

BARCODING MICRO- AND MESO-FAUNA

Barcoding diatoms: Is there a good marker?

MÓNICA B. J. MONIZ and IRENA KACZMARSKA

*Biology Department, Mount Allison University, 63B York Street, Sackville, NB, Canada E4L 1G7***Abstract**

The promise of DNA barcoding is based on a small DNA fragment divergence coinciding with biological species separation. Here we evaluated the performance of three markers as diatom barcodes, the small ribosomal subunit (1600 bp), a 5' end fragment of cytochrome *c* oxidase subunit 1 (430 bp), and the second internal transcribed spacer region combined with the 5.8S gene (5.8S + ITS-2, 300–400 bp). Forty-four sequences per marker representing 28 species from all diatom classes were analysed. Sequence alignment of the three genetic markers and uncorrected genetic distances (*P*) were calculated at the intra- and heterospecific level. All three markers correctly separated the species examined and had advantages which contribute to their feasibility as a DNA barcode. Small ribosomal subunit had the largest GenBank data set, its success rate in amplification and sequencing was assumed to be the highest of all three and was readily aligned. However, it required a long fragment to recover divergence sufficient for species separation and small genetic distances increased the potential for misidentifications. Cytochrome *c* oxidase subunit 1 demonstrated a substantial heterospecific divergence level and was also readily alignable, but it showed very low amplification and sequencing success rates with currently existing primers. 5.8S + ITS-2 was amplified and sequenced with high success rate and was the most variable of the three markers, but its secondary structure was needed to aid in alignment. However, since it has been recently suggested that ITS-2 may provide insight into sexual compatibility, this marker offers an additional advantage. We therefore propose that the 5.8S + ITS-2 fragment is the best candidate as a diatom DNA barcode.

Keywords: cytochrome *c* oxidase I, diatom, DNA barcoding, internal transcribed spacer, small ribosomal subunit

Received 15 October 2008; revision received 8 December 2008; accepted 17 December 2008

Introduction

Alpha (α)-taxonomy is not only an academic conceptual exercise. The importance of identifying species reliably, quickly and cost-effectively is essential to any scientific study using live subjects. Bortolus (2008) discussed the cascading negative effects of bad taxonomy when used in ecology. Correct identification of species may be especially difficult with certain microbiota. Diatoms, being ecologically and economically important protists, are the centre of many biological studies where taxonomic identification frequently is not performed by specialists. Archibald (1984) emphasized how many studies (e.g. water quality assessments and others) using diatoms resulted in scientific errors because identifications were not verified by experienced diatom taxonomists.

Recently, the concept of barcoding was introduced to diatom taxonomy (Evans *et al.* 2007; Kaczmarska *et al.* 2007). This type of molecular α -taxonomy is based on the premise that the divergence of a small DNA fragment coincides with biological separation of species. This DNA fragment becomes a tag or a DNA barcode for species. Once a comprehensive database of such DNA barcodes is available, it may be used as a first approach to flag new species, select optimal taxa for phylogenetic studies, or to signal the geographical extent of divergences in a population (Hajibabaei *et al.* 2007). DNA barcoding should only be used in a first approach for these different applications, followed by larger in-depth studies in the respective fields.

Different DNA regions within the nuclear, mitochondrial and chloroplast genomes have been considered for testing as a universal DNA barcode for diatoms. Within the nuclear DNA, the small ribosomal subunit (SSU) has the advantages of being a coding gene and has been extensively used

Correspondence: Mónica B.J. Moniz, Fax: 01-506-364-2505; E-mail: mbmoniz@mta.ca

in diatom phylogenetics (Kooistra & Medlin 1996; Medlin *et al.* 1996; Medlin & Kaczmarek 2004; Sarno *et al.* 2005; Sorhannus 2007). The existing diatom SSU database is one of the most extensive for any DNA region. At the time of writing, the sequences available in GenBank for SSU covered diversity close to 30 orders compared to the *coxI* sequences which only covered around five orders with 60% of these sequences belonging to the genus *Sellaphora* Mereschkowsky. Therefore, SSU provides a good initial platform for an identification database. In the case of the protist genus *Blastocystis* (Alexieff) Brumpt, a short fragment of the SSU region was proven sufficient to identify operational taxonomic units (OTU), which are probably equivalent to species in this group (Sciicluna *et al.* 2006).

Initially, the fragment of choice for the Barcode of Life Initiative (BOLI) was the mitochondrial gene cytochrome *c* oxidase 1 (*coxI*). The reasoning for the choice of this marker included: (i) the mitochondrial genome evolves at a faster rate than the nuclear genome and therefore there is a greater potential for species specific informative regions (Chantangsi *et al.* 2007); (ii) the triple code characteristic of any coding gene aids alignment (Cywinska *et al.* 2006); and (iii) since mitochondria reproduce by binary fission without sexual recombination, their genes are less subject to insertions, deletions or other large-scale rearrangements that introduce ambiguous variation into the sequence (Hebert *et al.* 2003; Saunders 2005). This marker has been used extensively and very successfully for barcoding various animal (Hebert *et al.* 2003, 2004; Ward *et al.* 2005; Cywinska *et al.* 2006; Hajibabaei *et al.* 2006; Ratnasingham & Hebert 2007) and protistean taxa such as red and brown algae (Saunders 2005; Kucera & Saunders 2008), strains of *Paramecium tetraurelia* Sonneborn and *P. caudatum* Ehrenberg (Barth *et al.* 2006), isolates from the genus *Tetrahymena* Furgason (Chantangsi *et al.* 2007) and 22 *Sellaphora* species together with three other raphid genera of diatoms (Evans *et al.* 2007).

We recently proposed a third DNA barcode alternative, 5.8S + ITS-2 (Moniz & Kaczmarek, unpublished data). The reasons behind our proposal included the growing number of studies successfully using the internal transcribed spacer (ITS) region for species resolution of closely related diatoms (Behnke *et al.* 2004; Casteleyn *et al.* 2008; Vanormelingen *et al.* 2008), including semi-cryptic species (Amato *et al.* 2007; Kaczmarek *et al.* 2008, and references therein) and the expanding reference data set for this region. Many of these works capitalized on recent advances in research on ITS-2 secondary structure that have improved alignments and showed that this marker provides insight into sexual compatibility (Wolf *et al.* 2005; Seibel *et al.* 2006; Coleman 2008; Müller *et al.* 2007). This is a significant benefit to a proposed diatom barcode, as traditional diatom taxonomy is mostly based on valve morphology. We also included 5.8S into our DNA barcode since it afforded resolution at the genus level and provided an additional anchor to aid alignment. The

purpose of this study was to evaluate and compare the efficacy of these three markers as potential DNA barcodes using existing primers for the three classes of diatoms.

Materials and methods

The sequences analysed are listed in Table 1 and were comprised of 65 sequences already available in GenBank (including all SSU sequences) and 67 sequences obtained in our laboratory (which include new *coxI* and 5.8S + ITS-2 sequences). In total, we analysed 44 sequences per marker, representing 28 species and 22 genera for which sequences were available for the three markers (Table 1). In order to include as much higher taxa diversity (genus and higher taxa) as possible, in some cases (*Coscinodiscus*, *Melosira*, *Minutocellulus*, *Chaetoceros*, *Eunotia*, *Nitzschia* and *Pseudonitzschia*; seven genera out of 22 examined), species included in analyses were not the same for all markers. In these cases, we selected congeners known to cluster in the same clade in SSU based polygenetic trees (Medlin & Kaczmarek 2004).

The sequences obtained in our laboratory included 11 clones established in our laboratory and 23 strains received from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP). The procedure for clone isolation by micropipetting followed Andersen (2005). Non-axenic cultures were maintained in ~20 mL tubes or 125 mL flasks in f/2 medium (Guillard 1983), a dilution thereof (f/4 to f/20), or in L1⁺ medium (Andersen 2005) at a temperature of approximately either 6, 12 or 20 °C depending on strain requirements, with a 12:12 h light:dark (L:D) photoperiod and a photon fluence rate of about 20–50 µE/m²/s. When necessary, identity of strains was verified using scanning electron microscopy (SEM), following Kaczmarek *et al.* (2005) and observed with a JEOL JSM-5600 SEM operating at 10 kV and 8 mm working distance, at the Mount Allison University Digital Microscopy Facility.

The UltraClean Soil DNA Kit (MoBio Laboratories) as well as the Power Plant DNA Isolation Kit (MoBio Laboratories) were used to obtain DNA from pelleted cultures, following manufacturer specifications.

CoxI sequences were obtained using four primer pair combinations: (i) *coxF* and *coxR* (Iwatani *et al.* 2005); (ii) *coxF* (Iwatani *et al.* 2005) and DM1R (AADHGACYATCAA-DACCWGAHTTWGCHA) (R. Stern, University of British Columbia, personal communication); (iii) GazF1 and GazR2 (Saunders 2005); and (iv) DiatcoxI-GAF1mod (TCMAC-MAAYCAYAAAGATATWGG) and DiatcoxIrev-bg (ATYAAAATRTAAACYTCWGGGTG) (B. Gemeinholzer, Botanischer Garten und Botanisches Museum Berlin-Dahlem, personal communication). The 25-µL polymerase chain reaction (PCR) mixes were carried out either in PuReTaq Ready-To-Go PCR Beads (GE Healthcare Biosciences) following Kaczmarek *et al.* (2008) or using JumpStart (Sigma) and included the same amounts of DNA extraction solution

Table 1 List of strains used in this study; species/strain names listed as they appear in publications, culture collections, and GenBank or BOLD accession numbers

Species/strain name	SSU accession no.	cox1 accession no.	5.8S + ITS-2 accession no.
<i>Coscinodiscus</i> sp. GGM-2004	AY485448	—	—
<i>C. wailesii</i> CCMP2513	—	DIAT001-06	—
<i>C. radiatus</i> CCMP310	—	—	DITS185-08
<i>Rhizosolenia setigera</i> CCMP1820	—	—	DITS146-08
<i>R. setigera</i> CCMP1330	AY485461	AB020226	—
<i>Hyalodiscus stelliger</i> CCMP454	AY485507	DIAT065-07	DITS219-08
<i>H. stelliger</i> CCMP1679	AY485519	ITSCO001-09	DITS284-08
<i>Melosira varians</i> CHMP7	AY569590	—	—
<i>M. varians</i> MV13	AJ243065	—	—
<i>M. nummuloides</i> CCMP484	—	DIAT069-07	DITS215-08
<i>M. nummuloides</i> clone B	—	DIAT005-06	DITS279-08
<i>Thalassiosira pseudonana</i> CCMP1335	AY485452	—	EF208793
<i>T. pseudonana</i>	—	DQ186202	—
<i>T. punctigera</i> NB02-22	DQ514893	—	—
<i>T. punctigera</i>	—	DIAT007-06	DITS277-08
<i>T. weissflogii</i> CCMP1587	EF585582	DIAT080-07	DITS204-08
<i>T. nordenskiöldii</i> CCMP997	DQ093365	DIAT159-07	DITS125-08
<i>Cyclotella meneghiniana</i> p567	AJ535172	—	—
<i>C. meneghiniana</i> CCMP332	—	DIAT101-07	DITS183-08
<i>Minutocellus</i> sp. RCC967	EU106801	—	—
<i>M. polymorphus</i> CCMP499	—	DIAT114-07	DITS170-08
<i>Ditylum brightwellii</i> CCMP358	AY485444	DIAT061-07	DITS223-08
<i>D. brightwellii</i> clone P2A4	AY188182	—	—
<i>D. brightwellii</i> clone 278	AY188181	—	—
<i>D. brightwellii</i> CCAP 1022/2	X85386	—	—
<i>D. brightwellii</i> SMDC02	EU364892	—	—
<i>D. brightwellii</i> SMDC01	EU364891	—	—
<i>D. brightwellii</i> CCMP357	—	DIAT105-07	DITS179-08
<i>D. brightwellii</i> CCMP1810	—	DIAT002-06	DITS150-08
<i>D. brightwellii</i> Martin	—	DIAT105-07	DITS282-08
<i>D. brightwellii</i> Marine Botany	—	DIAT188-07	—
<i>D. brightwellii</i> KM11E	—	DIAT192-07	—
<i>D. brightwellii</i> CCMP359	—	—	DITS222-08
<i>D. brightwellii</i> CCMP356	—	—	DITS180-08
<i>Chaetoceros calcitrans</i> PCC537	AY485449	—	—
<i>C. calcitrans</i> Ifremer-Argenton	DQ887756	—	—
<i>C. socialis</i> CCMP203	—	DIAT057-07	DITS227-08
<i>C. socialis</i> CCMP1579	—	DIAT131-07	DITS153-08
<i>Attheya longicornis</i> CCMP214	AY485450	—	—
<i>A. longicornis</i>	—	DIAT197-07	DITS087-08
<i>Grammatophora oceanica</i> CCMP410	AY485466	DIAT110-07	DITS174-08
<i>Asterionelopsis glacialis</i> WK20	AY216904	—	—
<i>A. glacialis</i> CCMP139	—	DIAT055-07	DITS229-08
<i>Grammonema striatula</i> CCMP1094	AY485474	—	—
<i>G. striatula</i>	X77704	—	—
<i>G. striatula</i> IIA2	—	DIAT003-06	DITS281-08
<i>G. striatula</i> IIA3	—	DIAT004-06	DITS280-08
<i>Eunotia bilunaris</i> EBIL1	AJ866995	—	—
<i>Eunotia</i> sp. EUN392T	—	EF164960	—
<i>E. bilunaris</i> DM22-5	—	—	AM747216
<i>Thalassionema nitzschioides</i> IKETn1206	—	—	DITS287-08
<i>T. nitzschioides</i> CCAP 1084-1	X77702	AB020228	—
<i>Entomoneis</i> cf. <i>alata</i> CCMP1522	—	ITSCO002-09	DITS356-08
<i>E. cf. alata</i> p540	AJ535160	—	—
<i>Phaeodactylum tricorutum</i> CCMP630	—	DIAT184-07	DITS100-08

Table 1 Continued

Species/strain name	SSU accession no.	cox1 accession no.	5.8S + ITS-2 accession no.
<i>P. tricornutum</i>	AY485459	—	—
<i>Sellaphora auldrekkie</i> BLUNT	AJ544654	—	AJ544676
<i>S. auldrekkie</i> clone BLA2	—	EF164932	—
<i>S. pupula</i> SM-BLCAP	AJ544653	—	AJ544675
<i>S. pupula</i> THR13	—	EF164954	—
<i>S. pupula</i> PSEUDOCAP-3	AJ544649	—	—
<i>S. pupula</i> clone F1	—	EF164942	—
<i>S. pupula</i> PSEUDOCAP-1 cL. d	—	—	AJ544670
<i>S. capitata</i> BLA10	EF151971	—	—
<i>S. capitata</i> CAP	—	—	AJ544651
<i>S. capitata</i> DUN6	—	EF164946	—
<i>S. blackfordensis</i> DUN5	—	EF164949	—
<i>S. blackfordensis</i> clone BLA6	EF151969	—	—
<i>S. blackfordensis</i> RECT-6	—	—	AJ544666
<i>S. laevissima</i> US	—	—	AJ544656
<i>S. laevissima</i> THR4	EF151981	EF164943	—
<i>Amphiprora paludosa</i> CCMP125	AY485468	ITSCO003-09	DITS288-08
<i>Cylindrotheca closterium</i> 46-3-B2-IF	AY485471	—	—
<i>C. closterium</i> CCMP1725	—	DIAT085-07	DITS199-08
<i>C. fusiformis</i> CCMP344	—	DIAT103-07	DITS181-08
<i>C. fusiformis</i> CCMP339	AY485457	—	—
<i>Nitzschia frustulum</i> p345	AJ535164	—	—
<i>N. frustulum</i> CCMP 558	—	AB020225	—
<i>N. laevis</i> WDCM NCC39	—	—	AY574378
<i>Pseudo-nitzschia multiseriis</i> PM-02	EU302796	—	—
<i>P. multiseriis</i> CLN-50	EU302795	—	—
<i>P. multiseriis</i> Pn-1	DQ445651	—	—
<i>P. multiseriis</i> POMX	AM235382	—	—
<i>P. multiseriis</i> TKA2-28	AM235381	—	—
<i>P. delicatissima</i> BBA4	—	DIAT283-07	DITS001-08
<i>P. delicatissima</i> W007c2	—	DIAT227-07	DITS057-08
<i>P. delicatissima</i> W007b2	—	DIAT226-07	DITS058-08
<i>P. delicatissima</i> A2A322F	—	DIAT256-07	DITS028-08
<i>P. delicatissima</i> W002	—	DIAT223-07	DITS061-08

and primer mixes, 12.5 µL of *Taq* Mix and 9 µL of ultrapure water (DEPC Treated water, Invitrogen). The amplification regime followed Amato *et al.* (2007). PCR products were visualized in a 1.3% agarose gel. Cleaning of PCR products and sequencing were conducted at Nanuq (McGill University and Genome Québec) following their standard procedure.

The nuclear ITS-1, ITS-2 and the 5.8S rDNA gene (hereinafter referred to collectively as the ITS region) sequences were obtained as described above using primers ITS1 and ITS4 (White *et al.* 1990) in most cases, except when primers ITS5 (White *et al.* 1990) or the pair NS7 m and LR1850 (White *et al.* 1990; Bhattacharya 1996) yielded better results.

Success rates for each marker were calculated for amplification and sequencing steps. We considered amplification unsuccessful when it failed on samples known to have good quality and quantity of DNA. DNA quality and quantity was judged to be good when PCR products for at least one

marker were recovered and, when in doubt, tested using Nanodrop 1000 (ThermoScientific). Amplicons significantly smaller than expected were disregarded and also considered failures. Sequencing of each sense was considered as one trial. All trials were included in success rate calculations for the *coxI* and ITS region equally. Only a small data set of SSU sequences were tested in our laboratory and never the whole region, so we relied on GenBank data for this marker. Success for SSU was assumed high as we mainly used SSU to test for presence of diatom DNA when the two other markers failed, and achieved 100% success rate in all those cases.

Electrophenograms for *coxI* and the 5.8S + ITS-2 fragments were edited and sequences were aligned using BioEdit version 7.0.5.3 (Hall 1999). All our sequences were based on high-quality bidirectional readings. Alignments were verified and/or improved manually. A gap open of 10 and gap extent of 1.2 were used for analyses. These parameters

were previously tested (Moniz & Kaczmarek, unpublished data) and found to coincide with alignments that improved low-scoring fragments without disrupting the known anchor-regions as recommended by Hall (2004). Distances were estimated using simple uncorrected pairwise (p) distances using the Visual Basic/Excel program DOINK (J.M. Ehrman, Digital Microscopy Facility, Mount Allison University, 2007). The program produces a nucleotide difference matrix identical to that produced by BioEdit but within an Excel spreadsheet. Genetic distances between sequences in each analysis were expressed as the number of substitutions, gaps or indels (differences) per site, following Litaker *et al.* (2007) for ease of comparison. When $p \leq 0.1$, uncorrected distances are very similar to corrected distances such as the Kimura 2-parameter model as in the current Barcode of Life Data System (BOLD, www.boldsystems.org, Ratnasingham & Hebert 2007). In all cases, intrageneric distances were less than 0.10.

Sequences were trimmed in order to include as much diatom diversity as possible using the sequences already available and those obtained in this study. The regions analysed were a fragment of *c.* 1600 bp near the 5' end for the nuclear encoded SSU region, a fragment of *c.* 430 bp near the 5' end for *coxI*, and a fragment of *c.* 300–400 bp for ITS, including most of nuclear encoded 5.8S and ITS-2 until the conserved motif in helix III of its transcript secondary structure.

Results

Success rates

Success rates of amplification and sequencing were calculated only for *coxI* and 5.8S + ITS-2 since these procedures were conducted in our laboratory. We assumed a high success rate for SSU sequences because we only tested a 500-bp fragment of SSU which was performed only in cases where amplification failed using primers for the other markers. In

these cases, the success rate for SSU amplification and sequencing was 100%. SSU sequences used for further evaluation were all obtained from GenBank. *CoxI* had the lowest overall success rate. Only 29% of good quality DNA amplified with *coxI* and of those only 30% were sequenced successfully and found to be diatom DNA. With the primers used in this study, 37% percent of the amplified DNA (with the correct amplicon size confirmed in agarose gel) was found to be bacterial contamination. The 5.8S + ITS-2 marker was easily amplified using existing primers (79% success) and sequenced reliably (84% success).

Divergence levels

For clarity, whenever the term intraspecific distance is used below, it refers to the maximum distance between two sequences within the same morpho-species. The term heterospecific (intrageneric or interspecific) distance refers to the minimum distance between two sequences from different species of the same morpho-genus. Intrafamilial distance refers to the minimum distance between two sequences from different genera but members of the same morpho-family. Finally, the distance between distant morpho-genera from the same class refers to the minimum distance between two sequences from species belonging to different genera and different families, but belonging to the same class of diatoms. In this last category, we include the species *Coscinodiscus wailesii* Gran & Angst, *Rhizosolenia setigera* Brightwell, *Cyclotella meneghiniana* Kützing, *Minutocellus polymorphus* (Hargraves & Guillard) Hasle, von Stosch, & Syvertsen, *Attheya longicornis* Crawford & Gardner, *Grammatophora oceanica* Ehrenberg, *Eunotia bilunaris* (Ehrenberg) Mills, *Thalassionema nitzschioides* Grunov, *Entomoneis alata* (Ehrenberg) Ehrenberg, *Phaeodactylum tricorutum* Bohlin and *Amphiprora paludosa* (Smith) Reimer which each one represent only one and separate family, and so cannot be integrated into the summaries presented in Fig. 1 and Table 2.

Table 2 Nucleotide sequence distance for the three markers between members of seven diatom orders at three levels of taxonomic affinity. The column n indicates how many sequences were analysed and within parenthesis how many species those sequences. Eleven species, listed in Table 1, are not included here since they belong to families only represented by one sequence

Order	n	Within species			n	Within genus			n	Within family		
		SSU	<i>coxI</i>	5.8S + ITS2		SSU	<i>coxI</i>	5.8S + ITS2		SSU	<i>coxI</i>	5.8S + ITS2
Meloseirales	4(2)	0.00	0.07	0.02	—	—	—	—	—	—	—	
Thalassiosirales	—	—	—	—	4(1)	0.03	0.17	0.15	—	—	—	
Lithodesmidales	6(1)	0.00	0.02	0.00	—	—	—	—	—	—	—	
Chaetocerotales	2(1)	0.00	0.04	0.00	—	—	—	—	—	—	—	
Fragilariales	2(1)	0.00	0.00	0.01	—	—	—	2(1)	0.07	0.18	0.26	
Naviculales	4(2)	0.01	0.05	0.07	4(1)	0.01	0.10	0.23	—	—	—	
Bacillariales	5(1)	0.01	0.00	0.00	2(1)	0.00	0.10	0.11	3(1)	0.04	0.16	0.21
Average	—	0.00	0.03	0.02	—	0.01	0.12	0.16	—	0.06	0.17	0.24

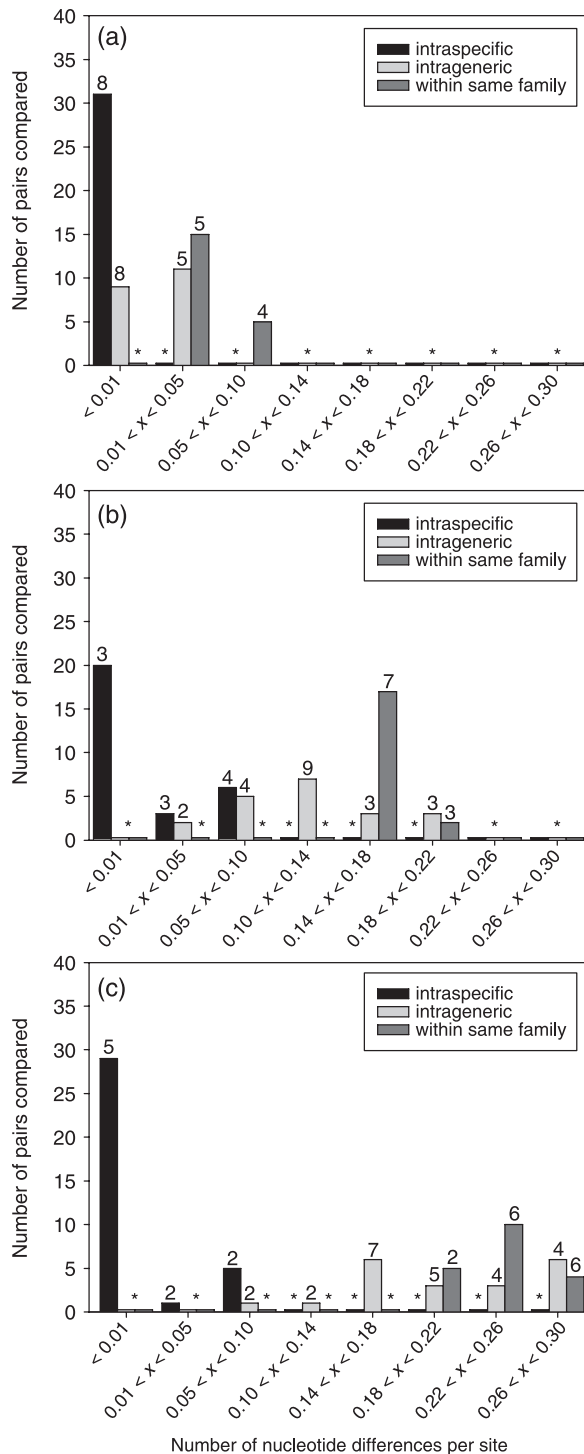


Fig. 1 Comparison of nucleotide sequence differences in a) SSU, b) *coxI*, c) 5.8S + ITS-2 among 18 species from three classes of diatoms; pairwise comparisons between 33 sequences are separated into three categories, intraspecific, intrageneric (heterospecific) and within a family. Numbers on top of bars indicate how many species the pairwise comparisons represent (* indicates that no pairwise comparisons were found with this distance). Eleven species, listed in Table 1, are not included here since they belong to families only represented by one sequence.

Comparing divergence levels of the three markers, it is apparent that SSU showed the lowest levels of divergence (Table 2). In the fragment examined, we found no intraspecific variability in the species studied. Heterospecific distances were on average $p = 0.01$ dif/site in our data set (Table 2), but varied between $p = 0.01$ – 0.04 dif/site (Fig. 1a), with the exception of *Cylindrotheca fusiformis* and *Cylindrotheca closterium* (which showed 100% identity) and known semi-cryptic species of *Sellaphora auldreekie* D. G. Mann & S. M. McDonald and *Sellaphora blackfordensis* D.G. Mann & S. Droop which also were not separated. Intrafamilial genera separated on average at $p = 0.06$ dif/site (Table 2) as did the distant genera from the same class (Fig. S1). Only the *A. paludosa* sequence showed an indel of 102 bp which was removed before alignment (Hall 2004).

For the *coxI* gene (Table 2, Fig. 1b), intraspecific distances were always below $p = 0.06$ dif/site and often non-existent with an average of $p = 0.03$ dif/site. Heterospecific distances varied between $p = 0.07$ – 0.17 dif/site (four to seven times higher than heterospecific distances recovered in SSU) with an average of $p = 0.12$ dif/site. Genera from the same family were separated at $p = 0.15$ – 0.20 dif/site (three to four times higher than the level found in SSU) with an average of $p = 0.17$ dif/site and distant genera from the same class were separated by $p > 0.20$ dif/site (c. three times higher compared to SSU) (Fig. S2).

The combined 5.8S + ITS-2 marker was heterospecifically the most variable marker of the three studied (Table 2, Fig. 1c). Intraspecific distances in the species examined were $p < 0.02$ dif/site with exceptions found in two *Sellaphora* species (*S. auldreekie* and *S. blackfordensis*) which presented higher intraspecific divergence. Heterospecific genetic distances within the same genus varied between $p = 0.06$ dif/site and $p = 0.30$ dif/site with an average of $p = 0.16$ dif/site; the minimum uncorrected distance being comparable to *coxI* and the maximum almost twice that found in *coxI*. Intrafamilial genera were separated at an average of $p = 0.24$ dif/site and distant genera from the same class were separated at $p > 0.40$ dif/site, twice as large as those recovered at the same taxa level in *coxI* (Fig. S3).

Discussion

All three markers correctly separate species in > 95% of our test set using existing primers. This data set was limited in size by the low success in recovering *coxI* sequences to match the species successfully sequenced using the other markers. The only exceptions in correct identification were the two species from the genus *Cylindrotheca* Rabenhorst and the two semi-cryptic species from the genus *Sellaphora* which were not separated by the SSU marker. However, each marker shows distinct features that factor into supporting their efficacy as a DNA barcode, when the barcode is regarded strictly as a means for species identification.

SSU is a widely used marker in phylogenetic studies of diatoms with a diverse and large database already available (Kooistra & Medlin 1996; Medlin *et al.* 1996; Medlin & Kaczmarek 2004; Sarno *et al.* 2005; Sorhannus 2007), both properties that made *coxI* an attractive candidate for a barcode in animal phyla. SSU is a readily amplifiable marker, with primers that successfully sequence diatom DNA. However, it requires a long fragment (five times longer than the other two markers) which is not recoverable in a single PCR and thus requires several primer pairs in the sequencing step in order to detect divergence sufficient to identify species, noted also by Alverson & Kolnick (2005). The small uncorrected distances, $p = 0.01\text{--}0.04$ dif/site between species, even when using long fragments (*c.* 1600 bp) increases the potential for sequencing errors which could affect the identification of closely related species (Fig. 1a). Since at least four species were not separated by this marker, it is possible that some other, closely related species would be missed when using this marker to assess the biodiversity of a sample.

CoxI has been tremendously successful in the animal kingdom, but has shown mixed results when applied to protists. It works well for red and brown algae (Robba *et al.* 2006; Kucera & Saunders 2008), but has limited application in others (Evans *et al.* 2007; Litaker *et al.* 2007). These studies demonstrated that appropriate thresholds for intraspecific vs. heterospecific distances should be defined for each taxon, a departure from the universal 10 \times ratio initially used in animal studies (Hebert *et al.* 2003), but adjusted subsequently (Hebert *et al.* 2004; Ward *et al.* 2005; Cywinska *et al.* 2006; Hajibabaei *et al.* 2006). Depending on the taxon, the minimum heterospecific distance varies. In red macro-algae using a 664-bp barcode, $p = 0.04$ was shown to be a sufficient and reliable inter-intraspecific threshold (Saunders 2005). This threshold was higher in *Tetrahymena* spp., where species indistinguishable by SSU were separated by a minimum of $p = 0.11$ over a 980-bp *coxI* fragment (Chantangsi *et al.* 2007). Two species of the ciliate *Paramecium* showed a distance of $p = 0.20$ using an 880-bp fragment (Barth *et al.* 2006). In our study using a still shorter fragment (430 bp), we found $p = 0.07$ (often $p > 0.10$) to be the minimum heterospecific distance between diatom species examined (Fig. 1b). We note that some of these studies use corrected distances to calculate these values, most commonly the K2p method. However, as we indicated in the material and methods section, with p distances close to 0.10 (as was the case of intra-heterospecific boundaries in our study and those referenced above), uncorrected distances are very close to other more complex calculated distance measures (Hall 2004), and so should be generally comparable. The list of protistean taxa that have successfully used *coxI* as a DNA barcode is not only short but also, with few exceptions, includes studies performed in only one genus, and largely involve the development of genus or group specific primers. Similar to our experience with diatoms, this appears to reflect

the difficulty in finding regions that are conserved enough to design universal primers for these divisions and phyla.

Within the *Sellaphora pupula* species complex, Evans *et al.* (2007) tested the effectiveness of several gene sequences (*coxI*, *rbcl*, SSU and ITS rDNA) as possible DNA barcodes to distinguish semi-cryptic diatom species. Similar to several protist groups discussed above, the *coxI* primers tested worked in raphid pennates (22 *Sellaphora* strains and two strains from the genera *Pinnularia* and *Eunotia*) but failed outside raphid diatoms, working only partially with the araphid pennate *Tabularia* sp. (not included in this study due to its small size) and not at all with centric diatoms. Our data also shows that this marker clearly separates diatom species thus far examined. Furthermore, as a coding gene, *coxI* has the advantage of being directly alignable with no need for parameter adjustment. However, as Evans *et al.* (2007) already indicated (and we concur), designing a universal diatom primer for this marker will be a challenging task. Not only was the overall success rate in amplification and sequencing low in our data set, but often we found that bacterial DNA was amplified instead of diatom DNA. Application of the techniques such as M13 tagging did not improve the success rate. In a recent study comparing different markers as plant DNA barcodes, regions that showed a success rate as low as that found for *coxI* in our study were disregarded (Fazekas *et al.* 2008). It is unlikely that existing primers hold promise of improvement since the available sequences show few regions that are sufficiently conserved. Still fewer are sequences with the appropriate AT/CG ratio at the 5' end of *coxI* across all extant diatom genera in the barcoding region that has been successfully used in the animal kingdom. This high divergence may reflect events of intron gain or loss and gene transfer from bacteria, highly pervasive in diatoms (Armbrust *et al.* 2004; Bowler *et al.* 2008). Also, the notion of diatom-wide uniparental transfer of mitochondria cannot be assumed since all pennates and many centrals show hologamic syngamy where the entire cytoplasmic contents (including mitochondria and plastids) from both parents contribute to the zygote (Medlin & Kaczmarek 2004), with possible mitochondrial fusion (Keeling & Palmer 2008).

ITS has already shown promise as a DNA barcode or additional locus in the identification of some non-animal kingdoms: fungi (Taylor *et al.* 2008), plant species of *Atractylodes* (Shiba *et al.* 2006), *Bupleurum* (Yang *et al.* 2007) and *Astragalus* (Ma *et al.* 2000) in Chinese and Japanese medicine, alone or together with other sequences (Gemeinholzer *et al.* 2006), dinoflagellates (Litaker *et al.* 2007) and recently in at least the two major classes of diatoms, Mediophyceae and Bacillariophyceae (Moniz & Kaczmarek, unpublished data). We also found that ITS sequences are not as readily aligned as coding markers such as *coxI* or SSU, but several recent studies demonstrated how to significantly improve alignability of ITS-2. Since the seminal studies of the ITS region as a potential marker for species delineation focused

on species from the family Volvocaceae (Coleman *et al.* 1994), several others (Coleman & Mai 1997; Coleman 2000, 2003, 2008; Schultz *et al.* 2005; Coleman 2007; Müller *et al.* 2007) included a wide range of biota and have shown how the secondary structure of the RNA transcript of ITS-2 contains subregions of significant conservation at different taxonomic levels, ranging from orders to species and populations. These more conserved regions are presumed necessary for production of a proper secondary structure of the ITS-2 transcript (Coleman & Mai 1997), which in turn is needed to guide the complex enzymatic processing that produces the ribosomal RNA subunits (Coleman 2007). In eukaryotes, ITS-2 typically has four helices. Helices II and III, which contain the regions of greatest conservation in the whole ITS-2 region, possibly pan-eukaryotically (Coleman 2007), are readily identifiable and aid in alignment. Furthermore, strains differing by even a single one-sided compensatory base change (CBC) within conserved regions have a high probability (93.11% for plants and fungi) of not being sexually compatible or of not producing viable progeny (Coleman 2007; Müller *et al.* 2007). These authors suggest that CBC is an indication that enough evolutionary time has passed as to be very probable that a speciation event has occurred (Müller *et al.* 2007). This is a serendipitous advantage in barcoding micro-organisms which are delineated primarily by morphology and where mating tests are not commonly practiced and often not even possible. Only two exceptions are known to date where two organisms with identical ITS-2 sequences failed to cross (Coleman 2008).

There are widely recognized concerns regarding the use of a nuclear multicopy, multigene families such as ribosomal RNA genes, especially in flowering plant phylogenetic studies (Álvarez & Wendel 2003) where allopolyploidization is the key mechanism in evolution (Feliner & Rosselló 2007). In such biota in particular, when concerted evolution is not fully operating, the genome may harbour a diversity of sequence types, including e.g. duplicate nonfunctional ribosomal loci (Feliner & Rosselló 2007). Hybridization and polyploidization does not, as far as we know, play such a role in the evolution of diatoms (Bowler *et al.* 2008) and intragenomic variability in the ITS region has been repeatedly shown to be very low (Behnke *et al.* 2004; Orsini *et al.* 2004; Casteleyn *et al.* 2008). Furthermore, recent plant and protist evolutionary studies have re-assessed these concerns and offer suggestions and guidelines for the wise use of such data (Feliner & Rosselló 2007; Thornhill *et al.* 2007; Coleman 2008). Thornhill *et al.* (2007), for example, tested three techniques to assess if intragenomic rDNA variation can confound estimates of microbial diversity. These authors evaluated direct sequencing of PCR products (used in our study), PCR-denaturing gradient gel electrophoresis fingerprinting and bacterial cloning, each with subsequent sequencing. All three techniques recovered the nearly iden-

tical, most common copy of the ITS sequence. This most common copy is understandably the functional one, and therefore the most desirable in phylogenetic studies. Not surprisingly, cloning of PCR products recovered additional, nonfunctional copies. These authors suggest that in order to avoid overestimating intragenomic variation, especially in environmental samples (an ultimate goal of barcoding), cloning should be avoided or if performed, secondary structure should be derived to ensure that the copy recovered is the functional one. They further concluded with the suggestion that it may be premature to abandon ITS as a species level marker in higher plants and provided guidelines to identify paralogous sequences, pseudogenes and contaminations that are applicable to any organism studied.

In the light of these advancements, Moniz & Kaczmarek (unpublished) proposed an alternative, nuclear encoded DNA barcode starting with the 5.8S region that serves as a useful anchor and ends at the pan-eukaryotically conserved motif on helix III of the initial transcript ITS-2 secondary structure. Direct comparison of the proposed DNA barcode successfully separated all biologically defined species, 88% of all 107 morpho-species tested in Moniz & Kaczmarek (unpublished) using the proposed $P = 0.07$ dif/site threshold, and separated all species in this study (Fig. 1c). Moreover, blasting within the existing online sequences and those included in this study correctly flagged mislabelled strains as reliably as SSU or *coxI*. Blasting against a good reference database is one of the most reliable tools of species identification, even more so than strict tree-based phylogenetic methods (Ross *et al.* 2008).

Conclusion

None of the three markers tested here fulfills the role that *coxI* seems to provide for animal phyla, where it distinguishes species at > 95%, while also being easy to amplify and align. However, capitalizing on recent advances in research on ITS-2 transcript secondary structure, the 5.8S + ITS-2 fragment emerges as a strong contender for a diatom DNA barcode. It is easily amplifiable and only a small fragment is required to distinguish even closely related species. Better understanding of its secondary structure has made alignments more straightforward and may additively provide valuable insight into sexual compatibility of ecologically important diatom morpho-species, using tools such as CBC analyser (Wolf *et al.* 2005). These insights may prove particularly useful in this most diverse division of photoautotrophic protists for which mating data exists for only 0.0015% to 0.15% of the estimated 10 000–1 000 000 species.

Acknowledgments

We thank J.M. Ehrman for all the SEM work involved in this study and for bioinformatics support; R.A. Andersen (CCMP) for

providing their cultures; N.A. Davidovich, J.L. Martin, M. LeGresley and K. Pauley for phytoplankton samples from which some clones were isolated; and A. Cockshutt for molecular troubleshooting early in this study. All colleagues in our laboratory also deserve thanks for lending help needed. We acknowledge funding through the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC, Mount Allison University and other sponsors listed at www.BOLNET.ca (to IK).

Conflict of interest statement

The authors have no conflict of interest to declare and note that the funders of this research had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Álvarez I, Wendell JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, **29**, 417–434.
- Alverson AJ, Kolnick L (2005) Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *Journal of Phycology*, **41**, 1248–1257.
- Amato A, Kooistra WHCF, Ghiron JHL *et al.* (2007) Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist*, **158**, 193–207.
- Andersen RA (2005) *Algal Culturing Techniques*. Elsevier Academic Press, Amsterdam, The Netherlands.
- Archibald R (1984) Diatom illustrations — an appeal. *Bacillaria*, **7**, 173–178.
- Armbrust E, Berges JA, Bowler C *et al.* (2004) The genome of *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science*, **306**, 79–86.
- Barth D, Krenek S, Fokin SI, Berendonk T (2006) Intraspecific genetic variation in *Paramecium* revealed by mitochondrial cytochrome *c* oxidase I sequences. *Journal of Eukaryotic Microbiology*, **53**, 20–25.
- Behnke A, Friedl T, Chepurinov VA, Mann DG (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology*, **40**, 193–208.
- Bhattacharya D (1996) Analysis of the distribution of bootstrap tree lengths using the maximum parsimony method. *Molecular Phylogenetic Evolution*, **6**, 339–350.
- Bortolus A (2008) Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio*, **37**, 114–118.
- Bowler C, Allen AE, Badger JH *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–244.
- Casteleyn G, Chepurinov VA, Leliaert F *et al.* (2008) *Pseudo-nitzschia pungens* (Bacillariophyceae): a cosmopolitan diatom species? *Harmful Algae*, **7**, 241–257.
- Chantangsi C, Lynn DH, Brandl MT *et al.* (2007) Barcoding ciliates: a comprehensive study of 75 isolates of the genus *Tetrahymena*. *International Journal of Systematic and Evolutionary Microbiology*, **57**, 2412–2425.
- Coleman AW (2000) The significance of a coincidence between evolutionary landmarks found in mating affinity and a DNA sequence. *Protist*, **151**, 1–9.
- Coleman AW (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics*, **19**, 370–375.
- Coleman AW (2007) Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic Acids Research*, **35**, 3322–3329.
- Coleman AW (2008) Is there a molecular key to the level of 'biological species' in eukaryotes? A DNA guide. *Molecular Phylogenetics and Evolution*. doi: 10.1016/j.ympev.2008.10.008.
- Coleman AW, Mai JC (1997) Ribosomal DNA ITS-1 and ITS-2 sequence comparisons as a tool for predicting genetic relatedness. *Journal of Molecular Evolution*, **45**, 168–177.
- Coleman AW, Suarez A, Goff L (1994) Molecular delineation of species and syngens in Volvocacean green algae (Chlorophyta). *Journal of Phycology*, **30**, 80–90.
- Cywinska A, Hunter FF, Hebert PDN (2006) Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*, **20**, 413–424.
- Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom 'barcode' genes (*cox1*, *rbcL*, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist*, **158**, 349–364.
- Fazekas A, Burgess KS, Kesanakurti PR *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802.
- Feliner GN, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*, **44**, 911–919.
- Gemeinholzer B, Oberprieler C, Bachman K (2006) Using Genbank data for plant identification: possibilities and limitations using the ITS1 of Asteraceae species belonging to the tribes Lactuceae and Anthemideae. *Taxon*, **55**, 173–187.
- Guillard RRL (1983) Culture of phytoplankton for feeding marine invertebrates. In: *Culture of Marine Invertebrates. Selected Readings* (ed. Berg Jr CJ), pp. 108–132. Hutchinson Ross Publishing Co, Stroudsburg, Pennsylvania.
- Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences, USA*, **103**, 968–971.
- Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.
- 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.
- Hall BG (2004) *Phylogenetic Trees Made Easy: A How-to Manual*, 2nd edn. Sinauer & Associates, Sunderland, Massachusetts.
- Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **270**, 313–321.
- Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.
- Iwatani N, Murakami S, Suzuki Y (2005) A sequencing protocol of some DNA regions in nuclear, chloroplastic and mitochondrial genomes with an individual colony of *Thalassiosira nordenskiöldii* Cleve (Bacillariophyceae). *Polar Biosciences*, **18**, 35–45.
- Kaczmarek I, LeGresley MM, Martin JL, Ehrman JM (2005) Diversity of the diatom genus *Pseudo-nitzschia* Peragallo in the Quoddy Region of the Bay of Fundy, Canada. *Harmful Algae*, **4**, 1–19.
- Kaczmarek I, Reid C, Moniz M (2007) Diatom taxonomy: morphology, molecules and barcodes.... In: *Proceedings of the 1st Central-European Diatom Meeting 2007, Berlin-Dahlem, 23 March–25*

- March 2007 (eds Kusber WH, Jahn R), pp. 69–72. Berlin-Dahlem, Germany.
- Kaczmarek I, Reid C, Martin JL, Moniz MJB (2008) Morphological, biological, and molecular characteristics of the diatom *Pseudo-nitzschia delicatissima* from the Canadian Maritimes. *Botany*, **86**, 763–772.
- Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nature*, **9**, 605–618.
- Kooistra WHCF, Medlin LK (1996) Evolution of the diatoms (Bacillariophyta): IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. *Molecular Phylogenetics and Evolution*, **6**, 391–407.
- Kucera H, Saunders GN (2008) Assigning morphological variants of *Fucus* (Fucales, Phaeophyceae) in Canadian waters to recognized species using DNA barcoding. *Botany*, **86**, 1065–1079.
- Litaker RW, Vandersea MW, Kibler SR *et al.* (2007) Recognizing dinoflagellate species using ITS rDNA sequences. *Journal of Phycology*, **43**, 344–355.
- Ma XQ, Duan JA, Zhu DY, Dong TTX, Tsim KWK (2000) Species identification of *Radix Astragali* (Huangqi) by DNA sequence of its 5S-rRNA spacer domain. *Phytochemistry*, **54**, 363–368.
- Medlin LK, Kaczmarek I (2004) Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia*, **43**, 245–270.
- Medlin LK, Kooistra WH, Gersonde R, Wellbrock U (1996) Evolution of the diatoms (Bacillariophyta). II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms. *Molecular Biology and Evolution*, **13**, 67–75.
- Müller T, Philipp N, Dandekar T, Schultz J, Wolf M (2007) Distinguishing species. *RNA*, **13**, 1469–1472.
- Orsini L, Procaccini G, Sarno D, Montresor M (2004) Multiple rDNA ITS-types within the diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae) and their relative abundances across a spring Bloom in the Gulf of Naples. *Marine Ecology Progress Series*, **271**, 87–98.
- Ratnasingham S, Hebert PDN (2007) BOLD: the Barcode of Life Data system. *Molecular Ecology Notes*, **7**, 355–364.
- Robba L, Russel SJ, Barker GL, Brodie J (2006) Assessing the use of the mitochondrial *cox1* marker for the use in DNA barcoding of red algae (Rhodophyta). *American Journal of Botany*, **93**, 1101–1108.
- Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, **57**, 216–230.
- Sarno D, Kooistra WHCF, Medlin LK, Percopo I, Zingone A (2005) Diversity in the genus *Skeletonema* (Bacillariophyceae). II: an assessment of the taxonomy of *S. costatum*-like species with the description of four new species. *Journal of Phycology*, **41**, 151–176.
- Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1879–1888.
- Schultz J, Maisel S, Gerlach D, Müller T, Wolf M (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, **11**, 361–364.
- Scicluna SM, Tawari B, Clark CG (2006) DNA barcoding of *Blastocystis*. *Protist*, **157**, 77–85.
- Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M (2006) 4SALE – A tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, **7**, 498.
- Shiba M, Kondo K, Miki E *et al.* (2006) Identification of medicinal *Atractylodes* based on ITS sequences of nrDNA. *Biological and Pharmaceutical Bulletin*, **29**, 315–320.
- Sorhannus U (2007) A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology*, **65**, 1–12.
- Taylor J, Bruns T, Lutzoni F (2008) ITS as the Fungal Barcode. Available from URL: <http://www.allfungi.com/its-barcode.php>.
- Thornhill DJ, Lajeunesse TC, Santos SR (2007) Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Molecular Ecology*, **16**, 5326–5340.
- Vanormelingen P, Chepuronov VA, Mann DG, Sabbe K, Vyverman W (2008) Genetic divergence and reproductive barriers among morphologically heterogeneous sympatric clones of *Eunotia bilunaris sensu lato* (Bacillariophyta). *Protist*, **159**, 73–90.
- Ward R, 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.
- White TJ, Bruns TD, Lee SB, Taylor JW (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: *PCR Protocols: A Guide to Methods and Applications* (eds Innis MA, Gelfand DH, Sninsky JJ, White TJ), pp. 315–322. Academic Press, London.
- Wolf M, Friedrich J, Dandekar T, Müller T (2005) CBC analyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures. *In Silico Biology*, **5**, 291–294.
- Yang Z, Chao Z, Huo K, Xie H, Tian Z, Pan S (2007) ITS sequence analysis used for molecular identification of the *Bupleurum* species from northwestern China. *Phytomedicine*, **14**, 416–423.

Supporting information

Additional supporting information may be found in the online version of this article:

Fig. S1 Matrix of nucleotide differences per site calculated for 44 sequences representing 28 species from the classes Coscinodisco-phyceae (up), Mediophyceae (middle) and Bacillariophyceae (down) using marker SSU.

Fig. S2 Matrix of nucleotide differences per site calculated for 44 sequences representing 28 species from the classes Coscinodisco-phyceae (up), Mediophyceae (middle) and Bacillariophyceae (down) using marker *cox I*.

Fig. S3 Matrix of nucleotide differences per site calculated for 44 sequences representing 28 species from the classes Coscinodisco-phyceae (up), Mediophyceae (middle) and Bacillariophyceae (down) using marker 5.8S+ITS-2.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.