


# DINOREF: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene

Solenn Mordret<sup>1</sup> | Roberta Piredda<sup>1</sup> | Daniel Vaultot<sup>2</sup> | Marina Montresor<sup>1</sup> |  
Wiebe H. C. F. Kooistra<sup>1</sup> | Diana Sarno<sup>1</sup> 

<sup>1</sup>Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Naples, Italy

<sup>2</sup>Sorbonne Université, CNRS, UMR Adaptation et Diversité en Milieu Marin, Station Biologique, Roscoff, France

## Correspondence

Diana Sarno, Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Naples, Italy.  
Email: diana.sarno@szn.it

## Funding information

Italian Ministry of Education, University and Research

## Abstract

Dinoflagellates are a heterogeneous group of protists present in all aquatic ecosystems where they occupy various ecological niches. They play a major role as primary producers, but many species are mixotrophic or heterotrophic. Environmental metabarcoding based on high-throughput sequencing is increasingly applied to assess diversity and abundance of planktonic organisms, and reference databases are definitely needed to taxonomically assign the huge number of sequences. We provide an updated 18S rRNA reference database of dinoflagellates: DINOREF. Sequences were downloaded from GENBANK and filtered based on stringent quality criteria. All sequences were taxonomically curated, classified taking into account classical morphotaxonomic studies and molecular phylogenies, and linked to a series of metadata. DINOREF includes 1,671 sequences representing 149 genera and 422 species. The taxonomic assignment of 468 sequences was revised. The largest number of sequences belongs to Gonyaulacales and Suessiales that include toxic and symbiotic species. DINOREF provides an opportunity to test the level of taxonomic resolution of different 18S barcode markers based on a large number of sequences and species. As an example, when only the V4 region is considered, 374 of the 422 species included in DINOREF can still be unambiguously identified. Clustering the V4 sequences at 98% similarity, a threshold that is commonly applied in metabarcoding studies, resulted in a considerable underestimation of species diversity.

## KEYWORDS

18S rRNA gene, dinoflagellates, diversity, phylogeny, sequence database, V4 region

## 1 | INTRODUCTION

Assessing global biodiversity constitutes an important and urgent task in the face of the currently unprecedented rate of climate change, but this task is fraught with major challenges. A large part of this biodiversity is composed of protists, and it is especially in these unicellular eukaryotes that taxonomists are confronted by the fact that cell morphology does not always allow discrimination of species, especially in small and relatively featureless taxa. When compared with morphological traits, sequence data usually provide more precise and apparently more objective ways to delineate species and therefore more

precise ways to enumerate them. DNA-based detection and enumeration methodologies, such as high-throughput sequencing (HTS) metabarcoding of environmental samples, now offer opportunities for assessing protistan diversity rapidly and precisely (Amaral-Zettler, McCliment, Ducklow, & Huse, 2009; Massana et al., 2015; Piredda et al., 2017; Stoeck et al., 2009; de Vargas et al., 2015). Yet, to translate these HTS data into species occurrences requires a comprehensive reference database. Curated databases of reference sequences linked to taxonomically identified specimens constitute important research infrastructures for the advancement of our knowledge of the protistan diversity (Decelle et al., 2015; Morard et al., 2015).

Dinoflagellates form a large phylum of protists distributed in various aquatic ecosystems (Hackett, Anderson, Erdner, & Bhattacharya, 2004; Not et al., 2012). The lineage includes autotrophs, heterotrophs and mixotrophs; most species are free-living, but parasites and symbionts are abundant as well, adding complexity to aquatic ecological networks (Gómez, 2012b; Jephcott et al., 2016; Stoecker, 1999; Weisse et al., 2016). Dinoflagellates are of socio-economic relevance because several species produce secondary metabolites, some of which are highly toxic to humans and marine organisms (Anderson, Cembella, & Hallegraeff, 2012; Berdalet et al., 2016). Therefore, dinoflagellates are monitored on a regular basis in many coastal regions. The assessment of dinoflagellate diversity is still largely based on light microscopy observations (LM). Yet, this approach has its limitations: it requires a high level of taxonomic expertise and is time-consuming, and many species, including minute, naked and parasitic dinoflagellates, are virtually impossible to identify. In some cases, toxic species are difficult to distinguish from nontoxic relatives in LM (de Salas, Laza-Martínez, & Hallegraeff, 2008; Montresor, John, Beran, & Medlin, 2004).

Despite its rather low variability (Murray, Flø Jørgensen, Ho, Patterson, & Jermini, 2005), the 18S ribosomal subunit (18S) is potentially a good DNA barcode region for dinoflagellates because of the large number of sequences deposited in public repositories as a result of its popular use in taxonomic and phylogenetic studies (Gómez, 2014; John et al., 2014), as well as in single-cell PCR amplification studies (e.g., Gómez, Moreira, & López-García, 2010; Hoppenrath, Murray, Sparmann, & Leander, 2012; Ki, Jang, & Han, 2005; Ruiz Sebastián & O'Ryan, 2001). The variable V4 or V9 regions in the 18S rRNA-encoding region are the most commonly used nucleotide markers in metabarcoding studies on environmental samples (e.g., Le Bescot et al., 2016; Onda et al., 2017).

To improve taxonomic annotation of environmental sequences generated by HTS approaches, curated reference barcode databases are needed. PR<sup>2</sup> was the first database that became available for protists (Guillou et al., 2013), which included 136,866 nuclear-encoded sequences that were taxonomically assigned to eight taxonomic fields. Few other specialized databases have been created to provide taxonomically validated sequences with updated nomenclature and contextual metadata for different groups of organisms (e.g., Decelle et al., 2015 for 16S of photosynthetic eukaryotes; Morard et al., 2015 for foraminifera).

The aim of this study was to provide a taxonomically curated database, called DINOREF, composed of the "core dinoflagellates," that is, the species with a dinokaryon, a modified nucleus containing permanently condensed fibrillar chromosomes, as defined by Orr, Murray, Stüken, Rhodes, and Jakobsen (2012). To populate the database, we gathered all dinoflagellate 18S rRNA sequences available in GENBANK, screened them against a set of quality criteria and verified their taxonomic assignments by means of phylogenetic positioning. For each sequence, taxonomy and nomenclature from GENBANK were updated, deploying a phylogenetic approach, and taking into account the most recent literature and taxonomy databases. We organized the 18S sequences into the same suprageneric ranks as those used in PR<sup>2</sup> (Guillou et al., 2013). Given the "dynamic status" of the

dinoflagellate classification at the suprageneric level, we also present an alternative organization of the sequence data, namely one based on results of 18S phylogenetic inferences integrated with the latest literature sources. Finally, we illustrated the degree to which the 18S V4 region, which is broadly used in metabarcoding surveys and has been proposed as the universal eukaryotic prebarcode (Hu et al., 2015; Pawlowski et al., 2012), is able to discriminate species.

## 2 | MATERIALS AND METHODS

### 2.1 | Sequence retrieval

Dinoflagellate 18S rRNA entries available on the 29th of August 2016 were downloaded from NCBI GENBANK (<https://www.ncbi.nlm.nih.gov/>) using the following text query: Dinophyceae[Organism] AND (small subunit ribosomal[titl] OR 18S[titl] OR SSU[titl]). The features associated with the GENBANK entries were extracted. Sequences of early branching dinoflagellates, not classified as Dinophyceae (sensu Orr et al. (2012)), were recovered genus by genus and were excluded from DINOREF, but provided as supplementary material (see below).

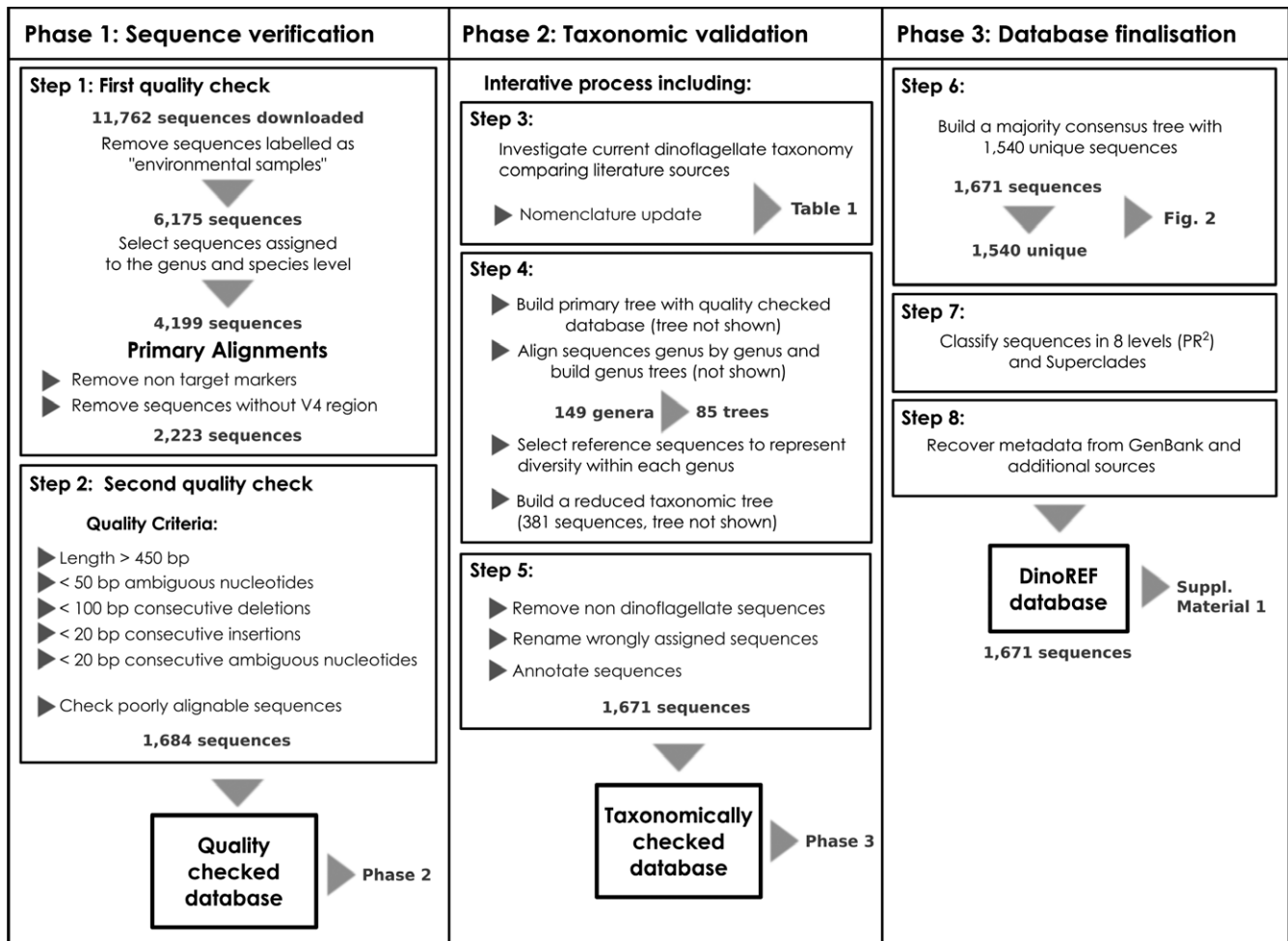
### 2.2 | Sequence verification

Sequences labelled as "environmental samples" in GENBANK (i.e., from metabarcoding, clone libraries, uncultured organisms—labelled as ENV in the GENBANK locus field) and those classified above the genus level were removed. Retained sequences were aligned with MAFFT version 7 (default parameters, Katoh & Standley, 2013) and manually checked. Sequences not meeting the following quality criteria were removed (Figure 1, step 1): (i) sequences deposited as 18S but not representing 18S at all or only the very 3'-end of it; and (ii) 18S sequences not including the full V4 region. Of the remaining sequences, the regions outside the 18S rRNA gene were removed. After a second filtration step (Figure 1, step 2), only those sequences that fulfilled all of the following criteria were retained: (i) sequence length  $\geq 450$  bp; (ii)  $< 50$  ambiguous nucleotides; (iii)  $< 20$  consecutive ambiguous bp; (iv) insertion  $< 20$  consecutive bp; and v. deletions  $< 100$  consecutive bp. Sequences that aligned poorly or not at all with any of the other sequences were blasted against GENBANK and were removed if BLAST results suggested placement outside dinoflagellates (i.e., BLAST assignment  $< 90\%$ ). The retained set of sequences constituted the "quality-checked database" (Figure 1).

### 2.3 | Taxonomic validation

The taxonomic validation was an iterative process including the following steps:

1. control on the validity of the nomenclature, based on Fensome et al. (1993) and Gómez (2012a); ALGAEBASE (Guiry & Guiry, 2017; <http://www.algaebase.org>), CEDIT (Hoppenrath & Elbrächter, 2015; <http://www.dinophyta.org>) and information from the literature of specific groups (Figure 1, step 3);



**FIGURE 1** Workflow describing the steps applied to generate the curated and annotated dinoflagellate 18S rRNA Reference Database DINOREF from the sequences downloaded from NCBI

2. phylogenetic evidences based on the primary tree, the genus-level trees and the taxonomic reference tree (Figure 1, step 4).

Species names were validated following taxonomically accepted names in ALGAEBASE (Guiry & Guiry, 2017; names marked as “C”).

Sequences included in the quality-checked database were aligned with MAFFT version 7, and a phylogenetic tree was built using FASTTREE (Price, Dehal, & Arkin, 2010) as implemented in the GENEIOUS software (Kearse et al., 2012). This primary tree provided information on the number, statistical support and position of the different terminal clades (Figure 1, step 4).

In order to identify sequences attributed to the wrong genus in GENBANK, we aligned the sequences labelled with the same genus name using MAFFT and visualized them in Seaview (Gouy et al., 2010). When possible (three or more sequences per genus), a maximum-likelihood tree was inferred using PHYML version 3.0 (Guindon et al., 2010). Branch support was established using 100 bootstrap replicates. Genera represented by less than three sequences were grouped with their closest phylogenetic groups after verifying their positioning in the primary tree. Outlying sequences in the generic trees were sought in the primary tree, and if they were found inside the clade of another genus, they were renamed accordingly.

A number of long sequences ( $\geq 1,700$  bp), representing the diversity within each genus, were selected to build a reduced taxonomic reference tree (Figure 1, step 4). In case the only reference for a given genus was represented by a shorter sequence, its phylogenetic placement was verified building a “rough tree” based on a shorter alignment, but these shorter sequences were excluded from the alignment used to build the reduced taxonomic reference tree. The data set was aligned using MAFFT and the tree built with RAXML (Stamatakis, 2014). Branch support was established using 100 bootstrap replicates. This reduced tree provided a clearer representation of dinoflagellate phylogeny and enabled checks on the phylogenetic relationships for the terminal clades, allowing the removal of any remaining nondinoflagellate sequence from the quality-checked database (Figure 1, step 5). This set of taxonomically curated sequences was called the “taxonomically checked database.”

## 2.4 | Database finalization

From the taxonomically checked database, a final nonredundant sequence alignment (i.e., only containing unique sequences) was produced using MAFFT and a majority-rule consensus tree (RAXML, 100 bootstraps) was built (Figure 1, step 6). Every sequence in the taxonomically

checked database was labelled using a standardized eight term-ranked taxonomy, that is, Kingdom, Supergroup, Division, Class, Order, Family, Genus and Species in the same way as in PR<sup>2</sup> (Guillou et al., 2013). The database includes also an additional classification of the sequences into different "Superclades" defined according to the results of phylogenetic inferences and specialized literature (Figure 1, step 7).

The standard metadata extracted from GENBANK (see Supplementary Material 1 for a complete list) has been supplemented by information on: (i) species habitat (Guiry & Guiry, 2017; Hoppentrath, Murray, Chomérat, & Horiguchi, 2014), (ii) potential toxicity (Moestrup et al., 2009) and (iii) symbiotic or parasitic lifestyle derived from original papers (Figure 1, step 8). The taxonomically checked database integrated with metadata constitutes "DINOREF (dinoflagellate reference database)".

## 2.5 | V4 analysis

To compare the resolution power of the V4 barcode region with that of the full-length 18S gene, the V4 fragments were extracted from the alignment using the V4 primers described in Piredda et al. (2017). Sequences of the V4 region were then dereplicated and split by genus and Superclade using Mothur (Schloss et al., 2009). Each group of sequences was then re-aligned using MAFFT and checked manually. For each Superclade and each genus, distance matrices of pairwise differences between sequences over the length of the V4 (p-distance) were generated using the software MEGA version 7 (Kumar, Stecher, & Tamura, 2016). Boxplots of distances were produced with R (R Development Core Team, 2016) using the "ggplot2" library (Wickham, 2009). V4 OTUs at 98% similarity were generated clustering the sequences with VSEARCH algorithm with "distance-based greedy clustering" (DGC, Rognes, Flouri, Nichols, Quince, & Mahé, 2016) as implemented in Mothur. The V4 unique sequences and the V4 OTUs at 98% similarity are provided as supplementary materials (see below). In those files, we also listed the 18S sequences sharing the same V4 region and the V4 sequences clustering in the same OTU at 98%.

DINOREF has been incorporated into PR<sup>2</sup> version 4.9.0 which is available at <https://doi.org/10.6084/m9.figshare.5913181>.

Supplementary data that include the DINOREF database as an Excel file (Supplementary Material 1), a full-length 18S sequences fasta text file (Supplementary Material 2) associated with three different taxonomy files, (i) the original taxonomy of GENBANK entry (Supplementary Material 3), (ii) the curated taxonomy (Supplementary Material 4) and (iii) the curated taxonomy including the Superclade classification used in this study (Supplementary Material 5), are available in flat file format on Figshare (<https://doi.org/10.6084/m9.figshare.5568454>). The format of the fasta and taxonomy text files is compatible with Mothur (Schloss et al., 2009) and Qiime (Caporaso et al., 2010) tools. Fasta and taxonomy files can be opened with a text editor. We also provide two Excel files with V4 sequences and V4 OTUs at 98% similarity (Supplementary Materials 6 and 7). Supplementary Material 8 includes a fasta file containing all sequences of early branching dinoflagellates recovered from GENBANK but not included in DINOREF.

## 3 | RESULTS

### 3.1 | The DINOREF database

#### 3.1.1 | Phase 1: Sequence verification

A total of 6,175 GENBANK sequences were downloaded with the given search criteria. Upon removal of sequences not assigned at a genus or species level, 4,199 were retained, and of these, 2,223 remained following removal of non-18S sequences as well as 18S sequences in which the V4 region was incomplete or lacking. The verification process resulted in a database of 1,684 aligned, good-quality sequences.

#### 3.1.2 | Phase 2: Taxonomic validation

The primary tree inferred from the quality-checked database revealed that genera constituted the best-supported taxonomic level; that is, they were usually monophyletic with high bootstrap support (data not shown). Three genera were found to be represented by more than 150 sequences each: *Alexandrium* with 210 sequences, *Gambierdiscus* with 169 sequences and *Symbiodinium* with 173 sequences (around 36% of the sequences in the database; Table 2). Some species within these three genera were represented by a large number of different 18S sequences. For instance, 54 slightly different sequences were attributed to *Gambierdiscus scabrosus*. The number of species included in other genera was much lower: a first group of 27 genera included each between 68 and 10 sequences (about 40% of the total number of sequences), a second group of 55 genera contained between nine and three sequences (18% of the total), whereas the remaining 63 genera were represented by one or two sequences only (6% of the total) (Figure 4).

The resulting taxonomically checked database contained 1,671 dinoflagellate sequences, corresponding to 1,540 unique sequences and belonging to 149 genera (Table 1) and 422 species (Table 2). Thirteen of the original 1,684 sequences had to be removed because they did not belong to dinoflagellates. The assignment of 468 sequences (28% of the total database) had to be revised because the names originally assigned to them were synonyms, invalid or the phylogenetic analyses revealed that they were attributed to the wrong taxon. Sequence length ranged from 579 to 1,764 bp (Figure S1), with 1,200 of them being full or nearly full length (between 1,600 and 1,764 bp). The majority of sequences between 1,100 and 1,400 bp originated from single-cell amplifications.

#### 3.1.3 | Phase 3: Database finalization

The curated sequences were organized hierarchically, following the 8-level taxonomic framework used in the PR<sup>2</sup> database (Columns E-L in Supplementary Material 1). For the assignment of the "order" level, we followed a conservative approach accepting the following six orders: Gonyaulacales, Peridinales, Dinophysiales, Prorocentrales, Suessiales and Gymnodinales ("Order" in Table 1; column I in

Supplementary Material 1). Some sequences could not be placed within any of these orders and were listed as Dinophyceae ordo incertae sedis. Assignment at the family level was problematic due to the still unresolved phylogenetic relationship among genera (see Section 4). We generally followed the classification provided by ALGAEBASE (Guiry & Guiry, 2017).

### 3.2 | Superclades: an attempt to depict the current organization of dinoflagellates

A majority-rule consensus tree built with the 1,540 unique sequences provided a phylogenetic representation of the 18S sequences contained in the DINOREF database. In this tree, a large number of clades with  $\geq 50\%$  bootstrap support collapsed onto a polytomy (Figure 2). The well-supported clades of the majority-rule consensus tree (Table 1; Figure 2) were grouped by us into higher taxonomic ranks ("Superclade" in Table 1; Column D in Supplementary Material 1) reflecting the current taxonomic organization based on previously published phylogenies and multigene phylogenies (see selected references in Table 1).

The largest Superclade included species of the order Gonyaulacales (Superclade 1) and was recovered from the tree in five well-supported clades. Within it, clade 1A (Table 1; Figure 2) encompassed a large number of sequences of the genera *Alexandrium* and *Gambierdiscus*, while the other contained either a single genus (clade 1D, *Gonyaulax*) or different but phylogenetically closely related genera (e.g., clade 1B, *Ceratium* and *Triplos*).

Sequences of the order Dinophysiales (Superclade 2) grouped into three clades, while Prorocentrales (Superclade 11) were distributed over five distinct clades.

Sequences attributed to the order Peridinales sensu lato were considerably diversified in the 18S phylogeny, and seven Superclades were identified: Thoracosphaeraceae (7 clades), Peridinales sensu stricto (4 clades), Kryptoperidiniaceae, Heterocapsaceae and Podolampadaceae, each with one clade; one Superclade only included a pair of genera (*Ensiculifera* and *Pentapharsodinium*), and only one included a single genus (*Blastodinium*).

All sequences of the order Suessiales (Superclade 3) grouped in a single clade and sequences of species classified within Gymnodinales sensu lato clustered into six distinct Superclades (Superclades 12–17), of which Superclade 13 included all Gymnodinales sensu stricto. Two Superclades of Gymnodinales sensu lato corresponded to Torodinales and Kareniaceae, respectively, and the other three Superclades contained sequences attributed to a single genus, which are *Amphidinium*, *Akashiwo* and *Gyrodinium*.

Two Superclades, that is, Tovelliaceae and Ptychodisciales, were classified as "Dinophyceae ordo incertae sedis" in ALGAEBASE (Guiry & Guiry, 2017). A series of dinoflagellate sequences, often including taxa of uncertain classification, were resolved in a whole series of small clades, and even single sequences, all emerging from the polytomy. We distributed these sequences into two categories: "Uncertain naked dinophyceae (UND)" and "Uncertain thecate dinophyceae (UTD)".

There was an incomplete and unbalanced 18S sequence representation of described taxa, and the number of available sequences differed considerably among Superclades, clades and genera (Table 2). For example, the order Suessiales included 14 genera represented by 18S sequences and 11 without reference 18S sequence, while Gonyaulacales were reasonably well covered (17 genera with sequences and three without). Overall, only 422 of the 2,342 (18%) taxonomically recognized dinoflagellate species were found to have a reference 18S sequence (Guiry & Guiry, 2017) (Table 2). When considering genera, only 149 of the 232 genera had a reference sequence.

The DINOREF database includes also a broad set of metadata retrieved from GENBANK and/or from the taxonomic literature (Supplementary Material 1). Most sequences (89%) originated from marine ecosystems, 8% from freshwater and 2% from brackish, estuarine or continental saline habitats. Specific lifestyle information (i.e., symbiont or parasite and their host) is available for 302 sequences (18%). In addition, 24% of the sequences were annotated as benthic, and 34% as belonging to potentially toxic species.

### 3.3 | The barcoding V4 region

If only the V4 region is considered, the number of unique DINOREF sequences shrunk from 1,540 to 946 (Supplementary Material 6). This decrease in sequence number was mostly due to slightly different intraspecific 18S sequences sharing identical V4 ribotypes, that is, haplotypes, unique sequences of ribosomal genes. Nonetheless, a large proportion of the intraspecific diversity, observed when the whole 18S was taken into account, was also detectable when only the V4 marker was considered. For example, the single morphologically defined species *Alexandrium fundyense* and *Gambierdiscus scabrosus* were represented by 103 and 54 18S and 36 and 39 V4 sequences, respectively. However, cases occurred in which different species shared the same V4 ribotype; for example, V4 failed to distinguish toxic *Azadinium spinosum* from nontoxic *A. trinitatum*, *Karenia brevis* from *K. mikimotoi* and seven *Dinophysis* species from one another (Supplementary Material 6). Moreover, there are cases in which sequences belonging to closely related genera shared identical V4 regions; as is the case of the V4 #788 shared among *Karlodinium veneficum*, *Takayama pulchellum* and *Takayama acrotrocha*, and the V4 #315 which is identical in several *Histioneis* spp. and *Ornithocercus* spp. (Supplementary Material 6). Overall, 374 of the 422 species included in DINOREF could be identified with the V4 barcode marker.

In general, Superclades with many sequences deriving from several genera showed large pairwise p-distances for the V4 region (Figure 3). Yet patterns differed profoundly among the Superclades. For example, Superclade 1 (Gonyaulacales) that contained 543 sequences from 17 genera and 92 species showed similar p-distance to the smaller Superclade 8 (Peridinales sensu stricto) that had only 116 sequences from 13 genera and 46 species (Table 2; Figure 3). Other Superclades, such as #16 (*Amphidinium*) or #18 (Tovelliaceae) and #19 (*Blastodinium*), showed high levels of variation even though there were only a small number of sequences (Table 2; Figure 3). Superclade 2 (Dinophysiales) showed similar p-distance patterns as



**TABLE 1** Organization of the dinoflagellate 18S sequences into genera, clades and Superclades

Order (AlgaeBase)	Superclade	Clade	Genera or species
Gonyaulacales	1. Gonyaulacales (Adl et al., 2012; Orr et al., 2012)	1A	<i>Alexandrium</i> , <i>Fragilidinium</i> , <i>Coolia</i> , <i>Ostreopsis</i> , <i>Fukuyoa</i> , <i>Gambierdiscus</i> , <i>Goniodoma</i> , <i>Pyrocystis</i> , <i>Pyrodinium</i> , <i>Pyrophacus</i>
		1B	<i>Ceratium</i> , <i>Tripos</i>
		1C	<i>Lingulodinium</i> , <i>Amylax</i> , <i>Gonyaulax verior</i>
		1D	<i>Gonyaulax</i>
		1E	<i>Ceratocorys</i> , <i>Protoceratium</i>
Dinophysiales	2. Dinophysiales (Adl et al., 2012; Hoppenrath, Chomérat, & Leander, 2013; Orr et al., 2012)	2A	<i>Amphisolenia</i> , <i>Dinophysis</i> , <i>Histioneis</i> <i>Ornithocercus</i> , <i>Phalacroma</i> , <i>Triposolenia</i>
		2B	<i>Sinophysis</i>
		2C	<i>Pseudophalacroma</i>
Suessiales	3. Suessiales (Adl et al., 2012; Orr et al., 2012)	3	<i>Ansanella</i> , <i>Asulcocephalium</i> , <i>Baldinia</i> , <i>Biecheleria</i> , <i>Biecheleriopsis</i> , <i>Borghiella</i> , <i>Cystodinium</i> , <i>Leiocephalium</i> , <i>Pelagodinium</i> , <i>Phytodinium</i> , <i>Piscinodinium</i> , <i>Polarella</i> , <i>Protodinium</i> , <i>Symbiodinium</i> , <i>Woloszynskia</i> (pro-parte)
Peridinales	4. Thoracosphaeraceae (Adl et al., 2012; Gottschling et al., 2012)	4A	<i>Amyloodinium</i> , <i>Cryptoperidiniopsis</i> , <i>Paulsenella</i> , <i>Pfiesteria</i>
		4B	<i>Scrippsiella sensu lato</i> : <i>Pernambugia</i> , <i>Duboscquodinium</i> , <i>Naiadinium</i> , <i>Scrippsiella</i> , <i>Theleodinium</i>
		4C	<i>Apocalathium</i>
		4D	<i>Crypthecodinium</i>
		4E	<i>Stoeckeria</i>
		4F	<i>Chimonodinium</i>
		4G	<i>Thoracosphaera</i> <u>Single sequences</u> : <i>Aduncodinium glandula</i> , <i>Tintinnophagus acutus</i>
Dinophyceae ordo incertae sedis	5. Amphidomataceae (Tillmann, Gottschling, Nézan, Krock, & Bilien, 2014)	5A	<i>Azadinium</i>
		5B	<i>Amphidoma</i>
		5C	<i>Azadinium dexteroporum</i>
		5D	<i>Azadinium polongum</i> , <i>Azadinium concinnum</i>
		5E	<i>Azadinium caudatum</i>
Peridinales	6. Kryptoperidiniaceae (Gottschling, Čalasan, Kretschmann, & Gu, 2017; Takano, Hansen, Fujita, & Horiguchi, 2008)	6	<i>Durinskia</i> , <i>Galeidinium</i> , <i>Kryptoperidinium</i> , <i>Unruhadinium</i> , <i>Blixaea</i>
		7	<i>Ensiculifera</i> , <i>Pentapharsodinium</i>
	8. Peridinales sensu stricto (Gu et al., 2015; Mertens et al., 2015)	8A	<u>Clade Monovela</u> : <i>Amphidiniopsis</i> , <i>Archaeperidinium</i> , <i>Herdmania</i> , <i>Islandinium</i> , <i>Protoperidinium americanum</i> , <i>P. fusiforme</i> , <i>P. fukuyoi</i> , <i>P. monovelum</i> , <i>P. parthenopes</i>
		8B	<u>Peridinium clade</u> : <i>Peridinium willei</i> , <i>P. volzii</i> , <i>P. cinctum</i> , <i>P. gatunense</i> , <i>P. bipes</i> , <i>P. limbatum</i>
		8C	<u>Protoperidinium sensu stricto</u> : <i>Protoperidinium abei</i> , <i>P. bipes</i> , <i>P. conicum</i> , <i>P. crassipes</i> , <i>P. divergens</i> , <i>P. denticulatum</i> , <i>P. elegans</i> , <i>P. excentricum</i> , <i>P. leonis</i> , <i>P. pallidum</i> , <i>P. pellucidum</i> , <i>P. pentagonum</i> , <i>P. punctulatum</i> , <i>P. thorianum</i> , <i>P. thulesense</i> , <i>Kolkwitzia</i>
		8D	<u>Diplopsalioideae III and Oceanica clade</u> : <i>Diplopsalopsis</i> , <i>Niea</i> , <i>Qia</i> , <i>Gotoius</i> , <i>Protoperidinium claudicans</i> , <i>Protoperidinium depressum</i> <u>Single sequences</u> : <i>Diplopsalis caspica</i> , <i>D. lenticula</i> , <i>Preperidinium meunieri</i>
	9. Heterocapsaceae (Salas, Tillmann, & Kavanagh, 2014)	9	<i>Heterocapsa</i>
	10. Podolampadaceae (Adl et al., 2012)	10A	<i>Blepharocysta</i> , <i>Podolampas</i> , <i>Roscoffia</i> <u>Single sequence</u> : <i>Lessardia elongata</i>

(Continued)

**TABLE 1** (Continued)

Order (AlgaeBase)	Superclade	Clade	Genera or species
Prorocentrales	11. Prorocentrales (Adl et al., 2012; Orr et al., 2012)	11A	<i>Prorocentrum dentatum</i> , <i>P. donghaiense</i> , <i>P. emarginatum</i> , <i>P. fukuyoi</i> , <i>P. mexicanum</i> , <i>P. micans</i> , <i>P. cordatum</i> , <i>P. rhathymum</i> , <i>P. shikokuense</i> , <i>P. texanum</i> , <i>P. triestinum</i> , <i>P. tsawwassenense</i>
		11B	<i>Prorocentrum hoffmannianum</i> , <i>P. bimaculatum</i> , <i>P. concavum</i> , <i>P. consutum</i> , <i>P. foraminosum</i> , <i>P. maculosum</i> , <i>P. leve</i> , <i>P. lima</i>
		11C	<i>Prorocentrum glenanicum</i> , <i>P. panamense</i> , <i>P. pseudopanamense</i>
		11D	<i>Plagiodinium</i>
		11E	<i>Prorocentrum cassubicum</i>
Gymnodiniales	12. Genus <i>Akashiwo</i> (Orr et al., 2012)	12	<i>Akashiwo</i>
	13. Gymnodiniales sensu stricto (Hoppenrath & Leander, 2007; Reñé, Camp, & Garcés, 2015)	13	<i>Chytriodinium</i> , <i>Dissodinium</i> , <i>Erythrospidinium</i> , <i>Gymnodinium</i> , <i>Gymnoxanthea</i> , <i>Gyrodiniellum</i> , <i>Lepidodinium</i> , <i>Nematodinium</i> , <i>Nusuttodinium</i> , <i>Paragymnodinium</i> , <i>Pellucidodinium</i> , <i>Pheopolykrikos</i> , <i>Polykrikos</i> , <i>Proterythropsis</i> , <i>Spiniferodinium</i> , <i>Warnowia</i>
	14. Kareniaceae (Adl et al., 2012)	14A	<i>Brachidinium</i> , <i>Karenia</i>
		14B	<i>Karlodinium</i> , <i>Takayama</i>
	15. Genus <i>Gyrodinium</i> (Reñé et al., 2015)	15	<i>Gyrodinium</i>
	16. Genus <i>Amphidinium</i> sensu stricto (Flø Jørgensen, Murray, & Daugbjerg, 2004)	16	<i>Amphidinium</i> Single sequences: <i>Amphidinium mootonorum</i> , <i>A. herdmanii</i> , <i>A. longum</i>
	17. Torodinales (Boutrup et al., 2016)	17A	<i>Torodinium</i>
17B		<i>Kapelodinium</i>	
Dinophyceae ordo incertae sedis	18. Tovelliaceae (Adl et al., 2012; Lindberg et al., 2005)	18A	<i>Esotropodinium</i>
		18B	<i>Jadwigia</i> (including #JQ639765 <i>Woloszynskia</i> sp.)
		18C	<i>Tovellia</i> (including #AY443025 <i>Woloszynskia leopoliensis</i> )
Peridinales	19. Genus <i>Blastodinium</i> (Skovgaard & Salomonsen, 2009)	19A	<i>Blastodinium navicula</i> , <i>B. mangini</i> , <i>B. galatheanum</i>
		19B	<i>Blastodinium spinulosum</i> , <i>B. crassum</i> , <i>B. pruvoti</i> , <i>B. inornatum</i>
		19C	<i>Blastodinium contortum</i> Single sequence: <i>Blastodinium oviforme</i>
Dinophyceae ordo incertae sedis	20. Ptychodiscales (Adl et al., 2012)	20	Single sequence: <i>Ptychodiscus noctiluca</i>
	UTD: «Uncertain Thecate Dinoflagellates»	UTD	<i>Adenoides</i> , <i>Ailadinium</i> , <i>Amphidiniella</i> , <i>Bysmatrum</i> , <i>Glenoaulax</i> , <i>Glenodiniopsis</i> , <i>Gloeodinium</i> , <i>Hemidinium</i> , <i>Heterodinium</i> , <i>Madanikinium</i> , <i>Oodinium</i> , <i>Palatinus</i> , <i>Parvodinium</i> , <i>Peridinium sociale</i> , <i>Peridiniopsis borgei</i> , <i>Pileidinium</i> , <i>Pseudadenoides</i> , <i>Rufusiella</i> , <i>Sabulodinium</i> , <i>Stylodinium</i> , <i>Thecadinium</i> , <i>Zooxanthea</i>
	UND: «Uncertain Naked Dinoflagellates»	UND	<i>Ankistrodinium</i> , <i>Apicoporus</i> , <i>Balechina</i> , <i>Bispinodinium</i> , <i>Ceratoperidinium</i> , <i>Margalefidinium</i> , <i>Cucumeridinium</i> , <i>Levanderina</i> , <i>Moestrupia</i> , <i>Testudodinium</i> , <i>Togula</i>

Clades represent the statistically supported (bootstrap values  $\geq 50$ ) larger clades of the majority-rule consensus tree. Clades are grouped into Superclades based on recent taxonomic literature; a selection of references supporting the identification of Superclades is reported. Dinoflagellate species included in DINOREF, but lacking morphological or phylogenetic evidence to be placed within any Superclade has been grouped in "Uncertain Thecate Dinophyceae" and "Uncertain Naked Dinophyceae." Names in bold represent genera included in the multigene dinoflagellate phylogeny by Orr et al. (2012). Colours identifying clades, and Superclades correspond to those used in the tree (Figure 2).

#3 (Suessiales), but included less than half as many sequences and a far lower number of genera and species. Within genera, the largest p-distance values were found for *Amphidinium*, *Coolia*, *Gambierdiscus*, *Gonyaulax*, *Protoperidinium* and *Togula* (Figure 4).

When clustered into OTUs at 98% similarity, the 946 unique V4 sequences were reduced to 313 OTUs (Supplementary Material 6); 33 of these OTUs included sequences from different species within

the same genus and 12 OTUs included sequences from different genera. Remarkably, a single OTU (OTU #126 in Supplementary material 7) contained sequences from 59 genera belonging to different Superclades (e.g., *Karlodinium*, *Prorocentrum*, *Gyrodinium*, *Podolampas*, *Duboscquodinium*). On the other hand, sequences belonging to the same species clustered in different OTUs (e.g., *Alexandrium fundyense* Group\_I and *Gambierdiscus scabrosus*).

**TABLE 2** Number of unique and total dinoflagellate 18S rRNA gene sequences by Superclade included in the database

Superclades	No. of sequences in DINOREF		No. of taxa in DINOREF		Total no. of described taxa		
	Unique	Total	Genera	Species	Genera	Species	
# 1	Gonyaulacales	507	543	17	85	20	296
# 2	Dinophysiales	97	97	8	38	13	358
# 3	Suessiales	223	240	14	29	26	91
# 4	Thoracosphaeraceae	71	82	16	22	19	66
# 5	Amphidomataceae	26	27	2	11	2	20
# 6	Kryptoperidiniaceae	21	23	5	10	6	16
# 7	<i>Pentapharsodinium-Ensiculifera</i>	6	6	2	4	2	6
# 8	Peridinales sensu stricto	106	116	13	46	25	475
# 9	Heterocapasaceae	18	20	1	7	1	16
# 10	Podolampadaceae	7	7	4	7	8	42
# 11	Prorocentrales	70	78	2	28	4	68
# 12	<i>Akashiwo</i>	8	13	1	1	1	1
# 13	Gymnodiniales sensu stricto	115	129	16	41	21	341
# 14	Kareniaceae	31	38	4	9	8	40
# 15	<i>Gyrodinium</i>	15	15	1	7	3	112
# 16	<i>Amphidinium</i>	36	40	1	10	3	101
# 17	Torodiniales	9	9	2	3	2	3
# 18	Tovelliaceae	6	10	4	4	4	19
# 19	<i>Blastodinium</i>	29	32	1	8	1	13
# 20	Ptychodiscals	1	1	1	1	1	2
UTD	Uncertain Thecate Dinophyceae	56	56	23	32	37	172
UND	Uncertain Naked Dinophyceae	82	89	11	19	25	84
Total	1,540	1,671	149	422	232	2,342	

Number of dinoflagellate genera and species represented in the database by at least one sequence. Sequences not assigned to the species level (annotated as "sp.") were not considered. Total number of genera and species described (based on Gómez (2012a), ALGAEBASE (Guiry & Guiry, 2017), CEDIT (Hoppenrath & Elbrächter, 2015).

## 4 | DISCUSSION

DINOREF adds to the already available 18S protists databases PR<sup>2</sup> (Guilou et al., 2013) and SILVA (Quast et al., 2013) a considerable number of new 18S sequences, provides an updated and validated identification for more than 400 entries, and places each sequence within a curated 8-level taxonomic framework. This is a "conservative" hierarchical system, largely reflecting the one presented in ALGAEBASE (Guiry & Guiry, 2017). In addition, we also present a tentative classification of dinoflagellates based on the 18S phylogeny and supported by recent taxonomic literature ("Superclade" in Table 1; column D in Supplementary Material 1). DINOREF also includes sequences of the V4 region with information on this marker's capability to discriminate among species.

The marked polytomy of the 18S tree built with the sequences included in DINOREF leaves the phylogenetic status of most of the higher taxa unresolved. These results are in line with what was known from previous studies carried out with this ribosomal marker (e.g., Bachvaroff et al., 2014; Murray et al., 2005; Saldarriaga, Taylor, Keeling, & Cavalier-Smith, 2001). Recovery of morphologically

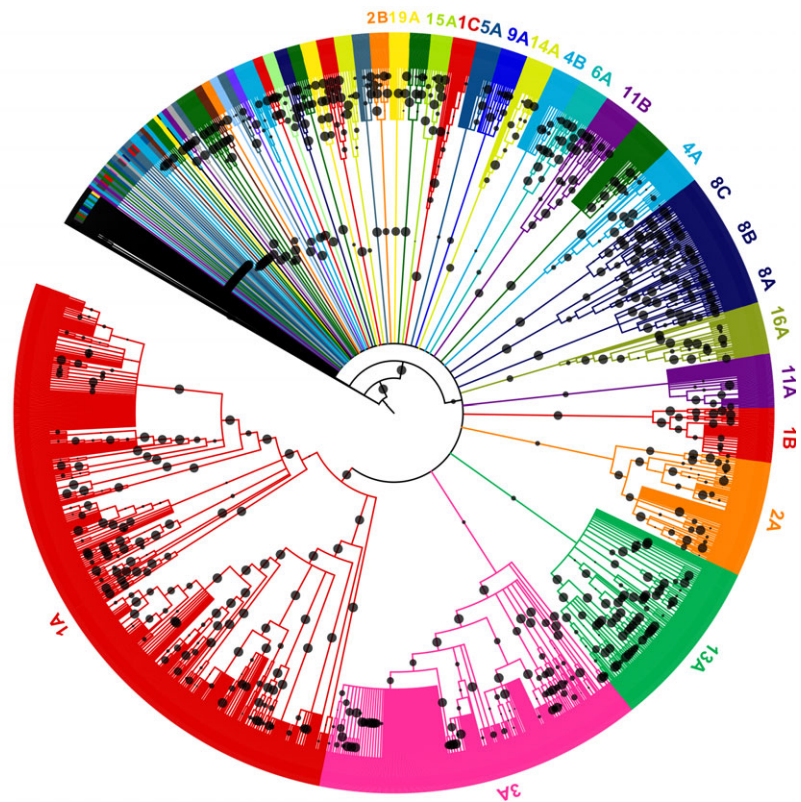
defined taxa in multiple clades emerging directly from a polytomy neither supports nor falsifies the natural status of these taxa; morphological evidence is simply not backed up by molecular evidence. Whether this polytomy is hard, that is, real, due to a brief period of rapid radiation, or soft, due to paucity of data, remains to be resolved. One promising approach is multigene phylogenies that start to provide a clearer picture of dinoflagellate evolutionary patterns (Janouškovec et al., 2017; Orr et al., 2012). Some groups, such as Gymnodiniales, appear to be polyphyletic and are in need of taxonomic revision. As a matter of fact, various papers have been recently published to clarify the taxonomic position of several groups (e.g., Boutrup, Moestrup, Tillmann, & Daugbjerg, 2016; Yuasa, Horiguchi, Mayama, & Takahashi, 2016). It is clear that the taxonomic treatment of the genera needs revision. For example, genera such as *Protoperdinium* are paraphyletic with daughter genera inside them (Gu, Liu, & Mertens, 2015; Liu et al., 2015; Mertens et al., 2015), suggesting that these paraphyletic genera need to be split.

The comparison between the number of dinoflagellate species estimated by recent checklists (Gómez, 2012a) and ALGAEBASE (Guiry & Guiry, 2017) and those for which sequences are available in

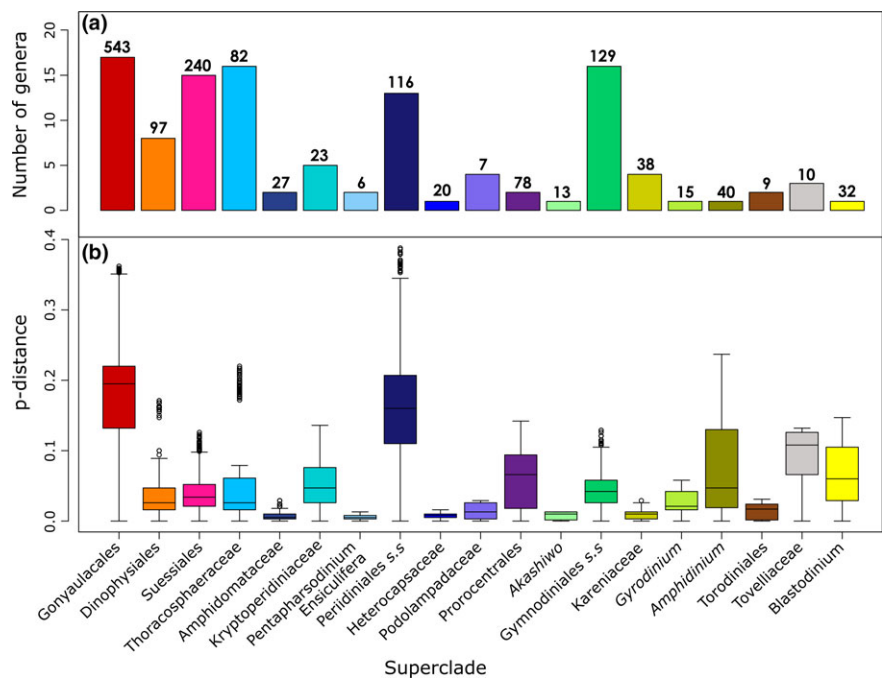


## Superclades

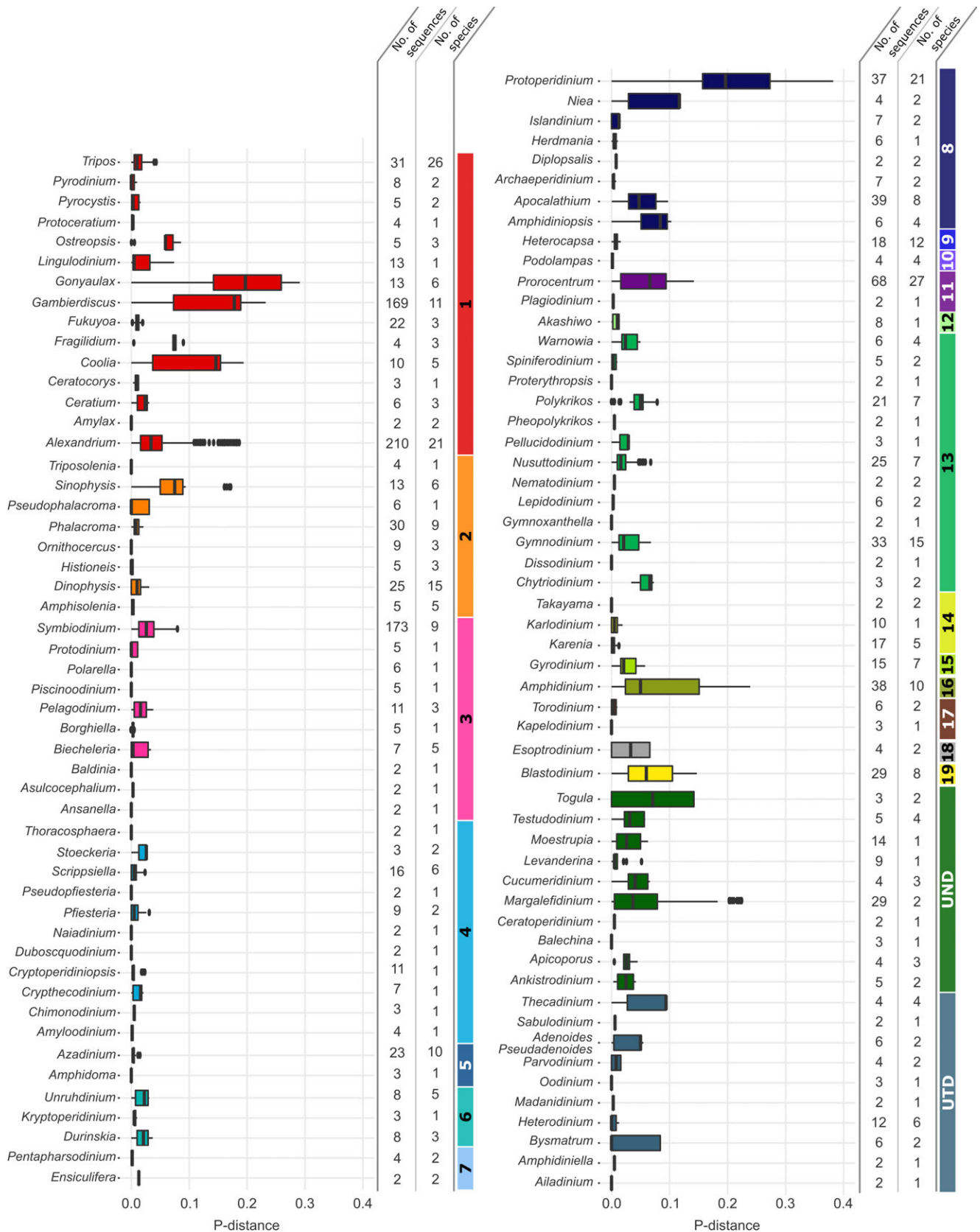
- # 1 - Gonyaulacales
- # 2 - Dinophysiales
- # 3 - Suessiales
- # 4 - Thoracosphaeraceae
- # 5 - Amphidomataceae
- # 6 - Kryptoperidiniaceae
- # 7 - genera *Pentapharsodinium-Enciculifera*
- # 8 - Peridinales sensu stricto
- # 9 - Heterocapsaceae
- # 10 - Podolampadaceae
- # 11 - Prorocentrales
- UTD - Uncertain Thecate Dinophyceae
- # 12 - genus *Akashiwo*
- # 13 - Gymnodinales sensu stricto
- # 14 - Kareniaceae
- # 15 - genus *Gyrodinium*
- # 16 - genus *Amphidinium*
- # 17 - Torodinales
- # 18 - Tovelliaceae
- # 19 - genus *Blastodinium*
- # 20 - Ptychodiscales
- UND - Uncertain Naked Dinophyceae
- OUTGROUPS



**FIGURE 2** Consensus phylogenetic tree (RAxML, GTR model) based on 1,540 unique 18S rRNA sequences in the DINOREF. Alignment of 2,153 bp with three sequences of Ciliates (U97109; X56165 and X03772) and three sequences of Apicomplexa (M97703; AF236097 and AF291427) used as outgroup. Clades are ordered according to their size and are supported by bootstrap values  $\geq 50\%$ ; black dots are proportional to bootstrap values. The colours of the Superclades and clades correspond to those in Table 1. Clades within each Superclade have been marked (A, B, C, etc.), along the outer rim of the tree, corresponding to their assignment in this figure. The Superclades “Uncertain Naked Dinophyceae” and “Uncertain Thecate Dinophyceae” have not been marked and neither have the small clades on the upper left of the tree. The tree can be visualized on *ITOL* version 3—Interactive Tree of Life (Letunic and Bork, 2016, at <https://itol.embl.de/tree/1932052318357911479398328>) in which all clades are marked



**FIGURE 3** (a) Barplot showing the number of genera with 18S rRNA information in 19 of the 20 Superclades depicted in Table 1 and Figure 2. Superclade 20 (Ptychodiscales) is not shown as it contains only one sequence. The number of sequences in each Superclade is reported on the top of each bar. (b) Boxplot showing the pairwise p-distances of the V4 region in the 19 dinoflagellate Superclades



**FIGURE 4** Boxplots showing the range of the pairwise p-distances over the V4 region of the 18S rRNA sequences within the genera included in each Superclade (indicated by number and colour code as in Table 1 and Figure 2). Number of sequences that have been used to calculate pairwise p-distance and number of species represented by those sequences are specified for each genus. Sequences with no species name (annotated “sp.”) were not accounted for in the number of species, but still used for the calculation of pairwise p-distance

DINOREF shows that all the taxonomic groups (defined here as Superclades) have representatives in the 18S data set, but not all of them are equally well covered. The coverage of the DINOREF database is strongly biased towards sequences of species in the focus of global research such as toxic species (e.g., *Alexandrium*, *Gambierdiscus*) and endosymbionts associated with coral reefs (*Symbiodinium*). In general, it is also biased towards autotrophs because these organisms are more easily grown in culture and therefore have been characterized more easily than heterotrophs and mixotrophs, which are difficult to isolate and may require alternative approaches such as single-cell sequencing. The DINOREF database is still far from covering dinoflagellate diversity. Only 18% of the 2,342 described species according to ALGAEBASE (Guiry & Guiry, 2017) are represented by an 18S reference sequence. The 422 dinoflagellate species with an 18S sequence present in DINOREF represent a marked increase from the around 150 species reported by Murray et al. (2005) and from the 345 by Gómez (2014) showing that sequence information is increasing rapidly as a result of the description of new species from different geographic areas. However, there are still 83 genera and 1,639 species for which no 18S sequences are available yet.

In some cases, however, the diversity of species defined based on morphological characters is probably overestimated. This may be the case of species attributed to the genus *Gymnodinium*, where 38% of the 270 described species have not been reported since their original description (Thessen, Patterson, & Murray, 2012). However, there are also genera such as *Symbiodinium* and *Scrippsiella* in which a marked intrageneric diversity has been uncovered by molecular approaches (Gottschling et al., 2012; Pochon, Putnam, & Gates, 2014).

The usefulness of metabarcoding using the V4 region of the 18S rRNA gene depends on how well the region is able to distinguish the various taxa of interest (Bendif et al., 2014; Hu et al., 2015). For dinoflagellates, it is considered variable enough to discern most of the species (Ki, 2012; Smith, Kohli, Murray, & Rhodes, 2017) although there are some exceptions. Our results show that the V4 region can unambiguously discriminate 374 of the 422 species included in DINOREF. On the upside, this outcome illustrates the capability of the V4 barcode region to resolve protist diversity (Hu et al., 2015; Ki, 2012), but on the downside, the V4 is unable to discriminate between some toxic and nontoxic close relatives within *Dinophysis*, *Karenia* and *Azadinium*. In metabarcoding studies, sequences are clustered at a given similarity level (98% or 97%) to avoid inflating diversity estimates (Massana et al., 2015; Onda et al., 2017; Smith et al., 2017; de Vargas et al., 2015). This may be appropriate for species exhibiting high intraspecific sequence variation (e.g., in *Symbiodinium*, *Gambierdiscus* and *Alexandrium*), yet clustering at the 98% level means that closely related species are lumped together in single OTU. We therefore recommend using ribotypes (unique sequences without any OTU clustering) rather than clustering for studies that focus on the species-level biodiversity of dinoflagellates.

The enormous diversity of unicellular organisms, the augmented knowledge we have about their morphology, ecology and life cycles together with the "revolution" of molecular approaches call for establishing a common taxonomic framework (e.g., Berney et al., 2017).

Phylogenetic studies and metabarcoding approaches will provide important information in this direction in the coming years. We have shown the good resolution capability of the V4 18S rDNA marker for dinoflagellates, but other barcode regions are worth exploring such as conserved regions of the LSU (e.g., Grzebyk et al., 2017). We recommend testing different potential markers in parallel with the production of new 18S reference sequences, because a larger amount of data is required to achieve the best possible solution for dinoflagellates and protists in general.

#### NOTE ADDED IN PROOF

Proposed taxonomic revisions are being incorporated in NCBI GENBANK.

#### ACKNOWLEDGEMENTS

SM has been supported by a PhD fellowship from Stazione Zoologica Anton Dohrn (SZN). This study was supported by the project FIRB Biodiversitalia (RBAP10A2T4) funded by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR). We are grateful to Dr. Chris Bowkett for improving the use of English in the manuscript.

#### AUTHOR CONTRIBUTION

S.M., D.S., R.P., W.H.C.F.K., D.V. and M.M. designed the research. S.M. and R.P. collected data and the designed framework for molecular and computational analyses. S.M. assembled the database and provided a first draft of the manuscript; all authors contributed to its final version.

#### DATA ACCESSIBILITY

DINOREF is available in flat file format on Figshare at <https://doi.org/10.6084/m9.figshare.5568454> and is incorporated in the Protist Ribosomal Reference database (PR<sup>2</sup>) version 4.9.0 which is available at <https://doi.org/10.6084/m9.figshare.5913181>.

#### ORCID

Diana Sarno  <http://orcid.org/0000-0001-9697-5301>

#### REFERENCES

- Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., ... Spiegel, F. W. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59(5), 429–493. <https://doi.org/10.1111/j.1550-7408.2012.00644.x>
- Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE*, 4(7), e6372. <https://doi.org/10.1371/journal.pone.0006372>



- Anderson, D. M., Cembella, A. D., & Hallegraeff, G. M. (2012). Progress in understanding harmful algal blooms: Paradigm shifts and new technologies for research, monitoring, and management. *Annual Review of Marine Science*, 4(1), 143–176. <https://doi.org/10.1146/annurev-marine-120308-081121>
- Bachvaroff, T. R., Gornik, S. G., Concepcion, G. T., Waller, R. F., Mendez, G. S., Lippmeier, J. C., & Delwiche, C. F. (2014). Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. *Molecular Phylogenetics and Evolution*, 70(1), 314–322. <https://doi.org/10.1016/j.ympev.2013.10.007>
- Bendif, E. M., Probert, I., Carmichael, M., Romac, S., Hagino, K., & de Vargas, C. (2014). Genetic delineation between and within the widespread coccolithophore morpho-species *Emiliana huxleyi* and *Gephyrocapsa oceanica* (Haptophyta). *Journal of Phycology*, 50(1), 140–148. <https://doi.org/10.1111/jpy.12147>
- Berdalet, E., Fleming, L. E., Gowen, R., Davidson, K., Hess, P., Backer, L. C., ... Enevoldsen, H. (2016). Marine harmful algal blooms, human health and wellbeing: Challenges and opportunities in the 21st century. *Journal of the Marine Biological Association of the United Kingdom*, 96(1), 61–91. <https://doi.org/10.1017/S0025315415001733>
- Berney, C., Ciuprina, A., Bender, S., Brodie, J., Edgcomb, V., Kim, E., ... de Vargas, C. (2017). UniEuk: Time to speak a common language in protistology!. *Journal of Eukaryotic Microbiology*, 64(3), 407–411. <https://doi.org/10.1111/jeu.12414>
- Boutrup, P. V., Moestrup, Ø., Tillmann, U., & Daugbjerg, N. (2016). *Katodinium glaucum* (Dinophyceae) revisited: Proposal of new genus, family and order based on ultrastructure and phylogeny. *Phycologia*, 55(2), 147–164. <https://doi.org/10.2216/15-138.1>
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Walters, W. A. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5), 335–336. <https://doi.org/10.1038/nmeth.f.303.QIIME>
- de Salas, M. F., Laza-Martínez, A., & Hallegraeff, G. M. (2008). Novel unarmored dinoflagellates from the toxigenic family Kareniaceae (Gymnodiniales): Five new species of *Karodinium* and one new *Takayama* from the Australian sector of the Southern Ocean. *Journal of Phycology*, 44(1), 241–257. <https://doi.org/10.1111/j.1529-8817.2007.00458.x>
- de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., ... Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1261605. <https://doi.org/10.1126/science.1261605>
- Decelle, J., Romac, S., Stern, R. F., Bendif, E. M., Zingone, A., Audic, S., ... Christen, R. (2015). PhytoREF: A reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources*, 15(6), 1435–1445. <https://doi.org/10.1111/1755-0998.12401>
- Fensome, R. A., Taylor, M. F. J. R., Norris, G., Sarjeant, W. A. S., Wharton, D. I., & Williams, G. L. (1993). *A classification of living and fossil dinoflagellates*. Hanover, Pennsylvania, USA: Sheridan Press.
- Flø Jørgensen, M. F., Murray, S., & Daugbjerg, N. (2004). *Amphidinium* revisited. I. Redefinition of *Amphidinium* (Dinophyceae) based on cladistic and molecular phylogenetic analyses. *Journal of Phycology*, 40(2), 351–365. <https://doi.org/10.1111/j.1529-8817.2004.03131.x>
- Gómez, F. (2012a). A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata). *Systematics and Biodiversity*, 10(3), 267–275. <https://doi.org/10.1080/14772000.2012.721021>
- Gómez, F. (2012b). A checklist and classification of living. *CICIMAR Océánides*, 27(1), 65–140.
- Gómez, F. (2014). Problematic biases in the availability of molecular markers in protists: The example of the Dinoflagellates. *Acta Protozoologica*, 53(1), 63. <https://doi.org/10.4467/16890027AP.13.0021.1118>
- Gómez, F., Moreira, D., & López-García, P. (2010). *Neoceratium* gen. nov., a new genus for all marine species currently assigned to *Ceratium* (Dinophyceae). *Protist*, 161(1), 35–54. <https://doi.org/10.1016/j.protis.2009.06.004>
- Gottschling, M., Čalasan, A. Ž., Kretschmann, J., & Gu, H. (2017). Two new generic names for dinophytes harbouring a diatom as an endosymbiont, *Blixaea* and *Unruhadinium* (Kryptoperidiniaceae, Peridinales). *Phytotaxa*, 306(4), 296–300. <https://doi.org/10.11646/phytotaxa.306.4.6>
- Gottschling, M., Soehner, S., Zinssmeister, C., John, U., Plötner, J., Schweikert, M., ... Elbrächter, M. (2012). Delimitation of the Thoracosphaeraeaceae (Dinophyceae), including the calcareous dinoflagellates, based on large amounts of ribosomal RNA sequence data. *Protist*, 163(1), 15–24. <https://doi.org/10.1016/j.protis.2011.06.003>
- Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4: A multi-platform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2), 221–224. <https://doi.org/10.1093/molbev/msp259>
- Grzebyk, D., Audic, S., Lasserre, B., Abadie, E., de Vargas, C., & Bec, B. (2017). Insights into the harmful algal flora in northwestern Mediterranean coastal lagoons revealed by pyrosequencing metabarcodes of the 28S rRNA gene. *Harmful Algae*, 68, 1–16. <https://doi.org/10.1016/j.hal.2017.06.003>
- Gu, H., Liu, T., & Mertens, K. N. (2015). Cyst–theca relationship and phylogenetic positions of *Protoperdinium* (Peridinales, Dinophyceae) species of the sections *Conica* and *Tabulata*, with description of *Protoperdinium shanghaiense* sp. nov. *Phycologia*, 54(1), 49–66. <https://doi.org/10.2216/14-047.1>
- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... Christen, R. (2013). The Protist Ribosomal Reference database (PR<sup>2</sup>): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 41(D1), D597–D604. <https://doi.org/10.1093/nar/gks1160>
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321. <https://doi.org/10.1093/sysbio/syq010>
- Guiry, M. D., & Guiry, G. M. (2017). AlgaeBase. World-wide electronic publication. <http://www.algaebase.org>; searched on 15 October 2017
- Hackett, J. D., Anderson, D. M., Erdner, D. L., & Bhattacharya, D. (2004). Dinoflagellates: A remarkable evolutionary experiment. *American Journal of Botany*, 91, 1523–1534. <https://doi.org/10.3732/ajb.91.10.1523>
- Hoppenrath, M., Chomérat, N., & Leander, B. S. (2013). Molecular phylogeny of *Sinophysis*: Evaluating the possible early evolutionary history of dinophysoid dinoflagellates. In *Biological and geological perspectives of dinoflagellates*. The Micropalaeontological Society, Special Publications. Geological Society, London (pp. 207–214). <https://doi.org/10.1144/TMS5>
- Hoppenrath, M., & Elbrächter, M. (2015). *Center of Excellence for Dinophyte Taxonomy (CEDiT)*. Retrieved from <http://www.dinophyta.org>
- Hoppenrath, M., & Leander, B. S. (2007). Morphology and phylogeny of the pseudocolonial dinoflagellates *Polykrikos lebourae* and *Polykrikos herdmanae* n. sp. *Protist*, 158(2), 209–227. <https://doi.org/10.1016/j.protis.2006.12.001>
- Hoppenrath, M., Murray, S. A., Chomérat, N., & Horiguchi, T. (2014). In M. Hoppenrath, S. A. Murray, N. Chomérat & T. Horiguchi (Eds.), *Marine benthic dinoflagellates - unveiling their worldwide biodiversity* (Kleine Sen). Stuttgart, Germany: Schweizerbart'sche Verlagsbuchhandlung.
- Hoppenrath, M., Murray, S., Sparrmann, S. F., & Leander, B. S. (2012). Morphology and molecular phylogeny of *Ankistrodinium* gen. nov. (Dinophyceae), a new genus of marine sand-dwelling dinoflagellates formerly classified within *Amphidinium*. *Journal of Phycology*,

- 48(5), 1143–1152. <https://doi.org/10.1111/j.1529-8817.2012.01198.x>
- Hu, S. K., Liu, Z., Lie, A. A. Y., Countway, P. D., Kim, D. Y., Jones, A. C., ... Caron, D. A. (2015). Estimating protistan diversity using high-throughput sequencing. *Journal of Eukaryotic Microbiology*, 62(5), 688–693. <https://doi.org/10.1111/jeu.12217>
- Janoušková, J., Gavelis, G. S., Burki, F., Dinh, D., Bachvaroff, T. R., Gornik, S. G., ... Saldarriaga, J. F. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proceedings of the National Academy of Sciences*, 114(2), E171–E180. <https://doi.org/10.1073/pnas.1614842114>
- Jephcott, T. G., Alves-de-Souza, C., Gleason, F. H., van Ogtrop, F. F., Sime-Ngando, T., Karpov, S. A., & Guillou, L. (2016). Ecological impacts of parasitic chytrids, syndiniales and perkinsids on populations of marine photosynthetic dinoflagellates. *Fungal Ecology*, 19, 47–58. <https://doi.org/10.1016/j.funeco.2015.03.007>
- John, U., Litaker, R. W., Montresor, M., Murray, S., Brosnahan, M. L., & Anderson, D. M. (2014). Formal revision of the *Alexandrium tamarense* species complex (Dinophyceae) taxonomy: The introduction of five species with emphasis on molecular-based (rDNA) classification. *Protist*, 165(6), 779–804. <https://doi.org/10.1016/j.protis.2014.10.001>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Ki, J.-S. (2012). Hypervariable regions (V1-V9) of the dinoflagellate 18S rRNA using a large dataset for marker considerations. *Journal of Applied Phycology*, 24(5), 1035–1043. <https://doi.org/10.1007/s10811-011-9730-z>
- Ki, J.-S., Jang, G. Y., & Han, M.-S. (2005). Integrated method for single-cell DNA extraction, PCR amplification, and sequencing of ribosomal DNA from harmful dinoflagellates *Cochlodinium polykrioides* and *Alexandrium catenella*. *Marine Biotechnology*, 6(6), 587–593. <https://doi.org/10.1007/s10126-004-1700-x>
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, 33(7), 1870–1874. <https://doi.org/10.1093/molbev/msw054>
- Le Bescot, N., Mahé, F., Audic, S., Dimier, C., Garet, M. J., Poulain, J., ... Siano, R. (2016). Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environmental Microbiology*, 18(2), 609–626. <https://doi.org/10.1111/1462-2920.13039>
- Lindberg, K., Moestrup, Ø., Daugbjerg, N., Woloszynska, I., Woloszynska, J., & Ehrenb, P. (2005). Studies on woloszynskioid dinoflagellates I: *Woloszynskia coronata* re-examined using light and electron microscopy and partial LSU rDNA sequences, with description of *Tovellia* gen. nov. and *Jadwigia* gen. nov. (Tovelliaceae fam. nov.). *Phycologia*, 44(4), 416–440. [https://doi.org/10.2216/0031-8884\(2005\)44\[416:sowdiw\]2.0.co;2](https://doi.org/10.2216/0031-8884(2005)44[416:sowdiw]2.0.co;2)
- Liu, T., Mertens, K. N., Ribeiro, S., Ellegaard, M., Matsuoka, K., & Gu, H. (2015). Cyst-theca relationships and phylogenetic positions of Peridiniales (Dinophyceae) with two anterior intercalary plates, with description of *Archaeperidinium bailongense* sp. nov. and *Protoperidinium fuzhouense* sp. nov. *Phycological Research*, 63(2), 134–151. <https://doi.org/10.1111/pre.12081>
- Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., ... de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, 17(10), 4035–4049. <https://doi.org/10.1111/1462-2920.12955>
- Mertens, K. N., Takano, Y., Yamaguchi, A., Gu, H., Bogus, K., Kremp, A., ... Matsuoka, K. (2015). The molecular characterization of the enigmatic dinoflagellate *Kolkwitzia acuta* reveals an affinity to the *Excentrica* section of the genus *Protoperidinium*. *Systematics and Biodiversity*, 13(6), 509–524. <https://doi.org/10.1080/14772000.2015.1078855>
- Moestrup, Ø., Akselmann, R., Fraga, S., Hansen, G., Hoppenrath, M., Iwataki, M., ... Zingone, A. (2009). IOC-UNESCO taxonomic reference list of harmful micro algae. Accessed at <http://www.marinespecies.org/hab> on 2017-03-13
- Montresor, M., John, U., Beran, A., & Medlin, L. K. (2004). *Alexandrium tamutum* sp. nov. (Dinophyceae): A new nontoxic species in the genus *Alexandrium*. *Journal of Phycology*, 40(2), 398–411. <https://doi.org/10.1111/j.1529-8817.2004.03060.x>
- Morard, R., Darling, K. F., Mahé, F., Audic, S., Ujiie, Y., Weiner, A. K. M., ... de Vargas, C. (2015). PFR2: A curated database of planktonic foraminifera 18S ribosomal DNA as a resource for studies of plankton ecology, biogeography and evolution. *Molecular Ecology Resources*, 15(6), 1472–1485. <https://doi.org/10.1111/1755-0998.12410>
- Murray, S., Flø Jørgensen, M., Ho, S. Y. W., Patterson, D. J., & Jermini, L. S. (2005). Improving the analysis of dinoflagellate phylogeny based on rDNA. *Protist*, 156(3), 269–286. <https://doi.org/10.1016/j.protis.2005.05.003>
- Not, F., Siano, R., Kooistra, W. H. C. F., Simon, N., Vaulot, D., & Probert, I. (2012). Diversity and ecology of eukaryotic marine phytoplankton. In G. Piganeau (Ed.), *Genomics insights into the biology of algae* (pp. 1–53). Amsterdam, The Netherlands: Elsevier Academic Press. <https://doi.org/10.1016/b978-0-12-391499-6.00001-3>
- Onda, D. F. L., Medrinal, E., Comeau, A. M., Thaler, M., Babin, M., & Lovejoy, C. (2017). Seasonal and interannual changes in ciliate and dinoflagellate species assemblages in the Arctic Ocean (Amundsen Gulf, Beaufort Sea, Canada). *Frontiers in Marine Science*, 4, 16. <https://doi.org/10.3389/fmars.2017.00016>
- Orr, R. J. S., Murray, S. A., Stüken, A., Rhodes, L., & Jakobsen, K. S. (2012). When naked became armored: An eight-gene phylogeny reveals monophyletic origin of theca in dinoflagellates. *PLoS ONE*, 7(11), e50004. <https://doi.org/10.1371/journal.pone.0050004>
- Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., ... de Vargas, C. (2012). CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, 10(11), e1001419. <https://doi.org/10.1371/journal.pbio.1001419>
- Piredda, R., Tomasino, M. P., D'Erchia, A. M., Manzari, C., Pesole, G., Montresor, M., ... Zingone, A. (2017). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiology Ecology*, 93(1), <https://doi.org/10.1093/femsec/fiw200>
- Pochon, X., Putnam, H. M., & Gates, R. D. (2014). Multi-gene analysis of *Symbiodinium* dinoflagellates: A perspective on rarity, symbiosis, and evolution. *PeerJ*, 2, e394. <https://doi.org/10.7717/peerj.394>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3), e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590–D596. <https://doi.org/10.1093/nar/gks1219>
- R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. <https://doi.org/10.1038/sj.hdy.6800737>
- Reñé, A., Camp, J., & Garcés, E. (2015). Diversity and phylogeny of Gymnodiniales (Dinophyceae) from the NW Mediterranean Sea



- revealed by a morphological and molecular approach. *Protist*, 166(2), 234–263. <https://doi.org/10.1016/j.protis.2015.03.001>
- Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, 4, e2584. <https://doi.org/10.7717/peerj.2584>
- Ruiz Sebastián, C., & O’Ryan, C. (2001). Single-cell sequencing of dinoflagellate (Dinophyceae) nuclear ribosomal genes. *Molecular Ecology Notes*, 1(4), 329–331. <https://doi.org/10.1046/j.1471-8278>
- Salas, R., Tillmann, U., & Kavanagh, S. (2014). Morphological and molecular characterization of the small armoured dinoflagellate *Heterocapsa minima* (Peridiniales, Dinophyceae). *European Journal of Phycology*, 49(4), 413–428. <https://doi.org/10.1080/09670262.2014.956800>
- Saldarriaga, J. F., Taylor, F. J. R., Keeling, P. J., & Cavalier-Smith, T. (2001). Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *Journal of Molecular Evolution*, 53(3), 204–213. <https://doi.org/10.1007/s002390010210>
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23), 7537–7541. <https://doi.org/10.1128/AEM.01541-09>
- Skovgaard, A., & Salomonsen, X. M. (2009). *Blastodinium galatheanum* sp. nov. (Dinophyceae) a parasite of the planktonic copepod *Acartia negligens* (Crustacea, Calanoida) in the central Atlantic Ocean. *European Journal of Phycology*, 44(3), 425–438. <https://doi.org/10.1080/09670260902878743>
- Smith, K. F., Kohli, G. S., Murray, S. A., & Rhodes, L. L. (2017). Assessment of the metabarcoding approach for community analysis of benthic-epiphytic dinoflagellates using mock communities. *New Zealand Journal of Marine and Freshwater Research*, 51(4), 555–576. <https://doi.org/10.1080/00288330.2017.1298632>
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M. J., Chistoserdov, A., ... Edgcomb, V. P. (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology*, 7(1), 72. <https://doi.org/10.1186/1741-7007-7-72>
- Stoecker, D. K. (1999). Mixotrophy among Dinoflagellates. *The Journal of Eukaryotic Microbiology*, 46(4), 397–401. <https://doi.org/10.1111/j.1550-7408.1999.tb04619.x>
- Takano, Y., Hansen, G., Fujita, D., & Horiguchi, T. (2008). Serial replacement of diatom endosymbionts in two freshwater dinoflagellates, *Peridiniopsis* spp. (Peridiniales, Dinophyceae). *Phycologia*, 47(1), 41–53. <https://doi.org/10.2216/07-36.1>
- Thessen, A. E., Patterson, D. J., & Murray, S. A. (2012). The taxonomic significance of species that have only been observed once: The genus *Gymnodinium* (Dinoflagellata) as an example. *PLoS ONE*, 7(8), e44015. <https://doi.org/10.1371/journal.pone.0044015>
- Tillmann, U., Gottschling, M., Nézan, E., Krock, B., & Bilien, G. (2014). Morphological and molecular characterization of three new *Azadinium* species (Amphidomataceae, Dinophyceae) from the Irminger Sea. *Protist*, 165(4), 417–444. <https://doi.org/10.1016/j.protis.2014.04.004>
- Weisse, T., Anderson, R., Arndt, H., Calbet, A., Hansen, P. J., & Montagnes, D. J. S. (2016). Functional ecology of aquatic phagotrophic protists – Concepts, limitations, and perspectives. *European Journal of Protistology*, 55, 50–74. <https://doi.org/10.1016/j.ejop.2016.03.003>
- Wickham, H. (2009). *ggplot2 Elegant graphics for data analysis*. Media (Vol. 35). <https://doi.org/10.1007/978-0-387-98141-3>
- Yuasa, T., Horiguchi, T., Mayama, S., & Takahashi, O. (2016). *Gymnoxanthella radiolariae* gen. et sp. nov. (Dinophyceae), a dinoflagellate symbiont from solitary polycystine radiolarians. *Journal of Phycology*, 52(1), 89–104. <https://doi.org/10.1111/jpy.12371>

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Mordret S, Piredda R, Vaultot D, Montresor M, Kooistra WHCF, Sarno D. DINOREF: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. *Mol Ecol Resour*. 2018;00:1–14. <https://doi.org/10.1111/1755-0998.12781>