to isolation, tissue samples were placed in a 30  $\mu$ L lysis buffer (containing 24  $\mu$ L Genfind lysis buffer, 3  $\mu$ L 18 mg/mL proteinase-K and 3  $\mu$ L 1 M DTT) and were shaken for 6–16 h at 57 °C and 1400 r.p.m. on a Thermomixer (Eppendorf). Convex-domed plastic tops (Bio-Rad) were used to seal each well individually and prevent cross-contamination unlike other sealing methods. After the samples were digested, the plates were centrifuged for 5 min at 3220 g to prevent cross-contamination during the removal of these tops.

The DNA isolation protocol for the liquid-handling robot can accommodate up to eight plates (768 samples) at a time. It requires approximately 1.5 h of unattended robotic handling time and 30 min of manual labour (excluding the placing of tissue samples in the plates) regardless of the number of plates. The first step in the protocol was the addition to each well of 30  $\mu$ L of Genfind lysis buffer, followed by 30  $\mu$ L of Genfind binding buffer (containing the magnetic beads) and the transfer of the plate to the stacker for 5 min. The plates were then moved to a magnet on the work platform for an additional 5 min and the liquid was removed, leaving the DNA bound to the beads in each well. The following steps were then repeated twice: (i) the plate was removed from the magnet, (ii) 200  $\mu$ L of Genfind protein wash was added, (iii) the plate remained off the magnet for 5 min before being transferred to the magnet for an additional 5 min, and (iv) the liquid was removed. Further purification was achieved (with the plate on the magnet) by washing the DNA bound to the beads four times with 100  $\mu$ L of 70% EtOH. The beads were then allowed to dry for a 15-min period before the DNA was eluted in 50  $\mu$ L of 0.5 mM Tris pH 7.5. During this elution, the plate was removed from the magnet for 10 min and then placed on the magnet for 5 min before the purified DNA was transferred to a new 96-well plate.

In order to test the effectiveness of the automated DNA isolation procedure, the yield and quality of DNA obtained from using this technique was compared to a standard ammonium acetate DNA isolation (Sambrook & Russell 2001) across a wide size range of larvae. From 20 larval tuna, DNA was isolated from one eyeball using the automated technique and the other eyeball using the ammonium acetate DNA isolation. Yield and quality of DNA were evaluated spectrophotometrically and by gel electrophoresis, respectively.

#### PCR and sequencing reactions

PCR amplification of a 715-bp fragment of the *cyt b* gene used the primer pair CBF-A (5'-CCCTCTAATATCTCQGTCTGATGAAA-3') and CB3-3 (5'-GCGTAGGCAAATAGGAARTATCAYTC-3' (modified from Palumbi 1996). The reactions had a final volume of 50  $\mu$ L and contained 12  $\mu$ L of the DNA template, 2 mM dNTP, 20 pmoles of each

#### 4 TECHNICAL NOTE

primer, 4  $\mu\text{L}$  of *Taq*, and a reaction buffer (final concentrations: 50 mM tris HCl pH 9.2, 16 mM  $(\text{NH}_4)_2\text{SO}_4$ , 2.25 mM  $\text{MgCl}_2$ , 2% DMSO, 0.1% Tween 20) (Paschall *et al.* 2004). Template DNA concentrations ranged from not detectable ( $< 1 \text{ ng}/\mu\text{L}$ ) for the smallest larvae to approximately 100  $\text{ng}/\mu\text{L}$  for the largest. Manual pipetting was used to make and aliquot the PCR mix, and then the liquid-handling robot was used to transfer DNA to the PCR plate. The thermocycling protocol was 50 cycles of 94 °C for 15 s, 55 °C for 30 s and 72 °C for 1 min, followed by 5 min at 72 °C and utilized a DNA tetrad Thermocycler (Bio-Rad). Samples were electrophoresized on a 1% agarose gel to verify results. PCR products were then purified using the standard AmPure kit (Agencourt) protocol on the liquid-handling robot, but with a 1:1 ratio of magnetic beads to PCR volume. For nonvoucher specimens, sequencing was unidirectional using 3.2 pmol of CBF-A primer, 5  $\mu\text{L}$  of BigDye Terminator (Applied Biosystems), reaction buffer diluted to 1 $\times$  and 2  $\mu\text{L}$  of template in a 10  $\mu\text{L}$  reaction. The sequencing thermocycling protocol consisted of 40 cycles at 94 °C for 15 s, 50 °C for 20 s and 60 °C for 4 min. For voucher specimens, both primers were used in bidirectional sequencing in order to increase the reliability of the voucher sequences. Sequences were purified on the liquid-handling robot using the standard CleanSeq kit protocol (Agencourt) and then run on a 3730 xl DNA Analyser (Applied Biosystems).

#### Sequence analysis

Sequence analysis was performed in MATLAB (MathWorks) using procedures in the publicly available MBEToolbox (Cai *et al.* 2005) and a custom-written script (available from drichardson@rsmas.miami.edu). A single run of this script produces the identifications and statistics for an entire 96-well plate. The input files are the \*.phd.1 files from the sequencer. One critical component of these files is an estimated probability of error of each base-call in the sample sequence. This value is termed the Phred score and is defined by the expression  $q = -10 * \log(p)$ , where  $p$  is the probability of error (e.g. a Phred score of 30 corresponds to a probability of error of  $10^{-3}$ ). A variety of parameters in the trace data of the DNA sequencer are used to calculate the Phred score, with empirical tests revealing that these scores are very accurate in predicting the probability of base-call errors (Ewing & Green 1998).

The script for species identification allows the user to choose a set of voucher sequences (e.g. all istiophorids, all scombrids), based on the taxonomic resolution of the morphological identification of the larvae. Sample sequences in groups of eight and all the voucher sequences are aligned using the alignSeqFile procedure in the MBEToolbox. Sample nucleotides that do not align with nucleotides in the voucher sequences are removed. Additionally, only

nonpolymorphic sites in the voucher sequences are used in all of the comparisons. The Phred score assigned to each sample nucleotide is the lower of the nucleotide specific Phred score and a five-nucleotide moving average Phred score.

Two sequence comparisons are used for species identification. The first is a full-sequence length comparison to each voucher sequence. As a measure of reliability, the number of correctly and incorrectly paired nucleotides are determined in each of three Phred score categories (1–20, 21–40 and 41+) for each voucher sequence to sample sequence pair. The second comparison uses only informative diagnostic characters. To obtain these informative sites, voucher sequences from each possible species pair are compared to extract the nucleotide positions that differ between the two species. At these informative sites, the sample sequence (Phred score  $> 20$ ) is compared to both of the voucher sequences. The number of nucleotides in the samples' sequence that matches each of the voucher sequences at these informative sites is quantified. The end result of this step of the script is an  $n \times n$  ( $n$  = number of voucher sequences) matrix of the number of matching diagnostic nucleotides obtained from each comparison between voucher sequences.

The results from these two comparisons are used to generate a most likely identification of the sample. Criteria for species identification (across the portion of the sample sequence with Phred scores  $> 20$ ) are: (i)  $> 300$  bp in length, (ii) less than a 5% difference with the voucher sequence, and (iii)  $> 80\%$  of informative nucleotides matching the identified species for every possible comparison with another species. These criteria are readily changed in the script, and can be tailored to different identification needs. For this specific study, more stringent values were chosen for the informative nucleotide comparison than the total sequence match criteria. This was done because each larva could be morphologically identified to a narrow set of species, all of which were included in the voucher sequence file. The implementation of more conservative total sequence match criteria would be desired in instances where the unknown sequence may represent a species not included in the voucher sequence file.

The final output of the script is a comma-delimited text file containing the results of these informative nucleotide and total sequence match comparisons for all 96 sequences. This file can be readily opened in Microsoft Excel and other spreadsheet programs. This allows for the manual inspection of sequence quality, comparison data, and suggested identifications. An example of this output from the larvae of two closely related species is shown in Table 1.

#### Results

A total of 493 tuna and billfish larvae were subjected to sequence analysis, with 438 being successfully sequenced

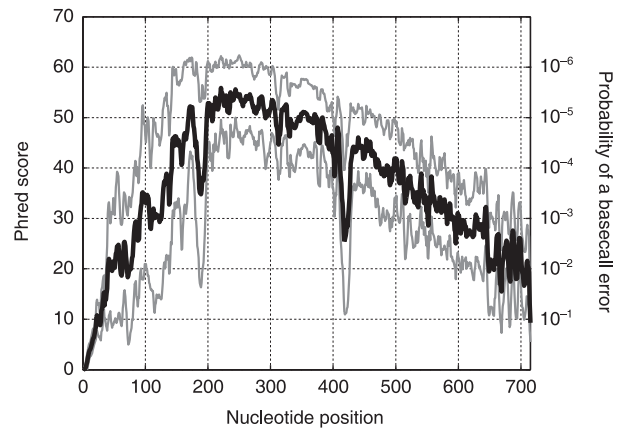
**Table 1** Output of the bioinformatics script for identifying individuals to species. For each sample two comparisons are made. The portion on the left contains the results of the comparison of the entire sample sequence against each of the voucher sequences, with the results (# wrong/total) broken down by Phred scores. The right box contains the results of comparisons at informative nucleotides. In parentheses is the total number of informative nucleotides that separate the two species. The first number is the percentage of informative nucleotides in the sample sequence matching the species in the row heading vs. the species in the column heading

Species	Phred score			Total (> 20)	Species	<i>T. atlanticus</i>	<i>T. albacares</i>	<i>T. thynnus</i>	<i>T. obesus</i>	<i>T. alalunga</i>
	0–20	21–40	41+							
<b>Sample: G2_C10</b>					<b>Most likely ID: <i>T. albacares</i></b>					
<i>T. atlanticus</i>	23/62	3/179	4/431	7/610	<i>T. atlanticus</i>	—	0% (3)	56% (9)	60% (5)	77% (13)
<i>T. albacares</i>	23/56	0/135	0/434	0/569	<i>T. albacares</i>	100% (3)	—	100% (7)	100% (5)	100% (11)
<i>T. thynnus</i>	24/69	4/201	7/434	11/635	<i>T. thynnus</i>	44% (9)	0% (7)	—	25% (4)	67% (15)
<i>T. obesus</i>	23/56	0/134	6/428	6/562	<i>T. obesus</i>	40% (5)	0% (5)	75% (4)	—	78% (9)
<i>T. alalunga</i>	24/69	5/200	10/431	15/631	<i>T. alalunga</i>	23% (13)	0% (11)	33% (15)	22% (9)	—
<b>Sample: G2_D10</b>					<b>Most likely ID: <i>T. atlanticus</i></b>					
<i>T. atlanticus</i>	19/49	2/119	0/503	2/622	<i>T. atlanticus</i>	—	100% (3)	89% (9)	100% (5)	100% (13)
<i>T. albacares</i>	19/47	0/74	5/503	5/577	<i>T. albacares</i>	0% (3)	—	71% (7)	60% (5)	91% (11)
<i>T. thynnus</i>	20/51	5/146	7/506	12/652	<i>T. thynnus</i>	11% (9)	29% (7)	—	25% (4)	73% (15)
<i>T. obesus</i>	19/47	0/74	6/496	6/570	<i>T. obesus</i>	0% (5)	40% (5)	75% (4)	—	89% (9)
<i>T. alalunga</i>	19/51	6/146	13/502	19/648	<i>T. alalunga</i>	0% (13)	9% (11)	27% (15)	11% (9)	—

and identified on the first try. The overall success rate for identifying these samples was 89%. This success rate did not vary between the billfish and tunas (chi-test  $P = 0.54$ ). The size of the larvae analysed did not play a role in the success rate of the technique (chi-test  $P = 0.40$ ). Low-quality sequences were obtained for the blanks from one of six plates. These sequences matched a taxonomic group not included on the plate, indicating that one of the reagents used in the technique was contaminated, but that well-to-well contamination did not occur. All of the other 20 blanks did not produce any sequence.

Total DNA yields ranged from not measurable (< 50 ng) to 9834 ng for the 20 larval tuna eyeballs used to test the magnetic bead extraction technique. In all but one of the 20 paired eyeball comparisons, the amount of DNA yielded using the GenFind kit automated on the liquid-handling robot was lower than the amount yielded using the manual ammonium acetate DNA isolation. On average, the GenFind kit yielded 43% as much DNA as the ammonium acetate DNA isolation. The quality of the DNA yielded by the two types of extractions appeared similar on an agarose gel. The extent to which the automated DNA yields were reduced relative to the ammonium acetate DNA yields did not appear to vary across the size range of larvae used, nor did it negatively impact the identification rate of these larvae.

The use of this technique resulted in high quality sequences. On average, 355 ( $\pm 93$ ) base calls had a Phred score > 40, 449 ( $\pm 82$ ) had a Phred score > 30, and 563 ( $\pm 61$ ) had a Phred score > 20 in successfully identified samples. Low-quality Phred scores generally occurred at the beginning of the



**Fig. 1** Average (black)  $\pm$  standard deviation (grey) of a 5-bp moving average of Phred scores and the associated probabilities of base call errors over the 715 bp sequence of successfully identified *Thunnus* and *Istiophorid* larvae ( $n = 311$ ).

sequence, the end of the sequence and at two locations within the sequence (Fig. 1).

Interspecific differences in the *cyt b* gene were sufficient for the identification of all of the species we tested. The number of diagnostic characters available to separate *Thunnus* species were low (3–15) but sufficient for species-level identifications (Table 2). For the billfishes, many more diagnostic characters (15–25) existed to separate species. Polymorphic sites ranged from three to 13 for billfish and two to 11 for the *Thunnus* species. These values included interocean differences for most species. As a

	<i>T. atlanticus</i>	<i>T. albacares</i>	<i>T. obsesus</i>	<i>T. alalunga</i>	<i>T. thynnus</i>	
<i>T. atlanticus</i>		3	5	13	9	cyt <i>b</i>
<i>T. albacares</i>	2		5	11	7	
<i>T. obsesus</i>	4	2		9	4	
<i>T. alalunga</i>	9	9	9		15	
<i>T. thynnus</i>	6	4	4	9		
			cox1			

**Table 2** Number of diagnostic characteristics separating Atlantic *Thunnus* species with the cytochrome *b* gene sequenced in this study (above the diagonal) and the cytochrome oxidase subunit 1 gene (below the diagonal). Both sequences are similar in length

**Table 3** Species composition of billfishes and tuna in the Straits of Florida obtained using morphological identification techniques vs. the molecular methodology presented in this study. The molecular identification numbers represent only those individuals which could not be identified to species through morphological means

Taxonomic group	Morphological identification (%)	Molecular identification (%)
Istiophoridae unknown	46	12
<i>Istiophorus platypterus</i>	42	18
<i>Makaira nigricans</i>	12	69
<i>Tetrapturus albidus</i>	—	1.3
<i>Tetrapturus pfluegeri</i>	—	0
Scombridae		
<i>Thunnus</i> unknown	100	10
<i>T. atlanticus</i>	—	81
<i>T. albacares</i>	—	12
<i>T. thynnus</i>	—	0.8
<i>T. obsesus</i>	—	0
<i>T. alalunga</i>	—	0

655-bp portion of the *cox1* gene has been sequenced for the *Thunnus* species, it was possible to make a between gene comparison of the number of diagnostic characteristics separating each species pairs. For all but one comparison the *cyt b* gene contained more diagnostic characteristics than the *cox1* gene (Table 2). These differences were generally relatively small. Polymorphic sites in the *Thunnus cox1* gene based on five specimens for each species were lower (0–3) than in the *cyt b* gene.

Versus morphological identification techniques, the larval composition of large pelagic species in our study area was more finely resolved with the molecular methodology (Table 3). In addition to pinpointing which species were most abundant as larvae in the Straits of Florida, the samples identified to date also indicate that some species which occur as adults in the region do not occur as larvae. This latter conclusion could not be reached with morphological identification techniques due to the large numbers of unidentifiable larvae. For the istiophorids, morphological and molecular techniques indicated a much different species composition. This is explained by the fact that sailfish can be morphologically identified to species over

a much broader size range than blue marlin, and is thus overrepresented in the species composition values.

## Discussion

The inability to identify individuals to species is one of the major factors limiting the questions that can be addressed in many ecological studies, including those focused on the early life stages of fishes. In this study, we have demonstrated a high-throughput species identification technique that should prove to be a critical tool in elucidating the life–history strategies and interactions of a wide range of species. To make this technique both time- and labour-efficient, we automated DNA isolation and other laboratory steps and reduced manual pipetting to the making and aliquoting of three master mixes (Lysis Buffer, PCR mix, sequencing mix). This automation allows multiple plates to be run simultaneously with only minimal increases in total time requirements, while in turn reducing the potential for pipetting errors and the misplacement of samples. In practice, the factor limiting the number of samples that can be identified, and the most labour-intensive portion of the technique, is the placement of tissue samples in individual wells. Importantly, the DNA isolation technique we used, while not as efficient in yielding DNA as traditional isolation techniques, did produce enough DNA for multiple reactions from even the smallest tissue samples. This makes the technique suitable for identifying small organisms, such as larval fishes, without compromising the integrity of the sample for use in additional studies.

In addition to requiring minimal labour, this technique was also very successful. The 89% success rate of the technique on the first try with larvae was higher than the 68% rate for billfish identification using an RFLP (restriction fragment length polymorphism) technique (Luthy *et al.* 2005), but lower than the 95% rate using a species-specific multiplex PCR assay (Hyde *et al.* 2005). Preservation methods used with the larvae may be the primary sources of these differences, as Luthy *et al.* (2005) included some larvae preserved in BHT-buffered ethanol, while Hyde *et al.* (2005) worked with larvae just after their collection. This 89% success rate was also equal to or better than five different DNA isolation methods evaluated using fish tissue samples (Hajibabaei *et al.* 2005). Within-plate

cross-contamination found in early runs was eliminated by switching to convex-domed plastic tops during the lysis buffer digestion and centrifuging the samples to remove condensation on the tops prior to their removal. However, an undetermined contaminated reagent was used in one plate of reactions. This resulted in low-quality sequences of a nontarget species in the blank wells. These contaminant sequences were readily identified during analysis. Highlighted by this particular contaminated run is the importance of the extensive control process used with this methodology.

The bioinformatics approach we used proved effective at species identification and reduced the time requirements of the technique in two different ways. The first is that the preprocessing of the sequencer data is not required; rather a single run of the MATLAB script automatically inputs all the files from a run of a 96-well plate and provides a readily interpreted output. The second is that by incorporating Phred scores into the analysis, less accurate, but more efficient unidirectional sequencing can be used. Assuming a comprehensive sequence library has been developed, misidentifications from the method described here, would require multiple sequencing errors at informative sites, an unlikely scenario given the Phred score cutoff of 20 ( $P = 0.01$  of an error for each nucleotide) used in the program. As with any analytical approach for species identification, sources of errors or ambiguities in identification are most likely to result from an insufficient number of individuals being used to develop a representative sequence for a species and closely related co-occurring species not being sampled. For the specific application described here, neither of these concerns appeared to be an issue.

Unlike most other studies that use distance- or tree-based methods to match sample sequences to species, this study utilized a character-based method. DeSalle *et al.* (2005) argued for the use of character-based methods for a number of reasons including the similarity to classical taxonomic methods and the ease vs. distance methods in establishing a fail to identify criteria. Additionally, character-based methods are very suitable for use with sequences containing errors, whereas tree-based methods will often group sample sequences containing random errors on more distant branches than species closely related to the one being identified. While this may not affect the suggested identification of tree-based programs, it will make the visualization of these results more ambiguous than the visualization of diagnostic characters. On the other hand, the character-based program used in this study is not designed for use with large voucher sequence databases as are most of the distance methods. This is not a factor in this specific application, as all the larvae can be morphologically identified to at least family level, but will be an issue to address in future modifications of the program.

The development of this methodology was focused on supporting a comparative study of the spawning strategies of large and medium size pelagic fishes. Because of this, voucher specimens were collected from only a limited number of families, and the degree to which this technique is useful for other species was not established. This issue is ultimately dependent upon the universality of primers being used and the extent of between species differences in sequences. Both these issues were addressed in the sequencing of the *cox1* gene from 207 Australian fish species (Ward *et al.* 2005). In that study, interspecific differences were sufficient for the identification of all the species, while combinations of two forward primers and two reverse primers were sufficient for successful PCR amplification. Of all the genera that Ward *et al.* (2005) considered, *Thunnus* contained the least between-species differences, and thus represents a critical test of the suitability of a gene for identification purposes. Our comparison indicates that the *cyt b* gene provides slightly more characters for species identification than the *cox1* gene, but also contains more polymorphic sites. Despite these differences, both genes can equally well be applied to species identification in the taxa considered here. Currently, the *cox1* gene is being set as the standard for species identification with large sequence databases being generated. Future applications of our methodology with additional species should capitalize on these efforts by switching to the *cox1* gene.

The development of a large sequence database as part of the Barcode of Life project, and more readily utilized sequencing techniques, such as the one described here, will greatly expand the questions that can be addressed through ichthyoplankton studies. Currently, the limitations of established identification systems necessitates that many studies use only family or genus level data, despite consistent indications that the ecology and distribution patterns of fish early life stages differ between even closely related species. As an example, within the wrasse (Labridae) family between-species differences exist in the timing and nature of spawning events (Robertson 1991), pelagic larval durations (Victor 1986) and temporal patterns of settlement (Sponaugle & Cowen 1997). Likewise, highly migratory pelagic species spawn at times and in locations that reflect a unique trade-off between the habitat preferences of adults and the need to place larvae in an area suitable for their survival. The species composition data presented in this study reflects this as some species present in the region as adults were not represented in the larval catches. A move towards species-level identifications in ichthyoplankton studies, especially in the tropics, will undoubtedly reveal many interesting strategies for spawning and surviving the larval stage that had previously been obscured by the available coarse resolution identification methods.

Over the long-term, one of the primary rationales for developing sequence databases and making identifications

more readily available is that these tools will enable us to more comprehensively understand and assess biodiversity (Hebert *et al.* 2003b; Blaxter 2004). Large-scale ichthyoplankton surveys, which commonly collect hundreds of thousands to millions of fishes, may have a unique role in accomplishing this goal. It is well recognized that many traditional means of quantifying diversity in the marine environment greatly underestimate species occupying cryptic habitats (Ackerman & Bellwood 2000) or deep water environments (Grassle & Maciolek 1992). As with nearly all teleost fishes, most species occupying these cryptic and deep-water habitats have a pelagic larval stage that occupies the upper couple hundred metres of the water column and is readily collected in net tows. For Elopomorph fishes (true eels, bonefish, tarpon, etc.) which possess a leptocephalus larval stage that can be identified to species, larval collections have proven to be a strong tool in assessments of regional diversity (Miller 1995; Minagawa *et al.* 2004; Richardson & Cowen 2004; Wouthuyzen *et al.* 2005), and have indicated that many species of fishes remain to be collected and described as adults (Bohlke 1989; Smith 2002). Biodiversity assessments of many other high-diversity taxonomic groups with similarly cryptic species would greatly benefit from the use of plankton tows and a readily utilized means of species identification.

Ultimately, an effective sequence-based identification system will require a well-developed database of high-quality sequences matched to voucher specimens, as well as a rapid means of generating identifications from unknown samples. Here we have demonstrated laboratory techniques that will make populating these databases and generating identifications from unknown samples more efficient. Additionally, we have shown that once quality voucher sequences are established, it is possible to use rapid analytical techniques that also reduce the amount of required laboratory work by allowing for unidirectional sequencing. When combined, these laboratory and analytical techniques make sequence-based identification time- and cost-effective for use in large-scale ecological studies.

### Acknowledgements

Special thanks to D. Williams for valuable advice on the methodologies, to E. Prince, D. Snodgrass and D. Lee for assisting with voucher specimen collections, and to C. Guigand, J. Llopiz, L. Buhmaster, S. Trbovich and the many cruise participants and plankton sorters for all their work in obtaining and processing the plankton samples. This study was supported by grants from the National Science Foundation (OCE-0136132), the Gulf States Marine Fisheries Commission (Billfish-2005-017) and the Large Pelagics Research Center at the University of New Hampshire through the NOAA award NAO4NMF4550391 to R.K.C. and a University of Miami Maytag Fellowship to D.E.R. Samples were collected with a NOAA permit (HMS-EFP-04-02, HMS-SRP-05-03)

and complied with University of Miami animal care protocols (05-134 & 05-135).

### References

- Ackerman JL, Bellwood DR (2000) Reef fish assemblages: a re-evaluation using enclosed rotenone stations. *Marine Ecology Progress Series*, **206**, 227–236.
- Bakun A (1996) *Patterns in the Ocean: Ocean Processes and Marine Population Dynamics*. California Sea Grant College System, La Jolla, California
- Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology*, **83**, 481–491.
- Blaxter ML (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **359**, 669–679.
- Bohlke EB (ed.) (1989) *Fishes of the Western North Atlantic: Leptocephali*. Sears Foundation for Marine Research, New Haven, Connecticut.
- Cai JJ, Smith DK, Xia X, Yuen K (2005) MBEToolbox: a MATLAB toolbox for sequence data analysis in molecular biology and evolution. *BMC Bioinformatics*, **6**, 64.
- Chapman DD, Abercrombie DL, Douady CJ *et al.* (2003) A streamlined, bi-organelle, multiplex PCR approach to species identification: application to global conservation and trade monitoring of the great white shark, *Charcharodon carcharias*. *Conservation Genetics*, **4**, 415–425.
- Chow S, Nohara K, Tanabe T *et al.* (2003) Genetic and morphological identification of larval and small juvenile tunas (Pisces: Scombridae) caught by a midwater trawl in the western Pacific. *Bulletin of Fisheries Research Agency*, **8**, 1–14.
- Cowen RK, Paris CB, Srinivasan A (2006) Scaling connectivity in marine populations. *Science*, **311**, 522–527.
- DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1905–1916.
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, **8**, 186–194.
- Finnerty JR, Block BA (1995) Evolution of the cytochrome *b* in the Scombroidei (Teleostei): molecular insights into billfish (Istiophoridae and Xiphiidae) relationships. *Fishery Bulletin*, **93**, 78–96.
- Govoni JJ, Laban EH, Hare JA (2003) The early life history of swordfish (*Xiphias gladius*) in the western North Atlantic. *Fishery Bulletin*, **101**, 778–789.
- Grassle JF, Maciolek NJ (1992) Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. *The American Naturalist*, **139**, 313–341.
- Guigand CM, Cowen RK, Llopiz JK, Richardson DE (2005) A coupled asymmetrical multiple opening closing net with environmental sampling systems. *Marine Technology Society Journal*, **39**, 22–24.
- Hajibabaei M, deWaard JR, Ivanova NV *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B*, **360**, 1959–1967.
- Hare JA, Cowen RK, Zehr JP, Juanes F, Day KH (1994) Biological and oceanographic insights from larval labrid (Pisces: Labridae) identification using mtDNA sequences. *Marine Biology*, **118**, 17–24.
- Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, 313–321.

- Hebert PDN, Ratnasingham S, deWaard JR (2003b) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, S596–S599.
- Hebert PDN, Stoeckle MY, Zemplak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, 1657–1663.
- Hunter JR, Lo NC-H (1993) Ichthyoplankton methods for estimating fish biomass introduction and terminology. *Bulletin of Marine Science*, **53**, 723–727.
- Hyde JR, Lynn E, Humphreys R *et al.* (2005) Shipboard identification of fish eggs and larvae by multiplex PCR, and description of fertilized eggs of blue marlin, shortbill spearfish, and wahoo. *Marine Ecology Progress Series*, **286**, 269–277.
- Janzen DH, Hajibabaei M, Burns JM *et al.* (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1835–1845.
- Kendall AW, Ahlstrom EH, Moser HG (1984) Early life history stages of fishes and their characters. In: *Ontogeny and Systematics of Fishes* (eds Moser HG, Richards WJ, Cohen DM, Fahay MP, Kendall AW, Richardson SL), pp. 12–22.
- Kendall AW, Matarese AC (1994) Status of early life history descriptions of marine teleosts. *Fishery Bulletin*, **92**, 725–736.
- Kiesling TL, Wilkinson E, Rabalais J *et al.* (2002) Rapid identification of adult and naupliar stages of copepods using DNA hybridization methodology. *Marine Biotechnology*, **4**, 30–39.
- Leis JM, Carson-Ewart BM (eds) (2000) *The Larvae of Indo-Pacific Coastal Fishes: an Identification Guide to Marine Fish Larvae*. Brill, Leiden, The Netherlands.
- Luthy SA, Cowen RK, Serafy JE, McDowell JR (2005) Toward identification of larval sailfin (*Istiophorus platypterus*) white marlin (*Tetrapturus albidus*), and blue marlin (*Makaira nigricans*) in the western North Atlantic Ocean. *Fishery Bulletin*, **103**, 588–600.
- McCleave JD (1993) Physical and behavioural controls on the oceanic distribution and migration of leptocephali. *Journal of Fish Biology*, **43**, 243–273.
- McDowell JR, Graves JE (2002) Nuclear and mitochondrial DNA markers for the specific identification of istiophorid and xiphiid billfishes. *Fishery Bulletin*, **100**, 537–544.
- Medeiros-Bergen DE, Olson RR, Conroy JA, Kocher TD (1995) Distribution of holothurian larvae determined with species-specific genetic probes. *Limnology and Oceanography*, **40**, 1225–1235.
- Miller MJ (1995) Species assemblages of leptocephali in the Sargasso Sea and Florida Current. *Marine Ecology Progress Series*, **121**, 11–26.
- Minagawa G, Miller MJ, Aoyama J, Wouthuyzen S, Tsukamoto K (2004) Contrasting assemblages of leptocephali in the western Pacific. *Marine Ecology Progress Series*, **271**, 245–259.
- Moser HG (ed.) (1996) *The Early Stages of Fishes in the California Current*. CALCOFI Atlas, La Jolla, California.
- Palumbi SR (1996) Nucleic Acids II: the polymerase chain reaction. In: *Molecular Systematics* (eds Hills DM, Moritz C, Mable BK), pp. 205–246. Sinauer Associates, Sunderland, Massachusetts.
- Paschall JE, Oleksiak MF, VanWye JD *et al.* (2004) FunnyBase: a system level functional annotation of *Fundulus* ESTs for the analysis of gene expression. *BMC Genomics*, **5**, 96.
- Ralston S, Bence JR, Eldridge MB, Lenarz WH (2003) An approach to estimating rockfish biomass based on larval production, with application to *Sebastes jordani*. *Fishery Bulletin*, **101**, 129–146.
- Richards WJ (ed.) (2006a) *Early Stages of Atlantic Fishes: an Identification Guide for Western Central North Atlantic*. Taylor & Francis, Boca Raton, Florida.
- Richards WJ (2006b) Scombridae: mackerels & tunas. In: *Early Stages of Atlantic Fishes: an Identification Guide for the Western Central North Atlantic* (ed. Richards WJ), pp. 2187–2229. Taylor & Francis, Boca Raton, Florida.
- Richards WJ, Luthy SA (2006) Istiophoridae: billfishes. In: *Early Stages of Atlantic Fishes: an Identification Guide for the Western Central North Atlantic* (ed. Richards WJ), pp. 2231–2240. Taylor & Francis, Boca Raton, Florida.
- Richards WJ, Baldwin CC, Ropke A (2006) Serranidae: sea basses. In: *Early Stages of Atlantic Fishes: an Identification Guide for the Western Central North Atlantic* (ed. Richards WJ), pp. 1225–1331. Taylor & Francis, Boca Raton, Florida.
- Richardson DE, Cowen RK (2004) Diversity of leptocephalus larvae around the island of Barbados (West Indies): relevance to regional distributions. *Marine Ecology Progress Series*, **282**, 271–284.
- Robertson DR (1991) The role of adult biology in the timing of spawning of tropical reef fishes. In: *The Ecology of Fishes on Coral Reefs* (ed. Sale PF), pp. 356–386. Academic Press, San Diego, California.
- Rocha-Olivares A (1998) Multiplex haplotype-specific PCR: a new approach for species identification of the early life stages of rockfishes of the species-rich genus *Sebastes* Cuvier. *Journal of Experimental Marine Biology and Ecology*, **231**, 279–290.
- Sambrook J, Russell DW (2001) *Molecular Cloning: a Laboratory Manual*, 3rd edn. Cold Springs Harbor Laboratory Press, Cold Springs Harbor, New York.
- Schmidt J (1922) The breeding places of the eel. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **211**, 179–211.
- Scott GP, Turner SC, Grimes CB, Richards WJ, Brothers EB (1993) Indices of larval bluefin tuna, *Thunnus thynnus*, abundance in the Gulf of Mexico; modelling variability in growth, mortality, and gear selectivity. *Bulletin of Marine Science*, **53**, 912–929.
- Serafy JE, Cowen RK, Paris CB, Capo TR, Luthy SA (2003) Evidence of blue marlin, *Makaira nigricans*, spawning in the vicinity of Exuma Sound, Bahamas. *Marine and Freshwater Research*, **54**, 299–306.
- Shivji M, Clarke S, Pank M *et al.* (2002) Genetic identification of pelagic shark body parts for conservation and trade monitoring. *Conservation Biology*, **16**, 1036–1047.
- Smith DG (2002) Larvae of the garden eel genus *Gorgasia* (Congridae, Heterocongrinae) from the Western Caribbean Sea. *Bulletin of Marine Science*, **70**, 831–836.
- Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1825–1834.
- Sponaugle S, Cowen RK (1997) Early life history traits and recruitment patterns of Caribbean wrasses (Labridae). *Ecological Monographs*, **67**, 177–202.
- Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1859–1868.
- Victor BC (1986) Duration of the planktonic larval stage of one hundred species of Pacific and Atlantic wrasses (family Labridae). *Marine Biology*, **90**, 317–326.
- Ward RD, Zemplak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **360**, 1847–1857.
- Wouthuyzen S, Miller MJ, Aoyama J *et al.* (2005) Biodiversity of anguilliform leptocephali in the central Indonesian seas. *Bulletin of Marine Science*, **77**, 209–224.