BARCODING VERTEBRATES

Identifying sharks with DNA barcodes: assessing the utility of a nucleotide diagnostic approach

EUGENE H.-K. WONG,* MAHMOOD S. SHIVJI† and ROBERT H. HANNER*

*Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, ON, Canada N1G 2W1, †Guy Harvey Research Institute and Save Our Seas Shark Center, Nova Southeastern University, 8000 North Ocean Drive, Dania Beach, FL 33004, USA

Abstract

Shark fisheries worldwide are mostly unmanaged, but the burgeoning shark fin industry in the last few decades has made monitoring catch and trade of these animals critical. As a tool for molecular species identification, DNA barcoding offers significant potential. However, the genetic distance-based approach towards species identification employed by the Barcode of Life Data Systems may oftentimes lack the specificity needed for regulatory or legal applications that require unambiguous identification results. This is because such specificity is not typically realized by anything less than a 100% match of the query sequence to an entry in the reference database using genetic distance. Although various divergence thresholds have been proposed to define acceptable levels of intraspecific variation, enough exceptions exist to cast reasonable doubt on many less than exact matches using a distance-based approach for the identification of unknowns. An alternative approach relies on the identification of discrete molecular characters that can be used to unambiguously diagnose species. The objective of this study was to assess the performance differences between these competing approaches by examining more than 1000 DNA barcodes representing nearly 20% of all known elasmobranch species. Our results demonstrate that a character-based, nucleotide diagnostic (ND) approach to barcode identification is feasible and also provides novel insights into the structure of haplotype diversity among closely related species of sharks. Considerations for the use of NDs in applied fields are also explored.

Keywords: DNA barcoding, sharks, single nucleotide diagnostics

Received 26 October 2008; revision received 4 January 2009; accepted 24 January 2009

Introduction

The demand for shark fins worldwide has significantly increased over the past two decades and is likely to be a major factor influencing shark mortality in fisheries and the health of shark populations (Clarke *et al.* 2004; Clarke *et al.* 2006). While the current state of the shark fin industry is lucrative, the low economic value of other shark products (meat, liver oil, cartilage, skin and teeth) in the past has led to historically unmanaged shark fisheries (Pank *et al.* 2001). More recent attempts to manage shark fisheries, prompted by conservation needs, have encountered numerous difficulties related to the identification of landed shark species. For example, when a shark is caught, the fins are often removed on-board and the remaining carcass is discarded back into the ocean. The net result of 'finning' is that the morphological features necessary for identification are never seen, which is a common problem for nearly all investigations involving commercial animal products. Identification is further complicated by the use of 30–45 market categories for shark fins in Hong Kong (85% of the world's shark imports) that do not necessarily correspond to taxonomic names (Vannuccini 1999). Therefore, tallies and proportions of landed shark species cannot be easily performed by tracing market activity alone.

Molecular diagnostic techniques are now commonly employed for species identification and circumvent the drawbacks of complex morphology-based identification keys. For shark species identification, recent studies (Shivji

Correspondence: Dr Robert H. Hanner, Fax: 519 767 1656; E-mail: rhanner@uoguelph.ca

et al. 2002; Abercrombie *et al.* 2005; Shivji *et al.* 2005; Clarke *et al.* 2006; Magnussen *et al.* 2007) have been based on successful multiplex polymerase chain reaction (PCR) assays using species-specific primers. This approach is rapid and inexpensive making it amenable for routine use in situations where resources are limited, but it is limited to testing for small groups of species at a time and species-specific primers have yet to be developed for many species.

Sequence analysis is an alternative approach to molecular species diagnosis in sharks (Greig *et al.* 2005; Quattro *et al.* 2006). Sequence data are often the most direct way to obtain a large amount of information, and have become faster and more affordable as sequencing technology progresses. However, with the increased accessibility of sequencing technology, sequence-based species identification techniques are being developed and employed independently in different laboratories without a set of reference standards or a common genetic marker. As a result, comparative power between sequences is limited and they cannot draw upon a comprehensive reference database. DNA barcoding (Hebert *et al.* 2003a; Hebert *et al.* 2003b) has successfully discriminated many regional shark species from Australian waters (Holmes *et al.* 2009; Ward *et al.* 2008). While identification via a DNA barcode is another sequence-based approach, the technique distinguishes itself because it is based on a standardized gene region. In addition, DNA barcode reference records are supported by a network of supplementary information, allowing barcode sequences to be independently reviewed (Ratnasingham & Hebert 2007).

Identification via DNA barcodes is based on the observation that intraspecific genetic divergence is usually lower than interspecific divergence (e.g. 'barcoding gap') (Meyer & Paulay 2005). The identification engine built into the Barcode of Life Data Systems (BOLD; www.barcodinglife.org) currently uses a phenetic-based approach by comparing unknown sequences to the reference database from the perspective of a sequence similarity estimate. Ferguson (2002) reasons that species identification through genetic divergences is based on the argument that genetic divergence suggests genetic incompatibility, which suggests reproductive isolation, which suggests a lack of gene flow, which culminates in an inference of separate species. However, Ferguson (2002) further argues that there are too many exceptions where substantial genetic divergence is not necessary for the development of reproductive isolation, or where the genetic bases for reproductive isolation are simply too varied for a predictive set of rules capable of differentiating species-level and population-level variation.

The best-case scenario for a phenetic-based approach is one in which an unknown sequence has 100% sequence similarity to a reference sequence; however, this is problematic as such a match is often not obtained, which raises the issue of a species threshold similarity value. Early DNA barcoding studies suggested that a 2% threshold typically represented a cut-off between species (Hebert *et al.* 2003a, 2004a, b), but it is known that the divergence between species can fall well below this value, depending on the taxa of interest (Ward *et al.* 2005). As with the arguments put forth by Ferguson (2002), interspecific levels of divergence are variable between taxa, and a generalized 2% rule cannot be applied across all species. The accuracy of the current distance measure for identifications via DNA barcodes is also heavily dependent on the sampling breadth of the database and is susceptible to false-positives under a number of scenarios, including incomplete taxonomic sampling and a lack of disparity between intra-specific and interspecific variation (Frézal & Leblois 2008).

Using a threshold value results in increased uncertainty as the sequence similarity between an unknown and reference sequence approaches this cut-off point (Rubinoff 2006). This approach may be suitable for academic purposes, but the current system is not structured to serve end users seeking to employ DNA barcoding in a regulatory setting, such as agents of conservation and wildlife forensic sciences who may be under strict scrutiny and involved in litigation (Ogden 2008).

The uncertainty of sequence similarity as a metric for species delineation has been recognized as a problem (Köhler 2007), and character-based methods are one of the proposed alternatives as they maintain information that is inherently lost in a distance approach (DeSalle *et al.* 2005; DeSalle 2007; Kelly *et al.* 2007; Waugh *et al.* 2007). A character-based approach is akin to the traditional morphology-based methods in that species diagnosis would be based on a binary signal — the presence or absence of a diagnostic character (in this case a DNA character), therefore bypassing the uncertainty found in an analogue measurement of sequence similarity. These diagnostic characters can be thought of as various single nucleotide polymorphisms that are fixed across a species but differ between species. However, the use of single nucleotide polymorphism (SNP) in this context would be a misnomer and an inappropriate use of the term. By definition, an SNP describes a variation that occurs at the population level. In order to avoid this confusion, the diagnostic sites of interest in this study are referred to as nucleotide diagnostics (ND).

NDs, or character attributes (CA) in the terminology of Sarkar *et al.* (2002a, b), have already been shown to be applicable and successful in species identification (Kelly *et al.* 2007; Rach *et al.* 2008), yet no direct comparisons of characters to distance have been made for identifications using DNA barcode sequence data. The study by Rach *et al.* (2008) focused only on 'simple' CAs. A simple CA is a stand alone diagnostic at a single nucleotide position. However, they do make note of 'compound' diagnostics that combine character states from multiple nucleotide sites. The addition of compound character states is the foundation

Table 1 Definitions and illustrated examples of terms used relating to nucleotide diagnostics

Sp.1

```
TTGATTCAGAGGAGGAGGAGACCCTATTCTTTACCAACCATCTTAGGAGGCCC
TTGATTCAGAGGAGGAGGAGACCCTCTTCTTTACCAACTATCTTAGGAGGCCC
TTGATTCAGAGGAGGAGGAGACCCTATTCTTTACCAACCATCTTAGGAGGCCC
TTGATTCAGAGGAGGAGGAGACCCTCTTCTTTACCAACTATCTTAGGAGGCCC
```

Sp.2

```
TTGATCCAGAGGAGGAGGAGACCCTATTCTCTACCAACCATCTTAGGAGGCCC
TTGATCCAGAGGAGGAGGAGACCCTATTATTTACCAACCATCTTAGGAGGCCC
TTGATCCAGAGGAGGAGGAGACCCTATTCTCTACCAACCATCTTAGGAGGCCC
TTGATCCAGAGGAGGAGGAGACCCTATTATTTACCAACCATCTTAGGAGGCCC
    (A)                     (B)      (C)
```

Nucleotide diagnostic (ND). NDs are characters that are unique (i.e. diagnostic) for a particular species, relative to the pool of species in which the ND was identified. They are either comprised of a single nucleotide position on its own (simple), or multiple single nucleotide positions combined (compound).

Simple. Sarkar *et al*. (2000b) used this term to describe diagnostic characters at single positions. An example of a simple ND (sND) can be seen at the nucleotide position marked (A) above. T is diagnostic for the top set of sequences (Sp.1), while C is diagnostic for the bottom set (Sp.2).

Compound. Sarkar *et al*. (2000b) used this term to describe diagnostic characters that are comprised of multiple nucleotide sites in conjunction. An example of a compound ND (cND) can be seen at the two nucleotide positions marked (B) above. C-T is diagnostic for Sp.1. A cND for a species is derived by the addition of one single nucleotide position at a time — each one of these being diagnostic for a different nested subset of the total species pool.

Conditional. A conditional ND is a special type of compound ND introduced in this study. Sarkar *et al*. (2000b) use the terms 'pure' and 'private' to describe the difference between the pattern seen at (A) and (C). At (A), these characters are considered pure diagnostics for each set. By definition, an sSND is always pure. At (C), the T exhibited in Sp.1 is considered private to that set. T would be diagnostic for Sp.1, while C may belong to either. Conditional NDs were used in this study when species pairs were not definitively differentiable, although a cND was capable of diagnosing down to the species pair. The addition of a private nucleotide site that was diagnostic for some haplotypes of a species to cND created a conditional ND. Therefore, a conditional ND is diagnostic for a subset of the species.

of an expandable character-based diagnostic system, as the power and resolution of diagnostic characters can be increased substantially by considering permutations of characters across multiple sites. Therefore, NDs can be broken down into two types: simple NDs (sNDs), which provide a diagnostic signal at a single nucleotide position, and compound NDs (cNDs), which combine several single nucleotide sites to become diagnostic. Table 1 provides an example of our use of these terms.

In this study, the potential use of NDs as an alternative approach to applied species identification was examined by generating a large data set of COI DNA barcodes for a diverse range of sharks, an ecologically, economically and culturally significant group of fishes. Unambiguous NDs will be a boon to DNA barcode applications that cannot afford the ambiguity inherent in the current BOLD distance–based identification engine, such as the monitoring of animal goods for conservation or legal purposes.

Methods

Sequence data preparation

1049 shark (superorder Selachimorpha) samples were acquired from Nova Southeastern University Oceanographic Center (NSUOC), Kansas University Natural History Museum, and the National Ocean Service Marine Forensics Archive. Sequences and sample records can be viewed on the Barcode of Life Data System or BOLD (at http://www.boldsystems.org) under project code EWSHK. Some samples from NSUOC were received as DNA extracts. All others were tissue stored in 95% ethanol. All samples were stored at –20 °C until processed.

As part of the supplementary information that comprises the DNA barcode data standard (Hanner 2005), locality data were maintained for the source of each specimen where available. The Fish Barcode of Life Initiative (FISH-BOL), a coordinated international campaign to compile reference barcodes for all fish, operates using the Food and Agriculture Organization's (FAO) 19 statistical marine regions. As a contribution to FISH-BOL, specimen sources were allocated to an FAO region in BOLD, and when possible specimens were chosen in order to represent regions across a species' distribution.

For all samples from NSUOC not received as DNA extracts, the DNeasy Tissue Kit (QIAGEN Inc.) was used to perform the genomic DNA extraction, following the instructions of the manufacturer. This same kit was also used to extract genomic DNA from all samples acquired from the NOSMF Archive. For all other samples, 1–2 mm^3 of tissue was used for DNA extraction via the protocols detailed by Ivanova *et al*. (2006).

A 652-bp fragment from the 5′ region of COI, corresponding to base positions 6474–7126 of the *Danio rerio* mitochondrial genome (Broughton *et al*. 2001), was PCR amplified using one of two sets of forward and reverse primer cocktails (Table 2), C_FishF1t1-C_FishR1t1 or C_VF1LF1t1-C_VR1LRt1 (Ivanova *et al*. 2007), appended with M13 tails to aid in sequencing (Messing 1983). Each PCR mixture consisted of 6.25 μL of 10% trehalose, 3.0 μL of ultrapure ddH_2O, 1.25 μL of 10× PCR buffer for Platinum *Taq* (Invitrogen, Inc.), 0.625 μL of 50 mM $MgCl_2$, 0.125 μL of each primer (10 μM), 0.0625 μL of 10 mM dNTP mix, 0.06 μL of Platinum *Taq* DNA polymerase (Invitrogen, Inc.), and 0.5–2.0 μL of template DNA (approximately 50–100 ng of DNA). PCR amplification reactions were conducted in Eppendorf Mastercycler gradient thermal cyclers (Brinkmann Instruments, Inc.) The reaction programme for samples using the C_FishF1t1-C_FishR1t1 primers consisted of

Table 2 PCR primer cocktail components and corresponding sequences. M13 tails are also listed below

Primer name			
Cocktail name	Component name	Sequence	
C_FishF1t1	VF2_t1	5′TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC3′	(Ward *et al*. 2005)
(1 : 1 ratio)	FishF2_t1	5′TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC3′	(Ward *et al*. 2005)
C_FishR1t1	FishR2_t1	5′CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA3′	(Ward *et al*. 2005)
(1 : 1 ratio)	FR1d_t1	5′CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA3′	(Ivanova *et al*. 2007)
C_VF1LFt1 (1 : 1 : 1 : 3 ratio)	LepF1_t1	5′TGTAAAACGACGGCCAGTATTCAACCAATCATAAAGATATTGG3′	(Hebert *et al*. 2004a)
	VF1_t1	5′TGTAAAACGACGGCCAGTTCTCAACCAACCACAAAGACATTGG3′	(Ivanova *et al*. 2006)
	VF1d_t1	5′TGTAAAACGACGGCCAGTTCTCAACCAACCACAARGAYATYGG3′	(Ivanova *et al*. 2006)
	VF1i_t1	5′TGTAAAACGACGGCCAGTTCTCAACCAACCAIAAIGAIATIGG3′	(Ivanova *et al*. 2006)
C_VR1LRt1 (1 : 1 : 1 : 3 ratio)	LepRI_t1	5′CAGGAAACAGCTATGACTAAACTTCTGGATGTCCAAAAAATCA3′	(Hebert *et al*. 2004a)
	VR1_t1	5′CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCRAARAAYCA3′	(Ivanova *et al*. 2006)
	VR1d_t1	5′CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCAAAGAATCA3′	(Ward *et al*. 2005)
	VR1i_t1	5′CAGGAAACAGCTATGACTAGACTTCTGGGTGICCIAAIAAICA3′	(Ivanova *et al*. 2006)
M13F		5′TGTAAAACGACGGCCAGT3′	(Messing 1983)
M13R		5′CAGGAAACAGCTATGAC3′	(Messing 1983)

Shaded portions highlight the M13 tails.

2 min at 94 °C, followed by 35 cycles of 30 s at 94 °C, 40 s at 52 °C, and 1 min at 72 °C. Upon completion of the 35 cycles, the programme concluded with 10 min at 72 °C and then held at 4 °C. The reaction programme for samples using the C_VFLF1t1-C_VRLR1t1 primers consisted of 2 min at 94 °C, followed by 5 cycles of 30 s at 94 °C, 40 s at 50 °C, and 1 min at 72 °C, followed by 35 cycles of 30 s at 94°C, 40 s at 54 °C, and 1 min at 72 °C. The programme concluded with 10 min at 72 °C and then held at 4 °C.

PCR products were visualized on 2% agarose E-gel 96 plates (Invitrogen, Inc.). PCR products were labelled using the BigDye Terminator version 3.1 Cycle Sequencing Kit (Applied Biosystems, Inc.). Each cycle sequencing reaction mixture consisted of 5.0 µL of 10% trehalose, 0.917 µL of ultrapure ddH$_2$O, 1.917 µL of 5× buffer (400 mm Tris-HCl ph 9.0 and 10 mm MgCl$_2$), 1.0 µL of primer (10 µm; M13F or M13R, Table 2), 0.167 µL of BigDye (Applied Biosystems, Inc.), and 1.5 µL of PCR product. Sequencing reaction clean-up was carried out using the Agencourt CleanSEQ system with SPRI paramagnetic bead technology (Agencourt Bioscience Corporation). Bi-directional sequencing reactions were resolved using an ABI 3730 capillary sequencer. Bi-directional contigue assembly was carried out using SeqScape version 2.1.1 (Applied Biosystems, Inc.).

All sequences in this study have been deposited in GenBank, via the use of BarSTool, under accession numbers FJ518910–FJ519800, FJ529802–FJ519955.

ND identification for shark species

Sequences for 74 independently identified shark species were used in the total species pool for the ND identification process. Sequences were run through Collapse1.2 (available at http://darwin.uvigo.es) in order to distill the sequences into unique haplotypes. The target was to identify NDs for the 64 species in the total species pool belonging to the orders Carcharhiniformes and Lamniformes (Table 3), which contain the most commercially relevant and utilized species, although the whale shark (*Rhincodon typus,* order Orectolobiformes) was also included due to its inclusion in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) indices. With the addition of the whale shark, all three shark species currently listed in CITES (Appendix II) were included in the analysis. The other two are the great white shark (*Carcharodon carcharias*) and the basking shark (*Cetorhinus maximus*). NDs for each species of interest were identified in the context of the entire pool of available species. As an example, in order to identify NDs for *Carcharhinus signatus*, haplotypes for this species would be compared to haplotypes from the other 73 shark species included in this study.

Identification and confirmation of NDs were accomplished in a straightforward manner by eye, using MEGA version 4 (Kumar *et al*. 2004) to display the aligned sequence data and highlight all variable sites. NDs for a target species were easy to identify with a visual scan when the 'use identical symbol' option was enabled in MEGA. Under this viewing format, base positions featuring a simple ND (sND) were characterized by a column in which the identical symbol (.) was absent for all other species. That is, all other species exhibited a character state not shared with the focus species.

Compound NDs (cND) were also easily identified using a simple manual algorithm (Fig. 1). When no sNDs are present in a species, an arbitrary site (base position 1) can

Table 3 NDs for each species are listed, with sample sizes included in parentheses after the species name

Order/family	Species	NDs	Notes
Carcharhiniformes/ Carcharhinidae	*Carcharhinus signatus* (9)	(22-C + **356-C**)	
	Carcharhinus sorrah (6)	(199-A + **350-A + 527-A**)	
	Carcharhinus albimarginatus (12)	(295-A + **413-G**)	
	Carcharhinus isodon (12)	(220-G + **371-C**)	
	Carcharhinus dussumieri (7)	(280-C + **413-G**)	
	Carcharhinus porosus (1)	(127-A + **167-A**)	
	Carcharhinus perezii (12)	(70-G + 76-G)	
	Carcharhinus brevipinna (10)	(58-T + **371-C**)	
	Carcharhinus brachyurus (5)	(334-C + 541-C)	
	Carcharhinus falciformis (11)	(70-G + 76-A)	
	Carcharhinus amboinensis (4)	(136-C + **350-A**)	
	Carcharhinus amblyrhynchos (27)	(298-C + **413-G**)	
	Carcharhinus longimanus (18)	(283-C + 307-G)	
	Carcharhinus cautus (1)	(343-G + **415-A**)	
	Carcharhinus melanopterus (8)	(350-A + 415-T + 484-G)	
	Carcharhinus leucas (19)	(40-T + 361-A)	
	Carcharhinus acronotus (11)	(139-C + 194-C + 565-G)	
	Nasolamia velox (1)	(217-G + 457-C)	
	Carcharhinus altimus (8)	(451-A + **542-T**)	cND only diagnostic down to the species pair.
	Carcharhinus plumbeus (12)		Conditional ND available for both species (Table 2).
	Carcharhinus limbatus (11)	(388-C + 397-T + 400-T)	cND only diagnostic down to the species pair.
	Carcharhinus tilstoni (8)		Conditional ND available for western Atlantic population of *C. limbatus* (Table 2).
	Carcharhinus obscurus (19)	(214-A + 475-T)	cND only diagnostic down to the species pair.
	Carcharhinus galapagensis (16)		Conditional ND available for *C. obscurus* (Table 2).
	Galeocerdo cuvier (160)	(388-A + **371-C**) 10-T	
	Prionace glauca (19)	(**371-C** + 565-T)	
	Negaprion acutidens (2)	(172-G + 319-A)	
	Negaprion brevirostris (11)	(145-G + **371-C**)	
	Triaenodon obesus (3)	496-G (346-G + **350-A**)	
	Rhizoprionodon acutus (1)	166-G (**350-A** + 493-C)	
	Rhizoprionodon lalandii (10)	**300-C** 379-C	
	Rhizoprionodon porosus (10)	(25-G + **356-A**)	cND only diagnostic down to the species pair.
	Rhizoprionodon terraenovae (15)		Conditional ND available for both species (Table 2).
Carcharhiniformes/ Sphyrnidae	*Eusphyra blochii* (4)	(43-T + **371-T**)	
	Sphyrna lewini (86)	(193-G + **35-A**)	
	Cryptic *Sphryna lewini* (4)	(502-C + **35-A**)	(Abercrombie *et al*. 2005; Quattro *et al*. 2006).
	Sphyrna mokarran (34)	(88-C + **371-C**)	
	Sphyrna tiburo (43)	(418-G + **542-T**)	
	Sphyrna tudes (1)	473-T (284-T + **371-C**)	Serine amino acid is diagnostic for this species.
	Sphyrna zygaena (21)	530-T (319-A + 415-A)	
Carcharhiniformes/ Scyliorhinidae	*Apristurus kampae* (2)	469-G 652-C	
	Apristurus brunneus (2)	45-G 124-C 301-C	Amino acid glycine (corresponding to sND 45-G) is diagnostic for this species.
	Apristurus profundorum (4)	(97-C + **371-C)**	
	Apristurus manis (4)	(604-C + **542-G**)	
	Parmaturus xaniurus (2)	8-A 251-G 414-C	Amino acids methionine (corresponding to sND 8-A), glycine (251-G), and alanine (414-C) are diagnostic for this species.
	Scyliorhinus retifer (1)	316-G 476-G 478-A	Amino acid valine (corresponding to sNDs 476-G and 478-A) is diagnostic for this species.

Table 3 *Continued*

Order/family	Species	NDs	Notes
Carcharhiniformes/ Triakidae	*Triakis semifasciata* (2)	(49-G + **86-T**)	
	Ghaleorhinus galeus (4)	(220-T + **371-C**)	
	Mustelus canis (15)	(280-C + **527-G**)	
	Mustelus henlei (3)	(529-C + 547-G)	
	Mustelus californicus (1)	(448-G + **371-C**)	
Lamniformes/ Lamnidae	*Lamna nasus* (80)	409-G (373-C + 196-A)	
	Lamna ditropis (12)	523-G (373-C + 181-A)	
	Carcharodon carcharias (6)	94-C 197-A 574-G	Amino acid isoleucine (corresponding to sND 197-A) is diagnostic for this species.
	Isurus oxyinchus (15)	628-T (418-G + 421-C)	
	Isurus paucus (13)	286-G 595-G	
Lamniformes/ Alopiidae	*Alopias pelagicus* (15)	(472-T + **356-T**)	
	Alopias superciliosus (18)	397-G (301-G + **107-G**)	
	Alopias vulpinus (18)	(262-A + **107-G**)	
Lamniformes/ Odontaspididae	*Carcharias taurus* (73)	577-T (244-C + 304-C)	
	Odontaspis ferox (1)	(42-T + 166-C)	
Lamniformes/ Pseudocarchariidae	*Pseudocarcharias kamoharai* (3)	187-T (334-A + 421-T)	
Lamniformes/ Cetorhinidae	*Cetorhinus maximus* (48)	466-G 520-C 589-G	
Orectolobiformes/ Rhincodontidae	*Rhincodon typus* (9)	41-C 367-G 640-G	Amino acid leucine (corresponding to sND 41-C) is diagnostic for this species.

For species with at least one sND, all sNDs are listed. If only one sND is available, a cND alternative is also provided. For species that require a cND for diagnosis, only one example is listed. Nucleotide positions and character states where point mutations would result in a nonsynonymous change are indicated in **bold**. Nucleotide positions and character states that differ from the species pool by a transversion are underlined.

be chosen to begin defining a cND, assuming that the character is fixed within the target species. The simplest site to choose would be the one in which the character state exhibited by the target species appears least often in other species, but this is not strictly necessary. Once this site is chosen, all other sequences that do not share the same character state are removed from the identification process and the remaining sequences of the reduced subset are scanned again for a sND. The site of any sND located in the nested scan, combined with base position 1 form a cND that is diagnostic for the species of interest. If an sND is still not located, repeated nested scans are conducted until the necessary base positions for species resolution are obtained.

Species pairs that could not be uniquely identified with NDs were examined using statistical parsimony networks (Templeton *et al*. 1992) to construct the relationship between haplotypes. These networks were generated using TCS 1.21 (available at http://darwin.uvigo.es/software/tcs.html) at the default 95% connection limit.

Results

General results

Sequences represented a total of 74 shark species across 33 genera, 17 families, and six orders. The six orders included were Carcharhiniformes, Lamniformes, Orectolobiformes, Squaliformes, Squatiniformes, and Heterdoniformes. The ND identification process focused on the Carcharhiniformes and Lamniformes, and one Orectolobiformes. The Carcharhiniformes in this data set were represented by 51 species across 15 genera and four families, while the

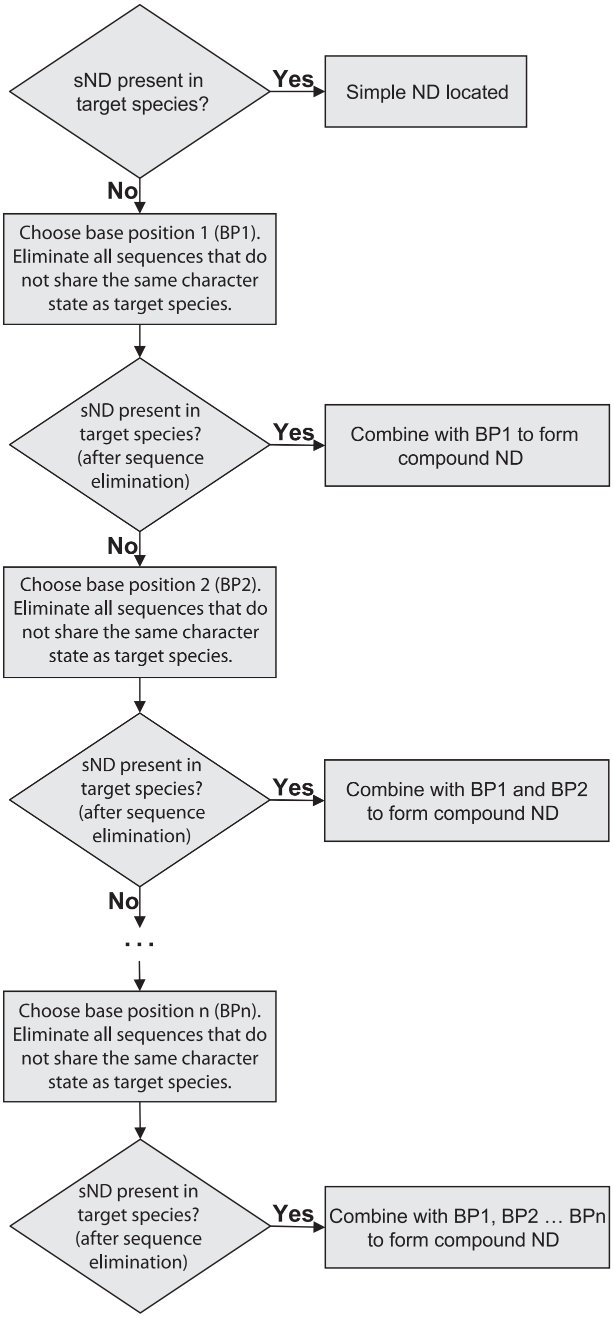

sND present in target species?
Yes
Simple ND located
No
Choose base position 1 (BP1). Eliminate all sequences that do not share the same character state as target species.
sND present in target species? (after sequence elimination)
Yes
Combine with BP1 to form compound ND
No
Choose base position 2 (BP2). Eliminate all sequences that do not share the same character state as target species.
sND present in target species? (after sequence elimination)
Yes
Combine with BP1 and BP2 to form compound ND
No
. . .
Choose base position n (BPn). Eliminate all sequences that do not share the same character state as target species.
sND present in target species? (after sequence elimination)
Yes
Combine with BP1, BP2 … BPn to form compound ND

Fig. 1 This flowchart describes the steps used in manually identifying NDs. All steps beyond the first decision box are designed to identify nucleotide positions that can be joined to form a compound ND. This assumes the species of interest contains at least one compound ND that is diagnostic for it. Chosen base positions (BP1, BP2 ... BPn) must have a consistent character state for all sequences of the target species.

Lamniformes were represented by 12 species across eight genera and five families.

Fourty-three of the 64 species targeted for ND identification were represented by five or more specimens. Overall, there were 16.03 specimens per species on average for the target 64 species, although this value is skewed upwards by a few extensively sampled species. The number of specimens per species ranged from one (eight species) to 161 (*Galeocerdo cuvier*).

Of the 56 species with more than one specimen, 35 (62.5%) were represented by samples from multiple FAO regions, 16 (28.6%) were represented by a single FAO region, and the remaining five were lacking geographical locality data. Specimens from extensively sampled species came from across their global distribution. Specimens from eight species came from at least five FAO regions, ranging up to eight regions. The number of FAO regions listed here encompassed by each species is a minimum estimate. Since not all specimens had locality data, it is possible that those specimens originate from an FAO region not accounted for by the specimens with locality data.

ND identification for shark species

Of the 64 shark species examined for NDs, 20 species had at least one sND, in the context of this study, and six of these also involved amino acid differences that were diagnostic for the species. Thirty-seven species without an sND could be identified using a cND of two to three nucleotide positions in combination (Table 3).

Eight species, forming four pairs of sister species (*Rhizoprionodon porosus*/*Rhizoprionodon terraenovae*, *Carcharhinus plumbeus*/*Carcharhinus altimus*, *Carcharhinus limbatus*/*Carcharhinus tilstoni*, and *Carcharhinus obscurus*/*Carcharhinus galapagensis*), could not be uniquely diagnosed, although cNDs were located that are capable of diagnosing down to each pair. Statistical parsimony networks are included for each species pair (Figs 2–5). The networks revealed nucleotide sites in six of these species such that their addition to the cND diagnostic for the pair would result in a new cND that could identify the species of some haplotypes, but not all of them (Table 4).

A recent re-analysis of two DNA barcoding studies using statistical parsimony networks indicated that species tended to dissociate into separate networks and that sequences from a single species remained together (Hart & Sunday 2007). An overall statistical parsimony network was constructed using the entire shark data set in this study to determine if the included species would form their own networks. Beyond the eight species that could not be diagnosed with an ND, nine more species did not separate into their own network. These nine species appeared as four more species pairs and one species (*Carcharhinus longimanus*) formed a network with the *C. obscurus*/*C. galapagensis* pair (Table 5). Of the 74 total species, 57 (77.0%) formed independent networks. *Sphyrna lewini* is the only species to separate into multiple (two) networks.

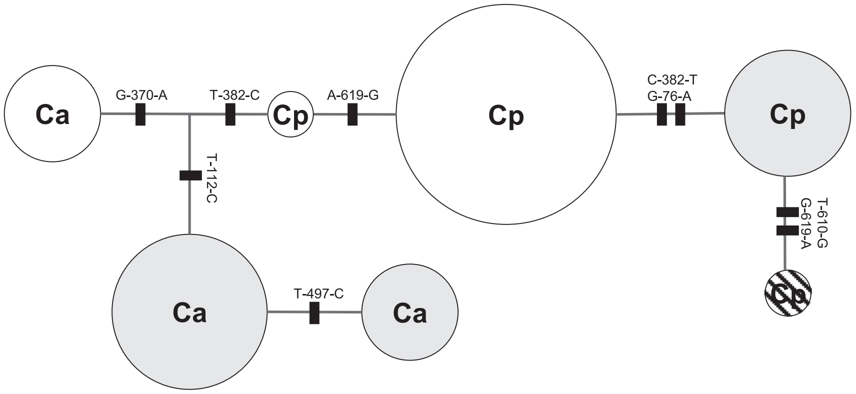


Fig. 2 Statistical parsimony network for haplotypes of *Carcharhinus altimus* (Ca) and *Carcharhinus plumbeus* (Cp). Haplotypes are indicated by the circular nodes, marked by the corresponding species designation. Each black hash mark represents a single character change. Unshaded nodes are from the Indo-Pacific, while shaded nodes are from the northwest Atlantic. Striped notes lacked locality data. The size of each haplotype node is proportional to the number of specimens representing each haplotype. The total number of specimens in this network is 20. The network was generated using TCS 1.21 (available at http://darwin.uvigo.es/software/tcs.html) at the default 95% connection limit.

Discussion

NDs as contextual entities

An immediate consideration when using character-based diagnostics is that the existence of NDs for a particular species is contextual and can only be applied with confidence in the appropriate context. That is, NDs are relative to the total reference sequence species pool examined because they are described as character states at certain nucleotide positions that distinguish the target species from the rest of the pool.

It would be a simple task to locate multiple NDs for a given species if the comparison was limited in scope to, for example, identifying the species of interest out of a pool of only its congeners. In an extreme example, it is easy to understand that NDs would be abundant if the user were only interested in distinguishing between two hypothetical species of one haplotype each. In this case, every single variable site between the two species would represent an sND. Obviously, the addition of another species or even a previously unrepresented haplotype will eliminate some sNDs. As the total specimen haplotype and species pool increases, sNDs are less likely to be present as the variability in character states reaches saturation, thus forcing a reliance on cNDs. However, since character-based diagnostics are being viewed as a solution to the ambiguities brought upon in an applied setting by the percentage similarity metric currently employed, it was important to retain other features of DNA barcoding that gave it an inherent advantage as an applied tool, such as its broad applicability to a large taxonomic scale. It is for this reason that NDs were identified for the shark species of interest relative to the entire pool of sequenced shark species available. While the species covered in this study do not constitute a comprehensive sampling of all shark species, the pool of 74 species represents approximately 17.5% of the total species diversity in sharks.

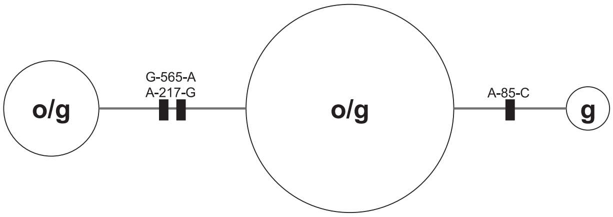


Fig. 3 Statistical parsimony network for haplotypes of *Carcharhinus obscurus* (o) and *Carcharhinus galapagensis* (g). Haplotypes are indicated by the circular nodes, marked by the corresponding species designation. Nodes that indicate both species are haplotypes that are shared between the species pair. Each black hash mark represents a single character change. Nodes do not correlate with a phylogeographical pattern. The size of each haplotype node is proportional to the number of specimens representing each haplotype. The total number of specimens in this network is 35. The network was generated using TCS 1.21 (available at http://darwin.uvigo.es/software/tcs.html) at the default 95% connection limit.

The addition of a closely related species is more likely to eliminate a previous ND. The samples from order Carcharhiniformes become a good test group, as the 51 species from that order contained 23 from the same genus (*Carcharhinus*). The effect of additional closely related species on the number of available sNDs was immediately apparent. In contrast to the Carcharhiniformes, the Lamniformes are a much smaller group, and in this study, contained fewer closely related species (even though 75% of the total Lamniformes diversity was represented, 12 of 16 species).

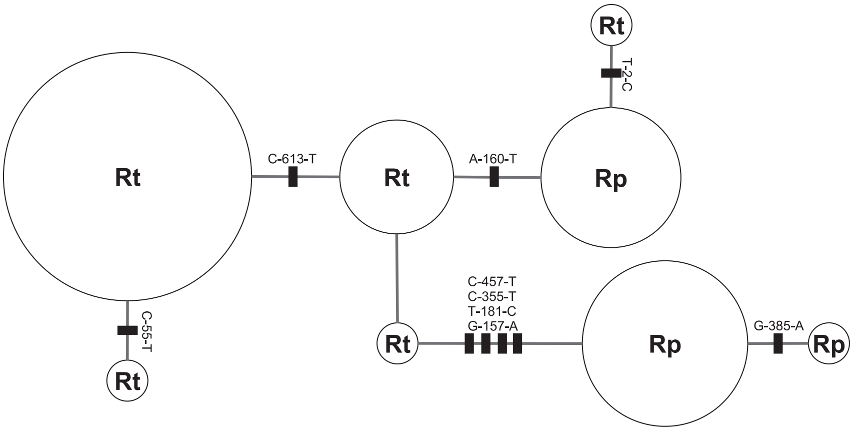


Fig. 4 Statistical parsimony network for haplotypes of *Rhizoprionodon terraenovae* (Rt) and *Rhizoprionodon porosus* (Rp). Haplotypes are indicated by the circular nodes, marked by the corresponding species designation. Each black hash mark represents a single character change. Nodes do not correlate with a phylogeographical pattern. The size of each haplotype node is proportional to the number of specimens representing each haplotype. The total number of specimens in this network is 25. The network was generated using TCS 1.21 (available at http://darwin.uvigo.es/software/tcs.html) at the default 95% connection limit.

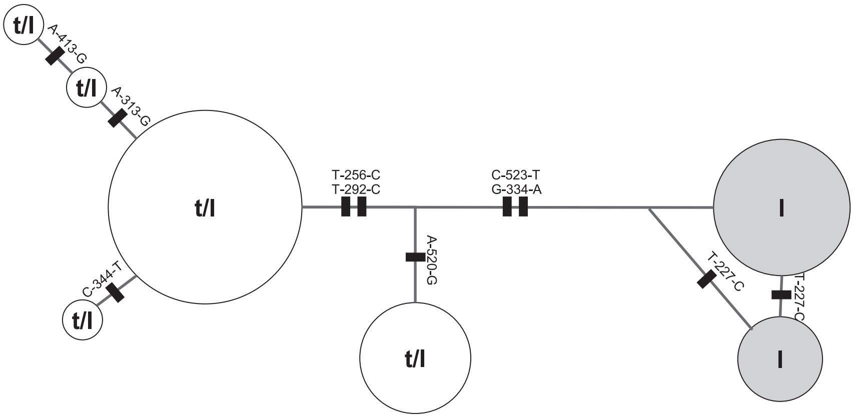


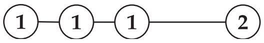
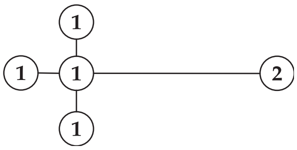
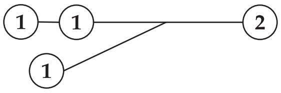
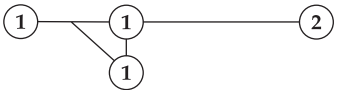
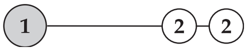
Fig. 5 Statistical parsimony network for haplotypes of *Carcharhinus tilstoni* (t) and *Carcharhinus limbatus* (l). Haplotypes are indicated by the circular nodes, marked by the corresponding species designation. Nodes that indicate both species are haplotypes that are shared between the species pair. Each black hash mark represents a single character change. Unshaded nodes are from the Pacific Ocean, while shaded nodes are from the west Atlantic Ocean. The size of each haplotype node is proportional to the number of specimens representing each haplotype. The total number of specimens in this network is 19. The network was generated using TCS 1.21 (available at http://darwin.uvigo.es/software/tcs.html) at the default 95% connection limit.

Table 4 List of conditional NDs for species pairs. The addition of the nucleotide sites and character states listed in column three to the previously established cND for the species pairs, in column two, creates a conditional ND for the corresponding species

Species	cND diagnostic for pair	Additional sites for conditional ND
Carcharhinus altimus *Carcharhinus plumbeus*	(451-A + 542-T)	(370-G) or (112-C) (76-A) or (382-C)
Carcharhinus limbatus *Carcharhinus tilstoni*	(388-C + 397-T + 400-T)	(334-A or 523-T) Corresponds to the western Atlantic population None. All *C. tilstoni* haplotypes are shared with at least one *C. limbatus* sequence.
Carcharhinus obscurus *Carcharhinus galapagensis*	(214-A + 475-T)	None. All *C. obscurus* haplotypes are shared with at least one *C. galapagensis* sequence. (85-C)
Rhizoprionodon porosus *Rhizoprionodon terraenovae*	(25-G + 356-A)	(457-T, 355-T, 181-C, or 157-A) (613-C) or (2-T)

Additional sites in column 3 are grouped in parentheses, with all sites within a group applying to the same branch in the statistical parsimony network.

Table 5 Statistical parsimony networks that contain multiple species, with at least one that is diagnosable with an ND

Species	Statistical parsimony network	Characters between species
(1) *Carcharhinus acronotus* (2) *Nasolamia velox*		3
(1) *Carcharhinus brevipinna* (2) *Carcharhinus brachyurus*		9
(1) *Carcharhinus melanopterus* (2) *Carcharhinus cautus*		8 or 11
(1) *Mustelus canis* (2) *Mustelus henlei*		10
(1) *Carcharhinus galapagensis*/ *Carcharhinus obscurus* (see Fig. 4) (2) *Carcharhinus longimanus*		10

Each unshaded circle represents a unique haplotype, and is numbered to correspond with the species name in column 1. The third column shows the number of character changes between each species. The shaded circle represents the *Carcharhinus galapagensis*/*Carcharhinus obscurus* species pair shown in Fig. 4.

Because of this, a larger proportion of the Lamniformes exhibited sNDs as compared to the Carcharhiniformes. Despite this, cNDs were still applicable within the Carcharhiniformes to identify species or, failing that, a species pair. Not surprisingly, none of the sNDs within the Carcharhiniformes were from the heavily represented genus *Carcharhinus*. All three CITES-listed species contained multiple sNDs, although all three are also the only extant species in their respective genera.

Software assisted ND identification

The contextual nature of NDs is a potential hurdle if the goal is to develop a broadly applicable and stable diagnostic system. As already discussed, ideal NDs would be relative to as large a taxonomic breadth as possible. However, no absolute certainty can be assigned to any ND without the impossible task of sequencing all individuals of all species. At best, the analytical work can move towards a comprehensive set of all species within the taxa of interest (e.g. all sharks) with a strong emphasis on appropriate representation across each species' geographical range. That is not to say that NDs cannot be identified for a data set that is less than complete, but users must be aware that the addition to the data set of a new species, or even a new haplotype of an existing species, may nullify the diagnostic status for a given nucleotide position and species, requiring a re-evaluation of previously defined NDs. If ND identification is to be a dynamic process, then the automation of ND identification would become necessary for application work. Several studies (Kelly *et al*. 2007; Rach *et al*. 2008) have already demonstrated the use of the characteristic attribute organization system (CAOS) algorithm (Sarkar *et al*. 2002a, b) to aid in the identification of diagnostic characters, but did not address compound diagnostics. Considering the simple nature of the manual approach used in this study, it stands that further development of existing or new software tools to allow for the recognition of nucleotide patterns and compound diagnostics is readily achievable.

A program designed to automate ND identification performs a very simple task. Ideally, the identification of NDs should be as user friendly. It could be as simple as submitting a text file, containing reference sequences for the group of interest, to the search program. The program would then be comparing strings of text, and need to be able to recognize which groups of sequences to treat as a single species. CAOS uses a guide tree to accomplish this, but sequence groups could also be established without one, thus simplifying the required steps. The most complicated and novel aspect of such a program will be to have it accurately assemble cNDs. One basic method would be to have the program mimic what was performed manually in this study, although there are undoubtedly more efficient, albeit more sophisticated, alternatives.

ND preference considerations

The manual method laid out in this study is essentially arbitrary. While the most direct way to identify a cND for a target species is to first choose a character that has the lowest presence in the overall species pool, thereby limiting the subset of species in the next step as much as possible, it is not strictly required. However, there should be some considerations regarding preferred nucleotide sites, especially if the process is to be automated. With saturation being a major source of character state overlap, the general rule of thumb in the selection of preferred NDs should be to focus on nucleotide sites that are less susceptible to random mutation, and therefore convergence (DeSalle 2006). Six species exhibited sNDs that corresponded to an amino acid difference, such that the amino acid itself was diagnostic. Similarly, the use of nonsynonymous nucleotide sites can also be applied to cNDs. In this case, the nucleotide positions that form a cND are chosen such that random mutations are under functional constraint. This typically means that changes at first or second codon positions are preferred, but are not necessarily available in all species, as some only exhibit third codon position variations or otherwise synonymous changes. A second consideration is whether or not existing variability involves a transitional or transversional mutation. Brown *et al*. (1982) detailed the predominance of transitional changes, noting that an average of 92% of all point mutational differences in their study on the rate and mode of mitochondrial DNA evolution were transitions. Under this observation, another useful guideline would be to look for nucleotide positions where the target species exhibits a character state that differs from the other species by a transversion substitution. Nucleotide positions characterized by a nonsynonymous change and/or transversion are marked in Table 3.

Species pairs and statistical parsimony networks

Four species pairs could not be cleanly diagnosed down to the individual species level. Although cNDs were found to diagnose each pair, further resolution could not be attained due to overlapping character states at all variable sites between species once a one to one comparison was made. For example, while the species pair of *Carcharhinus plumbeus* and *C. altimus* can be picked out of the overall species pool with the cND (451-A + 541-T), the variable sites between just these two species do not feature any characters that are fixed across all haplotypes of one species, while remaining different from the second species. The close mitochondrial DNA relationship between these two species has been noted previously (Greig *et al*. 2005) and the patterns of variability cannot provide definitive diagnostics, but they can still provide partial resolution. In some cases, there are nucleotide positions with character states that are unique to a species, but not all haplotypes of that species will exhibit it. This creates a conditional ND (Table 1), or a private character (Sarkar *et al*. 2002b; DeSalle *et al*. 2005). It is then possible to use a conditional ND to identify some individuals within one of these species pairs to the species level. The sampling for *C. plumbeus* and *C. altimus* broke down into four and three unique haplotypes for these species respectively (Fig. 2). At nucleotide position 112, all four *C. plumbeus* and one *C. altimus* haplotypes are characterized by a T, whereas the remaining two *C. altimus* are characterized by a C. Therefore, the addition of this nucleotide position to the previous cND would create a conditional ND. (451-A + 541-T + 112-C) would be diagnostic to *C. altimus*, whereas (451-A + 541-T + 112-T) could still be either *C. altimus* or *C. plumbeus*. The statistical parsimony network for this species pair (Fig. 2) centralizes the Indo-Pacific haplotypes of both species, with the haplotypes from the northwest Atlantic forming two distinct branches, one for each species. Both branches leading to the northwest Atlantic *C. altimus* and *C. plumbeus* experience the same character change events at nucleotide positions 382 and 619. However, both of these changes are third codon position transitions.

The lack of diagnostic characters may be the result of several factors. All four pairs are comprised of species that are closely related to one another and are morphologically very similar. In some cases, such as with *Carcharhinus obscurus* and *C. galapagensis*, the species are most easily distinguished by the number of vertebrae as morphometric ranges of characteristics like dorsal fin height overlap (Compagno 1984). It is possible that some misidentification of reference specimens has taken place. Misidentifications are inevitable, and provide justification for the link back to voucher specimens in DNA barcode records. In practice, these cases of suspected misidentification can be reviewed and corrected as necessary. The resulting statistical parsimony network for these two species is simplistic (Fig. 3). Two of the three haplotypes between these species were shared, with only a small group of *C. galapagensis* being distinguishable, although a group of *C. obscurus* did form their own branch save for the inclusion of a single *C. galapagensis* specimen. Mapping the source FAO regions for each sample with locality data onto the network did not yield any clear patterns, as specimens from multiple FAO regions were found across all haplotype nodes. The extent of haplotype sharing between these two species indicates more than just the occasional misidentified specimen. Assuming the taxonomy of two biologically valid species is correct, it may be that *C. obscurus* and *C. galapagensis* both retain ancestral polymorphisms and are simply too recently diverged to have accumulated fixed diagnostic differences. Alternatively, these 'species' may not be reproductively isolated.

The taxonomic status of *Rhizoprionodon terraenovae* and *R. porosus* has remained uncertain for decades, as there have been questions as to whether or not *R. porosus* is a separate species from *R. terraenovae* (Compagno 1984). Both species are from the western Atlantic, but the former is distributed through the Caribbean, Central and Southern America, while the latter is from the Gulf of Mexico and up along North America. Springer (1964) proposed the separation, but *R. porosus* has been considered a subspecies or clinal variant of *R. terraenovae* (Compagno 1984). The two entities have tentatively been left as separate species, but further taxonomic review has yet to take place. The *R. porosus* nodes on the statistical parsimony network (Fig. 4) appear to form two terminal branches, except for the position of a single sample labelled as *R. terraenovae*. As expected by their tentative species descriptions, all *R. terraenovae* samples were collected in the western central Atlantic to the northwestern Atlantic, while all *R. porosus* samples were collected in the western central Atlantic to the southwestern Atlantic. However, the majority of samples from each species were from the overlapping region of the western central Atlantic, containing the Caribbean, the Gulf of Mexico, and Central America. Samples from this region, for both species, are found throughout the various haplotype nodes in the network, except for a pair of terminal nodes, one for *R. terraenovae* and one for *R. porosus*, which are exclusively northwest Atlantic and southwest Atlantic respectively. These terminal haplotypes are not well represented though, and further sampling of these species from the north and west would be required to determine the predominant haplotype(s) in each region. The suite of changes between *R. terraenovae* and most of the *R. porosus* samples is not correlated with any geographical barrier.

Carcharhinus limbatus and *C. tilstoni* have had a similar unclear relationship, but the COI barcodes did exhibit a distinct phylogeographical pattern, which can be seen clearly in the statistical parsimony network (Fig. 5). In the Pacific, haplotypes were shared between both species, while the western Atlantic population of *C. limbatus* formed its own divergent branch. This separation allowed for the identification of a conditional ND (388-C + 397-T + 400-T + <u>334-A</u>) that is diagnostic for the western Atlantic population of *C. limbatus*. Originally the two species were thought to be a single species under the name *C. limbatus*, but *C. tilstoni* was distinguished and separated based on morphological and physiological differences (Stevens & Wiley 1986). The separation was later corroborated with allozyme electrophoresis and it was reported that the two species diverged approximately 200 000 years ago (Lavery & Shaklee 1991). However, the patterns described by the COI data suggest a scenario of incomplete lineage sorting. This is supported by a recent global phylogeographical study of *C. limbatus* (Keeney & Heist 2006) suggesting that the western Atlantic *C. limbatus* population has been isolated from the remainder of the *C. limbatus* populations as long as or longer than *C. limbatus* and *C. tilstoni* have been reproductively isolated.

Hart & Sunday (2007) suggest that statistical parsimony networks can provide an objective, non-arbitrary metric for genetic species differentiation. However, as with threshold values, the standard 95% connection limit of statistical parsimony networks cannot be applied generally to all species, and can be considered conservative with only 77.0% of the species in this study dissociating into a distinct network. Seventeen species did not form their own networks, including the eight species pairs that could not be diagnosed with NDs. Because the other nine species contained diagnostic NDs, networks involving these species appeared polarized, with haplotypes for each species clustering on opposite sides of the network (Table 4). Increasing the connection limit from 95% would reduce such cases of false-negatives, but will conversely increase the chances of false-positives (Hart & Sunday 2007). At the 95% limit, *Sphyrna lewini* (scalloped hammerhead) was the only case of a false-positive and compared with its congeners has the highest level of intraspecific variation for COI. It is worth noting that the cryptic lineage of hammerhead sharks detailed from within *S. lewini* (Abercrombie *et al*. 2005; Quattro *et al*. 2006) also forms its own network. *Sphyrna lewini*'s separation into two networks correlates with an Atlantic/East Indian (Madagascar) group and a Pacific group.

As an unexpected benefit, because nucleotide changes can be mapped onto the conservative statistical parsimony networks, they proved to be an effective way to identify nucleotide sites that could be used in a conditional ND. In some cases, the conditional NDs highlighted by the statistical parsimony networks apply to rarer haplotypes of species, represented by fewer specimens. Purposely merging distinct species networks by lowering the connection limit and observing mapped character changes may also be a viable approach for identifying NDs.

Complementary approaches

Traditional DNA barcoding phenetic-based measures and a character-based approach could have complementary roles in species identification. Although the concern with a distance measure is the potential for ambiguous results, it can provide an important service. The NDs in this study were meant to be as widely applicable as possible. That is, the context in which they were derived assumes that the availability of prior knowledge that could direct analysis may be minimal. While broad diagnostic characters are ideal when possible, the distance measure provided by BOLD is a fast and accurate way to narrow down the focus of a required ND if needed, assuming the BOLD identification engine doesn't immediately return an unambiguous 100% match to a reference DNA barcode. In other words, the added utility of the current BOLD identification engine would facilitate the possibility of using NDs that were identified in the context of a species pool with a narrower taxonomic breadth, thus creating a larger number of viable NDs.

NDs and development of diagnostic markers

Depending on their location within the DNA barcode, clear COI sNDs could also potentially be utilized for developing species-specific primers for use in multiplex PCR assays (Shivji *et al*. 2002; Chapman *et al*. 2003). They might also find application in nucleic acid hybridization probe construction and microarray development. Such nonsequencing assays are useful in situations where sequencing is problematic, but it is likely that there will be situations involving closely related species that require identification through cND evaluation of sequence haplotypes.

Conclusion

The generally successful identification of NDs in this study has highlighted the potential of a character-based diagnostic system for species identification. As a dynamic and scaleable approach, further studies and refinement will be necessary, especially the development of software tools capable of recognizing cNDs. In the cases where an ND is not available for a species, a conditional ND can sometimes be generated, but more importantly, these exceptional cases draw attention to groups that require further examination. In some situations an identification error may be uncovered, while in others the patterns seen within the DNA barcode data can set the groundwork that will support and feed into other studies, such as a more in depth exploration of the relationship between the unresolved species pairs encountered here.

For conservation efforts and other applied uses, DNA barcoding has already grown to encompass many elements that will allow it to be a powerful tool in those fields (Wong & Hanner 2008). The emergence of a standardized system is a significant innovation and step forward, and multiple international campaigns continue to expand the reference DNA barcode database for species across a wide range of taxa. Populating the reference database will take time, but BOLD does currently house DNA barcodes for approximately 43% of the world's shark diversity. In working with the ~17.5% subset of all shark species in this study alone, the framework and considerations of ND identification holds promise and could soon be applied to the much larger library of these important marine animals, and other groups of conservation interest.

An important step moving forward will be the verification of specimen identifications during database development. This is particularly problematic with sharks because voucher specimens are usually not retained given their large size. The effectiveness of a reference sequence database is dependent on its accuracy, and misidentified specimens can severely hinder attempts at identifying unknowns. However, as highlighted by NDs and the general DNA barcode sequences, there are several areas of potential taxonomic uncertainty within the sharks, requiring further investigation by shark taxonomists.

Acknowledgements

We would like to thank the following individuals for access to reference specimens included in this study: Andy Bentley and Ed Wiley from the Natural History Museum and Biodiversity Research Center at the University of Kansas, as well as Kathy Moore and Trey Knott from the Center for Coastal Environmental Health and Biomolecular Research at Charleston (NOAA/NOS). We thank our colleagues at the Canadian Centre for DNA Barcoding and the BOLD analytics staff at the Biodiversity Institute of Ontario for technical support. We thank Rob DeSalle, graduate students from Paul Hebert's lab at the University of Guelph and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript. This research was supported through funding to the Canadian Barcode of Life Network from Genome Canada (through the Ontario Genomics Institute), NSERC and other sponsors (listed at http://www.bolnet.ca/ps_overview.php).

Conflict of interest statement

The authors have declared the following competing interests: Robert Hanner receives funding from the Canadian Barcode of Life Network and serves as a member of its Secretariat. The remaining authors have no conflicts to declare.

References

Abercrombie DL, Clarke SC, Shivji MS (2005) Global-scale genetic identification of hammerhead sharks: application to assessment of the international fin trade and law enforcement. *Conservation Genetics*, **6**, 755–788.

Broughton RE, Milam JE, Roe BA (2001) The complete sequence of the zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. *Genome Research*, **11**, 1958–1967.

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequence of primates: tempo and mode of evolution. *Journal of Molecular Evolution*, **18**, 225–239.

Chapman D, Abercrombie D, Douady C, Pikitch E, Stanhope M, Shivji M (2003) A streamlined, bi-organelle, multiplex PCR approach to species identification: application to global conservation and trade monitoring of the great white shark, *Carcharodon carcharias*. *Conservation Genetics*, **4**, 415–425.

Clarke S, McAllister M, Michielsens C (2004) Estimates of shark species composition and numbers associated with the shark fin trade based on Hong Kong auction data. *Journal of Northwest Atlantic Fish Science*, **35**, 453–465.

Clarke SC, McAllister MK, Milner-Gulland EJ *et al*. (2006) Global estimates of shark catches using trade records from commercial markets. *Ecology Letters*, **9**, 1115–1126.

Compagno LJV (1984) FAO species catalogue Vol. 4, Part 2 Sharks of the World. An annotated and illustrated catalogue of shark species known to date. *FAO Fisheries Synopsis*, **125**, 251–655.

DeSalle R (2006) What's in a character? *Journal of Biomedical Informatics*, **39**, 6–17.

DeSalle R (2007) Phenetic and DNA taxonomy; a comment on Waugh. *Bioessays*, **29**, 1289–1290.

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.

Ferguson JWH (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*, **75**, 509–516.

Frézal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution*, **8**, 727–736.

Greig TW, Moore MK, Woodley CM, Quattro JM (2005) Mitochondrial gene sequences useful for species identification of western North Atlantic Ocean sharks. *Fishery Bulletin*, **103**, 516–523.

Hanner R (2005) Proposed standards for BARCODE records in INSDC. Database Working Group, Consortium for the Barcode of Life. http://www.barcoding.si.edu/PDF/DWG_data_standards-Final.pdf (last accessed 19 March 2009).

Hart MW, Sunday J (2007) Things fall apart: biological species form unconnected parsimony networks. *Biology Letters*, **3**, 509–512.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identification through DNA barcodes. *Proceedings of the Royal Society 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 B: Biological Sciences*, **270**, S96–S99.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences, USA*, **101**, 14812–14817.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.

Holmes BH, Steinke D, Ward RD (2009) Identification of shark and ray fins using DNA barcoding. *Fisheries Research*, **95**, 280–288.

Ivanova NV, DeWaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6**, 998–1002.

Ivanova N, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes*, **7**, 544–548.

Keeney DB, Heist EJ (2006) Worldwide phylogeography of the blacktip shark (*Carcharhinus limbatus*) inferred from mitochondrial DNA reveals isolation of western Atlantic populations coupled with recent Pacific dispersal. *Molecular Ecology*, **15**, 3669–3679.

Kelly RP, Sarkar IN, Eernisse DJ, DeSalle R (2007) DNA barcoding using chitons (genus *Mopalia*). *Molecular Ecology Notes*, **7**, 177–183.

Köhler F (2007) From DNA taxonomy to barcoding — how a vague idea evolved into a biosystematic tool. *Mitteilungen aus dem Museum für Naturkunde in Berlin. Zooligische Reihe*, **83**(Suppl.), 44–51.

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.

Lavery S, Shaklee JB (1991) Genetic evidence for separation of two sharks, *Carcharhinus limbatus* and *C. tilstoni*, from northern Australia. *Marine Biology*, **108**, 1–4.

Magnussen JE, Pikitch EK, Clarke SC *et al*. (2007) Genetic tracking of basking shark products in international trade. *Animal Conservation*, **10**, 199–207.

Messing J (1983) New M13 vectors for cloning. *Methods in Enzymology*, **101**, 20–78.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PloS Biology*, **3**, 2229–2238.

Ogden R (2008) Fish forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries*, **9**, 462–472.

Pank M, Stanhope M, Natanson L, Kohler N, Shivji M (2001) Rapid and simultaneous identification of body parts from the morphologically similar sharks *Carcharhinus obscurus* and *Carcharhinus plumbeus* (Carcharhinidae) using multiplex PCR. *Marine Biotechnology*, **3**, 231–240.

Quattro JM, Stoner DS, Driggers WB *et al*. (2006) Genetic evidence of cryptic speciation within hammerhead sharks (Genus *Sphyrna*). *Marine Biology*, **148**, 1143–1155.

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 237–247.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Rubinoff D (2006) Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology*, **20**, 1026–1033.

Sarkar IN, Planet PJ, Bael TE *et al*. (2002a) Characteristic attributes in cancer microarrays. *Journal of Biomedical Informatics*, **35**.

Sarkar IN, Thornton JW, Planet PJ *et al*. (2002b) An automated phylogenetic key for classifying homeoboxes. *Molecular Phylogenetics and Evolution*, **24**, 388–399.

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.

Shivji MS, Chapman DD, Pikitch EK, Raymond PW (2005) Genetic profiling reveal illegal international trade in fins of the great white shark, *Carcharodon carcharias*. *Conservation Genetics*, **6**, 1035–1039.

Springer VG (1964) A revision of the carcharhinid shark genera Scoliodon, Loxodon, and Rhizoprionodon. *Proceedings of the US National Museum*, **115**, 559–632.

Stevens JD, Wiley PD (1986) Biology of two commercially important carcharhinid sharks from northern Australia. *Australian Journal of Marine and Freshwater Research*, **37**, 671–688.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–633.

Vannuccini S (1999) Shark utilization, marketing and trade. In: *FAO Fisheries Technical Paper No. 389*. FAO, Rome, Italy.

Ward RD, Zemlak 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.

Ward RD, Holmes BH, White WT, Last PR (2008) DNA barcoding Australasian chondrichthyans: results and potential uses in conservation. *Marine and Freshwater Research*, **59**, 57–71.

Waugh J, Huynen L, Millar C, Lambert D (2007) DNA barcoding of animal species — response to DeSalle. *Bioessays*, **30**, 92–93.

Wong EH-K, Hanner RH (2008) DNA barcoding detects market substitution in North American seafood. *Food Research International*, **41**, 828–837.