Received: 3 November 2017 | Revised: 15 February 2018 | Accepted: 24 February 2018

DOI: 10.1111/1755-0998.12781

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

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

Solenn Mordret[1] | Roberta Piredda[1] | Daniel Vaulot[2] | Marina Montresor[1] | Wiebe H. C. F. Kooistra[1] | Diana Sarno[1]

[1]Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Naples, Italy

[2]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 assignation 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).