Received: 28 March 2022 | Revised: 18 May 2023 | Accepted: 18 May 2023

DOI: 10.1111/jeu.12986

ORIGINAL ARTICLE

Comparison between p-distance and single-locus species delimitation models for delineating reproductively tested strains of pennate diatoms (Bacillariophyceae) using *cox*1, *rbc*L and *ITS*

Andréa de O. da R. Franco[1] | Matt P. Ashworth[2] | Clarisse Odebrecht[1]

[1]Institute of Oceanography, Federal University of Rio Grande – FURG, Rio Grande, Brazil

[2]Department of Molecular Biosciences, UTEX Culture Collection of Algae, University of Texas at Austin, Austin, Texas, USA

Correspondence
Andréa de O. da R. Franco, Institute of Oceanography, Federal University of Rio Grande – FURG, Av. Itália, km 8, CEP 96203-900, Rio Grande, Brazil.
Email: andrea.fitoplancton@gmail.com

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 15/2016 and 203883/2017-9; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 15/2016; Fundação de Amparo à Ciência e Tecnologia do Estado do Rio Grande do Sul

Abstract

Several automated molecular methods have emerged for distinguishing eukaryote species based on DNA sequence data. However, there are knowledge gaps around which of these single-locus methods is more accurate for the identification of microalgal species, such as the highly diverse and ecologically relevant diatoms. We applied genetic divergence, Automatic Barcode Gap Discovery for primary species delimitation (ABGD), Assemble Species by Automatic Partitioning (ASAP), Statistical Parsimony Network Analysis (SPNA), Generalized Mixed Yule Coalescent (GMYC) and Poisson Tree Processes (PTP) using partial *cox*1, *rbc*L, *5.8S* + *ITS*2, *ITS*1 + *5.8S* + *ITS*2 markers to delineate species and compare to published polyphasic identification data (morphological features, phylogeny and sexual reproductive isolation) to test the resolution of these methods. ASAP, ABGD, SPNA and PTP models resolved species of *Eunotia*, *Seminavis, Nitzschia, Sellaphora* and *Pseudo-nitzschia* corresponding to previous polyphasic identification, including reproductive isolation studies. In most cases, these models identified diatom species in similar ways, regardless of sequence fragment length. GMYC model presented smallest number of results that agreed with previous published identification. Following the recommendations for proper use of each model presented in the present study, these models can be useful tools to identify cryptic or closely related species of diatoms, even when the datasets have relatively few sequences.

KEYWORDS
ABGD, ASAP, GMYC, microalgae species, molecular taxonomy, PTP, SPNA, systematics, taxonomy

INTRODUCTION

THE identification of eukaryotic microalgae is not trivial, considering that some species lack obvious features for identification. Diverse criteria have been used to identify species, these include the biological, ecological, evolutionary and phylogenetic species concepts. All these approaches assume that species are metapopulation lineages that evolved separately (De Queiroz, 2007; Leliaert et al., 2014).

Diatoms are one of the most species-rich eukaryotic algal groups, with an estimated 30,000–100,000 species (Mann & Vanormelingen, 2013). They can be found in many kinds of habitats: marine, freshwater, ice, terrestrial (soils, sand, mosses, rock walls) and can even be found as endosymbionts of dinoflagellates and epibionts of animals and macroalgae (Ashworth et al., 2022; Hehenberger et al., 2016; Mann & Vanormelingen, 2013; Round et al., 1990). The diatoms are also notable among the microalgae in their distinctive morphology; diatoms have a siliceous cell wall (the "frustule"), which can show high levels of morphological variability between taxa. Traditionally, the morphology of the diatom frustule has been used for species identification and taxonomical studies (Mann, 1999). However, the so-called crypitc, semicryptic or pseudocryptic species have not shown sufficient morphological distinctions to separate species even with advances in electron microscopy (Mann et al., 2010). Some of these species have shown some level of reproductive isolation, and several studies have distinguished species using molecular tools (Amato et al., 2007; Evans et al., 2007; Mann et al., 2008; Poulíčková et al., 2010; Quijano-Scheggia et al., 2009). With regard to reproductive isolation, this has been tested experimentally in only a few diatom taxa, due to the difficulties in inducing sexual reproduction in culture (Chepurnov et al., 2004; Poulíčková & Mann, 2019).

Molecular data can be applied in different ways to delimit algal taxa: (a) using thresholds of sequence similarity and genetic divergence to define species (diagnosable genotypic concept) (Evans et al., 2007; Hamsher et al., 2011; MacGillivary & Kaczmarska, 2011), (b) defining molecular clades as species based on one or more molecular markers (phylogenetic concept) (Leliaert et al., 2014), (c) recognizing significant changes in the pace of branching events in the phylogenetic tree as indicators of speciation (coalescent methods – phylogenetic concept) (Leliaert et al., 2014) and (d) occurrence of compensatory base pair changes (CBCs) in the secondary structure of ITS (Coleman, 2009). Regarding the diagnosable genotypic concept, there are several methods used to estimate the genetic divergence between putative species. The p-distance (simple uncorrected distance) is a method that quantifies the proportion (p) of nucleotide sites which differ between two sequences and is used to make pairwise comparisons of phylogenetically close species (Nei & Kumar, 2000). The p-distance has been used together with other methods (polyphasic approach) to identify pennate diatoms, e.g., *Eunotia bilunaris* (Ehrenberg) Schaarschmidt sensu lato (Vanormelingen et al., 2008), the *Nitzschia palea* (Kutzing) complex (Trobajo et al., 2009, 2010) and the *Asterionellopsis glacialis* (Castracane) Round species complex (Franco et al., 2016; Kaczmarska et al., 2014).

Automated single-locus computational methods have emerged to assist molecular taxonomy, such as the Automatic Barcode Gap Discovery for primary species delimitation (ABGD) (Puillandre et al., 2012), the Assemble Species by Automatic Partitioning (ASAP) (Puillandre et al., 2021) and the Statistical Parsimony Network Analysis (SPNA) (Clement et al., 2000; Templeton et al., 1992), which are based on genetic divergence (diagnosable genotypic concept). In addition, the Poisson Tree Processes (PTP) (Zhang et al., 2013) and General Mixed Yule Coalescent method (GMYC) (Fujisawa & Barraclough, 2013) follow the phylogenetic concept. The ABGD is an automated interactive process that sorts sequences into putative species and selects the reliable delimitations of species based on pairwise distance among sequences. The models search for a possible barcode gap, the individuals of each suggested species having smaller pairwise distances than among potential distinct species (Puillandre et al., 2012). The ASAP is a model that merges sequences in "groups" (or species) by ascending hierarchical clustering successively. After that, each putative species group is assigned a single ASAP score obtained by combination of barcode gap widths and the probability of groups to be panmictic species, to indicate the most reliable species delimitation (Puillandre et al., 2021). ABGD outperformed ASAP considering the species delimitation with the best ASAP score; however, when the ASAP species identifications with the first and second best ASAP scores were considered, both models showed similar results to real datasets (5–643 species of gastropod; decapod crustaceans; amphibians; cladocerans; mammals; insects; birds). In addition, both models also showed comparable results using simulated datasets (Puillandre et al., 2021). SPNA estimates species by comparing the genetic distance between haplotypes; the model calculates the maximum number of mutational steps constituting a parsimonious connection between two haplotypes. The haplotypes that constitute the same network are considered to represent the same species (Clement et al., 2000; Templeton et al., 1992).

The PTP and GMYC models, on the other hand, are tree-based species delimitation methods, which delineate species based on significant changes in the pace of branching events in a molecular phylogeny (Fujisawa & Barraclough, 2013; Pons et al., 2006; Zhang et al., 2013). While PTP directly uses the number of base pair substitutions between branching events (Zhang et al., 2013), the GMYC model considers time among branching events. For this reason, the GMYC model needs ultrametric trees as input data (Fujisawa & Barraclough, 2013; Pons et al., 2006). In general, both methods show comparable results using real datasets (species of *Gallotia* lizards; arthropod metabarcoding sequences; bears; bees) and simulated datasets (Luo et al., 2018; Zhang et al., 2013). However, the PTP outperformed the GMYC model when the genetic distance between species in the dataset was very small (more closely related species, Zhang et al., 2013) or in dataset with fewer species (Luo et al., 2018).

Few studies used automated species delimitation models to investigate diatom taxonomy and only five species complexes were analyzed by these methods: *N. palea* complex (Rimet et al., 2014); *Pinnularia subgibba* complex (Kollár et al., 2019); *Cylindrotheca closterium* complex (Stock et al., 2019); *Pinnularia borealis* complex (Pinseel et al., 2020) and *Achnanthidium minutissimum* complex (Rimet et al., 2023). GMYC failed to delimitate taxa of the *N. palea* complex (Rimet et al., 2014), and PTP indicated that *C. closterium* is a true complex of species with at least 12 species. The phylogenetic position of *C. closterium* strains showed a link with temperature, as strains from the same clade showed a similar thermal optimum, which was correlated to the water temperature of the place of origin (Stock et al., 2019). The SPNA and ABGD indicated the presence of five and seven putative species within the *A. minutissimum* complex using amplicon/metabarcode sequences (Rimet et al., 2023) of barcode region *rbc*L312 pb (Kermarrec et al., 2013; Vasselon et al., 2017).

The studies that address *Pinnularia* taxonomy used different models simultaneously. For *P. subgibba* Krammer, a species originally described by morphology, part of the strains formed a clade of reproductively compatible clones (Poulíčková et al., 2007), which was described as *Pinnularia lacustrigibba* (Poulíčková et al., 2018). The PTP, GMYC and Statistical Parsimony Network Estimation (SPNE) indicated that the *P. subgibba* complex in fact comprised 14 or 15 species (Kollár et al., 2019), one of them corresponded to the reproductively compatible *Pinnularia lacustrigibba* (Kollár et al., 2019; Poulíčková et al., 2018). The global study of diversification in the *P. borealis* complex used hundreds of samples and, based on the consensus of five automated delimitation models (GMYC, PTP-ML, bPTP-h, ABGD, SPNA/SPNE), indicated the presence of 126 species in this globally distributed complex.

Relatively few studies used automated single-locus computational methods as a tool for species identification within microalgae lineages as diverse as the dinoflagellates *Symbiodinium* and *Gymnodinium* (Annenkova et al., 2020; Correa & Baker, 2009), cryptophytes (Hoef-Emden, 2012), the rhodophytes of the Cyanidiales (Hsieh et al., 2015), *Choricystis*, *Coccomyxa* and *Chlorella*-like species in the Trebouxiophyceae (Kulakova et al., 2020; Malavasi et al., 2016; Zou, Fei, Song, et al., 2016), *Haematococcus* and *Scenedesmus* in the Chlorophyceae (Allewaert et al., 2015; Zou, Fei, Wang, et al., 2016) and the five aforementioned diatom species complexes (Kollár et al., 2019; Pinseel et al., 2020; Rimet et al., 2014, 2023; Stock et al., 2019). However, most automated species delimitation models (i.e., ABGD, ASAP, PTP, GMYC) were developed and tested using simulated datasets and/or real datasets with species of macroorganisms (Pons et al., 2006; Puillandre et al., 2012, 2021; Zhang et al., 2013). Macroorganisms in general have a relatively small population size and more restrictive dispersal capabilities than microorganisms, influencing gene flow between distinct populations and the speciation process and rate, which may influence the accuracy of these models (Luo et al., 2018; Puillandre et al., 2012, 2021). The present study is the first to compare the results of different automated molecular species delimitation methods or models using data from previously identified diatom species. The selected models are available free of charge and can be used easily through online web server or software. Our goal in the present study is to test the accuracy of single-locus methods (genetic divergence thresholds, ABGD, ASAP, SPNA, GMYC and PTP) using diatom species datasets. In order to avoid neglecting cryptic, semicryptic or pseudocryptic species and generate useful information to compare close species, we restricted the analysis to diatom species and complexes whose species delimitation was done a priori, based on polyphasic taxonomy studies that used morphological features, molecular information and the sexual reproductive isolation data.

MATERIALS AND METHODS

Bibliographic survey

We searched for sequences of *cox*1, *rbc*L, *5.8S+ITS*2 and *ITS*1*+5.8S+ITS*2 from diatoms, which were identified using both molecular information and morphological features and also had their sexual compatibility tested in previous studies. These molecular markers originate from distinct organelles, respectively, the chloroplast, mitochondria and nucleus, and are often used to identify closely related diatom species. Regarding reproductive criterion, the diatom strains were considered conspecific when experiments show that the crosses between them produced descendants through sexual reproduction. To guarantee that observed auxospores were the result of sexual reproduction among strains, we considered the results of crosses between heterothallic strains only, where the progeny was generated by crossing distinct parental strains with documented gametogenesis. Eight studies presented sequences available on GenBank from strains that fill the requirements cited above. These selected strains correspond to 17 taxa in the class Bacillariophyceae (Table 1).These sequences were used to estimate the intra- and interspecific genetic divergence (p-distance) and to delimit species by Automatic Barcode Gap Discovery for primary species delimitation (ABGD), Assemble Species by Automatic Partitioning (ASAP), Statistical Parsimony Network Analysis (SPNA), Generalized Mixed Yule Coalescent (GMYC) and Poisson Tree Processes (PTP) that were calculated in two different ways: PTP maximum likelihood (PTP-ML) and Bayesian PTP heuristic (bPTP-h).

Calculation of genetic divergence

The alignment of sequences and the calculation of genetic divergence (p-distance) were carried out in the MEGA6 Program (Nei & Kumar, 2000; Tamura et al., 2013); transitions and transversions were included in the calculation; gaps and missing data were treated by pairwise deletion. The alignments were performed using ClustalW (Thompson et al., 1997), with the standard parameters of MEGA6 for *cox*1 and *rbc*L (Gap Opening Penalty 10 and Gap Extension Penalty 6.66) (Tamura et al., 2013). The *5.8S+ITS*2 marker was aligned using standard parameters or the Gap Opening Penalty 10 and Gap Extension Penalty 1.2, following the recommendations of Moniz and Kaczmarska (2010). When necessary, the ITS alignments were manually corrected. Each diatom genus was aligned separately, since the genera were not compared to

TABLE 1 Genetic divergence calculated in the present study from strains with phylogeny and sexual compatibility tested by Vanormelingen et al. (2008), Trobajo et al. (2009), Amato et al. (2007), Quijano-Scheggia et al. (2009), Vanormelingen et al. (2013), Behnke et al. (2004), De Decker et al. (2018) and Kaczmarska et al. (2009).

Taxon	Provenance of sequences	Barcode marker	N (T)	Intra-specific genetic divergence	N (T)	Inter-specific genetic divergence (different clades)

Eunotia bilunaris (Ehrenberg) Schaarschmidt	Vanormelingen et al. (2008)	*5.8S+ITS2* 372 alignable positions with gaps	28 (1)
			10 (1)
				E. bilunaris 'robust': 0%–1% (0%±0.3);
				E. bilunaris 'slender': 0%–2% (1%±0.6);
					40 (2)	('robust', 'slender'): 12% (±0.2);

		*cox*1 400 bp	1 (1)	*E. bilunaris* 'robust' 0%	2 (2)	('robust', 'slender'): 4% (±0);

		rbcL 540 bp	21 (1)
			3 (1)
				E. bilunaris 'robust': 0% (±0);
				E. bilunaris 'slender': 0% (±0);
					21 (2)	('robust', 'slender'): 2% (±0);

Nitzschia palea (Kutzing) Smith	Trobajo et al. (2009)	*rbcL* 540 bp	6 (1)
			3 (1)
				N. palea 1 (Belgium): 0% (±0);
				N. palea 2 (Brazil, Paraguai, Spain A4): 0% (±0);
					12 (2)	('Belgium', 'Brazil, Paraguai, Spain A4'): 1% (±0);

		*cox*1 371 bp	1 (1)	*N. palea* 2 (Brazil, Paraguai, Spain A4): 0% (±0);		

Pseudo-nitzschia pseudodelicatissima (Hasle) Hasle complex	Amato et al. (2007)	*5.8S+ITS2* 363 alignable positions with gaps	10 (1)
			1 (1)
			1 (1)
				P. pseudodelicatissima: 0% (±0);
				P. mannii Amato & Montresor: 0% (±0);
				P. calliantha Lundholm, Moestrup & Hasle: 0% (±0);
					10 (2)
					10 (2)
					4(2)
						(*P. pseudodelicatissima, P. mannii*): 9% (±0);
						(*P. pseudodelicatissima, P. calliantha):* 9% (±0);
						(*P. calliantha, P. mannii):* 3% (±0)

Pseudo-nitzschia delicatissima (Cleve) Heiden complex	Quijano-Scheggia et al. (2009)	*5.8S+ITS2* 337 alignable positions with gaps	210 (1)
			6 (1)
			6 (1)
				P. delicatissima: 0% (±0); *P. arenysensis* Quijano-Scheggia, Garcés & Lundholm: 0% (±0);
				P. dolorosa Lundholm & Moestrup: 0% (±0)
					84 (2)
					84 (2)
					16 (2)
						(*P. delicatissima, P. arenysensis*): 4% (±0);
						(*P. delicatissima, P. dolorosa*): 9% (±0);
						(*P. arenysensis, P. dolorosa*): 9% (±0);

Sellaphora pupula (Kützing) Mereschkovsky, complex	Vanormelingen et al. (2013)	*cox*1	3 (1)
			15 (1)
			36 (1)
				Sellaphora auldreekie Mann & McDonald*:* 0% (±0);
				S. pupula agg. 'coarse *auldreekie*'*:* 0% (±0);
				S. pupula agg. 'southern *auldreekie*'*:* 0% (±0);
					54 (2)
					18 (2)
					27 (2)
						(Coarse, southern): 5% (±0);
						(coarse, *auldreekie*): 6% (±0);
						(*auldreekie*, southern): 5% (±0);

		*rbc*L *540 bp*		*S. auldreekie:* 0% (±0);
				S. pupula agg. 'coarse *auldreekie*'*:* 0% (±0);
				S. pupula agg. 'southern *auldreekie*':0%
					6 (2)
					9 (2)
					6 (2)
						(coarse, southern): 1% (±0);
						(coarse, *auldreekie*): 1% (±0);
						(*auldreekie*, southern): 1% (±0);

	Behnke et al. (2004)	*5.8S+ITS2* 418 alignable positions with gaps	21(1)
			21(1)
				S. pupula agg. 'pseudocapitate': 0% (±0.1);
				Sellaphora backfordenses Mann & Droop: 0%–1% (1%±0.4);
					49 (2)	*S. pupula* agg. 'pseudocapitate', *S. backfordenses:* 6%–7% (7%±0.4);

Seminavis robusta Danielidis & Mann	De Decker et al. (2018)	*5.8S+ITS2* 334 alignable positions with gaps	235(1)
			168(1)
				Same clade: 0–1% (0.1%±0.3);
				Different clades (1, 2): 0%–1% (0.3%±0.2);
					–	–

		rbcL 540 bp	1991(1)
			1107(1)
				Same clade: 0% (±0)
				Different clades (1, 2): 0%–1% (0.4%±0.03)
					–	–

Tabularia fasciculata (Agardh) Williams & Round	Kaczmarska et al. (2009)	*5.8S+ITS2* 388 alignable positions with gaps	9(1)
			12(1)
				Same clade: 0%–1% (0.3%±0.1)
				Different clades: 3% (±0)
					–	–

		rbcL 540 bp	1(1)	Different clades: 1%	–	–

Note: The numbers of pairwise comparisons (N) and taxa (T), minimum and maximum values of genetic divergence (mean±standard deviation) are shown.

each other for the calculation of the intra- and interspecific divergences. Therefore, the *5.8S+ITS2* alignments differed in length between genera, even though they corresponded to the same barcode region (the complete *5.8S* and up to the conserved region of the helix III motif in *ITS2*), as suggested by Moniz and Kaczmarska (2010). The length variation of *5.8S+ITS2* ranged from 300 to 500 bp. The *rbc*L alignments were trimmed to correspond with the proposed diatom barcode fragment: 540 bp starting 139 amino acids downstream from the start codon, a conserved region starting with an "AGA" triplet, and ending after 540 bp on a "TAA" triplet, as proposed by MacGillivary and Kaczmarska (2011). The *cox*1 fragment corresponded to the region near the 5′ end, with approximately 431 bp (Moniz & Kaczmarska, 2009). The trimmed alignments were used to calculate p-distance. So we calculated genetic divergence using barcode regions (Table 1) to compare our results with thresholds proposed to delimit diatom species.

Alignments and trees used as input for automated species delimitation models

The target clades used in the present study correspond to selected species or populations. The clades were proposed based on trees produced by previous polyphasic taxonomic studies of the selected species (Amato et al., 2007; Behnke et al., 2004; De Decker et al., 2018; Kaczmarska et al., 2009; Quijano-Scheggia et al., 2009; Trobajo et al., 2009; Vanormelingen et al., 2008, 2013) (see the Bibliographic survey).

The input alignments used in the ABGD, ASAP and SPNA analyses were built using ClustalW in the MEGA6 Program (Tamura et al., 2013; Thompson et al., 1997), with the same parameters cited above. However, we used the gene markers trimmed in two different ways: (A) as large as possible; (B) the barcode region only: *cox*1 431 bp (Moniz & Kaczmarska, 2009), *rbc*L 540 bp (MacGillivary & Kaczmarska, 2011) and *5.8S+ITS2* (Moniz & Kaczmarska, 2010). The barcode regions were used in the previously described calculation of genetic divergence. We considered gene markers with at least four available sequences per genus; thus, alignments of *Tabularia fasciculata*−*rbc*L (two sequences) were discarded.

In order to obtain the same number of sequences per target clade ("balanced" alignments), we randomly removed a few sequences from clades that had a higher number of sequences. In summary, 22 "balanced" alignments were constructed with each genus or species complex separately (Table 2). In addition, a second type of dataset was built, by removing all identical sequences from previous cited alignments, thus obtaining 21 alignments containing different haplotypes only (Table 3).

Despite the random deletion of sequences to obtain the same number of sequences per target clade, our dataset represents almost all the genetic variability available in the studied species. One sequence per diatom strain/individual was included. Therefore, some haplotypes from the same strain of *Eunotia* and *Sellaphora* (*ITS* region and *5.8S+ITS2*) were excluded from the analyses. Vanormelingen et al. (2008) suggested that this intraclonal variability in the *ITS* region is related to distinct copies of the molecular marker in the genome. The random exclusion of sequences removed one haplotype of *rbc*L in *Seminavis robusta* and *Sellaphora pupula* agg. 'southern auldreekie'.

The phylogenetic trees used as input for PTP and GMYC were built with the same 43 alignments used in the ABGD, ASAP and SPNA. The best-fit substitution model used to construct phylogenetic trees was estimated for each alignment separately, by the Akaike information criterion (Akaike, 1974) using JModeltest 2 (Darriba et al., 2012).

For the construction of the ultrametric trees used as GMYC input, the best-fit molecular clock and the branch length estimation model were selected based on the harmonic mean (Kass & Raftery, 1995) of models available in MrBayes 3.2.3 (Ronquist et al., 2012).

The Bayesian analyses to build the trees for PTP and the ultrametric trees for GMYC were performed using MrBayes 3.2.3 by two independent Markov Chain Monte Carlo (MCMC) runs, each with four chains of 1–5 million generations each. The convergence was initially verified when the average standard deviation in split frequencies was <0.01. The consensus trees were built from topologies sampled every 100 generations, with 25% of the generations discarded as burn-in (Ronquist et al., 2012). The MCMC performance and convergence between independent runs were confirmed with the likelihood plots for each run and the effective sample size (>200) using the software Tracer 1.5 (Rambaut et al., 2018). The obtained consensus trees were converted to NEXUS format by FigTree v.1.4.0 and used as input of PTP. The ultrametric trees were converted to NEWICK format by input of FigTree v.1.4.0 and used as GMYC.

Automatic barcode gap discovery for primary species delimitation (ABGD)

The analysis was performed using the default parameters in the ABGD webserver (Puillandre et al., 2012) available at https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html, except for the method to calculate the pairwise distance, for which we used simple distance (p-distance). We considered the species delimitation that occurred in a higher number of partitions (initial+recursive).

Assemble species by automatic partitioning (ASAP)

The ASAP model was carried out with the same input alignments used in the ABGD analysis, using the

TABLE 2 The number of species obtained by ABDG, ASAP, SPNA, GMYC, PTP-ML and bPTP-h models using "balanced" input dataset and the polyphasic identification (morphological, molecular and sexual criteria) according to previous studies: Vanormelingen et al. (2008), Trobajo et al. (2009), Amato et al. (2007), Quijano-Scheggia et al. (2009), Vanormelingen et al. (2013), Behnke et al. (2004), De Decker et al. (2018) and Kaczmarska et al. (2009).

Marker (number of base pair or alignable positions with gaps)	Input data: Number of sequences (seq) for species or population	ABGD	ASAP	SPNA	GMYC	PTP-ML and bPTP-h	Previously identified species
*rbc*L (1341 bp)	3 seq *E. bilunaris* 'robust' + 3 seq *E. bilunaris* 'slender'	2	2	2	3[a]	2 and 2	2
*rbc*L (540 bp)	3 seq *E. bilunaris* 'robust' + 3 seq *E. bilunaris* 'slender'	2	2	Y	4[a]	2 and 2	2
*ITS*1 + *5.8S* + *ITS*2 (630 bp)	3 seq *E. bilunaris* 'robust' + 3 seq *E. bilunaris* 'slender'	2	3[a]	2	3[a]	2 and 2	2
5.8S + *ITS*2 (362 bp)	3 seq *E. bilunaris* 'robust' + 3 seq *E. bilunaris* 'slender'	2	4[a]	2	3[a]	2 and 2	2
*rbc*L (1355 bp)	3 seq *N. palea* 1 + 3 sequences *N. palea* 2	2	2	1[a]	3[a]	2 and 2	2
*rbc*L (540 bp)	3 seq *N. palea* 1 + 3 seq *N. palea* 2	1[a]	2	1[a]	2	2 and 3	2
*ITS*1 + *5.8S* + *ITS*2 (770 bp)	4 seq *Pseudo-nitzschia arenysensis* + 4 seq *P. delicatissima* + 4 seq *P. dolorosa*	3	3	3	2[a]	3 and 3	3
5.8S + *ITS*2 (337 bp)	4 seq *P. arenysensis* + 4 seq *P. delicatissima* + 4 seq *P. dolorosa*	3	3	3	6[a]	3 and 3	3
*ITS*1 + *5.8S* + *ITS*2 (856 bp)	2 seq *P. pseudodelicatissima* + 2 seq *P. mannii* + 2 seq *P. calliantha*	1[a]	3	4[a]	3	3 and 3	3
5.8S + *ITS*2 (363 bp)	2 seq *P. pseudodelicatissima* + 2 seq *P. mannii* + 2 seq *P. calliantha*	1[a]	3	3	3	3 and 3	3
cox1 (624 bp)	3 seq *S. pupula* agg. 'southern *auldreekie*' + 3 seq *S. pupula* agg. 'coarse *auldreekie*' + 3 seq *Sellaphora auldreekie*	3	3	3	3	3 and 3	3
cox1 (422 bp)	3 seq *S. pupula* agg. 'southern *auldreekie*' + 3 seq *S. pupula* agg. 'coarse *auldreekie*' + 3 seq *Sellaphora auldreekie*	3	3	Y	5[a]	3 and 3	3
*rbc*L (1398 bp: ABGD, SPNA, GMYC and PTP; 1390 bp: ASAP)	2 seq *S. pupula* agg. 'southern *auldreekie*' + 2 seq *S. pupula* agg. 'coarse *auldreekie*' + 2 seq *Sellaphora auldreekie*	1[a]	3	2[a]	3	3 and 3	3
*rbc*L (540 bp)	2 seq *S. pupula* agg. 'southern *auldreekie*' + 2 seq *S. pupula* agg. 'coarse *auldreekie*' + 2 seq *Sellaphora auldreekie*	1[a]	3	1[a]	3	3 and 3	3
*ITS*1 + *5.8S* + *ITS*2 (889 bp)	4 seq *Sellaphora backfordensis* + 4 seq *S. pupula* agg. 'pseudocapitate'	2	2	4[a]	2	2 and 2	2
5.8S + *ITS*2 (418 bp)	4 seq *Sellaphora backfordensis* + 4 seq *S. pupula* agg. 'pseudocapitate'	2	2	3[a]	2	2 and 4[a]	2
*rbc*L (1096 bp)	27 seq *S. robusta* – clade 1 + 27 seq *S. robusta* – clade 2	1	2[a]	1	2[a]	2[a] and 54[a]	1
*rbc*L (540 bp)	27 seq *S. robusta* – clade 1 + 27 seq *S. robusta* – clade 2	1	2[a]	1	2[a]	2[a] and 54[a]	1
ITS1 + 5.8S + ITS2 (420 bp)	12 seq *S. robusta* – clade 1 + 12 seq *S. robusta* – clade 2	2[a]	2[a]	1	2[a]	17[a] and 21[a]	1
5.8S + *ITS*2 (334 bp)	12 seq *S. robusta* – clade 1 + 12 seq *S. robusta* – clade 2	1	2[a]	1	22[a]	17[a] and 21[a]	1
5.8S + *ITS*2 (388 bp)	3 seq *T. fasciculata* – clade Bw + 3 seq *T. fasciculata* – clade An	2[a]	2[a]	2[a]	2[a]	2[a] and 2[a]	1
5.8S + *ITS*2 (389 bp)	14 seq *T. fasciculata* – clade Bw + 14 seq *T. fasciculata* – clade An	2[a]	2[a]	2[a]	2[a]	2[a] and 2[a]	1

Note: The ABDG, ASAP, SPNA, GMYC, PTP-ML and bPTP-h results were obtained using input data with the same number of sequences per species and are shown for each barcode marker, with the number of sequences for each taxon and clade. 'Y' No run: Unable to calculate any connection among sequences: software does not provide any network and closes itself.

[a]Identification of automatic method does not match with previous study.

TABLE 3 The number of species obtained by ABDG, ASAP, SPNA, GMYC, PTP-ML and bPTP-h models using only haplotype data and the polyphasic identification (morphological, molecular and sexual criteria) according to previous studies: Vanormelingen et al. (2008), Trobajo et al. (2009), Amato et al. (2007), Quijano-Scheggia et al. (2009), Vanormelingen et al. (2013), Behnke et al. (2004), De Decker et al. (2018) and Kaczmarska et al. (2009).

Marker (number of base pair or alignable positions with gaps)	Input data: Number of sequences/haplotypes (haplo) for species or population	ABGD	ASAP	SPNA	GMYC	PTP-ML and bPTP-h	Previously identified species
*rbc*L (1341 bp)	3 haplo *E. bilunaris* 'robust' + 2 haplo *E. bilunaris* 'slender'	2	2	2	2	2 and 2	2
*rbc*L (540 bp)	1 haplo *E. bilunaris* 'robust' + 1 haplo *E. bilunaris* 'slender'	1[a]	X	Y	X	X	2
*ITS*1 + *5.8S* + *ITS*2 (628 bp)	3 haplo *Eunotia. bilunaris* 'robust' + 3 haplo *E. bilunaris* 'slender'	2	3[a]	2	3[a]	2 and 2	2
5.8S + *ITS*2 (362 bp)	3 haplo *E. bilunaris* 'robust' + 2 haplo *E. bilunaris* 'slender'	2	3[a]	2	2	2 and 2	2
*rbc*L (1355 bp)	1 haplo (belgium) *N. palea* 1 + 2 haplo *N. palea* 2	1[a]	2	1[a]	X	X	2
*rbc*L (540 bp)	1 haplo (belgium) *N. palea* 1 + 2 haplo *N. palea* 2	1[a]	2	1[a]	X	X	2
*ITS*1 + *5.8S* + *ITS*2 (770 bp)	2 haplo *P. arenysensis* + 1 haplo *P. delicatissima* + 3 haplo *P. dolorosa*	3	3	3	3	3 and 3	3
5.8S + *ITS*2 (337 bp)	1 haplo *P. arenysensis* + 1 haplo *P. delicatissima* + 2 haplo *P. dolorosa*	1[a]	3	3	3	3 and 3	3
*ITS*1 + *5.8S* + *ITS*2 (856 bp)	2 haplo *P. pseudodelicatissima* + 2 haplo *P. mannii* + 2 haplo *P. calliantha*	1[a]	3	4[a]	3	3 and 3	3
5.8S + *ITS*2 (363 bp)	1 haplo *P. pseudodelicatissima* + 2 haplo *P. mannii* + 1 haplo *P. calliantha*	1[a]	3	3	2[a]	3 and 3	3
cox1 (624 bp)	1 haplo *S. pupula* agg. 'southern *auldreekie*' + 2 haplo *S. pupula* agg. 'coarse *auldreekie*' + 1 haplo *Sellaphora auldreekie*	1[a]	2[a]	3	3	3 and 3	3
cox1 (422 bp)	1 haplo *S. pupula* agg. 'southern *auldreekie*' + 1 haplo *S. pupula* agg. 'coarse *auldreekie*' + 1 haplo *Sellaphora auldreekie*	1[a]	2[a]	Y	X	X	3
*rbc*L (1398 bp: ABGD, SPNA, GMYC and PTP; 1390 bp: ASAP)	1 haplo *S. pupula* agg. 'southern *auldreekie*' + 1 haplo *S. pupula* agg. 'coarse *auldreekie*' + 1 haplo *Sellaphora auldreekie*	1[a]	2[a]	2[a]	X	X	3
*rbc*L (540 bp)	1 haplo *S. pupula* agg. 'southern *auldreekie*' + 1 haplo *S. pupula* agg. 'coarse *auldreekie*' + 1 haplo *Sellaphora auldreekie*	1[a]	2[a]	1[a]	X	X	3
*ITS*1 + *5.8S* + *ITS*2 (888 bp)	4 haplo *Sellaphora backfordensis* + 3 haplo *S. pupula* agg. 'pseudocapitate'	2	2	4[a]	2	2 and 2	2
5.8S + *ITS*2 (418 bp)	4 haplo *Sellaphora backfordensis* + 3 haplo *S. pupula* agg. 'pseudocapitate'	2	2	3[a]	2	2 and 2	2
*rbc*L (1096 bp)	1 haplo *S. robusta* – clade 1 + 1 haplo *S. robusta* – clade 2	1	X	1	X	X	1
*rbc*L (540 bp)	1 haplo *S. robusta* – clade 1 + 1 haplo *S. robusta* – clade 2	1	X	1	X	X	1
ITS1 + 5.8S + ITS2 (420 bp)	4 haplo *S. robusta* – clade 1 + 5 haplo *S. robusta* – clade 2	1	3[a]	1	W	5[a] and 8[a]	1
5.8S + *ITS*2 (334 bp)	3 haplo *S. robusta* – clade 1 + 5 haplo *S. robusta* – clade 2	1	7[a]	1	W	2[a] and 8[a]	1
5.8S + *ITS*2 (389 bp)	2 haplo *T. fasciculata* – clade Bw + 4 haplo *T. fasciculata* – clade An	2[a]	2[a]	2[a]	2[a]	2[a] and 2[a]	1

Note: 'X': No run: ASAP not performed or input trees (PTP and GMYC) not built, due to the low number of sequences. 'W': No run: Input tree not bifurcated, impossible to correct, since all (multiple) branches originated from the same node. 'Y': No run: Unable to calculate any connection among sequences: software does not provide any network and closes itself.

[a] Identification of automatic method does not match with previous study.

ASAP webserver (Puillandre et al., 2021) available at https://bioinfo.mnhn.fr/abi/public/asap/asapweb.html. The method to calculate the pairwise distance chosen in the webserver was simple distance (p-distance). We considered the delimitation of species with the lowest ASAP score according to Puillandre et al. (2021); however, if two or more species delimitations presented the same lowest ASAP score, the one with the lowest *p*-value was chosen.

Statistical parsimony network analysis (SPNA)

The SPNA model was carried out with the same input alignments used in the ABGD and ASAP analyses, using TCS v.1.21 with default configuration: the connection limit was calculated with 95% significance and with gaps treated as the fifth state (Clement et al., 2000).

Poisson tree processes (PTP)

PTP is a model to infer putative species through a phylogenetic input tree based on the number of bp substitutions between branching (Zhang et al., 2013). The PTP analyses were performed on the Species Delimitation web server (https://species.h-its.org). The PTP analyses were based on two strategies: (A) PTP maximum likelihood (PTP-ML) and (B) Bayesian PTP heuristic (bPTP-h) (Zhang et al., 2013). Regarding the Bayesian implementation in PTP web server (https://species.h-its.org/), we applied the maximum number of generations (500,000), 0.25 burn-in and default sampled trees. The Bayesian likelihood plots were checked for convergence, following the recommendations of Zhang et al. (2013).

Generalized mixed yule coalescent (GMYC)

Ultrametric and fully bifurcating trees are required as input for GMYC analyses. When the ultrametric trees were not fully bifurcating, they were corrected using *ape* package with *multi2di* function in R (Fujisawa & Barraclough, 2013; Paradis et al., 2004). The GMYC analysis was performed using the *splits* package with *gmyc* function in R (Ezard et al., 2009; Fujisawa & Barraclough, 2013).

RESULTS

Genetic divergence in reproductively compatible strains

Intraspecific diatom genetic divergence values ranged from 0% to 1% (*rbc*L), 0% (*cox*1) and 0%–3% (*5.8S + ITS*2), while the interspecific divergence (with species defined by polyphasic taxonomy) was 1%–2% (*rbc*L), 4%–6% (*cox1*) and 3%–12% (*5.8S + ITS*2) (Table 1). All reproductively isolated species were monophyletic, and each species showed internal clades that likely represent population-level variation (Table 1).

ABGD, ASAP, SPNA, GMYC and PTP models in reproductively compatible strains

ABGD, ASAP, SPNA, GMYC and PTP were conducted on alignments and trees built with the same sequences used to calculate genetic divergence trimmed as barcode region and as the larger fragment with the highest number of bp possible. The sequences were organized in two different types of datasets: (A) 22 alignments and trees built using the same number of sequences per studied clade, tested species or population (Table 2); (B) 21 alignments and trees built using haplotypes only, identical sequences were removed of this dataset (Table 3). All alignments and trees used as input of models were built for each genus or species complex separately (Tables 2 and 3).

Overall, in the first type of input data (the same number of sequences per studied clade), the ABGD, ASAP and PTP showed similar results (Figure 1A), with a high number of analyses agreeing with previously-published species delimitation. The PTP-ML and bPTP-h models delineated the species in accordance with previous identification in 16 and 15 cases, respectively, followed by ASAP and ABGD that agreed with the polyphasic species identification in 14 cases. The SPNA and GMYC presented 11 and 8 analyses that delineated the species according to previous identification (Table 2; Figure 1A).

The use of dataset composed by haplotype only, affected the GMYC performance positively, as the number of "wrong" identifications reduced from 12 to 3 (Figure 1A,B). This happened due to the increase of the number of "no run" analyses (6), when models were not able to be used with dataset (see details for each case in Tables 2 and 3 – footnotes), and the GMYC analyses that agreed with the previous published identification using the datasets with haplotypes (4) (Tables 2 and 3).

On the other hand, when using dataset composed by haplotype only, the PTP-ML and bPTP-h showed a lower number of analyses that agreed with previous identification (Figure 1A,B) because the number of "no run" increased, since the alignments with less than four sequences cannot be used to build an input tree (Tables 2 and 3). ABGD and ASAP models were negatively affected by dataset comprising only haplotypes as the number of cases that the results disagree with the previous identification increased (Figure 1A,B). The SPNA does not show any difference between two types of datasets, since the software automatically removes identical sequences and builds networks only using the haplotypes. The ASAP, GMYC and PTP models were unable to resolve the identification when the dataset consisted of only one species (Tables 2 and 3: *S. robusta* or *T. fasciculata*), suggesting that at least two nominate species must be included in these analyses.

The ABGD and SPNA methods agreed with the previous polyphasic delimitation of *S. robusta*, in almost all cases (dataset consisting of a single species), but overestimated *T. fasciculata* in all analyses (Table 2). In fact, all

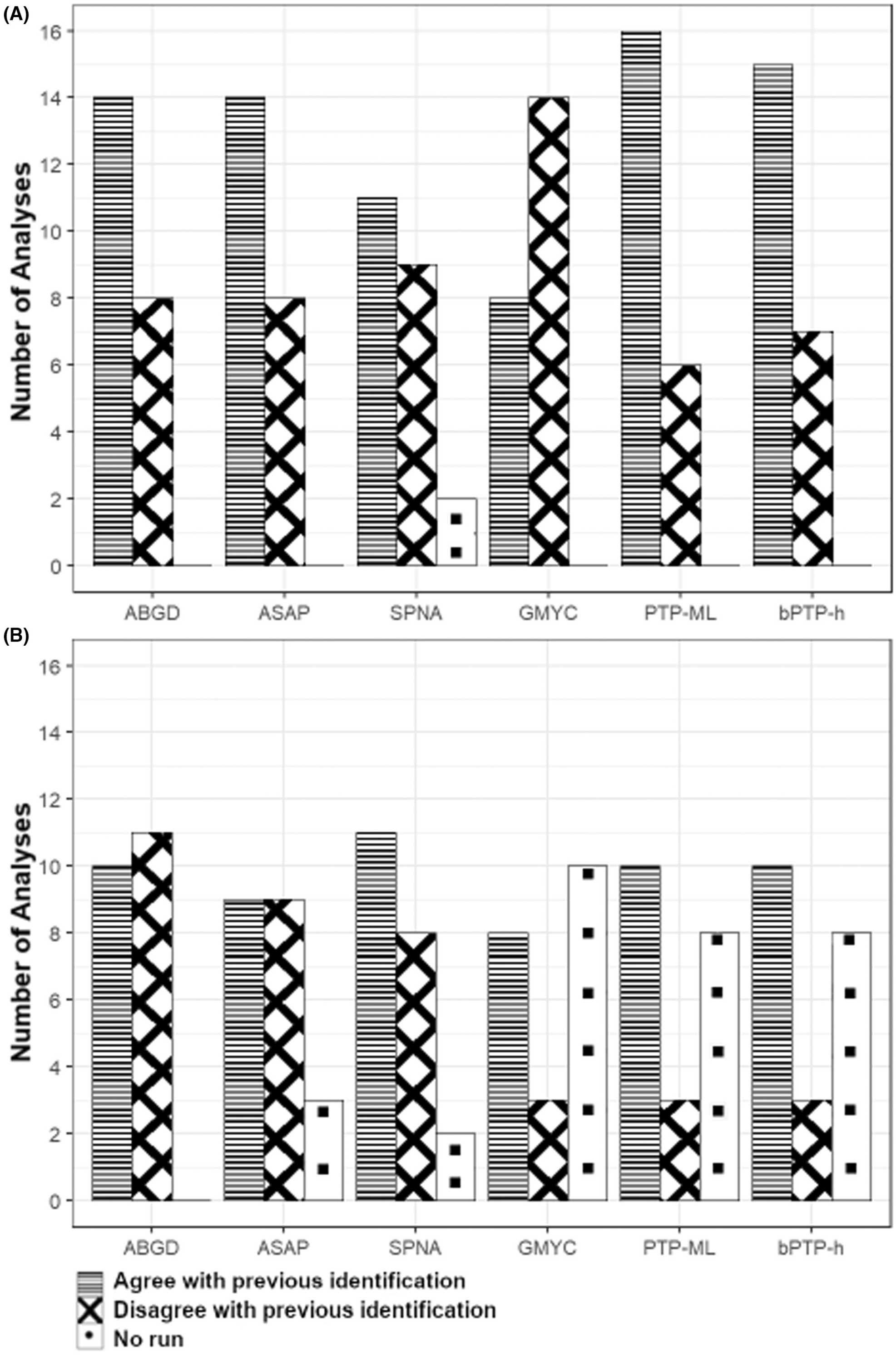

FIGURE 1 Number of analysis carried out with ABDG, ASAP, SPNA, GMYC, PTP-ML and bPTP-h models that agree with and disagree with polyphasic identification and that were not able to run (No run). Polyphasic identification according to Vanormelingen et al. (2008), Trobajo et al. (2009), Amato et al. (2007), Quijano-Scheggia et al. (2009), Vanormelingen et al. (2013), Behnke et al. (2004), De Decker et al. (2018) and Kaczmarska et al. (2009). (A) Balanced datasets – the same number of sequences per target clade. (B) Dataset composed only of haplotypes.

five studied models indicated that the sequences of the reproductively compatible *T. fasciculata* (Table 1) were two distinct species (Table 3).

When ABGD, SPNA, ASAP or PTP species delimitation did not agree with the previously published identification, the same result was generally observed in the short and large DNA fragments from the same DNA marker (*rbc*L, *5.8S+ITS2*, *ITS*1*+5.8S+ITS2* or *cox*1) and group of sequences (Tables 2 and 3). In these cases, the size of the molecular marker had little effect on the result.

DISCUSSION

The use of genetic divergence threshold for species identification

For the delimitation of closely related species, the intraspecific divergence is more useful as a threshold than interspecific divergence, since the latter varies according to the phylogenetic relationships between species. The values of intraspecific (0%–1%) and interspecific (1%–2%) divergences obtained for *rbc*L in the present study correspond close to the cut-off values proposed by MacGillivary and Kaczmarska (2011) for the class Bacillariophyceae, based on the classical morphological criteria (Table 1). The intraspecific divergence of *cox*1 (0%) corresponded to the previously observed values (Moniz & Kaczmarska, 2009), while the interspecific values of *cox*1 (4%–6%) were lower than found by Moniz and Kaczmarska (2009), probably, due to differences in the number of sequences and species analyzed. For the *5.8S+ITS2* marker, the intraspecific (0%–3%) and interspecific (3%–12%) divergence values showed a higher variation than other markers; however, these ranges may vary even more with the inclusion of more diatom species (Moniz & Kaczmarska, 2010) (Table 1). We should highlight, of course, that the genetic distance of *5.8S+ITS*2 from *T. fasciculata*, *S. pupula* complex, *E. bilunaris* and some species of both *Pseudo-nitzschia* complexes, as well as the genetic divergence of *rbc*L from *T. fasciculata* and *N. palea* were previously calculated (MacGillivary & Kaczmarska, 2011; Moniz & Kaczmarska, 2009, 2010); therefore, in these cases, the genetic divergence calculated here (Table 1) is not independent from these cited studies.

The values of intra- or interspecific divergence can present discrepancies from those shown in previous studies due to differences in the number of sequences and species analyzed. For example, the interspecific (but intrageneric) divergence value for taxa in the Bacillariophyceae using the *5.8S+ITS2* marker was 11%–23% based on nine sequences and three species (Moniz & Kaczmarska, 2009), but presented a larger range (1%–48%) in other analyses, with higher number of sequences (316 pairwise comparisons) and species (45) (Moniz & Kaczmarska, 2010).

There is little to support the idea of a "universal" genetic divergence threshold for species across diatoms. The *5.8S+ITS*2 showed a high variation of intraspecific divergence between families, mainly of the Bacillariophyceae (Moniz & Kaczmarska, 2010), and the percent of identified species with *5.8S+ITS*2 increased from 59% to 96% (Bacillariophyceae) using family-specific rather than class-specific thresholds (Moniz & Kaczmarska, 2010). Intraspecific divergence of *rbc*L in the Bacillariophyceae varies from 0% to 11% (237 pairwise comparisons and 39 species). However, 9 of 12 studied families showed an intraspecific mean ≤ 1% (MacGillivary & Kaczmarska, 2011). For this reason, MacGillivary and Kaczmarska (2011) suggested 2% divergence as the threshold for separate species within the Bacillariophyceae using the *rbc*L marker. However, this threshold was not able to separate all the studied species. *N. palea* complex and *S. pupula* complex showed interspecific divergence of 1%, and if a specific threshold of Sellaphoraceae (MacGillivary & Kaczmarska, 2011: 0%) were applied to the species of the *S. pupula* complex, the taxon delimitation would agree with the previous identification of Vanormelingen et al. (2013). Regarding *cox*1, the seven diatom orders that were examined by Moniz and Kaczmarska (2009), each showed a different intraspecific mean value, which can be used as specific order thresholds. However, the information for this marker is the least available than other markers, probably associated with low amplification and sequencing success rates (Moniz & Kaczmarska, 2009). There appears to be a general pattern of intra- and interspecific genetic divergence for each DNA marker (*cox*1, *rbcL* or *5.8S+ITS2*), but the values may vary among the family or order and probably among the genus of the same family. Therefore, the use of genetic divergence depends on reference data specific to the groups of species studied to inform a reliable threshold. It was shown that speciation and extinction rates vary across different groups of diatoms, based on metagenomic data from the global oceans of the 20 most abundant genera (Nakov et al., 2018a), and sequences from 1151 taxa, including bipolar diatoms with and without raphe (Nakov et al., 2018b). In addition, the diversification rates also vary within a diatom genus with lots of species, such as estimated in *Pinnularia* (Pinseel et al., 2020). These differences will affect genetic divergence, the rate of branching and the branch lengths in the phylogenetic tree, hampering the species identification of datasets that gather simultaneously phylogenetically distant diatoms, using threshold or automated molecular species delimitation methods.

Using ABGD, ASAP, SPNA, GMYC and PTP to identify reproductively compatible diatoms

The ASAP, PTP and GMYC models did not delimit diatom species in agreement with reproductive isolation studies when the dataset consisted of only one species (Table 2: *S. robusta* or *T. fasciculata*), which corroborates previous studies, e.g., conducted by Luo et al. (2018) using PTP and GMYC and simulated dataset. In these cases, the ASAP model is likely to overestimate the species due to a lack of information on intra- and interspecific levels of divergence. The PTP and GMYC models estimate species based on the significant changes in the pace of branching events on the tree (Fujisawa & Barraclough, 2013; Pons et al., 2006; Zhang et al., 2013). Therefore, the overestimation associated with the absence of distinct species in the input tree of GMYC and PTP analyses may be due to the fact that the dataset does not include reference to distinguish intra- and interspecific levels of variability, and the significant changes in the tree among populations are interpreted as species.

The ABGD and SPNA models identified *S. robusta* (in a dataset consisting of a single species) corresponding to previous polyphasic identification, but overestimated *T. fasciculata*. This overestimation of *T. fasciculata* species (Tables 2 and 3) by the five models could be explained by the measured lower diversification rate and branch lengths in bipolar araphid diatoms than bipolar raphid pennate ones (Nakov et al., 2018a, 2018b). According to these results, we would recommend that sequences from other species of the same genus or closely related species be included in alignments or trees to obtain more reliable species delimitation by all studied methods (Tables 2 and 3).

In many cases, the delimitation of species from alignments with two or more species using ASAP, ABGD, SPNA and PTP agreed with the previous studies of polyphasic taxonomy, which included reproductive compatibility tests through breeding experiments (Tables 2 and 3). Our results and those of previous studies indicate that GMYC is outperformed by PTP using datasets with few species (Luo et al., 2018) or with closely related species with relatively small genetic distance among them (Zhang et al., 2013). Therefore, we would propose that GMYC is not particularly useful for datasets with relatively few and closely related diatom species, such as our dataset. This model presented the worst performance among studied methods. However, GMYC performed better when using datasets comprised only of haplotypes than datasets with identical sequences.

It is recommended that the input alignments of the ASAP and ABGD analyses be balanced, i.e., the studied clades containing the candidate species or populations should be represented by the same or similar number of sequences.

The sequences must be from different individuals, but they can be the same haplotype. This was inferred from the ASAP and ABGD models, which agreed with the previously published polyphasic delimitation more frequently when the datasets were "balanced" than in datasets with only one haplotype (Figure 1A,B). This happens probably because these models are based on genetic divergence and the presence of sequences from different individuals with the same haplotype represents their real genetic variability.

The ABGD model can underestimate species with less than three to five sequences (Puillandre et al., 2012), explaining the higher number of misleading species delimitation by ABDG than in ASAP, SPNA and PTP in alignments composed of species that had one to two sequences (Tables 2 and 3).

When the results of the automated single-locus computational methods differed from the previous polyphasic identification, ABGD tended to underestimate the number of species, while ASAP, GMYC and PTP tended to overestimate this number. The SPNA presented both types of disagreement in the species delimitation. In most cases, at least three models agreed with previous polyphasic species identification (Tables 2 and 3). Therefore, the comparison of distinct automated single-locus computational methods improved the species delimitation, such as the identification of species belonging to the *P. subgibba* complex, *P. borealis* complex and *A. minutissimum* complex (Kollár et al., 2019; Pinseel et al., 2020; Rimet et al., 2023).

The ABGD, ASAP, SPNA and PTP analyses appear to have great potential for molecular diatom taxonomy, even using relatively small datasets (6–12 sequences; 2–3 species) using markers with a wide size range (~300–1000 bp). Most importantly, the advantage of these tools is associated with the fact that they do not depend on specific thresholds, such as genetic divergence. Therefore, these tools can be applied to diatom identification using sequence data generated for other research purposes, such as barcode genes (short DNA fragments) used in environmental DNA surveys or the larger DNA fragments and markers used for phylogenetic and molecular systematic studies.

Due to the paucity of sexual reproduction data available from most diatom species, our datasets were composed of a few species and only four genera (six species complexes), which had more than one species. To improve our knowledge on the accuracy of ABGD, ASAP, SPNA, GMYC and PTP identifications, further studies are needed using datasets with more geographically diverse strains of *Eunotia*, *Seminavis, Nitzschia*, *Sellaphora*, *Tabularia, Pseudo-nitzschia* and species from other diatom families.

CONCLUSION

Genetic divergence, ABGD, ASAP, SPNA and PTP methods delimited species of *Eunotia*, *Seminavis, Nitzschia, Sellaphora* and *Pseudo-nitzschia*, correlating to polyphasic taxonomy, including sexual reproductive compatibility. In summary, it is recommended to form a hypothesis before using automated molecular delimitation methods to address taxonomy questions with relatively few sequences or species, as our dataset. First, a molecular phylogeny based on all available sequence data should be constructed and considered together with other available information (e.g. morphology, ecological preferences and/or sexual reproductive data) for the target clades (tested species or populations) to be defined. For reliable species delimitation using the ABGD, ASAP, SPNA, GMYC and PTP models, the alignments and trees used as input should include sequences of at least one closely related species to the studied clade or taxa of interest, preferably from the same genus. For the ABGD and ASAP models, the input data should include alignments without a major discrepancy in the number of sequences among the studied clades (tested species or populations), and only of sequences from different individuals; however, distinct individuals with the same haplotype need not be removed. The ABGD also requires that each target clade comprises at least three sequences to test the hypothesis if the clade is a species. In the case of SPNA, the software will remove identical sequences and build the network with haplotypes only. For the PTP model, the tree built with the same alignment as in the ABGD and ASAP analyses can be used (with a balanced dataset – the same or similar number of sequences per studied clade), or the tree built with the alignment containing only haplotypes. It can be useful to compare the results of the PTP analysis using both types of datasets. The GMYC model is not recommended for small datasets, and when this model is used, it requires an input tree built only with haplotypes and should include sequences of at least one closely related species to the target clades. We highlight the great potential of automated methods and that they should be used to augment, but not replace traditional taxonomy approaches. Also, we would like to encourage the diatom community to apply a wide range of models/methods and not just simply one model/method when using molecular data to automate species delimitation and define a consensus among them to improve the species delineation.

ACKNOWLEDGMENTS

This work was supported by Brazilian funding agencies: CNPq Proc. 203883/2017-9; Brazilian Long Term Ecological Research: CNPq/CAPES/FAPERGS/BC – Fundo Newton/PELD n° 15/2016.

DATA AVAILABILITY STATEMENT

Zip file containing all alignments (.FAS or .NEXUS) and phylogenetic trees (.TRE or .NEXUS or .NEWICK) used as input in the ABDG, ASAP, SPNA, GMYC and PTP analyses and an example of a phylogenetic tree figure showing the studied individuals/sequences and the results of the species delimitation models were deposited in the repository FigShare with the doi: 10.6084/m9.figshare.23069618.

ORCID

Andréa de O. da R. Franco https://orcid.org/0000-0002-6656-8358
Matt P. Ashworth https://orcid.org/0000-0002-4162-2004
Clarisse Odebrecht https://orcid.org/0000-0001-7159-4713

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Allewaert, C.C., Vanormelingen, P., Pröschold, T., Gómez, P.I., González, M.A., Bilcke, G. et al. (2015) Species diversity in European *Haematococcus pluvialis* (Chlorophyceae, Volvocales). *Phycologia*, 54(6), 583–598.

Amato, A., Kooistra, W.H.C.F., Ghiron, J.H.L., Mann, D.G., Pröschold, T. & Montresor, M. (2007) Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist*, 158, 193–207.

Annenkova, N.V., Hansen, G. & Rengefors, K. (2020) Closely related dinoflagellate species in vastly different habitats – an example of a marine–freshwater transition. *European Journal of Phycology*, 55(4), 478–489.

Ashworth, M.P., Majewska, R., Frankovich, T.A., Sullivan, M., Bosak, S., Filek, K. et al. (2022) Cultivating epizoic diatoms provides insights into the evolution and ecology of both epibionts and hosts. *Scientific Reports*, 12, 15116.

Behnke, A., Friedl, T., Chepurnov, V.A. & Mann, D.G. (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupul*a species complex (Bacillariophyta). *Journal of Phycology*, 40(1), 193–208.

Chepurnov, V.A., Mann, D.G., Sabbe, K. & Vyverman, W. (2004) Experimental studies on sexual reproduction in diatoms. *International Review of Cytology*, 237, 91–154.

Clement, M., Posada, D. & Crandall, K.A. (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, 9(10), 1657–1659.

Coleman, A.W. (2009) Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. *Molecular Phylogenetics and Evolution*, 50, 197–203.

Correa, A.M.S. & Baker, A.C. (2009) Understanding diversity in coral-algal symbiosis: a cluster-based approach to interpreting fine-scale genetic variation in the genus *Symbiodinium*. *Coral Reefs*, 28, 81–93.

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9, 77.

de Decker, S., Vanormelingen, P., Pinseel, E., Sefbom, J., Audoor, S., Sabbe, K. et al. (2018) Incomplete reproductive isolation between genetically distinct sympatric clades of the pennate model diatom *Seminavis robusta*. *Protist*, 169, 569–583.

de Queiroz, K. (2007) Species concepts and species delimitation. *Systematic Biology*, 56, 879–886.

Evans, M.K., Wortley, A.H. & Mann, G.D. (2007) An assessment of potential diatom "barcode" genes (*cox*1, *rbc*L, 18S and *ITS* rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist*, 158, 349–364.

Ezard, T., Fujisawa, T. & Barraclough, T.G. (2009) SPLITS: SPecies' LImits by threshold statistics. R package version 1.0-18/r45. Available: http://R-Forge.R-project.org/projects/splits/ (last accessed March 2023).

Franco, A.O.R., They, N.H., Canani, L.G.C., Maggioni, R. & Odebrecht, C. (2016) *Asterionellopsis tropicalis* (Bacillariophyceae): a new tropical species found in diatom accumulations. *Journal of Phycology*, 52, 888–895.

Fujisawa, T. & Barraclough, T.G. (2013) Delimiting species using single-locus data and the generalized mixed yule coalescent

(GMYC) approach: a revised method and evaluation on simulated datasets. *Systematic Biology*, 62, 707–724.

Hamsher, S.E., Evans, K.M., Mann, D.G., Poulíčková, A. & Saunders, G.W. (2011) Barcoding diatoms: exploring alternatives to COI-5P. *Protist*, 162, 405–422.

Hehenberger, E., Burki, F., Kolisko, M. & Keeling, P.J. (2016) Functional relationship between a dinoflagellate host and its diatom endosymbiont. *Molecular Biology and Evolution*, 33(9), 2376–2390.

Hoef-Emden, K. (2012) Pitfalls of establishing DNA barcoding systems in protists: the Cryptophyceae as a test case. *PLoS One*, 7(8), e43652.

Hsieh, C.J., Zhan, S.H., Lin, Y., Tang, S.L. & Liu, S.L. (2015) Analysis of *rbc*L sequences reveals the global biodiversity, community structure, and biogeographical pattern of thermoacidophilic red algae (Cyanidiales). *Journal of Phycology*, 51(6), 682–694.

Kaczmarska, I., Ehrman, J.M., Moniz, M.B.J. & Davidovich, N. (2009) Phenotypic and genetic structure of interbreeding populations of the diatom *Tabularia fasciculata* (Bacillariophyta). *Phycologia*, 48, 391–403.

Kaczmarska, I., Mather, L., Luddington, I.A., Muise, F. & Ehrman, J.M. (2014) Cryptic diversity in a cosmopolitan diatom known as *Asterionellopsis glacialis* (Fragilariaceae): implications for ecology, biogeography, and taxonomy. *American Journal of Botany*, 101, 267–286.

Kass, R.E. & Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F. & Bouchez, A. (2013) Nextgeneration sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, 13, 607–619.

Kollár, J., Pinseel, E., Vanormelingen, P., Poulíčková, A., Souffreau, C., Dvořák, P. et al. (2019) A polyphasic approach to the delimitation of diatom species: a case study for the genus *Pinnularia* (Bacillariophyta). *Journal of Phycology*, 55, 365–379.

Kulakova, N.V., Kashin, S.A. & Bukin, Y.S. (2020) The genetic diversity and phylogeny of green microalgae in the genus *Choricystis* (Trebouxiophyceae, Chlorophyta) in Lake Baikal. *Limnology*, 21, 15–24.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J.M., Zuccarello, G.C. et al. (2014) DNA-based species delimitation in algae. *European Journal of Phycology*, 49(2), 179–196.

Luo, A., Ling, C., Ho, S.Y.W. & Zhu, C.D. (2018) Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Systematic Biology*, 67, 830–846.

MacGillivary, M.L. & Kaczmarska, I. (2011) Survey of the efficacy of a short fragment of the *rbc*L gene as a supplemental DNA barcode for diatoms. *Journal of Eukaryotic Microbiology*, 58, 529–536.

Malavasi, V., Škaloud, P., Rindi, F., Tempesta, S., Paoletti, M. & Pasqualetti, M. (2016) DNA-based taxonomy in ecologically versatile microalgae: a Re-evaluation of the species concept within the coccoid green algal genus *Coccomyxa* (Trebouxiophyceae, Chlorophyta). *PLoS One*, 11(3), e0151137.

Mann, D.G. (1999) The species concept in diatoms. *Phycologia*, 38(6), 437–495.

Mann, D.G., Sato, S., Trobajo, R., Vanormelingen, P. & Souffreau, C. (2010) DNA barcoding for species identification and discovery in diatoms. *Cryptogamie Algologie*, 31, 557–577.

Mann, D.G., Thomas, S.J. & Evans, K.M. (2008) Revision of the diatom genus *Sellaphora*: a first account of the larger species in the British Isles. *Fottea*, 9, 15–78.

Mann, D.G. & Vanormelingen, P. (2013) An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, 60, 414–420.

Moniz, M.B.J. & Kaczmarska, I. (2009) Barcoding diatoms: is there a good marker? *Molecular Ecology Resources*, 9, 65–74.

Moniz, M.B.J. & Kaczmarska, I. (2010) Barcoding of diatoms: nuclear encoded *ITS* revisited. *Protist*, 161, 7–34.

Nakov, T., Beaulieu, J.M. & Alverson, A.J. (2018a) Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (diatoms, Bacillariophyta). *New Phytologist*, 219(1), 462–473.

Nakov, T., Beaulieu, J.M. & Alverson, A.J. (2018b) Insights into global planktonic diatom diversity: the importance of comparisons between phylogenetically equivalent units that account for time. *The ISME Journal*, 12, 2807–2810.

Nei, M. & Kumar, S. (2000) *Molecular evolution and phylogenetics*. New York. p: Oxford University Press, p. 333.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289–290.

Pinseel, E., Janssens, S.B., Verleyen, E., Vanormelingen, P., Kohler, T.J., Biersma, E.M. et al. (2020) Global radiation in a rare biosphere soil diatom. *Nature Communications*, 11, 2382.

Pons, J., Barraclought, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S. et al. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55, 595–609.

Poulíčková, A., Kollár, J., Hašler, P., Dvořák, P. & Mann, D.G. (2018) A new species *Pinnularia lacustrigibba* sp. nov. within the *Pinnularia subgibba* group (Bacillariophyceae). *Diatom Research*, 33, 273–282.

Poulíčková, A. & Mann, D.G. (2019) Diatom sexual reproduction and life ycles. In: Seckbach, J. & Gordon, R. (Eds.) *Diatoms: fundamentals and applications*. Beverly: Scrivener Publishing LLC, pp. 245–272.

Poulíčková, A., Mayama, S., Chepurnov, V.A. & Mann, D.G. (2007) Heterothallic auxosporulation, incunabula and perizonium in *Pinnularia* (Bacillariophyceae). *European Journal of Phycology*, 42, 367–390.

Poulíčková, A., Veselá, J., Neustupa, J. & Škaloud, P. (2010) Pseudocryptic diversity versus cosmopolitanism in diatoms: a case study on *Navicula cryptocephala* Kütz (Bacillariophyceae) and morphologically similar taxa. *Protist*, 161, 353–369.

Puillandre, N., Brouillet, S. & Achaz, G. (2021) ASAP: assemble species by automatic partitioning. *Molecular Ecology Resources*, 21(2), 609–620.

Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, 21, 1864–1877.

Quijano-Scheggia, S.I., Garcés, E., Lundholm, N., Moestrup, Ø., Andree, K. & Camp, J. (2009) Morphology, physiology, molecular phylogeny and sexual compatibility of the cryptic *Pseudo-nitzschia delicatissima* complex (Bacillariophyta), including the description of *P. arenysensis* sp. nov. *Phycologia*, 48, 492–509.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. (2018) Posterior summarisation in Bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67, 901–904.

Rimet, F., Pinseel, E., Bouchez, A., Japoshvili, B. & Levan Mumladze, L. (2023) Diatom endemism and taxonomic turnover: assessment in high-altitude alpine lakes covering a large geographical range. *Science of the Total Environment*, 871, 161970.

Rimet, F., Trobajo, R., Mann, D.G., Kermarrec, L., Franc, A., Domaizon, I. et al. (2014) When is sampling complete? The effects of geographical range and marker choice on perceived diversity in *Nitzschia palea* (Bacillariophyta). *Protist*, 65(3), 245–259.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S. et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), 539–542.

Round, F.E., Crawford, R.M. & Mann, D.G. (1990) *The diatoms*. Biology and Morphology of the Genera: Cambridge University Press, Cambridge, p. 747.

Stock, W., Vanelslander, B., Rüdiger, F., Sabbe, K., Vyverman, W. & Karsten, U. (2019) Thermal niche differentiation in the benthic diatom *Cylindrotheca closterium* (Bacillariophyceae) complex. *Frontiers Microbiology*, 10, 1395.

Tamura, K., Steche, G., Peterson, D., Filipski, A. & Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30, 2725–2729.

Templeton A.R., Crandall K.A. & Sing C.F. 1992.A Cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 132: 619–633.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25, 4876–4882.

Trobajo, R., Clavero, E., Chepurnov, V.A., Sabbe, K., Mann, D.G., Ishihara, S. et al. (2009) Morphological, genetic and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyceae). *Phycologia*, 48, 443–459.

Trobajo, R., Mann, D.G., Clavero, E., Evans, K.M., Vanormelingen, P. & McGregor, R.C. (2010) The use of partial cox1, *rbc*L and *LSU* rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). *European Journal of Phycology*, 45, 413–425.

Vanormelingen, P., Chepurnov, V.A., Mann, D.G., 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.

Vanormelingen, P., Evans, K., Chepurnov, V.A., Vyverman, W. & Mann, D.G. (2013) Molecular species discovery in the diatom *Sellaphora* and its congruence with mating trials. *Journal of the Czech Phycological Society*, 13(2), 133–148.

Vasselon, V., Rimet, F., Tapolczai, K. & Bouchez, A. (2017) Assessing ecological status with diatoms DNA metabarcoding: scaling-up a WFD monitoring network (Mayotte Island, France). *Ecological Indicators*, 82, 1–12.

Zhang, J., Kapli, P., Pavlidis, P. & Stamatakis, A. (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29(22), 2869–2876.

Zou, S., Fei, C., Song, J., Bao, Y., He, M. & Wang, C. (2016) Combining and comparing coalescent, distance and character-based approaches for barcoding microalgae: a test with *chlorella*-like species (Chlorophyta). *PLoS One*, 11(4), e0153833.

Zou, S., Fei, C., Wang, C., Gao, Z., Bao, Y., He, M. et al. (2016) How DNA barcoding can be more effective in microalgae identification: a case of cryptic diversity revelation in *Scenedesmus* (Chlorophyceae). *Scientific Reports*, 6, 36822.

How to cite this article: Franco, A.d.O.d.R., Ashworth, M.P. & Odebrecht, C. (2023) Comparison between p-distance and single-locus species delimitation models for delineating reproductively tested strains of pennate diatoms (Bacillariophyceae) using *cox*1, *rbc*L and *ITS*. *Journal of Eukaryotic Microbiology*, 70, e12986. Available from: https://doi.org/10.1111/jeu.12986