DNA BARCODING

Identifying gastropod spawn from DNA barcodes: possible but not yet practicable

N. PUILLANDRE,* E. E. STRONG,† P. BOUCHET,‡ M.-C. BOISSELIER,* A. COULOUX§ and S. SAMADI*

*UMR 7138, Systématique, adaptation, évolution (UPMC/IRD/MNHN/CNRS), Université Pierre et Marie Curie (UPMC), CP26, 57 rue Cuvier, 75231 Paris cedex 05, France, †Department of Invertebrate Zoology, Smithsonian Institution, National Museum of Natural History, MRC 163, PO Box 37012, Washington, DC 20013-7012, USA, ‡Muséum National d'Histoire Naturelle, 57 rue Cuvier, 75231 Paris cedex 05, France, §GENOSCOPE, Centre National de Séquençage, 91000 Evry, France

Abstract

Identifying life stages of species with complex life histories is problematic as species are often only known and/or described from a single stage. DNA barcoding has been touted as an important tool for linking life-history stages of the same species. To test the current efficacy of DNA barcodes for identifying unknown mollusk life stages, 24 marine gastropod egg capsules were collected off the Philippines in deep water and sequenced for partial fragments of the COI, 16S and 12S mitochondrial genes. Two egg capsules of known shallow-water Mediterranean species were used to calibrate the method. These sequences were compared to those available in GenBank and the Barcode of Life Database (BOLD). Using COI sequences alone, only a single Mediterranean egg capsule was identified to species, and a single Philippine egg capsule was identified tentatively to genus; all other COI sequences recovered matches between 76% and 90% with sequences from BOLD and Gen-Bank. Similarity-based identification using all three markers confirmed the Mediterranean specimens' identifications. A phylogenetic approach was also implemented to confirm similarity-based identifications and provide a higher-taxonomic identification when species-level identifications were not possible. Comparison of available GenBank sequences to the diversity curve of a well-sampled coral reef habitat in New Caledonia highlights the poor taxonomic coverage achieved at present in existing genetic databases, emphasizing the need to develop DNA barcoding projects for megadiverse and often taxonomically challenging groups such as mollusks, to fully realize its potential as an identification and discovery tool.

Keywords: Barcode of Life Database, DNA barcoding, egg capsules, GenBank, Neogastropoda

Received 23 October 2008; revision accepted 28 December 2008

Introduction

Identification of different life stages for species with complex life histories has long posed a challenge to biologists of almost every discipline as species are routinely known and/or described based only on a single stage. Identifying early life stages is particularly difficult for marine invertebrates including mollusks that often undergo dramatic metamorphoses from the egg to one or more larval stages before finally reaching juvenile and adult form. The issue is trivial in mollusk species with parental care, but these represent only a small minority, the vast majority dispersing their larvae or their spawn (e.g. Lebour 1937; Fretter & Graham 1962; Robertson 1974) sometimes vast distances through ocean currents.

Mollusk spawn are highly conservative within species and present features that are diagnostic at many taxonomic levels, providing another character set useful for understanding evolutionary relationships (e.g. Habe 1960; Robertson 1974; Bandel 1976a, b). But due to the difficulties inherent in identification, spawn have been described for

Correspondence: N. Puillandre, Fax: +33 140 79 37 43,
E-mail: puillandre@mnhn.fr

only several hundred species at best, with the tropical mollusk fauna particularly poorly known. This limits the utility of spawn in systematic and ecological studies.

The process of identifying mollusk spawn typically relies on serendipitous discoveries of individuals actively depositing eggs, painstaking and time-consuming breeding and rearing of larvae *ex situ*, or circumstantial evidence (species presence, abundance) and a process of elimination using criteria including substrate, spawn morphology, developmental mode, ovum size, ovary/ovum colour, etc. (e.g. D'Asaro 1970; Winner 1987; Gustafson *et al*. 1991). However, *in situ* observations or *ex situ* rearing of juveniles are often impractical or may be impossible, for example in vent (Lutz *et al*. 1986) and other deep-water species (present study).

Classic surveys of molluscan spawn have focused on northern temperate biomes with low numbers of species and are almost exclusively limited to species with intra-capsular development (crawl-away juveniles) allowing comparisons between near-hatching embryos and the protoconchs of known adults. Such conditions provide an adequately simple system facilitating the identification of larvae and spawn (Thorson 1940b). But even after careful scrutiny of the protoconchs, this approach may still be fallible and lead to incorrect identifications in a certain number of cases (e.g. Knudsen 1950) or will be inconclusive in poorly known systems and especially in species with planktonic larval development (e.g. Gustafson *et al*. 1991). In an extreme case, a parallel nomenclature was devised for egg capsules that could not be recognizably linked to benthic species (Tokioka 1950). Moreover, identification to species level can be complicated in some cases where spawn of several species are morphologically indistinguishable (e.g. simple gelatinous egg masses of some vetigastropods, Lebour 1937; Fretter & Graham 1962; some species of *Conus,* Kohn 1961).

DNA barcodes are a promising and expedient new tool for accurately identifying and linking the varied life-history stages of single species (Schindel & Miller 2005). Indeed, DNA barcoding has demonstrated its capacity to do so successfully in several animal groups (Blaxter 2004; Steinke *et al*. 2005; Thomas *et al*. 2005; Vences *et al*. 2005; Pegg *et al*. 2006; Ahrens *et al*. 2007), and has already been used for marine fauna to link larvae and adults (Victor 2007). Such approaches have also been used to explore biodiversity, leading to the discovery of previously unrecognized species collected from their planktonic larvae while the difficult-to-collect adults are still unknown (Barber & Boyce 2006). Among mollusks, and especially bivalves, molecular approaches have been used to identify larvae using rRNA probes for *in situ* hybridization (Le Goff-Vitry *et al*. 2007; Pradillon *et al*. 2007; Jones *et al*. 2008) or polymerase chain reaction–single strand conformation polymorphism (PCR-SSCP) combined with sequencing (Livi *et al*. 2006) of

taxonomically informative fragments of 18S rDNA. Such methods have been used primarily to confirm the identity of target species using species-specific probes developed from known adults and as such, have mostly practical utility in, for example, biomonitoring of target, commercial or invasive species. Although it has been claimed that these methods may have broader utility for species identification by overcoming some of the limitations of mitochondrial markers (introgression, pseudogenes) (Livi *et al*. 2006), 18S rDNA faces the limitation that it is a highly conserved gene and will have decreased ability to accurately distinguish between closely related species. Recently, DNA barcodes in combination with other gene fragments (H3, 16S, 18S) were used to link egg masses, larvae and adults in one family of gastropods (Naticidae) in the western Mediterranean — a geographically circumscribed area with a well documented fauna (Huelsken *et al*. 2008).

Here we assess the capacity of the animal barcode (COI), two additional mitochondrial markers (16S, 12S) and the existing genetic databanks [GenBank, Barcode of Life Database (BOLD)] to identify a set of unknown gastropod egg capsules collected in the megadiverse region of the Indo-Pacific as compared to identified gastropod egg cases from the well-documented Mediterranean fauna. We use a two-tiered process first involving similarity-based methods (the identification engine of BOLD and the BLAST algorithm on GenBank), previously used in DNA barcode identification (e.g. Wong & Hanner 2008), followed by phylogenetic reconstruction (using all available sequences in GenBank) to tentatively determine the sister taxa of unidentified egg capsules, when similarity-based methods are not effective.

Materials and methods

Sampling

Twenty-four egg capsules were collected in the Philippines on scattered hard substrates (e.g. stones, shells) from soft bottoms by trawling, between 150 and 1450 m, during the Aurora 2007 deep sea cruise off the east coast of Luzon. All were tentatively recognized as neogastropod capsules as most share the lenticular shape typical of many neogastropod egg cases (Bandel 1976b) (Fig. 1c–h); one egg capsule was identified as that of a species of *Conus* (Conidae, Neogastropoda, Fig. 1b) with the flask-like shape characteristic of the genus. All capsules were first photographed on the substrate, then placed in 95% ethanol. In addition, two identified egg capsules (EC1, EC2) from the French Mediterranean coast were used as a control, allowing us to assess the capacity of DNA barcodes to identify known samples. Both were readily identifiable to species as *Coralliophila meyendorffii* broods egg capsules in the mantle cavity, and adults of *Erosaria spurca* were found near the egg capsules (Fig. 1a).
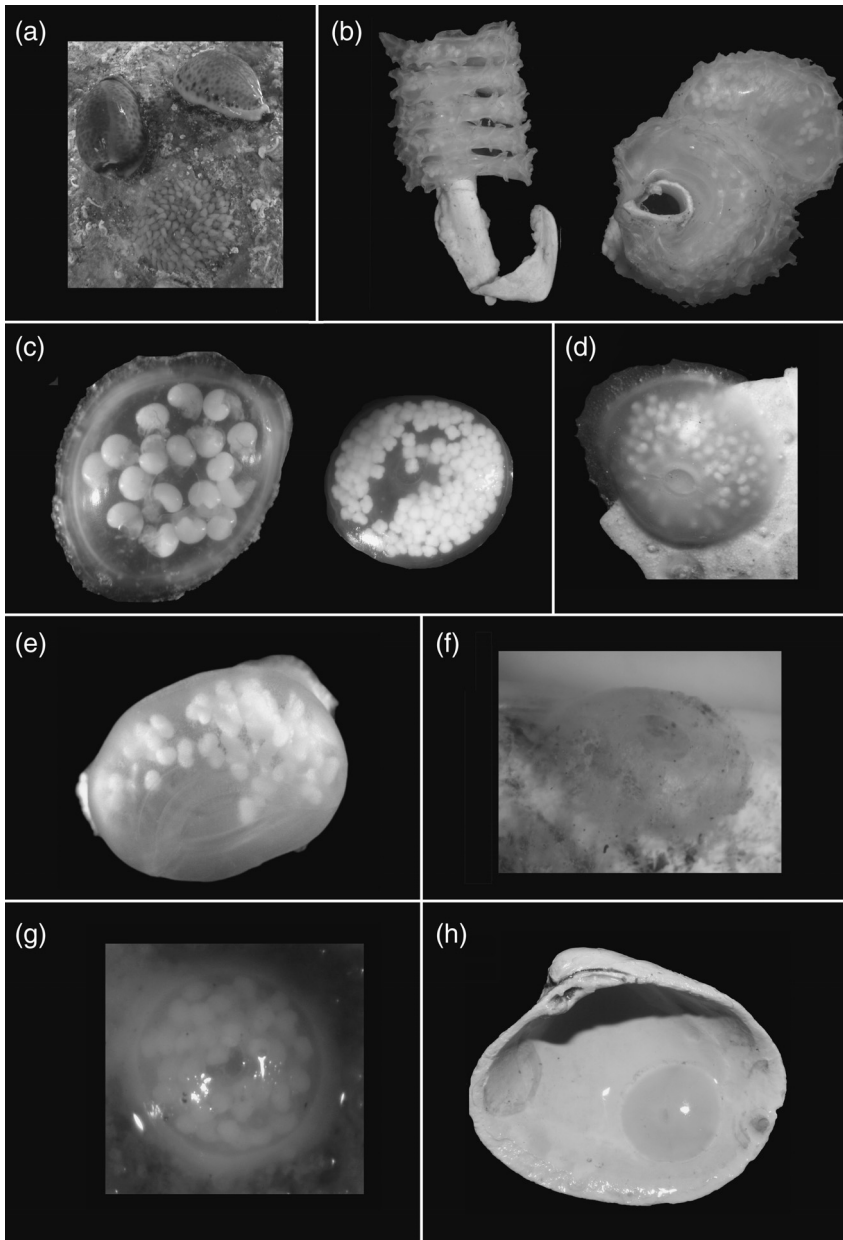
Fig. 1 Egg capsules used in molecular analyses and resulting identifications. (a) EC2 (*Erosaria spurca,* egg capsule with adults) (b) EC3 (*Conus* sp.) (c) EC9 (left) and EC11 (*Granulifusus* sp., at two different stages) (d) EC8 (*Comitas* sp.) (e) EC5 (*Belomitra* sp.) (f) EC14 (unidentified caenogastropod) (g) EC23 (Conidae, Clathurellinae) (h) EC26 (Conidae, Raphitominae) fixed on a bivalve shell. Photo credits: André Hoareau (a) and Barbara Buge (MNHN) (b–h). Colour photos are also published in Morpho-Bank (http://morphobank.geongrid.org/permalink/?P232).

Sequencing

DNA was extracted from the whole egg capsule, using 6100 Nucleic Acid Prepstation system (Applied Biosystem). Three gene fragments were amplified, corresponding to some of the most represented molluscan genes in GenBank, and also to genes commonly used at the species level (Hebert *et al*. 2003; Remigio & Hebert 2003): (i) a 658-bp fragment of the cytochrome oxidase I (COI) mitochondrial gene using universal primers LCO1490 and HCO2198 (Folmer *et al*. 1994), (ii) a 550-bp fragment of the 16S mitochondrial gene using primers 16SH (CGTGATCTGAG-TTCAGACCGG) and 16SL (GTTTACCAAAAACATGGC-TTC), and (iii) a 600-bp fragment of the 12S mitochondrial gene using primers 12SI (TGCCAGCAGYCGCGGTTA) and 12SIII (AGAGYGRCGGGCGATGTGT). All PCRs were performed in 25 µL, containing 3 ng of DNA, 1× reaction buffer, 2.5 mM MgCl$_2$, 0.26 mM dNTP, 0.3 µM of each primer, 5% DMSO and 1.5 U of Q-Bio *Taq*, QBiogene for all genes. Thermocycles used for COI gene are those described in Hebert *et al*. (2003); for 16S and 12S genes, they consisted in an initial denaturation step at 94 °C for 4′, followed by 30 cycles of denaturation at 94 °C for 30″, annealing at 52 °C for 16S and 54 °C for 12S. The final extension was at 72 °C for 10′. PCR products were purified and sequenced by the Genoscope. In all cases,

both directions were sequenced to confirm accuracy of each haplotype (GenBank Accession nos EU870520–EU870589).

Species identification

A two-tiered approach was employed to identify egg capsules, the first step involving similarity-based methods (identification engine of BOLD and BLAST search routine in GenBank) and a second step involving tree-based methods using Bayesian analysis of neogastropod sequences in GenBank. In the first step, each sequence was compared to available GenBank sequences using the BLASTN search routine as implemented in GenBank (default parameters). The best hit, as measured by percentage maximum sequence identity, was retained. In addition, each sequence was compared to all available barcode records in BOLD, using the identification engine BOLD-IDS, with the option 'searching all barcode records in BOLD'. This provides a list of similar sequences with the associated taxon name and the percentage sequence similarity. Contrary to the BLAST algorithm, identification in BOLD is based on genetic distances, and is not influenced by sequence length (Ratnasingham & Hebert 2007). A cut off value of 1% sequence divergence was used for considering two sequences conspecific (Ratnasingham & Hebert 2007).

In the second step, all egg capsule sequences were included in phylogenetic analyses to assess which sequences form molecular operational taxonomic units (OTU) and to evaluate the higher taxonomic placement of the OTUs. All neogastropod COI, 16S and 12S sequences were initially included, but to limit the total number of sequences, only one sequence per species was retained. An alignment with egg capsule sequences was generated using ClustalW multiple alignment implemented in BioEdit version 7.0.5.3 (Hall 1999) and only those sequences corresponding to the fragments sequenced for the egg capsules were retained. Ultimately, 159, 127 and 54 (for COI, 16S and 12S, respectively) GenBank sequences were used for the phylogenetic analyses. Phylogenetic trees were constructed using Bayesian inference with MrBayes (Huelsenbeck et al. 2001) (two Markov chains, 2 000 000 generations each with a sampling frequency of one tree each hundred generations, four parallel analyses).

Similarity-based methods (BLAST, BOLD-IDS) followed by phylogeny reconstruction were implemented for each gene, except for BOLD-IDS identifications as BOLD only contains COI sequences. Match scores are provided for BOLD searches with COI, and for BLAST searches with all three genes. For each sample, a final identification, corresponding to the best similarity score for the three genes (match score superior to 95%) and/or to the name of the sister taxa in the tree, is proposed based on the results of these analyses (see details in Table 1).

Results

Similarity-based identifications

Of the 26 samples (24 unknown, 2 known), only one of the known Mediterranean samples was identified to species using the BOLD identification engine (specimen EC2: 99.51% similarity with a sequence of *Erosaria spurca*), confirming the field-based identification (Table 1). Of the Philippine samples, only specimen EC8 was tentatively identified to genus, returning a match with a sequence of *Comitas* sp. (Turridae) at 97.84% similarity. All other COI sequences returned matches at 84% to 89% similarity with one or several sequences in BOLD, far exceeding the genetic distance considered to separate species (Hebert et al. 2003). Similarly, only the specimen of *E. spurca* was identified to species with COI alone using the BLAST search routine in GenBank (98% identity); all other first hits returned matches at 76% to 90% sequence identity.

The two additional mitochondrial markers sequenced in this analysis (16S, 12S) did not fare much better. Again, identification of specimen EC2 (*E. spurca*) was confirmed with the 16S gene (99% identity) and the 12S sequence of specimen EC1 returned a match with a sequence of *Coralliophila meyendorffii* at 99% identity, confirming the field-based identification of the second Mediterranean sample. These were the only matches at species level. However, twelve 16S egg capsule sequences produced matches with a 16S sequence of *Granulifusus niponicus* at between 95% and 98% identity, and the 16S sequence of specimen EC3 matched with a sequence of *Conus radiatus* at 95% identity; these results tentatively suggest a higher taxonomic identification at genus level (but see Discussion, below). All other matches for the 16S gene ranged from 82% to 89% identity, while those for the 12S gene ranged from only 79% to 94% identity.

Tree-based identifications

Among the 26 samples, 11 molecular OTUs were recognized using phylogeny reconstruction (Table 1). As shown for the 16S gene (Fig. 2), the phylogenetic placement of these OTUs can be used to confirm similarity-based identifications, but also to suggest supra-specific identifications when identification to species level is not possible based on sequence similarity. For example, with the 16S gene, the 12 sequences comprising OTU 4 form a clade with the GenBank sequence of *G. niponicus* (posterior probability PP = 0.98) (Fig. 2). EC23 (OTU 10) and EC26 (OTU 11) form a clade with *Raphitoma linearis* (Conidae, Raphitominae) with a PP of 1. With the 12S gene (results not shown), OTU 8 (EC 12, 15) clusters in a group with EC8 composed exclusively of Turridae species (PP = 0.97). With the COI gene (results not shown), OTU 5 (EC5, 6 and 7) is closely
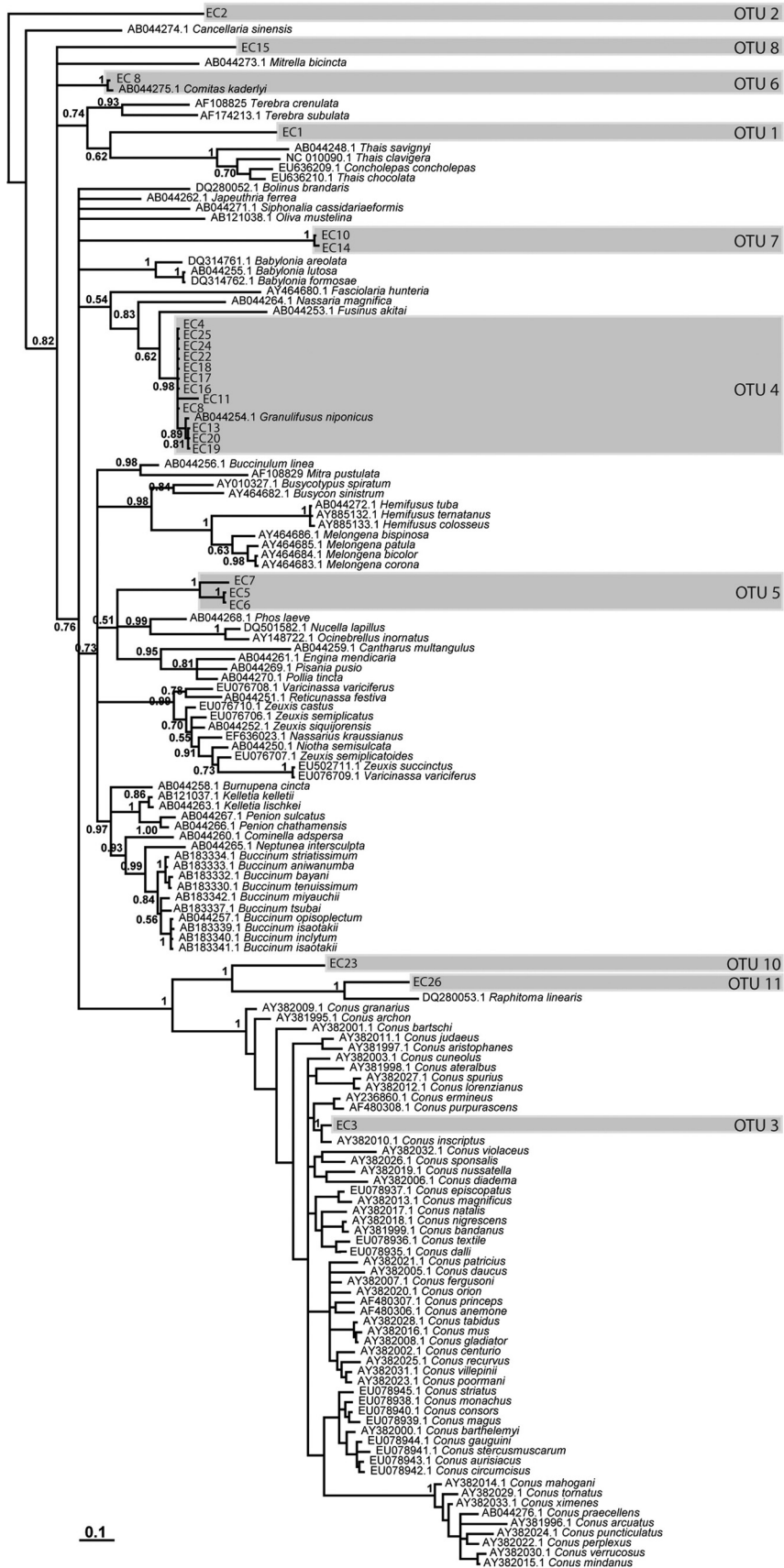
Table 1 Identification of egg capsules with COI, 16S and 12S genes using similarity- and tree-based methods. Match scores ≥ 95% sequence identity highlighted in bold. Tree-based identification for EC2 (*Erosaria spurca*, Cypraeidae) was not possible because only neogastropods sequences were included in the phylogenetic analyses. x: no match in Bold

| | | COI | | | | | 16S | | | 12S | | | |
| | | BOLD | | Genbank | | | Genbank | | | Genbank | | | |
No.	OTU	ID	Percentage	ID	Percentage	Tree ID	ID	Percentage	Tree ID	ID	Percentage	Tree ID	Final ID
EC1	1	x	x	*Fluminicola virens*	76		*Conus circumcisus*	82		*Coralliophila meyendorffii*	**99**	*C. meyendorffii*	*C. meyendorffii*
EC2	2	*Erosaria spurca*	**99.51**	*Erosaria spurca*	**98**		*Erosaria spurca*	**99**		*Cypraea annulus*	94		*Erosaria spurca*
EC3	3	*Conus sulcatus*	88.94	*Conus venulatus*	90	*Conus* sp.	*Conus radiatus*	**95**	*Conus* sp.	*Conus textile*	90	*Conus*	*Conus* sp.
EC4	4	*Busycon carica*	87.19	*Busycon sinistrum*	86		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC5	5	*Batillaria multiformis*	87.63	*Belomitra* sp.	89	*Belomitra* sp.	*Buccinum opisoplectum*	89		*Ilyanassa obsoleta*	84		*Belomitra* sp.
EC6	5	*Batillaria multiformis*	87.45	*Belomitra* sp.	89	*Belomitra* sp.	*Buccinum opisoplectum*	89		*Ilyanassa obsoleta*	83		*Belomitra* sp.
EC7	5	*Busycon sinistrum*	85.74	*Belomitra* sp.	89	*Belomitra* sp.	*Burnupena cincta*	89		*Gemmula sogodensis*	83		*Belomitra* sp.
EC8	6	*Comitas*	**97.84**	*Pyrgulopsis glandulosa*	85	*Lophiotoma*	*Comitas kaderlyi*	**98**	*C. kaderlyi*	*Gemmula rosario*	84		*Comitas* sp.
EC9	4	*Busycon carica*	87.26	*Nucella lapillus*	87		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	87		*Granulifus* sp.
EC10	7	*Lacuna pallidula*	84.64	*Ilyanessa obsoleta*	84		*Buccinulum linea*	86		*Ilyanassa obsoleta*	79		?
EC11	4						*Granulifusus niponicus*	**95**	*G. niponicus*	*Ilyanassa obsoleta*	87		*Granulifus* sp.
EC12	8									*Gemmula rosario*	83	Turridae	Turridae
EC13	4	*Busycon carica*	87.61	*Ilyanessa obsoleta*	86		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	85		*Granulifus* sp.
EC14	7						*Buccinulum linea*	85					?
EC15	8						*Penion chathamensis*	85		*Lophiotoma unedo*	79	Turridae	Turridae
EC16	4	*Busycon carica*	87.09	*Busycon sinistrum*	86		*Granulifusus niponicus*	**97**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC17	4	*Busycon carica*	87.3	*Busycon sinistrum*	86		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC18	4	*Busycon carica*	86.99	*Busycon sinistrum*	85		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC19	4	*Busycon carica*	87.26	*Acanthinucella punctulata*	84		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC20	4	*Busycon carica*	87.26	*Acanthinucella punctulata*	84		*Granulifusus niponicus*	**98**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC21	9									*Lophiotoma polytropa*	84	Cypraeidae	?
EC22	4	*Busycon carica*	87.37	*Busycon sinistrum*	86		*Granulifusus niponicus*	**97**	*G. niponicus*	*Ilyanassa obsoleta*	86		*Granulifus* sp.
EC23	10	*Nannodiella*	87.4	*Gymnobela* sp.	86	Clathurellinae	*Conus consors*	86	Conidae	*Ilyanassa obsoleta*	82		Conidae (Clathurellinae)
EC24	4	*Busycon carica*	87.19	*Busycon sinistrum*	86		*Granulifusus niponicus*	**97**	*G. niponicus*	*Ilyanassa obsoleta*	87		*Granulifus* sp.
EC25	4	*Busycon carica*	87.19	*Busycon sinistrum*	86		*Granulifusus niponicus*	**97**	*G. niponicus*	*Ilyanassa obsoleta*	87		*Granulifus* sp.
EC26	11	*Gymnobela*	87.52	*Nannodiella*	87	Raphitominae	*Raphitoma linearis*	87	Raphitominae	*Lophiotoma acuta*	81		Conidae (Raphitominae)

Fig. 2 Bayesian phylogram generated from all 16S neogastropod sequences in GenBank and 23 neogastropod egg capsule specimens from the Philippines and France. *Erosaria spurca* (EC2), a non-neogastropod, is the outgroup.

No. of egg cases	OTUs	Family ID?	Genus ID?	Species ID?
a 24	9	no → 2; yes → 7	no → 3; yes → 4	no → 4; yes → 0
b (MNHN sequences excluded) 24	9	no → 5; yes → 4	no → 1; yes → 3	no → 3; yes → 0
c (COI only) 19	7	no → 2; yes → 5	no → 2; yes → 3	no → 3; yes → 0
d (COI only, MNHN sequences excluded) 19	7	no → 5; yes → 2	no → 0; yes → 2	no → 2; yes → 0

Fig. 3 Number of positive identifications of unknown Philippine egg capsules at different taxonomic levels based on all genes (a, b) or COI only (c, d). (a) and (c) include all GenBank sequences; (b) and (d) exclude sequences produced by our own research group, demonstrating the impact of a single research group on the global data set.

related to a sequence of *Belomitra* sp. (Buccinidae) (PP = 1). This identification was already suggested with the BLAST search routine in GenBank, but only with 89% identity. The COI gene tree also identifies EC23 as a member of the Raphitominae (Conidae) and EC26 as a member of the Clathurellinae (Conidae), thereby refining the 16S familial placement of these two egg capsules.

Discussion

Egg capsule identification

Identification of the egg capsules of *Erosaria spurca* collected in France illustrates the capacity of the animal DNA barcode to successfully link the different life stages of a single gastropod species. However, only one of the two Mediterranean samples was identified by the COI barcode with a high level of precision (i.e. species level at > 98%

similarity). None of the Philippine egg capsules were identified with the same level of precision using COI alone. The threshold routinely used to consider two COI sequences as belonging to the same OTU ranges from 1% to 2% (Hebert *et al.* 2004; Bichain *et al.* 2007; Ratnasingham & Hebert 2007), and except for the specimen of *Erosaria spurca*, match scores for all other samples exceed this threshold. Similarly, only the Mediterranean samples retrieved matches at > 98% sequence similarity for the 16S and 12S markers; species identification was confirmed for specimen EC2 (*Erosaria spurca*) with 16S (99% identity) and for specimen EC1 (*Coralliophila meyendorffii*) with 12S (also 99% identity).

For many of the remaining samples, a supra-specific identification was possible based on match scores or using a combination of similarity- and tree-based methods. For example, although a genus-level genetic threshold is difficult to specify and will vary greatly between taxa (Holland *et al.* 2004), a low genetic distance (≤ 5%) tentatively supports a genus-level identification for OTU6 (97.84% — *Comitas* sp.) based on COI and for OTU3 (95% – *Conus radiatus*) and OTU4 (95%–98% — *Granulifusus niponicus*) based on 16S (see Fig. 3). The positioning of OTUs in the phylogenetic tree confirms these similarity-based identifications, but can also refine them in some cases. For example, the COI sequence for EC23 returned matches with *Nannodiella* (Conidae, Clathurellinae) in BOLD (87.4%) and with *Gymnobela* (Conidae, Raphitominae) in GenBank (86%), suggesting a family-level identification (Conidae). In the COI gene tree, EC23 is more closely related to *Nannodiella*, thereby suggesting that EC23 is a Clathurellinae.

Three egg capsules (OTU7, 9) could not be assigned even a tentative supra-specific identification based on the combination of similarity- and tree-based methods (see Table 1). Hit scores for these samples for all three genes are inferior to 90%; BLAST and BOLD results indicate that when the first matching sequence displays low sequence similarity (i.e. < 90%), the following matching sequences correspond (with lower or similar percentage) to completely different taxa. For example, BLAST results indicate that the COI sequence of EC3 matches first with *Lacuna pallidula* (Littorinidae) with a sequence identity of 85%, but then with *Nucella lamellosa* (Neogastropoda, Muricidae) with the same score, with *Ilyanassa obsoleta* (Neogastropoda, Nassaridae) at 84% identity and with *Urosalpinx perrugata* (Neogastropoda, Muricidae) also at 84%. Consequently, such scores are not useful for hypothesizing even superfamily-level identifications with a high degree of certainty. Moreover, the relationships of these samples were not robustly resolved in the phylogenetic analyses, preventing a family-level identification. However, this is not true in all cases. For example, OTU5 returned matches with *Belomitra* sp. (Buccinidae) in GenBank at 89% sequence identity,

and this identification is supported by the phylogenetic analyses. In the case of specimens EC12 and EC15 (OTU8), BOLD and GenBank best hits for 12S correspond to turrid species with low sequence similarity (*Gemmula rosario* for EC12, 83%, and *Lophiotoma unedo* for EC15, 79%), but the 12S gene tree tentatively supports this identification at family rank (Turridae).

Morphology of egg capsules

Specimen EC3 has a morphology typical of *Conus* species with its flask-like shape (Fig. 1b), and DNA analysis confirms this hypothesis. Specimen EC5, identified as *Belomitra* sp. (Buccinidae) is globular (Fig. 1e). Although *Belomitra* egg morphology has never been described, this result is consistent with the literature, as similar egg capsule shapes have been already reported for several buccinid genera (Thorson 1935, 1940a). All other egg capsules examined in the present study share a similar lenticular shape, with an escape aperture at the apex of the capsule, and containing a variable number of eggs. This morphology is not common among fasciolarids (see, e.g. Knudsen 1950) which instead usually have flask or vase-shaped capsules, sometimes with an undulating apical ridge or keel that surrounds the escape aperture. Here, lenticular capsules were tentatively identified as *Granulifusus* sp. (Fig. 1d) which would be the first record of an egg case for this genus.

Database completeness

Accuracy of DNA-based identification is dependent, of course, on the maturity of existing genetic databases. The impediment that incomplete databases pose for accurate and precise species identification has already been acknowledged, but how incomplete are they, and how does existing taxonomic coverage accurately mirror known diversity? The present analysis clearly indicates that even well-known faunas are inadequately represented in existing databases as evidenced by the fact only one of two samples from the French Mediterranean fauna could be identified to species using COI alone. It is surprising that even a well-known species as *C. meyendorffii* is not represented in databases for the COI and 16S genes; indeed, there is not a single 16S coralliophiline sequence and only one for the COI gene in these databases — a clade with roughly 250 species worldwide that form ecologically important associations with Cnidaria. Phylogenetic analysis can provide a supra-specific identification in cases where similarity-based methods return matches with only low similarity, and is the best alternative when no sequences of the same species are present in the databases. However, tree-based identifications should be treated with an appropriate amount of caution. The weak phylogenetic signal of each gene alone does not allow a robust resolution of the phylogenetic relationships among neogastropods and most family-level relationships are not well supported. Concatenation of all three genes to improve resolution of the tree is not possible due to high levels of terminal mismatch — yet another measure of the incompleteness of existing databases.

Tallying the number of sequences available provides a more concrete measure of database completeness, although we must assume that all specimens have been identified correctly. Despite this caveat, by this measure taxonomic diversity of mollusks is severely under-represented, as less than 16 000 molluscan COI sequences are currently published in GenBank, ostensibly corresponding to 3688 species, or less than 2% of the roughly 80 000 valid species (of which 53 000 are marine; Bouchet 2006) already described. But equally important is the taxonomic distribution of available sequences. In order to quantify taxonomic coverage, we compared gastropod species richness by family at a well-sampled tropical site in New Caledonia (Koumac, Bouchet *et al.* 2002) with the number of COI sequences available in GenBank over the last three years (Fig. 4). The results illustrate several important biases in family representation in GenBank. For example, five of the six most speciose families at Koumac (Triphoridae, Eulimidae, Pyramidellidae, Cerithiopsidae, Costellariidae) are each represented by less than three COI sequences. These families are all highly diverse and taxonomically complex with large numbers of minute species that classically have been overlooked in systematic studies. To complicate matters, they are almost exclusively predatory and typically occur in low abundance, with as many as 20% of the species known only from single specimens at any given site (Bouchet *et al.* 2002).

Conversely, some families are conspicuously over-represented, with the number of sequences available disproportionate to their known global diversity. Among the marine families, these over-represented groups tend to be large, charismatic 'sea shells' of interest to collectors and hobby naturalists (Cypraeidae, Muricidae). For example, the Cypraeidae comprises ~230 species worldwide (Lorenz 2002) and are represented by 682 COI sequences in GenBank. Whereas for Turridae *s.l.* (Conoideans except Terebridae and *Conus*), the most speciose assemblage at Koumac and with about 4000 named valid species (Tucker 2004) and perhaps as many as 10–20 000 in reality (Kantor *et al.* in press), there are only 92 COI sequences available. The other well-represented marine families (e.g. Littorinidae, Neritidae) are similarly disproportionately represented given their known diversity, and tend to be easily accessible and abundant in shallow, near-shore habitats.

A comparison of the number of sequences available in GenBank between 2006 and 2008 shows a slight increase in the number of sequences for most families, with several
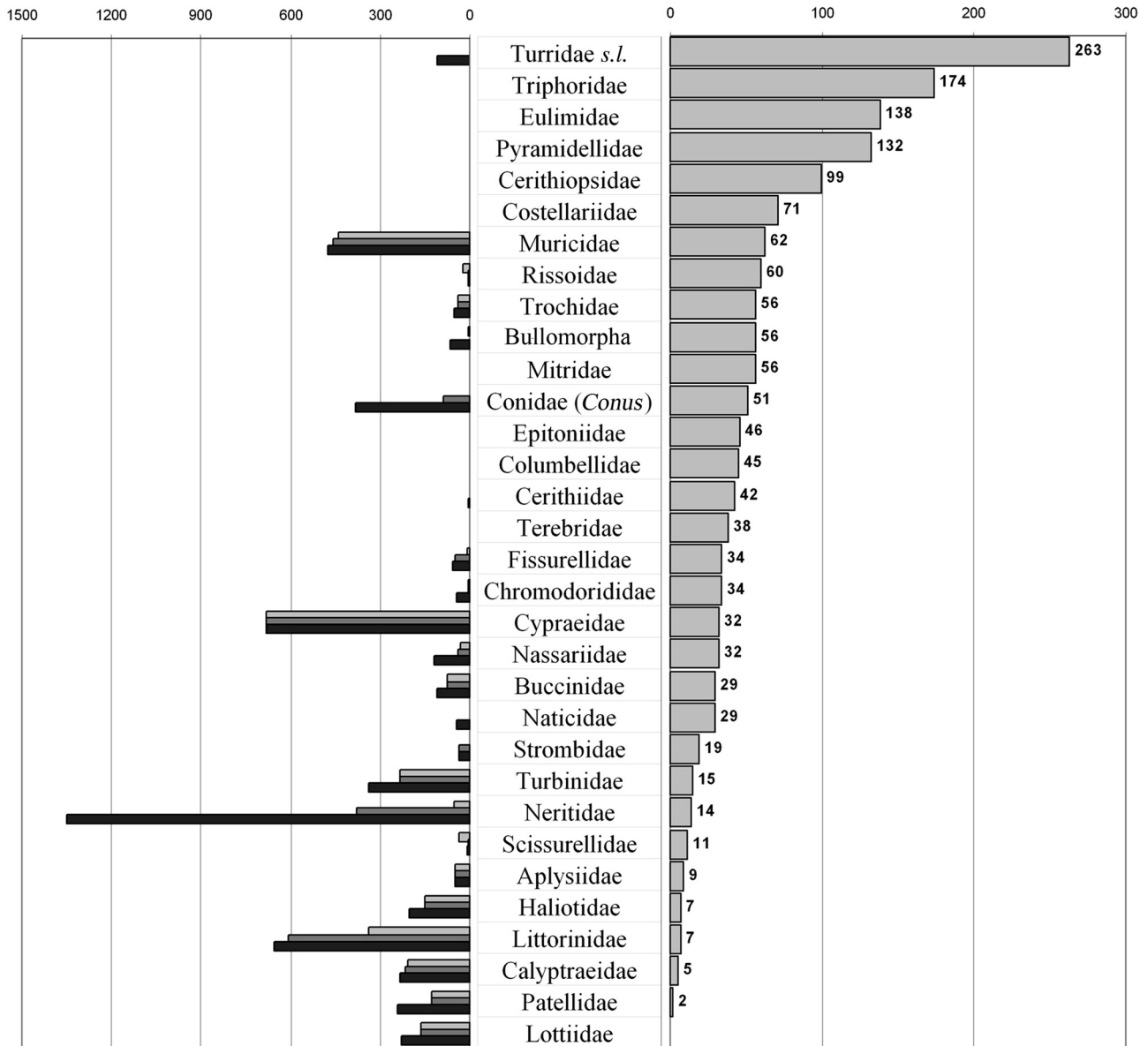
Fig. 4 GenBank coverage and actual marine gastropod diversity for gastropod families. Histogram on the left indicates number of COI sequences in GenBank: light grey, July 28, 2006; dark grey, April 6, 2007; black, July 4, 2008. Histogram on the right shows species richness of gastropod families at a well-sampled tropical site (Koumac, New Caledonia; Bouchet *et al.* 2002). Only groups with more than 10 species in Koumac or more than 100 sequences in GenBank (in 2006) are represented here. The decrease of the number of sequences for some groups through time is because some sequences were removed from GenBank.

showing more significant increases (e.g. *Conus*, Littorinidae, Neritidae, Turridae *s.l.*). The overall number of sequences and sudden increases in sequence availability can often be traced to the contributions of a single individual or research group. These may represent large numbers of sequences for a limited number of species (e.g. population genetic studies of Neritidae) or targeted systematic studies of particular taxonomic groups. For example, of the 92 COI turrid *s.l.* sequences available, all but one (109/110) were generated by our own research group. The potential impact of such targeted studies is evident when comparing the resolving power of existing databases including and excluding our own sequences (see Fig. 3). However, the majority of the most speciose families remain enormously under-represented.

Conclusion

The COI gene, the universal barcode for animals, in association with two other mitochondrial markers, has

demonstrated its ability to identify gastropod spawn, and thus constitutes an important tool for taxonomic identification of various animal life stages. However, our results indicate that sequence availability in existing genetic databases is disproportionately low given the known and estimated diversity of gastropods, and that the taxonomic coverage is highly biased towards shallow water species and/or highly collectible macro-mollusks. Consequently, barcodes are currently unable to provide species-level identifications for most unknowns, even for well-known, shallow water European species. In cases when species-level identification is impossible, tree-based methods are useful for refining the higher taxonomic placement of unidentified samples, but should be treated with caution. Barcoding efforts under development for targeted vertebrates (birds, fishes), 'charismatic' invertebrates or economically important groups currently allow positive identification to species in ~99% of unknowns in some cases (e.g. Wong & Hanner 2008), but similar projects should be developed for megadiverse groups such as mollusks, not only to facilitate taxonomic expertise but also to enhance species discovery in such poorly known groups.

Acknowledgements

References

Ahrens D, Monaghan MT, Vogler AP (2007) DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). *Molecular Phylogenetics and Evolution*, **44**, 436–449.

Bandel K (1976a) Observations on spawn, embryonic development and ecology of some Caribbean lower Mesogastropoda. *The Veliger*, **18**, 249–270.

Bandel K (1976b) Spawning, development and ecology of some higher Neogastropoda from the Carribean Sea of Colombia (South America). *The Veliger*, **19**, 176–193.

Barber P, Boyce SL (2006) Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae. *Proceedings of the Royal Society B: Biological Sciences*, **273**, 2053–2061.

Bichain JM, Boisselier MC, Bouchet P, Samadi S (2007) Delimiting species in the genus *Bythinella* (Mollusca: Caenogastropoda: Rissooidea): molecular and morphometric approachs. *Malacologia*, **49**, 291–311.

Blaxter ML (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **359**, 669–679.

Bouchet P (2006) The magnitude of marine biodiversity. In: *The Exploration of Marine Biodiversity — Scientific and Technological Challenges* (ed. Duarte C), pp. 31–62. Fundacion BBVA, Bilbao, Spain.

Bouchet P, Lozouet P, Maestrati P, Héros V (2002) Assessing the magnitude of species richness in tropical marine environments: exceptionnally high numbers of molluscs at a new Caledonian site. *Biological Journal of the Linnean Society*, **75**, 421–436.

D'Asaro CN (1970) Egg capsules of prosobranch mollusks from south Florida and the Bahamas, and notes on spawning in the laboratory. *Bulletin of Marine Science*, **20**, 414–440.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Fretter V, Graham A (1962) *British Prosobranch Molluscs; Their Functional Anatomy and Ecology*. Ray Society, London.

Gustafson RG, Littlewood DTJ, Lutz RA (1991) Gastropod egg capsules and their contents from deep-sea hydrothermal vent environments. *Biological Bulletin*, **180**, 34–55.

Habe T (1960) Egg masses and egg capsules of some Japanese marine prosobranchiate gastropod. *Bulletin of the Marine Biological Station of Asamushi*, **10**, 121–126.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA Barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, 1657–1663.

Holland BS, Dawson MN, Crow GL, Hofmann DK (2004) Global phylogeography of *Cassiopea* (Scyphozoa: Rhizostomeae): molecular evidence for cryptic species and multiple invasions of the Hawaiian Islands. *Marine Biology*, **145**, 1119–1128.

Huelsenbeck JP, Ronquist F, Hall B (2001) MrBayes: bayesian inference of phylogeny. *Bioinformatics*, **17**, 754–755.

Huelsken T, Marek T, Schreiber S, Schmidt I, Hollmann M (2008) The Naticidae (Mollusca: Gastropoda) of Giglio Island (Tuscany, Italy): shell characters, live animals, and a molecular analysis of egg masses. *Zootaxa*, **1770**, 1–40.

Jones WJ, Preston CM, Marin R, Scholin CA, Vrijenhoek RC (2008) A robotic molecular method for in situ detection of marine invertebrate larvae. *Molecular Ecology Resources*, **8**, 540–550.

Kantor YI, Puillandre N, Olivera BM, Bouchet P (2008) Morphological proxies for taxonomic decisions in turrids (Mollusca, Neogastropoda): a test of the value of shell and radula characters using molecular data. *Zoological Science*, **25**, 1156–1170.

Knudsen J (1950) Egg capsules and development of some marine prosobranchs from tropical West Africa. *Atlantide Report*, **1**, 85–130.

Kohn AJ (1961) Studies on spawning behavior, egg masses, and larval development in the gastropod genus Conus, Part I. Observation on nine species in Hawaii. *Pacific Science*, **15**, 163–179.

Le Goff-Vitry MC, Chipman AD, Comtet T (2007) *In situ* hybridization on whole larvae: a novel method for monitoring bivalve larvae. *Marine Ecology Progress Series*, **343**, 161–172.

Lebour MV (1937) The eggs and larvae of the British prosobranchs with special reference to those living in the plankton. *Journal of the Marine Biological Association of the United Kingdom*, **22**, 105–166.

Livi S, Cordisco C, Damiani C, Romanelli M, Crosetti D (2006) Identification of bivalve species at an early developmental stage through PCR-SSCP and sequence analysis of partial 18S rDNA. *Marine Biology*, **149**, 1149–1161.

Lorenz F (2002) *New Worldwide Cowries. Descriptions of New Taxa and Revisions of Selected Groups of Living Cypraeidae (Mollusca: Gastropoda)*. Conchbooks, Hackenheim, Germany.

Lutz R, Bouchet P, Jablonski D, Turner R, Warén A (1986) Larval ecology of mollusks at deep-sea hydrothermal vents. *American Malacological Bulletin*, **4**, 49–54.

Pegg GG, Sinclair B, Briskey L, Aspden WJ (2006) MtDNA barcode identification of fish larvae in the southern Great Barrier Reef, Australia. *Scientia Marina*, **70**, 7–12.

Pradillon F, Schmidt A, Peplies J, Dubilier N (2007) Species identification of marine invertebrate early stages by whole-larvae *in situ* hybridisation of 18S ribosomal RNA. *Marine Ecology Progress Series*, **333**, 103–116.

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

Remigio EA, Hebert PDN (2003) Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Molecular Phylogenetics and Evolution*, **29**, 641–647.

Robertson R (1974) Marine prosobranch gastropods: larval studies and systematics. *Thalassia Jugoslavica*, **10**, 213–238.

Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature*, **435**, 17.

Steinke D, Vences M, Salzburger W, Meyer A (2005) TaxI: a software tool for DNA barcoding using distance methods. *Proceedings of the Royal Society B: Biological Sciences*, **360**, 1975–1980.

Thomas M, Raharivolololoniaina M, Glaw F, Vences M, Vireites DR (2005) Montane Tadpoles in Madagascar: Molecular Identification and Description of the Larval Stages of *Mantidactylus elegans*, *Mantidactylus madecassus*, and *Boophis laurenti* from the Andringitra Massif. *Copeia*, **1**, 174–183.

Thorson G (1935) Studies on the egg-capsules and development of Arctic marine prosobranchs. *Meddelelser Om Grønland*, **100**, 20–73.

Thorson G (1940a) Notes on the egg-capsules of some North-Atlantic prosobranchs of the genus *Troschelia*, *Chysodomus*, *Volutopsis*, *Sipho* and *Trophon*. *Videnskabelige Meddelelser fra Dansk naturhistorisk Forening i Kobenhavn*, **104**, 251–265.

Thorson G (1940b) *Studies on the Egg Masses and Larval Development of Gastropoda from the Iranian Gulf. Danish Scientific Investigations in Iran, Part 2.* Ejnar Munksgaard, Copenhagen, Denmark.

Tokioka R (1950) Droplets from the plankton net. V. New names for egg capsules of littorinid gastropods. *Publication of the Seto Marine Biological Laboratory*, **1**, 151–152.

Tucker JK (2004) Catalog of Recent and fossil turrids (Mollusca: Gastropoda). *Zootaxa*, **682**, 1–1295.

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 (2007) *Coryphopterus kuna*, a new goby (Perciformes: Gobiidae: Gobiinae) from the western Caribbean, with the identification of the late larval stage and an estimate of the pelagic larval duration. *Zootaxa*, **1526**, 51–61.

Winner BE (1987) *A Field Guide to Molluscan Spawn*, Vol. 1. EBM, North Palm Beach, Florida.

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