



HAL
open science

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

N. Puillandre, E. E Strong, P. Bouchet, M.-C. Boisselier, A. Couloux, S
Samadi

► **To cite this version:**

N. Puillandre, E. E Strong, P. Bouchet, M.-C. Boisselier, A. Couloux, et al.. Identifying gastropod spawn from DNA barcodes: possible but not yet practicable. *Molecular Ecology Resources*, 2009, 9 (5), pp.1311-1321. 10.1111/j.1755-0998.2009.02576.x . hal-02458063

HAL Id: hal-02458063

<https://hal.science/hal-02458063v1>

Submitted on 28 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Identifying gastropod spawn from DNA barcodes:**

2 **possible but not yet practicable**

3
4 N. Puillandre*¹, E.E. Strong², P. Bouchet³, M.-C. Boisselier¹, A. Couloux⁴ & S. Samadi¹

5
6 ¹ Université Pierre et Marie Curie (UPMC), UMR 7138, Systématique, adaptation, évolution
7 (UPMC/IRD/MNHN/CNRS), CP26, 57 rue Cuvier, 75231 Paris Cedex 05, France.

8 ² Department of Invertebrate Zoology, Smithsonian Institution, National Museum of Natural
9 History, MRC 163, P.O. Box 37012, Washington, DC 20013-7012, USA.

10 ³ Muséum National d'Histoire Naturelle, 57 rue Cuvier, 75231 Paris Cedex 05, France.

11 ⁴ GENOSCOPE, Centre National de Séquençage, Evry, France.

12
13 Keywords: DNA barcoding, egg capsules, Neogastropoda, Barcode of Life Database,
14 GenBank

15
16
17
18 * Corresponding author: UMR 7138, Systématique, adaptation, évolution, UPMC; IRD;
19 MNHN; CNRS, Service de systématique moléculaire (CNRS, IFR 101), Département
20 systématique et évolution, Muséum National d'Histoire Naturelle, CP26, 57 rue Cuvier,
21 75231 Paris Cedex 05, France.; puillandre@mnhn.fr; +33 1 40 79 37 43

22
23 **Running title:** Barcoding gastropod spawn

24

25 **Abstract**

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

45

46 **Introduction**

47 Identification of different life stages for species with complex life histories has long posed a
48 challenge to biologists of almost every discipline as species are routinely known and/or
49 described based only on a single stage. Identifying early life stages is particularly difficult for
50 marine invertebrates including mollusks that often undergo dramatic metamorphoses from the
51 egg to one or more larval stages before finally reaching juvenile and adult form. The issue is
52 trivial in mollusk species with parental care, but these represent only a small minority, the
53 vast majority dispersing their larvae or their spawn (e.g. Lebour 1937; Fretter & Graham
54 1962; Robertson 1974) sometimes vast distances through ocean currents.

55

56 Mollusk spawn are highly conservative within species and present features that are diagnostic
57 at many taxonomic levels, providing another character set useful for understanding
58 evolutionary relationships (e.g. Habe 1960; Robertson 1974; Bandel 1976a,b). But due to the
59 difficulties inherent in identification, spawn have been described for only several hundred
60 species at best, with the tropical mollusk fauna particularly poorly known. This limits the
61 utility of spawn in systematic and ecological studies.

62

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

70

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

84

85 DNA barcodes are a promising and expedient new tool for accurately identifying and linking
86 the varied life history stages of single species (Schindel & Miller 2005). Indeed, DNA
87 barcoding has demonstrated its capacity to do so successfully in several animal groups
88 (Blaxter 2004; Steinke et al. 2005; Pegg et al. 2006; Thomas et al. 2005; Vences et al. 2005;
89 Ahrens et al. 2007), and has already been used for marine fauna to link larvae and adults
90 (Victor 2007). Such approaches have also been used to explore biodiversity, leading to the
91 discovery of previously unrecognized species collected from their planktonic larvae while the
92 difficult-to-collect adults are still unknown (Barber & Boyce 2006). Among mollusks, and
93 especially bivalves, molecular approaches have been used to identify larvae using rRNA

94 probes for *in situ* hybridization (Le Goff-Vitry et al. 2007; Pradillon et al. 2007; Jones et al.
95 2008) or polymerase chain reaction-single strand conformation polymorphism (PCR-SSCP)
96 combined with sequencing (Livi et al. 2006) of taxonomically informative fragments of 18S
97 rDNA. Such methods have been used primarily to confirm the identity of target species using
98 species-specific probes developed from known adults and as such, have mostly practical
99 utility in, e.g., biomonitoring of target, commercial or invasive species. Although it has been
100 claimed that these methods may have broader utility for species identification by overcoming
101 some of the limitations of mitochondrial markers (introgression, pseudogenes) (Livi et al.
102 2006), 18S rDNA faces the limitation that it is a highly conserved gene and will have
103 decreased ability to accurately distinguish between closely related species. Recently, DNA
104 barcodes in combination with other gene fragments (H3, 16S, 18S) were used to link egg
105 masses, larvae and adults in one family of gastropods (Naticidae) in the western
106 Mediterranean – a geographically circumscribed area with a well documented fauna
107 (Huelsenken et al. 2008).

108

109 Here, we assess the capacity of the animal barcode (COI), two additional mitochondrial
110 markers (16S, 12S) and the existing genetic databanks (GenBank, Barcode of Life Database -
111 BOLD) to identify a set of unknown gastropod egg capsules collected in the megadiverse
112 region of the Indo-Pacific as compared to identified gastropod egg cases from the well-
113 documented Mediterranean fauna. We use a two-tiered process first involving similarity-
114 based methods (the identification engine of BOLD and the BLAST algorithm on GenBank),
115 previously used in DNA barcode identification (e.g. Wong & Hanner 2008), followed by
116 phylogenetic reconstruction (using all available sequences in GenBank) to tentatively

117 determine the sister taxa of unidentified egg capsules, when similarity-based methods are not
118 effective.

119

120 **Material and methods**

121 *Sampling*

122 Twenty-four egg capsules were collected in the Philippines on scattered hard substrates (e.g.
123 stones, shells) from soft bottoms by trawling, between 150 and 1450 meters, during the
124 Aurora 2007 deep sea cruise off the east coast of Luzon. All were tentatively recognized as
125 neogastropod capsules as most share the lenticular shape typical of many neogastropod egg
126 cases (Bandel 1976b) (Fig. 1C-H); one egg capsule was identified as that of a species of
127 *Conus* (Conidae, Neogastropoda – Fig. 1B) with the flask-like shape characteristic of the
128 genus. All capsules were first photographed on the substrate, then placed in 95% ethanol. In
129 addition, two identified egg capsules (EC1, EC2) from the French Mediterranean coast were
130 used as a control, allowing us to assess the capacity of DNA barcodes to identify known
131 samples. Both were readily identifiable to species as *Coralliophila meyendorffii* broods egg
132 capsules in the mantle cavity, and adults of *Erosaria spurca* were found near the egg capsules
133 (Fig. 1A).

134

135 *Sequencing*

136 DNA was extracted from the whole egg capsule, using 6100 Nucleic Acid Prepstation system
137 (Applied Biosystem). Three gene fragments were amplified, corresponding to some of the
138 most represented molluscan genes in GenBank, and also to genes commonly used at the
139 species level (Hebert et al. 2003; Remigio & Hebert 2003): (i) a 658 bp fragment of the
140 Cytochrome Oxidase I (COI) mitochondrial gene using universal primers LCO1490 and

141 HCO2198 (Folmer et al. 1994), (ii) a 550 bp fragment of the 16S mitochondrial gene using
142 primers 16SH (CGTGATCTGAGTTCAGACCGG) and 16SL
143 (GTTTACCAAAAACATGGCTTC) and (iii) a 600 bp fragment of the 12S mitochondrial
144 gene using primers 12SI (TGCCAGCAGYCGCGGTTA) and 12SIII
145 (AGAGYGRCGGGCGATGTGT). All PCR reactions were performed in 25 µl, containing 3
146 ng of DNA, 1X reaction buffer, 2.5 mM MgCl₂, 0.26 mM dNTP, 0.3 µM of each primer, 5%
147 DMSO and 1.5 units of Q-Bio Taq, QBiogene for all genes. Thermocycles used for COI gene
148 are those described in Hebert et al. (2003); for 16S and 12S genes, they consisted in an initial
149 denaturation step at 94°C for 4', followed by 30 cycles of denaturation at 94°C for 30'',
150 annealing at 52°C for 16S and 54°C for 12S. The final extension was at 72°C for 10'. PCR
151 products were purified and sequenced by the Genoscope. In all cases, both directions were
152 sequenced to confirm accuracy of each haplotype (GenBank Accession Numbers: EU870520-
153 EU870589).

154

155 *Species identification*

156 A two-tiered approach was employed to identify egg capsules, the first step involving
157 similarity-based methods (identification engine of BOLD and BLAST search routine in
158 GenBank) and a second step involving tree-based methods using Bayesian analysis of
159 neogastropod sequences in GenBank. In the first step, each sequence was compared to
160 available GenBank sequences using the BLASTn search routine as implemented in GenBank
161 (default parameters). The best hit, as measured by percent maximum sequence identity, was
162 retained. In addition, each sequence was compared to all available barcode records in BOLD,
163 using the identification engine BOLD-IDS, with the option “searching all barcode records in
164 BOLD”. This provides a list of similar sequences with the associated taxon name and the

165 percent sequence similarity. Contrary to the BLAST algorithm, identification in BOLD is
166 based on genetic distances, and is not influenced by sequence length (Ratnasingham & Hebert
167 2007). A cut off value of 1% sequence divergence was used for considering two sequences
168 conspecific (Ratnasingham & Hebert 2007).

169

170 In the second step, all egg capsule sequences were included in phylogenetic analyses to assess
171 which sequences form molecular operational taxonomic units (OTUs) and to evaluate the
172 higher taxonomic placement of the OTUs. All neogastropod COI, 16S and 12S sequences
173 were initially included, but to limit the total number of sequences, only one sequence per
174 species was retained. An alignment with egg capsule sequences was generated using ClustalW
175 multiple alignment implemented in BioEdit version 7.0.5.3 (Hall 1999) and only those
176 sequences corresponding to the fragments sequenced for the egg capsules were retained.
177 Ultimately, 159, 127 and 54 (for COI, 16S and 12S, respectively) GenBank sequences were
178 used for the phylogenetic analyses. Phylogenetic trees were constructed using Bayesian
179 inference with MrBayes (Huelsenbeck et al. 2001) (two Markov chains, 2000000 generations
180 each with a sampling frequency of one tree each hundred generations, four parallel analyses).

181

182 Similarity-based methods (BLAST, BOLD-IDS) followed by phylogeny reconstruction were
183 implemented for each gene, except for BOLD-IDS identifications as BOLD only contains
184 COI sequences. Match scores are provided for BOLD searches with COI, and for BLAST
185 searches with all three genes. For each sample, a final identification, corresponding to the best
186 similarity score for the three genes (match score superior to 95%) and/or to the name of the
187 sister taxa in the tree, is proposed based on the results of these analyses (see details in Table
188 1).

189

190 **Results**

191 *Similarity-based identifications*

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

202

203 The two additional mitochondrial markers sequenced in this analysis (16S, 12S) did not fare
204 much better. Again, identification of specimen EC2 (*Erosaria spurca*) was confirmed with the
205 16S gene (99% identity) and the 12S sequence of specimen EC1 returned a match with a
206 sequence of *Coralliophila meyendorffii* at 99% identity, confirming the field-based
207 identification of the second Mediterranean sample. These were the only matches at species
208 level. However, twelve 16S egg capsule sequences produced matches with a 16S sequence of
209 *Granulifusus niponicus* at between 95% and 98% identity, and the 16S sequence of specimen
210 EC3 matched with a sequence of *Conus radiatus* at 95% identity; these results tentatively
211 suggest a higher taxonomic identification at genus level (but see Discussion, below). All other

212 matches for the 16S gene ranged from 82% to 89% identity, while those for the 12S gene
213 ranged from only 79% to 94% identity.

214

215 *Tree-based identifications*

216 Among the 26 samples, 11 molecular OTUs were recognized using phylogeny reconstruction
217 (Table 1). As shown for the 16S gene (Fig. 2), the phylogenetic placement of these OTUs can
218 be used to confirm similarity-based identifications, but also to suggest supra-specific
219 identifications when identification to species level is not possible based on sequence
220 similarity. For example, with the 16S gene, the 12 sequences comprising OTU 4 form a clade
221 with the GenBank sequence of *G. niponicus* (posterior probability PP = 0.98) (Fig. 2). EC23
222 (OTU 10) and EC26 (OTU 11) form a clade with *Raphitoma linearis* (Conidae,
223 Raphitominae) with a PP of 1. With the 12S gene (results not shown), OTU 8 (EC 12, 15)
224 clusters in a group with EC8 composed exclusively of Turridae species (PP = 0.97). With the
225 COI gene (results not shown), OTU 5 (EC5, 6 and 7) is closely related to a sequence of
226 *Belomitra* sp. (Buccinidae) (PP = 1). This identification was already suggested with the
227 BLAST search routine in GenBank, but only with 89% identity. The COI gene tree also
228 identifies EC23 as a member of the Raphitominae (Conidae) and EC26 as a member of the
229 Clathurellinae (Conidae), thereby refining the 16S familial placement of these two egg
230 capsules.

231

232 **Discussion**

233 *Eggs capsule identification*

234 Identification of the egg capsules of *Erosaria spurca* collected in France illustrates the
235 capacity of the animal DNA barcode to successfully link the different life stages of a single

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

246

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

259

260 Three egg capsules (OTU7, 9) could not be assigned even a tentative supra-specific
261 identification based on the combination of similarity- and tree-based methods (see Table 1).
262 Hit scores for these samples for all three genes are inferior to 90%; BLAST and BOLD results
263 indicate that when the first matching sequence displays low sequence similarity (i.e. <90%),
264 the following matching sequences correspond (with lower or similar percentage) to
265 completely different taxa. For example, BLAST results indicate that the COI sequence of EC3
266 matches first with *Lacuna pallidula* (Littorinidae) with a sequence identity of 85%, but then
267 with *Nucella lamellosa* (Neogastropoda, Muricidae) with the same score, with *Ilyanassa*
268 *obsoleta* (Neogastropoda, Nassaridae) at 84% identity and with *Urosalpinx perrugata*
269 (Neogastropoda, Muricidae) also at 84%. Consequently, such scores are not useful for
270 hypothesizing even superfamily-level identifications with a high degree of certainty.
271 Moreover, the relationships of these samples were not robustly resolved in the phylogenetic
272 analyses, preventing a family-level identification. However, this is not true in all cases. For
273 example, OTU5 returned matches with *Belomitra* sp. (Buccinidae) in GenBank at 89%
274 sequence identity, and this identification is supported by the phylogenetic analyses. In the
275 case of specimens EC12 and EC15 (OTU8), BOLD and GenBank best hits for 12S
276 correspond to turrid species with low sequence similarity (*Gemmula rosario* for EC12, 83%,
277 and *Lophiotoma unedo* for EC15, 79%), but the 12S gene tree tentatively supports this
278 identification at family rank (Turridae).

279

280 *Morphology of eggs capsules*

281 Specimen EC3 has a morphology typical of *Conus* species with its flask-like shape (Fig. 1 B),
282 and DNA analysis confirms this hypothesis. Specimen EC5, identified as *Belomitra* sp.
283 (Buccinidae) is globular (Fig. 1 E). Although *Belomitra* egg morphology has never been

284 described, this result is consistent with the literature, as similar egg capsule shapes have been
285 already reported for several buccinid genera (Thorson 1935, 1940a). All other egg capsules
286 examined in the present study share a similar lenticular shape, with an escape aperture at the
287 apex of the capsule, and containing a variable number of eggs. This morphology is not
288 common among fasciolarids (see e.g. Knudsen 1950) which instead usually have flask or
289 vase-shaped capsules, sometimes with an undulating apical ridge or keel that surrounds the
290 escape aperture. Here, lenticular capsules were tentatively identified as *Granulifusus* sp. (Fig.
291 1 D) which would be the first record of an egg case for this genus.

292

293 *Database completeness*

294 Accuracy of DNA-based identification is dependent, of course, on the maturity of existing
295 genetic databases. The impediment that incomplete databases pose for accurate and precise
296 species identification has already been acknowledged, but how incomplete are they, and how
297 does existing taxonomic coverage accurately mirror known diversity? The present analysis
298 clearly indicates that even well-known faunas are inadequately represented in existing
299 databases as evidenced by the fact only one of two samples from the French Mediterranean
300 fauna could be identified to species using COI alone. It is surprising that even a well-known
301 species as *Coralliophila meyendorffii* is not represented in databases for the COI and 16S
302 genes; indeed, there is not a single 16S coralliophiline sequence and only one for the COI
303 gene in these databases – a clade with roughly 250 species worldwide that form ecologically
304 important associations with Cnidaria. Phylogenetic analysis can provide a supra-specific
305 identification in cases where similarity-based methods return matches with only low
306 similarity, and is the best alternative when no sequences of the same species are present in the
307 databases. However, tree-based identifications should be treated with an appropriate amount

308 of caution. The weak phylogenetic signal of each gene alone does not allow a robust
309 resolution of the phylogenetic relationships among neogastropods and most family-level
310 relationships are not well supported. Concatenation of all three genes to improve resolution of
311 the tree is not possible due to high levels of terminal mismatch – yet another measure of the
312 incompleteness of existing databases.

313

314 Tallying the number of sequences available provides a more concrete measure of database
315 completeness, although we must assume that all specimens have been identified correctly.
316 Despite this caveat, by this measure taxonomic diversity of mollusks is severely
317 underrepresented, as less than 16,000 molluscan COI sequences are currently published in
318 GenBank, ostensibly corresponding to 3,688 species, or less than 2% of the roughly 80,000
319 valid species (of which 53,000 are marine; Bouchet 2006) already described. But equally
320 important is the taxonomic distribution of available sequences. In order to quantify taxonomic
321 coverage, we compared gastropod species richness by family at a well-sampled tropical site in
322 New Caledonia (Koumac - Bouchet et al. 2002) with the number of COI sequences available
323 in GenBank over the last three years (Fig. 4). The results illustrate several important biases in
324 family representation in GenBank. For example, five of the six most speciose families at
325 Koumac (Triphoridae, Eulimidae, Pyramidellidae, Cerithiopsidae, Costellariidae) are each
326 represented by less than three COI sequences. These families are all highly diverse and
327 taxonomically complex with large numbers of minute species that classically have been
328 overlooked in systematic studies. To complicate matters, they are almost exclusively
329 predatory and typically occur in low abundance, with as many as 20% of the species known
330 only from single specimens at any given site (Bouchet et al. 2002).

331

332 Conversely, some families are conspicuously over-represented, with the number of sequences
333 available disproportionate to their known global diversity. Among the marine families, these
334 over-represented groups tend to be large, charismatic “sea shells” of interest to collectors and
335 hobby naturalists (Cypraeidae, Muricidae). For example, the Cypraeidae comprises ~230
336 species worldwide (Lorenz 2002) and are represented by 682 COI sequences in GenBank.
337 Whereas for Turridae *s.l.* (Conoideans except Terebridae and *Conus*), the most speciose
338 assemblage at Koumac and with about 4,000 named valid species (Tucker 2004) and perhaps
339 as many as 10-20,000 in reality (Kantor et al. in press), there are only 92 COI sequences
340 available. The other well-represented marine families (e.g. Littorinidae, Neritidae) are
341 similarly disproportionately represented given their known diversity, and tend to be easily
342 accessible and abundant in shallow, near shore habitats.

343

344 A comparison of the number of sequences available in GenBank between 2006 and 2008
345 shows a slight increase in the number of sequences for most families, with several showing
346 more significant increases (e.g. *Conus*, Littorinidae, Neritidae, Turridae *s.l.*). The overall
347 number of sequences and sudden increases in sequence availability can often be traced to the
348 contributions of a single individual or research group. These may represent large numbers of
349 sequences for a limited number of species (e.g. population genetic studies of Neritidae) or
350 targeted systematic studies of particular taxonomic groups. For example, of the 92 COI turrid
351 *s.l.* sequences available, all but one (109/110) were generated by our own research group. The
352 potential impact of such targeted studies is evident when comparing the resolving power of
353 existing databases including and excluding our own sequences (see Fig. 3). However, the
354 majority of the most speciose families remain enormously underrepresented.

355

356 **Conclusion**

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

372

373 **References**

- 374 Ahrens D, Monaghan MT, Vogler AP (2007) DNA-based taxonomy for associating adults
375 and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae).
376 *Molecular Phylogenetics and Evolution* 44, 436–449.
- 377 Bandel K (1976a) Observations on spawn, embryonic development and ecology of some
378 Caribbean lower Mesogastropoda. *The Veliger* 18, 249-270.

379 Bandel K (1976b) Spawning, Development and ecology of some higher neogastropoda from
380 the Carribean Sea of Colombia (South America). *The Veliger* 19, 176-193.

381 Barber P, Boyce SL (2006) Estimating diversity of Indo-Pacific coral reef stomatopods
382 through DNA barcoding of stomatopod larvae. *Proceedings of the Royal Society B*
383 273, 2053-2061.

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

387 Blaxter ML (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the*
388 *Royal Society of London B* 359, 669-679.

389 Bouchet P (2006) The magnitude of marine biodiversity. In: *The Exploration of Marine*
390 *Biodiversity - Scientific and Technological Challenges* (ed. BBVA F).

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

394 D'Asaro CN (1970) Egg capsules of prosobranch mollusks from south Florida and the
395 Bahamas, and notes on spawning in the laboratory. *Bulletin of Marine Science* 20,
396 414-440.

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

400 Fretter V, Graham A (1962) *British Prosobranch Molluscs; their functional anatomy and*
401 *ecology* Ray Society, London.

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

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

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

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

410 Hebert PDN, Stoeckle MY, Zemplak TS, Francis CM (2004) Identification of birds through
411 DNA barcodes. *PLOS Biology* 2, 1657-1663.

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

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

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

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

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

- 426 Knudsen J (1950) Egg capsules and development of some marine prosobranchs from tropical
427 West Africa. *Atlantide Report* 1, 85-130.
- 428 Kohn AJ (1961) Studies on spawning behavior, egg masses, and larval development in the
429 gastropod genus *Conus*, Part I. Observation on nine species in Hawaii. *Pacific Science*
430 15, 163-179.
- 431 Le Goff-Vitry MC, Chipman AD, Comtet T (2007) *In situ* hybridization on whole larvae: a
432 novel method for monitoring bivalve larvae. *Marine Ecology progress Series* 343,
433 161-172.
- 434 Lebour MV (1937) The eggs and larvae of the British prosobranchs with special reference to
435 those living in the plankton. *Journal of the Marine Biological Association of the*
436 *United Kingdom* 22, 105-166.
- 437 Livi S, Cordisco C, Damiani C, Romanelli M, Crosetti D (2006) Identification of bivalve
438 species at an early developmental stage through PCR-SSCP and sequence analysis of
439 partial 18S rDNA. *Marine Biology* 149, 1149-1161.
- 440 Lorenz F (2002) *New Worldwide cowries. Descriptions of new taxa and revisions of selected*
441 *groups of living Cypraeidae (Mollusca: Gastropoda)* Conchbooks.
- 442 Lutz R, Bouchet P, Jablonski D, Turner R, Warén A (1986) Larval ecology of mollusks at
443 deep-sea hydrothermal vents. *American Malacological Bulletin* 4, 49-54.
- 444 Pegg GG, Sinclair B, Briskey L, Aspden WJ (2006) MtDNA barcode identification of fish
445 larvae in the southern Great Barrier Reef, Australia. *Scientia Marina* 70, 7-12.
- 446 Pradillon F, Schmidt A, Peplies J, Dubilier N (2007) Species identification of marine
447 invertebrate early stages by whole-larvae *in situ* hybridisation of 18S ribosomal RNA.
448 *Marine Ecology progress Series* 333, 103-116.

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

451 Remigio EA, Hebert PDN (2003) Testing the utility of partial COI sequences for phylogenetic
452 estimates of gastropod relationships. *Molecular Phylogenetics and Evolution* 29, 641-
453 647.

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

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

457 Steinke D, Vences M, Salzburger W, Meyer A (2005) TaxI: a software tool for DNA
458 barcoding using distance methods. *Proceedings of the Royal Society B* 360, 1975-
459 1980.

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

464 Thorson G (1935) Studies on the egg-capsules and development of Arctic marine
465 prosobranchs. *Medd. Om Gronland* 100, 20-73.

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

469 Thorson G (1940b) *Studies on the egg masses and larval development of Gastropoda from the*
470 *Iranian Gulf. Danish Scientific Investigations in Iran, Part 2* Ejnar Munksgaard,
471 Copenhagen.

- 472 Tokioka R (1950) Droplets from the plankton net. V. New names for egg capsules of littorinid
473 gastropods. *Publication of the Seto Marine Biological Laboratory* 1, 151-152.
- 474 Tucker JK (2004) Catalog of Recent and fossil turrids (Mollusca: Gastropoda). *Zootaxa* 682,
475 1-1295.
- 476 Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity
477 through DNA barcoding: chances and challenges. *Philosophical Transactions of the*
478 *Royal Society B: Biological Sciences* 360, 1859-1868.
- 479 Victor BC (2007) *Coryphopterus kuna*, a new goby (Perciformes: Gobiidae: Gobiinae) from
480 the western Caribbean, with the identification of the late larval stage and an estimate
481 of the pelagic larval duration. *Zootaxa* 1526, 51-61.
- 482 Winner BE (1987) *A field guide to molluscan spawn. Vol. 1*. EBM, North Palm Beach,
483 Florida.
- 484 Wong EH-K, Hanner RH (2008) DNA barcoding detects market substitution in North
485 American seafood. *Food Research International* 41, 828-837.

486

487

488 **Acknowledgments**

489 The AURORA 2007 cruise (Principal Investigators: Marivene Manuel, NMP, and Philippe
490 Bouchet, MNHN) on board M/V DA-BFAR was a joint project between the Philippine
491 Bureau of Fisheries and Aquatic Resources (BFAR), National Museum of the Philippines
492 (NMP), and Muséum National d'Histoire Naturelle (MNHN). It was made possible by a grant
493 from the Lounsbery Foundation, and was carried under a *Census of Marine Life / Census of*
494 *Margins* umbrella. This work was supported by the "Consortium National de Recherche en
495 Génomique", and the "Service de Systématique Moléculaire" of the Muséum National

496 d'Histoire Naturelle (IFR 101). It is part of the agreement n°2005/67 between the Genoscope
497 and the Muséum National d'Histoire Naturelle on the project "Macrophylogeny of life"
498 directed by Guillaume Lecointre". We are grateful to André Hoareau and Jacques Pelorce for
499 the two specimens collected in France. We are also pleased to thank Barbara Buge (MNHN)
500 for the pictures of the egg capsules and Yuri Kantor (Russian Academy of Sciences) for
501 discussions about egg capsule morphology.

502

503 **Figure Legends**

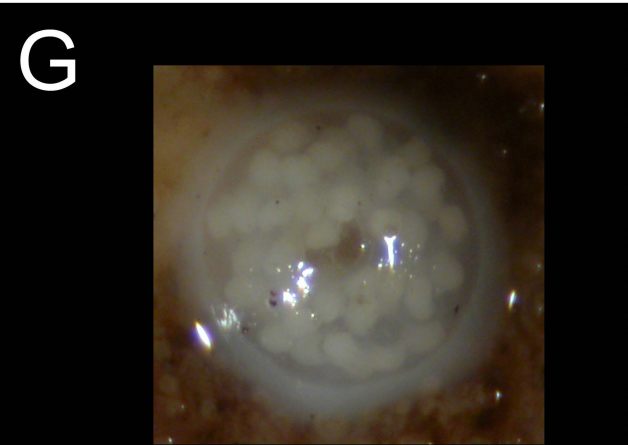
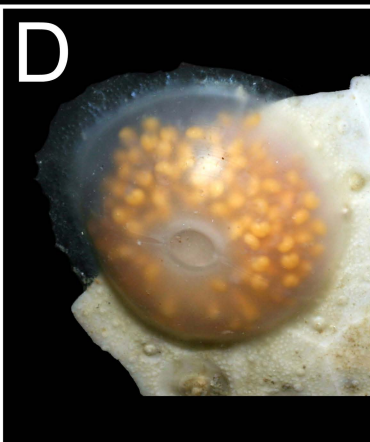
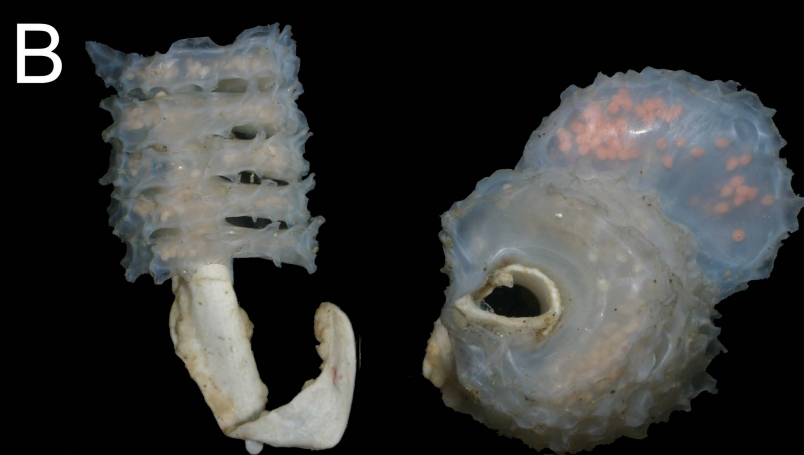
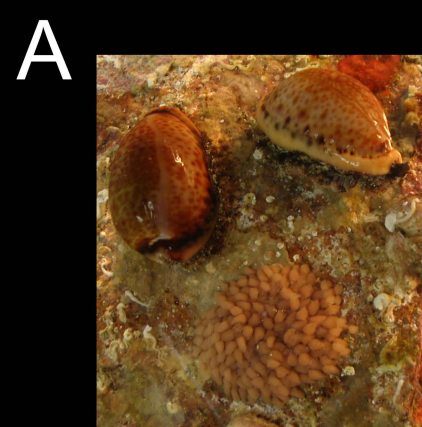
504 Fig. 1: Egg capsules used in molecular analyses and resulting identifications. A. EC2
505 (*Erosaria spurca* – egg capsule with adults) B. EC3 (*Conus* sp.) C. EC9 (left) and EC11
506 (*Granulifusus* sp., at two different stages) D. EC8 (*Comitas* sp.) E. EC5 (*Belomitra* sp.) F.
507 EC14 (unidentified caenogastropod) G. EC23 (Conidae, Clathurellinae) H. EC26 (Conidae,
508 Raphitominae) fixed on a bivalve shell. Photo credits: André Hoareau (A) and Barbara Buge
509 (MNHN) (B-H). Color photos are also published in MorphoBank
510 (<http://morphobank.geongrid.org/permalink/?P232>)

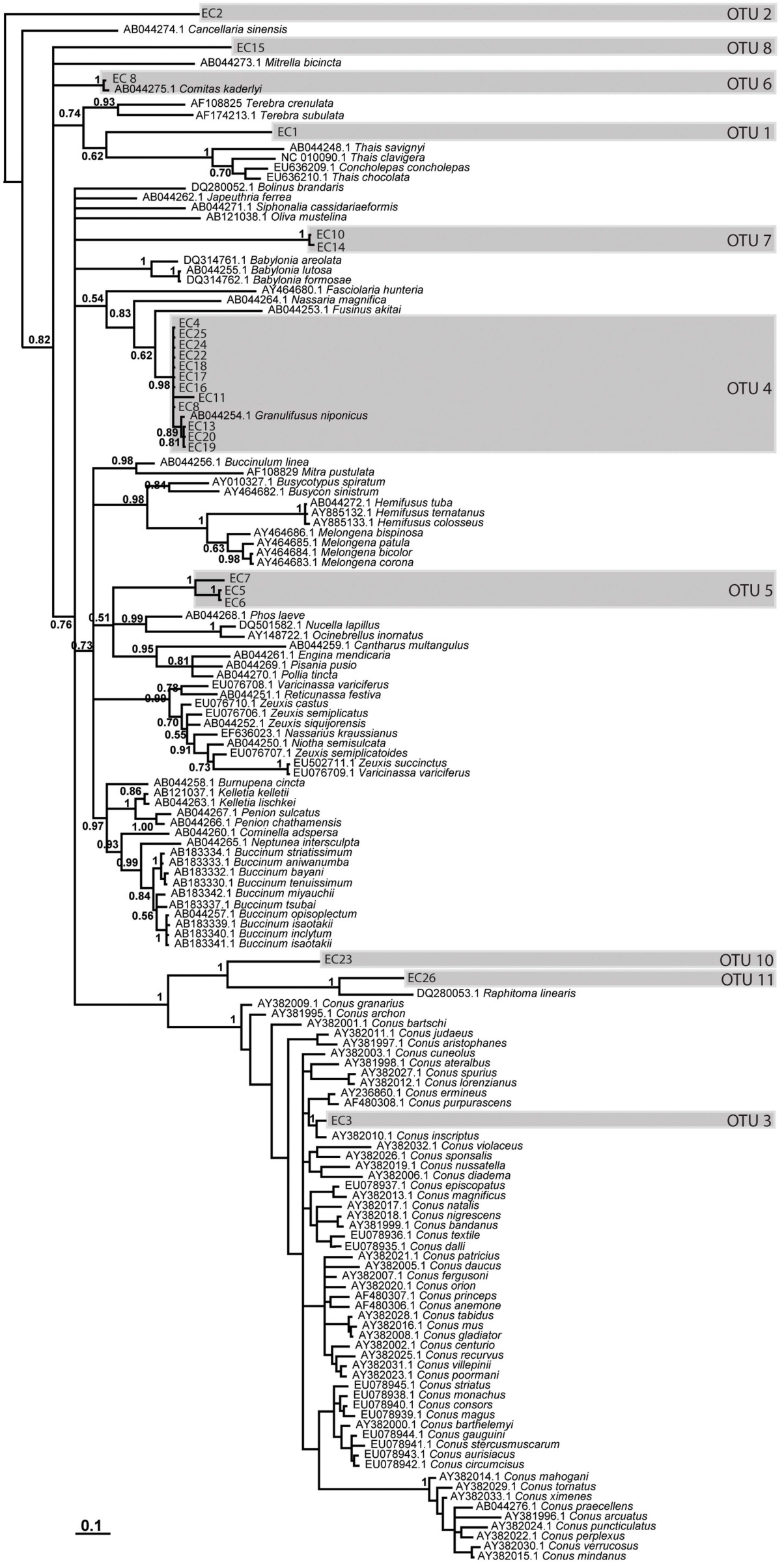
511
512 Fig. 2: Bayesian phylogram generated from all 16S neogastropod sequences in GenBank and
513 23 neogastropod egg capsule specimens from the Philippines and France. *Erosaria spurca*
514 (EC2), a non-neogastropod, is the outgroup.

515
516 Fig. 3: Number of positive identifications of unknown Philippine egg capsules at different
517 taxonomic levels based on all genes (A, B) or COI only (C, D). A and C include all GenBank
518 sequences; B and D exclude sequences produced by our own research group, demonstrating
519 the impact of a single research group on the global dataset.

520
521 Fig. 4: GenBank coverage and actual marine gastropod diversity for gastropod families.
522 Histogram on the left indicates number of COI sequences in GenBank: light gray – July 28,
523 2006; dark gray – April 6, 2007; black – July 4, 2008. Histogram on the right shows species
524 richness of gastropod families at a well-sampled tropical site (Koumac, New Caledonia;
525 Bouchet *et al.* 2002). Only groups with more than 10 species in Koumac or more than 100
526 sequences in GenBank (in 2006) are represented here. The decreasing of the number of

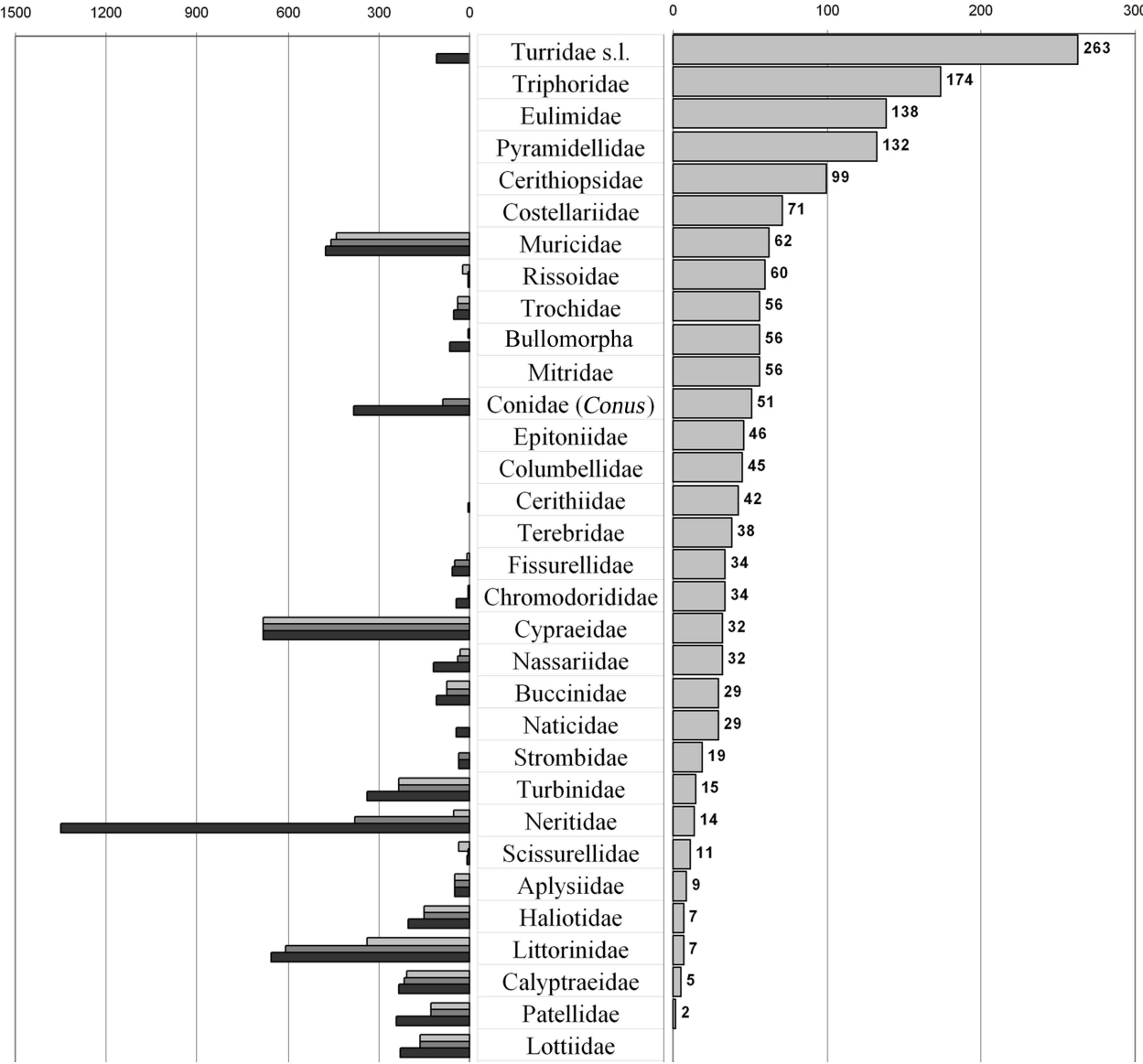
527 sequences for some groups through time is due to the fact that some sequences were removed
528 from GenBank.





0.1

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



							COI		
							BOLD		Genbank
ID	Station	COI	16S	12S	N°	OTU	ID	%	ID
P3		EU870569	EU870545	EU870520	EC1	1	x	x	<i>Fluminicola virens</i>
P4		EU870570	EU870546	EU870521	EC2	2	<i>Erosaria spurca</i>	99,51	<i>Erosaria spurca</i>
496	CP2665	EU870571	EU870547	EU870522	EC3	3	<i>Conus sulcatus</i>	88,94	<i>Conus venulatus</i>
648	CP2678	EU870572	EU870548	EU870523	EC4	4	<i>Busycon carica</i>	87,19	<i>Busycon sinistrum</i>
694	CP2684	EU870573	EU870549	EU870524	EC5	5	<i>Batillaria multiformis</i>	87,63	<i>Belomitra sp.</i>
695	CP2684	EU870574	EU870550	EU870525	EC6	5	<i>Batillaria multiformis</i>	87,45	<i>Belomitra sp.</i>
696	CP2684	EU870575	EU870551	EU870526	EC7	5	<i>Busycon sinistrum</i>	85,74	<i>Belomitra sp.</i>
861	CC2700	EU870576	EU870552	EU870527	EC8	6	<i>Comitas</i>	97,84	<i>Pyrgulopsis glandulosa</i>
898	CP2707	EU870577	EU870553	EU870528	EC9	4	<i>Busycon carica</i>	87,26	<i>Nucella lapillus</i>
1057	CP2712	EU870578	EU870554	EU870529	EC10	7	<i>Lacuna pallidula</i>	84,64	<i>Ilyanessa obsoleta</i>
1285	CC2724		EU870555	EU870530	EC11	4			
1402	CP2727			EU870531	EC12	8			
1404	CP2727	EU870579	EU870556	EU870532	EC13	4	<i>Busycon carica</i>	87,61	<i>Ilyanessa obsoleta</i>
1405	CP2727		EU870557		EC14	7			
1409	CP2727		EU870558	EU870533	EC15	8			
1646	CP2735	EU870580	EU870559	EU870534	EC16	4	<i>Busycon carica</i>	87,09	<i>Busycon sinistrum</i>
1647	CP2735	EU870581	EU870560	EU870535	EC17	4	<i>Busycon carica</i>	87,3	<i>Busycon sinistrum</i>
1648	CP2735	EU870582	EU870561	EU870536	EC18	4	<i>Busycon carica</i>	86,99	<i>Busycon sinistrum</i>
1681	CC2743	EU870583	EU870562	EU870537	EC19	4	<i>Busycon carica</i>	87,26	<i>Acanthinucella punctulata</i>
1682	CC2743	EU870584	EU870563	EU870538	EC20	4	<i>Busycon carica</i>	87,26	<i>Acanthinucella punctulata</i>
1689	CC2745			EU870539	EC21	9			
1690	CC2745	EU870585	EU870564	EU870540	EC22	4	<i>Busycon carica</i>	87,37	<i>Busycon sinistrum</i>
1691	CC2745	EU870586	EU870565	EU870541	EC23	10	<i>Nannodiella</i>	87,4	<i>Gymnobela sp.</i>
1692	CC2745	EU870587	EU870566	EU870542	EC24	4	<i>Busycon carica</i>	87,19	<i>Busycon sinistrum</i>
1693	CC2745	EU870588	EU870567	EU870543	EC25	4	<i>Busycon carica</i>	87,19	<i>Busycon sinistrum</i>
1775	CP2751	EU870589	EU870568	EU870544	EC26	11	<i>Gymnobela</i>	87,52	<i>Nannodiella</i>

		16S			12S		
%	Tree	Genbank		Tree	Genbank		Tree
		ID	%		ID	%	
76		<i>Conus circumciscus</i>	82		<i>Coralliophila meyendorffii</i>	99	<i>C. meyendorffii</i>
98		<i>Erosaria spurca</i>	99		<i>Cypraea annulus</i>	94	
90	<i>Conus sp.</i>	<i>Conus radiatus</i>	95	<i>Conus sp.</i>	<i>Conus textile</i>	90	<i>Conus</i>
86		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
89	<i>Belomitra sp.</i>	<i>Buccinum opisoplectum</i>	89		<i>Ilyanassa obsoleta</i>	84	
89	<i>Belomitra sp.</i>	<i>Buccinum opisoplectum</i>	89		<i>Ilyanassa obsoleta</i>	83	
89	<i>Belomitra sp.</i>	<i>Burnupena cincta</i>	89		<i>Gemmula sogodensis</i>	83	
85	<i>Lophiotoma</i>	<i>Comitas kaderlyi</i>	98	<i>C. kaderlyi</i>	<i>Gemmula rosario</i>	84	
87		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	87	
84		<i>Buccinulum linea</i>	86		<i>Ilyanassa obsoleta</i>	79	
		<i>Granulifusus niponicus</i>	95	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	87	
					<i>Gemmula rosario</i>	83	Turridae
86		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	85	
		<i>Buccinulum linea</i>	85				
		<i>Penion chathamensis</i>	85		<i>Lophiotoma unedo</i>	79	Turridae
86		<i>Granulifusus niponicus</i>	97	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
86		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
85		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
84		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
84		<i>Granulifusus niponicus</i>	98	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
					<i>Lophiotoma polytropia</i>	84	Cypraeidae
86		<i>Granulifusus niponicus</i>	97	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	86	
86	Raphitominiae	<i>Conus consors</i>	86	Conidae	<i>Ilyanassa obsoleta</i>	82	
86		<i>Granulifusus niponicus</i>	97	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	87	
86		<i>Granulifusus niponicus</i>	97	<i>G. niponicus</i>	<i>Ilyanassa obsoleta</i>	87	
87	Clathurellinae	<i>Raphitoma linearis</i>	87	Conidae	<i>Lophiotoma acuta</i>	81	

Final ID
<i>C. meyendorffii</i>
<i>Erosaria spurca</i>
<i>Conus sp.</i>
<i>Granulifus sp.</i>
<i>Belomitra sp.</i>
<i>Belomitra sp.</i>
<i>Belomitra sp.</i>
<i>Comitas sp.</i>
<i>Granulifus sp.</i>
?
<i>Granulifus sp.</i>
Turridae
<i>Granulifus sp.</i>
?
Turridae
<i>Granulifus sp.</i>
<i>Granulifus sp.</i>
<i>Granulifus sp.</i>
<i>Granulifus sp.</i>
<i>Granulifus sp.</i>
?
<i>Granulifus sp.</i>
Conidae (Raphitominae)
<i>Granulifus sp.</i>
<i>Granulifus sp.</i>
Conidae (Clathurellinae)