Exploration de la diversité et du potentiel pour la reproduction sexuée au sein du complexe d'espèces Amoebophrya ceratii (Syndiniales), parasites de dinoflagellés marins

Ruibo Cai

Sorbonne Université

École doctorale 227: Sciences de la Nature et de l'Homme: Écologie et Évolution

Station Biologique de Roscoff / Adaptation et diversité en milieu marin, UMR 7144

Hidden species diversity and the potential for sexual reproduction in the species complex *Amoebophrya ceratii* (Syndiniales), parasites of marine dinoflagellates

Exploration de la diversité et du potentiel pour la reproduction sexuée au sein du complexe d'espèces *Amoebophrya ceratii* (Syndiniales), parasites de dinoflagellés marins

Ruibo CAI

Thèse de doctorat en Evolution moléculaire et Génomique comparative

Dirigée par **Laure Guillou**

Présentée et soutenue publiquement le 5 Novembre 2019

Devant un jury composé de :

Georges Barbier, Pr HDR, ESIAB	Rapporteur
Ramon Massana, Pr, CSIC, Barcelone, Espagne	Rapporteur
Aurélie Chambouvet, CR, CNRS, LEMAR UMR 6539	Examinateur
Christophe Destombe, Pr, HDR, SU, UMI3614	Examinateur
Christine Paillard, DR HDR, CNRS, LEMAR UMR 6539	Examinateur
Eric Pelletier, CR, Genoscope, CEA	Examinateur
Laure Guillou, DR HDR, CNRS, UMR7144	Directeur de Thèse

Contents

General Introduction

Parasitism is a frequent lifestyle in nature and a major source of evolutionary pressure for both hosts and their parasites. Given the ubiquity of host-parasite interactions, understanding the factors that generate, maintain, and constrain these associations is of primary interest with implications for a wide range of ecological issues, including dynamics of emerging infectious diseases and invasions (Daszak et al., 2000; Keane and Crawley, 2002). Although there is a long history in studying marine parasites, in particular with respect to commercially exploited species and aquaculture, little is known on parasites of marine microbes.

Given the diversity and abundance of marine protists, their parasites would be a particularly promising area of studies. Although little studied, many extremely virulent microeukaryotic parasites infecting microalgae have been detected in the marine plankton. Among them are Syndiniales, which constitute a diverse and highly widespread group (Guillou et al., 2008). Because of their virulence and abundant offspring, such parasites have the potential to control dinoflagellate populations, and therefore toxic microalgal blooms (Montagnes et al., 2008; Chambouvet et al., 2008; Alves-de-Souza et al., 2012).

1 Diversity and Evolution in Alveolata

Alveolata is a large and diverse assemblage of protists and has been considered as a major clade across eukaryotes (Adl et al., 2012; Adl et al., 2019) (**Fig** 1). It, together with Stramenopiles and Rhizaria, forms the SAR lineage (Adl et al., 2019). Stramenopiles is a very diverse group ranging from members of the human gut flora, plant pathogens, to the photosynthetic diatoms and the giant kelps (Baldauf, 2003; Burki et al., 2007; Parfrey et al., 2010), while Rhizaria is the least studied supergroup but has started to draw more attention from scientists (Burki and Keeling, 2014). The alveolates were named based on the cortical alveoli just beneath the outer cell membrane (i.e. membranous sacs subtending the plasma membrane).

Ciliates, dinoflagellates and apicomplexans are three well-defined and relatively well-studied groups in Alveolata (**Fig** 1; Cavalier-Smith and Chao, 2004; Gajadhar et al., 1991; Tikhonenkov et al., 2014; Bachvaroff et al., 2014). The dinoflagellates are notable primary producers, especially in marine environments, and the apicomplexans are known as parasites, particularly the malaria agents *Plasmodium*. The ciliates are most notable for the diversity of their habitats and unusual cell biology including dual nuclei, one germinal and the other somatic. The other alveolate groups encompass a number of species that display alveolate features (e.g. cortical alveoli), but lack features that would ally them specifically with any one of these three subgroups. For instance, *Chromera velia* and *Vitrella brassicaformis* (classified under the phylum Chromerida) are both close relatives of the parasitic apicomplexan lineage but have photosynthetic plastids (Janouškovec et al., 2010; Khadka et al., 2015). At the base of the dinoflagellates are the Syndiniales (**Fig** 1), a group of parasitic dinoflagellates well represented by *Amoebophrya* spp. (Cachon and Cachon, 1987; Fensome, 1993). The

motile *Amoebophrya* sp. dinospores have a recognizable dinoflagellate cell shape but lack some of the more exotic features of the dinoflagellate nucleus, including the high DNA content and condensed chromosomes characteristic of dinophycean dinoflagellates (i.e. core dinoflagellates) (Adl et al., 2005; Cachon and Cachon, 1970). Interestingly, the intracellular trophont and sporont stages of *Amoebophrya* resemble some apicomplexans (Cachon and Cachon, 1987; Bachvaroff et al., 2011; Miller et al., 2012). Between the syndinian dinoflagellates and the apicomplexans are a suite of difficult species to assign including the parasites *Perkinsus marinus*, *Parvilucifera infectans* (classified under the phylum Perkinsozoa; Noren and Moestrup, 1999), and the heterotroph *Oxyrrhis marina*, all placed with or within the dinoflagellates (Dinoflagellata) (Bachvaroff et al., 2014).

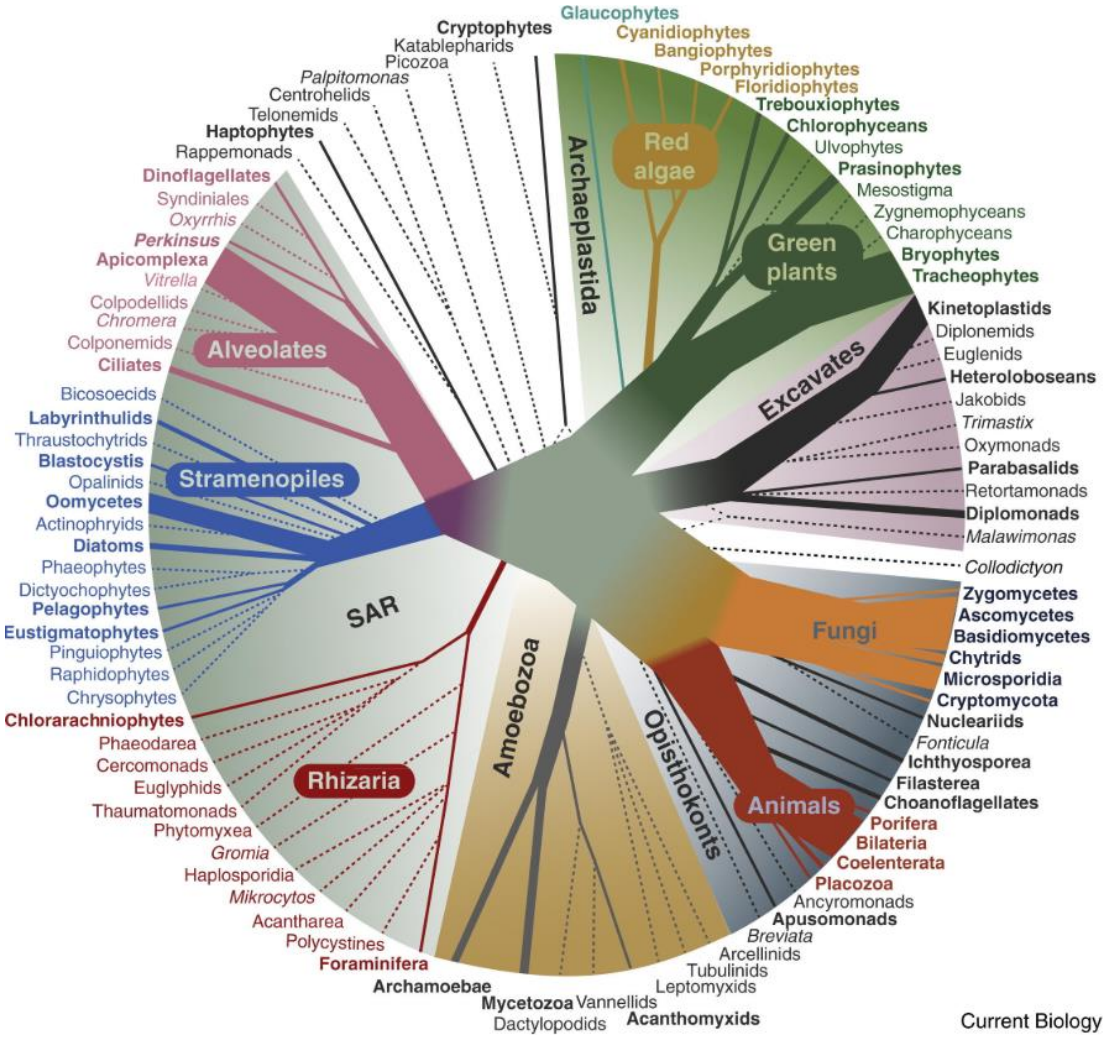


Fig 1. Evolutionary relationships among eukaryotes. (Burki and Keeling, 2014)

Members of Alveolata groups are related by various ultrastructural and genetic similarities (Fig 2A). However, the evolutionary relationship among them is really complicate and remains to be completely understood yet. Apicomplexans, chromerids and peridinin dinoflagellates share a monophyletic plastid lineage with heterokont algae, implying that they may have acquired their plastids from a red alga

(Janouskovec et al., 2010; Moore et al., 2008). So it seemed likely that the ancestor of the alveolate group was photosynthetic (Reyes-Prieto et al., 2008). Furthermore, it's suggested that the common ancestor of dinoflagellates, apicomplexans, Colpodella and Chromerida was a myzocytotic predator with two heterodynamic flagella, micropores, trichocysts, rhoptries, micronemes, a polar ring and a coiled open sided conoid (Fig 2B; Kuvardina et al., 2002). As ciliates ingest prey by a different mechanism (Tikhonenkov et al., 2014), it has been argued that myzocytosis was acquired after their emergence, and gave rise to other alveolates.

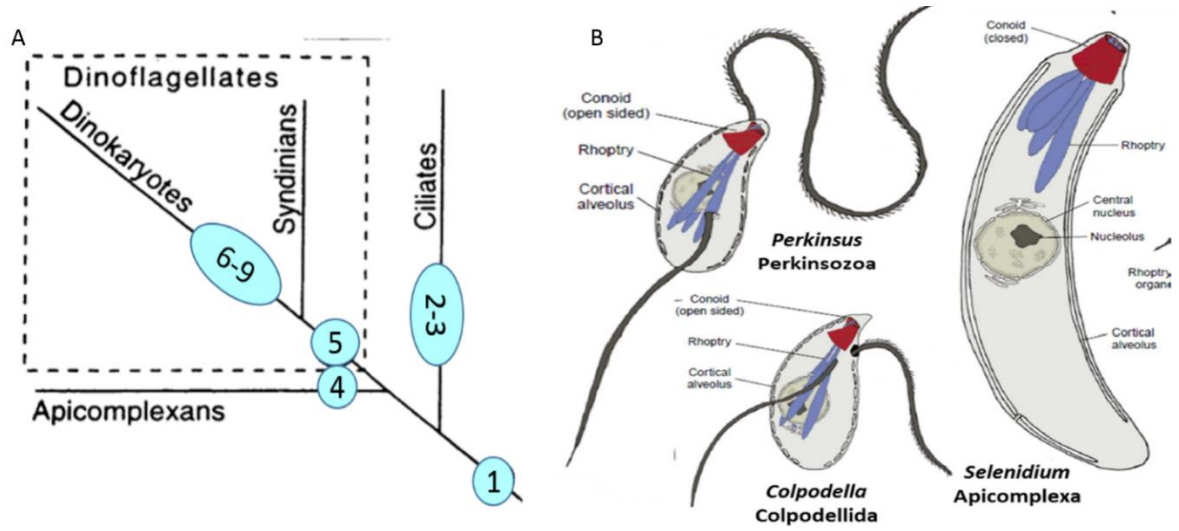


Fig 2. (A) Relationships of alveolate lineage based mainly on ultrastructure. Numbers indicate points in phylogeny where selected significant features appeared: 1 - alveolae; 2 - polykineties; 3 - nuclear dimorphism; 4 - apical complex; 5 - dinokont flagella; 6 - extranuclear mitotic spindle; 7 -temporary dinokaryon; 8 - permanent dinokaryon; 9 - loss of histones. (modified from Fensome et al. 1999.) (B) Major cytological features in several alveolate lineages (all with cortical alveoli). Red: conoid or open conoid, blue: rhoptries. Conoid and rhoptries are important components of the apical system in apicomplexans. Similar structures have been detected in Syndiniales (Miller et al. 2012). From Leander and Keeling (2003).

Dinoflagellates appear to have diverged from ciliates and apicomplexans around 900 million years ago [MYA] (Escalante and Ayala, 1995) and then showed a tremendous evolutionary radiation at the beginning of the Mesozoic (~250 MYA) (**Fig** 3; Fensome et al., 1999). However, dinoflagellates appear to be more closely related to apicomplexa than to the ciliates evolutionarily (Bachvaroff et al., 2011; Hoppenrath, 2017). Dinoflagellates and apicomplexa both have plastids, and most share a bundle or cone of microtubules at the top of the cell. In apicomplexans, this forms part of a complex used to enter host cells, while in some colorless dinoflagellates it forms a peduncle used to ingest prey.

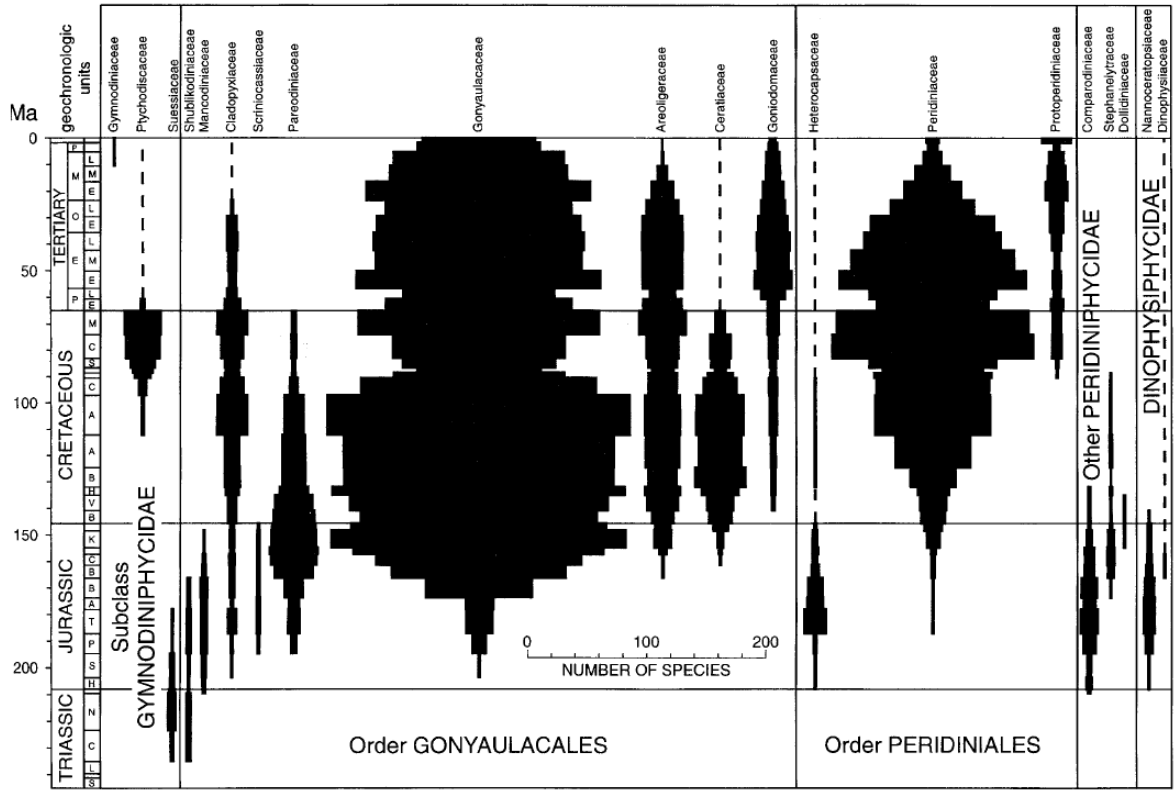


Fig 3. Spindle plots showing the number of species per family per time interval. Stages (Mesozoic) and epochs and subepochs (Tertiary) are indicated as follows, in ascending order. Triassic stages: S = Scythian, unlabelled = Anisian, L = Ladinian, C = Carnian, N = Norian, unlabelled = Rhaetian. Jurassic stages: H = Hettangian, S = Sinemurian, P = Pliensbachian, T = Toarcian, A = Aalenian, B = Bajocian, C = Callovian, unlabelled = Oxfordian, K = Kimmeridgian, unlabelled = Portlandian. Cretaceous stages: B = Berriasian, V =Valanginian, unlabelled = Hauterivian, B = Barremian, A = Aptian, A = Albian, C = Cenomanian, unlabelled = Turonian unlabelled = Coniacian, S = Santonian, C = Campanian, M = Maastrichtian. Tertiary epochs: P = Paleocene, E = Eocene, 0 = Oligocene, M = Miocene, P=Pliocene. Tertiary epochs are divided into Early (E) and Late (L) or Early (E), Mid (M), and Late (L). (Fensome et al., 1999)

2 Dinoflagellates

Dinoflagellates are morphologically distinct from other eukaryotes in the structure of their (dinokont) flagellar apparatus and (dinokaryotic) nucleus (i.e., permanently condensed chromosomes and with an extranuclear spindle that passes through cytoplasmic channels) (Taylor, 1987; Fensome, 1993; Hoppenrath and Saldarriaga, 2008 Hoppenrath et al., 2010). In terms of species number, dinoflagellates are one of the largest groups of marine eukaryotes (Guiry, 2012). The latest estimates suggest a total of 2200-2500 living dinoflagellate species (Hoppenrath, 2017).

2.1 Biology of dinoflagellates

Most (but not all) dinoflagellates have a dinokaryon, a unique eukaryotic nucleus structure in which the chromosomes are fibrillar in appearance, more or less continuously condensed (Gómez, 2012) and attached to the nuclear membrane. Dinoflagellate nuclei contain a novel, dominant family of nuclear

proteins that appear to be of viral origin, and thus are called dinoflagellate/viral nucleoproteins (DVNPs) (Gornik et al., 2012).

Dinoflagellates possess two dissimilar flagella (**Fig** 4) arising from the ventral cell side: a ribbon-like transverse flagellum that beats to the cell's left and a longitudinal flagellum that beats posteriorly (Gaines and Taylor, 1985). The flagellar movement produces forward propulsion and also a turning force. The flagella lie in surface grooves: the transverse one in the cingulum and the longitudinal one in the sulcus.

Dinoflagellates synthesize secondary metabolites including sterols, polyketides, toxins, and dimethylsulfide, which are of ecological importance (reviewed in Janouškovec et al., 2017).

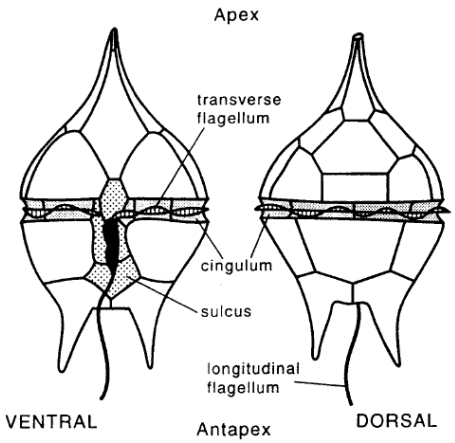


Fig 4. General morphology of a typical motile dinoflagellate. Cingulum and sulcus shaded; flagellar pore black. (Fensome et al., 1999).

2.2 Diversity of dinoflagellates

Dinoflagellates reveal extraordinary diversity in cell morphology and nutritional modes (e.g., phagotrophy, 'klepto-phototrophy', photoautotrophy, mixotrophy, and parasitism) (Taylor, 1987; Taylor et al., 2008; Hackett et al., 2004). Interactions between dinoflagellates and other organisms are very diverse, including symbioses (Decelle et al, 2015), predation (Jeong et al., 2010), kleptoplasty (Gast et al., 2007).

Both heterotrophic and autotrophic members of dinoflagellates are ecologically important components of marine planktonic communities (reviewed in Hoppenrath and Leander, 2010). About half of the extant species are photosynthetic (Gómez, 2012) and they constitute the dominant marine primary producers. Phagotrophic species play an important role in the microbial loop through predation and nutrient recycling. Some of the fast blooming species (e.g. *Alexandrium* species) can make up episodic blooms (red tides or harmful algal blooms) and produce toxins that do harm to fisheries or aquaculture (Flewelling et al., 2005; Kohli et al., 2016; Orr et al., 2013). Symbiotic genera like *Symbiodinium* participate in interactions with metazoans and are essential for the formation and sustaining of reef ecosystems in the oceans worldwide (Goodson et al., 2001; Lin et al., 2015). Parasitic species like

Amoebophrya spp. play a central role in the collapse of harmful algal blooms (Chambouvet et al., 2008; Velo-Suárez et al., 2013).

Dinoflagellates' ecological significance befits their abundance (Janouškovec et al., 2017). Environmental metabarcoding based on high-throughput sequencing is increasingly applied to assess diversity and abundance of planktonic organisms (de Vargas et al., 2015; Le Bescot et al., 2016; Massana et al., 2015). Dinoflagellates have been highlighted as important members with high abundance in both coastal and open-ocean protistan communities based on environmental molecular barcoding surveys (Le Bescot et al., 2016; Massana et al., 2015; de Vargas et al., 2015). The Massive metabarcoding sequencing from plankton communities collected across the world's surface oceans stressed that dinoflagellate diversity has been largely underestimated, representing overall ~1/2 of protistan rDNA (V4) metabarcode richness in the world's surface oceans (**Fig** 5, Le Bescot et al., 2016).

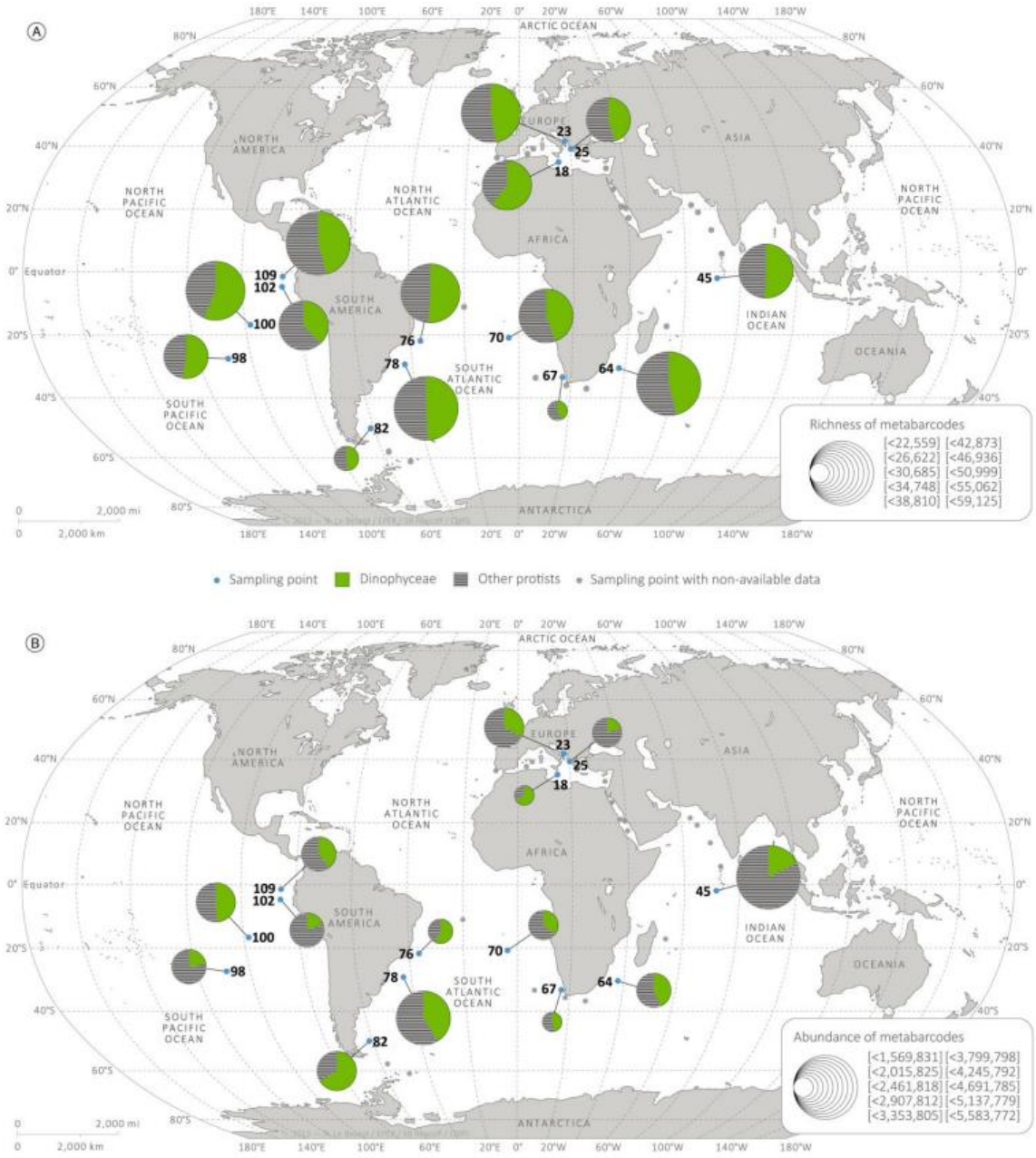


Fig 5. Metabarcode dinoflagellate richness (A) and abundance (B) (green) over total protist community (grey with stripes). Pie chart sizes are proportional to total metabarcode numbers analyzed per sample. (Le Bescot et al., 2016)

Along with metabarcoding surveys, environmental investigations of protists also entail genomic and transcriptomic data. Interpretation of such large data sets is limited by the lack of reference data from the host organisms, resulting in a high proportion of unknown sequences (Caron et al., 2016; Sibbald and Archibald, 2017). This is particularly significant for dinoflagellates as this taxon remains poorly explored at the sequence level (Meng et al., 2018). Presently, sequence data is only available for a small proportion (around 10% or even less) of the known dinoflagellate species diversity (Murray et al., 2005; Orr et al., 2012). Furthermore, a bias towards the photosynthetic taxa also exists, as a large proportion

of heterotrophic species, which make up about 50% of the true dinoflagellate (core dinoflagellate) lineage, are difficult or impossible to culture (Orr et al., 2012).

Noteworthy, the diversity in dinoflagellates has been greatly extended by the discovery of an astonishing breadth and abundance of sequences assigned to‘marine alveolates’ from marine environmental clone libraries (Bachvaroff et al., 2014; Lopez Garcia et al., 2001; Moon-van der Staay et al., 2001). Many of these sequences are placed with known syndinean dinoflagellates in phylogenies and the raw abundance of such sequences dwarfs the tens of sequences attributed to described syndinean species or genera (Bachvaroff et al., 2012). The identification of the marine alveolate lineages (MALVs) gives an insight into the large parasitic Syndiniales diversity (Lopez Garcia et al., 2001; Moon-van der Staay et al., 2001; Skovgaard et al., 2009; Brate et al., 2012; Harada et al., 2007; Guillou et al., 2008).

2.3 Taxonomy of dinoflagellates

Dinoflagellates (Dinoflagellata) are divided into around seven classes (**Table** 1; Gómez, 2012), among which Dinophyceae/Dinokaryota is the most diverse. Dinoflagellate taxonomy is based on morphological characters such as the presence of a dinokaryon, and the arrangement and shape of thecal plate-containing amphiesmal vesicles (i.e. tabulation).

Table 1. The hierarchical classification for Dinoflagellata (Gómez, 2012; Adl, et al., 2019).

Phylum	Class	Order	Family	Example of species images
Dinoflagellata	Ellobiopsea?			
		Thalassomycetales		
	Oxyrrhea			
		Oxyrrhida		 *Oxyrrhis marina*[1]
	Pronoctiluca			
	Duboscquella? (MALV- I)			
		Duboscquodinida		
	Syndinea			
		Syndiniales		
			Euduboscquellidae	
			Syndinidae (MALV-IV)	
			Sphaeriparaceae	
			Amoebophyridae (MALV-II)	 *Amoebophrya* sp. dinospores [2]
	Noctilucea			
		Noctilucales		 *Noctiluca scintillans*[3]
	Dinophyceae (Dinokaryota)			
		Haplozooidea/ Haplozoonales		
		Dinotrichales		

		Dinococcales		
		Akashiwo		
		Brachidiniales		
		Gymnodiniales s.s.		 Gymnodiniales[4]
		Gymnodiniales s.l.		
		Ptychodiscales		
		Thoracosphaerales		
		Peridiniales s.s.		
		Peridiniales s.l.		
		Peridiniales incertae sedis		
		Actiniscales		
		Amphilothales		
		Prorocentrales		
		Dinophysales		
		Blastodiniales		
		Gonyaulacales		
		Gonyaulacales incertae sedis		
		Uncertain orders		

Ps: [1] https://en.wikipedia.org/wiki/Oxyrrhis#/media/File:Oxyrrhis_marina.jpg

[2]Coats et al 2012.

[3] https://en.wikipedia.org/wiki/Dinoflagellate#/media/File:Noctiluca_scintillans_varias.jpg

[4] https://en.wikipedia.org/wiki/Gymnodiniales

? indicates the uncertainty of taxonomy classification. Both Duboscquella and Ellobiopsea have been placed within Syndiniales in the recent classification by Adl et al., (2019).

A dinokaryon, a modified nucleus containing permanently condensed fibrillar chromosomes, is present in the core dinoflagellates (i.e. Dinokaryota) (Taylor, 1987; Dodge, 1987; Saldarriaga et al., 2004; Fensome, 1993; Lin et al., 2010; Roy and Morse, 2012), but lacking from the lineages Oxyrrhinaceae and the Syndiniales (Taylor et al., 2008; Okamoto et al., 2012). The Noctilucales also lack a dinokaryon during particular life cycle stages. These lineages are thought to be basal to dinokaryotes (Fensome, 1993; Fukuda and Endoh, 2006; Fukuda and Endoh 2008; Ki, 2010).

The arrangement of the thecal plate bearing amphiesmal vesicles is an important character in distinguishing clades of dinoflagellates (Hoppenrath and Leander, 2010; Orr et al., 2012). The thecate orders (Dinophysiales, Gonyaulacales, Peridiniales, Prorocentrales and Suessiales) have comparatively fewer, large amphiesmal vesicles in distinctive patterns, with cellulosic material in the vesicles. Six major tabulation types have been recognized (**Fig** 6), and traditionally used to define higher level taxa in Dinokaryota. These types include the gymnodinioid, suessioid, gonyaulacoid-peridinioid, dinophysioid, nannoceratopsioid and prorocentroid types (Fensome et al., 1999).

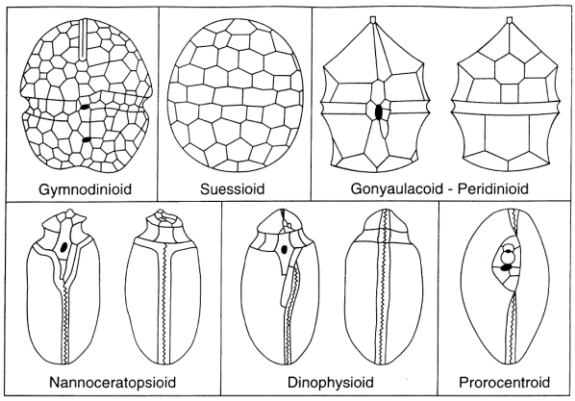


Fig 6. Dinoflagellate tabulation types. Flagellar pores are black. (Fensome et al. 1999)

In contrast, athecate taxa (Gymnodiniales, Noctilucales and Syndiniales) usually contain hundreds of alveoli lacking cellulosic material, and therefore relationships are determined based on other features, such as the presence and shape of grooves on the cell surface or on the cell apex, and the shape of the epicone (Daugbjerg et al., 2000; Takayama, 1985; Jørgensen et al., 2004; Dodge and Crawford, 1968).

2.4 Phylogeny and evolution of dinoflagellates

The monophyly of dinoflagellates and their sister relationships to the Apicomplexa have been established from a number of early phylogenies (Hoppenrath and Leander, 2010; Leander and Keeling, 2004; Saldarriaga et al., 2003; Zhang et al., 2007; Shalchian-Tabrizi et al., 2006; Burki et al., 2008). These studies have been dedicated to infer the dinoflagellate phylogenetic relationships, based on different molecular markers including ribosomal DNA (rDNA) (Daugbjerg et al., 2000; Leander and Keeling, 2004; Murray et al., 2005; Tillmann et al., 2012; Gribble et al., 2006; Saldarriaga et al., 2001; Yamaguchi et al., 2006; Hoppenrath et al., 2009; Shalchian-Tabrizi et al., 2006) and protein-coding genes, such as actin, alpha- and beta-tubulin (Saldarriaga et al., 2003), hsp90 (Hoppenrath and Leander et al., 2010; Shalchian-Tabrizi et al., 2006), and the mitochondrial cytochrome genes (Zhang et al., 2007). However, the phylogenetic relationship between the different dinoflagellate orders, has been a longstanding issue, with a lack of statistical support for the phylogenetic backbone (example shown in **Fig** 7; reviewed in Orr et al., 2012).

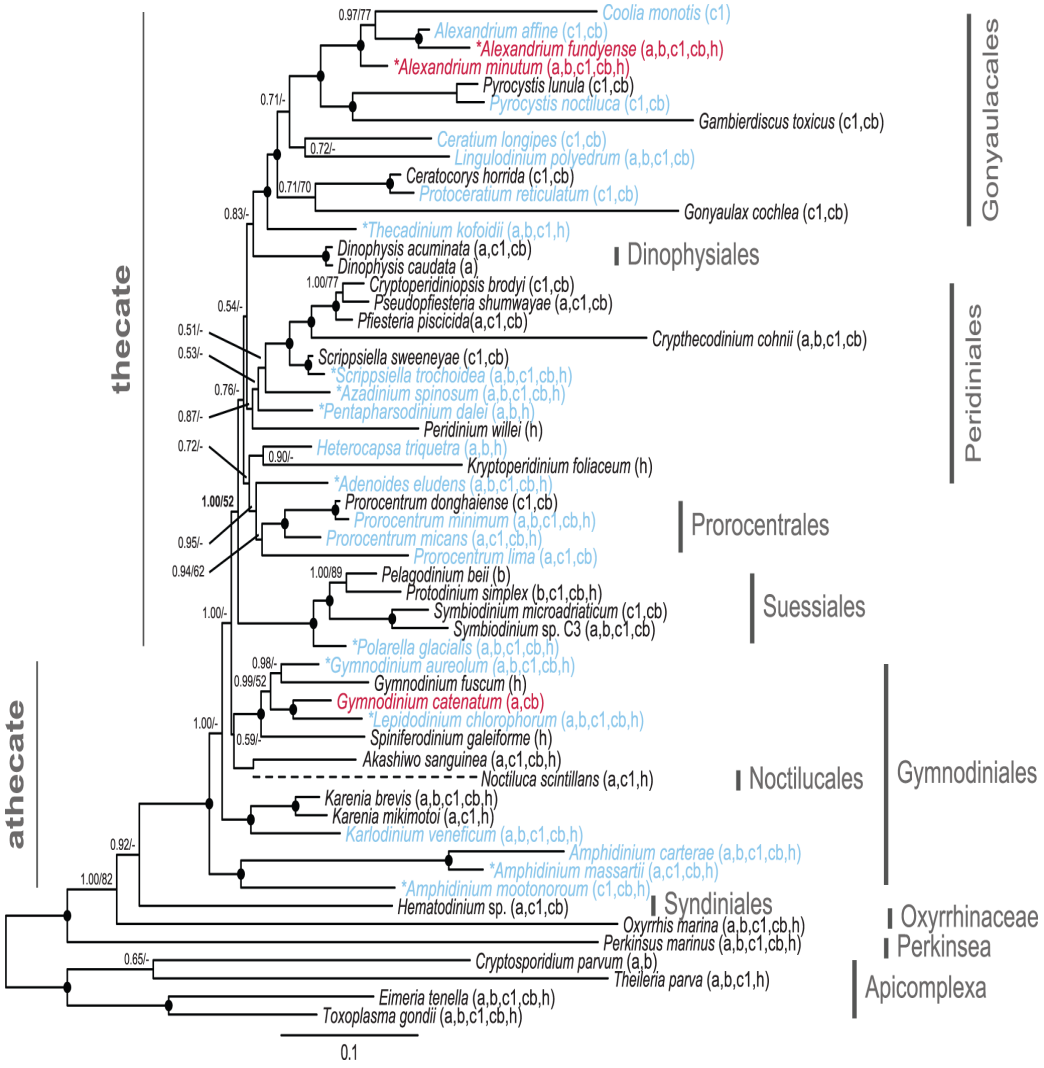


Fig 7. Concatenated phylogeny inferred from 18S+5.8S+28S+cob+cox1+ actin+ beta-tubulin+hsp90 (7138 characters). The tree is reconstructed with Bayesian inference (MrBayes). Numbers on the internal nodes represent posterior probability and bootstrap values (>50%) for MrBayes and RAxML (ordered; MrBayes/RAxML). Black circles indicate a posterior probability value of 1.00 and bootstrap >90%. *N. scintilans* is represented with a dashed branch as this taxon was excluded from the inference; alternatively, its most "probable" placement was determined from a parallel Bayesian analysis. * Denotes taxa sequences generated from this study. Red font indicates sxtA presence and blue font indicates no sxtA detection. Non-ribosomal gene presence for each taxon is represented in brackets behind each species name (a: actin, b: beta-tubulin, c1: cox1, cb: cob, h: hsp90). The phylogenetic support for the thecate/athecate split is highlighted with bold type. (Orr et al 2012)

More recently, molecular phylogenies using concatenated multiple ribosomal proteins have established the deep-branching positions of *Oxyrrhis marina* and the Syndiniales (Bachvaroff et al., 2014). To date, a large dataset of dinoflagellate transcriptomes has clearly demonstrated that the core dinoflagellates are monophyletic and provided the best resolution of internal phylogenetic relationships of dinoflagellates (**Fig** 8, Janouškovec et al., 2017).

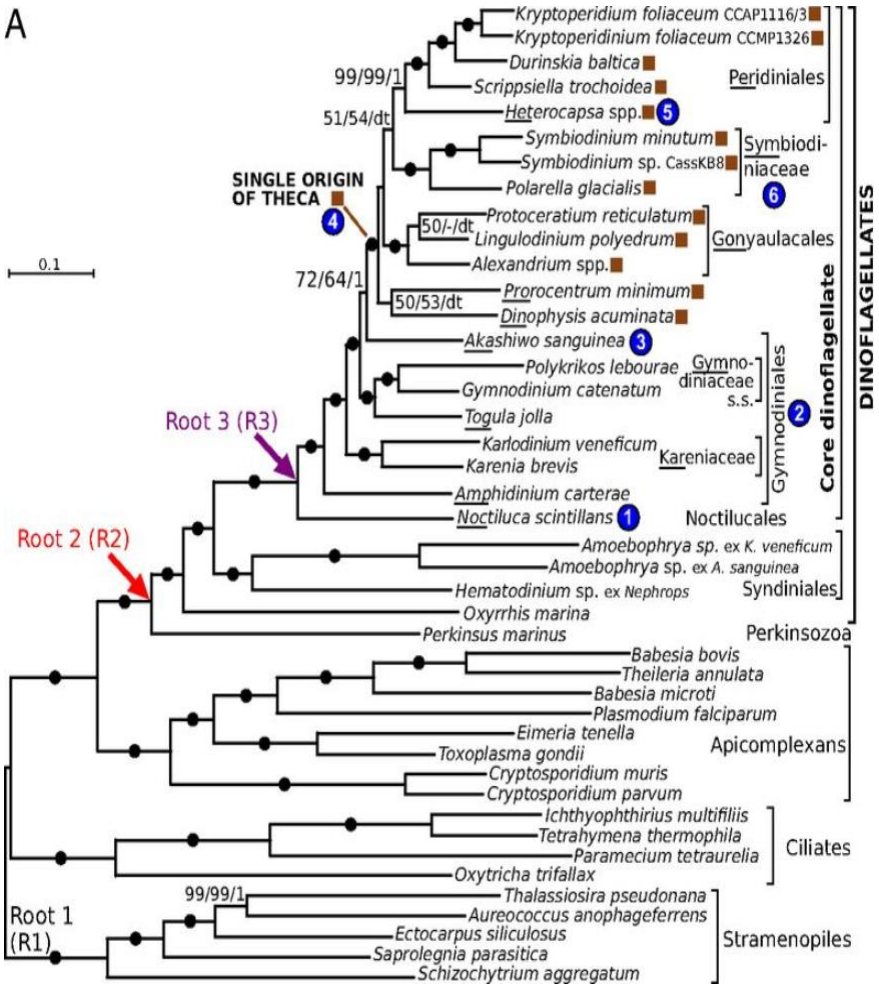


Fig 8. Best maximum-likelihood tree (IQ-Tree) of dinoflagellates and relatives based on 101-protein dataset (root 1 matrix, 43 species, 29,400 sites). Branches show ultrafast bootstraps (IQ-Tree)/nonparametric bootstraps (RAxML)/posterior probabilities (PhyloBayes) (dash indicates <50/50/0.5 support; filled circles indicate 100/100/1 support; dt indicates a different topology). Roots of alternative matrices (Perkinsus, root 2, 30,780 sites; and Noctiluca, root 3, 30,988 sites) are shown by arrows. (Janouškovec et al 2017)

Using a taxonomically representative dataset of dinoflagellate transcriptomes, Janouškovec et al. (2017) inferred a strongly supported phylogeny to map major morphological and molecular transitions in dinoflagellate evolution. The results showed an early branching position of *Noctiluca*, monophyly of thecate dinoflagellates, and paraphyly of athecate ones (**Fig** 8B), which provided unambiguous phylogenetic evidence for a single origin of the group's cellulosic theca (**Fig** 8B). It's suggested that all living thecate dinoflagellates originated from ancestors with a gonyaulacoid–peridinoid tabulation (**Fig** 9A, Janouškovec et al., 2017). And also the late acquisition of dinosterol in the group is inconsistent with dinoflagellates being the source of this biomarker in pre-Mesozoicstrata (**Fig** 9C; Janouškovec et al., 2017). Three distantly related non-photosynthetic dinoflagellates, *Noctiluca*, *Oxyrrhis*, and *Dinophysis*, contain cryptic plastidial metabolisms, suggesting that all free-living (but not all parasitic) dinoflagellates metabolically rely on plastids, which are very likely derived from the ancestral peridinin plastid (**Fig 10D**). It's also suggested that the evolutionary origin of bioluminescence in

nonphotosynthetic dinoflagellates may be linked to plastidic tetrapyrrole biosynthesis. Finally, dinoflagellate nuclei may have recruited DNA-binding proteins in three distinct evolutionary waves, which included two independent acquisitions of bacterial histone-like proteins (**Fig** 11; Janouškovec et al., 2017).

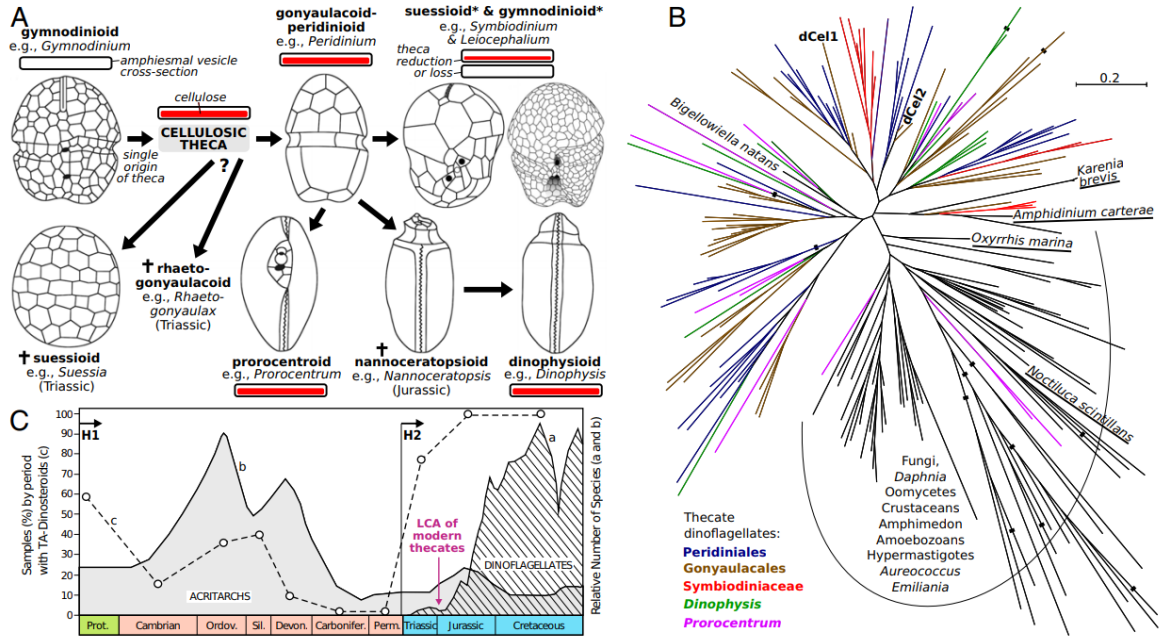


Fig 9. Thecal evolution and dinoflagellate paleohistory. (A) Phylogeny-driven model of changes between major modern and fossil (crosses) tabulational types. Gymnodinoid tabulation with numerous small, empty amphiesmal vesicles is ancestral and gave rise to the gonyaulacoid–peridinioid tabulation with a few large, cellulose-rich thecal plates. Suessioid and gymnodinioid tabulations in modern Symbiodiniaceae and Borghiellaceae (asterisk) are derived independently of the standard gymnodiniod and Triassic suessioid tabulations (Suessia), and are characterized by decrease or loss of cellulose content. Prorocentroid and dinophysioid tabulations are derived from the gonyaulacoid–peridiniod tabulation (the latter probably via a nannoceratopsioid intermediate). Triassic suessioid and rhaetogonyaulacoid tabulations may represent evolutionary intermediates or independent experiments in thecal plate reduction. (B) Maximum-likelihood phylogeny (IQ-Tree) of 184 eukaryotic GH7 proteins reveals cellulases in athecate dinoflagellates (underlined) and their radiation in the thecate (color-coded). Black rectangles indicate 50% reduction in branch length. Known GH7 cellulases in P. lunula (dCel1) and Lingulodinium polyedrum (dCel2) are shown. (C) Alternative hypotheses (H1 and H2) on the first emergence of triaromatic dinosteranes attributable to dinoflagellates or their direct ancestors. Relative species numbers of dinoflagellates (a) and acritarchs (b) and percentage of dinosterane-positive samples from the Proterozoic (green), Paleozoic (red), and Mesozoic (blue) are shown together with the predicted emergence of the last common ancestor (LCA) of modern thecates. (Janouškovec et al., 2017)

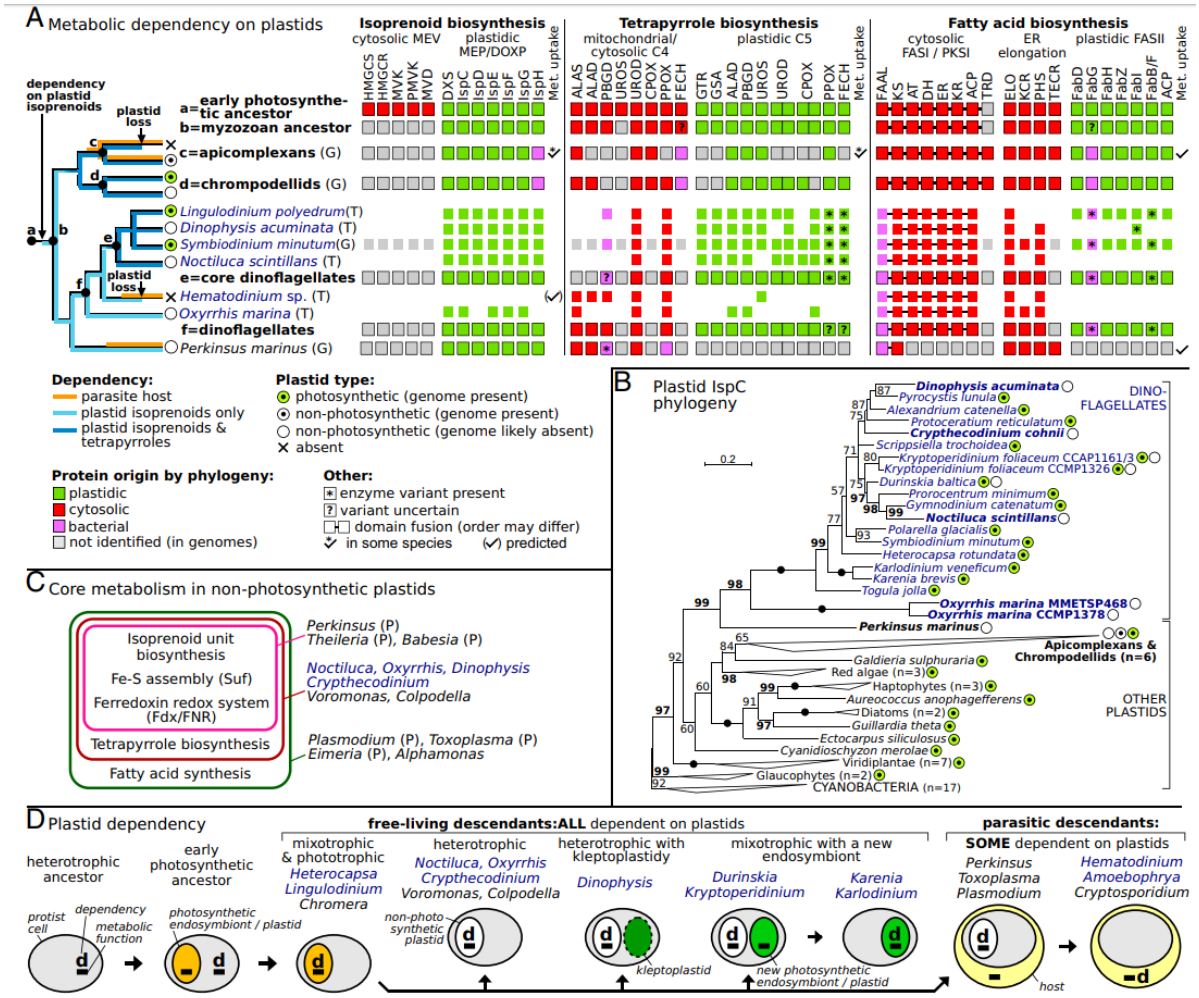


Fig 10. Plastid metabolism and dependency in nonphotosyntetic dinoflagellates. (A) Phylogeny-driven reconstruction of plastid and nonplastid variants of core metabolism (isoprenoid, tetrapyrrole, and fatty acid biosynthesis) in genomes (marked as "G") or transcriptomes ("T") of dinoflagellates and relatives. Individual enzymes were classified by protein phylogenies and color-coded as to their presence/absence and origin. The data suggest that Oxyrrhis, Noctiluca, and Dinophysis are metabolically dependent on plastids. (B) Maximum-likelihood phylogeny (IQ-Tree) reveals IspCs of cyanobacterial origin in nonphotosynthetic dinoflagellates and relatives (bold); ultrafast bootstraps at branches are shown (>50 shown; ≥95 highlighted; filled circles, 100). (C) Three grades in functional organization of core metabolic pathways in nonphotosynthetic plastids in dinoflagellates (blue) and relatives ("P" represents parasites). (D) Model for evolutionary dependency on plastids in dinoflagellates and relatives. Ancestral dependency (marked as "d") on plastid metabolism (loss of cytosolic isoprenoid biosynthesis; later reinforced by the loss of C4 tetrapyrrole biosynthesis in some taxa) led to retention of plastids in all free-living and many parasitic descendants. The dependency can be transferred onto a new plastidial symbiont (Kareniaceae) or host organism (in parasites dependent solely on host-derived metabolites); only the latter leads to an outright loss of the plastid. (Janouškovec et al., 2017)

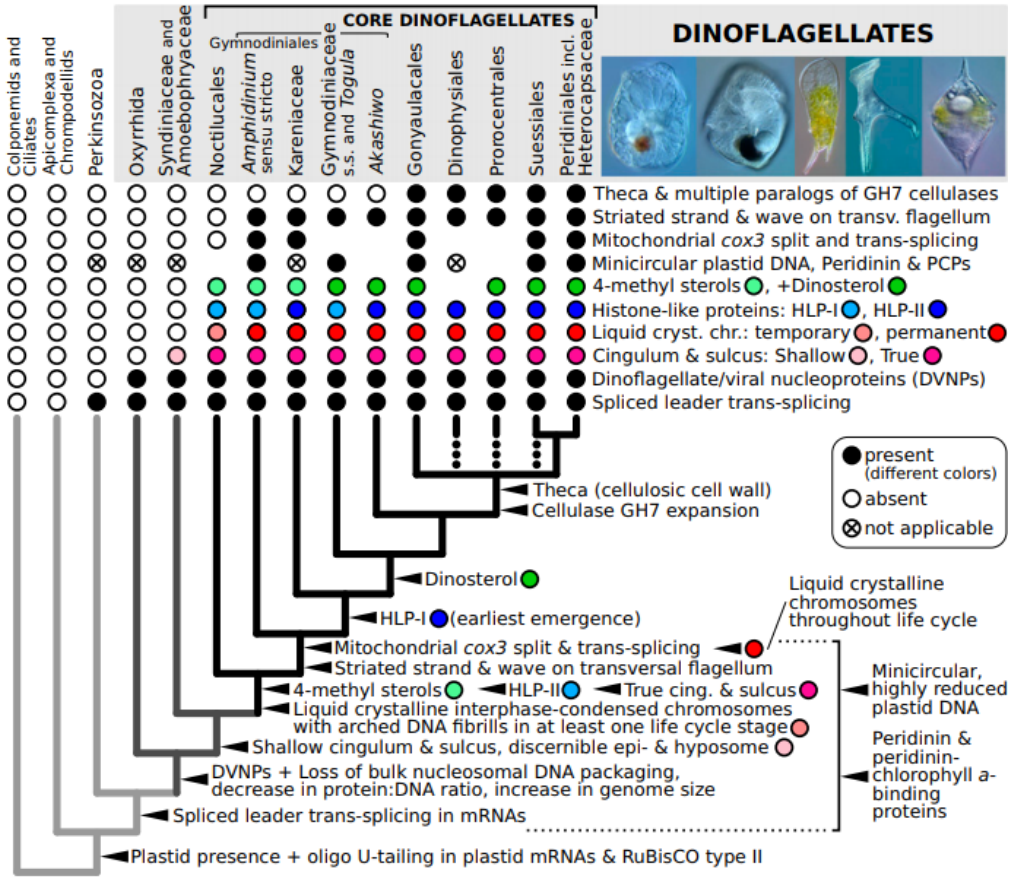


Fig 11. Model for character evolution in dinoflagellates. Ancestral character states (filled circles) of conserved traits are reconstructed on the consensus phylogeny of dinoflagellates and their relatives by parsimony (arrowheads). Dotted branches in the thecate lineages indicate uncertain placement. Gaps indicate missing data, and "not applicable" denotes plastid genome absence or the presence of a different plastid genome type (Kareniaceae). The vertical square bracket indicates an evolutionary range in which traits emerged. Photos of dinoflagellates, left to right: *Kofoidinium* sp. (Noctilucales), *Nematodinium* sp. (Gymnodiniaceae s.s.), *Neoceratium praelongum* (Gonyaulacales), *Dinophysis miles* (Dinophysiales), and *Heterocapsa* sp. (Peridiniales). (Janouškovec et al., 2017)

Taken together, the phylogeny of dinoflagellates has been greatly advanced and showed a better view of dinoflagellate evolution. However, Syndinea mostly occurred as a small group with few branches in the phylogeny of dinoflagellates. The syndinian sequences were generally from few and only representatives of syndinian genera, *Amoebophrya* and *Hematodinium*.

2.5 Life cycles of dinoflagellates

Generally, dinoflagellates spend most of their life cycles as haploid cells (i.e. a haplontic life cycle) and proliferate by mitosis (Von Stosch, 1973). Mitosis in dinoflagellates shows distinctive features than a typical one observed in most higher eukaryotes, including endomitosis and nucleolar disassembly (i.e the nuclear envelope and nucleoli remain complete throughout the whole cell division process) (Soyer-Gobillard and Geraud, 1992). Population growth of dinoflagellate normally occurs through asexual division, which allows for rapid proliferation and thus leads to dense blooms.

Sexual reproduction also occurs in dinoflagellates. But clearly established sexual fusion has been observed only in a small proportion of dinoflagellates so far (von Stosch 1973; Pfiester and Anderson, 1987), although sexual reproduction is probably widespread (gametes resemble regular motile cells, and fusion occurs at night in photosynthetic species) (Hoppenrath and Saldarriaga, 2008). Sexual fusion (syngamy, **Fig** 12A) may involve equal (isogamy) or unequal (anisogamy) motile gametes. Both homothallism (gamete fusion in clonal strains) and heterothallism (no fusion in clonal strains) are known to occur in dinoflagellates (Hoppenrath and Saldarriaga, 2008). Also known is complex heterothallism with more than two sexual types in some dinoflagellate species like *Alexandrium* spp. (Figueroa et al., 2007). In homothally, the gametes can be genetically identical such as in *Alexandrium taylori* (Giacobbe and Yang, 1999). In heterothally, genetically determined factors in the gametes allow successful mating, sexual fusion and meiosis. In the latter case, the sexual compatibility can comprise only two different mating types (simple heterothallism), such as in *Lingulodinium polyedrum* (Figueroa and Bravo, 2005) or more mating types (complex heterothallism), such as in *Alexandrium minutum* (Figueroa et al., 2007). However, different mating systems can occur within the same species, and a continuum between homothally and heterothally has been observed for example in the species *Gymnodinium catenatum* (Figueroa et al., 2010).

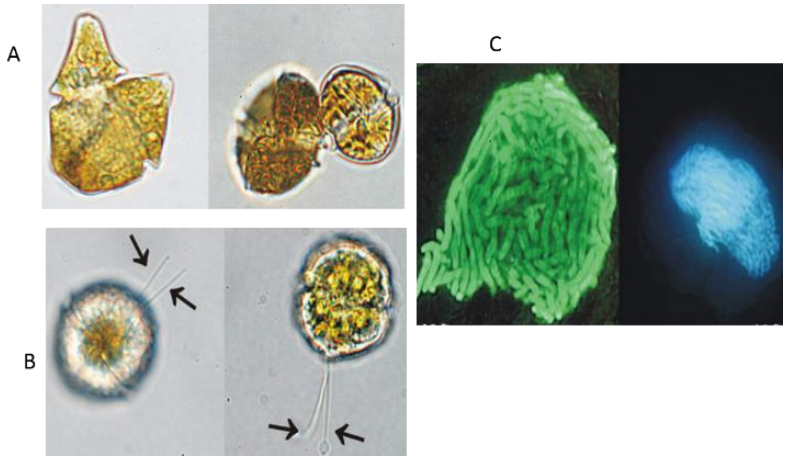


Fig 12 (A) Isogamous (left, *Gymnodinium nolleri*) and anisogamous (right, *Alexandrium tamutum*) gamete pairs. (B) Examples of planozyotes, which are characterized by two longitudinal flagella (arrows) instead of one, in *Alexandrium minutum* (left) and *Alexandrium taylori*. (C) Planozygote nuclei in *G. nolleri* (left) and *A. minutum* (right, undergoing meiosis). (Hoppenrath and Saldarriaga, 2008)

Sexuality can be induced in dinoflagellates by endo- and exogenous factors, and in many cases, results in a diploid resting cyst with environmental resistance and dispersal functions (Pfiester and Anderson, 1987). During sexual reproduction, 2 gametes fuse and form a diploid mobile zygote called planozygote (**Fig** 12B), which may remain motile for hours or a few days. The zygote may later enter a resting stage and form a nonmotile thick-walled hypnozygote then called a cyst or dinocyst. Either a cyst or a planozygote is able to start the process of meiosis and produce new haploid cells (von et al., 2011;

Figueroa et al., 2018; Cuadrado et al., 2019). Excystment of the resting cyst occurs after a varying length of time of dormancy, which lasts from hours to days or months and is species-specific (e.g. 50 days for *Alexandrium taylori*, 1 month for *Alexandrium minutum* or only days for *Kryptoperidinium foliaceum*) (Hoppenrath and Saldarriaga, 2008). Meiosis in resting cysts or in planozygotes (**Fig 12C**) is heralded by a peculiar swirling and rotation of the nucleus, a process termed nuclear cyclosis associated with the pairing of homologous chromosomes (von Stosch, 1972). Meiosis may precede or follow excystment and is normally accomplished in two conventional, successive divisions. The complete life cycle combining sexual phase and asexual phase is summarized in **Fig** 13.

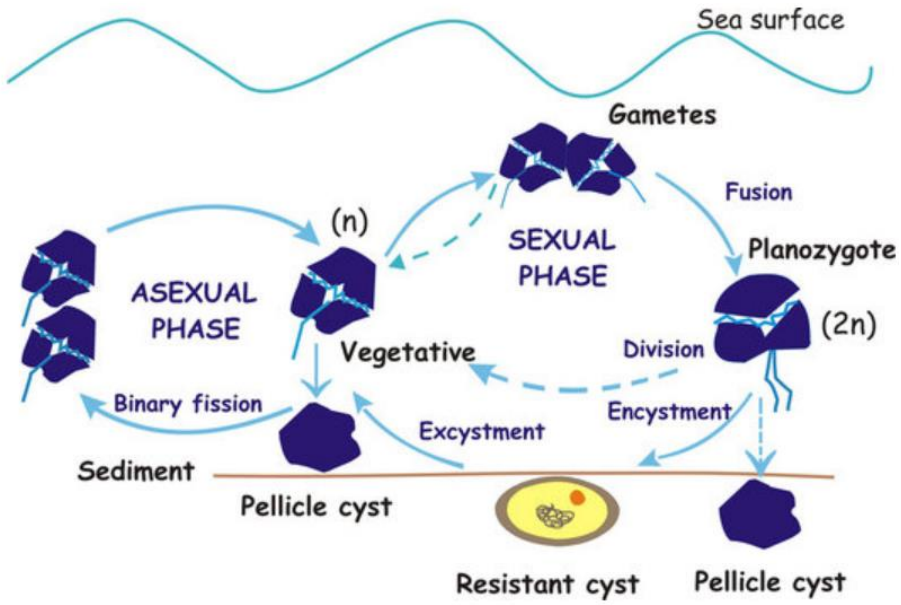


Fig 13. Dinoflagellate life cycle. (Hoppenrath and Saldarriaga, 2008)

Dinoflagellate life cycles can be much more complicated than described above. For instance, Figueroa et al. (2006) found multiple routes of sexuality in *Alexandrium taylori*. The mobile zygotes can follow three different routes (Hoppenrath and Saldarriaga, 2008): direct division by desmoschisis, short-term encystment (temporary, pellicle or ecdysal cysts), and long-term encystment (resting cysts). At least 10% of all known dinoflagellates are resting cyst producers (Persson et al., 2000). Sexual reproduction is thought to be essential for seasonal survival of these species, although asexual resting cysts are also known in *Scrippsiella hangoei* (Kremp and Parrow, 2006). Dinoflagellate cysts in the sediment can provide the inoculum for future blooms after a long time. Dormancy and maturation of resting cysts are biological processes essential for the dinoflagellate population (Hoppenrath and Saldarriaga, 2008).

Traditionally, the presence of sexual reproduction in dinoflagellates is mostly determined by the test for meiosis-related morphological features, such as the pairing of 2 cells (mating gametes) and the presence of resting cysts (Figueroa et al., 2018; Persson et al., 2013). However, vegetative cells, gametes and planozygotes are morphologically very similar and therefore difficult to study individually, which to

some extent sets an obstacle to a broader study of the sexual life stage in dinoflagellates (Cuadrado et al., 2019).

2.6 Genomics

Dinoflagellates generally have dramatically large genomes, which range from 1.5 to 112 Gbp in size (Wisecaver and Hackett, 2011; Murray et al., 2016) and are comparable to or even larger than those of higher organisms, such as human and the hexaploid Triticum wheat (3.2 Gb and 17 Gb respectively, from: https://plants.ensembl.org/Triticum_aestivum/Info/Annotation/).

Gene expression in dinoflagellates involves trans-splicing of mRNAs through the addition of a 5'-end splice leader sequence (Lidie and Van Dolah, 2007; Zhang et al., 2007). These trans-spliced mRNA harbor unusual GC/GA dinucleotide pairs at their 5' donor splice site (Shoguchi et al., 2013). In terms of organelles, dinoflagellate plastid genomes occur as plasmid-like minicircles (Zhang et al., 1999). Even in photosynthetic species the plastid genome is highly reduced and fragmented (14 genes as compared to a typical plastid genome containing more than 100 genes), as most plastid genes have been transferred to the nucleus (Janouškovec et al., 2017). The mitochondrial genomes of dinoflagellates typically harbor only three protein-coding genes and fragments of rRNA genes (Jackson et al., 2012; Waller and Jackson, 2009), which represent the minimal mitochondrial genomes in aerobic species (Flegontov et al., 2015). Furthermore, both mitochondrial and nuclear transcripts are extensively edited (Liew et al., 2017; Liu et al., 2018).

The genomes of parasites are often characterized by a reduction in size, loss of genes, and loss of functions due to the dependence on the host. It turns out the genome sizes of Syndiniales (hundreds of Mb, John et al. 2019) are much smaller than that of a typical dinoflagellate genome (the smallest ones are about 1.5 Gb). The reduced genomes of Syndiniales occur with the loss of endosymbiotic organelle genomes (either mitochondria or plastids). The syndiniales *Hematodinium* sp. and *Amoebophrya* spp. represent a rare case of plastid elimination among Myzozoan (Gornik et al., 2015; John et al., 2019). The ancestral mitochondrial genome of dinoflagellates is already highly reduced as mitosomes (Waller and Jackson, 2009) while one *Amoebophrya ceratii* strain appeared to have lost its mitochondrial genome completely but still have its functional mitochondria (John et al., 2019).

Next-generation sequencing technologies have made genomic sequences available for some dinoflagellates such as *Symbiodinium* spp. [e.g. *S. kawagutii* (Lin et al., 2015), *S. minutum* (Shoguchi et al., 2013), *S. microadiaticum* (Aranda et al., 2016)] and parasitic forms [e.g. *Hematodinum sp.* (Gornik et al., 2015) and *Amoebohpyra sp.* (John et al., 2019)], which contributes to a better understanding of the biology and evolution of these enigmatic organisms. For example, by sequencing the whole genome of *S. kawagutii*, Lin et al. (2015) found the evidence for gene family expansion involved in processes important for successful symbiosis with corals. The microRNA system potentially regulating gene expression in both symbiont and coral and the biochemical complementarity between

genomes of *S. kawagutii* and the anthozoan *Acropora*, together provide insights into genome evolution and regulation of gene expression in dinoflagellates and the molecular basis of coral-*Symbiodinium* symbiosis. Adopting a comparative approach, Liu et al. (2018) identified genes and functions with evidence of adaptive selection in *Symbiodinium* and genes for meiosis and response to light stress. These results indicate adaptive selection in *Symbiodinium* gene functions that are related to the establishment of symbiosis, and provide genomic evidence that *Symbiodinium* is capable of meiosis (based on gene repertoire).

However, the functional capacity of dinoflagellate genes is poorly understood when relying on the commonly used annotation approach, whereby predicted proteins are compared against a set of curated proteins with known functions that are largely derived from model organisms. Whereas genome data from dinoflagellates are limited and difficult to deal with even when available, transcriptome data provide an avenue for the exploration of genes with unknown function, so-called dark genes (Stephens et al., 2018; Meng et al., 2018). In this way, researchers showed that dark genes account for a substantial proportion (15%-64%) of overall gene numbers and are prevalent in the investigated dinoflagellate taxa (**Fig** 14b, Stephens et al., 2018). Also, these dark genes are largely lineage-specific and discovered in multiple taxa and may represent lineage-specific features (Stephens et al., 2018).

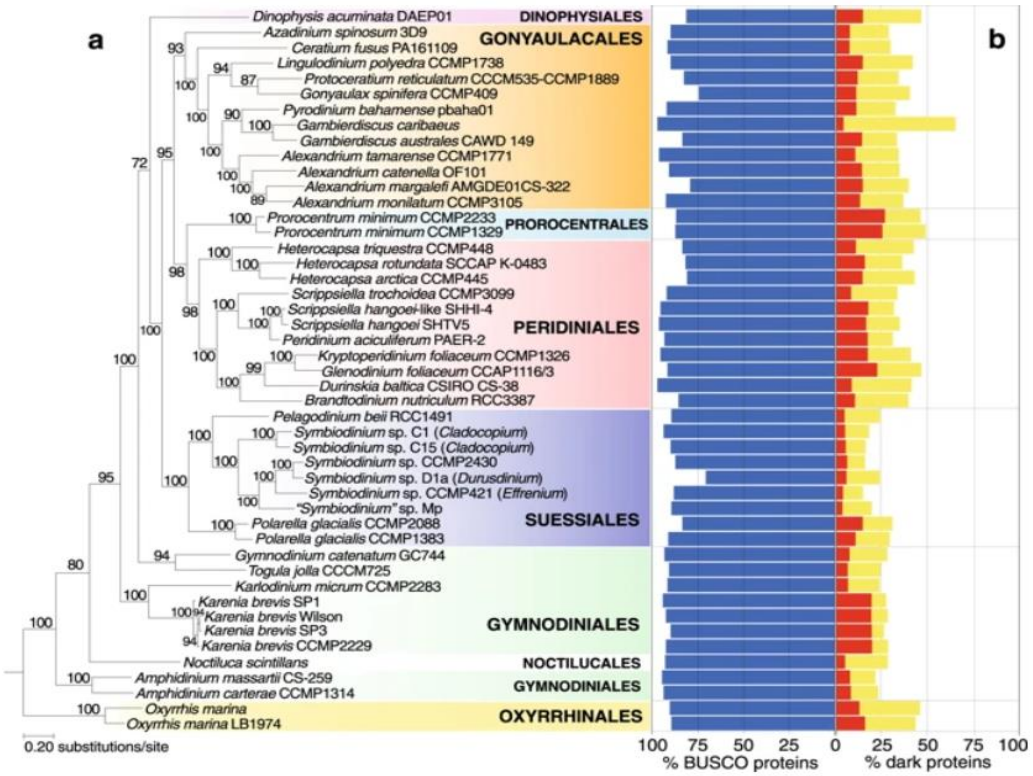


Fig 14. (a) Maximum-likelihood phylogeny inferred using the 1043 orthologous protein sets. Support values, based on 2000 ultrafast bootstrap approximations, are shown at the internal nodes. The unit of branch length is the number of substitutions per site. (b) The percentage of recovered alveolate + stramenopile BUSCO (Benchmarking Universal Single-Copy Orthologs) proteins and of dark proteins in each dataset. High- and low-confidence dark proteins are shown in red and yellow bars, respectively. (Stephens et al., 2018)

3 The parasite in dinoflagellates: Syndiniales

3.1 Taxonomy and phylogeny of Syndiniales

The Syndiniales are an order of early branching dinoflagellates (**Fig** 15; Strassert et al., 2018). This order is known to include exclusively parasites, which target hosts covering fish eggs, crustaceans, algae, cnidarians, and protists (including ciliates, radiolarians, and other dinoflagellates) (van den Hoek et al., 1995; Bråte et al., 2012; Strassert et al., 2018). The latest estimate reported 11 genera and 46 species exist in this order (Gómez, 2012), among which the genera *Hematodinium*, *Syndinium*, *Amoebophrya* are better known and studied.

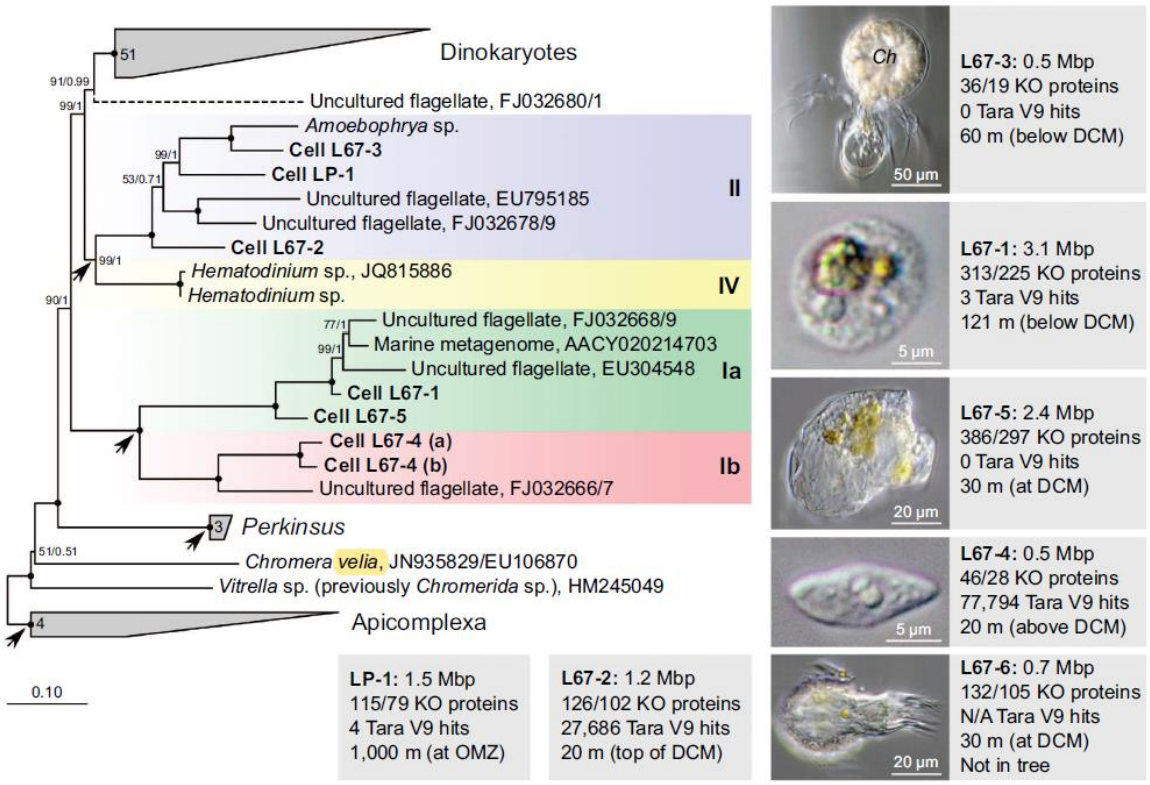


Fig 15. Morphology and phylogenetic relationships of MALVs. Tree topology is based on maximum likelihood analysis of concatenated SSU and LSU rRNA gene sequences (43900 nucleotide positions). Node support is shown by RAxML bootstraps and MrBayes posterior probabilities. Black circles indicate maximum support in both analyses. Numbers in polygons refer to the number of grouped taxa, roman numerals signify the different MALV groups (compare with Guillou et al., 2008), and the arrows indicate parasitic clades. (Strassert et al., 2018)

Marine alveolate (MALV) lineages, have been discovered in marine planktonic communities (Díez et al., 2001; Moon-van der Staay et al., 2001; Lopez-garcia et al., 2002) and are restricted to marine waters. Several MALV groups (e.g. MALV groups I and II) have been presumably assigned to Syndiniales according to the close relationship with few formally described species previously sequenced (such as the genera *Amoebophrya*, *Syndinium* and *Hematodinium*) revealed by 18S rRNA gene phylogeny (**Fig** 16, Guillou et al., 2008). In particular, MALV-I corresponds to *Euduboscquella*, MALV-II to

Amoebophrya and MALV-IV to the genera *Syndinium* and *Hematodinium* (Harada et al., 2007; Guillou et al., 2008). Wide diversity and occurrence of these organisms in marine waters led to a large increase in the number of MALV sequences deposited in GenBank. However, it is still unknown how to delimit the species boundaries within the newly discovered lineages. For example, MALV-II is believed to be synonymous to the family Amoebophryidae. This lineage groups more than 44 genetic clusters based on the 18S rRNA gene, and each cluster is subdivided into several sub-clusters (Guillou et al., 2008).

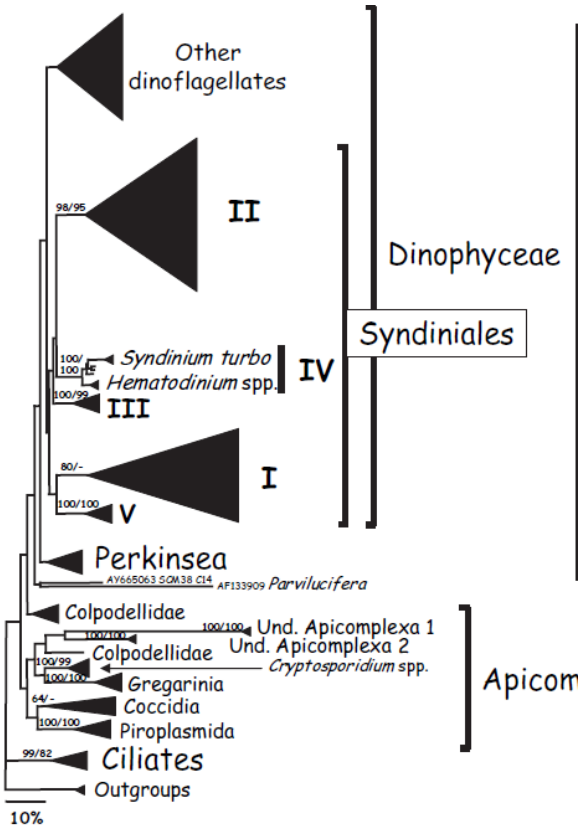


Fig 16. Phylogeny of dinoflagellates using near-complete 18S rRNA gene sequences. Bayesian phylogeny of alveolates based on the analysis of 291 near full-length 18S rRNA gene sequences. Five sequences of Bolidophyceae (stramenopiles) were used as outgroup. Bootstrap values, given at the principal nodes of the tree, correspond to neighbour-joining and maximum parsimony analyses respectively (1000 replicates, values >60% shown). The scale bar corresponds to 10% sequence divergence. (Guillou et al., 2008)

3.2 Biology of Syndiniales

Syndiniales lack several features common to dinokaryotes although they are the closest lineage to dinokaryotes. For example, unlike all other orders in dinoflagellates, Syndiniales lack a theca and chloroplasts (or plasts), and their nucleus is never a dinokaryon (described above). Their chromosomes remain attached to the nuclear membrane and form a "V" shape, which is characteristic of Syndiniales. The chromatin is packed into a low number of chromosomes (5-7) compared to the approximately 20-270 chromosomes found in free-living dinoflagellates (Cachon, 1964). Syndiniales have DVNPs. They have two flagella which resemble dinoflagellates during the free-living stages.

3.3 Ecology of Syndiniales

All Syndiniales known to date have a parasitic lifestyle and are parasitoids (*i.e.* obligatorily killing their host to complete their life cycle). They are all biotrophic, meaning that they keep their host living during infection. Most of the representatives infecting dinoflagellates belong to the species complex

Amoebophrya ceratii, which have the potential to end red tides (reviewed in Kim, 2006; Coats, 1999). It is demonstrated that such parasites are able to thrive and infect a wide range of dinoflagellate species both in coastal and ultra-oligotrophic open waters (Siano et al., 2011).

A recent large-scale investigation on marine planktonic protist diversity in global oceans highlighted the ubiquitous occurrence and remarkable diversity of parasites in marine environment (de Vargas et al., 2015). The interaction networks (**Fig** 17) constructed from metabarcoding datasets showed most of the parasite associations involve the Syndiniales MALV-I and MALV-II groups, both associated with zooplankton and to a lesser extent, with micro-phytoplankton (Lima-mendez et al., 2015). This emphasizes Syndiniales as major top-down drivers of plankton population structure and functioning.

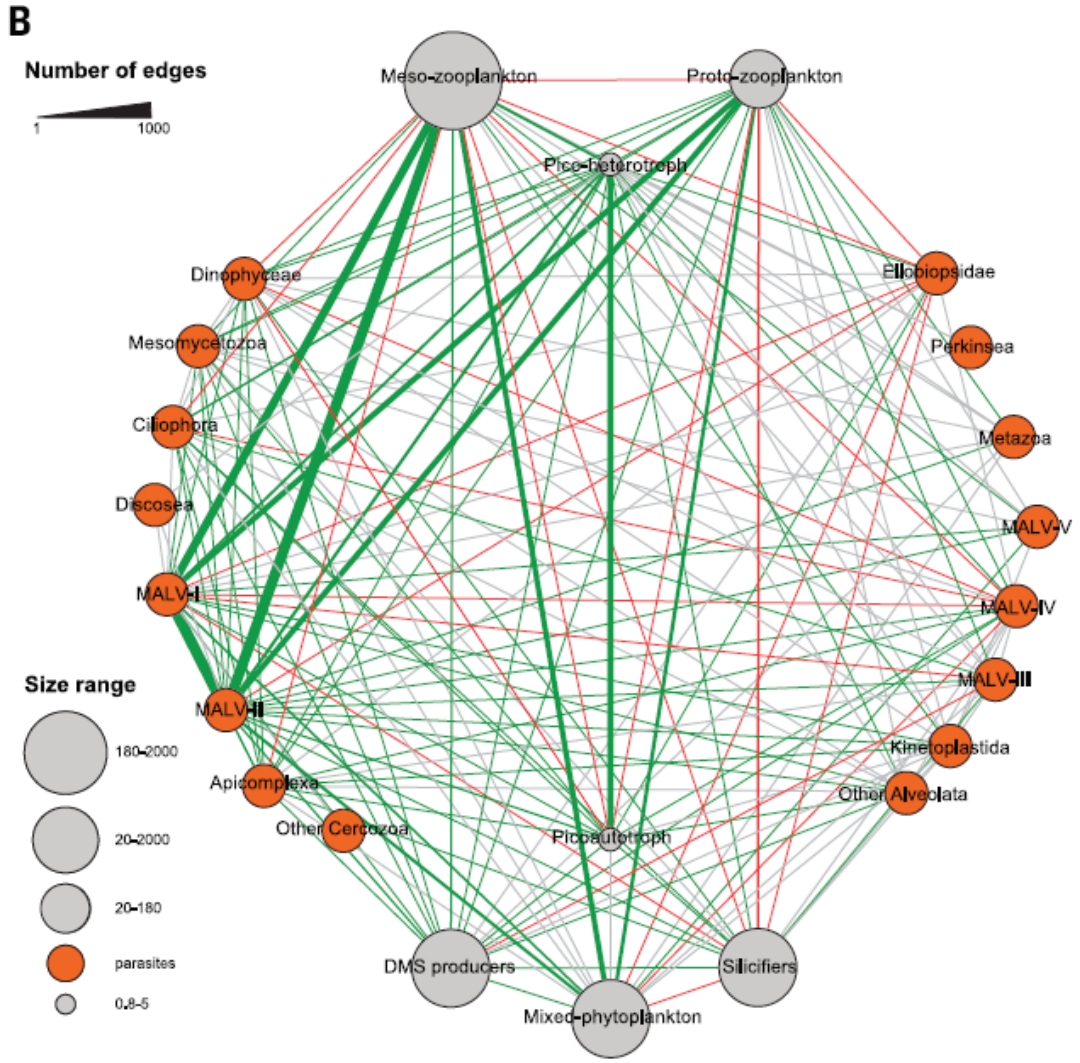


Fig 17. Top-down interactions in plankton. Subnetwork of metanodes that encapsulate barcodes affiliated to parasites or PFTs (plankton functional types). The PFTs mapped onto the network are: phytoplankton DMS producers, mixed phytoplankton, phytoplankton silicifiers, pico-eukaryotic heterotrophs, proto-zooplankton and meso-zooplankton. Edge width reflects the number of edges in the taxon graph between the corresponding metanodes. Over-represented links (multiple-test corrected $P < 0.05$) are colored in green if they represent copresences and in red if they represent exclusions; gray means non-overrepresented combinations. When both copresences and exclusions were significant, the edge is shown as copresence. (Lima-mendez et al., 2015)

3.4 Life cycles of Syndiniales

The asexual cycle of marine parasitic dinoflagellates typically involves a biflagellate zoospore (i.e. dinospore) that is widely accepted as the infective agent, a growth stage (i.e. trophont) that rarely resembles a dinoflagellate in gross appearance, and a sporogenic reproductive phase that generates large numbers of dinospores. In Syndiniales, growth of the trophont is accompanied by nuclear division with little or no cytoplasmic fission, thus forming a multinucleate cell (e.g. *Amoebophrya* infecting dinoflagellates) or a plasmodium/plasmodia (e.g. *Syndinium* and *Hematodinium* in crustacea). At maturity, these multinucleate trophonts undergo cytokinesis, perhaps accompanied by additional nuclear divisions, to produce dinospores (reviewed in Coats, 1999).

The life cycle of *Amoebophrya sp.* infecting the dinoflagellate *Prorocentrum triestrinum* has been described in detail (**Fig** 18; Guillou et al., 2010): Infection initiates when a small flagellate cell called dinospore, penetrates inside the host cell. Usually, only one dinospore develops per host cell, but multiple infections are frequently observed in culture. The trophont (the feeding stage) first settles inside the nucleus in most cases. The dinospore thus needs to cross host membranes systems (external cell wall and the nuclear membrane) to initiate its maturation. The host nucleus is then rapidly digested, and the trophont extends to the cytoplasm. Intracellular trophont usually matures within 2-4 days into a multicellular free-living, 'vermiform' stage. This vermiform stage is released into the seawater and all attached cells have a synchronous swim. Hundreds of new flagellates are then liberated in hours, each cell of the vermiform becoming a dinospore. Dinospores survive only few days when released into water (Coats and Park, 2002).

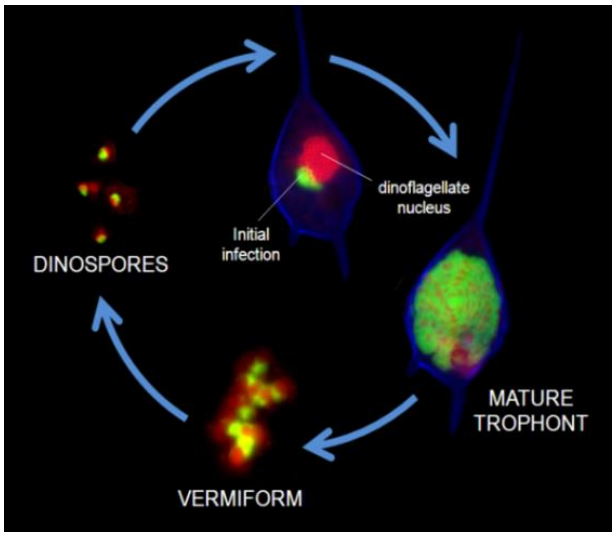


Fig 18. Life cycle of *Amoebophrya* (Syndiniales) infecting the dinoflagellate *Prorocentrum triestrinum*. Green: cytoplasm of the parasite stained by FISH. Red: nuclei stained by propidium iodide. Blue: Host theca stained by calcofluor. (Guillou et al, 2010)

Sporulation in some Syndiniales species results in the formation of either large or small dinospores (macrospores and microspores, respectively), with only one type arising from each host individual (Coats, 1999). The existence of morphologically distinct spore-types has long been considered as evidence of a sexual cycle (Cachon and Cachon, 1987). Recently, Coats et al. (2012) reported cell fusion

(syngamy) in *Eduboscquella sp*. between two different spore types, followed by successive division into four daughter cells, suggesting that sexual reproduction occurs.

Encystment is a survival strategy to contend with periods of low host abundance. However, cysts or cyst-like stages have only been reported for one syndidian species *Duboscquella cachoni* to date (Coats, 1988). Intriguingly, the interplay between the parasite *Amoebophrya sp*. and the resting cyst of *Scrippsiella trochoidea* has been observed, in which parasite and host simultaneously enter dormancy, emerging months later to propagate both species (**Fig** 19; Chambouvet et al., 2011).

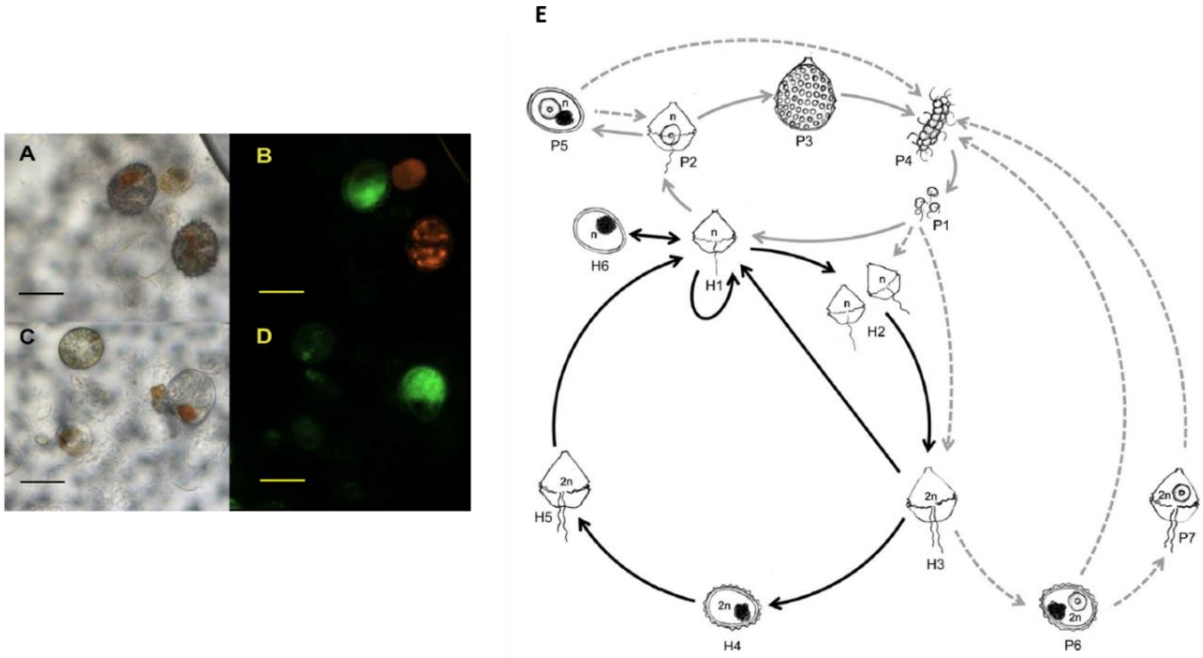


Fig 19. *Amoebophrya* sp. infecting *Scrippsiella trochoidea* cysts. A-B. Calcified cysts. The resting cyst with green autofluorescence at the left is infected. C-D. Infected non-calcified cysts. A and C: Cells observed under white light (phase contrast). B and D. Same cells observed under blue light excitation (parasite revealed by its natural green autofluorescence). Scale bars = 20 um. (E) Interactions between *Scrippsiella trochoidea* and *Amoebophrya* sp. life cycles. Black arrows indicate *S. trochoidea* life cycle with haploid vegetative cells (H1), gametes (H2), diploid planozygote (H3), diploid calcified resting cyst (H4), diploid planomeiocyte (H5) and haploid non-calcified cyst (H6). *Amoebophrya* sp.life cycle (lines in grey) with free-living stage of the parasite (dinospores, P1), able to infect vegetative cells of *S.trochoidea* (P2), mature trophont of *Amoebobophrya* (typical beehive stage, P3), and the vermiform stage (P4). The parasite was additionally detected in non-calcified (P5) and calcified cysts (P6) of its host. Dotted lines illustrated uncertain routes for the parasite. For example, infected non-calcified (P5) and calcified (P6) cysts eventually give rise to infected vegetative cells (P2) and infected planomeiocyte (P7) respectively or directly to the vermiform stage (P4) and dinospores (P1). (Chambouvet et al., 2011)

3.5 Genomics

In Syndiniales, the *Hematodinium sp.* was the first to be sequenced for genomes (partial). It's reported to have a drastically reduced mitochondrial genome that contains only three protein-coding genes and two ribosomal RNA (rRNA) genes, which is similar to Apicomplexa (Jackson et al., 2012). Further sequencing on *Hematodinium* transcriptomes and genome revealed that the parasite likely has lost the plastid organelle (Gornik et al., 2015). This independence on known plastid functions has been achieved

through retention of ancestral anabolic pathways, enzyme relocation from the plastid to the cytosol, and metabolic scavenging from the parasite's host (**Fig** 20).

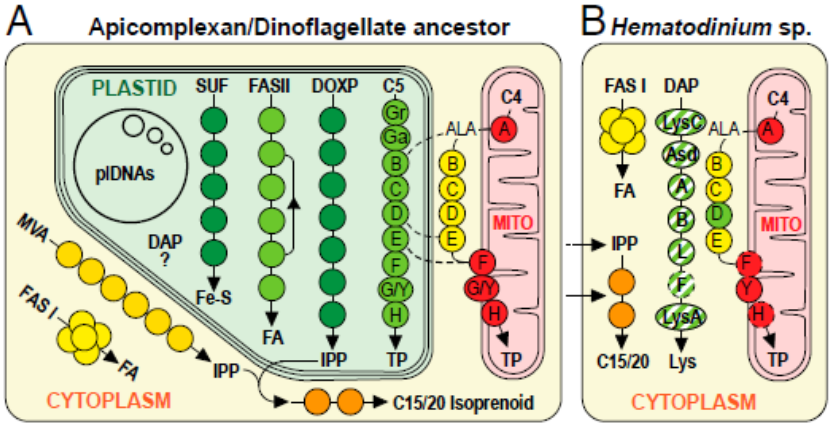


Fig 20. Reconstructed metabolic pathways in the common ancestor of apicomplexans and dinoflagellates at the time of plastid gain (A) and in *Hematodinium* sp. (B) from transcriptomic data. Cytosolic MVA pathway is not present in any apicomplexan or dinoflagellate, but is present in ciliates and is inferred to be present at the time of plastid gain. Dashed lines indicate apicomplexan hybrid tetrapyrrole pathways; dashed circles indicate enzymes currently unidentified in transcriptomes. Enzyme color represents typical location and origin as follows: green, plastid; yellow, cytosol; red, mitochondrion. Hatched (green/white) DAP pathway indicates uncertain origin of this typically plastid-located pathway in *Hematodinium*. MVA, mevalonate IPP pathway; DOXP, 1-deoxy-D-xylulose-5-phosphate IPP pathway; C15/20, isoprene chains 15 and 20 carbons long derived from IPP (an external source of IPP/isoprenoids for *Hematodinium* sp. is predicted); SUF, plastid-type iron-sulfur cluster pathway; DAP, diaminopimelate lysine pathway; C4/C5 pathways for tetrapyrrole (TP) synthesis differ only by the reactions to δ-aminolevulinic acid (ALA) and their location. (Gornik et al., 2015)

Recently, one strain belonging to *Amoebophrya* has been sequenced for the whole genome (**Fig** 21 A and B; John et al., 2019). Despite a reduction in genome size (around 100 Mb), loss of genes, and loss of functions due to dependence on the host, this strain has retained most of the genome functionality of a free-living species. The most noteworthy characteristic of this genome is the transfer of all essential functional mitochondrial genes to the nucleus, resulting in complete loss of the mitochondrial genome. But the mitochondria remains with a similar electron transport system and oxidative phosphorylation as found in *Chromera velia* (**Fig** 21D; John et al., 2019). Besides, the parasite has lost the plastid organelle as found in the *Hematodinium*.

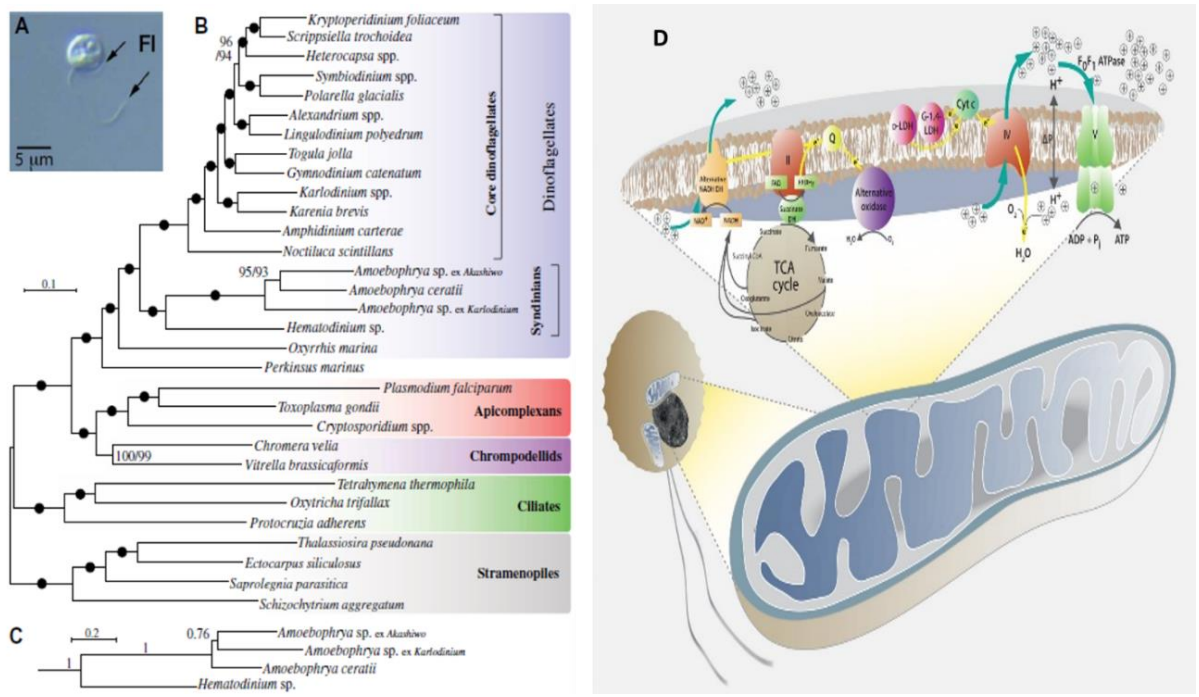


Fig 21. Multiprotein phylogeny of *Amoebophrya* isolated from three separate hosts, 15 other dinoflagellates, and 13 related eukaryotes. (A) Free-living stage of the parasite *Amoebophrya*. Fl, flagellum. (B) The best maximum likelihood tree (IQ-TREE) under the LG + G4 + I + F model with nonparametric bootstrap supports at branches (black circles denote 100/100 support). (C) Relationships among *Amoebophrya* isolates in a PhyloBayes GTR + CAT + G4 inference with posterior probabilities at branches; the rest of the tree is identical to (B) and is fully supported at all branches. (D) Model of mitochondrial functions in *A. ceratii* based on the genome gene content. The *C. velia* model was taken as a template. Mitochondrial complex I has been replaced by an alternative NADH dehydrogenase (DH), which reduced the NADH from the tricarboxylic acid (TCA) cycle. Both alternative NADH dehydrogenase and succinate dehydrogenase (complex II) channel electrons through the carrier ubiquinone (Q) to the alternative oxidase (yellow arrows). Electrons may also be passed by other sources, such as D-lactate: cytochrome c oxidoreductase (D-LDH) and galacto-1,4-lactone: cytochrome c oxidoreductase (G-1,4-LDH) to cytochrome c (yellow arrows), which passes them on to complex IV (cytochrome c oxidase). Stippled yellow arrows indicate alternative pathways of electron flow as proposed in *Chromera*. (John et al., 2019)

4 Research models in this study

Several species of dinoflagellates can form red tides or harmful algal blooms that sometimes cause illness and death in humans and large scale mortality of fish and shellfish, resulting in serious economic losses (Kim et al., 2019). The recent increase in the frequency of massive inshore blooms may originate from geographical and temporary disruptions between dinoflagellate and their natural enemies (Chambouvet et al., 2008). *Amoebophrya ceratii* (Syndiniales) are parasitoids and predators of dinoflagellates. Infection by *Amoebophrya ceratii* may retard or prevent the formation of dinoflagellates red tides and facilitate the decline of blooms (reviewed in Coats, 1999). By causing the destruction of bloom-forming hosts, these parasites effectively recycle undergrazed phytoplankton production through the "microbial loop", which represents an important aspect of marine planktonic food webs (Coats, 1999).

Interestingly, both the bloom-causing dinoflagellates and *Amoebophrya* belong to phylum Dinoflagellata in the Alveolata lineage.

4.1 Our targeted hosts

Our targeted dinoflagellate species, *Heterocapsa triquetra*, *Scrippsiella acuminata* (previously called *S. trochoidea*, see: http://aquasymbio.fr/) and *Scrippsiella donghaienis*, belong to Dinophyceae and to the Order Peridiniales.

The genus *Scrippsiella* includes approximately 30 known species, comprising a diverse group of thecate photosynthetic marine phytoplankton (reviewed in Kim et al., 2019). Within this genus, *Scrippsiella acuminata* (previously *S. trochoidea* STR1+STR2, Kretschmann et al, 2015) is ecologically known as a bloom-forming species (Gu et al., 2008) and has caused red tides in the waters of many countries around the world (Hallegraeff, 1992, Villarino et al., 1995; Frehi et al., 2007; Wang et al., 2008; Gárate-Lizárraga et al., 2009; Park et al., 2013; Hameed and Saburova, 2015; Tas and Yilmaz, 2015). It can be found in a broad range of temperatures from 10°C to 30°C and salinities from 5 to 55‰. It is therefore recognized as being well adapted to a wide variety of environmental conditions (Kim and Han, 2000). Moreover, this species is homothallic. The resting cyst is formed by sexual reproduction (see http://www.aquasymbio.fr/fr/node/111). This dormant stage is covered by calcareous spines and structures (**Fig** 22 C and D).

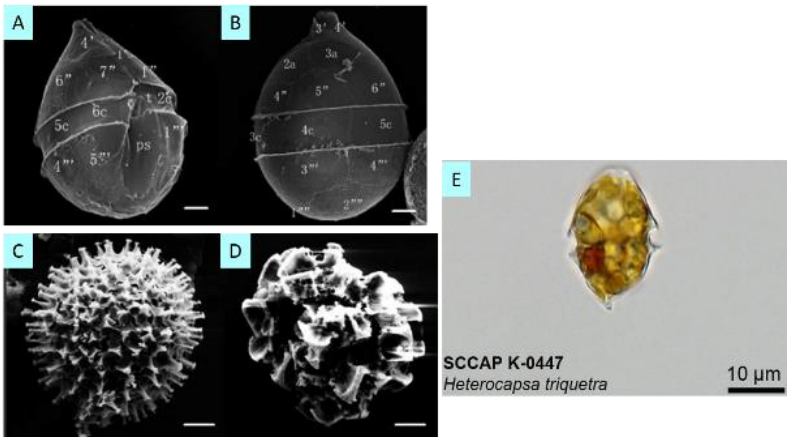


Fig 22. Ventral (A) and dorsal (B) view of *S. donghaienis* vegetative cell. Calcareous cysts of *S. acuminata* with typical (C) and unusual spines (D) (Gu et al 2008). Scale bars: 5 µm. (E) The view of *H. triquetra* vegetative cell (From http://nordicmicroalgae.org/taxon/Heterocapsa%20triquetra)

Scrippsiella donghaienis is a species very similar to *S. acuminata* morphologically (Gu et al., 2008). The vegetative cells of these two species are indistinguishable solely based on theca shape and plate patterns. However, *S. donghaienis* produces non-calcareous cysts while *S. acuminata* does not. These 2 species also show distinct internal transcribed spacer (ITS) sequences (**Fig** 23; Gu et al., 2008). The

presence of *S. donghaienis* has been reported in the waters of many countries (Lewis, 1991; Olli and Anderson, 2002; Hoppenrath, 2004; Joyce et al., 2005; Gu et al., 2008; Soehner et al., 2012; Lee et al., 2018). This species also has the potential to cause red tides or blooms.

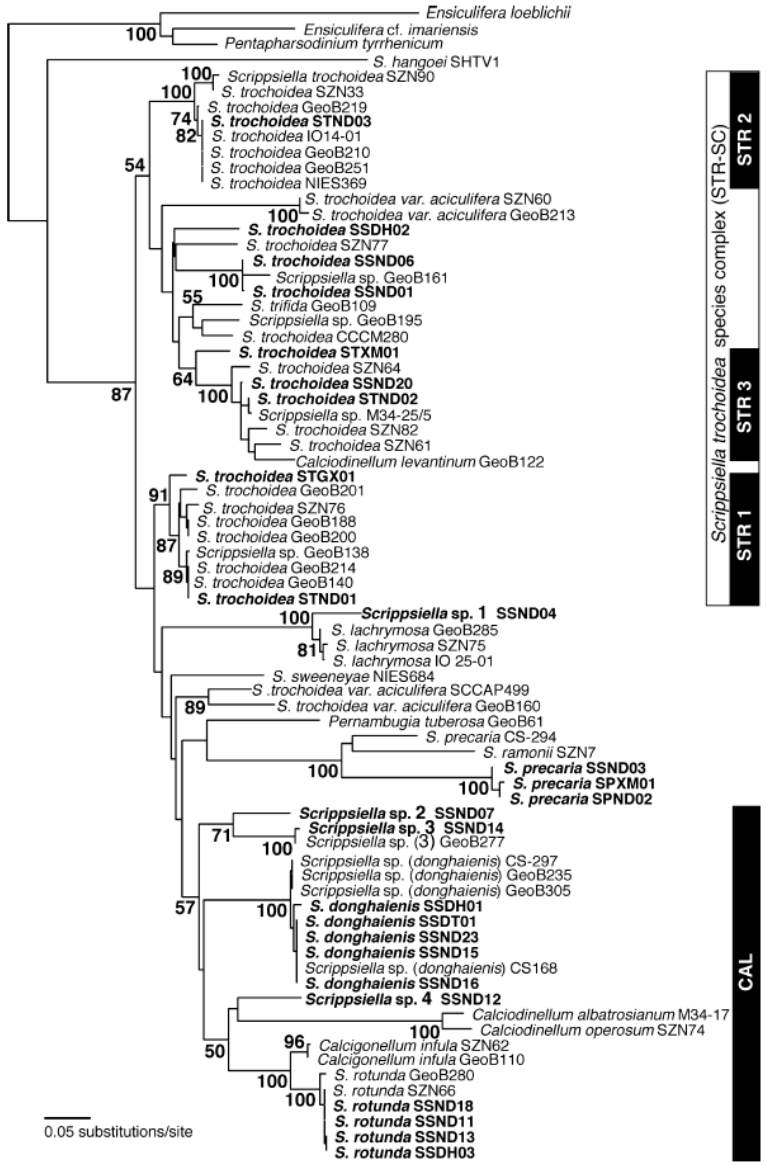


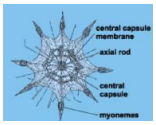
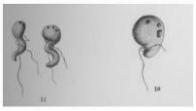
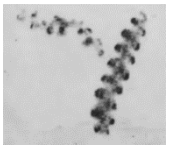
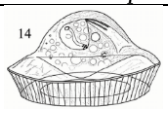
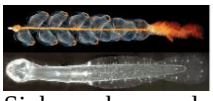
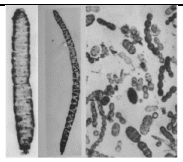
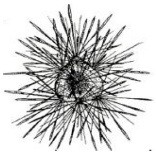
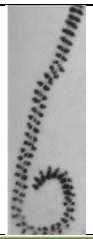
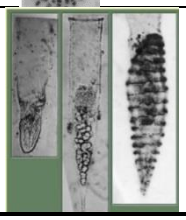
Fig 23. Neighbor-joining (NJ) phylogeny of pair-wise maximum-likelihood distances between ITS sequences from 71 ingroup taxa of *Scrippsiella* spp. and related genera, and three outgroup sequences (*S. hangoei, Ensiculifera, and Pentapharsodinium*). All sequences have been marked with species name and strain identification number. Bootstrap values have been indicated left of the clades which they refer to. Lack of a value indicates <50% support. (Gu et al., 2008)

H. triquetra is also a cosmopolitan dinoflagellate species known for forming blooms (Havskum and Hansen, 2006). It has a large pyrenoid and thecal plate with a characteristic protrusion (see http://nordicmicroalgae.org/taxon/Heterocapsa%20triquetra). Resting stage and reproduction mode is unknown for this species.

4.2 Our targeted parasites: *Amoebophrya*

Amoebophrya is one genus of Syndiniales and belongs to the family Amoebophryidae. Seven species have been formally described in this genus, including *A. acanthometrae*, *A. ceratii*, *A. grassei*, *A. leptodisci*, *A. rosei*, *A. sticholonchae*, *A. tintinni* (Cachon, 1964). They are distinguished from each other by virtue of their trophont morphology, spore morphology and the pattern of sporogenesis (**Table** 2).

Table 2. Seven *Amoebophrya* species described by Chaton (1964). Emended from Chambouvet thesis (2009). (from: http://www.theses.fr/136877354)

Species	Host	Dinospores	Trophont/ /vermiform
A. acanthometrae Koeppen 1894	 *Acanthometra pellucida*	 Microspores (left) and macrospores (right) of *Amoebophrya acanthometrae*. (Cachon 1964)	 Vermiform of *Amoebophrya acanthometrae*. Nucleus in telophase, first mitotic division. (Cachon 1964)
A. ceratii Cachon 1964	Dinoflagellates	-	-
A. grassei Cachon 1964	*Oodinium sp.*	 Microspores (by two: still ending their mitotic division, right) and macrospores (left) of *Amoebophrya grassei*. (Cachon 1964)	 Vermiform stage (fragmented) of *Amoebophrya grassei*. (Cachon 1964)
A. leptodisci Cachon 1964	 *Leptodiscus medusoides* (Dynophyceae)	-	-
A. rosei	 Siphonophore and Chaetognathe	-	 Development of the vermifrom stage of *Amoebophrya rosei*; left: just after the release (600 µm), middle: elongation 5 min after the release (1500 µm), right: fragmentation, 2 days after the release. (Cachon 1964)
A. sticholonchae Koeppen 1894	 Sticholonche	-	 First observation of *Amoebophrya sticholonchae* by Hertwig in 1879.
A. tintinni Cachon 1964	 (tintinnid)	-	 Mature trophont of *Amoebophrya tintinni* (x 2000). Sagittal view, right and left half. Note the presence of a single nucleus. (Cachon 1964)

The host photographs are taken from Chambouvet thesis (2009). The photographs for dinospores, vermiforms and trophonts are taken from the website: http://www.aquasymbio.fr

Amoebophrya is able to infect a wide range of marine organisms including ciliates, radiolarians, dinoflagellates, and even other dinoflagellate parasites (Cachon, 1964; Coats, 1999; Park et al., 2013). *A. ceratii* is known to infect numerous free-living dinoflagellate species including some toxic and harmful algal bloom species and thus was once thought as a biological control agent (Cachon, 1964; Coats, 1999). Infections by *A. ceratii* have been found in coastal waters over the world (examples in **Fig** 24; Park et al., 2013). Environmental sequences attributed to the Amoebophridae have been obtained from oceanic surface and deep waters (Guillou et al., 2008).

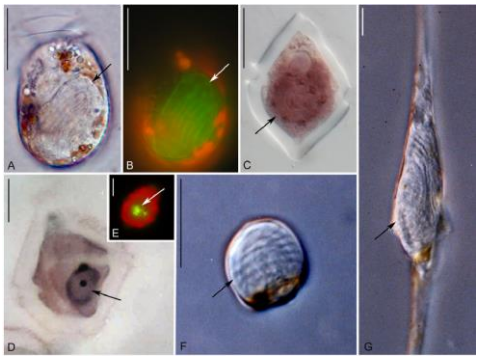


Fig 24. Representative dinoflagellates infected by eukaryotic parasites in coastal waters. (A) *Dinophysis ovum* with late, beehive stage of *Amoebophrya* sp. (B) Same cell as in A, but viewed with epifluorescence microscopy. (C) *Heterocapsa triquetra* with a mid-stage infection of *Amoebophrya* sp. (D) Very early infection in the nucleus of *Gonyaulax polygramma* caused by *Amoebophrya* sp. (E) *G. polygramma* with *Amoebophrya* sp. in early infection viewed with epifluorescence microscopy. (F) *Prorocentrum minimum* with very late, beehive stage of *Amoebophrya* sp. (G) *Neoceratium fusus* containing the beehive stage of *Amoebophrya* sp. (Park et al., 2013)

Among the 7 described species in *Amoebophrya, A. ceratii* is the better-known model as its members could be easily observed from natural samples and isolated in culture. Over two decades, however, a varying degree of host specificity and considerable sequence differences among *A. ceratii* strains have been revealed by molecular studies and laboratory experiments, suggesting *A. ceratii* is actually a species complex (Coats et al., 1996; Gunderson et al., 2002; Kim, 2006; Kim et al., 2008; Park et al., 2013). Based on 18S rDNA sequences from cultures or single cells isolated from the field, five distinct clades were identified in the phylogeny of *A. ceratii* (**Fig** 25; Park et al., 2013), which all have been named in the form of their host species, such as *Amoebophrya sp. ex Alexandrium affine, Amoebophrya sp ex. Ceratium tripos.*

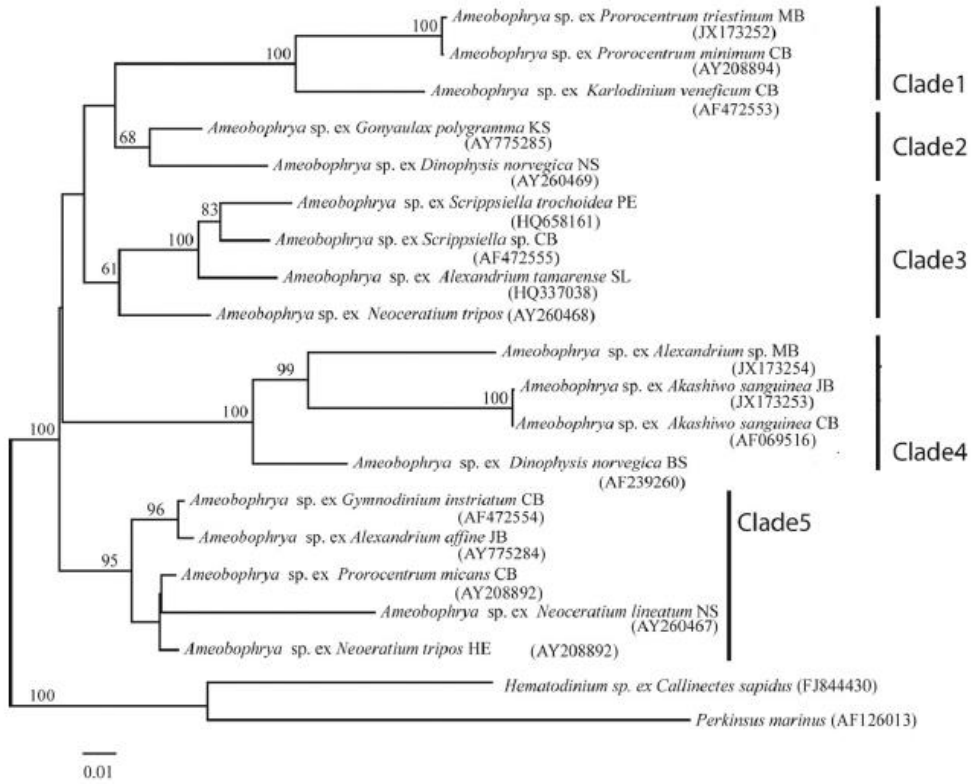


Fig 25. Phylogeny of *Amoebophrya* 18S rRNA gene. The maximum-likelihood phylogeny for the *Amoebophrya* parasites is shown. Bootstrap values are shown above the branches when greater than 50%. Origins of the *Amoebophrya* strains are indicated in abbreviation as follows: CB: Chesapeake Bay; NS: North Sea; SP: Salt Pond, USA; PE: Penze´ estuary, France; KS: Kunsan of Korea; JB: Jinhae Bay of Korea; MB: Masan Bay of Korea; BS: Baltic Sea; HE: Helsingor of Denmark. (Park et al., 2013)

However, this is far to reflect the diversity of *Amoebophrya* revealed by metabarcoding approach (de Vargas et al., 2015). MALV-II (the Family Amoebophryidae) revealed from environmental clone libraries was identified with 44 distinct clades (Guillou et al., 2008). Such high genetic diversity within the Amoebophryidae and/or *Amoebophrya* group raises a question as to whether the sequences indeed represent species diversity (Park et al., 2013).

As many *Amoebophrya* strains have been established in the lab, deeper studies have been performed and more aspects of *Amoebophrya* have been revealed. Ultrastructure of *Amoebophrya sp.* and its changes during the course of infection have been elucidated in very detail by transmission electron microscope (TEM) and scanning electron microscope (SEM) (**Fig** 26; Miller et al., 2012).

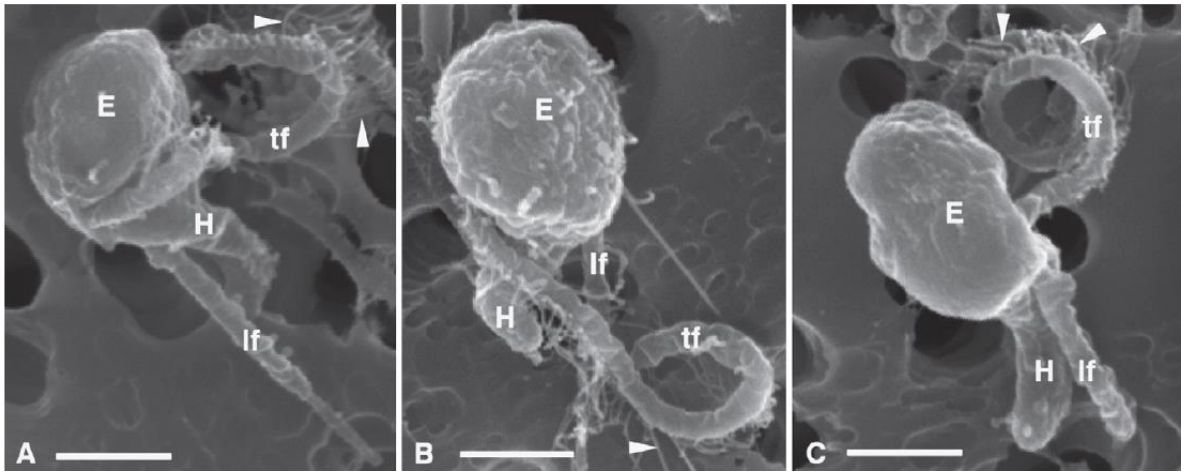


Fig 26. Scanning electron microscope (SEM) images of *Amoebophrya* sp. dinospores from *Akashiwo sanguinea* showing the episome (E), hyposome (H), longitudinal flagellum (lf), transverse flagellum (tf), and flagellar hairs (arrowheads) viewed from several perspectives. A. From the left displaying the constriction at the distal end of the longitudinal flagellum and the transverse flagellum emerging into the girdle, which at this point is an indentation of the plasma membrane. B. Dorsal showing the continuation of the girdle as a groove formed by the bulbous episome and twist of the narrow hyposome. C. From the right. Scale bars = 1 μm. (Miller et al., 2012)

A comparative transcriptomics analysis has been performed to investigate the host infection process by two *Amoebophrya* strains, one being a specialist (infecting a single host over the tested strain collection) and the other being considered as a more generalist parasite (Farhat et al., 2018). The analysis of the time-scale gene expression along a complete infection cycle revealed a set of genes involved in parasite development and host-parasitoid interaction at each stage. Intriguingly, these 2 strains showed a contrasting difference in gene expression mode involved in the defense process (oxidative stress response, **Fig** 27), suggesting the establishment of different strategies for parasite protection related to host specificity.

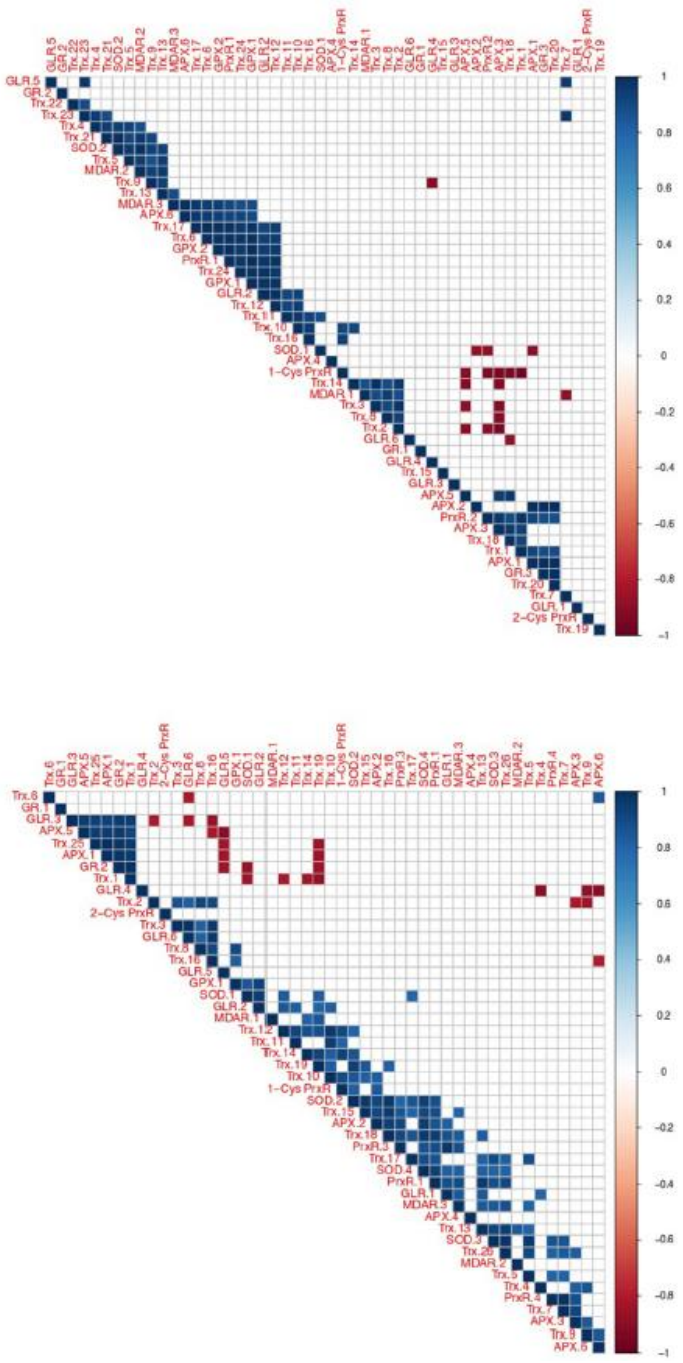


Fig 27. Correlation of oxidative stress response gene expression in *Amoebophrya*. Correlated gene expression of *Amoebophrya* anti-ROS family members. The correlation coefficient for each anti-ROS pair of gene expression values is shown (up, A120; down, A25). (Farhat et al., 2018)

Objectives of this thesis

The aim of my thesis is: 1) to determine if the sequence diversity of *Amoebophrya* observed in marine environments and in cultures represents species diversity; 2) to evaluate the possibility of sexual reproduction in *Amoebophrya*.

To determine if the sequence diversity observed in *Amoebophrya* represents the species diversity, we provided a comprehensive description of the diversity of strains maintained in the lab and single cells isolated from natural samples (chapter 1). The use of different species concepts (biological species concept, ecological species concept, phylogenetic species concept, morphological species concept and biological species concept (realized by CBC test)) in discriminating species was tested. To achieve this goal, we paid a large effort and established numerous strains in culture, collected natural samples from the field, recorded environmental parameters in time series, sequenced the rDNA by Sanger sequencing and the whole genomes by high through-output sequencing, tested for host spectrum using strains and for ecological niches using metabarcoding data, and described morphological traits that can be recognised by flow cytometry.

In chapter 2, I took advantage of available genomic data (recently sequenced but not published) to investigate the existence of a set of genes specifically involved in sexual reproduction and thereby predicted the possibility of sexual reproduction in *Amoebophrya*.

At the end, based on this contextual dataset, the species richness at the global scale was evaluated (general discussion), providing a comprehensive view of variations of these parasites. The genomic approach to discovering the genetic diversity of protists was also assessed.

Chapter 1

Cryptic species in the parasitic *Amoebophrya* species complex revealed by a polyphasic approach

Cryptic species in the parasitic *Amoebophrya* species complex revealed by a polyphasic approach

Ruibo Cai[a], Ehsan Kayal[b], Catharina Alves-de-Souza[c], Estelle Bigeard[a], Erwan Corre[b], Christian Jeanthon[a], Dominique Marie[a], Betina M. Porcel[d], Raffaele Siano[e], Jeremy Szymczak[a], Matthias Wolf[f], Laure Guillou[a,1]

[a] Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France

[b] Sorbonne Université, CNRS, FR2424 ABIMS, Station Biologique de Roscoff SBR, 29680 Roscoff, France

[c] Algal Resources Collection, MARBIONC, Center for Marine Sciences, University of North Carolina Wilmington, 5600 Marvin K. Moss Lane, Wilmington, NC 28409, US

[d] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

[e] Ifremer-Centre de Bretagne, Département/Unité/Laboratoire ODE/DYNECO/Pelagos, Z.I. Technopôle Brest-Iroise, Pointe du Diable BP70, 29280 Plouzané, France

[f] Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

[1]To whom correspondence should be addressed. Email:lguillou@sb-roscoff.fr. Laure Guillou, Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France. +33667972473.

Keywords: cryptic species, marine alveolates, dinoflagellates, environmental sequences, planktonic parasites

Running title: Species boundaries in the *Amoebophrya* species complex

Abstract

As critical primary producers and recyclers of organic matter, the diversity of marine protists has been extensively explored by high-throughput barcode sequencing. However, classification of short metabarcoding sequences into traditional taxonomic units is not trivial, especially for lineages mainly known by their genetic fingerprints. This is the case for the widespread *Amoebophrya ceratii* species complex (Syndiniales), parasites of their dinoflagellate congeners. Based on a polyphasic approach involving genetic and phenotypic characters applied to 119 strains (from cultures and environmental "single-cell" sequencing) isolated from two French estuaries, we defined 8 ribotypes (here considered as cryptic species) using ITS2 sequence-structure phylogenies, compensatory base changes (CBCs) and genomic *k*-mer comparison. All these ribotypes are able to infect the same host (*Scrippsiella acuminata*) and strains belonging to the same ribotype display similar host ranges. The closest relative ribotypes share 99.77% SSU rDNA sequence identity, suggesting that unique sequences (i.e., 100% threshold sequences similarity) rather than OTUs should be used in barcoding studies. We followed genetic traces of infections by *Amoebophrya*-like parasites during a three-year survey of summer dinoflagellate blooms in the Penzé estuary. Even though most of these *Amoebophrya* ribotypes co-occur and share the same ecological niche, we observed a unimodal pattern between population fitness and host range, where the maximal fitness values were observed for *Amoebophrya* ribotypes having an intermediate number of hosts. This case study highlights the need for a better delimitation of species boundaries in protist ecology.

Introduction

The accurate estimation of the diversity of protists (i.e., eukaryotic microbes) is crucial for gaining a better understanding of their ecological roles in world oceans [1][2]. However, traditional methods for species delimitations are challenging to apply to single-cell organisms where morphological features are frequently not discriminative enough [3][4]. The inventory of planktonic protist diversity in marine systems has recently expanded thanks to culture-independent, DNA barcode-based methods directly applied in the field over large geographic scales [5][6]. While this avalanche of environmental sequences is generally classified into manageable operational taxonomical units (OTUs), the correct assessment of the quantitative contribution and functional roles of marine pelagic protists is however hindered by the uncertainty of real species richness. In other words, intraspecific sequence variation within morphospecies needs to be differentiated from "true" species diversity [7]. So far, there are no universal rules linking molecular data to species richness partially due to the high incidence of asexuality, morphological and evolutionary convergence, and sometimes high discordance between genetic and phenotypic characters [8].

Parasitism is an essential ecological process contributing to the resilience of ecosystems, while acting as an evolutionary pressure for both hosts and parasites [9]. Given the parasitic genetic diversity and ubiquity, understanding the factors that generate, maintain, and constrain host-parasite interactions is of primary interest in ecology and evolution. Achieving a reliable delimitation of cryptic species within parasitic protistan lineages becomes then critical for gaining a better knowledge of their ecological niches and host range. The problem of species delimitation is pervasive for parasitic lineages almost exclusively composed of environmental sequences, such as the Marine ALVeolate lineages (MALVs) [10][11]. MALVs represents one of the most hyperdiverse lineage (> 1,000 estimated OTUs) recovered in the metabarcoding dataset collected during the Tara Oceans expedition [5][12]. However, only a handful species representatives of the different MALV lineages have been formally described, all of them obligatory aplastidial parasites occurring as intracellular biotrophs (i.e., the host is maintained alive during the infection but eventually killed) and belonging to the order Syndiniales [11]. Among them, Amoebophryidae (or MALV-II) were observed to have the highest rate of cladogenesis

(i.e., speciation minus extinction rates) among 65 marine protist lineages [13], making their classification even more challenging.

The *Amoebophrya ceratii* morphospecies is a MALV-II clade with a worldwide distribution that could be isolated in culture, and likely constitutes a species complex [14][15]. All *A. ceratii* populations described to date were reported to infect a broad range of marine dinoflagellates [16][11]. After a generation time lasting a couple of days, a single infected host produces hundreds of dinospores (i.e., free-living, flagellated infective propagules) with a very short life span [17]. Those dinospores frequently account for a substantial proportion (>25%) of the nanoplanktonic fraction (2-20 µm) in coastal waters [18] and can be readily consumed by microzooplankton grazers (20-200 µm) [19]. Consequently, such parasites potentially constitute key trophic links between different compartments of the marine food web in the oceanic carbon cycle [20], notably through population control of dinoflagellate blooms [21][22].

Here, we explored the diversity of the *A. ceratii* species complex thanks to an expanded isolation and sequencing effort of 76 strains in culture and 43 environmental single-cells from two close localities (the Penzé and Rance estuaries, France). We followed a polyphasic approach to provide the first comprehensive species boundaries delimitation within the *A. ceratii* species complex. To do so, we combined (i) ribotyping (both the SSU rDNA and ITS1-5.8S-ITS2 regions), (ii) *k*-mer analysis from whole-genome sequencing, (iii) analysis of the ITS2 compensatory base changes (CBCs), (iv) the assessment of phenotypic characteristics of dinospores by flow cytometry, and (v) their host range through cross-infection culture experiments. Finally, we applied our novel species boundaries (considered here as cryptic species until formal description is performed) to answer the following questions: do these *Amoebophrya* cryptic species share the same ecological niches? Can we explain their fitness (maximal abundance and persistence in time) by their host range? We explored the population dynamics of the *Amoebophrya* newly-defined cryptic species during a three-year summer metabarcoding survey of dinoflagellate blooms in the Penzé estuary, a site well known for its high diversity of *Amoebophrya* ribotypes infecting a wide range of dinoflagellate species, and where parasitic prevalence can reach 40% of total cell abundance [21]. This study constitutes the first evaluation of the interannual variability *Amoebophrya* strains, their ecological niches, and population fitness in the field.

Material and Methods

Amoebophrya-like cell isolation and genome sequencing

Amoebophrya-like individuals infecting different dinoflagellate species were collected from the Penzé (48°37'37.57"N, 3°57'13.17"W) and the Rance estuaries (48°31'49.61"N, 1°58'21.81"W), both located in the western Channel Sea (France) and separated by ~150 km. We isolated 76 *Amoebophrya*-like strains in culture as well as 43 infected dinoflagellate cells at the late stage of infection directly in a tube that were flash freeze (hereafter referred to as "single-cells") (Table S1). Our strategy to discriminate individuals (i.e., strains and single-cells) was to find fundamental units that formed separate branches on rRNA phylogenetic trees (i.e., ribotypes) and then check whether these fundamental units shared a unique combination of phenotypic characters as the first backbone for their taxonomy. For that, individuals were screened by sequencing the ITS1-5.8S-ITS2 region of the ribosomal operon as explained in Blanquart et al. [9]. Then, Illumina whole genome sequencing was performed for a selection of 50 cultivated strains (where the flow cytometry-estimated bacterial contamination was <10%) and 17 single-cells by trying to maximize the number of representative ribotypes. The methodology for cell harvesting for genomic analysis is detailed in the protocole.io dx.doi.org/10.17504/protocols.io.vrye57w. Whole genome amplification from single-cells was performed using a multiple displacement amplification (MDA) approach with RepliG (QIAGEN) according to the manufacturer's instructions. Paired-end libraries were prepared individually and sequenced on an Illumina HiSeq2000 platform and a draft genome was assembled for each of the strains. More detail regarding cultivation, isolation, sequencing and genome assembly are described in the Supplementary Methods. Raw data are available upon request or using the following link: http://application.sb-roscoff.fr/project/hapar.

Ribosomal operons analyses

We estimated the average number of ribosomal operons per *Amoebophrya* genome by comparing the read coverage to that of a list of putatively single-copy genes (starting list of 67 genes) (unpublished data). To do so, we first used a BLASTn (e-value < 0.0001) search against the draft

genome assemblies capture the ribosomal operon and the genes of interest. A gene was discarded from the putative single-copy gene list either if 1) it was detected in multiple copies using a reciprocal BLAST approach, or 2) had no hit. Genomic reads were then mapped to each of the best hits using Bowtie2 [23]. Only the aligned region (i.e., high-scoring pairings as reported by BLASTn) was used for calculating the average coverage of reference genes, and then used to estimate the number of repeated ribosomal operon per genome.

We annotated the full-length ITS2 sequences of the ribosomal operons using either the Hidden Markov Models (HMMs) [24] implemented in the ITS2 database [25] or by alignment to annotated ITS2 sequences. We then predicted the secondary structures by homology modeling using a relevant template (e.g., [26][27]) or by RNA structure using energy minimization and constraint folding [28][29] by forcing the inner circle from which the four helices emanate (i.e., no base pairs were allowed between areas separating helices). We used the common core structure for eukaryotes ITS2 RNA fold to build the alignment, which consist of four helices, considering than helix III was the longest, helix IV short and divergent [30][25], and helix IV was lost in some *Amoebophrya* individuals. The inner circle from which the four helices emanate was forced, i.e., no base pairs were allowed between areas separating helices. We used the procedures outlined in [31] for phylogenetic analysis of the ITS2 dataset. In short, a global multiple sequence-structure alignment was generated in 4SALE v1.7 where ITS2 sequences and their respective secondary structures were simultaneously aligned using a 12×12 ITS2 sequence-structure specific scoring-matrix [32][33][34]. Based on the simultaneous consideration of the primary sequence and the secondary structure, phylogenetic relationships were reconstructed by neighbor-joining (NJ) through the use of an ITS2 sequence-structure specific Jukes-Cantor correction (JC) and an ITS2 sequence-structure specific general time reversible (GTR) substitution model, both implemented in ProfDistS v0.9.9 [35]. Using the ITS2 sequence and secondary structure simultaneously (encoded by a 12-letter alphabet, [34]), a maximum parsimony tree (MP) was reconstructed by PAUP [36] based on default settings. A sequence-structure maximum likelihood tree (ML) was calculated using the "phangorn" package [37] in R [38]-[34]. Bootstrap support for the sequence-structure trees was estimated based on 100 replicates. A compensatory base change (CBC) table was transferred from 4SALE [33].

Genome comparison using SIMKA *k*-mer analysis

We used filtered reads as input in order to estimate the k-mer distribution of the various genomes with SIMKA (k = 21 pb; minimum read size ≥90 bp) [39] after discarding low complexity (i.e., Shannon index < 1.5) reads. Due to inherent differences in the genome coverage obtained from cultivated strains and single-cells, we based the cluster analysis upon the presence/absence of *k*-mers by considering only the distance indexes (based on the formulas given by [39]) that give more weight to the double presence of *k*-mers (i.e., Kulczynski, Ochiai, and Chord/Hellinger distances) [40]. Statistical supports for clusters were checked by bootstrap analysis after 100 permutations using the *clusterboot* function from the "fpc" R package. The permutations were directly performed on the distance matrix output by SIMKA with 'clusterCBI' as clustering method, considering the above-estimated number of ribotypes as the desired number of clusters.

Cell morphology analyses using flow cytometry

We estimated some of the morphological cell signatures of the cultured strains by flow cytometry using the SSC (side scatter) and the FSC (forward scatter) parameters, as well as the natural green autofluorescence of *Amoebophrya* spp dinospores when excited by light at 405 nm wavelength [41]. For that, we used 500 µl of fresh cultures directly loaded on a FACsAria flow cytometer (Becton Dickinson, New Jersey, USA). At the same time, we estimated the genome size of each strain following the procedure explained in [42], where the ratio between the mean distribution of the dinospores and the internal reference *Micromonas pusilla* RCC299 cells (1C = 20.9 fg) was used for the evaluation of the nuclear DNA content.

Host range

We monitored the host range of the parasites in culture in the laboratory through cross-infecting experiments using various dinoflagellate strains isolated from similar geographic area (or close), than the *Amoebophrya* strains used in this study (Table S1). Freshly produced dinospores were filtered through a 5-µm cellulose acetate filter (Minisart, Sartorius, Germany) and 100 µl were inoculated into

1 ml of different exponentially growing dinoflagellate cultures into 24-well plates. Infections by *Amoebophrya* strains were determined based on the detection of their natural green fluorescence under fluorescent microscopy (see above) between 2 and 5 days after inoculation. Hosts were classified either as resistant (no trace of infection) or sensitive (at least one infected host cell observed). For the same couple, cross-infections were processed 3 to 5 times at different dates.

Environmental metabarcoding survey

We obtained environmental rDNA metabarcode sequences of samples collected in the Penzé estuary during summer of three consecutive years (2010-2012). The DNA extraction method was based on a phenol-chloroform protocol [43]; the universal TAReuk454FWD1 (5'-CCAGCASCYGCGGTAATTCC-3') and the modified reverse BioMarKs (5'-ACTTTCGTTCTTGATYRATGA-3') primers [44] were used to amplify the V4 region (~380 bp) of the eukaryotic 18S rDNA of the >10-µm size-fraction. PCR amplifications were performed in duplicates for each sample using 5 µM of each primer, 5 µl of 5x buffer, 37.5 mM of magnesium chloride, 6.25 mM of dNTPs, 0.5 unit of GoTaq Flexi (Promega, Wisconsin, USA), approximately 2 ng of DNA, and pure water to obtain a final volume of 25 µl. Amplifications were performed using the following thermal conditions: a first denaturation at 95°C for 3 min, followed by 22 to 25 cycles of denaturation at 95°C for 45s, primer ligation at 50°C for 45s, and extension at 68°C for 90s, and a final extension at 68°C for 5 min. The size and quality of amplicons were checked on a 1% agarose gel before being sent to the GeT-PlaGe platform in Toulouse (France) for Illumina Miseq library preparation and paired-end sequencing. Taxonomic annotations were performed on unique sequences (100% threshold sequences similarity) observed in at least two different libraries using Mothur [45] implemented by the PR2 reference database [46] modified to take into account the *Amoebophrya* species boundary thresholds detected here.

Statistical analyses

All the statistical analyses described below were performed in R using packages freely available on the CRAN repository (http://www.cran-r-project.org).

Comparison of ribotypes based on flow cytometry features, number of operons and host range. We first used Pearson correlations to establish whether the different morphological variables monitored here (excluding host range) were related to one another. Then, differences between ribotypes were assessed by pairwise Mann-Whitney analysis using the *cor.test* and *wilcox.test* functions from the 'stats' package based on [log (x+1)] transformed data. For comparison of *Amoebophrya* ribotypes based on their host range, results from the cross-infections were organized into a presence/absence matrix (i.e., infection = 1; no infection = 0) with parasites in the columns and dinoflagellate host strains in the rows. This matrix was then used to generate a heatmap using the function *heatmap.2* of the 'gplots' package [47]. Finally, we assessed the relative importance of all characters in the differentiation of the ribotypes using NMDS analysis with the function *metaMDS* of the 'vegan' package [48] on standardized variables (between 0 and 1) based on their minimum and maximum values [49]. Then, we used the function *envfit* from the same package to fit the tested variables to the two first NMDS axes.

Niche analysis. The Outlying Mean Index (OMI) analysis [50] was first performed to determine the niche position and niche breadth of *Amoebophrya* ribotypes using the function *niche* in the 'ade4' package [51]. We included all 1,153 unique sequences detected in the metabarcodes (distributed into different phylogenetic lineages) to get a better resolution in the niche position of the *Amoebophrya* ribotypes. Relative read abundances (compared to the total number of reads) and several environmental parameters [i.e., water temperature, salinity, precipitation, tide coefficient, NO_3, PO_4 and $Si(OH)_4$] were included in two separate matrixes (number of samples = 48). Before analysis, relative read abundances were Hellinger transformed [52] whereas the environmental variables were standardized to values between 0 and 1 [49]. The function *envfit* was used to fit the environmental variables to the first two OMI axes. Sample scores from the first two OMI axes were then used to estimate the kernel density weighted by abundance [53][54] of *Amoebophrya* ribotypes using the *kde* function from the 'ks' package [55]. The niche overlap was then estimated by the comparison of the realized niches (i.e., kernel densities) through the calculation of the *D* metric [56] for each pair of *Amoebophrya* ribotypes using the *ecospat.niche.overlap* function from the 'ecospat' package [57]. Pair-wise *D* metrics were then used to generate a heatmap to detect clustering of the ribotypes related to their niche overlap, following the same procedure described previously for the analysis of cross-infection results.

Relationship between ribotypes' population fitness and host range. We first obtained a more precise estimation of the quantitative contribution of the different ribotypes by dividing the relative abundance of each ribotype in a given metabarcoding sample by their average number of operons estimated from the genome analysis of the strain. We used this normalized abundance to estimate the population fitness of the six *Amoebophrya* ribotypes that could be discriminated in the metabarcodes through their V4 sequences, in each one of the three years (number of samples = 18), based on 1) their maximal normalized relative read abundance and 2) persistence in the sample (e.g., the number of consecutive days the non-normalized relative contribution of the ribotype to the total number of reads was higher than 10%). The relationship between these two fitness indicators and the host range for each ribotype (maximal number of infected host species in the cross-infection experiment) was then assessed by polynomial regressions using the *poly* function in the 'stats' package following [log (x+1)] transformation.

Results and discussion

Ribotypes as cryptic species

We amplified and sequenced part of the ITS1-5.8S-ITS2 region from 76 strains in culture and 43 environmental "single-cell" samples (Table S1). The alignment based on the secondary structure of the ITS2 region clustered *Amoebophrya*-like individuals into eight main ribotypes (RIBs 1-8, Fig.1A-C). These ribotypes displayed low intra-group sequence variations (<3 single-nucleotide polymorphism or SNPs) in the ITS1-5.8S-ITS2 region and none in the SSU rDNA region, with the notable exception of RIB1 that contained one SNP in the V1-V2 region. Following the nomenclature proposed by Guillou et al. [11], members of RIB2 belonged to the MALV-II clade 4, whereas the remaining ribotypes were members of the MALV-II clade 2 (Fig. S1). Individuals belonging to ribotypes in MALV-II clade 2 (RIBs 1 and 3-8) shared 95.96-99.77% pairwise sequence identities, but only 92.89-93.91% with those from the RIB2 clade (Table S3). RIB3 and RIB8 were the most similar ribotypes (four SNPs in their SSU rDNA, no SNP in V4 region and one in V9 region, Table S3).

We investigated whether the observed rDNA sequence variability reflected species-level or intraspecific diversity by analyzing compensatory base changes (CBCs) between the ribotypes ITS2

sequences. CBCs are mutations impacting both nucleotides of a paired region in the folded RNA transcript that maintains the pairing (e.g., A-U to G-C) and the secondary hairpin structure of the ITS2 [58]. According to Müller et al. [59], CBCs found in the ITS2 region of the rDNA correlate (with a probability of 0.93) to the biological species concept (interbreeding populations generating fertile offspring and reproductively isolated from others) of species [60] and the absence of CBC, suggesting that the two ITS2 belong to the same species with a probability of 0.76. As a consequence, the CBC species concept stands as a valuable and pragmatic alternative for indicating the potential for novel species into protistan lineages (e.g., [61][62][63][64]). We observed no CBC within ribotypes, whereas 1-9 CBCs were observed between different ribotypes (Fig. 1D). The phylogenetically closest ribotypes RIB3 and RIB8 displayed 2 CBCs, while RIB 1 and 6 only diverged by one CBC despite being further apart on the rDNA tree (Fig. 1D).

Considering that CBCs and ribotypes are targeting the same genomic region (ribosomal operon), we aimed to determine if a comparison at the genome level should be a more appropriate approach for determining species, considering that two genomes should be similar enough in size and sequence to pair during sexual reproduction. Genome sizes of strains estimated by flow cytometry oscillated between 121 and 250 Mb (Fig. 2A). Overall, we observed a somewhat consistent genome size within ribotypes that clustered into two main groups with no significant genome size intra-group variability (Mann-Whitney pairwise tests; $p > 0.01$): the group made of RIBs 2, 5 and 6 displayed larger estimated genome size values than the group composed of RIBs 1, 3, 4, and 7. Such level of genome size disparity likely prevents any sexual reproduction between these two groups. We additionally estimated the number of ribosomal operons per genome ranging between 58 (strain A151 belonging to RIB4) and 270 (strain A147 belonging to RIB2), with no correlation between the number of operons and the genome size ($R = 0.22$; $p = 0.71$) (Fig. 2B). Thanks to genomic sequence reads acquired for 67 individuals (17 of which were environmental "single-cell" samples) we estimated the *k*-mer distribution (Table S2) from strains sharing the same ribotype to be part of the same cluster with high bootstrap support (>90%; Fig. 1A, E). All of these results suggest a low gene flow, if any, between ribotypes, consistent with placing them into separate cryptic species, awaiting for more formal description. As a consequence, we consider all ribotypes to belong to different cryptic species.

Correlation between "molecular" and "phenotypic" species boundaries in *Amoebophrya*

We explored whether these eight ribotypes displayed distinguishable phenotypes. Flow cytometer data showed a significant correlation between SSC and FSC parameters (R = 0.81; $p > 0.01$) as well as green autofluorescence (R = 0.71 and 0.94, for SSC and FSC respectively; $p > 0.01$). We frequently observed different populations of dinospores per strain illustrated by distinct flow cytometry signatures, suggesting that dinospores could be still engaged in cell divisions occurring during sporulation, as previously reported for Syndinids [17][65]. FSC, SSC and green autofluorescence differentiated strains belonging to the RIB2 from the rest, as their dinospores seemed to be brighter and larger when compared to other ribotypes (Mann-Whitney pairwise tests; $p > 0.01$) (Fig. 2C-E). We observed no significant differences among the other ribotypes for these three parameters. The separation of RIB2 (MALV-II clade 4) from the other ribotypes suggests that flow cytometry signatures can be useful for discriminating strains belonging to different higher taxonomic levels such as various MALV-II clades as previously proposed [11].

We explored the host range of representative *Amoebophrya* strains in culture (Fig. 2F). All strains can infect the same strain of *Scrippsiella acuminata* STR1, an autotrophic dinoflagellate species ubiquitous in both localities and used as host in cultures. We defined as specialists ribotypes (RIBs 1, 3, 6 and 7) the ones that only infected a single dinoflagellate species (i.e., *Scrippsiella acuminata* STR1), while those capable of infecting several species in the same *Scrippsiella* genus (RIB5) or from different genera (RIB2 and RIB4 infecting both *Scrippsiella* and *Heterocapsa*; Fig. 2F) were considered as generalists. We found that the capacity to infect more than one host species correlated with ribotype boundaries, where the strains belonging to the same ribotype displayed similar host ranges (Fig. 2F). The overall consistency in the host spectrum observed within the different ribotypes might suggest a genetic determinism underlying host specialization. Host spectrum is often considered as more permissive in culture experiments compared to the natural environment [66], while higher genomic diversity exists and potentially extends or reduces the host range from that observed in the laboratory. Using microscopical host identification of infected single-cells from environmental samples, we isolated RIBs 2, 4, 5 and 8 from both Scrippsielloids and *H. triquetra*, allowing us to enlarge previous

observations made in the laboratory (Table S1). Interestingly, RIBs 3 and 8 (the closest ribotypes by their rRNA sequences) differed by their host range. The intra-ribotype host spectrum variability observed in the field suggests a potential for rapid shifts in the host spectrum depending on the availability of target species following bloom cycles.

We performed a non-metric multidimensional scaling (NMDS) analysis to assess the relative importance of (i) the phenotypic characters assessed by flow cytometry (genome size and phenotypic features) and (ii) the number of hosts, in discriminating the eight ribotypes defined above (Fig. 2G). The *envfit* test indicated that the number of hosts and the genome size were the main features explaining the phenotypic discrimination of the strains into three clusters (R^2 = 0.97 and 0.96, respectively; p > 0.001). Strains from RIB4 separated from the other ribotypes based upon the highest number of potential hosts whereas the remaining strains separated into two groups based on their genome sizes. Overall, our results suggest that biological features such as most phenotypic characters analyzed here are not sufficient to distinguish *Amoebophrya* ribotypes which should be considered as cryptic species.

Application of cryptic species boundaries to environmental data

As a case study, we applied the newly defined cryptic species boundaries for *Amoebophrya* to a metabarcoding survey performed during dinoflagellate blooms in the Penzé estuary over three consecutive summers (2010-2012). Using a 100% threshold SSU rDNA sequences similarity (i.e., unique sequences) except for RIBs 3 and 8 that cannot be differentiated using the V4 region (referred to as RIB3/8 hereafter), we found all *Amoebophrya* ribotypes coexist in the Penzé estuary during most of the survey period, but with contrasting patterns among the different years (Fig. 3A). While the proportion of *Amoebophrya*-like reads did not exceed 6% of the total reads for any given ribotype, ribotypes RIB3/8 and RIB5 were the most ubiquitous during the survey periods. The niche analysis based on the outlying mean index (OMI) pointed out a strong interannual variability (Fig. 3B) mainly due to NO_3 concentration and temperature levels (*envfit* test; R^2 = 0.92 and 0.63, respectively; p < 0.05), both showing higher values in 2010 and 2011 than in 2012. Kernel density plots on the first two OMI axes (Fig 3C) indicated that most ribotypes showed similar realized niches during the entire sampling period. Exceptions to this pattern were however observed for RIB2 and RIB4, whose occurrences were

more restricted to 2010 and 2011 for RIB2 and to 2012 for RIB4. These differences were highlighted by the heatmap analysis based on the *D* metric (i.e., niche overlap) calculated using the Kernel densities (Fig. 3D) indicating a clear separation of RIBs 2 and 4 from the other ribotypes. The heatmap considering the niche overlap between parasites and other dinoflagellates unique sequences further indicated that RIBs 2 and 4 co-occurred with different dinoflagellate assemblages when compared to the other ribotypes (Fig. 3D). By contrast, the other ribotypes (RIBs 1 and 3-8) were in sympatry, i.e. sharing the same environment and potentially the same hosts during the same period of the year. In other words, these cryptic species naturally co-occur in the Penzé estuary and potentially compete for the same resources, as they can infect the same host species.

Finally, we investigated whether the host spectrum of each ribotype was related to its population fitness, taking into account the normalized relative abundance of reads based on the average number of operons in each ribotype. We observed a unimodal response between the two fitness estimators for each year (i.e., maximal normalized relative read abundance and persistence in the system) and the number of potential hosts infected by each ribotype (Fig. 3E-F), where ribotypes having medium host range (3 different host species; $p < 0.05$) displayed significantly higher sequence abundances and marginally longer persistence time in the environment. Although this outcome needs to be interpreted with care due to the low sampling size (N = 18), this result suggests a putative ecological advantage for *Amoebophrya* to infect more than one host leveraged by a lower fitness for the more generalistic parasitic species/strains.

Concluding remarks

Here, we provide molecular evidence for the presence of at least eight ribotypes likely reflecting cryptic *Amoebophrya* species. Our results indicate that the ITS2 region of the ribosomal operon is a better proxy than phenotypic characters (such as size and behavior) for species delimitation in the Amoebophryidae clade and that CBC is a statistically robust test to differentiate putative species. Thanks to this new species definition, we observed that most of these cryptic species co-occurred during dinoflagellate blooms over a three-year monitoring survey in the Penzé estuary, likely competing for similar ecological niches and host resource. We also suggest a maximal fitness for parasites having a medium

host range, reflecting an elevated cost either for infecting a large host range or being highly specialized. This study suggests that a complete taxonomic revision of parasitic dinoflagellates is long overdue in order to understand their role in plankton population dynamics.

Acknowledgements

RC and EK were funded by the Agence Nationale de la Recherche ANR-14-CE02-0007 (HAPAR project) and the Région Bretagne (ARED PARASITE-9450 and SAD HAPAR-S15JRCT024 grants). We warmly thank Karen Lebret for her help with metabarcoding sequencing, and Ramon Massana for providing unpublished MALASPINA sequences used in Figure S1. LG conceived this study. LG, CAdS, EB, CJ, DM, RF participated in the sample cruises and strain isolation. LG performed cross-infection. DM and RC performed the flow cytometry analyses. RC, EB, LG, JS prepared the material for sequencing. RC, EK, EC, & BP performed genetic analyses. MW did ITS2 secondary structure predictions, sequence-structure phylogenetics and the CBC analysis. CAdS performed statistical analyses. LG, EK & RC wrote the paper. All authors edited and approved the final version of this paper.

Competing Interests

The authors declare no competing financial interests

References

1. Sherr BF, Sherr EB, Caron DA, Vaulot D, Worden AZ. Oceanic Protists. *Oceanography* 2007; **20**: 130–134.
2. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. Protists are microbes too: A perspective. *ISME J* 2009; **3**: 4–12.
3. Ruhl MW, Wolf M, Jenkins TM. Compensatory base changes illuminate morphologically difficult taxonomy. *Mol Phylogenet Evol* 2010; **54**: 664–669.
4. Wolf M, Chen S, Song J, Ankenbrand M, Müller T. Compensatory base changes in ITS2 secondary structures correlate with the biological species concept despite intragenomic variability in ITS2 sequences - A proof of concept. *PLoS One* 2013; **8**: e66726.
5. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. *Science (80-)* 2015; **348**: 1–11.
6. Villarino E, Watson JR, Jönsson B, Gasol JM, Salazar G, Acinas SG, et al. Large-scale ocean

connectivity and planktonic body size. *Nat Commun* 2018; **9**.

7. Caron DA, Hu SK. Are we overestimating protistan diversity in nature? *Trends Microbiol* 2019; **27**: 197–205.
8. Boenigk J, Ereshefsky M, Hoef-Emden K, Mallet J, Bass D. Concepts in protistology: Species definitions and boundaries. *Eur J Protistol* 2012; **48**: 96–102.
9. Blanquart F, Valero M, Alves-De-Souza C, Dia A, Lepelletier F, Bigeard E, et al. Evidence for parasite-mediated selection during short-lasting toxic algal blooms. *Proc R Soc B Biol Sci* 2016; **283**.
10. López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 2001; **409**: 603–607.
11. Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, et al. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* 2008; **10**: 3349–3365.
12. Clarke LJ, Bestley S, Bissett A, Deagle BE. A globally distributed Syndiniales parasite dominates the Southern Ocean micro-eukaryote community near the sea-ice edge. *ISME J* 2019; **13**: 734–737.
13. Pernice MC, Logares R, Guillou L, Massana R. General Patterns of Diversity in Major Marine Microeukaryote Lineages. *PLoS One* 2013; **8**: e57170.
14. Gunderson JH, John SA, Boman II WC, Coats DW. Multiple strains of the parasitic dinoflagellate Amoebophrya exist in Chesapeake Bay. *J Eukaryot Microbiol* 2002; **49**: 469–474.
15. Kim S, Park MG, Kim KY, Kim CH, Yih W, Park JS, et al. Genetic diversity of parasitic dinoflagellates in the genus Amoebophrya and its relationship to parasite biology and biogeography. *J Eukaryot Microbiol* 2008; **55**: 1–8.
16. Cachon J. Contribution à l'étude des péridiniens parasites. Cytologie, cycles évolutifs. *Ann des Sci Nat Zool Paris* 1964; 1–158.
17. Coats DW, Park MG. Parasitism of photosynthetic dinoflagellates by three strains of Amoebophrya (Dinophyta): parasite survival, infectivity, generation time, and host specificity. *Aquat Microb Ecol* 2002; **528**: 520–528.

18. Siano R, Alves-de-Souza C, Foulon E, M. Bendif E, Simon N, Guillou L, et al. Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences* 2011; **8**: 267–278.

19. Johansson M, Coats DW. Ciliate grazing on the parasite Amoebophrya sp. decrease infection of the red-tide dinoflagellate Akashiwo sanguinea. *Aquat Microb Ecol* 2002; **28**: 69–78.

20. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 2016; **532**: 465–470.

21. Chambouvet A, Morin P, Marie D, Guillou L. Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* 2008; **322**: 1254–1257.

22. Alves-de-Souza C, David P, Emilie LF, Sébastien M, Cécile R, Behzad M, et al. Significance of plankton community structure and nutrient availability for the parasitic control of dinoflagellate blooms by parasites: a modeling approach. *PLoS One* 2015; e0127623.

23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–9.

24. Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M. 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene* 2009; **430**: 50–57.

25. Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F. ITS2 database V: Twice as much. *Mol Biol Evol* 2015; **32**: 3030–3032.

26. Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *Bioinformatics* 2005; **11**: 1616–1623.

27. Selig C, Wolf M, Müller T, Dandekar T, Schultz J. The ITS2 Database II: Homology modelling RNA structure for molecular systematics. *Nucleic Acids Res* 2008; **36**: 377–380.

28. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of {RNA}. *J Mol Biol* 1999; **288**: 911–940.

29. Reuter JR, Mathews DM. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 2010; **11**: 129.

30. Schultz J, Maisel S, Gerlach D, Müller T, Wolf M. A common core of secondary structure of

the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. 2005; **2**: 361–364.

31. Schultz J, Wolf M. ITS2 sequence-structure analysis in phylogenetics: A how-to manual for molecular systematics. *Mol Phylogenet Evol* 2009; **52**: 520–523.

32. Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M. Structure alignment and editing. *BMC Bioinformatics* 2006; **7**: 498.

33. Seibel PN, Müller T, Dandekar T, Wolf M. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Res Notes* 2008; **1**: 91.

34. Wolf M, Koetschan C, Müller T. ITS2, 18S, 16S or any other RNA - simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. *Gene* 2014; **546**: 145–149.

35. Wolf M, Ruderisch B, Dandekar T, Schultz J, Müller T. ProfDistS: (profile-) distance based phylogeny on sequence - Structure alignments. *Bioinformatics* 2008; **24**: 2401–2402.

36. Swofford DL. PAUP*. Phylogenetic analysis using parsimony (*and other methods). 2003. Sinauer Associates, Sunderland, Massachusetts.

37. Schliep KP. phangorn: Phylogenetic analysis in R. *Bioinformatics* 2011; **27**: 592–593.

38. R Development Core Team. R: A language and environment for statistical computing. 2014. Foundation for Statistical Computing, Vienna, Austria.

39. Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, et al. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ* 2016; **2**: e94.

40. Legendre P, Legendre LF. Numerical ecology, vol 24. 2012. Elsevier.

41. Coats DW, Bockstahler KR, Berg GM, Sniezek JH. Dinoflagellate infections of Favella panamensis from two North American estuaries. *Mar Biol* 1994; **119**: 105–113.

42. Marie D, Partensky F, Simon N, Guillou L, Vaulot D. Flow cytometry analysis of marine picoplankton. In: Diamond RA, DeMaggio S (eds). *Living Colors: protocols in cytometry and cell sorting*. 2000. Springer Verlag, pp 421–455.

43. Díez B, Pedrós-Alió C, Marsh TL, Massana R. Application of denaturing gradient gel electrophoresis (DGGE) to study the diversity of marine picoeukaryotic assemblages and comparison of DGGE with other molecular techniques. *Appl Environ Microbiol* 2001; **67**: 2942–

2951.

44. Piredda R, Tomasino MP, D'Erchia AM, Manzari C, Pesole G, Montresor M, et al. Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean long term ecological research site. *FEMS Microbiol Ecol* 2017; **93**: fiw200.

45. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537–7541.

46. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 2013; **41**: 597–604.

47. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. Package 'gplots'. 2016. Available online: https://cran.r-project.org/web/packages/gplots/gplots.pdf (accessed on 17 September 2018).

48. Oksanen J, Kindt R, Legendre P, O'Hara B, Henry M, Stevens H. The vegan package. *Community Ecol Packag* 2007; **10**: 631–637.

49. Alves-de-Souza C, Benevides TS, Santos JBO, Von Dassow P, Guillou L, Menezes M. Does environmental heterogeneity explain temporal β diversity of small eukaryotic phytoplankton? Example from a tropical eutrophic coastal lagoon. *J Plankton Res* 2017; **39**: 698–714.

50. Dolédec S, Chessel D, Gimaret-Carpentier C. Niche separation in community analysis: A new method. *Ecology* 2000; **81**: 2914–2927.

51. Dray S, Dufour AB. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* 2007; **22**: 1–20.

52. Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. *Oecologia* 2001; **129**: 271–280.

53. Broennimann O, Fitzpatrick MC, Pearman PB, Petitpierre B, Pellissier L, Yoccoz NG, et al. Measuring ecological niche overlap from occurrence and spatial environmental data. *Glob Ecol Biogeogr* 2012; **21**: 481–497.

54. Hernandez-Fariñas T, Bacher C, Soudant D, Belin C, Barillé L. Assessing phytoplankton

realized niches using a French national phytoplankton monitoring network. *Estuar Coast Shelf Sci* 2015; **159**: 15–27.

55. Duong T. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J Stat Softw* 2007; **21**: 10.18637/jss.v021.i07.

56. Schoener TW. Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology* 1970; **51**: 408–418.

57. Broennimann O, Di Cola V, Petitpierre B, Breiner F, Scherrer D, D'Amen M, et al. Package 'ecospat'. 2018. Available online: https://cran.r-project.org/web/packages/ecospat/ecospat.pdf (accessed on 21 August 2018).

58. Gutell RR, Larsen N, Woese CRR. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Mol Biol Rev* 1994; **58**: 10–26.

59. Müller T, Philippi N, Dandekar T, Schultz RG, Wolf M. Distinguishing species. *Rna* 2007; **13**: 1469–1472.

60. Mayr E. The growth of biological thought diversity, evolution, and inheritance. 1982. Belknap Press.

61. Amato A, Kooistra WHCF, Levialdi Ghiron JH, Mann DG, Pröschold T, Montresor M. Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* 2007; **158**: 193–207.

62. Rodríguez-Martínez R, Rocap G, Salazar G, Massana R. Biogeography of the uncultured marine picoeukaryote MAST-4: Temperature-driven distribution patterns. *ISME J* 2013; **7**: 1531–1543.

63. Annenkova N V., Hansen G, Moestrup Ø, Rengefors K. Recent radiation in a marine and freshwater dinoflagellate species flock. *ISME J* 2015; **9**: 1821–1834.

64. Simon N, Foulon E, Grulois D, Six C, Desdevises Y, Latimier M, et al. Revision of the genus *Micromonas* Manton et Parke (Chlorophyta, Mamiellophyceae) of the species M. pusilla (Butcher) Manton et Parke, of the species M. commoda van Baren, Bachy et Worden and description of two new species. *Protist* 2017; **168**: 612–635.

65. Shadrin AM, Simdyanov TG, Pavlov DS, Nguyen THT. Free-living stages of the life cycle of the parasitic dinoflagellate Ichthyodinium chabelardi Hollande et J. Cachon, 1952 (Alveolata:

Dinoflagellata). *Dokl Biol Sci* 2015; **461**: 104–107.

66. Poulin R, Keeney DB. Host specificity under molecular and experimental scrutiny. *Trends Parasitol* 2008; **24**: 24–28.

Figure and Table legends

Figure 1: The eight *Amoebophrya* ribotypes (RIBs 1-8) defined by ITS2 secondary structures and SIMKA *k*-mer genome comparison.

(A) Secondary structure neighbor-joining (NJ) tree rooted with ribotype 2 (RIB2) derived from a multiple sequence-structure alignment of the ITS2 region with a 12x12 JC correction. Bootstrap values >50 are mapped to nodes. (B) Secondary structure NJ tree rooted with ribotype 2 (RIB2) derived from subset of the multiple sequence-structure alignment of the ITS2 region (A) using a GTR substitution model. Bootstrap values >50 derived from NJ, maximum parsimony (MP) and maximum likelihood (ML) analyses are mapped to above, below, and to the right of the nodes. (C) An example of ITS2 secondary structure from the *Amoebophrya* RIB2 clade. (D) Compensatory base changes (CBCs) between the eight *Amoebophrya* ribotypes (RIBs 1-8). (E) SIMKA k-mer genome comparison analysis based on Kulczynski distance. Bootstrap values for terminal nodes are shown.

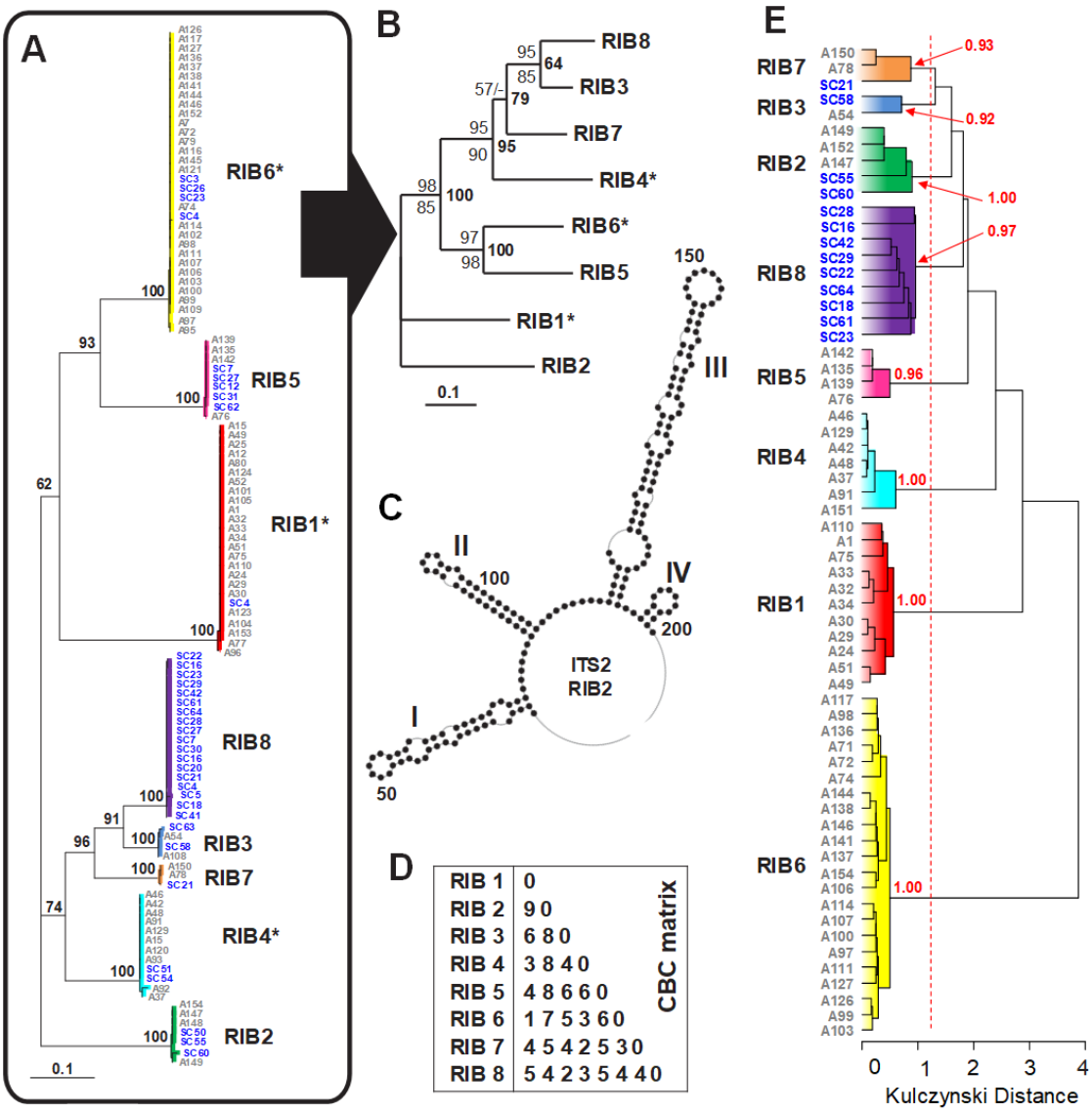


Figure 2: **Phenotypic characteristics of the seven (RIBs 1-7) *Amoebophrya* ribotypes in culture.**

(A-E) Boxplots showing predicted genome sizes (A), estimated number of ribosomal operons (B), and flow cytometry signatures (based on FSC (C), SSC (D), and green autofluorescence at 405 nm (E) for the seven cultivated *Amoebophrya* ribotypes. Horizontal lines in the boxplots indicate the median values. (F) Heatmap showing the results of the crossing infection experiments where 36 strains of *Amoebophrya* were exposed to 54 host strains belonging to 9 species (see Table S2 and Figure S3 for details on the host strains). (G) Non-metric multidimensional scaling (NMDS) ordination diagram assessing the relative importance of six phenotypic characters (blue vectors) in differentiating various *Amoebophrya* strains. The three clusters of *Amoebophrya* strains defined by *k*-mean are depicted by dashed grey lines. The main characters contributing to the separation of strains (establish by *envfit* function) are indicated with asterisks. Operon = number of ribosomal operons; Green = green fluorescence; Genome = genome size; Host = maximal number of infected hosts per strain in cross-infection experiments.

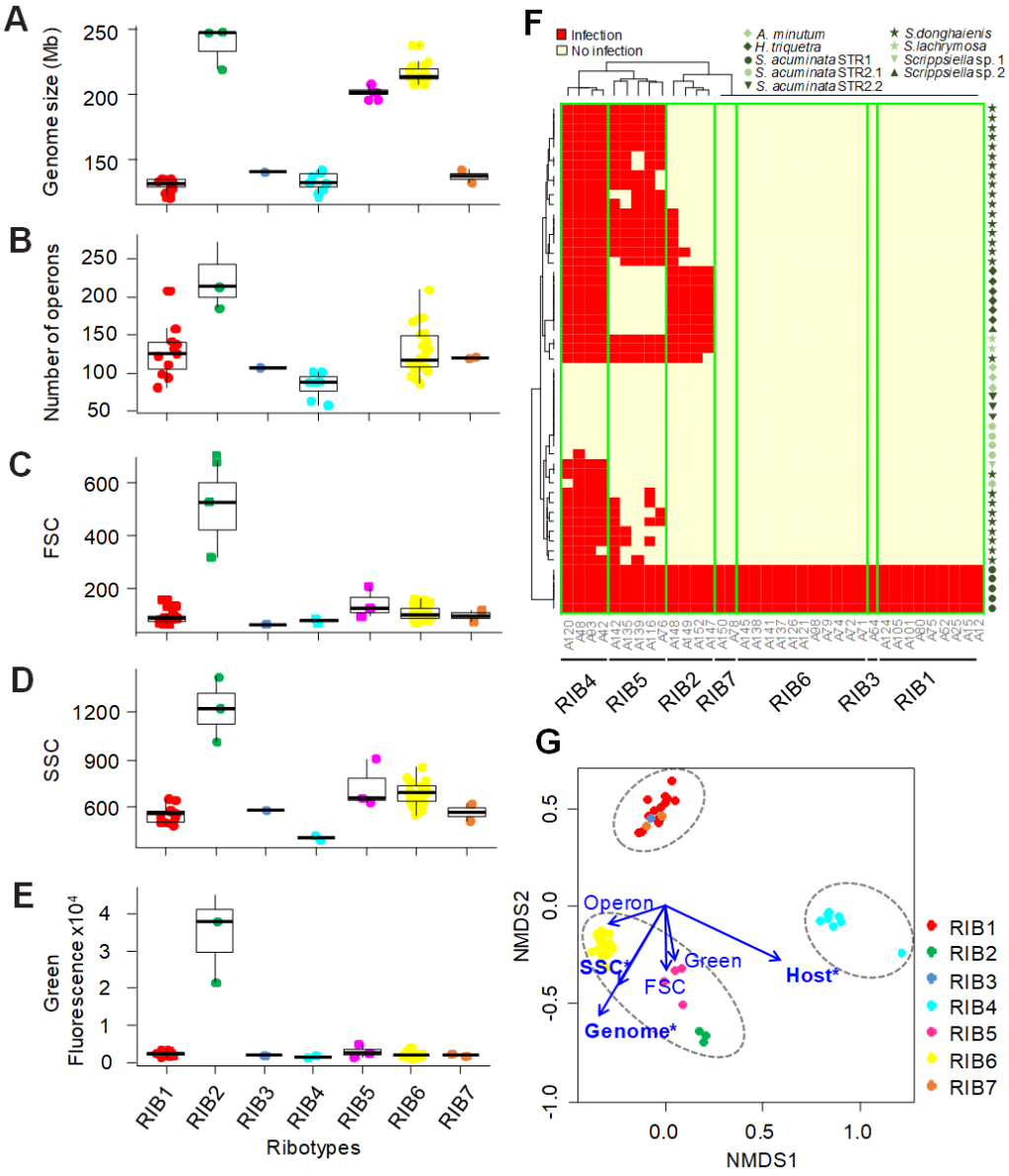


Figure 3: **Environmental monitoring of the eight ribotypes in the Penzé estuary during a three-year survey of summer dinoflagellate blooms.**

(A) Relative abundance (in % of total reads) of *Amoebophrya* ribotypes in the Penzé Estuary (summers of 2010, 2011, and 2012) based on the V4 SSU rDNA metabarcoding analysis. RIBs 3 and 8 were jointly quantified as they could not be differentiated using this marker. (B) Ordination diagram originated from the outlying mean index (OMI) analysis showing the distribution of the samples from the three years in the environmental space determined by the abiotic variables (blue vectors): temperature (Temp), salinity (Sal), precipitation (Prec), tide coefficient (Coef), and nutrients (NO_3, PO_4, $SiOH_4$). (C) Distribution of the kernel densities of the different ribotypes in the OMI multivariate space. The color gradient from yellow to red represents the density (from low to high, respectively), whereas the black dots correspond to the environmental samples shown in (B). (D) Heatmap showing similarities between ribotypes based on the pairwise *D* metric (i.e., niche overlap) calculated using the kernel densities showed in C. (E-F) Relationship between the host range (number of host infected by each ribotype detected in the cross-infection experiments) and the field population fitness, such as the normalized maximal abundance of ribotypes (E) and their permanence in days in the ecosystem (F). Horizontal lines indicate the median for the different parameters. The red brackets indicate the significant differences between clusters pointed out by the Dunn test. (* $p < 0.05$).

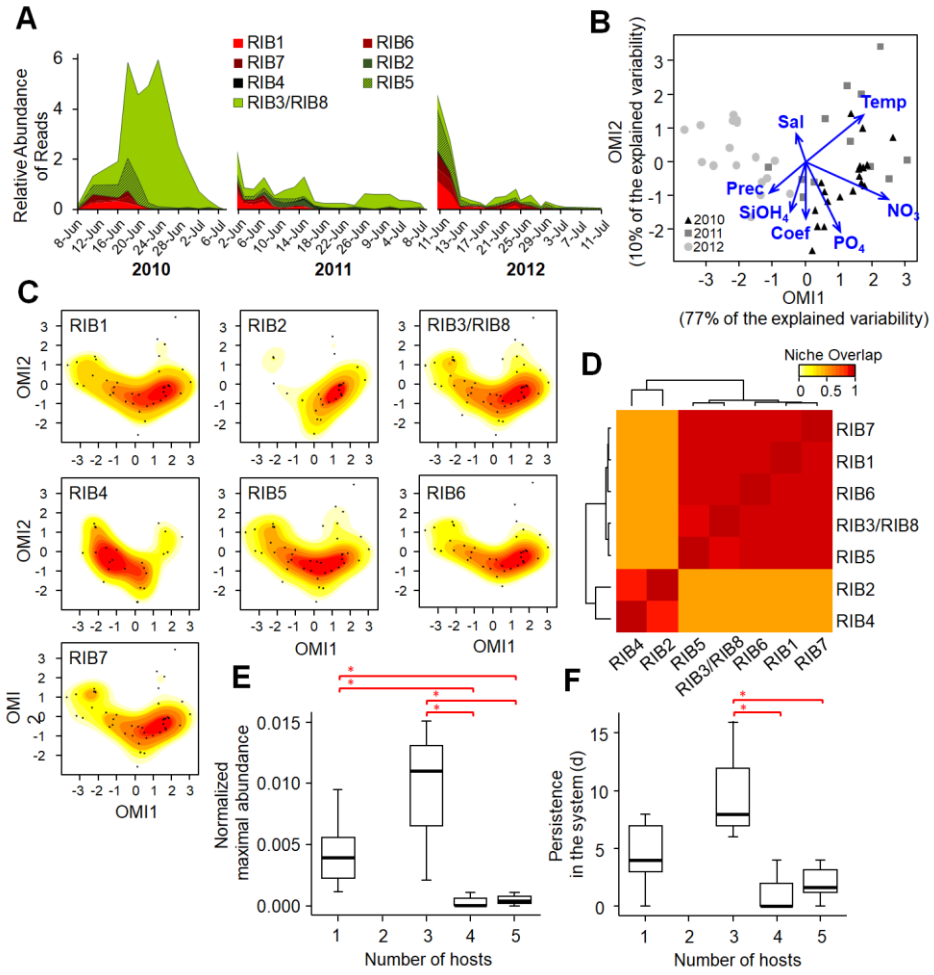


Figure S1: Phylogeny of Syndiniales Group II (MALV-II) based upon the V4 region of the SSU rDNA gene.

Collapsed PhyML phylogenetic tree of 1,550 sequences 511 bp in length of the SSU rDNA corresponding to the V4 region using the GTR + G model. All selected strains from this study (in blue) belong to the MALVII (Amoebophryidae) clades 2 and 4 nomenclature following the one from Guillou et al. (2008). Terminal values represent the number of sequences per clade. Bootstraps values based on 100 replicates of terminal nodes are shown for the MALVII clades 2 and 4.

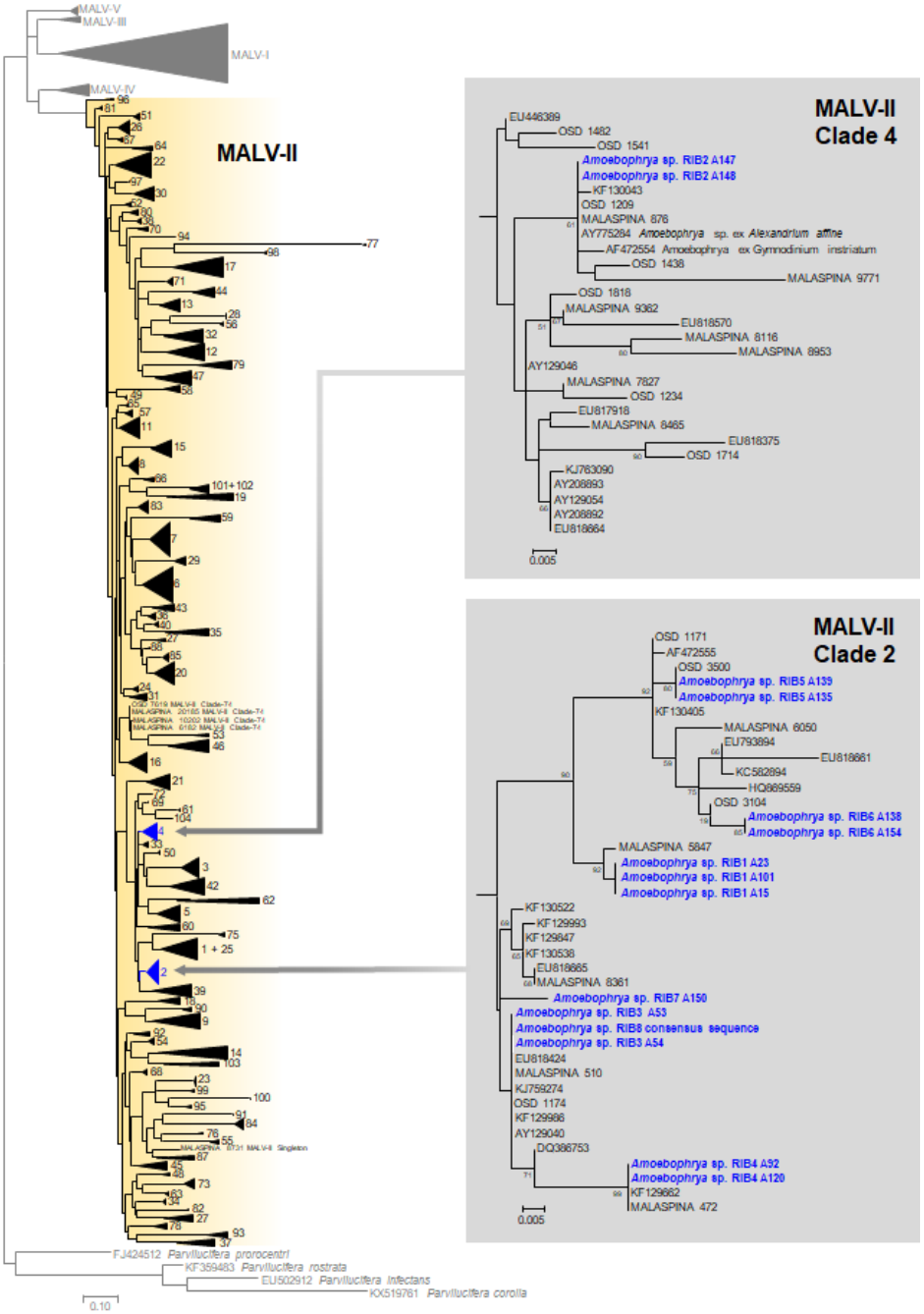


Figure S2: Phylogeny of the *Scrippsiella* spp. strains used in this study based on the D1-D2 domains of the LSU rDNA gene.

PhyML phylogeny based upon analysis of sequences 705 bp in length of the D1-D2 region of the LSU rDNA gene using the GTR + G model. Bootstrap values (> 70%) based on 100 replicates for the main clades is shown.

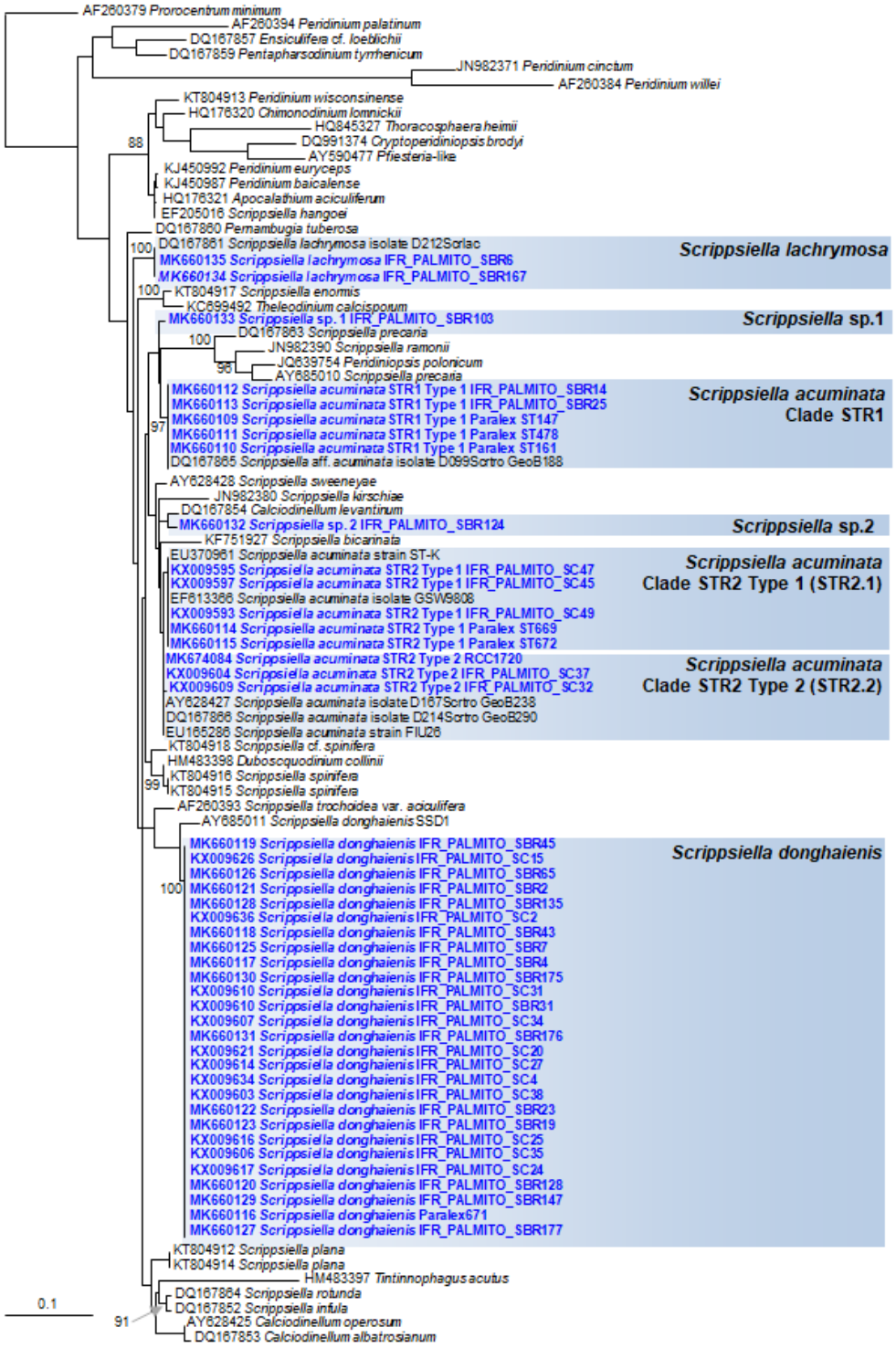


Figure S3: Phylogeny of *Scrippsiella* strains used in the cross-infection experiments, based on the D1-D2 domains of the LSU rDNA gene. Strains used in this study are depicted in blue (see Table S2 for details).

Table S1: List of strains used in this study, including species identity RCC number (Roscoff Culture Collection, http://roscoff-culture-collection.org/), GenBank accession numbers for marker genes, host used for parasitic strains in culture, geographical origin, and date of isolation. The three last columns indicate the analysis that used the different strains.

Species	Mode	ID_strain/SC	Host_during isolation	Host_present	Roscoff Culture Collection (RCC)	GenBank acc. Number	Origin	Date of isolation (datation of sediment)
Amoebophrya RIB1	Strain	A1	ST147	ST161	RCC5984	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	23/06/2007
Amoebophrya RIB6	Strain	A100	ST161	ST161	RCC5997	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB1	Strain	A101	ST161	ST161	RCC4387	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB6	Strain	A102	ST147	ST161	RCC6079	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB6	Strain	A103	ST161	ST161	RCC5998	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB1	Strain	A104	ST147	ST147	LOST	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	03/06/2011
Amoebophrya RIB1	Strain	A105	ST147	ST161	RCC4388	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	03/06/2011
Amoebophrya RIB6	Strain	A106	ST147	ST161	RCC5999	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	03/06/2011
Amoebophrya RIB6	Strain	A107	ST161	ST161	RCC6000	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	03/06/2011
Amoebophrya RIB3	Strain	A108	ST147	ST147	LOST	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	07/06/2011
Amoebophrya RIB6	Strain	A109	ST161	ST161	RCC6080	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	07/06/2011
Amoebophrya RIB1	Strain	A110	ST147	ST161	RCC6001	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	07/06/2011
Amoebophrya RIB6	Strain	A111	ST161	ST161	RCC6002	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB6	Strain	A112	ST147	ST161	RCC6081	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB6	Strain	A114	ST147	ST161	RCC6003	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	13/06/2011
Amoebophrya RIB6	Strain	A116	ST147	ST161	RCC4399	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	13/06/2011
Amoebophrya RIB6	Strain	A117	ST147	ST161	RCC6004	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	13/06/2011
Amoebophrya RIB1	Strain	A12	ST147	ST161	RCC4382	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	23/06/2007
Amoebophrya RIB4	Strain	A120	HT150	HT150	RCC4398	XXXXXXXXX (18S, ITS1, 5.8S, ITS2) from Matthieu	Penzé estuary	13/06/2011
Amoebophrya RIB6	Strain	A121	ST161	ST161	RCC4409	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB1	Strain	A123	ST147	ST161	RCC6082	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB1	Strain	A124	ST161	ST161	RCC4389	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB6	Strain	A126	ST161	ST161	RCC4410	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB6	Strain	A127	ST161	ST161	RCC6005	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB4	Strain	A129	HT150	HT150	RCC6006	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB5	Strain	A135	ST147	ST161	RCC4402	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Rance estuary	29/05/2011
Amoebophrya RIB6	Strain	A136	ST147	ST161	RCC6007	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	01/06/2011
Amoebophrya RIB6	Strain	A137	ST147	ST161	RCC4411	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	01/06/2011
Amoebophrya RIB6	Strain	A138	ST147	ST161	RCC4413	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB5	Strain	A139	ST147	ST161	RCC4403	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	01/06/2011
Amoebophrya RIB6	Strain	A141	ST147	ST161	RCC4412	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	08/06/2011
Amoebophrya RIB5	Strain	A142	ST147	ST161	RCC4401	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	22/06/2011
Amoebophrya RIB6	Strain	A144	ST147	ST161	RCC6008	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	01/06/2011
Amoebophrya RIB6	Strain	A145	ST161	ST161	RCC4414	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB6	Strain	A146	ST161	ST161	RCC6009	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB2	Strain	A147	ST147	ST161	RCC4390	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	08/07/2011
Amoebophrya RIB2	Strain	A148	ST147	ST161	RCC4391	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	08/07/2011
Amoebophrya RIB2	Strain	A149	ST147	ST161	RCC4392	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	08/07/2011

Amoebophrya RIB1	Strain	A15	ST147	ST161	RCC4381	HQ658161 (18S), XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	23/06/2007
Amoebophrya RIB7	Strain	A150	ST161	ST161	RCC4416	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/07/2011
Amoebophrya RIB4	Strain	A151	HT150	HT150	LOST	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	08/07/2011
Amoebophrya RIB6	Strain	A152	ST161	ST161	RCC4393	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/07/2011
Amoebophrya RIB1	Strain	A153	ST161	ST161	RCC6083	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB2	Strain	A154	ST161	ST161	RCC6010	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB1	Strain	A24	ST147	ST161	RCC5985	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB1	Strain	A25	ST147	ST161	RCC4383	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB1	Strain	A29	ST147	ST161	RCC5986	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB1	Strain	A30	ST147	ST161	RCC5987	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB1	Strain	A32	ST147	ST161	LOST	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB1	Strain	A33	ST147	ST161	RCC5988	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB1	Strain	A34	ST147	ST161	RCC5989	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	15/06/2009
Amoebophrya RIB4	Strain	A37	ST147	ST161	RCC5990	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB4	Strain	A42	ST147	ST161	RCC4395	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB4	Strain	A46	ST147	HT150	RCC5991	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB4	Strain	A48	ST147	ST161	RCC4396	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB1	Strain	A49	ST147	ST161	RCC5992	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	19/06/2009
Amoebophrya RIB1	Strain	A51	ST147	ST161	RCC5993	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	19/06/2009
Amoebophrya RIB1	Strain	A52	ST147	ST161	RCC4384	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	19/06/2009
Amoebophrya RIB3	Strain	A54	ST147	ST161	RCC4394	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	19/06/2009
Amoebophrya RIB6	Strain	A71	ST161	ST161	RCC4404	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	11/06/2010
Amoebophrya RIB6	Strain	A72	ST161	ST161	RCC4405	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	11/06/2010
Amoebophrya RIB6	Strain	A74	ST161	ST161	RCC4406	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	11/06/2010
Amoebophrya RIB1	Strain	A75	ST161	ST161	RCC4385	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	11/06/2010
Amoebophrya RIB5	Strain	A76	ST161	ST161	RCC4400	XXXXXXXXX (18S, ITS1, 5.8S, ITS2)	Penzé estuary	11/06/2010
Amoebophrya RIB1	Strain	A77	ST161	ST161	RCC6084	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	18/06/2010
Amoebophrya RIB7	Strain	A78	ST161	ST161	RCC4415	XXXXXXXXX (18S,ITS1-5.8S-ITS2)	Penzé estuary	03/07/2010
Amoebophrya RIB6	Strain	A79	ST161	ST161	RCC4407	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	03/07/2010
Amoebophrya RIB1	Strain	A80	ST161	ST161	RCC4386	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	03/07/2010
Amoebophrya RIB4	Strain	A91	HT150	HT150	RCC5994	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB4	Strain	A92	HT150	HT150	RCC6085	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB4	Strain	A93	HT150	HT150	RCC4397	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	18/06/2009
Amoebophrya RIB6	Strain	A95	ST147	ST161	RCC6096	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB1	Strain	A96	ST147	ST161	RCC6087	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB6	Strain	A97	ST147	ST161	RCC5995	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB6	Strain	A98	ST161	ST161	RCC4408	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	01/06/2011
Amoebophrya RIB6	Strain	A99	ST161	ST161	RCC6088	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	01/06/2011
Scrippsiella acuminata STR1	Strain	Paralex 147	NA	NA	RCC1627	MK660109 (LSU)	Penzé estuary	2005

Heterocapsa triquetra	Strain	Paralex 150	NA	NA	RCC3596	MK660139 (LSU)	Penzé estuary	06/07/2007
Scrippsiella acuminata STR1	Strain	Paralex 161	NA	NA	RCC6094	MK660110 (LSU)	Penzé estuary	2005
Alexandrium minutum	Strain	Paralex 176	NA	NA	RCC3018	MK660136 (LSU)	Morlaix Bay	1989
Alexandrium minutum	Strain	Paralex 331	NA	NA	RCC3145	MK660137 (LSU)	Penzé estuary	02/06/2010
Heterocapsa triquetra	Strain	Paralex 36	NA	NA	LOST	MK660140 (LSU)	Penzé estuary	28/06/2007
Scrippsiella acuminata STR1	Strain	Paralex 478	NA	NA	RCC3048	MK660111 (LSU)	Penzé estuary	22/06/2010
Heterocapsa triquetra	Strain	Paralex 668	NA	NA	RCC3044	MK660141 (LSU)	Penzé estuary	11/07/2011
Scrippsiella acuminata STR2 Type 1	Strain	Paralex 669	NA	NA	RCC3049	MK660114 (LSU)	Penzé estuary	11/07/2011
Heterocapsa triquetra	Strain	Paralex 670	NA	NA	RCC3043	MK660142 (LSU)	Penzé estuary	11/07/2011
Scrippsiella donghaienis	Strain	Paralex 671	NA	NA	RCC3047	MK660116 (LSU)	Penzé estuary	11/07/2011
Scrippsiella acuminata STR2 Type 1	Strain	Paralex 672	NA	NA	LOST	MK660115 (LSU)	Penzé estuary	11/07/2011
Heterocapsa triquetra	Strain	Paralex 694	NA	NA	LOST	MK660143 (LSU)	Penzé estuary	15/06/2011
Heterocapsa triquetra	Strain	Paralex 836	NA	NA	LOST	MK660144 (LSU)	Penzé estuary	11/06/2011
Alexandrium minutum	Strain	Paralex 873	NA	NA	RCC3278	MK660138 (LSU)	Rance estuary	05/06/2011
Scrippsiella sp.2	Strain	IFR_PALMITO_SBR103	NA	NA	RCC6113	MK660133 (LSU)	Penzé estuary	08/04/2014 (2000 +/- 4.1)
Scrippsiella sp. 1	Strain	IFR_PALMITO_SBR124	NA	NA	LOST	MK660132 (LSU)	Penzé estuary	14/04/2014 (1998 +/- 4,6)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR128	NA	NA	RCC6100	MK660120 (LSU)	Penzé estuary	16/05/2014 (1998 +/- 4.6)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR135	NA	NA	RCC6115	MK660128 (LSU)	Penzé estuary	08/04/2014 (2000 +/- 4.1)
Scrippsiella acuminata STR1	Strain	IFR_PALMITO_SBR14	NA	NA	RCC6116	MK660112 (LSU)	Penzé estuary	24/02/2014 (2006 +/-2.3)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR147	NA	NA	RCC6101	MK660129 (LSU)	Penzé estuary	05/06/2014 (2000 +/- 4.1)
Scrippsiella lachrymosa	Strain	IFR_PALMITO_SBR167	NA	NA	LOST	MK660134 (LSU)	Brest rade	16/05/2014 (1993 +/- 1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR175	NA	NA	RCC6117	MK660130 (LSU)	Penzé estuary	27/05/2014 (1999 +/- 4,3)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR176	NA	NA	RCC6118	MK660131 (LSU)	Penzé estuary	27/05/2014 (1999 +/- 4,4)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR177	NA	NA	RCC6102	MK660127 (LSU)	Penzé estuary	27/05/2014 (1998 +/- 4.6)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR19	NA	NA	RCC6103	MK660123 (LSU)	Penzé estuary	24/02/2014 (2006 +/-2.3)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR2	NA	NA	RCC6104	MK660121 (LSU)	Penzé estuary	24/02/2014 (2000 +/- 4.1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR23	NA	NA	RCC6105	MK660122 (LSU)	Penzé estuary	28/02/2014 (2006 +/-2.3)
Scrippsiella acuminata STR1	Strain	IFR_PALMITO_SBR25	NA	NA	RCC6106	MK660113 (LSU)	Penzé estuary	28/02/2014 (2006 +/-2.3)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR31	NA	NA	RCC6107	MK660124 (LSU)	Penzé estuary	28/02/2014 (2002 +/- 3.5)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR4	NA	NA	RCC6108	MK660117 (LSU)	Penzé estuary	24/02/2014 (2000 +/- 4.1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR43	NA	NA	RCC6109	MK660118 (LSU)	Penzé estuary	28/02/2014 (2002 +/- 3.5)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR45	NA	NA	RCC6110	MK660119 (LSU)	Penzé estuary	28/02/2014 (2002 +/- 3.5)
Scrippsiella lachrymosa	Strain	IFR_PALMITO_SBR6	NA	NA	LOST	MK660135 (LSU)	Penzé estuary	18/03/2014 (2006 +/- 2,3)

Scrippsiella donghaienis	Strain	IFR_PALMITO_SBR65	NA	NA	RCC6111	MK660126 (LSU)	Penzé estuary	03/03/2014 (2006 +/-2.3)
-*Scrippsiella donghaienis*	Strain	IFR_PALMITO_SBR7	NA	NA	RCC6112	MK660125 (LSU)	Penzé estuary	24/02/2014 (2000 +/- 4.1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC15	NA	NA	RCC4734	KX009626 (LSU)	Brest rade	24/02/2014 (1986 +/- 2)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC2	NA	NA	RCC4733	KX009636 (LSU)	Brest rade	24/02/2014 (1991 +/- 1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC20	NA	NA	RCC4722	KX009621 (LSU)	Brest rade	24/02/2014 (1986 +/- 2)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC24	NA	NA	RCC4715	KX009617 (LSU)	Brest rade	28/02/2014 (1995 +/- 1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC25	NA	NA	RCC4716	KX009616 (LSU)	Brest rade	28/02/2014 (1995 +/- 1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC27	NA	NA	RCC4723	KX009614 (LSU)	Brest rade	28/02/2014 (1978 +/- 2)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC31	NA	NA	RCC4726	KX009610 (LSU)	Brest rade	28/02/2014 (2002 +/-3.5)
Scrippsiella acuminata STR2_Type2	Strain	IFR_PALMITO_SC32	NA	NA	RCC6120	KX009609 (LSU)	Brest rade	28/02/2014 (2003 +/-1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC34	NA	NA	RCC4711	KX009607 (LSU)	Brest rade	28/02/2014 (2010 +/-1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC35	NA	NA	RCC4712	KX009606 (LSU)	Brest rade	28/02/2014 (2010 +/-1)
Scrippsiella acuminata STR2_Type2	Strain	IFR_PALMITO_SC37	NA	NA	RCC4732	KX009604 (LSU)	Brest rade	28/02/2014 (2006 +/-1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC38	NA	NA	RCC4713	KX009603 (LSU)	Brest rade	28/02/2014 (2006 +/-1)
Scrippsiella donghaienis	Strain	IFR_PALMITO_SC4	NA	NA	RCC6119	KX009634 (LSU)	Brest rade	24/02/2014 (1991 +/- 1)
Scrippsiella acuminata STR2_Type1	Strain	IFR_PALMITO_SC45	NA	NA	RCC4728	KX009597 (LSU)	Brest rade	28/02/2014 (2006 +/-1)
Scrippsiella acuminata STR2_Type1	Strain	IFR_PALMITO_SC47	NA	NA	RCC6121	KX009595 (LSU)	Brest rade	28/02/2014 (2001 +/-1)
Scrippsiella acuminata STR2_Type1	Strain	IFR_PALMITO_SC49	NA	NA	RCC4729	KX009593 (LSU)	Brest rade	28/02/2014 (1997 +/- 1)
Amoebophrya RIB5	Single cell	RIB12	Heterocaspa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	17/06/2010
Amoebophrya RIB8	Single cell	RIB16	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	24/06/2010
Amoebophrya RIB8	Single cell	RIB16	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB8	Single cell	RIB18	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	24/06/2010
Amoebophrya RIB8	Single cell	RIB20	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB7	Single cell	RIB21	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	04/06/2011
Amoebophrya RIB8	Single cell	RIB21	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB8	Single cell	RIB22	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB6	Single cell	RIB23	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	04/06/2011
Amoebophrya RIB8	Single cell	RIB23	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	08/06/2011
Amoebophrya RIB6	Single cell	RIB26	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB5	Single cell	RIB27	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB8	Single cell	RIB27	Heterocapsa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	11/06/2011
Amoebophrya RIB8	Single cell	RIB28	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Rance estuary	05/06/2011

Amoebophrya RIB8	Single cell	RIB29	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	11/06/2011
Amoebophrya RIB6	Single cell	RIB3	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	06/06/2011
Amoebophrya RIB8	Single cell	RIB30	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB5	Single cell	RIB31	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	05/06/2011
Amoebophrya RIB8	Single cell	RIB36	Heterocapsa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	11/06/2011
Amoebophrya RIB2	Single cell	RIB39	Heterocapsa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	11/06/2011
Amoebophrya RIB1	Single cell	RIB4	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	06/06/2011
Amoebophrya RIB8	Single cell	RIB4	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	28/05/2011
Amoebophrya RIB6	Single cell	RIB41	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	09/06/2011
Amoebophrya RIB8	Single cell	RIB41	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB8	Single cell	RIB42	Scrippsielloid	NA	NA	XXXXXXXXX (18S,ITS1-5.8S-ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB6	Single cell	RIB45	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB3	Single cell	RIB46	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB8	Single cell	RIB5	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	28/05/2011
Amoebophrya RIB2	Single cell	RIB50	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB4	Single cell	RIB51	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB8	Single cell	RIB53	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB4	Single cell	RIB54	Heterocapsa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB2	Single cell	RIB55	Heterocapsa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB8	Single cell	RIB56	Heterocapsa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB3	Single cell	RIB58	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB2	Single cell	RIB59	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB2	Single cell	RIB60	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	14/06/2011
Amoebophrya RIB8	Single cell	RIB61	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	17/06/2011
Amoebophrya RIB5	Single cell	RIB62	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	17/06/2011
Amoebophrya RIB3	Single cell	RIB63	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	17/06/2011
Amoebophrya RIB8	Single cell	RIB64	Scrippsielloid	NA	NA	XXXXXXXXX (18S, ITS1-5.8S-ITS2)	Penzé estuary	17/06/2011
Amoebophrya RIB5	Single cell	RIB7	Heterocaspa triquetra	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Penzé estuary	16/06/2010
Amoebophrya RIB8	Single cell	RIB7	Scrippsielloid	NA	NA	XXXXXXXXX (ITS1, 5.8S, ITS2)	Rance estuary	28/05/2011
Scrippsiella acuminata STR2_Type2	Strain		NA	NA	RCC1720	MK674084 (LSU)	SOMLIT ASTAN	13/05/2008

Table S2: Statistic of the genome assemblies: read number (total and after filtration), N50 (Total and > 1000 kb), remaping rate of reads and average coverage.

Strain or Single cell	Total reads	Filtered reads count	N50	N50 (>1000bp)	Remapping rate	Average coverage
A1	32 597 788	27545422	40522	41612	0.9316	41.2111
A24	22 170 180	15938544	26540	28275	0.9189	25.4989
A29	65979782	52436534	47986	48999	0.9561	85.8073
A30	51615038	40469950	47467	48447	0.9561	67.6153
A32	40 027 570	31811814	47946	49352	0.9373	47.0513
A33	53163770	42648628	46543	48069	0.954	69.5386
A34	70966608	58662234	38412	40462	0.9504	88.9921
A37	64406860	53939788	23421	24880	0.9552	80.2298
A42	41 309 688	37943270	16486	17556	0.9009	34.3543
A46	30 299 444	26409052	18302	18900	0.9378	38.527
A48	53343078	45069552	23195	23880	0.9525	69.1362
A49	73049494	57191260	49348	52032	0.9554	92.5387
A51	56891848	45454372	42243	48507	0.9438	67.2662
A54	41345204	31884946	34790	36036	0.9373	64.1275
A71	45778458	41698404	16451	17719	0.9199	38.1075
A72	32 203 188	29695764	14883	15657	0.8859	26.4288
A74	40 360 306	16201658	7055	8488	0.9266	27.0515
A75	56756818	45081158	47672	49291	0.9542	73.9866
A76	28 310 694	26250730	16308	16969	0.9003	23.5293
A78	63266228	20037572	19235	23889	0.9294	67.8865
A91	49042824	43141252	15147	19085	0.8651	49.2374
A97	46 036 804	39440992	22806	23423	0.9461	58.8624
A98	46 570 476	42506420	16426	17535	0.9091	37.7321
A99	42878088	39009314	16915	17947	0.9156	36.1088
A100	66401074	59685208	17822	18875	0.9216	54.6545
A103	57958210	52046724	18496	19384	0.9237	49.3881
A106	57259114	51905410	18289	19244	0.9201	48.6952
A107	56306672	52267272	9044	10905	0.8852	42.0832
A110	40100050	32197968	38645	42651	0.9383	45.1752
A111	61902422	56453154	16001	18152	0.9122	48.8841
A114	54334936	49506978	14849	16764	0.9161	41.067
A117	52766046	48775460	12750	14522	0.9006	41.2465

A126	52654516	48383588	14574	16522	0.9032	41.9721
A127	43 000 160	39483346	16318	17649	0.9051	34.7821
A129	38 361 080	33170086	20514	21069	0.9441	49.2397
A135	41 192 256	37716018	21052	21710	0.9081	34.9305
A136	44 945 232	40804012	17800	18824	0.9004	38.0065
A137	68183234	62126772	16138	17436	0.9026	55.786
A138	61799720	55524646	18723	19606	0.922	53.5879
A139	39 978 988	36850220	18178	18919	0.9057	33.5654
A141	58564830	51076030	16969	19812	0.9251	42.9193
A142	43673922	39927002	19750	20451	0.9231	37.511
A144	42098046	38433110	17111	18065	0.9162	36.6419
A146	42 121 926	38857576	17072	18094	0.8963	35.2477
A147	49472368	39955422	30574	33333	0.922	33.1948
A149	49604200	40403686	30784	34095	0.9219	32.9939
A150	70304300	58422836	22880	25568	0.9319	95.1714
A151	39 005 356	6542732	3123	12835	0.9469	20.1562
A152	57278004	53076884	13648	15424	0.9044	46.579
A154	71473328	55056178	36810	38011	0.9433	51.4394
PZ10_SC18	31961110	3038564	3429	5790	0.9587	46.4045
PZ11_SC16	33792376	18817268	2420	3793	0.9156	128.088
PZ11_SC20	34209206	26460326	1520	2926	0.5371	300.497
PZ11_SC22	34306130	29688056	2688	4088	0.8808	37.6085
PZ11_Sc23	33718280	30468090	2290	3486	0.8777	181.152
PZ11_SC29	33697940	31584928	4370	5568	0.9336	55.9334
PZ11_SC41	49321478	46204634	1540	3407	0.8537	141.487
PZ11_SC42	48094994	40317924	5334	6703	0.9074	59.5274
PZ11_SC55	33797508	29474132	4631	6693	0.851	74.1605
PZ11_SC58	33633210	25078280	2495	3674	0.9368	128.145
PZ11_SC60	34306696	31837950	2311	3757	0.8637	89.619
PZ11_SC61	33555166	30037224	3230	4638	0.9487	136.417
PZ11_SC64	33404478	27752464	3225	4660	0.924	68.2133
RC11_SC21	34943860	17228812	1699	4692	0.8715	88.4824
RC11_SC28	37527660	35263616	1906	3851	0.5883	106.311
RC11_SC30	42337260	41869542	1985	3339	0.5277	245.002
RC11_SC5	29459360	13402234	1363	2901	0.8861	482.56

Table S3: Nucleotide differences (top right) and percent identity (bottom left) of the complete SSU rDNA gene (top in each cell), 18S-V4 (middle in each cell) and 18S-V9 (bottom in each cell) regions between the eight ribotypes defined in this study.

Table S3. Variation of the SSU rDNA between ribotypes.

	RIB1	RIB2	RIB3	RIB4	RIB5	RIB6	RIB7	RIB8
RIB1		**125** 29 25	**54** 10 8	**61** 13 9	**50** 10 6	**56** 13 5	**60** 13 9	**55** 10 9
RIB2	92.89% 92.39% 80.92%		**111** 20 25	**122** 28 26	**109** 28 23	**111** 26 24	**112** 21 25	**107** 20 24
RIB3	96.93% 97.38% 93.89%	93.68% 94.75% 80.92%		**43** 10 5	**55** 12 7	**54** 13 6	**18** 5 1	**4** 0 1
RIB4	96.53% 96.59% 93.13%	93.06% 92.65% 80.15%	97.55% 97.38% 96.18%		**65** 17 11	**71** 18 10	**41** 11 4	**43** 10 6
RIB5	97.15% 97.38% 95.42%	93.80% 92.65% 82.44%	96.87% 96.85% 94.66%	96.30% 95.54% 91.60%		**22** 6 1	**60** 16 8	**53** 12 8
RIB6	96.81% 96.59% 96.18%	93.68% 93.18% 81.68%	96.93% 96.59% 95.42%	95.96% 95.28% 92.37%	98.75% 98.43% 99.24%		**59** 17 7	**52** 13 7
RIB7	96.59% 96.59% 93.13%	93.63% 94.49% 80.92%	98.98% 98.69% 99.24%	97.67% 97.11% 96.95%	96.59% 95.80% 93.89%	96.64% 95.54% 94.66%		**20** 5 2
RIB8	96.87% 97.38% 93.13%	93.91% 94.75% 81.68%	99.77% 100.00% 99.24%	97.55% 97.38% 95.42%	96.98% 96.85% 93.89%	97.04% 96.59% 94.66%	98.86% 98.69% 98.47%	

Supplementary methods.

This document contain supplementary informations on the sampling strategy, culturing effort, single-cells isolation, identification tehcniques of parasites and hosts and genome analysis.

SUPPLEMENTARY METHODS

1. SAMPLING STRATEGY

We sampled two estuaries distant of each other by approximately ~150 km; the Penzé Estuary (48°37’37.57"N, 3°57’13.17"W) and the Rance Estuary (48°31’49.61"N, 1°58’21.81"W), both located in the western Channel (France). Planktonic communities were monitored every 1-2 days during early summer (May to July) over 8 years (2004-2007, 2009, 2010-2012) for the Penzé Estuary and in 2011 for the Rance Estuary. A portable probe was used to measure *in situ* temperature and salinity. Samples were rapidly (less than 2 hours) filtrated through a series of different-size filters (10 µm, 3 µm, 0.2 µm), flash-frozen in liquid nitrogen, and stored at -80°C for further genetic analyses. Abiotic parameters recorded included salinity, temperature (air and water), nutrients (among the most important are NO_3, NH_4, and PO_4), rainfall and light intensity. Biotic parameters include Lugol-fixed cells (> 10 µm) and flow cytometry to count bacteria, viruses, cyanobacteria, picoeukaryotes and phototrophic cryptophytes (based on their pigment and DNA contents). Detailed information on the sampling strategy and data acquisition can be found in [1][2][3].

2. CULTURING

Isolation of dinoflagellates hosts

We isolated dinoflagellates in culture by micropipetting during the whole monitoring period and later used these strains to maintain the parasites in culture and to screen their host range.

Isolation of *Amoebophrya* strains

To isolate *Amoebophrya* strains, we used 24-well plates that we incubated with 1 ml of a healthy host strain (or a mix of strains) supplemented with either 1) 1 ml of field sample filtered through polycarbonate filters (3-5 µm pore size) (fraction presumably containing dinospores), or 2) one infected host cell isolated by micropipetting (Box 1). We obtained similar percentages of infection success using both methods (8-10 %), with more chance of success observed when starting with a mix of host species.

Box 1: Percentage of successful infections, either by using direct incubation of field filtered sea water or isolated single infected host cell

Host species	Number of incubations	Number of starting infections	Success rate (%)
Alexandrium minutum	438	0	0
Gymnodinium sp.	7	0	0
Heterocapsa rotundata	12	0	0
Heterocapsa triquetra	194	5	2.58
Prorocentrum micans	9	0	0
Scrippsiella donghaienis	166	17	10.24
Scrippsiella acuminata STR1	180	50	27.78
*Scrippsiella sp.*1	2	0	0
Mix of species	60	25	41.67
Total	**1068**	**97**	**9.08**
Methodology			
by incubation of field water	976	83	8.50
by isolating one infected host	195	21	10.77

Plates were checked for *Amoebophrya*-like parasites through their natural green autofluorescence using an epifluorescence microscope (BX51, Olympus) equipped with the U-MWB2 cube (450- to 480-nm excitation, 500-nm emission [4]). Overall, we successfully observed newly infections after 3-7 days in 9.08 % of cases (deduced from the 1068 incubations processed in 2010 and 2011), with success rates that depend on host (no infection in *A. minutum*, 2.6% in *H. triquetra*, 10.2% in *S. donghaienis*, 27.8% in *S. acuminata* STR1 type 1, 41.7% using a mix of species). For the strain establishment, a single infected host cell was isolated from those incubations (only one kept per well) by micropipeting, washed three times and newly transferred in the original healthy host. Clonality of strain was ensured by repeating this step 2-5 times.

Maintenance of strains

Host and parasite strains were grown in F/2 medium (Marine Water Enrichment Solution, Sigma), using 0.2 µm-filtered and autoclaved natural seawater from the Penzé Estuary (27 practical salinity units) and stored in the dark for > 3 months. The medium was supplemented with 5% (v/v) soil extract [5]. A final filtration (0.22 µm) was processed under sterile conditions. Cultures were grown at 21°C under continuous light at 100 µEinstein m^2 s^{-1} in ventilated flasks. To maintain parasitic strains, infected hosts were regularly transferred (every 3-7 days) into a healthy host culture on 15 ml culture tubes using a 1/10 dilution rate.

3. SINGLE-CELLS

For single-cells, hosts infected by *Amoebophrya*-like organisms at late-stage infection from freshly collected field samples (less than 3 hours) were sorted individually by micropipeting, and washed three times into filtered sterilized (< 0.2 µm) freshly prepared medium. Hosts were identified according to their morphology, and single-cells were transferred in cryovials with a minimum of medium (3-5 µL), flash-frozen, and store at -80°C. DNA extraction and purification were performed both on pelleted strains and single-cells using the MasterPure kit (Epicentre).

4. IDENTIFICATION OF PARASITES AND HOSTS

Amoebophrya-like individuals

In its initial description, Cachon [6] defined species boundaries within Amoebophryidae based on the specific configuration of the cytopharynx, a structure responsible for the transit of particles from the host to the parasite during the internal development (trophont) stages. The ultrastructure of intracellular stages in dinoflagellate parasites is however highly dependent upon the physiology of the host and the number of co-infections [7]. Moreover, we observed that the *Amoebophrya* sp. strain A120 (RIB 4), which consistently develops in the nucleus of *S. acuminata* in culture conditions, starts its development in the cytoplasm when infecting an alternative host (*Heterocapsa triquetra*) (data not shown). These observations argue against the use of internal parasitic features as stable criteria for taxonomical description, as they are highly dependent on the nature and physiology of the host. We therefore opted for the use of the free-living (dinospore) stage for taxonomic purposes as what is done for other groups

such as Rhizophydiales (see [8]). To do so, we used the forward versus side scatter (FSC vs. SSC) gating from flow cytometric analyses of dinospores to identify cells based on their granularity (complexity) and relative size. We also used the level of the natural green autofluorescence signal emitted by *Amoebophrya* spp. when excited at 405 nm, as a third parameter.

Here, we initially placed all *Amoebophrya*-like individuals based upon the V4 region of their SSU rDNA sequences compared to the initial classification of MALV-II published by [9]. For that, sequences were aligned using MAFFT and the FFT-NS-i refined method [10].

Dinoflagellate hosts

The identity of the hosts was confirmed by sequencing the D1 and D2 domains of the LSU rDNA gene following the procedure explained in [11]. We based our nomenclature of scrippsielloids upon a phylogeny using the D1 and D2 domains of the LSU rDNA genes, after taxonomy of Luo et al. 2016. However, the taxonomy of scrippsielloids is still under construction by experts. The most common species of *Scrippsiella*, i.e. *S. trochoidea* (F.Stein) A.R.Loebl., comprises three genetically diverse clades, designated as STR1, STR2 and STR3 ([12][13]). These genetic clusters should be considered as distinct species. Recently, strains of *S. trochoidea* from the type locality proved to be in STR2, thus STR1 and STR3 might not be true *S. trochoidea* at all ([14][15]). The true *S. trochoidea* (STR2) is now considered a heterotypic synonym of *S. acuminata* (Ehrenb.) Kretschmann, Elbr., Zinssmeister, S. Soehner, Kirsch, Kusber & Gottschling [16], a change that we take into account here, in complementary to the assignation to genetic clade, awaiting for formal description.

5. GENOME ANALYSIS

Sequencing steps

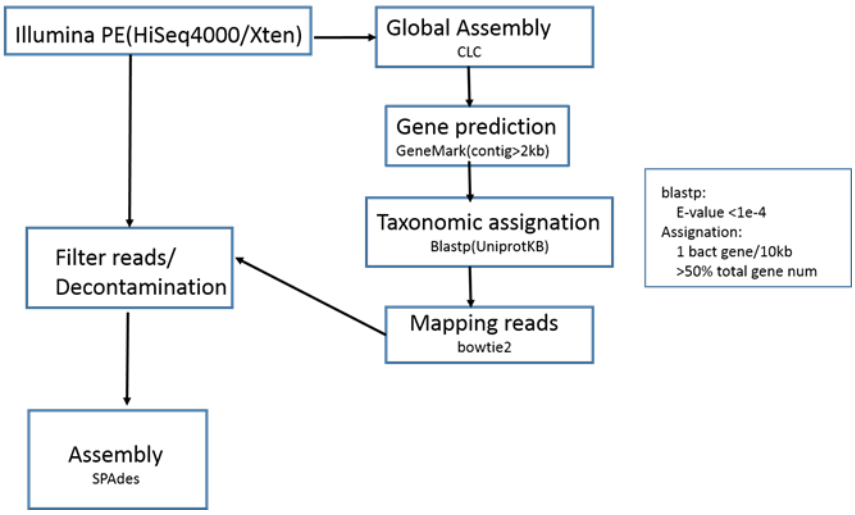
Total DNA was quantified on a Qubit Fluorometer. DNA quality was checked by electrophoresis on a 0.7% agarose gel. We prepared overlapping paired-end libraries from 250 ng of total DNA using a semi-automated protocol. Briefly, DNA was sheared on the Covaris E210 instrument (Covaris, Inc., Woburn, Massachusetts, USA) in order to generate fragments of 150-400 bp in size. End repair, A-tailing and ligation with Illumina compatible adaptors (Bio Scientific Austin, Texas, USA)

were performed using the SPRIWorks Library Preparation System and a SPRI TE instrument (Beckmann Coulter, Danvers, Massachusetts, USA) according to the manufacturer's protocol. The 200-400 bp size fragments were amplified by 12 cycles of PCR with the Pfx Platinum Taq polymerase (ThermoFisher, Waltham, Massachusetts, USA) and Illumina adapter-specific primers. Amplified library fragments were size selected on a 3% agarose gel around 300 bp and purified. We prepared mate-pair (MP) library for A25 using 10 µg of fragmented DNA according to the Illumina protocol (Illumina Mate Pair library kit, Illumina, San Diego, CA). For strain A120, the MP library was prepared with the Nextera Mate Pair Sample Preparation Kit (Illumina) using 4 µg of fragmented DNA.

We evaluated the size of all Illumina libraries on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) machine and quantified them by qPCR with the KAPA Library Quantification Kit (KapaBiosystems Inc., Woburn, MA, USA) on a MxPro instrument (Agilent Technologies). Libraries were then sequenced using the 101 bp paired-end reads chemistry on a HiSeq2000 Illumina sequencer. Few more individuals have been sequenced on an Illumina HiSeq XTEN or BGISEQ-500 platform in BGI (Box 2). After filtering off duplicated, low quality reads and reads with adaptor sequences, 3 - 6 Gb (~15-30 X genome sequencing depth high-quality clean reads were retained for each sample.

Assembling of genomes

A first assembly was processed using CLC assembler, clc_mapper, with the options (-p fb ss 200 800 –q). The bioinformatics pipeline was then customized to remove bacterial contamination in the chart flow below:



We additionally confirmed the identity of each individual by comparing the partial ribosomal operon (SSU rDNA, ITS1, 5.8S, ITS2) extracted from contigs from the one obtained by PCR.

REFERENCES

1. Chambouvet A, Morin P, Marie D, Guillou L. Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* 2008; **322**: 1254–1257.
2. Dia A, Guillou L, Mauger S, Bigeard E, Marie D, Valero M, et al. Spatiotemporal changes in the genetic diversity of harmful algal blooms caused by the toxic dinoflagellate [i]Alexandrium minutum[/i]. *Mol Ecol* 2014; **23**: 549–560.
3. Blanquart F, Valero M, Alves-De-Souza C, Dia A, Lepelletier F, Bigeard E, et al. Evidence for parasite-mediated selection during short-lasting toxic algal blooms. *Proc R Soc B Biol Sci* 2016; **283**.
4. Coats DW, Bockstahler KR. Occurrence of the parasitic dinoflagellate [i]Amoebophrya ceratii[/i] in Chesapeake Bay populations of [i]Gymnodinium sanguineum[/i]. *J Eukaryot Microbiol* 1994; **41**: 586–593.
5. Starr R, Zeikus J. UTEX-The culture collection of algae at the University of Texas at Austin. *J Phycol* 1993; **29**: 1–106.
6. Cachon J. Contribution à l'étude des péridiniens parasites. Cytologie, cycles évolutifs. *Ann des Sci Nat Zool Paris* 1964; 1–158.

7. Figueroa RI, Garcés E, Massana R, Camp J. Description, Host-specificity, and Strain Selectivity of the Dinoflagellate Parasite Parvilucifera sinerae sp. nov. (Perkinsozoa). *Protist* 2008; **159**: 563–578.

8. Lepelletier F, Karpov S a., Alacid E, Le Panse S, Bigeard E, Garcés E, et al. [i]Dinomyces arenysensis[/i] gen. et sp. nov. (Rhizophydiales, Dinomycetaceae fam. nov.), a chytrid infecting marine dinoflagellates. *Protist* 2014; **165**: 230–244.

9. Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, et al. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* 2008; **10**: 3349–3365.

10. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002; **30**: 3059–66.

11. Klouch KZ, Schmidt S, Andrieux-Loyer F, Le Gac M, Hervio-Heath D, Qui-Minet ZN, et al. Historical records from dated sediment cores reveal the multidecadal dynamic of the toxic dinoflagellate Alexandrium minutum in the Bay of Brest (France). *FEMS Microbiol Ecol* 2016; **92**: 1–16.

12. Montresor M, Sgrosso S, Procaccini G, Kooistra WHCF. Intraspecific diversity in [i]Scrippsiella trochoidea[/i] (Dinophyceae): evidence for cryptic species. *Phycologia* 2003; **42**: 56–70.

13. Gottschling M, Knop R, Plötner J, Kirsch M, Willems H, Keupp H. A molecular phylogeny of Scrippsiella sensu lato (Calciodinellaceae, Dinophyta) with interpretations on morphology and distribution. *Eur J Phycol* 2005; **40**: 207–220.

14. Zinssmeister C, Soehner S, Kirsch M, Facher E, Sebastian Meier KJ, Keupp H, et al. Same but different: Two novel bicarinate species of extant calcareous dinophytes (thoracosphaeraceae, peridiniales) from the mediterranean sea. *J Phycol* 2012; **48**: 1107–1118.

15. Soehner S, Zinssmeister C, Kirsch M, Gottschling M. Who am I – and if so, how many? Species diversity of calcareous dinophytes (Thoracosphaeraceae, Peridiniales) in the Mediterranean Sea. *Org Divers Evol* ; **12**: 339–348.

16. Kretschmann J, Elbrächter M, Zinssmeister C, Soehner S, Kirsch M, Kusber WH, et al.

Taxonomic clarification of the dinophyte Peridinium acuminatum Ehrenb. *Phytotaxa* 2015; **17**: 239–256.

Chapter 2

Potential for sexual reproduction in *Amoebophrya* spp. (Syndiniales, dinoflagellates), parasites of dinoflagellates

Chapter 2 Potential for sexual reproduction in *Amoebophrya* spp. (Syndiniales, dinoflagellates), parasites of dinoflagellates

Ruibo Cai, Ehsan Kayal, Erwan Corre, Laure Guillou

Abstract

Sexual reproduction is a hallmark for all the major eukaryotic lineages. Meiosis is an essential step in sexual reproduction and meiotic genes are found to be widely present in eukaryotes and conserved among animals, fungi, plants and protists. The presence of most, or all, of these genes in a genome suggests they have been maintained for meiosis and implicitly sexual reproduction may occur in that species. In contrast, their absence would be consistent with the loss of meiosis and asexuality. In this study, we found that *Amoebophrya* possesses 4 of 9 genes which have been known to function in meiosis specifically. Since many of these genes are members of large gene families, a combination of phylogenetic reconstructions and analysis of sequence domains is used to identify gene identities. This strategy uncovered some putative errors in the previously assigned genes from the literature and provided a better-annotated database for genes putatively involved in sexual reproduction. We also discussed the use of meiosis-specific gene toolkit in a broad range and the possible stage when sexual reproduction occurs in *Amoebophrya*.

Introduction

Sexual reproduction is the dominant mode of reproduction in eukaryotes (Schurko and Logsdon, 2008). The universality of sexual reproduction across the eukaryotic tree of life accentuates its importance in eukaryotic evolution (Dacks and Roger, 1999; Ramesh et al., 2005). The maintenance of sex is crucial for the long-term survival of most eukaryotic lineages (Ramesh et al., 2005; Signorovitch et al., 2005). According to ecological models, sex introduces novel gene combinations that facilitate adaptation to changing environments (Colegrave, 2002; Goddard et al., 2005). Epidemiological studies also suggest that sex may help pathogens spread competitive alleles, allowing them to quickly respond to environmental changes (e.g. host immune response) and to expand their geographic range. For example, one of the most virulent *Toxoplasma* strains arose from sexual recombination between two distinct clonal strains (Grigg et al., 2001; Boyle, 2006). However, the phylogenetic distribution of sexual reproduction has not been widely determined for protists, which represent most of eukaryotic diversity.

Large-scale investigations on marine planktonic protist diversity in global oceans highlighted the ubiquitous occurrence and marked diversity of parasites in marine environment (de Vargas et al., 2015). A large portion of the parasite associations in marine environment involve the Syndiniales MALV-I and MALV-II groups, suggesting Syndiniales may be the major top-down driver for plankton population structuring and functioning (Lima-mendez et al., 2015). All known Syndiniales so far are parasitic dinoflagellates, which form a basal group to core dinoflagellates (dinokaryotes) (Strassert et al., 2018). Nevertheless, little is known about biology of these groups.

Most dinoflagellates (including Syndiniales) appear to exhibit a haplontic life cycle. They multiply by mitotic divisions at the haploid vegetative stage and restore diploidy during the transient sexual stage (Parrow and Burkholder, 2004; Tillmann and Hoppenrath, 2013). Sexual reproduction generally results (but not necessarily) in the production of a resting diploid cyst. Given sexual reproduction is widespread in Alveolata and occurs in both free-living (e.g. phototrophic dinoflagellates) and parasitic lineages (e.g. Apicomplexa), the maintenance of a sexual reproduction in Syndiniales deserves to be explored. Indeed, production of different types of zoospores, having different sizes, is widespread in Syndiniales (Coats, 1999; Skovgaard et al., 2009). A putative sexual reproduction was yet observed in the syndinian *Euduboscquella* sp. (MALV-I), where microscopy revealed cell fusion (syngamy) between two different spore types, followed by successive division into four daughter cells, suggesting meiotic recombination (Coats et al., 2012).

Amoebophrya spp. are one of the representatives of Syndiniales and widespread endoparasites. They infect a wide variety of marine organisms, such as ciliates, radiolarians, free-living dinoflagellates, and even other parasitic relatives (Cachon, 1964; Coats, 1999). *A. ceratii* strains infecting dinoflagellates have revealed a varying degree of host specificity and marked sequence differences, and the parasite is now widely believed to be a species complex (Coats et al., 1996; Gunderson et al., 2002; Kim, 2006; Kim et al., 2008; Park et al., 2013; Cai et al., submitted).

In previous work, we characterized 8 putative species in *Amoebophrya*, all waiting for formal description (Cai et al., submitted). Although a polyphasic approach has been adopted to discriminate species, the capability of sexual reproduction in *Amoebophrya* has not been determined nor evaluated yet. Two of the investigated *Amoebophrya* strains have been subjected to whole-genome sequencing and annotation. The big difference between them in terms of gene contents confirmed that they may represent biologically distinct species despite their nearly identical cellular morphologies (Farhat et al., in prep). These two genomes share a high level of synteny, but no sexual life cycle has been described. The observed diversity and applications of species concepts in *Amoebophrya* have been complicated by a long-standing uncertainty of whether or not sexual reproduction occurs in *Amoebophrya*. Ecological success and species diversification beg the question of the maintenance and prevalence of sexual reproduction in Syndiniales.

Meiosis is an essential step in the process of sexual reproduction. In diplontic model organisms, it serves to reduce the chromosome number from diploidy in the germline to haploidy in the gametes. The exact mechanism of meiosis is far from complete understanding. However, the discovery of genes involved in meiosis of model organisms opened a door for the exploration in non-model species. Although some of the meiosis-specific gene functions have not been accurately determined yet, the use of sequence similarities, phylogeny and domain conservation to predict functions of unknown genes have greatly paved the way towards functional annotation. Such studies revealed that many meiotic genes are

conserved among animals, fungi and plants and some eukaryotic microorganisms (protists) (reviewed in Schurko and Logsdon, 2008). A set of meiotic genes that represent the best markers for the presence of meiosis has thereafter been established (Table1; Schurko and Logsdon, 2008; Malik et al., 2008). Some genes in this set have their counterparts involved in mitosis, and together they make up big gene families. For example, meiosis-specific genes MSH4 and MSH5 are members of MutS homologs (MSH) family, which includes six major paralogous eukaryotic groups (MSH1–MSH6) (Lin et al., 2007). REC8 is the meiotic homolog of RAD21 (Parisi et al., 1999). DMC1 is the meiosis-specific homolog of RAD51 (Malik et al., 2008). Some of these genes have been reported to be present in ciliates (Chi et al., 2013), Apicomplexa (Schurko and Logsdon, 2008; Malik et al., 2008), chromerid (Füssy et al., 2017), *Symbiodinium* dinoflagellates (Chi et al., 2014), indicating that sexual reproduction is likely widespread in Alveolata. In this study, we used this meiotic gene inventory approach in the available genomes of two *Amoebophrya* strains to infer the genetic capacity for canonical eukaryotic sex. Then we traced back the gene expression of those genes over a complete infection cycle to predict when sexual reproduction may occur in this parasitic lineage.

Materials and methods

Search for homology

A list of meiosis-specific genes was compiled from Fussy et al. (2016) or key word search in NCBI and then expanded with *Symbiodinium minutum* genes reported by Chi et al. (2014). Protein sequences (Table 1) of these genes from different species were used as queries to search for homologues in the proteomes of two *Amoebophrya* strains, A25 and A120, provided by the ORCAE website (https://bioinformatics.psb.ugent.be/orcae/). The local BLASTP was used as a search tool with E-value cutoff of 10^{-4} and BLOSUM62 alignment matrix. The E-value cutoff was adjusted to 0.1 for HOP2 homologue search considering the short length of this gene (138 residues in humans). For REC8, potential homologues were further searched using HMMER 3.0 (http://hmmer.org/) based on the two characteristic conserved PFAM domains PF04825 and PF04824 found at the N- and C- terminus of its protein, respectively (Howard-Till et al., 2013). Domain structures of potential homologs were analyzed using InterproScan (Jones et al., 2014) against the PFAM (El-Gebali et al., 2018) and SMART (Letunic and Bork, 2017) domain databases, except for HOP2 where both PANTHER (Thomas et al., 2003) and PFAM databases were used. BLAST searches against NCBI NR database were also used to remove false positive hits.

Table 1 query sequences used in this study

Gene	*Homo sapiens*	*Saccharomyces cerevisiae*	*Vitrella brassicaformis*	*Symbiodinium minutum*	Function
	NCBI	NCBI	Fussy et al. (2016)	Chi et al. (2014)	
DMC1	CAG30372	NP_011106.1	Vbra_4727 Vbra_17182	symbB.v1.2.008353.t1 symbB.v1.2.000608.t1	Homolog of RAD51, promotes interhomolog recombination (Malik et al., 2008)
HOP1	NP_001186758	NP_012193.1	Vbra_223.t1 Vbra_1541.t1	-	Binds double strand breaks (DSB) and forms axial and lateral elements of the synaptonemal complex (Malik et al., 2008)
HOP2	NP_001242945	NP_011482.2	Vbra_4454.t1	symbB.v1.2.026766.t1	MND1 and HOP2 form a heterodimeric complex that interacts with RAD51 and DMC1 to promote meiotic recombination and to reduce synapsis and recombination of non-homologous chromosomes (Schurko et al., 2009; Petukhova et al., 2005)
MND1	NP_115493	NP_011332.2	Vbra_4074.t1, Vbra_6181.t1	symbB.v1.2.036043.t1	
MER3	NP_001017975	NP_011263.2	Vbra_14058.t1	-	DNA helicase that promotes holliday junction resolution (Malik et al., 2008)
MSH4	AAB72039	P40965.1	Vbra_13067.t1	symbB.v1.2.013503.t1	Members of MutS homolog (MSH) families. MSH4 and MSH5 form a heterodimer and participate in meiotic crossing-over and chromosome segregation (Hollingsworth et al 1995; Pochart et al., 1997; Ross-Macdonald and Roeder, 1994).
MSH5	BAB63375	NP_010127.1	Vbra_4012.t1	symbB.v1.2.033801.t1	
REC8	NP_001041670	NP_015332.1	-	-	Homolog of RAD21. Critical for meiotic sister chromatid cohesion and correct chromosome segregation (klein, 1998; Watanabe and Nurse, 1999).
SPO11	AAD52562	NP_011841.1	Vbra_16613.t1, Vbra_16614.t1	symbB.v1.2.038121.t1 symbB1.v1.2.012520.t1	Transeseterase, creats DSBs in homologous chromosomes (Malik et al., 2008)

Phylogeny inference

Taxa were sampled from animals (*Homo sapiens*), fungi (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe)*, and plants (*Arabidopsis thaliana*; *Oryza sativa japonica)* and alveolates (e.g. *Plasmodium*, *Symbiodium*) and some other microbial eukaryotes such as *Dictyostelium*, and *Trypanosoma*. For REC8, references were chosen from manually curated sequences from UniprotKB/SWISSprot database (all available for the sampled taxon). For MSH4 and MSH5, reference protein sequences verified by phylogeny were taken from Lin et al. (2007) in order to make a clear distinction between members of the gene family MutS.

Multiple amino acid sequence alignments were constructed using MAFFT v6.240 (Katoh et al., 2002). For the whole length of sequence alignments, we used trimAl (Capella-Gutierrez et al., 2009) to remove gappy positions (≥20% of the sequences) in the alignment. For the alignments of domain sequences, no trimming was performed in order to keep their integrity. Identification of visually recognizable conserved regions in these alignments was performed in Seaview (Galtier et al., 1996).

FastTree v2.1.10 (Price et al., 2010) was used to infer approximately-maximum-likelihood phylogenetic trees based on the alignments of these protein sequences. In some cases, Maximum likelihood trees were also generated in PhyML (Guindon et al., 2010) for reference with bipartition support from 100 bootstrap replicates. The amino acid substitution model WAG+I+Γ was used in the ML analysis. Trees were visualized with FigTree v1.4.3 (Rambaut, 2016).

Results

We targeted nine key genes known to participate exclusively in meiosis (**Table** 1). Hits were obtained for all proteins except for HOP1, which had no homologue in either of the two *Amoebophrya* strains. All significant hits have been screened by phylogeny and for the presence/absence of important protein domains.

SPO11

SPO11 gets a single hit by blastp, which is verified to be homologue of SPO11 in *Amoebophrya* by phylogeny (**Fig** 1). In this tree, the *Amoebophrya* sequences cluster with Apicomplexan SPO11 proteins with strong support (bootstrap value 95%).

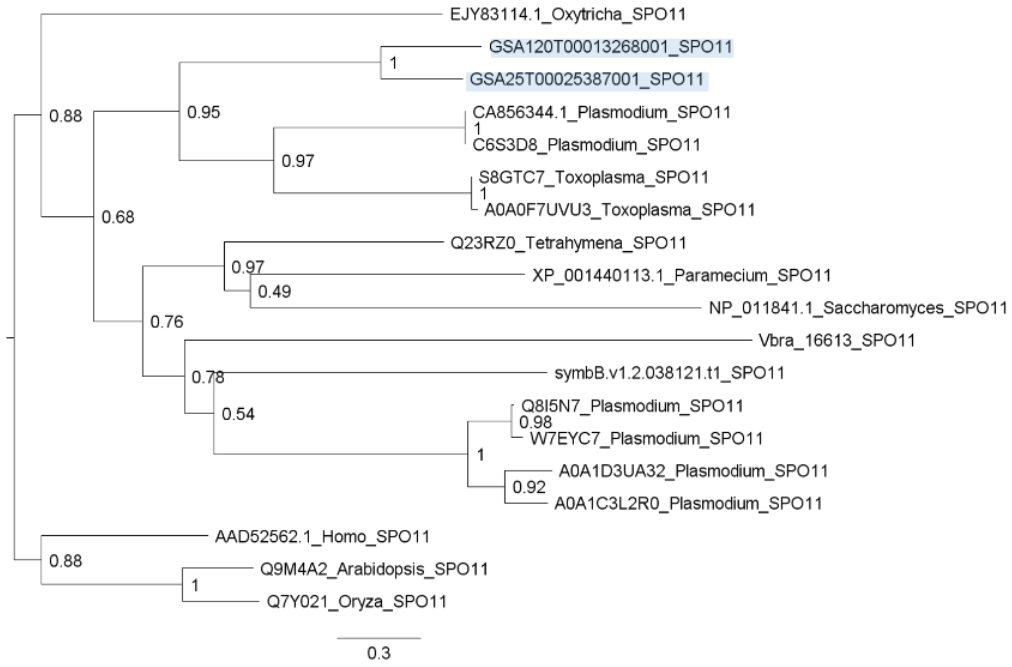


Fig 1 Phylogeny of SPO11 protein homologs based on the whole length alignment trimmed by trimal (Capella-Gutierrez et al., 2009). The tree was built using FastTree and rooted with the animals plus plant group (Homo+Arabidopsis+Oryza at the bottom in the figure). Blue shaded are *Amoebophrya* sequences. Scale bar: 0.3

HOP2/MND1

By domain search, GSA25T00018581001 and GSA120T00008837001 are annotated as MND1 proteins directly by PFAM. However, the two hits of HOP2 cannot be annotated by PFAM as the TBPIP domain

(PFAM: PF07106) required for the function of this protein is absent when examined by searching against PFAM database. Further search in PANTHER database showed both of them are homologous-pairing protein 2 with the structure PTHR15938 (**Table** S1). Phylogenetic analysis showed these *Amoebophrya* sequences produced long branches in the tree (**FIG** 2). But the enclosure of the four *Amoebophrya* proteins by HOP2 and MND1 class respectively, supported the idea that they are homologs of HOP2 and MND1 respectively in *Amoebophrya* strains.

It's worth noting that XP_001024593.2, one of the two MND1 homologs from *Tetrahymena* from previous studies (Chi et al., 2013), is found not to contain the characteristic Pfam domain (PF03962) of this gene (Table S1). In this study, this protein appears similarly related to either HOP2 or MND1 group when the tree is rooted with MND1 class or HOP2 class (**FIG** 2 and **Fig** S1). The protein alignment also shows it badly aligns to MND1 and HOP2 classes (**Fig** S2). Given two copies of MND1 gene in *Tetrahymena* species, this protein may have altered its function in meiosis or have been mistakingly assigned as MND1.

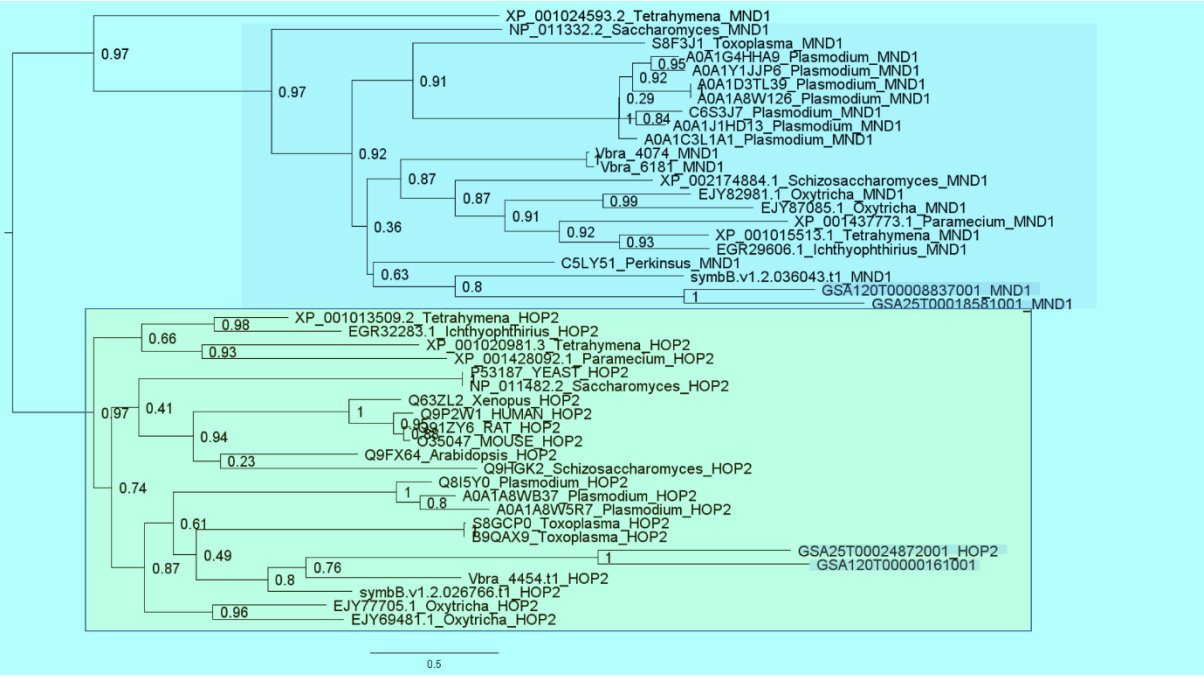


Fig 2. Phylogeny of HOP2/MND1 protein homologs based on the whole length alignment trimmed by trimal (Capella-Gutierrez et al., 2009). The tree was built using FastTree and rooted with HOP2 group. Blue shaded are *Amoebophrya* sequences. Scale bar: 0.5

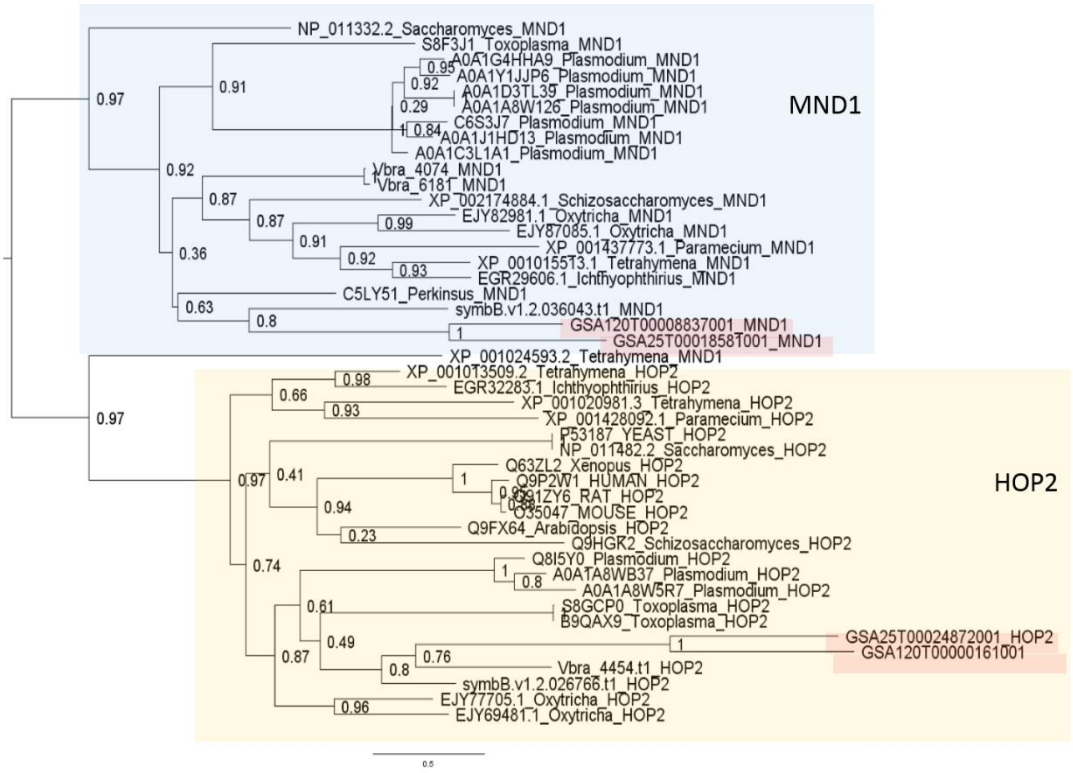


Fig S1 Phylogeny of HOP2/MND1 protein homologs based on the whole length alignment trimmed by trimal (Capella-Gutierrez et al., 2009). The tree was built using FastTree and rooted with MND1 group excluding XP_001024593.2. Red shaded are *Amoebophrya* sequences. Scale bar: 0.5.

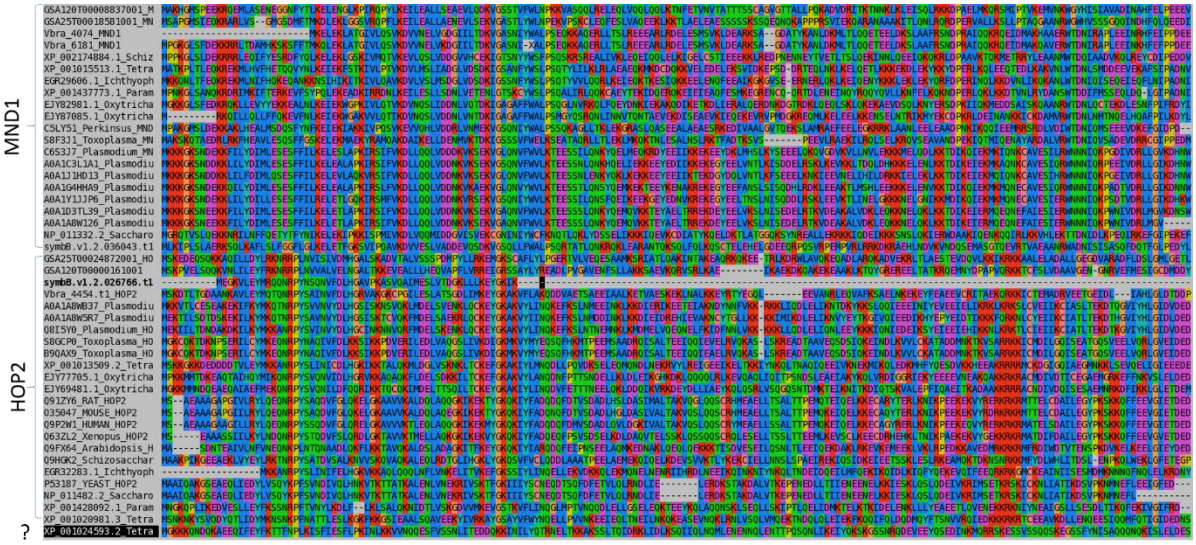


Fig S2 Whole length alignment of HOP2/MND1 homologs trimmed by trimal (Capella-Gutierrez et al., 2009). The question mark (?) shows the previous classification of this sequence may be wrong.

DMC1/RAD51

Two hits were obtained for DMC1 in each *Amoebophrya* strains using blastp. These two hits get the same PFAM annotation as DNA recombination and repair protein Rad51-like proteins (PFAM: PF08423) and have similar domains (**Table** S1). As in Malik et al. (2008), DMC1 and RAD51 reference

proteins were well separated by phylogeny, so we were able to identify DMC1 from RAD51 homologues in the two *Amoebophrya* strains (**Fig**. 3).

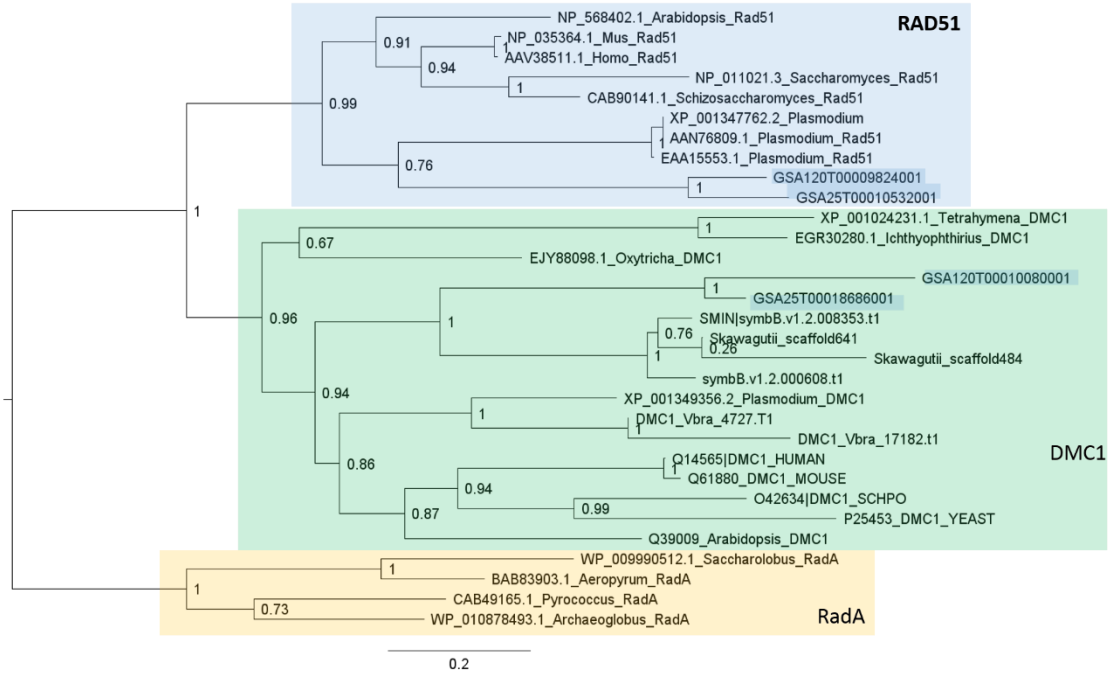


Fig 3 Phylogenies of protein DMC1/RAD51 homologs based on the whole length alignment trimmed by trimal (Capella-Gutierrez et al., 2009). The tree was built using FastTree and rooted with the archaeal RadA sequences taken from Malik et al (2008). Blue shaded are *Amoebophrya* sequences. Scale bar: 0.2.

MSH 4/5 homologs

For MSH4 and MSH5, both genes got three best hits by blastp in two *Amoebophrya* strains. All best hits contain MutS domains, which constitute the structural features of MutS homologs (MSH) (**Table** S1). But the phylogenetic tree (**Fig** 4) constructed based on the shared domain, MutSac, suggests that these hits are the homologs of the sub-families MSH1, MSH2 and MSH6, respectively. Previously, MSH1 was only found in Fungi and plants (Lin et al., 2007; Ogata et al., 2011.). The homology that 2 *Amoebophrya* sequences (GSA120T00012118001 and GSA25T00025231001) show to the plant-specific MSH1 (termed 'plt-MSH1' by Ogata et al., 2011) suggests the distribution of MSH1 could be extended to Alveolata.

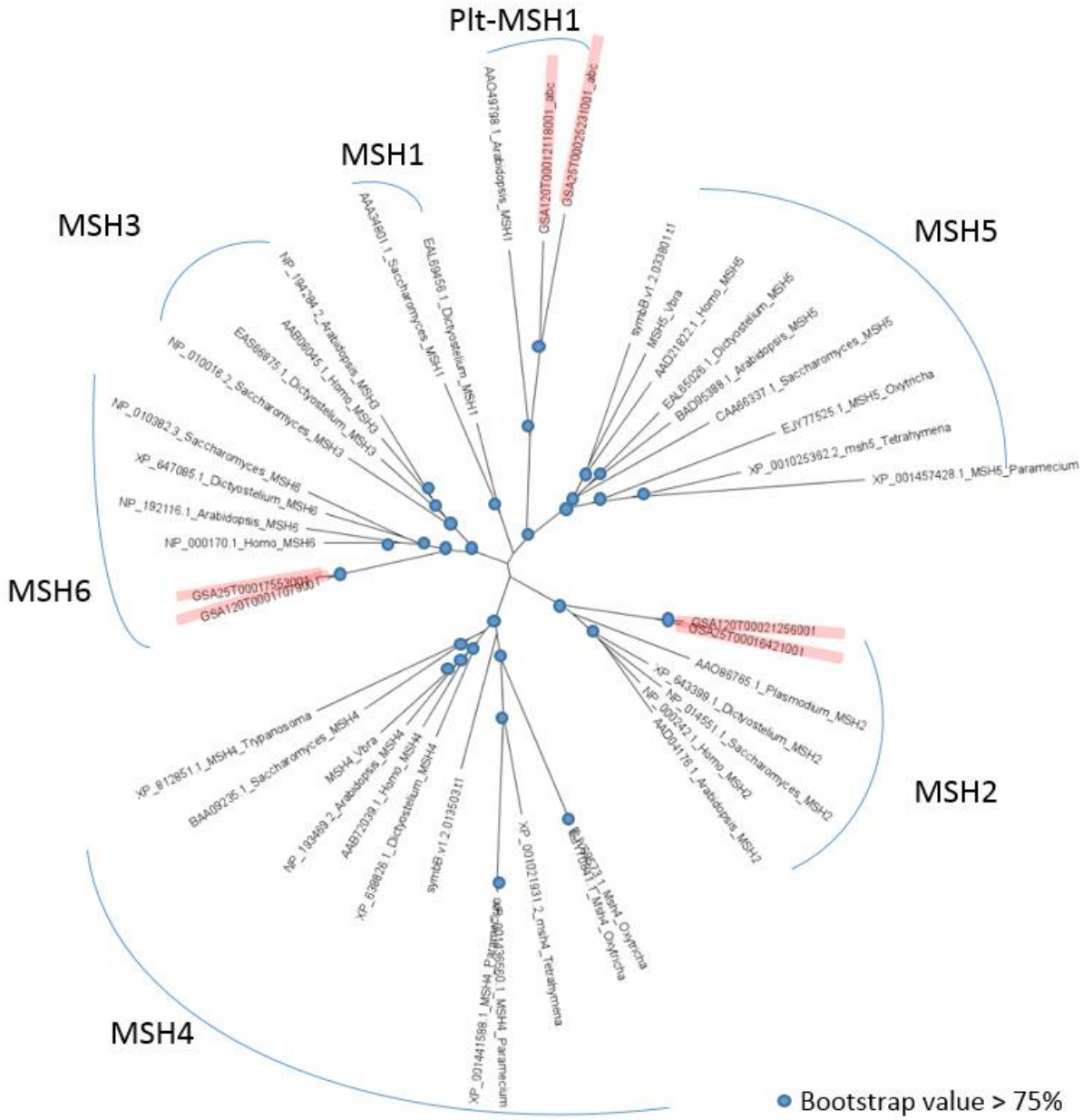


Fig 4 Phylogeny of MutS homologs based on the whole length alignment trimmed by trimal (Capella-Gutierrez et al., 2009). The tree was built using FastTree and unrooted. Red shaded are *Amoebophrya* sequences.

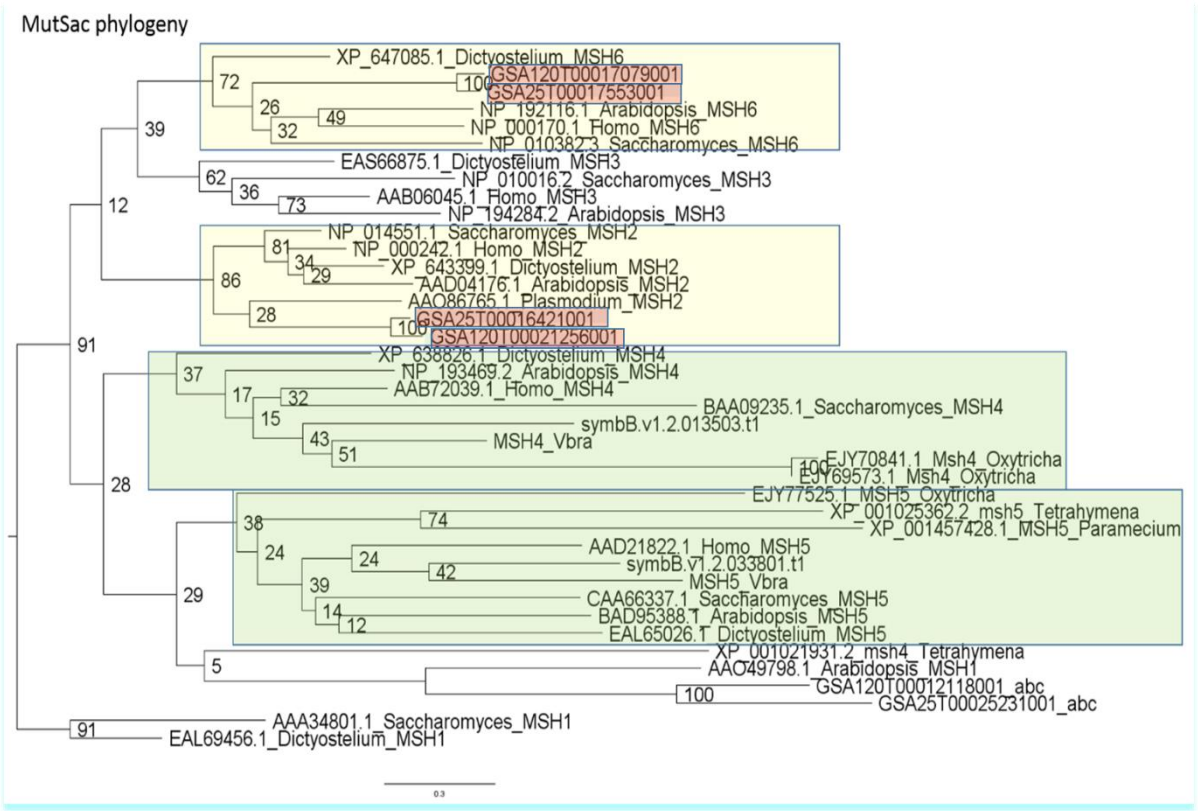


Fig S3. Phylogeny of MutS homologs based on the domain MutSac (SMART: SM00534) sequences. The tree was built using phyML with the evolution model WAG+I+8F and rooted with MSH1. Node values are support from 100 bootstrap running. Scale bar: 0.3. Red shaded are *Amoebophrya* sequences.

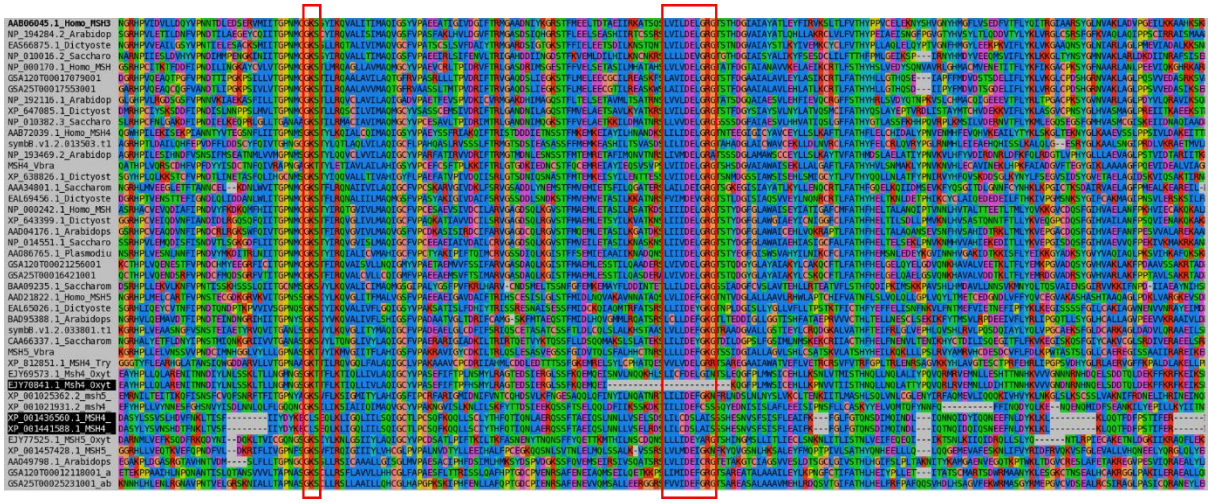


Fig S4. Alignment of the conserved domain MutSac across the MutS homolog family. The red rectangles mark out the shared short motifs by most of the sequences.

Noteworthy, in *Oxytricha trifallax* and *Paramecium tetraurelia*, there were two genes found to be the orthologues of MSH4, respectively (Chi et al., 2013). But in all the species tested experimentally, only one copy of MSH4 was detected (Lin et al., 2007). This renders us to consider the true identity of these proteins more carefully. In *O. trifallax*, both putative MSH4 orthologues, EJY70841.1 and EJY69573.1,

are clustered with MSH4 subfamily in the phylogenetic tree based on full-length alignment (**Fig** S3), but in the MutSac domain, EJY70841.1 lacks the very conserved motif, (F/L)(V/I)(M/L/V))(I/M)DE(I/L/F)G(K/R)(G/R) (**Fig** S4), suggesting this protein could be pseudogene product, or have altered its function as a MutS homolog. In *P. tetraurelia*, both putative MSH4 orthologues, XP_001441588.1 and XP_001436560.1, are found to have lost the MutSd domain and replace MutSac domain with a very similar MutS domain V (PFAM: PF00488). As a result, both proteins lack several unique features of the MutS gene family, including the GKS motif. Altogether, these two proteins may not have kept the same function as other MSH4 orthologues. We also noticed that the putative MSH4 othologue in *T. thermophila*, XP_001021931.1, does not cluster with the MSH4 class but with MSH1 class with a low support value in the tree constructed with MutSac domain sequences (**Fig** 4). It could be due to the absence of one of the conserved motifs mentioned here in the MutSac region (**Fig** S4).

MER3 homolog

MER3 encodes a DNA helicase (Mazina et al., 2004; Nakagava and kolodner, 2002) and is a member of helicase superfamily. To make accurate prediction and comparison, all references included in this analysis were annotated as helicase superfamily 1 or 2 members. They have been categorized into 2 classes: U5 snrnp 200kD RNA helicases and MER3 DNA helicase, which is supported by the previous phylogenetic study (Malik et al., 2008).

We concluded all MER3 hits from *Amoebophrya* are not MER3 homologs for a number of reasons. All curated MER3 references have three domains: DExDc (SMART: SM00487), HELICc (SMART: SM00490) and sec63 (SMART: SM00973) (**Fig** 5A and Table S1) (For A0A1Q9D472_Symbiodinium_MER3, there are only two domains detected, but this protein is not curated by SWISS-PROT yet). All the *Amoebophrya* MER3 hits have more domains than MER3 references or less in some cases. Given that domains are the essential elements of protein functions, we identified two proteins from each *Amoebophrya* strain with all three domains making up the structure of MER3 proteins (**Fig** 5A). For instance, for A120, GSA120T00004033001 and GSA120T00018887001 are found to encompass all three domains. Interestingly, all 3 domains are replicated with 2 copies and tandemly arranged in GSA120T00004033001 while GSA120T00018887001 has one extra copy of the sec63 domain at the N-terminus. Likewise, for GSA25T00003083001 in A25, all three domains have been doubled and GSA25T00014432001 contains one extra DexDc domain. However, the same cases occur to RNA-helicases. For example, O48534.1_Arabidopsis_RNA_helicase, a U5 small nuclear ribonucleoprotein helicase, is annotated as containing DExDc, HELICc and sec63, each with two copies. P32639.2_YEAST_RNA_helicase has one extra DExDc and sec63 apart from those three domains.

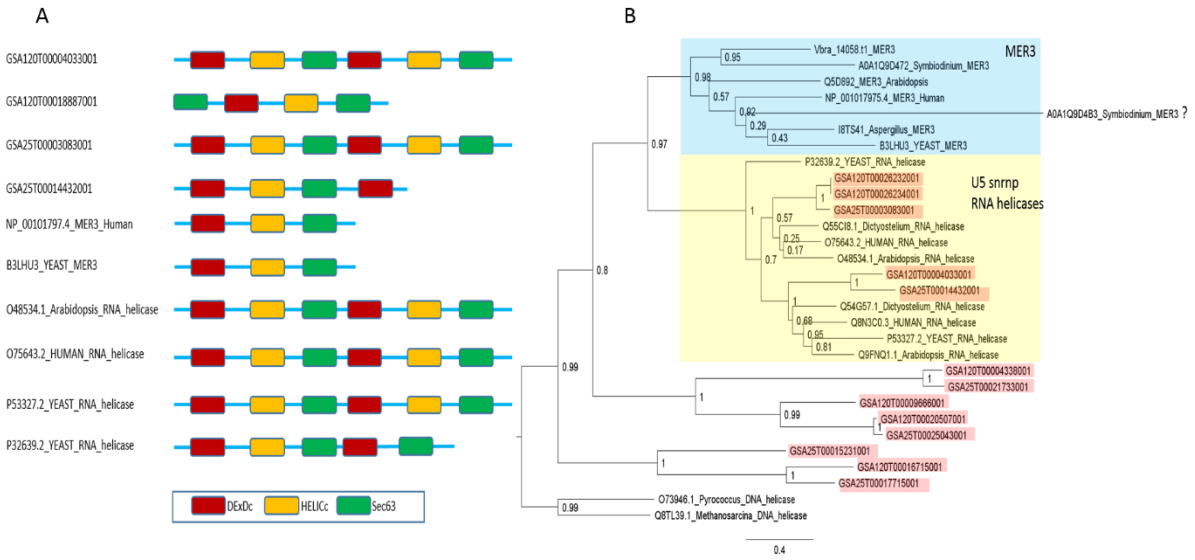


Fig 5. (A) Domain architecture of REC8 and RNA-helicase proteins. Colored rectangles represent conserved domains. (B) Phylogenies of DNA/RNA helicases based on whole length alignment trimmed by trimAl (Capella-Gutierrez et al., 2009). The U5 snrnp RNA helicase sequences are taken from Malik et al (2008). The Tree was built using FastTree and rooted with Archaeal Ski DNA helicase. Red shaded are *Amoebophrya* sequences. Scale bar: 0.4

Further phylogenetic analysis suggests these *Amoebophrya* sequences are not likely the homologues of MER3. The trees based on the domain HELICc (**Fig** S5), the consecutive three domains(**Fig** S6), and the whole-length sequence (**Fig** 5B) all show that no *Amoebophrya* sequences reveal homology closer to MER3 than to reference RNA helicases.

Despite these evidences, we cannot rule out the possibility that MER3 homologs with a different domain organization from model species MER3 occur in *Amoebophrya*, with a conserved function in meiotic recombination.

Noteworthy, one sequence from *Symbiodium* previously annotated as MER3 homolog in UniProtKB database, A0A1Q9D4B3, appears as a long branch in the phylogenetic tree based on the full-length sequence alignment. The fact that we could not annotate this protein using our annotation process and this sequence barely align to other helicases, suggests this protein could not be MER homolog.

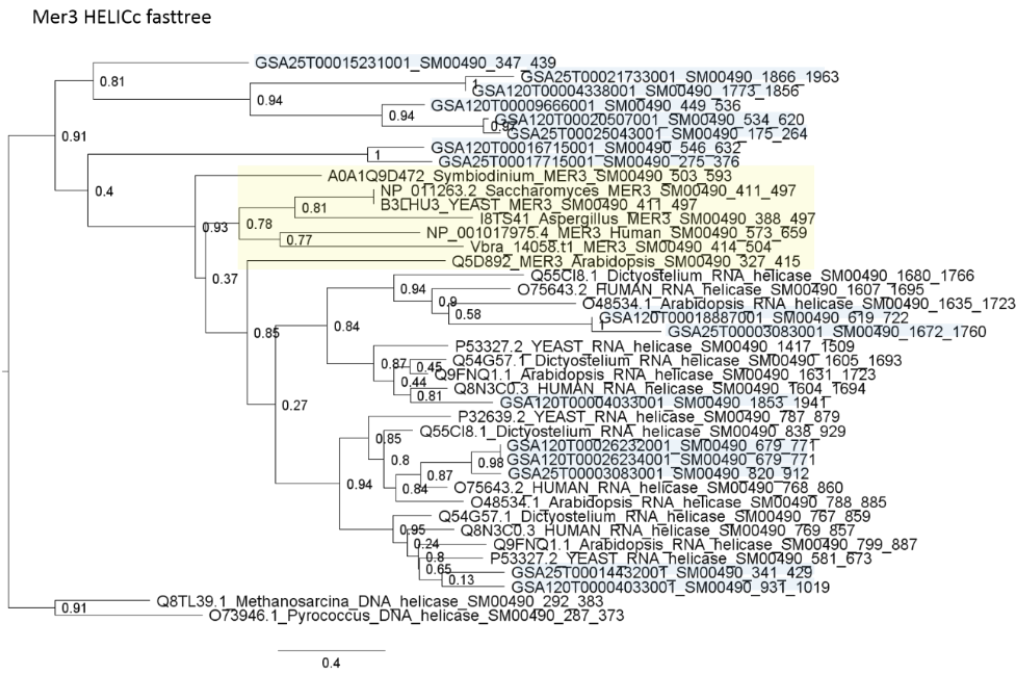


Fig S5. Phylogeny of DNA/RNA helicases based on the conserved domain HELICc (SMART: SM00490). The tree was built using FastTree and rooted with Archaeal Ski DNA helicase. Blue shaded are *Amoebophrya* sequences. Scale bar: 0.4

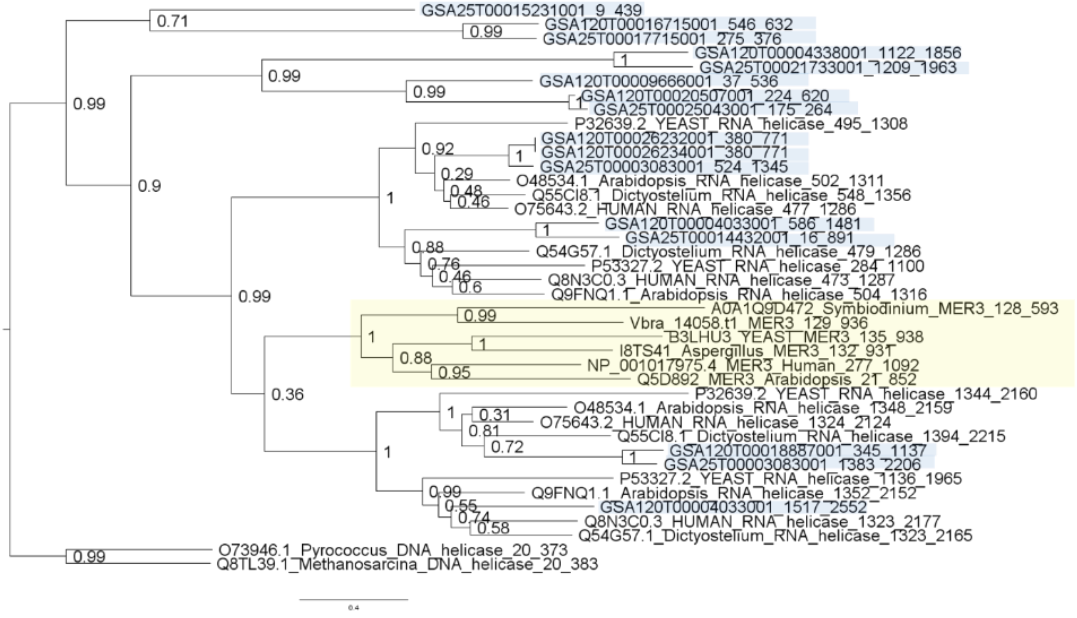


Fig S6 Phylogeny of 3 consecutive conserved protein domains of DNA/RNA helicases. The tree was built using FastTree and rooted with Archaeal Ski DNA helicase (2 sequences at the bottom). Blue shaded are *Amoebophrya* sequences. Scale bar: 0.4.

REC8

REC8 and RAD21 are the mitotic and meiotic members, respectively, of a gene family involved in DNA metabolism (Parisi et al., 1999). In this study, we performed a comprehensive comparison of the REC8 and RAD21 proteins from different eukaryotes using curated sequences from uniprotKB. We assumed that all *Amoebophrya* REC8 candidates are RAD21 homologs for the following reasons.

Firstly, the absence of one of the major structural motifs indicates the biological function might have been shifted to some extent from that of REC8. In particular, the isoelectric point of GSA120T00017233001 is 6.4, which is very different than the curated REC8/RAD21 proteins (pI: 5.0-5.5). All REC8 or RAD21 references have been found to contain two conserved domains, N-terminus domain (PF04825) and C-terminus domain (PF04824) except for REC8 homolog in Yeast (Q12188_REC8_YEAST, see **table** S1). However, for all the candidates of REC8/RAD21-like proteins from the two *Amoebophrya* strains, only one of the conserved domains was detected (**Table** S1).

Secondly, at the residue level in the conserved domains, there are small motifs shared by these candidates with RAD21 sequences other than with REC8 sequences (**Fig** 6). For example, GSA120T00017233001 shared common motifs HWDK(K/R) and GHLLL with other verified RAD21 group in the N-terminal domain, which is a unique feature distinct from in REC8 class. GSA25T00010409001 and GSA120T00009901001 possess (M/F/L)LVLK as in other sequences of the RAD21 group.

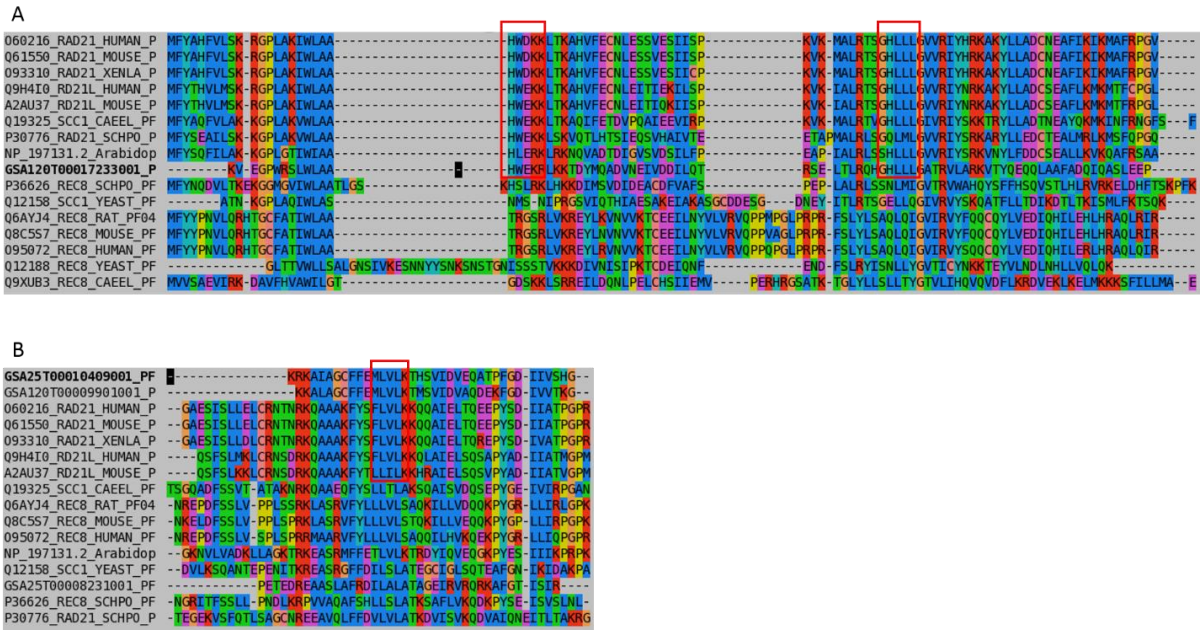


Fig 6. Alignments of the 2 conserved regions detected from the Rec8/Rad21 like proteins sequences used for the construction of phylogenetic trees showed above. (A) phylogeny of N-termimus of REC8/RAD21 like proteins. (B) phylogeny of C-terminus of REC8/RAD21 like proteins. The red rectangles mark out the motifs indicating the identity to RAD21 class.

Thirdly, at the phylogenetic level, these candidates from *Amoebophrya* are evolutionarily closer to RAD21 class than to REC8 class. In the trees constructed with full sequences (**Fig** 6) and conserved domains (**Fig** S7) respectively, all candidates cluster with RAD21 class although there is no clear distinction between the RAD21 and REC8 members of the gene family based on these phylogenies.

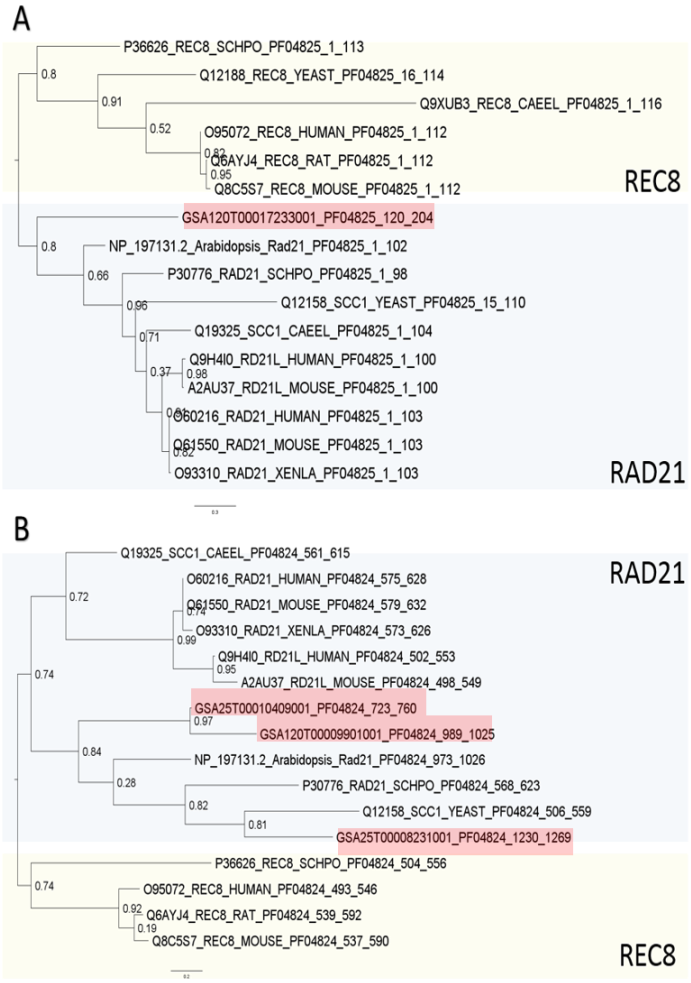


Fig S7. Phylogenies of two conserved domains of REC8/RAD21 like proteins. A was constructed using the N-terminus conserved domain (PFAM: PF04825) amino acid sequences and rooted with rec8 class, and B using the conserved region (PFAM: PF04824) amino acid sequences, rooted with REC8 class. Trees were built using FastTree. Blue shaded are *Amoebophrya* sequences. Scale bar: 0.3 for A and 0.2 for B.

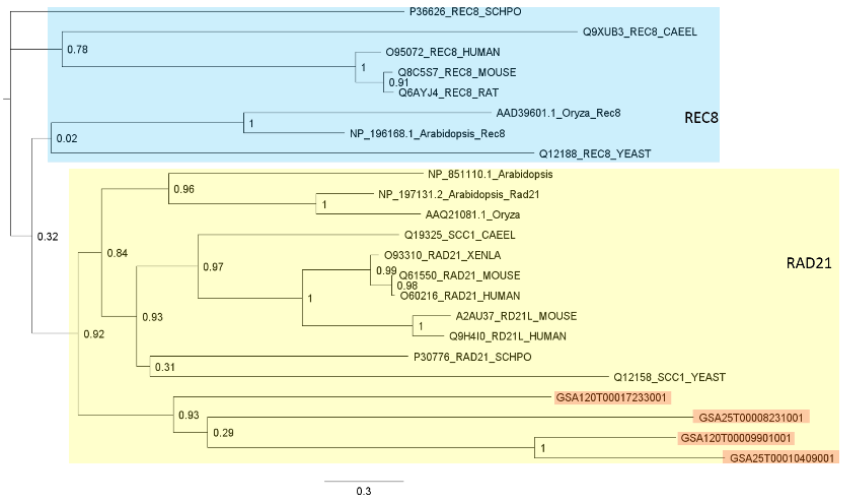


Fig 7. Phylogeny of REC8/RAD21-like proteins based on the whole length alignment trimmed by trimal (Capella-Gutierrez et al., 2009). The tree was built using FastTree and unrooted. Red shaded are *Amoebophrya* sequences.

Noteworthy, experimental studies demonstrated that *Tetrahymena* uses a single version of α-kleisin (Rad21/Rec8 family) for both mitosis and meiosis (Howard-Till et al., 2013). So it's likely that these *Amoebophrya* sequences could be also used for meiosis.

Expression of meiosis-specific genes

To assess if the meiosis-specific genes detected in *Amoebophrya* are functional, we queried the transcriptomes along a complete infection cycle and found evidence of transcription of some genes. 3 of 4 meiosis-specific genes detected in the two *Amoebophrya* strains showed higher expression at dinospore stage than in other intra-host stages (Table 2), indicating sex, if exists, may occur at the free-living stage.

Table 2 the meiosis-specific genes in the 2 strains of *Amobophrya*

	A120	A25	Differentially expressed in stage*
DMC1	GSA120T00010080001	GSA25T00018686001	URG1
HOP1	-	-	
HOP2	GSA120T00000161001	GSA25T00024872001	ND
REC8	-	-	
MER3	-	-	
MND1	GSA120T00008837001	GSA25T00018581001	URG1
MSH4	-	-	
MSH5	-	-	
SPO11	GSA120T00013268001	GSA25T00025387001	URG1

ND: NOT detected. * the developmental stage corresponds to the differential expression analysis of Farhat et al. (2018). URG1: at the dinospore stage.

Discussion

The use of meiosis detection toolkit in alveolates

Meiosis was predominantly studied in model species, such as *Saccharomyces cerevisiae* and *Arabidopsis*, that have advanced our understanding of the details of this process. This basic scheme is conserved in eukaryotes but the mechanism has been complicated by the presence of multiple paralogs of the genes involved. On the other hand, when the species investigated is far related to model species,

in many cases the comparison to model species is complicated. MER3 is a good example to illustrate that purpose.

A MER3 homologue in plants, RCK protein, is first found in *Arabidopsis* (Chen et al., 2005). The high sequence similarity between these proteins and mutant phenotypes suggest that RCK function similarly in meiosis as MER3 in yeast. In animals, HFM1 (the human homologue of yeast MER3) encodes a putative DNA helicase expressed specifically in germ-line tissues (Tanaka et al., 2006). Sequence similarity searches of databases uncovered many more putative MER3 homologs in other plants (e.g. rice) and animals (e.g. mice). Although the function of these putative MER3 homologs is not known yet, the high degree of sequence identity suggests that they share a conserved function that is likely important for meiotic recombination (Chen et al., 2005). However, when the species (e.g. alveolates) in question is distantly related to these model organisms, the sequence similarity-based approach becomes highly questionable in practice.

Moreover, the mechanisms associated with helicase superfamilies vary considerably. Helicases are classified into 6 superfamilies according to the sequence and structure characteristics (reviewed by Singleton et al., 2007). RNA helicases are found in the helicase superfamilies 1-5. The superfamily 2 (SF2) is subdivided into at least 10 families, based on phylogenetic analysis of the sequences of the helicase core domains (Jankowsky and Margaret, 2010). Five of these SF2 families, DEAD-box, DEAH/RHA, Ski2-like, RIG-I-like, and viral DExH proteins, the NS3/NPH-II family, are comprised mainly of RNA helicases and are thus termed "RNA helicase families".

The lack of a clear distinction between DNA and RNA helicases in families and superfamilies is hard for discrimination between DNA and RNA substrates. Instead, mechanistic features of proteins from the respective families appear to be utilized in both RNA- and DNA- related processes. While each helicase family has distinct or sometimes subtle structural characteristics, structures of DNA and RNA helicases within each family are highly similar, and it is thus not clear which structural features dictate functions on DNA, RNA or both (reviewed by Jankowsky and Margaret, 2010). In our case, the analysis of domain structures shows that some of *Amoebophrya* proteins contain conserved domains (**Fig** S5): DEXDc, HELICc and a SEC63. The DEXDc domain is found in members of the DEAD-like helicases, a diverse superfamily of helicases that use ATP hydrolysis to unwind DNA or RNA (Nakagawa et al., 2001). The HELICc domain is also found in a wide variety of helicases that contain DEXDc-, DEAD-, or DEAH-box domains (Shibata et al., 1999; Theis et al., 1999).

Proteins have been frequently designated as MER3/RNA helicases based on sequence similarity but without direct evidence of a DNA/RNA-related function. As pointed by Jankowsky and Margaret (2010), the classification of a given protein as DNA/RNA helicase based solely on sequence is problematic, because several "RNA helicase families" also contain enzymes that function on both DNA and RNA substrate.

Insights into sexual reproduction in *Amoebophrya*

The expression of SPO11, DMC1 and MND1 orthologs suggests that meiotic recombination may occur in *Amoebophrya*. It's worth noting that some genes thought to be meiosis-specific may also function in parthenogenetic organisms. For example, the expression of SPO11 was detected in the parthenogenesis of *monogonont rotifers* and *Daphnia pulex* (Hanson et al., 2013; Schurko et al., 2009). SPO11, DMC1 and MND1 have been reported to serve parasexual genetic recombination in *Candida albicans* and *Giardia intestinalis* (Forche et al., 2008; Carpenter et al., 2012). Therefore, the expression of these genes could also result from parthenogenesis or parasexual process. *Amoebophrya* has some characteristics common to asexual life cycles, including short generation times and efficient dispersal abilities, which often contribute to the survival of parthenogenetic lineages (Simon et al., 2003). However, 9 genes included in this study are "meiosis-specific" since they are only known to function in meiosis in animals, fungi and plants and thus hypothesized to only be present in organisms with canonical sexual machinery. Since alveolates are phylogenetically far from animals, fungi and plants, they may undergo non-canonical meiosis, which explains our inability to detect other genes.

In dinoflagellate, the typical life cycles include a haplontic stage and a short diploid phase. Sexual stages have not been observed in the genetically diverse genus *Amoebophrya* yet. Given the evolutionary closeness to free-living dinoflagellates, sex in *Amoebophrya* appears to be facultative. In such an instance, a sexual cycle may actually exist but rarely happen and thus is difficult to observe.

In apicomplexa, a sexual stage is essential to complete its life cycle. *Plasmodium* parasites have a dimorphic sexual stage that is closely linked to the transmission cycle (the sexual stage is the transmissible stage) (reviewed in Weedall and Hall, 2015). In the environmentally transmitted parasites (e.g. *Eimeria*), gametocytogenesis and gametogenesis form a continuous process that takes place inside one infected host cell and appears to be programmed to occur after approximately three asexual cycles. This is similar to *Amoebophrya* in that it must experience a free-living stage to find a new host. Given novel results obtained from gene expression, if sexual reproduction exists, it likely occurs during this free-living stage. Successive division right before producing final dinospores has been observed in culture. Given this information, sexual reproduction is more likely to be common in *Amoebophrya*.

Conclusions and Perspective

In microeukaryotes, the direct observation of sex is an arduous task due to their size, morphological diversity and paucity of knowledge regarding their life histories (reviewed in Schurko et al 2008). In this study, we provide evidence that a set of meiosis-specific genes exist in *Amoebophrya* genomes and expressed at some period of its life cycle, suggesting the potentiality for sex and meiosis in this group. Sexual fusion (syngamy) of spores has been reported in a number of syndinean species, such as *Coccidinium mesnili*, *Duboscquella anisospora* and *Euduboscquella Crenulata* (Coats et al., 2012). Taken together, sexuality in Syndiniales potentially involves syngamy of anisospores. Zoospores

development was well followed in *Ichthyodinium chabelardi* (MALV-I) (Shadrin et al., 2015). Observations demonstrated that zoospores dimorphism proceeds by two distinct developmental pathways, one with two divisions leading to the formation of large macrospores, and the other with three divisions leading to small microspore formation. In *Amoebophrya*, zoospores with different sizes and the successive division right before the release of final dinospores have been seen frequently in the lab but fusion of these zoospores needs to be verified. Haploid, diploid, triploid phases have so far been observed in *Amoebophrya* by flow cytometry, but whether these stages correspond to distinct developmental pathways or the end of the sporulation needs to be better explored.

There are numerous methods for detecting sexual reproduction. However, many are difficult to apply to protists. Observation for detecting sexual reproductive structures, mating or production of males provides the strongest evidence of sex. However, recreating conditions that induce sex under laboratory conditions is often difficult and mating of gametes may be rare and/or very short events. Methods for detecting sexual reproduction based on the genetic consequences of sex and meiotic recombination would be more informative and potentially supportive in the near future.

References

Boyle, J. P., Rajasekar, B., Saeij, J. P., Ajioka, J. W., Berriman, M., Paulsen, I., ... & Boothroyd, J. C. (2006). Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like *Toxoplasma gondii*. Proceedings of the National Academy of Sciences, 103(27), 10514-10519.

Cachon, J. (1964). Contribution a l'étude des Péridinies parasites. Cytologie, cycles évolutifs. Ann. Sci. Nat., 12 ser. 6: 1-158.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972-1973.

Carpenter, M. L., Assaf, Z. J., Gourguechon, S., & Cande, W. Z. (2012). Nuclear inheritance and genetic exchange without meiosis in the binucleate parasite *Giardia intestinalis*. Journal of Cell Science, 125(10), 2523-2532.

Chen, C., Zhang, W., Timofejeva, L., Gerardin, Y., & Ma, H. (2005). The *Arabidopsis* ROCK-N-ROLLERS gene encodes a homolog of the yeast ATP-dependent DNA helicase MER3 and is required for normal meiotic crossover formation. The Plant Journal, 43(3), 321-334.

Chi, J., Mahé, F., Loidl, J., Logsdon, J., & Dunthorn, M. (2013). Meiosis gene inventory of four ciliates reveals the prevalence of a synaptonemal complex-independent crossover pathway. Molecular biology and evolution, 31(3), 660-672.

Chi, J., Parrow, M. W., & Dunthorn, M. (2014). Cryptic sex in *Symbiodinium* (Alveolata, Dinoflagellata) is supported by an inventory of meiotic genes. Journal of Eukaryotic Microbiology, 61(3), 322-327.

Coats, D. W. (1999). Parasitic Life Styles of Marine Dinoflagellates. Journal of Eukaryotic Microbiology, 46(4), 402-409.

Coats, D. W., Adam, E. J., Gallegos, C. L., & Hedrick, S. (1996). Parasitism of photosynthetic dinoflagellates in a shallow subestuary of Chesapeake Bay, USA. Aquatic Microbial Ecology, 11(1), 1-9.

Coats, D. W., Bachvaroff, T. R., & Delwiche, C. F. (2012). Revision of the Family Duboscquellidae with Description of *Euduboscquella crenulata* n. gen., n. sp. (Dinoflagellata, Syndinea), an Intracellular Parasite of the Ciliate *Favella panamensis* Kofoid & Campbell. Journal of Eukaryotic Microbiology, 59(1), 1-11.

Colegrave, N. (2002). Sex releases the speed limit on evolution. Nature, 420(6916), 664.

Dacks, J., & Roger, A. J. (1999). The first sexual lineage and the relevance of facultative sex. Journal of Molecular Evolution, 48(6), 779-783.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... & Carmichael, M. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science, 348(6237), 1261605.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... & Sonnhammer, E. L. L. (2018). The Pfam protein families database in 2019. Nucleic acids research, 47(D1), D427-D432.

Forche, A., Alby, K., Schaefer, D., Johnson, A. D., Berman, J., & Bennett, R. J. (2008). The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. PLoS biology, 6(5), e110.

Füssy, Z., Masařová, P., Kručinská, J., Esson, H. J., & Oborník, M. (2017). Budding of the alveolate alga *Vitrella brassicaformis* resembles sexual and asexual processes in apicomplexan parasites. Protist, 168(1), 80-91.

Galtier, N., Gouy, M., & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Bioinformatics, 12(6), 543-548.

Goddard, M. R., Godfray, H. C. J., & Burt, A. (2005). Sex increases the efficacy of natural selection in experimental yeast populations. Nature, 434(7033), 636.

Grigg, M. E., Bonnefoy, S., Hehl, A. B., Suzuki, Y., & Boothroyd, J. C. (2001). Success and virulence in *Toxoplasma* as the result of sexual recombination between two distinct ancestries. Science, 294(5540), 161-165.

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.

Gunderson, J. H., John, S. A., Chanson Boman, W., & Coats, D. W. (2002). Multiple strains of the parasitic dinoflagellate *Amoebophrya* exist in Chesapeake Bay. Journal of Eukaryotic Microbiology, 49(6), 469-474.

Hanson, S. J., Schurko, A. M., Hecox-Lea, B., Mark Welch, D. B., Stelzer, C. P., & Logsdon Jr, J. M. (2013). Inventory and phylogenetic analysis of meiotic genes in *monogonont rotifers*. Journal of Heredity, 104(3), 357-370.

Hollingsworth, N. M., Ponte, L., & Halsey, C. (1995). MSH5, a novel MutS homolog, facilitates meiotic reciprocal recombination between homologs in *Saccharomyces cerevisiae* but not mismatch repair. Genes & development, 9(14), 1728-1739.

Howard-Till, R. A., Lukaszewicz, A., Novatchkova, M., & Loidl, J. (2013). A single cohesin complex performs mitotic and meiotic functions in the protist *Tetrahymena*. PLoS genetics, 9(3), e1003418.

Jankowsky, E., & Fairman-Williams, M. E. (2010). An introduction to RNA helicases: superfamilies, families, and major themes. RNA helicases, 19, 1-31.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14), 3059-3066.

Kim, S. (2006). Patterns in host range for two strains of *Amoebophrya* (Dinophyta) infecting thecate dinoflagellates: *Amoebophrya* spp. ex *Alexandrium affine* and ex *Gonyaulax polygramma*. Journal of phycology, 42(6), 1170-1173.

Kim, S., Park, M. G., KIM, K. Y., KIM, C. H., Yih, W., Park, J. S., & Coats, D. W. (2008). Genetic diversity of parasitic dinoflagellates in the genus *Amoebophrya* and its relationship to parasite biology and biogeography. Journal of eukaryotic microbiology, 55(1), 1-8.

Klein, F., Mahr, P., Galova, M., Buonomo, S. B., Michaelis, C., Nairz, K., & Nasmyth, K. (1999). A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. Cell, 98(1), 91-103.

Letunic, I., & Bork, P. (2017). 20 years of the SMART protein domain annotation resource. Nucleic acids research, 46(D1), D493-D496.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., ... & Bittner, L. (2015). Determinants of community structure in the global plankton interactome. Science, 348(6237), 1262073.

Lin, Z., Nei, M., & Ma, H. (2007). The origins and early evolution of DNA mismatch repair genes—multiple horizontal gene transfers and co-evolution. Nucleic acids research, 35(22), 7591-7603.

Malik, S. B., Pightling, A. W., Stefaniak, L. M., Schurko, A. M., & Logsdon Jr, J. M. (2008). An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. PloS one, 3(8), e2879.

Mazina, O. M., Mazin, A. V., Nakagawa, T., Kolodner, R. D., & Kowalczykowski, S. C. (2004). *Saccharomyces cerevisiae* MER3 helicase stimulates 3′–5′ heteroduplex extension by Rad51: implications for crossover control in meiotic recombination. Cell, 117(1), 47-56.

Nakagawa, T., & Kolodner, R. D. (2002). *Saccharomyces cerevisiae* Mer3 is a DNA helicase involved in meiotic crossing over. Molecular and cellular biology, 22(10), 3281-3291.

Nakagawa, T., Flores-Rozas, H., & Kolodner, R. D. (2001). The MER3 helicase involved in meiotic crossing over is stimulated by single-stranded DNA-binding proteins and unwinds DNA in the 3′ to 5′ direction. Journal of Biological Chemistry, 276(34), 31487-31493.

Ogata, H., Ray, J., Toyoda, K., Sandaa, R. A., Nagasaki, K., Bratbak, G., & Claverie, J. M. (2011). Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. The ISME journal, 5(7), 1143.

Parisi, S., McKay, M. J., Molnar, M., Thompson, M. A., Van Der Spek, P. J., van Drunen-Schoenmaker, E., ... & Kohli, J. (1999). Rec8p, a meiotic recombination and sister chromatid cohesion phosphoprotein of the Rad21p family conserved from fission yeast to humans. Molecular and cellular biology, 19(5), 3515-3528.

Park, M. G., Kim, S., Shin, E. Y., Yih, W., & Coats, D. W. (2013). Parasitism of harmful dinoflagellates in Korean coastal waters. Harmful algae, 30, S62-S74.

Parrow, M. W., & Burkholder, J. M. (2004). The sexual life cycles of *Pfiesteria piscicida* and cryptoperidiniopsoids (Dinophyceae). Journal of Phycology, 40(4), 664-673.

Petukhova, G. V., Pezza, R. J., Vanevski, F., Ploquin, M., Masson, J. Y., & Camerini-Otero, R. D. (2005). The Hop2 and Mnd1 proteins act in concert with Rad51 and Dmc1 in meiotic recombination. Nature structural & molecular biology, 12(5), 449.

Pochart, P., Woltering, D., & Hollingsworth, N. M. (1997). Conserved properties between functionally distinct MutS homologs in yeast. Journal of Biological Chemistry, 272(48), 30345-30349.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one, 5(3), e9490.

Rambaut, A. (2016). FigTree v1. 4.3 software. Institute of Evolutionary Biology, University of Edinburgh.

Ramesh, M. A., Malik, S. B., & Logsdon Jr, J. M. (2005). A phylogenomic inventory of meiotic genes: evidence for sex in Giardia and an early eukaryotic origin of meiosis. Current biology, 15(2), 185-191.

Ross-Macdonald, P., & Roeder, G. S. (1994). Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. Cell, 79(6), 1069-1080.

Schurko, A. M., & Logsdon Jr, J. M. (2008). Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. Bioessays, 30(6), 579-589.

Schurko, A. M., Logsdon, J. M., & Eads, B. D. (2009). Meiosis genes in *Daphnia pulex* and the role of parthenogenesis in genome evolution. BMC evolutionary biology, 9(1), 78.

Shadrin, A. M., Simdyanov, T. G., Pavlov, D. S., & Nguyen, T. (2015, March). Free-living stages of the life cycle of the parasitic dinoflagellate *Ichthyodinium chabelardi* Hollande et J. Cachon, 1952 (Alveolata: Dinoflagellata). In Doklady Biological Sciences (Vol. 461, No. 1, p. 104). Springer Science & Business Media.

Shibata, A., Nakagawa, N., Sugahara, M., Masui, R., Kato, R., Kuramitsu, S., & Fukuyama, K. (1999). Crystallization and preliminary X-ray diffraction studies of a DNA excision repair enzyme, UvrB, from *Thermus thermophilus* HB8. Acta Crystallographica Section D: Biological Crystallography, 55(3), 704-705.

Signorovitch, A. Y., Dellaporta, S. L., & Buss, L. W. (2005). Molecular signatures for sex in the Placozoa. Proceedings of the National Academy of Sciences, 102(43), 15518-15522.

Simon, J. C., Delmotte, F., Rispe, C., & Crease, T. (2003). Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals. Biological Journal of the Linnean Society, 79(1), 151-163.

Singleton, M. R., Dillingham, M. S., & Wigley, D. B. (2007). Structure and mechanism of helicases and nucleic acid translocases. Annu. Rev. Biochem., 76, 23-50.

Skovgaard, A., Meneses, I., & Angélico, M. M. (2009). Identifying the lethal fish egg parasite *Ichthyodinium chabelardi* as a member of Marine Alveolate Group I. Environmental microbiology, 11(8), 2030-2041.

Strassert, J. F., Karnkowska, A., Hehenberger, E., del Campo, J., Kolisko, M., Okamoto, N., ... & Hallam, S. J. (2018). Single cell genomics of uncultured marine alveolates shows paraphyly of basal dinoflagellates. The ISME journal, 12(1), 304.

Tanaka, K., Miyamoto, N., Shouguchi-Miyata, J., & Ikeda, J. E. (2006). HFM1, the human homologue of yeast Mer3, encodes a putative DNA helicase expressed specifically in germ-line cells. DNA Sequence, 17(3), 242-246.

Theis, K., Chen, P. J., Skorvaga, M., Van Houten, B., & Kisker, C. (1999). Crystal structure of UvrB, a DNA helicase adapted for nucleotide excision repair. The EMBO journal, 18(24), 6899-6907.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., ... & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome research, 13(9), 2129-2141.

Tillmann, U., & Hoppenrath, M. (2013). Life cycle of the pseudocolonial dinoflagellate *Polykrikos kofoidii* (Gymnodiniales, Dinoflagellata). Journal of phycology, 49(2), 298-317.

Watanabe, Y. and Nurse, P. (1999). Cohesin Rec8 is required for reductional chromosome segregation at meiosis. Nature 400,461 -464.

Weedall, G. D., & Hall, N. (2015). Sexual reproduction and genetic exchange in parasitic protists. Parasitology, 142(S1), S120-S127.

Table S1. Annotation of the references and Amoebobophyra protein sequences of meiotic genes.

Sequence	Database	Domain	domain anotation	start	end	InterproScan	InterproScan annotation
DMC1+RAD51							
SMIN|symbB.v1.2.008353.t1	Pfam	PF08423	Rad51	537	787	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
SMIN|symbB.v1.2.008353.t1	Pfam	PF14520	Helix-hairpin-helix domain	483	532		
GSA120T00010080001	Pfam	PF08423	Rad51	119	309	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
GSA120T00010080001	SMART	SM00382		153	318	IPR003593	AAA+ ATPase domain
Q39009_Arabidopsis_DMC1	Pfam	PF08423	Rad51	91	342	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
Q39009_Arabidopsis_DMC1	SMART	SM00382		125	311	IPR003593	AAA+ ATPase domain
Skawagutii_1	Pfam	PF08423	Rad51	3	68	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
GSA25T00010532001	SMART	SM00382		224	410	IPR003593	AAA+ ATPase domain
GSA25T00010532001	Pfam	PF08423	Rad51	190	440	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
NP_035364.1_Mus_Rad51	SMART	SM00382		119	306	IPR003593	AAA+ ATPase domain
NP_035364.1_Mus_Rad51	Pfam	PF14520	Helix-hairpin-helix domain	32	79		
NP_035364.1_Mus_Rad51	Pfam	PF08423	Rad51	84	336	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
XP_001347762.2_Plasmodium_DMC1	SMART	SM00382		129	316	IPR003593	AAA+ ATPase domain
XP_001347762.2_Plasmodium_DMC1	Pfam	PF08423	Rad51	95	346	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
AAN76809.1_Plasmodium_Rad51	SMART	SM00382		129	316	IPR003593	AAA+ ATPase domain
AAN76809.1_Plasmodium_Rad51	Pfam	PF08423	Rad51	95	346	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
Q61880_DMC1_MOUSE	Pfam	PF08423	Rad51	84	337	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
Q61880_DMC1_MOUSE	SMART	SM00382		118	307	IPR003593	AAA+ ATPase domain
P25453_DMC1_YEAST	Pfam	PF08423	Rad51	79	331	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
P25453_DMC1_YEAST	SMART	SM00382		113	301	IPR003593	AAA+ ATPase domain
P25453_DMC1_YEAST	Pfam	PF14520	Helix-hairpin-helix domain	26	74		
Skawagutii_2	Pfam	PF08423	Rad51	1	41	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
CAB90141.1_Schizosaccharomyces_Rad51	Pfam	PF14520	Helix-hairpin-helix domain	53	100		
CAB90141.1_Schizosaccharomyces_Rad51	Pfam	PF08423	Rad51	106	357	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
CAB90141.1_Schizosaccharomyces_Rad51	SMART	SM00382		141	327	IPR003593	AAA+ ATPase domain
AAV38511.1_Homo_Rad51	Pfam	PF14520	Helix-hairpin-helix domain	32	79		
AAV38511.1_Homo_Rad51	Pfam	PF08423	Rad51	84	336	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
AAV38511.1_Homo_Rad51	SMART	SM00382		119	306	IPR003593	AAA+ ATPase domain
DMC1_Vbra_4727.T1	SMART	SM00382		124	310	IPR003593	AAA+ ATPase domain
DMC1_Vbra_4727.T1	Pfam	PF08423	Rad51	90	341	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
EGR30280.1_Ichthyophthirius_DMC1	SMART	SM00382		99	287	IPR003593	AAA+ ATPase domain
EGR30280.1_Ichthyophthirius_DMC1	Pfam	PF08423	Rad51	66	317	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
O42634|DMC1_SCHPO	SMART	SM00382		112	300	IPR003593	AAA+ ATPase domain
O42634|DMC1_SCHPO	Pfam	PF08423	Rad51	78	330	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
XP_001024231.1_Tetrahymena_DMC1	Pfam	PF08423	Rad51	95	346	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
XP_001024231.1_Tetrahymena_DMC1	SMART	SM00382		128	316	IPR003593	AAA+ ATPase domain
GSA120T00009824001	SMART	SM00382		253	439	IPR003593	AAA+ ATPase domain
GSA120T00009824001	Pfam	PF08423	Rad51	218	469	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
GSA25T00018686001	Pfam	PF14520	Helix-hairpin-helix domain	41	88		
GSA25T00018686001	SMART	SM00382		133	321	IPR003593	AAA+ ATPase domain
GSA25T00018686001	Pfam	PF08423	Rad51	99	350	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
symbB.v1.2.000608.t1	SMART	SM00382		427	616	IPR003593	AAA+ ATPase domain

symbB.v1.2.000608.t1	Pfam	PF08423	Rad51	393	643	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
symbB.v1.2.000608.t1	Pfam	PF14520	Helix-hairpin-helix domain	339	388		
XP_001349356.2_Plasmodium_DMC1	Pfam	PF08423	Rad51	93	345	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
EAA15553.1_Plasmodium_Rad51	SMART	SM00382		131	318	IPR003593	AAA+ ATPase domain
EAA15553.1_Plasmodium_Rad51	Pfam	PF08423	Rad51	97	348	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
NP_568402.1_Arabidopsis_Rad51	SMART	SM00382		122	309	IPR003593	AAA+ ATPase domain
NP_568402.1_Arabidopsis_Rad51	Pfam	PF08423	Rad51	87	340	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
NP_011021.3_Saccharomyces_Rad51	Pfam	PF08423	Rad51	142	394	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
NP_011021.3_Saccharomyces_Rad51	Pfam	PF14520	Helix-hairpin-helix domain	86	136		
NP_011021.3_Saccharomyces_Rad51	SMART	SM00382		177	364	IPR003593	AAA+ ATPase domain
Q14565|DMC1_HUMAN	SMART	SM00382		118	307	IPR003593	AAA+ ATPase domain
Q14565|DMC1_HUMAN	Pfam	PF08423	Rad51	84	337	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
DMC1_Vbra_17182.t1	Pfam	PF08423	Rad51	62	135	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal
EJY88098.1_Oxytricha_DMC1	SMART	SM00382		66	253	IPR003593	AAA+ ATPase domain
EJY88098.1_Oxytricha_DMC1	Pfam	PF08423	Rad51	31	284	IPR013632	DNA recombination and repair protein Rad51-like; C-terminal

HOP2+MND1

EGR32283.1_Ichthyophthirius_HOP2	PANTHER	PTHR15938		1	184	IPR010776	Homologous-pairing protein 2
EGR32283.1_Ichthyophthirius_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	1	161	IPR010776	Homologous-pairing protein 2
Q9HGK2_Schizosaccharomyces_HOP2	PANTHER	PTHR15938		1	210	IPR010776	Homologous-pairing protein 2
Q9HGK2_Schizosaccharomyces_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	16	184	IPR010776	Homologous-pairing protein 2
O35047_MOUSE_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	13	180	IPR010776	Homologous-pairing protein 2
O35047_MOUSE_HOP2	PANTHER	PTHR15938		2	216	IPR010776	Homologous-pairing protein 2
A0A1A8WB37_Plasmodium_HOP2	PANTHER	PTHR15938		509	732	IPR010776	Homologous-pairing protein 2
A0A1A8WB37_Plasmodium_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	532	698	IPR010776	Homologous-pairing protein 2
B9QAX9_Toxoplasma_HOP2	PANTHER	PTHR15938		392	601	IPR010776	Homologous-pairing protein 2
B9QAX9_Toxoplasma_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	402	566	IPR010776	Homologous-pairing protein 2
XP_001013509.2_Tetrahymena_HOP2	PANTHER	PTHR15938		182	417	IPR010776	Homologous-pairing protein 2
XP_001013509.2_Tetrahymena_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	223	389	IPR010776	Homologous-pairing protein 2
A0A1A8W5R7_Plasmodium_HOP2	Pfam	PF00581	Rhodanese-like domain	49	181	IPR001763	Rhodanese-like domain
A0A1A8W5R7_Plasmodium_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	907	1073	IPR010776	Homologous-pairing protein 2
A0A1A8W5R7_Plasmodium_HOP2	SMART	SM00450		48	185	IPR001763	Rhodanese-like domain
A0A1A8W5R7_Plasmodium_HOP2	SMART	SM00450		237	368	IPR001763	Rhodanese-like domain
A0A1A8W5R7_Plasmodium_HOP2	PANTHER	PTHR11364:SF17		846	846		
A0A1A8W5R7_Plasmodium_HOP2	PANTHER	PTHR11364		846	846		
NP_011482.2_Saccharomyces_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	18	187	IPR010776	Homologous-pairing protein 2
NP_011482.2_Saccharomyces_HOP2	PANTHER	PTHR15938		5	197	IPR010776	Homologous-pairing protein 2
Vbra_4454.t1_HOP2	PANTHER	PTHR15938		73	328	IPR010776	Homologous-pairing protein 2
Vbra_4454.t1_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	133	304	IPR010776	Homologous-pairing protein 2
Vbra_6181_MND1	Pfam	PF03962	Mnd1 family	24	205	IPR005647	Meiotic nuclear division protein 1
A0A1D3TL39_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	16	198	IPR005647	Meiotic nuclear division protein 1
XP_002174884.1_Schizosaccharomyces_MND1	Pfam	PF03962	Mnd1 family	16	203	IPR005647	Meiotic nuclear division protein 1
XP_001020981.3_Tetrahymena_HOP2	PANTHER	PTHR15938		126	342	IPR010776	Homologous-pairing protein 2
XP_001020981.3_Tetrahymena_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	143	310	IPR010776	Homologous-pairing protein 2
EJY77705.1_Oxytricha_HOP2	PANTHER	PTHR15938		214	486	IPR010776	Homologous-pairing protein 2

EJY77705.1_Oxytricha_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	270	450	IPR010776	Homologous-pairing protein 2
GSA120T00000161001	PANTHER	PTHR15938		34	248	IPR010776	Homologous-pairing protein 2
S8GCP0_Toxoplasma_HOP2	PANTHER	PTHR15938		392	601	IPR010776	Homologous-pairing protein 2
S8GCP0_Toxoplasma_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	402	566	IPR010776	Homologous-pairing protein 2
S8F3J1_Toxoplasma_MND1	Pfam	PF03962	Mnd1 family	18	193	IPR005647	Meiotic nuclear division protein 1
P53187_YEAST_HOP2	PANTHER	PTHR15938		5	212	IPR010776	Homologous-pairing protein 2
P53187_YEAST_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	18	187	IPR010776	Homologous-pairing protein 2
XP_001428092.1_Paramecium_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	75	240	IPR010776	Homologous-pairing protein 2
XP_001428092.1_Paramecium_HOP2	PANTHER	PTHR15938		59	267	IPR010776	Homologous-pairing protein 2
GSA25T00018581001_MND1	Pfam	PF03962	Mnd1 family	66	235	IPR005647	Meiotic nuclear division protein 1
A0A1C3L1A1_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	16	200	IPR005647	Meiotic nuclear division protein 1
GSA120T00008837001_MND1	Pfam	PF03962	Mnd1 family	113	310	IPR005647	Meiotic nuclear division protein 1
Q8I5Y0_Plasmodium_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	448	614	IPR010776	Homologous-pairing protein 2
Q8I5Y0_Plasmodium_HOP2	PANTHER	PTHR15938		369	650	IPR010776	Homologous-pairing protein 2
Q91ZY6_RAT_HOP2	PANTHER	PTHR15938		2	216	IPR010776	Homologous-pairing protein 2
Q91ZY6_RAT_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	13	180	IPR010776	Homologous-pairing protein 2
symbB.v1.2.026766.t1_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	2	56	IPR010776	Homologous-pairing protein 2
symbB.v1.2.026766.t1_HOP2	PANTHER	PTHR15938		2	56	IPR010776	Homologous-pairing protein 2
A0A1Y1JJP6_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	16	200	IPR005647	Meiotic nuclear division protein 1
C5LY51_Perkinsus_MND1	Pfam	PF03962	Mnd1 family	19	203	IPR005647	Meiotic nuclear division protein 1
A0A1G4HHA9_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	16	200	IPR005647	Meiotic nuclear division protein 1
symbB.v1.2.036043.t1_MND1	Pfam	PF03962	Mnd1 family	393	453	IPR005647	Meiotic nuclear division protein 1
Q9FX64_Arabidopsis_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	8	176	IPR010776	Homologous-pairing protein 2
Q9FX64_Arabidopsis_HOP2	PANTHER	PTHR15938		5	213	IPR010776	Homologous-pairing protein 2
A0A1A8W126_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	43	210	IPR005647	Meiotic nuclear division protein 1
Q9P2W1_HUMAN_HOP2	PANTHER	PTHR15938		8	216	IPR010776	Homologous-pairing protein 2
Q9P2W1_HUMAN_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	12	180	IPR010776	Homologous-pairing protein 2
EJY87085.1_Oxytricha_MND1	Pfam	PF03962	Mnd1 family	15	195	IPR005647	Meiotic nuclear division protein 1
Q63ZL2_Xenopus_HOP2	PANTHER	PTHR15938		4	213	IPR010776	Homologous-pairing protein 2
Q63ZL2_Xenopus_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	9	177	IPR010776	Homologous-pairing protein 2
A0A1J1HD13_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	17	200	IPR005647	Meiotic nuclear division protein 1
C6S3J7_Plasmodium_MND1	Pfam	PF03962	Mnd1 family	16	199	IPR005647	Meiotic nuclear division protein 1
XP_001437773.1_Paramecium_MND1	Pfam	PF03962	Mnd1 family	17	201	IPR005647	Meiotic nuclear division protein 1
XP_001024593.2_Tetrahymena_MND1	PANTHER	PTHR31398		985	1119		
GSA25T00024872001_HOP2	PANTHER	PTHR15938		146	277	IPR010776	Homologous-pairing protein 2
EGR29606.1_Ichthyophthirius_MND1	Pfam	PF03962	Mnd1 family	16	201	IPR005647	Meiotic nuclear division protein 1
EJY69481.1_Oxytricha_HOP2	Pfam	PF07106	Tat binding protein 1(TBP-1)-interacting protein (TBPIP)	223	392	IPR010776	Homologous-pairing protein 2
EJY69481.1_Oxytricha_HOP2	PANTHER	PTHR15938		172	423	IPR010776	Homologous-pairing protein 2
Vbra_4074_MND1	Pfam	PF03962	Mnd1 family	1	172	IPR005647	Meiotic nuclear division protein 1
EJY82981.1_Oxytricha_MND1	Pfam	PF03962	Mnd1 family	18	202	IPR005647	Meiotic nuclear division protein 1
XP_001015513.1_Tetrahymena_MND1	Pfam	PF03962	Mnd1 family	17	202	IPR005647	Meiotic nuclear division protein 1
NP_011332.2_Saccharomyces_MND1	Pfam	PF03962	Mnd1 family	17	213	IPR005647	Meiotic nuclear division protein 1
MER3							
GSA25T00021733001	SMART	SM00490		1866	1963	IPR001650	Helicase; C-terminal

GSA25T00021733001	Pfam	PF00270	DEAD/DEAH box helicase	1216	1382	IPR011545	DEAD/DEAH box helicase domain
GSA25T00021733001	SMART	SM00487		1209	1414	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00026232001	SMART	SM00487		380	599	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00026232001	SMART	SM00490		679	771	IPR001650	Helicase; C-terminal
GSA120T00026232001	Pfam	PF00271	Helicase conserved C-terminal domain	625	771	IPR001650	Helicase; C-terminal
GSA120T00026232001	Pfam	PF00270	DEAD/DEAH box helicase	386	569	IPR011545	DEAD/DEAH box helicase domain
GSA120T00026234001	SMART	SM00487		380	599	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00026234001	SMART	SM00490		679	771	IPR001650	Helicase; C-terminal
GSA120T00026234001	Pfam	PF00271	Helicase conserved C-terminal domain	625	771	IPR001650	Helicase; C-terminal
GSA120T00026234001	Pfam	PF00270	DEAD/DEAH box helicase	386	569	IPR011545	DEAD/DEAH box helicase domain
GSA120T00018887001	SMART	SM00490		619	722	IPR001650	Helicase; C-terminal
GSA120T00018887001	Pfam	PF00270	DEAD/DEAH box helicase	351	521	IPR011545	DEAD/DEAH box helicase domain
GSA120T00018887001	SMART	SM00487		345	557	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00018887001	SMART	SM00973		839	1137	IPR004179	Sec63 domain
GSA120T00018887001	SMART	SM00973		1	307	IPR004179	Sec63 domain
GSA120T00018887001	Pfam	PF02889	Sec63 Brl domain	1	304	IPR004179	Sec63 domain
GSA120T00018887001	Pfam	PF02889	Sec63 Brl domain	840	1080	IPR004179	Sec63 domain
Vbra_14058.t1_MER3	SMART	SM00487		129	339	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Vbra_14058.t1_MER3	SMART	SM00973		618	936	IPR004179	Sec63 domain
Vbra_14058.t1_MER3	SMART	SM00490		414	504	IPR001650	Helicase; C-terminal
Vbra_14058.t1_MER3	Pfam	PF02889	Sec63 Brl domain	618	933	IPR004179	Sec63 domain
Vbra_14058.t1_MER3	Pfam	PF00270	DEAD/DEAH box helicase	136	306	IPR011545	DEAD/DEAH box helicase domain
Vbra_14058.t1_MER3	Pfam	PF00271	Helicase conserved C-terminal domain	361	502	IPR001650	Helicase; C-terminal
NP_011263.2_Saccharomyces_MER3	SMART	SM00490		411	497	IPR001650	Helicase; C-terminal
NP_011263.2_Saccharomyces_MER3	SMART	SM00973		616	938	IPR004179	Sec63 domain
NP_011263.2_Saccharomyces_MER3	SMART	SM00487		135	334	IPR014001	Helicase superfamily 1/2; ATP-binding domain
NP_011263.2_Saccharomyces_MER3	Pfam	PF02889	Sec63 Brl domain	616	911	IPR004179	Sec63 domain
NP_011263.2_Saccharomyces_MER3	Pfam	PF00270	DEAD/DEAH box helicase	141	309	IPR011545	DEAD/DEAH box helicase domain
NP_011263.2_Saccharomyces_MER3	Pfam	PF00271	Helicase conserved C-terminal domain	360	495	IPR001650	Helicase; C-terminal
GSA25T00003083001	SMART	SM00382		543	769	IPR003593	AAA+ ATPase domain
GSA25T00003083001	Pfam	PF02889	Sec63 Brl domain	1036	1342	IPR004179	Sec63 domain
GSA25T00003083001	Pfam	PF02889	Sec63 Brl domain	1878	2204	IPR004179	Sec63 domain
GSA25T00003083001	Pfam	PF00270	DEAD/DEAH box helicase	1389	1557	IPR011545	DEAD/DEAH box helicase domain
GSA25T00003083001	Pfam	PF00270	DEAD/DEAH box helicase	530	710	IPR011545	DEAD/DEAH box helicase domain
GSA25T00003083001	SMART	SM00490		820	912	IPR001650	Helicase; C-terminal
GSA25T00003083001	SMART	SM00490		1672	1760	IPR001650	Helicase; C-terminal
GSA25T00003083001	Pfam	PF00271	Helicase conserved C-terminal domain	822	912	IPR001650	Helicase; C-terminal
GSA25T00003083001	SMART	SM00973		1035	1345	IPR004179	Sec63 domain
GSA25T00003083001	SMART	SM00973		1877	2206	IPR004179	Sec63 domain
GSA25T00003083001	SMART	SM00487		1383	1590	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA25T00003083001	SMART	SM00487		524	740	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00973		1812	2177	IPR004179	Sec63 domain
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00973		978	1287	IPR004179	Sec63 domain
Q8N3C0.3_HUMAN_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	704	855	IPR001650	Helicase; C-terminal
Q8N3C0.3_HUMAN_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	1607	1692	IPR001650	Helicase; C-terminal
Q8N3C0.3_HUMAN_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1812	2175	IPR004179	Sec63 domain

Q8N3C0.3_HUMAN_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	978	1283	IPR004179	Sec63 domain
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00490		769	857	IPR001650	Helicase; C-terminal
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00490		1604	1694	IPR001650	Helicase; C-terminal
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00382		1341	1491	IPR003593	AAA+ ATPase domain
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00382		491	685	IPR003593	AAA+ ATPase domain
Q8N3C0.3_HUMAN_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	479	653	IPR011545	DEAD/DEAH box helicase domain
Q8N3C0.3_HUMAN_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1329	1494	IPR011545	DEAD/DEAH box helicase domain
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00487		1323	1527	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q8N3C0.3_HUMAN_RNA_helicase	SMART	SM00487		473	685	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00009666001	SMART	SM00490		449	536	IPR001650	Helicase; C-terminal
GSA120T00009666001	Pfam	PF00271	Helicase conserved C-terminal domain	460	536	IPR001650	Helicase; C-terminal
GSA120T00009666001	Pfam	PF00270	DEAD/DEAH box helicase	43	194	IPR011545	DEAD/DEAH box helicase domain
GSA120T00009666001	SMART	SM00487		37	223	IPR014001	Helicase superfamily 1/2; ATP-binding domain
P53327.2_YEAST_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	796	1099	IPR004179	Sec63 domain
P53327.2_YEAST_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1631	1937	IPR004179	Sec63 domain
P53327.2_YEAST_RNA_helicase	SMART	SM00973		1626	1965	IPR004179	Sec63 domain
P53327.2_YEAST_RNA_helicase	SMART	SM00973		795	1100	IPR004179	Sec63 domain
P53327.2_YEAST_RNA_helicase	SMART	SM00487		284	500	IPR014001	Helicase superfamily 1/2; ATP-binding domain
P53327.2_YEAST_RNA_helicase	SMART	SM00487		1136	1336	IPR014001	Helicase superfamily 1/2; ATP-binding domain
P53327.2_YEAST_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1142	1307	IPR011545	DEAD/DEAH box helicase domain
P53327.2_YEAST_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	290	469	IPR011545	DEAD/DEAH box helicase domain
P53327.2_YEAST_RNA_helicase	SMART	SM00490		581	673	IPR001650	Helicase; C-terminal
P53327.2_YEAST_RNA_helicase	SMART	SM00490		1417	1509	IPR001650	Helicase; C-terminal
P53327.2_YEAST_RNA_helicase	SMART	SM00382		302	489	IPR003593	AAA+ ATPase domain
P53327.2_YEAST_RNA_helicase	SMART	SM00382		1154	1388	IPR003593	AAA+ ATPase domain
P53327.2_YEAST_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	530	672	IPR001650	Helicase; C-terminal
P53327.2_YEAST_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	1422	1507	IPR001650	Helicase; C-terminal
Q5D892_MER3_Arabidopsis	SMART	SM00973		534	852	IPR004179	Sec63 domain
Q5D892_MER3_Arabidopsis	Pfam	PF00271	Helicase conserved C-terminal domain	262	413	IPR001650	Helicase; C-terminal
Q5D892_MER3_Arabidopsis	SMART	SM00490		327	415	IPR001650	Helicase; C-terminal
Q5D892_MER3_Arabidopsis	SMART	SM00487		21	241	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q5D892_MER3_Arabidopsis	Pfam	PF00270	DEAD/DEAH box helicase	28	215	IPR011545	DEAD/DEAH box helicase domain
Q5D892_MER3_Arabidopsis	Pfam	PF02889	Sec63 Brl domain	534	849	IPR004179	Sec63 domain
Q55CI8.1_Dictyostelium_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	1624	1767	IPR001650	Helicase; C-terminal
Q55CI8.1_Dictyostelium_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	784	927	IPR001650	Helicase; C-terminal
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00487		548	759	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00487		1394	1596	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q55CI8.1_Dictyostelium_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1892	2213	IPR004179	Sec63 domain
Q55CI8.1_Dictyostelium_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1050	1355	IPR004179	Sec63 domain
Q55CI8.1_Dictyostelium_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	554	730	IPR011545	DEAD/DEAH box helicase domain
Q55CI8.1_Dictyostelium_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1400	1566	IPR011545	DEAD/DEAH box helicase domain
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00973		1050	1356	IPR004179	Sec63 domain
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00973		1892	2215	IPR004179	Sec63 domain
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00490		1680	1766	IPR001650	Helicase; C-terminal
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00490		838	929	IPR001650	Helicase; C-terminal
Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00382		1412	1580	IPR003593	AAA+ ATPase domain

Q55CI8.1_Dictyostelium_RNA_helicase	SMART	SM00382		566	882	IPR003593	AAA+ ATPase domain
Q8TL39.1_Methanosarcina_DNA_helicase	SMART	SM00487		20	216	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q8TL39.1_Methanosarcina_DNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	26	182	IPR011545	DEAD/DEAH box helicase domain
Q8TL39.1_Methanosarcina_DNA_helicase	SMART	SM00490		292	383	IPR001650	Helicase; C-terminal
Q8TL39.1_Methanosarcina_DNA_helicase	Pfam	PF14520	Helix-hairpin-helix domain	651	700		
Q8TL39.1_Methanosarcina_DNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	241	381	IPR001650	Helicase; C-terminal
GSA120T00020507001	SMART	SM00490		534	620	IPR001650	Helicase; C-terminal
GSA120T00020507001	Pfam	PF13234	rRNA-processing arch domain	674	1142	IPR025696	rRNA-processing arch domain
GSA120T00020507001	SMART	SM00487		224	407	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00020507001	Pfam	PF00270	DEAD/DEAH box helicase	231	376	IPR011545	DEAD/DEAH box helicase domain
GSA120T00020507001	SMART	SM01142		1252	1476	IPR012961	ATP-dependent RNA helicase Ski2; C-terminal
GSA120T00020507001	Pfam	PF08148	DSHCT (NUC185) domain	1254	1471	IPR012961	ATP-dependent RNA helicase Ski2; C-terminal
GSA120T00020507001	Pfam	PF00271	Helicase conserved C-terminal domain	543	620	IPR001650	Helicase; C-terminal
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00487		1323	1523	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00487		479	689	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q54G57.1_Dictyostelium_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1810	2163	IPR004179	Sec63 domain
Q54G57.1_Dictyostelium_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	981	1280	IPR004179	Sec63 domain
Q54G57.1_Dictyostelium_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1329	1494	IPR011545	DEAD/DEAH box helicase domain
Q54G57.1_Dictyostelium_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	485	661	IPR011545	DEAD/DEAH box helicase domain
Q54G57.1_Dictyostelium_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	1605	1691	IPR001650	Helicase; C-terminal
Q54G57.1_Dictyostelium_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	709	857	IPR001650	Helicase; C-terminal
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00490		767	859	IPR001650	Helicase; C-terminal
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00490		1605	1693	IPR001650	Helicase; C-terminal
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00973		1810	2165	IPR004179	Sec63 domain
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00973		980	1286	IPR004179	Sec63 domain
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00382		1341	1507	IPR003593	AAA+ ATPase domain
Q54G57.1_Dictyostelium_RNA_helicase	SMART	SM00382		497	686	IPR003593	AAA+ ATPase domain
GSA25T00015231001	Pfam	PF00270	DEAD/DEAH box helicase	27	184	IPR011545	DEAD/DEAH box helicase domain
GSA25T00015231001	SMART	SM00487		9	207	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA25T00015231001	Pfam	PF00271	Helicase conserved C-terminal domain	357	439	IPR001650	Helicase; C-terminal
GSA25T00015231001	SMART	SM00490		347	439	IPR001650	Helicase; C-terminal
GSA25T00014432001	SMART	SM00490		341	429	IPR001650	Helicase; C-terminal
GSA25T00014432001	Pfam	PF00271	Helicase conserved C-terminal domain	303	429	IPR001650	Helicase; C-terminal
GSA25T00014432001	Pfam	PF00270	DEAD/DEAH box helicase	934	1097	IPR011545	DEAD/DEAH box helicase domain
GSA25T00014432001	Pfam	PF00270	DEAD/DEAH box helicase	22	213	IPR011545	DEAD/DEAH box helicase domain
GSA25T00014432001	SMART	SM00487		928	1131	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA25T00014432001	SMART	SM00487		16	242	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA25T00014432001	SMART	SM00973		556	891	IPR004179	Sec63 domain
GSA25T00014432001	Pfam	PF02889	Sec63 Brl domain	556	889	IPR004179	Sec63 domain
GSA120T00004338001	SMART	SM00487		1122	1324	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00004338001	SMART	SM00490		1773	1856	IPR001650	Helicase; C-terminal
GSA120T00004338001	Pfam	PF00271	Helicase conserved C-terminal domain	1777	1856	IPR001650	Helicase; C-terminal
GSA120T00004338001	Pfam	PF00270	DEAD/DEAH box helicase	1130	1295	IPR011545	DEAD/DEAH box helicase domain
I8TS41_Aspergillus_MER3	Pfam	PF00271	Helicase conserved C-terminal domain	366	495	IPR001650	Helicase; C-terminal
I8TS41_Aspergillus_MER3	SMART	SM00973		612	931	IPR004179	Sec63 domain
I8TS41_Aspergillus_MER3	SMART	SM00487		132	331	IPR014001	Helicase superfamily 1/2; ATP-binding domain

I8TS41_Aspergillus_MER3	Pfam	PF00270	DEAD/DEAH box helicase	138	303	IPR011545	DEAD/DEAH box helicase domain
I8TS41_Aspergillus_MER3	Pfam	PF02889	Sec63 Brl domain	612	921	IPR004179	Sec63 domain
I8TS41_Aspergillus_MER3	SMART	SM00490		388	497	IPR001650	Helicase; C-terminal
NP_001017975.4_MER3_Human	SMART	SM00487		277	492	IPR014001	Helicase superfamily 1/2; ATP-binding domain
NP_001017975.4_MER3_Human	SMART	SM00973		777	1092	IPR004179	Sec63 domain
NP_001017975.4_MER3_Human	Pfam	PF00270	DEAD/DEAH box helicase	284	462	IPR011545	DEAD/DEAH box helicase domain
NP_001017975.4_MER3_Human	SMART	SM00490		573	659	IPR001650	Helicase; C-terminal
NP_001017975.4_MER3_Human	Pfam	PF00271	Helicase conserved C-terminal domain	525	657	IPR001650	Helicase; C-terminal
NP_001017975.4_MER3_Human	Pfam	PF02889	Sec63 Brl domain	777	1090	IPR004179	Sec63 domain
O48534.1_Arabidopsis_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1354	1521	IPR011545	DEAD/DEAH box helicase domain
O48534.1_Arabidopsis_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	508	681	IPR011545	DEAD/DEAH box helicase domain
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00487		502	714	IPR014001	Helicase superfamily 1/2; ATP-binding domain
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00487		1348	1554	IPR014001	Helicase superfamily 1/2; ATP-binding domain
O48534.1_Arabidopsis_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1840	2157	IPR004179	Sec63 domain
O48534.1_Arabidopsis_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1007	1309	IPR004179	Sec63 domain
O48534.1_Arabidopsis_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	741	883	IPR001650	Helicase; C-terminal
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00973		1006	1311	IPR004179	Sec63 domain
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00973		1840	2159	IPR004179	Sec63 domain
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00490		788	885	IPR001650	Helicase; C-terminal
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00490		1635	1723	IPR001650	Helicase; C-terminal
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00382		1366	1559	IPR003593	AAA+ ATPase domain
O48534.1_Arabidopsis_RNA_helicase	SMART	SM00382		520	698	IPR003593	AAA+ ATPase domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00973		1008	1316	IPR004179	Sec63 domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00973		1839	2152	IPR004179	Sec63 domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00490		799	887	IPR001650	Helicase; C-terminal
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00490		1631	1723	IPR001650	Helicase; C-terminal
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00382		522	703	IPR003593	AAA+ ATPase domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00382		1370	1545	IPR003593	AAA+ ATPase domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	1633	1721	IPR001650	Helicase; C-terminal
Q9FNQ1.1_Arabidopsis_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	737	885	IPR001650	Helicase; C-terminal
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00487		1352	1552	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	SMART	SM00487		504	715	IPR014001	Helicase superfamily 1/2; ATP-binding domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1358	1522	IPR011545	DEAD/DEAH box helicase domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	510	684	IPR011545	DEAD/DEAH box helicase domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1839	2150	IPR004179	Sec63 domain
Q9FNQ1.1_Arabidopsis_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1008	1315	IPR004179	Sec63 domain
O73946.1_Pyrococcus_DNA_helicase	SMART	SM00278		646	665	IPR003583	Helix-hairpin-helix DNA-binding motif; class 1
O73946.1_Pyrococcus_DNA_helicase	SMART	SM00278		679	698	IPR003583	Helix-hairpin-helix DNA-binding motif; class 1
O73946.1_Pyrococcus_DNA_helicase	SMART	SM00487		20	204	IPR014001	Helicase superfamily 1/2; ATP-binding domain
O73946.1_Pyrococcus_DNA_helicase	Pfam	PF14520	Helix-hairpin-helix domain	648	700		
O73946.1_Pyrococcus_DNA_helicase	SMART	SM00490		287	373	IPR001650	Helicase; C-terminal
O73946.1_Pyrococcus_DNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	235	372	IPR001650	Helicase; C-terminal
O73946.1_Pyrococcus_DNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	26	180	IPR011545	DEAD/DEAH box helicase domain
GSA25T00025043001	Pfam	PF13234	rRNA-processing arch domain	318	381	IPR025696	rRNA-processing arch domain
GSA25T00025043001	Pfam	PF13234	rRNA-processing arch domain	398	724	IPR025696	rRNA-processing arch domain
GSA25T00025043001	SMART	SM00490		175	264	IPR001650	Helicase; C-terminal

GSA25T00025043001	SMART	SM01142		832	1062	IPR012961	ATP-dependent RNA helicase Ski2; C-terminal
GSA25T00025043001	Pfam	PF00271	Helicase conserved C-terminal domain	185	264	IPR001650	Helicase; C-terminal
GSA25T00025043001	Pfam	PF08148	DSHCT (NUC185) domain	836	1057	IPR012961	ATP-dependent RNA helicase Ski2; C-terminal
P32639.2_YEAST_RNA_helicase	SMART	SM00487		1344	1549	IPR014001	Helicase superfamily 1/2; ATP-binding domain
P32639.2_YEAST_RNA_helicase	SMART	SM00487		495	707	IPR014001	Helicase superfamily 1/2; ATP-binding domain
P32639.2_YEAST_RNA_helicase	SMART	SM00490		787	879	IPR001650	Helicase; C-terminal
P32639.2_YEAST_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1848	2158	IPR004179	Sec63 domain
P32639.2_YEAST_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	998	1306	IPR004179	Sec63 domain
P32639.2_YEAST_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	727	877	IPR001650	Helicase; C-terminal
P32639.2_YEAST_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1350	1515	IPR011545	DEAD/DEAH box helicase domain
P32639.2_YEAST_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	501	677	IPR011545	DEAD/DEAH box helicase domain
P32639.2_YEAST_RNA_helicase	SMART	SM00973		998	1308	IPR004179	Sec63 domain
P32639.2_YEAST_RNA_helicase	SMART	SM00973		1846	2160	IPR004179	Sec63 domain
GSA120T00004033001	Pfam	PF00271	Helicase conserved C-terminal domain	1863	1939	IPR001650	Helicase; C-terminal
GSA120T00004033001	SMART	SM00487		1517	1721	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00004033001	SMART	SM00487		586	836	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA120T00004033001	SMART	SM00973		1139	1481	IPR004179	Sec63 domain
GSA120T00004033001	SMART	SM00973		2092	2552	IPR004179	Sec63 domain
GSA120T00004033001	Pfam	PF02889	Sec63 Brl domain	2469	2550	IPR004179	Sec63 domain
GSA120T00004033001	Pfam	PF02889	Sec63 Brl domain	2092	2386	IPR004179	Sec63 domain
GSA120T00004033001	Pfam	PF02889	Sec63 Brl domain	1139	1478	IPR004179	Sec63 domain
GSA120T00004033001	Pfam	PF00270	DEAD/DEAH box helicase	593	808	IPR011545	DEAD/DEAH box helicase domain
GSA120T00004033001	Pfam	PF00270	DEAD/DEAH box helicase	1523	1687	IPR011545	DEAD/DEAH box helicase domain
GSA120T00004033001	SMART	SM00490		1853	1941	IPR001650	Helicase; C-terminal
GSA120T00004033001	SMART	SM00490		931	1019	IPR001650	Helicase; C-terminal
GSA120T00004033001	SMART	SM00382		1535	1685	IPR003593	AAA+ ATPase domain
GSA120T00004033001	SMART	SM00382		604	873	IPR003593	AAA+ ATPase domain
GSA120T00016715001	Pfam	PF00271	Helicase conserved C-terminal domain	557	632	IPR001650	Helicase; C-terminal
GSA120T00016715001	SMART	SM00490		546	632	IPR001650	Helicase; C-terminal
B3LHU3_YEAST_MER3	SMART	SM00490		411	497	IPR001650	Helicase; C-terminal
B3LHU3_YEAST_MER3	SMART	SM00973		616	938	IPR004179	Sec63 domain
B3LHU3_YEAST_MER3	SMART	SM00487		135	334	IPR014001	Helicase superfamily 1/2; ATP-binding domain
B3LHU3_YEAST_MER3	Pfam	PF02889	Sec63 Brl domain	616	911	IPR004179	Sec63 domain
B3LHU3_YEAST_MER3	Pfam	PF00270	DEAD/DEAH box helicase	141	309	IPR011545	DEAD/DEAH box helicase domain
B3LHU3_YEAST_MER3	Pfam	PF00271	Helicase conserved C-terminal domain	360	495	IPR001650	Helicase; C-terminal
O75643.2_HUMAN_RNA_helicase	SMART	SM00487		477	690	IPR014001	Helicase superfamily 1/2; ATP-binding domain
O75643.2_HUMAN_RNA_helicase	SMART	SM00487		1324	1528	IPR014001	Helicase superfamily 1/2; ATP-binding domain
O75643.2_HUMAN_RNA_helicase	Pfam	PF00271	Helicase conserved C-terminal domain	714	858	IPR001650	Helicase; C-terminal
O75643.2_HUMAN_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	982	1285	IPR004179	Sec63 domain
O75643.2_HUMAN_RNA_helicase	Pfam	PF02889	Sec63 Brl domain	1812	2123	IPR004179	Sec63 domain
O75643.2_HUMAN_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	1330	1496	IPR011545	DEAD/DEAH box helicase domain
O75643.2_HUMAN_RNA_helicase	Pfam	PF00270	DEAD/DEAH box helicase	483	658	IPR011545	DEAD/DEAH box helicase domain
O75643.2_HUMAN_RNA_helicase	SMART	SM00490		1607	1695	IPR001650	Helicase; C-terminal
O75643.2_HUMAN_RNA_helicase	SMART	SM00490		768	860	IPR001650	Helicase; C-terminal
O75643.2_HUMAN_RNA_helicase	SMART	SM00973		981	1286	IPR004179	Sec63 domain
O75643.2_HUMAN_RNA_helicase	SMART	SM00973		1812	2124	IPR004179	Sec63 domain

A0A1Q9D472_Symbiodinium_MER3	Pfam	PF00270	DEAD/DEAH box helicase	263	400	IPR011545	DEAD/DEAH box helicase domain
A0A1Q9D472_Symbiodinium_MER3	Pfam	PF00271	Helicase conserved C-terminal domain	456	591	IPR001650	Helicase; C-terminal
A0A1Q9D472_Symbiodinium_MER3	Pfam	PF02889	Sec63 Brl domain	739	819	IPR004179	Sec63 domain
A0A1Q9D472_Symbiodinium_MER3	SMART	SM00490		503	593	IPR001650	Helicase; C-terminal
A0A1Q9D472_Symbiodinium_MER3	SMART	SM00487		128	432	IPR014001	Helicase superfamily 1/2; ATP-binding domain
GSA25T00017715001	SMART	SM00490		275	376	IPR001650	Helicase; C-terminal
GSA25T00014431001	Pfam	PF00271	Helicase conserved C-terminal domain	2	65	IPR001650	Helicase; C-terminal

MSH

NP_000242.1_Homo_MSH2	SMART	SM00534		662	849	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_000242.1_Homo_MSH2	Pfam	PF05190	MutS family domain IV	474	568	IPR007861	DNA mismatch repair protein MutS; clamp
NP_000242.1_Homo_MSH2	Pfam	PF01624	MutS domain I	18	131	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_000242.1_Homo_MSH2	Pfam	PF05188	MutS domain II	156	289	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_000242.1_Homo_MSH2	Pfam	PF00488	MutS domain V	665	852	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_000242.1_Homo_MSH2	Pfam	PF05192	MutS domain III	305	609	IPR007696	DNA mismatch repair protein MutS; core
NP_000242.1_Homo_MSH2	SMART	SM00533		321	645	IPR007696	DNA mismatch repair protein MutS; core
AAD21822.1_Homo_MSH5	Pfam	PF05192	MutS domain III	226	536	IPR007696	DNA mismatch repair protein MutS; core
AAD21822.1_Homo_MSH5	Pfam	PF00488	MutS domain V	589	776	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAD21822.1_Homo_MSH5	Pfam	PF05190	MutS family domain IV	398	496	IPR007861	DNA mismatch repair protein MutS; clamp
AAD21822.1_Homo_MSH5	SMART	SM00534		585	776	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAD21822.1_Homo_MSH5	SMART	SM00533		249	569	IPR007696	DNA mismatch repair protein MutS; core
AAB06045.1_Homo_MSH3	Pfam	PF05188	MutS domain II	357	513	IPR007860	DNA mismatch repair protein MutS; connector domain
AAB06045.1_Homo_MSH3	Pfam	PF00488	MutS domain V	883	1085	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAB06045.1_Homo_MSH3	Pfam	PF01624	MutS domain I	221	332	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
AAB06045.1_Homo_MSH3	Pfam	PF05192	MutS domain III	531	829	IPR007696	DNA mismatch repair protein MutS; core
AAB06045.1_Homo_MSH3	SMART	SM00534		880	1082	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAB06045.1_Homo_MSH3	SMART	SM00533		546	861	IPR007696	DNA mismatch repair protein MutS; core
NP_010016.2_Saccharomyces_MSH3	SMART	SM00533		435	768	IPR007696	DNA mismatch repair protein MutS; core
NP_010016.2_Saccharomyces_MSH3	Pfam	PF01624	MutS domain I	133	260	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_010016.2_Saccharomyces_MSH3	Pfam	PF05192	MutS domain III	423	735	IPR007696	DNA mismatch repair protein MutS; core
NP_010016.2_Saccharomyces_MSH3	Pfam	PF05188	MutS domain II	276	403	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_010016.2_Saccharomyces_MSH3	Pfam	PF05190	MutS family domain IV	612	689	IPR007861	DNA mismatch repair protein MutS; clamp
NP_010016.2_Saccharomyces_MSH3	Pfam	PF00488	MutS domain V	788	978	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_010016.2_Saccharomyces_MSH3	SMART	SM00534		784	976	IPR000432	DNA mismatch repair protein MutS; C-terminal
EAL65026.1_Dictyostelium_MSH5	Pfam	PF00488	MutS domain V	592	724	IPR000432	DNA mismatch repair protein MutS; C-terminal
EAL65026.1_Dictyostelium_MSH5	SMART	SM00534		589	826	IPR000432	DNA mismatch repair protein MutS; C-terminal
EAL65026.1_Dictyostelium_MSH5	Pfam	PF05192	MutS domain III	230	541	IPR007696	DNA mismatch repair protein MutS; core
EAL65026.1_Dictyostelium_MSH5	SMART	SM00533		253	574	IPR007696	DNA mismatch repair protein MutS; core
XP_638826.1_Dictyostelium_MSH4	SMART	SM00534		789	977	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_638826.1_Dictyostelium_MSH4	Pfam	PF05190	MutS family domain IV	600	691	IPR007861	DNA mismatch repair protein MutS; clamp
XP_638826.1_Dictyostelium_MSH4	Pfam	PF00488	MutS domain V	793	979	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_638826.1_Dictyostelium_MSH4	SMART	SM00533		416	768	IPR007696	DNA mismatch repair protein MutS; core
XP_638826.1_Dictyostelium_MSH4	Pfam	PF05192	MutS domain III	401	733	IPR007696	DNA mismatch repair protein MutS; core
XP_638826.1_Dictyostelium_MSH4	Pfam	PF05188	MutS domain II	249	374	IPR007860	DNA mismatch repair protein MutS; connector domain
GSA120T00017079001	SMART	SM00533		616	993	IPR007696	DNA mismatch repair protein MutS; core
GSA120T00017079001	Pfam	PF05192	MutS domain III	603	951	IPR007696	DNA mismatch repair protein MutS; core
GSA120T00017079001	Pfam	PF00488	MutS domain V	1012	1192	IPR000432	DNA mismatch repair protein MutS; C-terminal

GSA120T00017079001	Pfam	PF05188	MutS domain II	441	575	IPR007860	DNA mismatch repair protein MutS; connector domain
GSA120T00017079001	SMART	SM00534		1008	1194	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA120T00017079001	Pfam	PF01624	MutS domain I	306	428	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_192116.1_Arabidopsis_MSH6	Pfam	PF01624	MutS domain I	380	495	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_192116.1_Arabidopsis_MSH6	SMART	SM00533		716	1056	IPR007696	DNA mismatch repair protein MutS; core
NP_192116.1_Arabidopsis_MSH6	Pfam	PF00488	MutS domain V	1080	1270	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_192116.1_Arabidopsis_MSH6	Pfam	PF05188	MutS domain II	506	667	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_192116.1_Arabidopsis_MSH6	SMART	SM00534		1076	1268	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_192116.1_Arabidopsis_MSH6	Pfam	PF05192	MutS domain III	701	1017	IPR007696	DNA mismatch repair protein MutS; core
NP_192116.1_Arabidopsis_MSH6	Pfam	PF05190	MutS family domain IV	886	977	IPR007861	DNA mismatch repair protein MutS; clamp
NP_192116.1_Arabidopsis_MSH6	SMART	SM00333		121	179	IPR002999	Tudor domain
AAA34801.1_Saccharomyces_MSH1	Pfam	PF05188	MutS domain II	220	340	IPR007860	DNA mismatch repair protein MutS; connector domain
AAA34801.1_Saccharomyces_MSH1	SMART	SM00534		764	959	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAA34801.1_Saccharomyces_MSH1	SMART	SM00533		383	745	IPR007696	DNA mismatch repair protein MutS; core
AAA34801.1_Saccharomyces_MSH1	Pfam	PF05192	MutS domain III	369	713	IPR007696	DNA mismatch repair protein MutS; core
AAA34801.1_Saccharomyces_MSH1	Pfam	PF01624	MutS domain I	81	194	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
AAA34801.1_Saccharomyces_MSH1	Pfam	PF00488	MutS domain V	767	958	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_812851.1_MSH4_Trypanosoma	Pfam	PF05192	MutS domain III	111	604	IPR007696	DNA mismatch repair protein MutS; core
XP_812851.1_MSH4_Trypanosoma	Pfam	PF05190	MutS family domain IV	378	443	IPR007861	DNA mismatch repair protein MutS; clamp
XP_812851.1_MSH4_Trypanosoma	SMART	SM00534		653	859	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_812851.1_MSH4_Trypanosoma	SMART	SM00533		109	636	IPR007696	DNA mismatch repair protein MutS; core
XP_812851.1_MSH4_Trypanosoma	Pfam	PF00488	MutS domain V	657	858	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAB72039.1_Homo_MSH4	SMART	SM00533		330	657	IPR007696	DNA mismatch repair protein MutS; core
AAB72039.1_Homo_MSH4	Pfam	PF00488	MutS domain V	676	868	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAB72039.1_Homo_MSH4	Pfam	PF05188	MutS domain II	155	292	IPR007860	DNA mismatch repair protein MutS; connector domain
AAB72039.1_Homo_MSH4	SMART	SM00534		673	866	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAB72039.1_Homo_MSH4	Pfam	PF05190	MutS family domain IV	493	584	IPR007861	DNA mismatch repair protein MutS; clamp
AAB72039.1_Homo_MSH4	Pfam	PF05192	MutS domain III	316	627	IPR007696	DNA mismatch repair protein MutS; core
MSH4_Vbra	SMART	SM00533		232	564	IPR007696	DNA mismatch repair protein MutS; core
MSH4_Vbra	SMART	SM00534		579	767	IPR000432	DNA mismatch repair protein MutS; C-terminal
MSH4_Vbra	Pfam	PF05192	MutS domain III	217	528	IPR007696	DNA mismatch repair protein MutS; core
MSH4_Vbra	Pfam	PF00488	MutS domain V	583	769	IPR000432	DNA mismatch repair protein MutS; C-terminal
MSH4_Vbra	Pfam	PF05188	MutS domain II	66	193	IPR007860	DNA mismatch repair protein MutS; connector domain
EJY70841.1_Msh4_Oxytricha	SMART	SM00533		243	570	IPR007696	DNA mismatch repair protein MutS; core
EJY70841.1_Msh4_Oxytricha	Pfam	PF00488	MutS domain V	592	659	IPR000432	DNA mismatch repair protein MutS; C-terminal
EJY70841.1_Msh4_Oxytricha	Pfam	PF05192	MutS domain III	227	525	IPR007696	DNA mismatch repair protein MutS; core
EJY70841.1_Msh4_Oxytricha	Pfam	PF05190	MutS family domain IV	420	482	IPR007861	DNA mismatch repair protein MutS; clamp
EJY70841.1_Msh4_Oxytricha	SMART	SM00534		588	740	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_000170.1_Homo_MSH6	SMART	SM00533		753	1102	IPR007696	DNA mismatch repair protein MutS; core
NP_000170.1_Homo_MSH6	Pfam	PF00488	MutS domain V	1131	1323	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_000170.1_Homo_MSH6	Pfam	PF05190	MutS family domain IV	932	1024	IPR007861	DNA mismatch repair protein MutS; clamp
NP_000170.1_Homo_MSH6	Pfam	PF05192	MutS domain III	739	1064	IPR007696	DNA mismatch repair protein MutS; core
NP_000170.1_Homo_MSH6	Pfam	PF01624	MutS domain I	407	524	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_000170.1_Homo_MSH6	Pfam	PF05188	MutS domain II	538	699	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_000170.1_Homo_MSH6	SMART	SM00534		1127	1321	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_000170.1_Homo_MSH6	Pfam	PF00855	PWWP domain	90	183	IPR000313	PWWP domain

NP_000170.1_Homo_MSH6	SMART	SM00293		90	152	IPR000313	PWWP domain
CAA66337.1_Saccharomyces_MSH5	SMART	SM00533		274	609	IPR007696	DNA mismatch repair protein MutS; core
CAA66337.1_Saccharomyces_MSH5	SMART	SM00534		636	843	IPR000432	DNA mismatch repair protein MutS; C-terminal
CAA66337.1_Saccharomyces_MSH5	Pfam	PF05192	MutS domain III	252	576	IPR007696	DNA mismatch repair protein MutS; core
CAA66337.1_Saccharomyces_MSH5	Pfam	PF00488	MutS domain V	639	843	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAD04176.1_Arabidopsis_MSH2	SMART	SM00533		314	642	IPR007696	DNA mismatch repair protein MutS; core
AAD04176.1_Arabidopsis_MSH2	SMART	SM00534		659	855	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAD04176.1_Arabidopsis_MSH2	Pfam	PF05190	MutS family domain IV	468	564	IPR007861	DNA mismatch repair protein MutS; clamp
AAD04176.1_Arabidopsis_MSH2	Pfam	PF05188	MutS domain II	144	283	IPR007860	DNA mismatch repair protein MutS; connector domain
AAD04176.1_Arabidopsis_MSH2	Pfam	PF00488	MutS domain V	662	858	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAD04176.1_Arabidopsis_MSH2	Pfam	PF05192	MutS domain III	299	606	IPR007696	DNA mismatch repair protein MutS; core
AAD04176.1_Arabidopsis_MSH2	Pfam	PF01624	MutS domain I	23	128	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
EAL69456.1_Dictyostelium_MSH1	Pfam	PF05188	MutS domain II	172	291	IPR007860	DNA mismatch repair protein MutS; connector domain
EAL69456.1_Dictyostelium_MSH1	SMART	SM00534		693	880	IPR000432	DNA mismatch repair protein MutS; C-terminal
EAL69456.1_Dictyostelium_MSH1	SMART	SM00533		337	675	IPR007696	DNA mismatch repair protein MutS; core
EAL69456.1_Dictyostelium_MSH1	Pfam	PF05192	MutS domain III	323	624	IPR007696	DNA mismatch repair protein MutS; core
EAL69456.1_Dictyostelium_MSH1	Pfam	PF05190	MutS family domain IV	493	582	IPR007861	DNA mismatch repair protein MutS; clamp
EAL69456.1_Dictyostelium_MSH1	Pfam	PF00488	MutS domain V	696	883	IPR000432	DNA mismatch repair protein MutS; C-terminal
EAL69456.1_Dictyostelium_MSH1	Pfam	PF01624	MutS domain I	1	114	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
symbB.v1.2.033801.t1	Pfam	PF05192	MutS domain III	216	549	IPR007696	DNA mismatch repair protein MutS; core
symbB.v1.2.033801.t1	Pfam	PF00488	MutS domain V	608	805	IPR000432	DNA mismatch repair protein MutS; C-terminal
symbB.v1.2.033801.t1	SMART	SM00534		605	803	IPR000432	DNA mismatch repair protein MutS; C-terminal
symbB.v1.2.033801.t1	SMART	SM00533		239	583	IPR007696	DNA mismatch repair protein MutS; core
GSA25T00016421001	SMART	SM00534		703	893	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA25T00016421001	Pfam	PF05192	MutS domain III	250	593	IPR007696	DNA mismatch repair protein MutS; core
GSA25T00016421001	Pfam	PF00488	MutS domain V	707	895	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA25T00016421001	SMART	SM00533		264	683	IPR007696	DNA mismatch repair protein MutS; core
NP_193469.2_Arabidopsis_MSH4	Pfam	PF05192	MutS domain III	171	492	IPR007696	DNA mismatch repair protein MutS; core
NP_193469.2_Arabidopsis_MSH4	SMART	SM00534		546	733	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_193469.2_Arabidopsis_MSH4	Pfam	PF05190	MutS family domain IV	363	452	IPR007861	DNA mismatch repair protein MutS; clamp
NP_193469.2_Arabidopsis_MSH4	Pfam	PF00488	MutS domain V	550	735	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_193469.2_Arabidopsis_MSH4	SMART	SM00533		190	531	IPR007696	DNA mismatch repair protein MutS; core
BAA09235.1_Saccharomyces_MSH4	SMART	SM00534		627	813	IPR000432	DNA mismatch repair protein MutS; C-terminal
BAA09235.1_Saccharomyces_MSH4	Pfam	PF00488	MutS domain V	630	815	IPR000432	DNA mismatch repair protein MutS; C-terminal
BAA09235.1_Saccharomyces_MSH4	Pfam	PF05192	MutS domain III	270	580	IPR007696	DNA mismatch repair protein MutS; core
BAA09235.1_Saccharomyces_MSH4	Pfam	PF05190	MutS family domain IV	448	541	IPR007861	DNA mismatch repair protein MutS; clamp
BAA09235.1_Saccharomyces_MSH4	SMART	SM00533		284	611	IPR007696	DNA mismatch repair protein MutS; core
BAA09235.1_Saccharomyces_MSH4	Pfam	PF05188	MutS domain II	105	250	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_194284.2_Arabidopsis_MSH3	Pfam	PF05190	MutS family domain IV	632	713	IPR007861	DNA mismatch repair protein MutS; clamp
NP_194284.2_Arabidopsis_MSH3	SMART	SM00533		440	793	IPR007696	DNA mismatch repair protein MutS; core
NP_194284.2_Arabidopsis_MSH3	SMART	SM00534		810	1006	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_194284.2_Arabidopsis_MSH3	Pfam	PF05192	MutS domain III	425	758	IPR007696	DNA mismatch repair protein MutS; core
NP_194284.2_Arabidopsis_MSH3	Pfam	PF00488	MutS domain V	813	1009	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_194284.2_Arabidopsis_MSH3	Pfam	PF01624	MutS domain I	105	215	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_194284.2_Arabidopsis_MSH3	Pfam	PF05188	MutS domain II	258	324	IPR007860	DNA mismatch repair protein MutS; connector domain
EAS66875.1_Dictyostelium_MSH3	Pfam	PF01624	MutS domain I	454	566	IPR007695	DNA mismatch repair protein MutS-like; N-terminal

EAS66875.1_Dictyostelium_MSH3	Pfam	PF00488	MutS domain V	1175	1372	IPR000432	DNA mismatch repair protein MutS; C-terminal
EAS66875.1_Dictyostelium_MSH3	SMART	SM00533		805	1152	IPR007696	DNA mismatch repair protein MutS; core
EAS66875.1_Dictyostelium_MSH3	Pfam	PF05192	MutS domain III	790	1117	IPR007696	DNA mismatch repair protein MutS; core
EAS66875.1_Dictyostelium_MSH3	Pfam	PF05188	MutS domain II	614	727	IPR007860	DNA mismatch repair protein MutS; connector domain
EAS66875.1_Dictyostelium_MSH3	SMART	SM00534		1172	1370	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA120T00012118001_abc	Pfam	PF00488	MutS domain V	1054	1197	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA120T00012118001_abc	SMART	SM00534		1050	1288	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA120T00012118001_abc	Pfam	PF01624	MutS domain I	238	325	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
AAO49798.1_Arabidopsis_MSH1	Pfam	PF00488	MutS domain V	765	945	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAO49798.1_Arabidopsis_MSH1	SMART	SM00534		761	947	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAO49798.1_Arabidopsis_MSH1	Pfam	PF01624	MutS domain I	129	216	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
AAO49798.1_Arabidopsis_MSH1	Pfam	PF01541	GIY-YIG catalytic domain	1025	1059	IPR000305	GIY-YIG nuclease superfamily
XP_001021931.2_msh4_Tetrahymena	SMART	SM00534		1359	1539	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001021931.2_msh4_Tetrahymena	SMART	SM00533		958	1311	IPR007696	DNA mismatch repair protein MutS; core
XP_001021931.2_msh4_Tetrahymena	Pfam	PF00488	MutS domain V	1355	1426	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001021931.2_msh4_Tetrahymena	Pfam	PF05192	MutS domain III	944	1272	IPR007696	DNA mismatch repair protein MutS; core
MSH5_Vbra	SMART	SM00534		701	904	IPR000432	DNA mismatch repair protein MutS; C-terminal
MSH5_Vbra	SMART	SM00533		315	663	IPR007696	DNA mismatch repair protein MutS; core
MSH5_Vbra	Pfam	PF05192	MutS domain III	291	630	IPR007696	DNA mismatch repair protein MutS; core
MSH5_Vbra	Pfam	PF00488	MutS domain V	703	904	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_647085.1_Dictyostelium_MSH6	SMART	SM00534		1020	1209	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_647085.1_Dictyostelium_MSH6	Pfam	PF05190	MutS family domain IV	837	921	IPR007861	DNA mismatch repair protein MutS; clamp
XP_647085.1_Dictyostelium_MSH6	Pfam	PF00488	MutS domain V	1023	1212	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_647085.1_Dictyostelium_MSH6	Pfam	PF05192	MutS domain III	651	963	IPR007696	DNA mismatch repair protein MutS; core
XP_647085.1_Dictyostelium_MSH6	Pfam	PF01624	MutS domain I	362	480	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
XP_647085.1_Dictyostelium_MSH6	SMART	SM00533		664	1001	IPR007696	DNA mismatch repair protein MutS; core
XP_647085.1_Dictyostelium_MSH6	Pfam	PF05188	MutS domain II	490	623	IPR007860	DNA mismatch repair protein MutS; connector domain
XP_001025362.2_msh5_Tetrahymena	SMART	SM00534		676	846	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001025362.2_msh5_Tetrahymena	Pfam	PF00488	MutS domain V	680	787	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001025362.2_msh5_Tetrahymena	SMART	SM00533		220	638	IPR007696	DNA mismatch repair protein MutS; core
GSA120T00021256001	Pfam	PF00488	MutS domain V	669	881	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA120T00021256001	SMART	SM00533		257	645	IPR007696	DNA mismatch repair protein MutS; core
GSA120T00021256001	SMART	SM00534		665	879	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA120T00021256001	Pfam	PF05192	MutS domain III	242	591	IPR007696	DNA mismatch repair protein MutS; core
GSA25T00017553001	Pfam	PF00488	MutS domain V	162	342	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA25T00017553001	SMART	SM00534		158	343	IPR000432	DNA mismatch repair protein MutS; C-terminal
BAD95388.1_Arabidopsis_MSH5	Pfam	PF05192	MutS domain III	3	253	IPR007696	DNA mismatch repair protein MutS; core
BAD95388.1_Arabidopsis_MSH5	SMART	SM00533		2	285	IPR007696	DNA mismatch repair protein MutS; core
BAD95388.1_Arabidopsis_MSH5	Pfam	PF00488	MutS domain V	303	497	IPR000432	DNA mismatch repair protein MutS; C-terminal
BAD95388.1_Arabidopsis_MSH5	SMART	SM00534		300	495	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001457428.1_MSH5_Paramecium	SMART	SM00533		328	654	IPR007696	DNA mismatch repair protein MutS; core
XP_001457428.1_MSH5_Paramecium	SMART	SM00534		671	862	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001457428.1_MSH5_Paramecium	Pfam	PF00488	MutS domain V	675	857	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_014551.1_Saccharomyces_MSH2	Pfam	PF05190	MutS family domain IV	491	586	IPR007861	DNA mismatch repair protein MutS; clamp
NP_014551.1_Saccharomyces_MSH2	SMART	SM00533		333	664	IPR007696	DNA mismatch repair protein MutS; core
NP_014551.1_Saccharomyces_MSH2	Pfam	PF00488	MutS domain V	684	879	IPR000432	DNA mismatch repair protein MutS; C-terminal

NP_014551.1_Saccharomyces_MSH2	Pfam	PF01624	MutS domain I	19	112	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_014551.1_Saccharomyces_MSH2	SMART	SM00534		681	877	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_014551.1_Saccharomyces_MSH2	Pfam	PF05188	MutS domain II	141	285	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_014551.1_Saccharomyces_MSH2	Pfam	PF05192	MutS domain III	301	627	IPR007696	DNA mismatch repair protein MutS; core
ESS64221.1_MSH5_Trypanosoma	SMART	SM00534		541	751	IPR000432	DNA mismatch repair protein MutS; C-terminal
ESS64221.1_MSH5_Trypanosoma	Pfam	PF00488	MutS domain V	545	689	IPR000432	DNA mismatch repair protein MutS; C-terminal
ESS64221.1_MSH5_Trypanosoma	Pfam	PF05192	MutS domain III	177	491	IPR007696	DNA mismatch repair protein MutS; core
ESS64221.1_MSH5_Trypanosoma	SMART	SM00533		200	524	IPR007696	DNA mismatch repair protein MutS; core
EJY69573.1_Msh4_Oxytricha	Pfam	PF00488	MutS domain V	632	792	IPR000432	DNA mismatch repair protein MutS; C-terminal
EJY69573.1_Msh4_Oxytricha	Pfam	PF05192	MutS domain III	227	565	IPR007696	DNA mismatch repair protein MutS; core
EJY69573.1_Msh4_Oxytricha	Pfam	PF05190	MutS family domain IV	444	522	IPR007861	DNA mismatch repair protein MutS; clamp
EJY69573.1_Msh4_Oxytricha	SMART	SM00533		243	610	IPR007696	DNA mismatch repair protein MutS; core
EJY69573.1_Msh4_Oxytricha	SMART	SM00534		628	844	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAO86765.1_Plasmodium_MSH2	Pfam	PF00488	MutS domain V	620	808	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAO86765.1_Plasmodium_MSH2	Pfam	PF05192	MutS domain III	169	561	IPR007696	DNA mismatch repair protein MutS; core
AAO86765.1_Plasmodium_MSH2	SMART	SM00534		616	805	IPR000432	DNA mismatch repair protein MutS; C-terminal
AAO86765.1_Plasmodium_MSH2	SMART	SM00533		197	597	IPR007696	DNA mismatch repair protein MutS; core
XP_643399.1_Dictyostelium_MSH2	Pfam	PF05192	MutS domain III	355	642	IPR007696	DNA mismatch repair protein MutS; core
XP_643399.1_Dictyostelium_MSH2	SMART	SM00534		700	887	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_643399.1_Dictyostelium_MSH2	Pfam	PF05188	MutS domain II	170	301	IPR007860	DNA mismatch repair protein MutS; connector domain
XP_643399.1_Dictyostelium_MSH2	SMART	SM00533		355	683	IPR007696	DNA mismatch repair protein MutS; core
XP_643399.1_Dictyostelium_MSH2	Pfam	PF05190	MutS family domain IV	509	599	IPR007861	DNA mismatch repair protein MutS; clamp
XP_643399.1_Dictyostelium_MSH2	Pfam	PF00488	MutS domain V	703	890	IPR000432	DNA mismatch repair protein MutS; C-terminal
symbB.v1.2.013503.t1	Pfam	PF00488	MutS domain V	61	263	IPR000432	DNA mismatch repair protein MutS; C-terminal
symbB.v1.2.013503.t1	SMART	SM00534		58	263	IPR000432	DNA mismatch repair protein MutS; C-terminal
EJY77525.1_MSH5_Oxytricha	SMART	SM00534		662	868	IPR000432	DNA mismatch repair protein MutS; C-terminal
EJY77525.1_MSH5_Oxytricha	Pfam	PF05192	MutS domain III	270	614	IPR007696	DNA mismatch repair protein MutS; core
EJY77525.1_MSH5_Oxytricha	SMART	SM00533		289	648	IPR007696	DNA mismatch repair protein MutS; core
EJY77525.1_MSH5_Oxytricha	Pfam	PF00488	MutS domain V	666	774	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA25T00025231001_abc	Pfam	PF00488	MutS domain V	530	680	IPR000432	DNA mismatch repair protein MutS; C-terminal
GSA25T00025231001_abc	SMART	SM00534		526	764	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_010382.3_Saccharomyces_MSH6	Pfam	PF01624	MutS domain I	312	426	IPR007695	DNA mismatch repair protein MutS-like; N-terminal
NP_010382.3_Saccharomyces_MSH6	Pfam	PF05188	MutS domain II	463	596	IPR007860	DNA mismatch repair protein MutS; connector domain
NP_010382.3_Saccharomyces_MSH6	Pfam	PF05190	MutS family domain IV	780	870	IPR007861	DNA mismatch repair protein MutS; clamp
NP_010382.3_Saccharomyces_MSH6	SMART	SM00533		634	956	IPR007696	DNA mismatch repair protein MutS; core
NP_010382.3_Saccharomyces_MSH6	Pfam	PF00488	MutS domain V	979	1167	IPR000432	DNA mismatch repair protein MutS; C-terminal
NP_010382.3_Saccharomyces_MSH6	Pfam	PF05192	MutS domain III	620	910	IPR007696	DNA mismatch repair protein MutS; core
NP_010382.3_Saccharomyces_MSH6	SMART	SM00534		975	1164	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001436560.1_MSH4_Paramecium	Pfam	PF00488	MutS domain V	544	663	IPR000432	DNA mismatch repair protein MutS; C-terminal
XP_001441588.1_MSH4_Paramecium	Pfam	PF00488	MutS domain V	544	663	IPR000432	DNA mismatch repair protein MutS; C-terminal
REC8+RAD21							
NP_197131.2_Arabidopsis_Rad21	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	973	1026	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
NP_197131.2_Arabidopsis_Rad21	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	102	IPR006910	Rad21/Rec8-like protein; N-terminal
A2AU37_RD21L_MOUSE	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	498	549	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
A2AU37_RD21L_MOUSE	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	100	IPR006910	Rad21/Rec8-like protein; N-terminal
EAA16145.1_Plasmodium	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	46	137	IPR006910	Rad21/Rec8-like protein; N-terminal

O95072_REC8_HUMAN	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	112	IPR006910	Rad21/Rec8-like protein; N-terminal
O95072_REC8_HUMAN	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	493	546	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
P30776_RAD21_SCHPO	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	98	IPR006910	Rad21/Rec8-like protein; N-terminal
P30776_RAD21_SCHPO	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	568	623	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q6AYJ4_REC8_RAT	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	539	592	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q6AYJ4_REC8_RAT	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	112	IPR006910	Rad21/Rec8-like protein; N-terminal
NP_851110.1_Arabidopsis	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	752	804	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
NP_851110.1_Arabidopsis	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	93	IPR006910	Rad21/Rec8-like protein; N-terminal
Q12188_REC8_YEAST	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	16	114	IPR006910	Rad21/Rec8-like protein; N-terminal
Q9H4I0_RD21L_HUMAN	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	100	IPR006910	Rad21/Rec8-like protein; N-terminal
Q9H4I0_RD21L_HUMAN	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	502	553	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
GSA120T00009901001	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	989	1025	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
P36626_REC8_SCHPO	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	504	556	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
P36626_REC8_SCHPO	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	113	IPR006910	Rad21/Rec8-like protein; N-terminal
Q61550_RAD21_MOUSE	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	103	IPR006910	Rad21/Rec8-like protein; N-terminal
Q61550_RAD21_MOUSE	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	579	632	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
AAD39601.1_Oryza_Rec8	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	26	96	IPR006910	Rad21/Rec8-like protein; N-terminal
O60216_RAD21_HUMAN	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	575	628	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
O60216_RAD21_HUMAN	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	103	IPR006910	Rad21/Rec8-like protein; N-terminal
GSA120T00017234001	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	2029	2067	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
AAQ21081.1_Oryza	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	1001	1050	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
AAQ21081.1_Oryza	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	102	IPR006910	Rad21/Rec8-like protein; N-terminal
GSA120T00017233001	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	120	204	IPR006910	Rad21/Rec8-like protein; N-terminal
Q12158_SCC1_YEAST	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	506	559	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q12158_SCC1_YEAST	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	15	110	IPR006910	Rad21/Rec8-like protein; N-terminal
GSA25T00010409001	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	723	760	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
GSA25T00008231001	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	1230	1269	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q19325_SCC1_CAEEL	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	561	615	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q19325_SCC1_CAEEL	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	104	IPR006910	Rad21/Rec8-like protein; N-terminal
O93310_RAD21_XENLA	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	573	626	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
O93310_RAD21_XENLA	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	103	IPR006910	Rad21/Rec8-like protein; N-terminal
NP_196168.1_Arabidopsis_Rec8	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	104	IPR006910	Rad21/Rec8-like protein; N-terminal
NP_196168.1_Arabidopsis_Rec8	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	565	617	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q9XUB3_REC8_CAEEL	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	116	IPR006910	Rad21/Rec8-like protein; N-terminal
Q8C5S7_REC8_MOUSE	Pfam	PF04824	Conserved region of Rad21 / Rec8 like protein	537	590	IPR006909	Rad21/Rec8-like protein; C-terminal; eukaryotic
Q8C5S7_REC8_MOUSE	Pfam	PF04825	N terminus of Rad21 / Rec8 like protein	1	112	IPR006910	Rad21/Rec8-like protein; N-terminal

SPO11

Q9M4A2_SPO11_Arabidopsis	PANTHER	PTHR10848:SF6		6	357		
Q9M4A2_SPO11_Arabidopsis	Pfam	PF04406	Type IIB DNA topoisomerase	75	135	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Q9M4A2_SPO11_Arabidopsis	PANTHER	PTHR10848		6	357	IPR002815	Spo11/DNA topoisomerase VI subunit A
EJY83114.1_Oxytricha	PANTHER	PTHR10848		23	287	IPR002815	Spo11/DNA topoisomerase VI subunit A
W7EYC7_Plasmodium	PANTHER	PTHR10848		25	207	IPR002815	Spo11/DNA topoisomerase VI subunit A
W7EYC7_Plasmodium	PANTHER	PTHR10848:SF7		25	207		
P23179_SPO11_Saccharomyces	Pfam	PF04406	Type IIB DNA topoisomerase	106	168	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
P23179_SPO11_Saccharomyces	PANTHER	PTHR10848:SF7		16	391		
P23179_SPO11_Saccharomyces	PANTHER	PTHR10848		16	391	IPR002815	Spo11/DNA topoisomerase VI subunit A

NP_011841.1_Saccharomyces	Pfam	PF04406	Type IIB DNA topoisomerase	106	168	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
NP_011841.1_Saccharomyces	PANTHER	PTHR10848:SF7		16	391		
NP_011841.1_Saccharomyces	PANTHER	PTHR10848		16	391	IPR002815	Spo11/DNA topoisomerase VI subunit A
XP_001440113.1_Paramecium	Pfam	PF09983	Uncharacterized protein conserved in bacteria C-term(DUF2220)	143	284	IPR024534	Domain of unknown function DUF2220
XP_001440113.1_Paramecium	PANTHER	PTHR10848		42	310	IPR002815	Spo11/DNA topoisomerase VI subunit A
XP_001440113.1_Paramecium	Pfam	PF04406	Type IIB DNA topoisomerase	54	114	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
GSA25T00025387001	Pfam	PF04406	Type IIB DNA topoisomerase	982	1048	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
GSA25T00025387001	PANTHER	PTHR10848		962	1289	IPR002815	Spo11/DNA topoisomerase VI subunit A
GSA120T00013268001	PANTHER	PTHR10848		1234	1589	IPR002815	Spo11/DNA topoisomerase VI subunit A
GSA120T00013268001	Pfam	PF04406	Type IIB DNA topoisomerase	1276	1342	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
A0A0F7UVU3_Toxoplasma	PANTHER	PTHR10848		42	339	IPR002815	Spo11/DNA topoisomerase VI subunit A
A0A0F7UVU3_Toxoplasma	Pfam	PF04406	Type IIB DNA topoisomerase	63	124	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Vbra_16613_SPO11	PANTHER	PTHR10848		296	340	IPR002815	Spo11/DNA topoisomerase VI subunit A
Vbra_16613_SPO11	Pfam	PF04406	Type IIB DNA topoisomerase	164	199	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Vbra_16613_SPO11	Pfam	PF04406	Type IIB DNA topoisomerase	80	141	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
A0A0G4G0S1_Vitrella	PANTHER	PTHR10848		296	340	IPR002815	Spo11/DNA topoisomerase VI subunit A
A0A0G4G0S1_Vitrella	Pfam	PF04406	Type IIB DNA topoisomerase	164	199	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
A0A0G4G0S1_Vitrella	Pfam	PF04406	Type IIB DNA topoisomerase	80	141	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
A0A1D3UA32_Plasmodium	PANTHER	PTHR10848:SF7		28	318		
A0A1D3UA32_Plasmodium	PANTHER	PTHR10848		28	318	IPR002815	Spo11/DNA topoisomerase VI subunit A
A0A1D3UA32_Plasmodium	Pfam	PF04406	Type IIB DNA topoisomerase	40	98	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
CA856344.1_Plasmodium	PANTHER	PTHR10848		8	189	IPR002815	Spo11/DNA topoisomerase VI subunit A
Q8I5N7_Plasmodium	Pfam	PF04406	Type IIB DNA topoisomerase	38	97	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Q8I5N7_Plasmodium	PANTHER	PTHR10848		31	320	IPR002815	Spo11/DNA topoisomerase VI subunit A
Q8I5N7_Plasmodium	PANTHER	PTHR10848:SF7		31	320		
symbB.v1.2.038121.t1	Pfam	PF04406	Type IIB DNA topoisomerase	115	173	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
symbB.v1.2.038121.t1	PANTHER	PTHR10848		102	401	IPR002815	Spo11/DNA topoisomerase VI subunit A
AAD52562.1_Homo_SPO11	PANTHER	PTHR10848		43	393	IPR002815	Spo11/DNA topoisomerase VI subunit A
AAD52562.1_Homo_SPO11	PANTHER	PTHR10848:SF7		43	393		
AAD52562.1_Homo_SPO11	Pfam	PF03533	SPO11 homologue	2	44	IPR004084	Meiosis-specific protein Spo11
AAD52562.1_Homo_SPO11	Pfam	PF04406	Type IIB DNA topoisomerase	109	170	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Q23RZ0_Tetrahymena	PANTHER	PTHR10848		27	435	IPR002815	Spo11/DNA topoisomerase VI subunit A
Q23RZ0_Tetrahymena	Pfam	PF04406	Type IIB DNA topoisomerase	127	187	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Q23RZ0_Tetrahymena	PANTHER	PTHR10848:SF7		27	435		
symbB.v1.2.012520.t1	PANTHER	PTHR10848		2	237	IPR002815	Spo11/DNA topoisomerase VI subunit A
S8GTC7_Toxoplasma	PANTHER	PTHR10848		41	341	IPR002815	Spo11/DNA topoisomerase VI subunit A
S8GTC7_Toxoplasma	Pfam	PF04406	Type IIB DNA topoisomerase	63	124	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
S8GTC7_Toxoplasma	Pfam	PF09664	Protein of unknown function C-terminus (DUF2399)	169	237	IPR024465	Domain of unknown function DUF2399
A0A1C3L2R0_Plasmodium	Pfam	PF04406	Type IIB DNA topoisomerase	41	99	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
A0A1C3L2R0_Plasmodium	PANTHER	PTHR10848:SF7		27	320		
A0A1C3L2R0_Plasmodium	PANTHER	PTHR10848		27	320	IPR002815	Spo11/DNA topoisomerase VI subunit A
Q7Y021_SPO11_Oryza	PANTHER	PTHR10848		22	377	IPR002815	Spo11/DNA topoisomerase VI subunit A
Q7Y021_SPO11_Oryza	Pfam	PF04406	Type IIB DNA topoisomerase	95	155	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
Q7Y021_SPO11_Oryza	PANTHER	PTHR10848:SF6		22	377		
C6S3D8_Plasmodium	Pfam	PF04406	Type IIB DNA topoisomerase	55	114	IPR013049	Spo11/DNA topoisomerase VI; subunit A; N-terminal
C6S3D8_Plasmodium	PANTHER	PTHR10848		50	333	IPR002815	Spo11/DNA topoisomerase VI subunit A

General discussion and perspective

The main task of this thesis was to characterize the inter- and intra-clade diversity in *Amoebophrya*. We addressed the difference in morphology, ecology and genome characteristics between strains (for single cells, the morphology and ecology data were not available). The genetic marker, 18S-ITS, separated these strains/single cells into eight groups (referring to eight subclades in the following text), which correspond to eight ribotypes. In the first chapter, we revealed that strains with the same ribotypes displayed similar host spectrum. Cell morphology (FSC and SSC) only distinguishes one subclade from all others. Likewise, ecological niches serve to discriminate one certain ribotype from all others. Genome sizes and short-fragment composition are conserved within a cluster although similar patterns may be shared between subclades. However, the finding of CBC in the ITS2 region differentiates all ribotypes and suggests each subclade is very likely a cryptic species. In the second chapter, we showed that 4 genes known to specifically participate in meiosis are present in two genomes belonging to two of the previously investigated ribotypes. Taken together, our results suggest *Amoebophrya* is likely a sexual lineage with high species diversity.

A polyphasic approach to delimiting species

In the first study (chapter 1), we performed a comprehensive comparison of characters between strains with respect to morphology, ecology and genome characteristics, which correspond to different species concepts respectively (morphological species concept, ecological species concept, phylogenetic species concept, and biological species concept (realized by CBC test)). These species concepts work as species criteria to recognize and delimit species by determining the boundaries between species units (biological units) (Zhao et al., 2018; De Queiroz, 2007; Giraud et al., 2008).

The morphological species concept uses discontinuities in morphological variation to distinguish species (Leliaert et al., 2014). Traditionally, researchers have most often relied on morphology to define a new species. When the species' characteristics are distinct and easy to observed, morphology-based method is feasible. The genus *Amoebophrya* was at first described as encompassing 7 morphospecies based on microscopic observation. According to the description of Cachon (1964), these species share most morphological characteristics and are mainly distinguished from other members of the genus by the structure of its trophont, the pattern of sporogenesis, and spore morphology. The high similarity in morphologies and very small cell sizes of dinospores (3-5 μm in general) make it difficult to carry out these standards in practice.

Flow cytometry was explored in this study to discriminate subclades. It turned out only 1 subclade (RIB2) can be differentiated from others based on FSC and SSC signatures, which provides support for the conclusion that striking and consistent differences in morphology are meaningful for delimiting species (Coats et al., 2012). However, the morphology-based methods have their limits in the use for distinguishing such microorganisms. In the lab, we frequently observed that even for the same strain, the distribution pattern based on FSC and SSC might shift from day to day. This is consistent with

previous observation that environmental conditions during laboratory incubations have a strong influence on the morphology of dinokont spores produced by parasitic dinoflagellates (Cachon, 1964). Thus taxonomic inferences based on subtle variation in spore morphology must be approached with caution.

The ecological species concept emphasizes adaptation to a particular ecological niche. We first tested the host spectrum of parasitic strains available in the lab. By cross-infection experiments, we proved that the host spectrum of strains is conserved within subclades, but not a character distinguishable between subclades. Furthermore, ecological niches were determined from meta-barcoding data, by which RIB2 and RIB4 strains are separated from others. Although the type and number of variables comprising the dimensions of an environmental niche vary from one species to another, the ecological niches based method is not highly reliable/efficient for delimiting species, considering the relative importance of particular environmental variables for a species may vary according to the geographic and biotic contexts (Peterson et al., 2011).

The phylogenetic species concept emphasizes nucleotide divergence and regards species as descent that can be identified based on reciprocal monophyly (Mayden, 1997). The analysis of sequence similarity in the 18S-ITS region showed that each subclade possesses one representative haplotype. As a result, 8 unique ribotypes are present in all strains and single cells. Comparison between ribotypes showed ITS is more variable than 18S region and has the potential to be a good species indicator. Moreover, the clustering pattern from the genome scale k-mer analysis is generally consistent with that from 18S-ITS phylogeny, which, to our knowledge, for the first time provides the strong evidence that the evolution of a short region of the genome (e.g. the 18S-ITS) can be indicative of that of the whole genome arising from speciation events. However, one should be wary of the use of 18S-ITS in delimiting species. This is because the extent to which ribosomal DNA varies differs from one lineage to another.

The biological species concept emphasizes reproductive isolation and is the golden standard for species delimitation for eukaryotes. Due to the small cell size (the dinospore is generally 3-5 um in size) and complex life cycle (endoparasites) of *Amoebophrya*, the sexual reproduction in this parasite remains unknown. This is a common question for microorganisms and even more complicated by a parasitic lifestyle. We circumvented this question by counting CBC numbers in ITS2 region, which indicates 2 different species with a probability of 0.93 when the number is not less than one (Müller et al., 2007). The presence of CBC (ranging from 1 to 9) demonstrated reproduction isolation (if it exists in *Amoebophrya*) occurs between different ribotypes and thus all subclades are very likely different species. And the CBC numbers between subclades generally reflect the genetic distances between different ribotypes when taking into account the drastic distinction in ITS2 secondary structures of some ribotypes from others (e.g. RIB1, RIB4 and RIB6 have 3 loops while the others 4 loops in their ITS2).

Taken together, all the above-mentioned species criteria correspond to different events of processes that occur during lineage separation and divergence and deal with the issue of speciation in different scales (**Fig** 1). An integration of species criterion would provide the most effective discriminatory tool although it may take more time and efforts. In this study, a polyphasic approach combining all species criteria used in this study together and proved that each subclade is very likely a species. We believe this method will continue to serve in the future to describe and delimit protistan species which are morphologically similar or the same. The value of a polyphasic approach is that it permits a comprehensive description of a species and allows researchers to gain a broader view of the living organisms so that one is able to draw conclusions more confidently. It seems the more methods are used, the less uncertainty we have. However, in this study, none of these criteria but CBC approach, solely determine each subclade corresponds to one species. On the other hand, there is obviously a cost incurred with the use of more methods – should we allocate cost to obtain more information on morphology, or simply rely on more methods? Whatever method is chosen, cautions should be paid that these strategies need to be worked out rigorously.

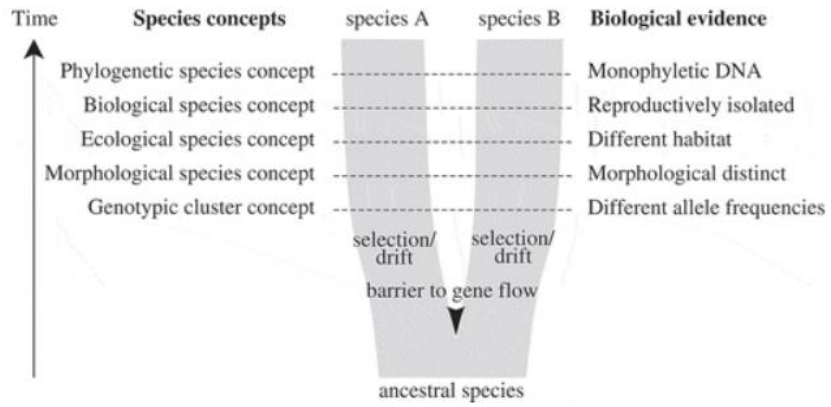


Fig 1. Simplified diagram of speciation, species concepts and corresponding biological properties of species. As populations separate by a barrier to gene flow, selection and drift will result in two daughter lineages with separate evolutionary trajectories. Through time, these daughter lineages will acquire different properties, which have traditionally served as biological evidence for species delimitation, corresponding to different species concepts. During the process of speciation, these secondary properties do not necessarily arise at the same time or in a regular order, and therefore different species concepts may come into conflict, especially during early stages of speciation. (Leliaert et al., 2014)

Use of V4/V9 in environmental investigations

Metabarcoding is a powerful tool for exploring microbial diversity in the environment. 18S sequences have been widely used in environmental studies. A small segment of 18S sequences, e.g. V4 or V9, has been used as DNA markers to investigate protistan diversity in a global scale (e.g. de Vargas et al., 2015; Le Bescot et al., 2016; Massana et al., 2015). But this has been done mostly at the OTU (operational taxonomic units) or swarm level, but not at the real "species level".

In this study, we distinguished 8 putative species, some of which shared very high level of similarities in their 18S, 18S-V4 and 18S-V9 regions, respectively. For example, the similarity between subclade 3 and 8 reached 99.77% (4 point mutations) in their whole 18S, 100% (no point mutation) in 18S-V4 region and 99.24% (1 point mutation) in the 18S-V9 region, respectively. Clearly, the use of the 97% cut-off in either of these regions is not able to distinguish these 2 putative species. Interestingly, we found that even with a sequence identity above 99% in 18S rDNA, V9 can still show variations that indicate speciation while V4 cannot although V9 (131 bp) is shorter than V4 region (381bp). This could be an occasion resulting from small sample sizes. However, Pernice et al. (2013) took advantage of a large number of sequences from clone libraries and demonstrated that the V4 region (and part of V5) represented the variability of the complete 18S rDNA better than the V9 region. For alveolates (**Fig** 2B), the V4-V5 distance correlates with the 18S full gene distance in a linear relationship starting from the origin or near the origin, which means the V4 region will gain more similarity (or lose more variations) when the full length 18S rDNA is getting more similar. In contrast, the linear equation which depicts how V9 distances vary with the full gene distances has an intercept of a positive value. So the V9 region still shows variations even in the case where the full length 18S rDNA gene reveals nearly 100% percent identity. This explains well why we saw variations in V9 but no variation in the V4 region in our case (RIB3 vs. RIB8). Therefore, V9 has a better resolution than V4 region when the sequence similarity is extremely high or the sequence distance is extremely low. This is generally true even for the other SAR lineages, such as Stramenopiles (**Fig** 2A) and Rhizaria (**Fig 2**C). Moreover, it's worth noting that the number of nucleotide differences directly indicates speciation while the genetic distance (in percentage) between sequences barely tells the differences when the sequence similarity is very high (e.g. around 99%).

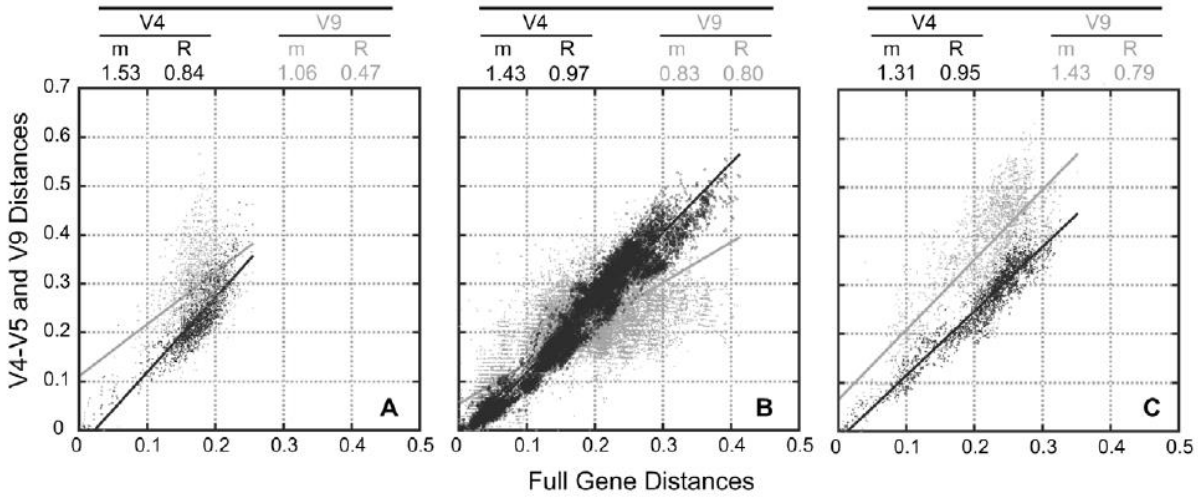


Fig 2. Comparison of partial and full-length 18S rDNA sequences to infer genetic distances. The three panels show pairwise genetic distances (Jukes Cantor corrected) of the complete gene against partial regions (V4–V5 in dark grey or V9 in light grey) for sequences within Stramenopiles (A), Alveolata (B), and Rhizaria (C). Slopes (m) and coefficients (R) of the correlations are shown at the top of the graphs. From Pernice et al., (2013).

Generally, unicellular lineages have especially low 18S divergence relative to their protein sequence divergence, suggesting that 18S ribosomal genes are too conservative to assess planktonic eukaryotic diversity (Piganeau et al., 2011). But the recognition of many novel planktonic organisms is based solely on their 18S rDNA sequence at present. A species delimited by its 18S rDNA sequence, or simply by its V4 or V9 region, might contain many cryptic species. Therefore, there is a trade-off between using genes that are easy to amplify in all species, but underestimate the true number of species, and using genes that are more difficult to amplify, but give a better indication of species numbers.

Highly underestimated species richness in Syndiniales

To date, Amoebophryidae is composed of only one genus, *Amoebophrya*, in which 7 morphospecies are included according to Cachon (1964). *Amoebophrya ceratii*, which is able to infect a wide range of dinoflagellates, is now believed to be a species complex (reviewed in Park et al., 2013). In this study, 8 putative species belonging to *Amoebophrya ceratii* have been identified. Noteworthy, the subclade 3 and subclade 8 representing 2 putative species differ by <1% in their 18S rDNA. This separation at the species level was reported likewise in between *Duboscquodinium collini* and *Scrippsiella trochoidea* (0.2% in 18S rDNA; Coats et al., 2010). Both studies suggested that the extremely high identity in SSU sequences may hide genetic distinction in dinoflagellates. This is further expected to be a common phenomenon for microeukaryotes, given that unicellular lineages have especially low 18S divergences relative to their protein sequence divergences in general (Piganeau et al., 2011).

Although 18S rDNA is too conservative to assess protistan diversity (Piganeau et al., 2011), the current available information about most of investigated uncultured microeukaryotes is from metadata owing to the powerful next-generation sequencing technique. In MALASPINA (18S-V4) and TARA (18S-V9) datasets, which both focused on the biodiversity in the global oceans, the diversity of some groups in marine protists is uncovered unprecedentedly high (e.g. MALVs, **Table** 1). Based on the fact that one nucleotide difference in V9 region could differentiate species (subclade 3 and subclade8 in our study), the unique sequence counts may represent the number of real species detected in the datasets if the sequence diversity is real (i.e. not from PCR or sequencing steps). The observation from the OSD (Ocean Sampling Day) dataset, for which both the V4 and V9 sequences are available, is in agreement with our prediction that V4 region is less indicative of speciation than V9. Thus, there are less V4 unique sequences than V9 unique sequences observed in this dataset. Furthermore, the observation that the unique sequences from MALASPINA dataset are much less than that from TARA dataset could be partly explained by the same reason, apart from that MALASPINA dataset (285 samples) is smaller than TARA dataset (334 samples) and covers smaller areas in the global scale.

Table 1. The counts of unique sequences assigned to different lineages from TARA, MALASPINA and OSD datasets. To be conservative, the values are the counts of sequences that are present in more than one sample. The numbers in parenthesis are the counts of sequences without this filtration step.

	TARA (V9)	MALASPINA (V4)	OSD(V4)	OSD(V9)
Eukaryote	2049725	-	2641(8922)	5615(18681)
Alveolata	667755	9652(17195)	1002(3197)	1603(4786)
Syndiniales	225549	-	293(840)	485(997)
MALV-I	107044	1352(2314)	58(136)	90(145)
MALV-II	95145	5204(8671)	215(658)	341(746)
MALV-III	4338	277(477)	17(40)	8(11)
MALV-IV	3364	142(257)	3(5)	9(16)
MALV-V	534	29(47)	0(1)	3(7)
Amoebophrya	6001	4927(8199)	0(0)	2(2)
Apicomplexa	26099	3(15)	31(249)	129(632)

Ps: '-' indicates the absence of data. TARA data shown here are taken from metabarcodes published in de Vargas et al (2015), in which every metabarcode was observed in two distinct samples (one confirming the other) and present in at least three copies among all samples. MALASPINA data and OSD data shown here were both processed with DADA2, which produced an amplicon sequence variant (ASV) table recording the number of times each exact amplicon sequence variant was observed in each sample.

The MALV-II lineage is found to be Syndiniales (Guillou et al 2008) and frequently detected with high abundance in marine environmental investigations (e.g. de Vargas et al., 2015; Massana et al., 2015). The accurate interpretation of metadata, especially on 18S rDNA, is impeded by varieties of technical (e.g. PCR and sequencing errors) and biological biases (e.g. intra-individual polymorphism) (reviewed in Decelle et al., 2014). Traditionally, metabarcoding based approaches assigned sequences into OTUs. A cut-off value at 0.97 has been frequently used in environmental investigations (Edgar, 2018; Hao et al., 2011). Although this has been proven to be conservative to delimit species, following the same way, MALV-II got 836 OTUs out of 5204 sequences from the MALASPINA V4 dataset (**Fig** 3). This is a conservative estimation excluding all possible local sequence/species which only occurred in one single sampling site.

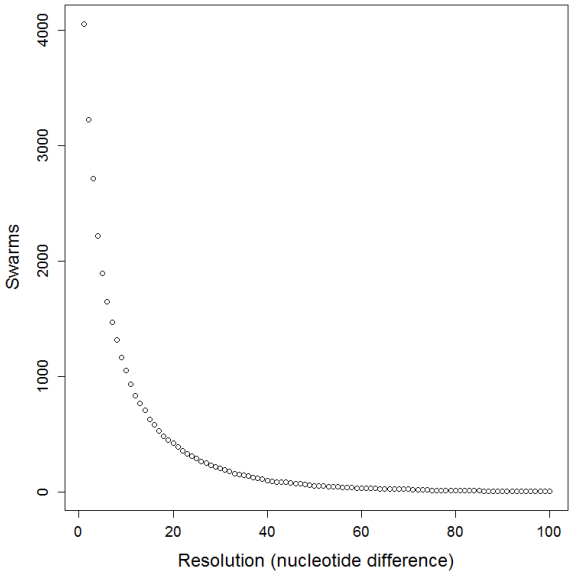


Fig 3. OTU (or swarm) counts calculated from all unique sequences (ASVs obtained by DADA2) assigned to MALV-II group in MALASPINA V4 dataset. The ASVs were clustered into OTUs using the 'Swarm' approach (Mahé et al. 2014).

In contrast, the most recent and compelling investigations on the classification of the extant dinoflagellates reported dinokaryotic dinoflagellates comprised 2,294 species belonging to 238 genera. Dinoflagellates *sensu lato* (Ellobiopsea, Oxyrrhea, Syndinea and Dinokaryota) comprised 2,377 species belonging to 259 genera, in which MALV I and MALV II together occupied only 11 genera and 46 species (Gómez, 2012). Our study proved that the species richness in Syndiniales at the global scale is far more than documented ever (over 100 times if the 5204 sequences assigned to MALV-II represent the real species). This may be true for the phylum Dinoflagellata and even for the supergroup Alveolata.

Pernice et al (2013) observed the maximal corrected distances of 18S rDNA sequences in both MALV-I and MALV-II are high and inferred these could represent higher taxonomic ranks. In our study, the maximal nucleotide difference between subclades (representing putative species) in the V4 region is 29 with a mean value of 15 while in MALASPINA V4 dataset, this value is 161 with a mean value of 57 (**Fig** 4), which corroborates higher taxonomic ranks exist in MALV-II. A large number of OTUs detected for MALV-II at 0.05 (421 OTUs) or 0.10 (101 OTUs) clustering distance also suggests the presence of many high-rank lineages in MALASPINA dataset. Last but not the least, the difference in the secondary structures of ITS2 from our study (3 loops for RIB1, 4, 6 contrast with 4 loops for RIB2) provides more evidence for high-rank variations in *Amoebophrya*, as observed in green algae and flowering plants (Mai and Coleman, 1997).

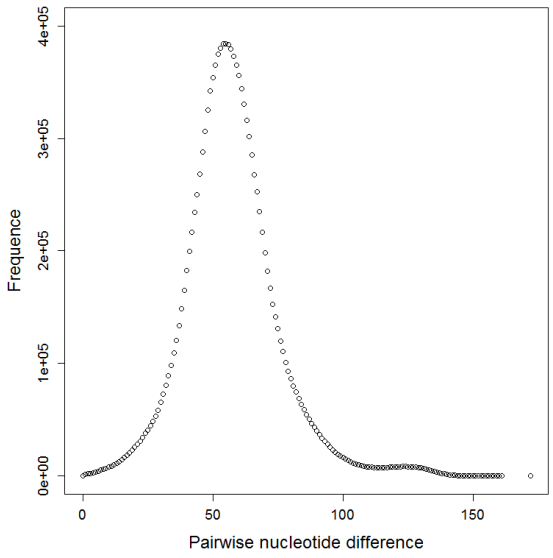


Fig 4. Frequence of pairwise differences between all unique sequences (ASVs) assigned to MALV-II group in MALASPINA V4 dataset.

A genomic approach for the discovery of genetic diversity in protists

It has been suggested that cryptic species complexes are very common in the marine environment (Knowlton, 1993). Members of a species complex share similar morphology, behavior and habitat and thus are difficult to identify morphologically. Their complex life-cycles also make identification of species challenging but important. *Amoebophrya ceratii*, for example, has been proven to be a cryptic species (reviewed in Park et al., 2013). Previous molecular studies, however, simply rely on a short sequence (e.g. 18S), based on which ecological relevance is inferred (e.g. Lima-mendez et al., 2015). The true role/identity of this complex's members is barely investigated yet. Culture-based method allowed us to study this cryptic species more closely and comprehensively. On this base, a genomic approach further contributed to the discovery of high genetic diversity hidden in this cryptic species.

Two strains (A25 and A120) belonging to two different subclades (subclade 1 and 4 respectively) have been fully sequenced recently and annotated, which showed drastic differentiation in their protein-coding sequences (Farhat et al., in preparation). Only 9,490 orthologs were identified between them, which represented less than 36% of the total number of predicted proteins that shared 48.2% sequence identity on average. The level of sequence identity between the two *Amoebophrya* proteomes was similar to that observed between any one of the sequenced *Amoebophrya* strains and *Symbiodinium* spp., *Perkinsus marinus*, or *Plasmodium falciparum.* Despite the low interspecific protein sequence similarity, a high level of synteny between the two genomes was observed. Most strikingly, both genomes possessed a high proportion (66-67%) of non-canonical introns (NCIs) which are unique among eukaryotes.

Further sequencing on more genomes of strains and single cells revealed that the differences between subclades are remarkable in terms of genomics as demonstrated by SNPs (**Fig** 5A) and remapping rates of whole genome reads (**Fig** 6). Despite the conservation in the 18S rDNA, the perfect consistency between the topology of 18S tree and that of whole genome tree has been observed (**Fig** 5B and 5C). More intriguingly, repetitive sequences were found being shared within subclades but not between subclades (data not shown). This is evidence that repetitive sequences, which take up a large portion of a genome (e.g. 23.8% in A120 and 13.1% in A25), are lineage-specific and could be an indicator for species delimitation.

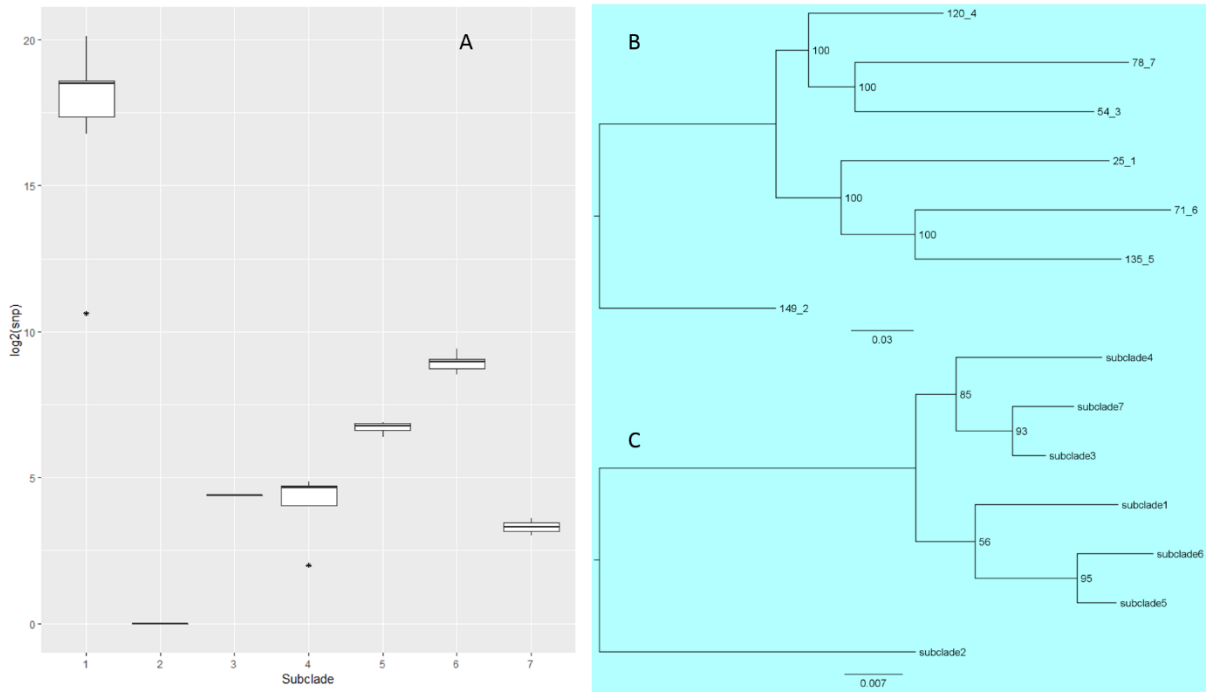


Fig 5. (A) Boxplot of SNP counts from strains belonging to different subclades against the annotated genes from the strain A25 (subclade 1). Stars in the figure indicate outliers in the dataset. As we can see, the SNP numbers between A25 and all subclades (including subclade 1 itself) order in this way: subclade1> subclade6> subclade5> subclade3/4> subclade7> subclade2. This can be explained by the reasoning that when the evolutionary distances between species enlarge with time, the gene similarity diminishes at the genome level, resulting in less SNPs sites shared by subclades. This explanation is supported by the species tree in Fig B and C. Notably, the subclade2 got no SNPs when compared to A25, which means the signature of genetic similarity that can be reflected by gene SNPs has been completely eliminated due to the over large evolutionary distance between them. (B and C) ML tree of strains. B is made based on the whole genome alignment and C based on 18S rDNA. Trees are built using RAXML with the evolution model GTRGAMMA and rooted with the subclade2 sequence. Node values are support from 100 bootstrap running.

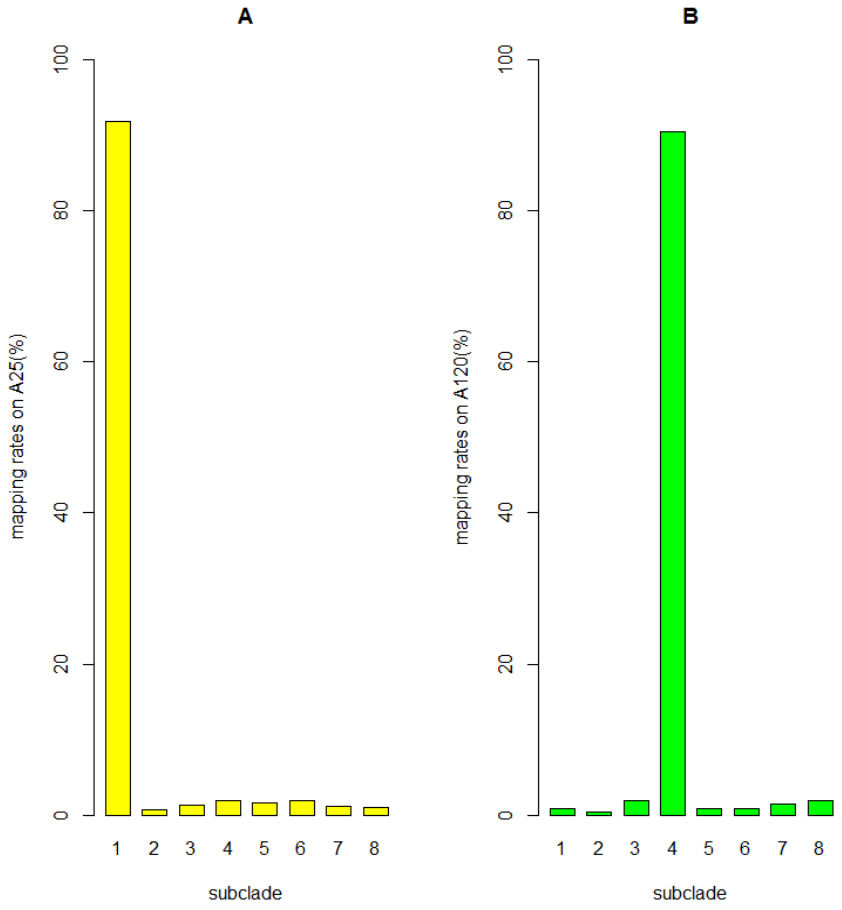


Fig 6. Barplot of average remapping rates of whole genome reads from strains/single cells belonging to different subclades against the complete genomes A25 (A) and against A120 (B). As A25 and A120 belong to subclade 1 and 4 respectively, the remapping rates from subclade 1 and 4 on A25 and A120 respectively are both high (over 90%). In contrast, the remapping rates from the subclades 2, 3, 5, 6, 7 and 8 are much lower (less than 2%).

In this thesis, a large effort has been made to predict introns for all the sequenced genomes. However, the classic way of genome annotation based on mRNA and/or proteins from close relatives as evidence is not working. Having applied the commonly used approach of gene prediction and genome annotation on the strains investigated in chapter 1, I found the detected BUSCOs (Benchmarking Universal Single-Copy Orthologs) (**Fig** 7A) are generally low (<75%) and the predicted gene numbers significantly deviated from those of the fully sequenced two genomes (**Fig** 7B). This is because the noncanonical architecture of gene structures in *Amoebophrya* is still an enigma, thus hampering a sound gene structure prediction.

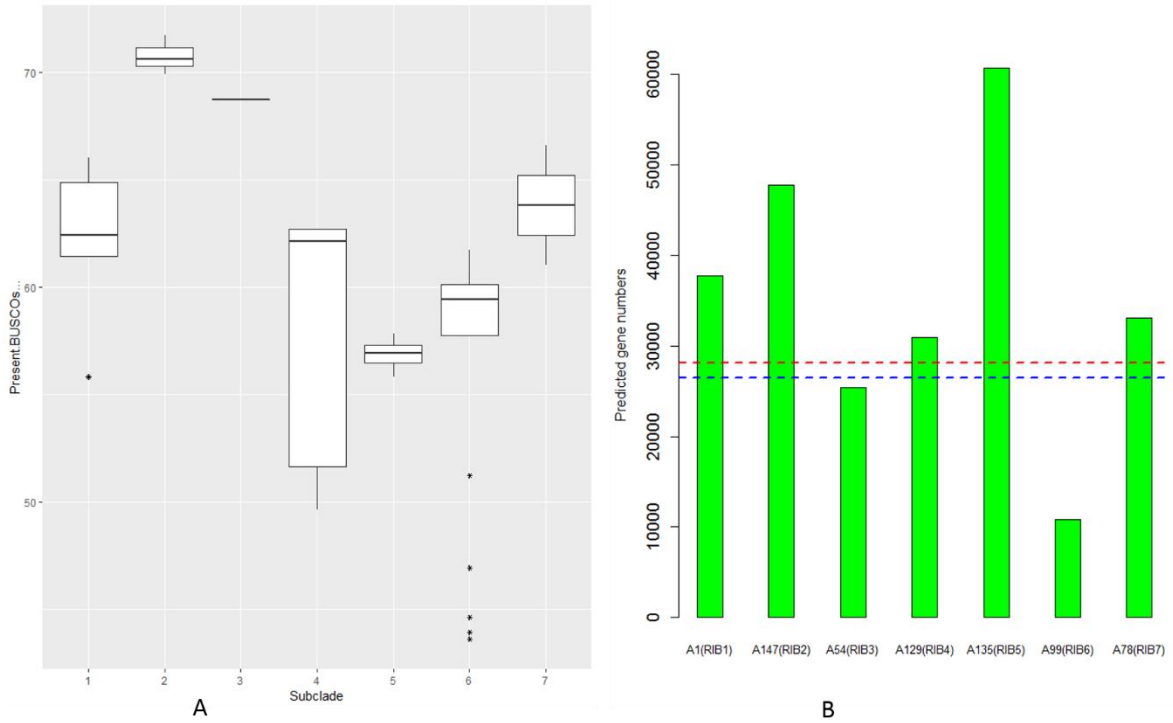


Fig 7. (A) Boxplot of BUSCOs (%) detected from strains belonging to different subclades. Stars in the figure indicate outliers in the data. 303 BUSCOs genes are used as the database. The analysis is done for the draft assemblies by the CLC assembler. (B) Barplot of predicted gene numbers from strains belonging to different subclades. This is done for the best assembly from each subclade, which was chosen by taking into consideration both N50 values and detected BUSCOs numbers. The dash line indicated the predicted protein-coding genes for A25 (red) and A120 (blue), both of which have been fully sequenced.

Looking forward, for dinoflagellates, an ancient alveolate group of about 2000 described extant species, DNA barcoding studies have revealed a large amount of unrecognized species diversity, most of which is not represented in culture collections. We expect that the biodiversity of microbial single-celled eukaryotic species will be evaluated more properly and that cryptic species will become more discernable with the era of genomics coming.

Glossary

BUSCO:

A method for quantitative assessment of genome assembly and annotation completeness based on the evolutionarily informed expectation that sets of Benchmarking Universal Single-Copy Orthologs (BUSCOs) should be present in any genome as single-copy orthologs (Simão et al., 2015). If the BUSCOs are not found in a genome assembly or annotated gene set, the assembly and/or annotation may have failed to reveal the complete expected gene content.

DVNP:

Dinoflagellate chromosomes are permanently condensed with nucleoproteins originating from phycodnaviruses called DVNPs (Dinoflagellate/Viral NucleoProteins). These proteins bind DNA in place of histones and are found throughout all dinoflagellates, including early- and late-branching taxa (Gornik et al., 2012).

TARA:

Tara is a 36-metre-long schooner. Started in September 2009, the schooner's expeditions (Tara Oceans and Tara Oceans Polar Circle) lasted for three years. Its purpose was to investigate planktonic and coral ecosystems around the world in the perspective of climate changes. (https://oceans.taraexpeditions.org/en/m/about-tara/les-expeditions/tara-oceans/)

MALASPINA:

The Malaspina expedition was a research project to assess the impact of global change on the oceans and explore their biodiversity. It started in December 2010 and lasted for eight-month. The project was supported by the Spanish Ministry of Science and Innovation and the Spanish Navy and led by the Spanish National Research Council (CSIC). (https://en.wikipedia.org/wiki/Malaspina_Expedition_2010)

OSD:

The Ocean Sampling Day (OSD) was a concurrent sampling campaign in the world's oceans which occurred on June 21st in the year 2014 and 2015. These samples provided insights into fundamental rules for describing microbial diversity and function. (https://www.microb3.eu/osd.html)

General References

General References (for both general introduction and general discussion)

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., ... & Cárdenas, P. (2019). Revisions to the classification, nomenclature, and diversity of eukaryotes. Journal of Eukaryotic Microbiology, 66(1), 4-119.

Adl, S. M., Simpson, A. G., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., ... & Heiss, A. (2012). The revised classification of eukaryotes. Journal of Eukaryotic Microbiology, 59(5), 429-514.

Alves-de-Souza, C., Varela, D., Iriarte, J. L., González, H. E., & Guillou, L. (2012). Infection dynamics of Amoebophryidae parasitoids on harmful dinoflagellates in a southern Chilean fjord dominated by diatoms. Aquatic Microbial Ecology, 66(2), 183-197.

Aranda, M., Li, Y., Liew, Y. J., Baumgarten, S., Simakov, O., Wilson, M. C., ... & Ryu, T. (2016). Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Scientific Reports, 6, 39734.

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, 314-322.

Bachvaroff, T. R., Handy, S. M., Place, A. R., & Delwiche, C. F. (2011). Alveolate phylogeny inferred using concatenated ribosomal proteins. Journal of Eukaryotic Microbiology, 58(3), 223-233.

Baldauf, S. L. (2003). The deep roots of eukaryotes. Science, 300(5626), 1703-1706.

Brate J, Krabberod AK, Dolven JK, Ose RF, Kristensen T, et al. (2012) Radiolaria associated with large diversity of marine alveolates. Protist 163: 767–777.

Burki, F., Shalchian-Tabrizi, K., & Pawlowski, J. (2008). Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biology letters, 4(4), 366-369.

Burki, F., & Keeling, P. J. (2014). Rhizaria. Current Biology, 24(3), 103-107.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S. I., Jakobsen, K. S., & Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. PloS one, 2(8), e790.

Cachon, J. & Cachon, M. (1987). Parasitic dinoflagellates. In : F. J. R. Taylor (ed.), The Biology of Dinoflagellates. Blackwell Sci. Publ., Oxford. p. 571-610.

Cachon, J. (1964). Contribution a l'étude des Péridinies parasites. Cytologie, cycles évolutifs. Ann. Sci. Nat., 12 ser. 6: 1-158

Cachon, J., & Cachon, M. (1970). Ultrastructure des Amoebophryidae (Péridiniens Duboscquodinida) II Systèmes atractophoriens et microtubulaires; leur intervention dans la mitose. Protistologica, 6, 57-70.

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., … Worden, A. Z. (2016). Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nature Reviews Microbiology, 15, 6– 20.

Cavalier-Smith, T., & Chao, E. E. (2004). Protalveolate phylogeny and systematics and the origins of Sporozoa and dinoflagellates (phylum Myzozoa nom. nov.). European Journal of Protistology, 40(3), 185-212.

Chambouvet A, Morin P, Marie D, & Guillou L (2008). Control of Toxic Marine Dinoflagellate Blooms by Serial Parasitic Killers. Science, 322(5905), 1254–1257.

Chambouvet, A., Alves-de-Souza, C., Cueff, V., Marie, D., Karpov, S., & Guillou, L. (2011). Interplay between the parasite *Amoebophrya* sp.(Alveolata) and the cyst formation of the red tide dinoflagellate *Scrippsiella trochoidea*. Protist, 162(4), 637-649.

Coats DW & Park MG (2002). Parasitism of photosynthetic dinoflagellates by three strains of *Amoebophrya* (Dinophyta): parasite survival, infectivity, generation time, and host specificity. Journal of Phycology, 38(3), 520-528.

Coats, D. W. (1988). Duboscquella cachoni N. Sp., a Parasitic Dinoflagellate Lethal to Its Tintinnine Host *Eutintinnus pectinis*. The Journal of protozoology, 35(4), 607-617.

Coats, D. W., Adam, E. J., Gallegos, C. L., & Hedrick, S. (1996). Parasitism of photosynthetic dinoflagellates in a shallow subestuary of Chesapeake Bay, USA. Aquatic Microbial Ecology, 11(1), 1-9.

Coats, D. W., Bachvaroff, T. R., & Delwiche, C. F. (2012). Revision of the Family Duboscquellidae with Description of *Euduboscquella crenulata* n. gen., n. sp. (Dinoflagellata, Syndinea), an Intracellular Parasite of the Ciliate *Favella panamensis* Kofoid & Campbell. Journal of Eukaryotic Microbiology, 59(1), 1-11.

Coats, D. W., Kim, S., Bachvaroff, T. R., Handy, S. M., & Delwiche, C. F. (2010). Tintinnophagus acutus ng, n. sp. (Phylum Dinoflagellata), an ectoparasite of the ciliate *Tintinnopsis cylindrica* Daday 1887, and its relationship to *Duboscquodinium collini* Grassé 1952. Journal of Eukaryotic Microbiology, 57(6), 468-482.

Cuadrado, Á., De Bustos, A., & Figueroa, R. I. (2019). Chromosomal markers in the genus *Karenia*: Towards an understanding of the evolution of the chromosomes, life cycle patterns and phylogenetic relationships in dinoflagellates. Scientific reports, 9(1), 3072.

Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2000). Emerging infectious diseases of wildlife--threats to biodiversity and human health. science, 287(5452), 443-449.

Daugbjerg, N., Hansen, G., Larsen, J., & Moestrup, Ø. (2000). Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates. Phycologia, 39(4), 302-317.

De Queiroz, K. (2007). Species concepts and species delimitation. Systematic biology, 56(6), 879-886.

De Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, ... & Carmichael M (2015). Eukaryotic plankton diversity in the sunlit ocean. Science, 348(6237), 1261605.

Decelle, J., Romac, S., Sasaki, E., Not, F., & Mahe, F. (2014). Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. PLoS One, 9(8), e104297.

Díez, B., Pedrós-Alió, C., & Massana, R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Applied and environmental microbiology, 67(7), 2932-2941.

Dodge, J. D. (1987). Dinoflagellate ultrastructure. In: Taylor FJR, editor. The Biology of Dinoflagellates. Oxford: Blackwell Scientific Publications. pp93–119.

Dodge, J. D., Crawford, R. M. (1968). Fine structure of the dinoflagellate *Amphidinium carteri Hulburt*. Protistologica, 4(231), 468.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics, 34(14), 2371-2375.

Escalante, A. A., & Ayala, F. J. (1995). Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. Proceedings of the National Academy of Sciences, 92(13), 5793-5797.

Farhat, S., Florent, I., Noël, B., Kayal, E., Da Silva, C., Bigeard, E., ... & Rombauts, S. (2018). Comparative time-scale gene expression analysis highlights the infection processes of two Amoebophrya strains. Frontiers in microbiology, 9, 2251.

Fensome, R. A., Saldarriaga, J. F., & Taylor, M. F. (1999). Dinoflagellate phylogeny revisited: reconciling morphological and molecular based phylogenies. Grana, 38(2-3), 66-80.

Fensome, R. A. (1993). A classification of living and fossil dinoflagellates. Micropaleontology, special publication, 7, 1-351.

Figueroa, R. I., & Bravo, I. (2005). Sexual reproduction and two different encystment strategies of *Lingulodinium polyedrum* (Dinophyceae) in culture. Journal of Phycology, 41(2), 370-379.

Figueroa, R. I., Bravo, I., & Garcés, E. (2006). The multiple routes of sexuality in *Alexandrium tailori* (Dinophyceae) in culture. Journal of Phycology, 42(5), 1028-1039.

Figueroa, R. I., Estrada, M. & Garcés, E. (2018). Life histories of microalgal species causing harmful blooms: Haploids, diploids and the relevance of benthic stages. Harmful Algae 73, 44–57.

Figueroa, R. I., Garces, E., & Bravo, I. (2007). Comparative study of the life cycles of *Alexandrium tamutum* and *Alexandrium minutum* (Gonyaulacales, Dinophyceae) in culture. Journal of Phycology, 43(5), 1039-1053.

Figueroa, R. I., Rengefors, K., Bravo, I., & Bensch, S. (2010). From homothally to heterothally: mating preferences and genetic variation within clones of the dinoflagellate *Gymnodinium catenatum*. Deep Sea Research Part II: Topical Studies in Oceanography, 57(3-4), 190-198.

Flegontov, P., Michálek, J., Janouškovec, J., Lai, D. H., Jirků, M., Hajdušková, E., ... & Oborník, M. (2015). Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. Molecular Biology and Evolution, 32(5), 1115-1131.

Flewelling, L. J., Naar, J. P., Abbott, J. P., Baden, D. G., Barros, N. B., Bossart, G. D., Landsberg, J. H. (2005). Brevetoxicosis: Red tides and marine mammal mortalities. Nature, 435, 755–756.

Frehi, H., Couté, A., Mascarell, G., Perrette-Gallet, C., Ayada, M., & Kara, M. H. (2007). Harmful and red-tide dinoflagellates in the Annaba bay (Algeria). Comptes rendus biologies, 330(8), 615-628.

Fukuda, Y., Endoh, H. (2006). New details from the complete life cycle of the red-tide dinoflagellate *Noctiluca scintillans* (Ehrenberg) McCartney. European Journal of Protistology 42: 209–219.

Fukuda, Y., Endoh, H. (2008). Phylogenetic analyses of the dinoflagellate *Noctiluca scintillans* based on beta-tubulin and Hsp90 genes. European Journal of Protistology 44: 27–33.

Gaines, G., & Taylor, F. J. R. (1985). Form and Function of the Dinoflagellate Transverse Flagellum. The Journal of protozoology, 32(2), 290-296.

Gajadhar, A. A., Marquardt, W. C., Hall, R., Gunderson, J., Ariztia-Carmona, E. V., & Sogin, M. L. (1991). Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Molecular and biochemical parasitology, 45(1), 147-154.

Gárate-Lizárraga, I., Band-Schmidt, C. J., López-Cortés, D. J., & Muñetón-Gómez, M. D. S. (2009). Bloom of *Scrippsiella trochoidea* (Gonyaulacaceae) in a shrimp pond in the southwestern Gulf of California, Mexico. Marine pollution bulletin, 58(1), 145-149.

Gast, R. J., Moran, D. M., Dennett, M. R., & Caron, D. A. (2007). Kleptoplasty in an Antarctic dinoflagellate: Caught in evolutionary transition? Environmental Microbiology, 9, 39–45.

Giacobbe, M. G., & Yang, X. (1999). The life history of *Alexandrium taylori* (Dinophyceae). Journal of Phycology, 35(2), 331-338.

Giraud, T., Refrégier, G., Le Gac, M., de Vienne, D. M., & Hood, M. E. (2008). Speciation in fungi. Fungal Genetics and Biology, 45(6), 791-802.

Gómez F (2012). A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata). Systematics and Biodiversity, 10(3): 267–275.

Gómez, F. (2012). A checklist and classification of living dinoflagellates (Dinoflagellata, Alveolata). Cicimar Oceánides, 27(1), 65-140.

Goodson, M. S., Whitehead, L. F., & Douglas, A. E. (2001). Symbiotic dinoflagellates in marine Cnidaria: Diversity and function. Hydrobiologia,461, 79–82.

Gornik, S. G., Cassin, A. M., MacRae, J. I., Ramaprasad, A., Rchiad, Z., McConville, M. J., ... & Waller, R. F. (2015). Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. Proceedings of the National Academy of Sciences, 112(18), 5767-5772.

Gornik, S. G., Ford, K. L., Mulhern, T. D., Bacic, A., McFadden, G. I., & Waller, R. F. (2012). Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. Current Biology, 22(24), 2303-2312.

Gribble, K. E., Anderson, D. M. (2006). Molecular phylogeny of the heterotrophic dinoflagellates, Protoperidinium, Diplopsalis and Preperidinium (Dinophyceae), inferred from large subunit rDNA. Journal of Phycology 42: 1081–1095.

Gu, H., Sun, J., Kooistra, W. H., & Zeng, R. (2008). Phylogenetic position and morphology of thecae and cysts of *Scrippsiella* (dinophyceae) species in the East China Sea. Journal of Phycology, 44(2), 478-494.

Guillou, L., Alves-de Souza, C., Siano, R., & Gonzalez, H. (2010). The ecological significance of small eukaryotic parasites in marine ecosystems. Microbiol Today, 92-95.

Guillou, L., Viprey, M., Chambouvet, A., Welsh, R. M., Kirkham, A. R., Massana, R., ... & Worden, A. Z. (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environmental microbiology, 10(12), 3349-3365.

Guiry, M. D. (2012). How many species of algae are there? Journal of phycology, 48(5), 1057-1063.

Gunderson, J. H., JOHN, S. A., Chanson Boman, W., & Coats, D. W. (2002). Multiple strains of the parasitic dinoflagellate *Amoebophrya* exist in Chesapeake Bay. Journal of Eukaryotic Microbiology, 49(6), 469-474.

Hackett, J. D., Anderson, D. M., Erdner, D. L., & Bhattacharya, D. (2004). Dinoflagellates: a remarkable evolutionary experiment. American Journal of Botany, 91(10), 1523-1534.

Hallegraeff, G. M. (1992). Harmful algal blooms in the Australian region. Marine pollution bulletin, 25(5-8), 186-190.

Hameed, H. A., & Saburova, M. (2015). First record of *Scrippsiella trochoidea* (Dinophyceae) in Shatt Al-Arab River (Southern Iraq). Marine Biodiversity Records, 8.

Hao, X., Jiang, R., & Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics, 27(5), 611-618.

Harada, A., Ohtsuka, S., Horiguchi, T. (2007). Species of the parasitic genus *Duboscquella* are members of the enigmatic Marine Alveolate Group I. Protist158, 337–347.

Havskum, H., & Hansen, P. J. (2006). Net growth of the bloom-forming dinoflagellate *Heterocapsa triquetra* and pH: why turbulence matters. Aquatic Microbial Ecology, 42(1): 55–62

Hoek, C., Mann, D., Jahns, H. M., & Jahns, M. (1995). Algae: an introduction to phycology. Cambridge university press. pp. 277–280.

Hoppenrath & Saldarriaga. (2008). Available: http://tolweb.org/Dinoflagellates/2445.

Hoppenrath, M. (2004). A revised checklist of planktonic diatoms and dinoflagellates from Helgoland (North Sea, German Bight). Helgoland Marine Research, 58(4), 243.

Hoppenrath, M. (2017). Dinoflagellate taxonomy—a review and proposal of a revised classification. Marine Biodiversity, 47(2), 381-403.

Hoppenrath, M., & Leander, B. S. (2010). Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences. PloS one, 5(10), e13220.

Hoppenrath, M., Bachvaroff, T. R., Handy, S. M., Delwiche, C. F., & Leander, B. S. (2009). Molecular phylogeny of ocelloid-bearing dinoflagellates (Warnowiaceae) as inferred from SSU and LSU rDNA sequences. BMC Evolutionary Biology, 9(1), 116.

Jackson, C. J., Gornik, S. G., & Waller, R. F. (2011). The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genome biology and evolution, 4(1), 59-72.

Janouškovec, J., Gavelis, G. S., Burki, F., Dinh, D., Bachvaroff, T. R., Gornik, S. G., ... & Waller, R. F. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proceedings of the National Academy of Sciences, 114(2), E171-E180.

Janouškovec, J., Horák, A., Oborník, M., Lukeš, J., & Keeling, P. J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proceedings of the National Academy of Sciences, 107(24), 10949-10954.

Jeong, H. J., Yoo, Y. D., Kim, J. S., Seong, K. A., Kang, N. S., & Kim, T. H. (2010). Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Science Journal, 45, 65–91.

John, U., Lu, Y., Wohlrab, S., Groth, M., Janouškovec, J., Kohli, G. S., ... & Frickenhaus, S. (2019). An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. Science advances, 5(4), eaav1110.

Jørgensen, M. F., Murray, S., & Daugbjerg, N. (2004). A new genus of athecate interstitial dinoflagellates, *Togula* gen. nov., previously encompassed within *Amphidinium* sensu lato: Inferred from light and electron microscopy and phylogenetic analyses of partial large subunit ribosomal DNA sequences. Phycological Research, 52(3), 284-299.

Joyce, L. B., Pitcher, G. C., Du Randt, A., & Monteiro, P. M. S. (2005). Dinoflagellate cysts from surface sediments of Saldanha Bay, South Africa: an indication of the potential risk of harmful algal blooms. Harmful algae, 4(2), 309-318.

Keane, R. M., & Crawley, M. J. (2002). Exotic plant invasions and the enemy release hypothesis. Trends in ecology & evolution, 17(4), 164-170.

Khadka, M., Salem, M., & Leblond, J. D. (2015). Sterol Composition and Biosynthetic Genes of *Vitrella brassicaformis*, a Recently Discovered Chromerid: Comparison to Chromera velia and Phylogenetic Relationship with Apicomplexan Parasites. Journal of Eukaryotic Microbiology, 62(6), 786-798.

Ki, J. S. (2010). Nuclear 28S rDNA phylogeny supports the basal placement of *Noctiluca scintillans* (Dinophyceae; Noctilucales) in dinoflagellates. European journal of protistology, 46(2), 111-120.

Kim, S. (2006). Patterns in host range for two strains of *Amoebophrya* (Dinophyta) infecting thecate dinoflagellates: *Amoebophrya* spp. ex *Alexandrium affine* and ex *Gonyaulax polygramma*. Journal of phycology, 42(6), 1170-1173.

Kim, S. J., Jeong, H. J., Kang, H. C., You, J. H., & Ok, J. H. (2019). Differential feeding by common heterotrophic protists on four *Scrippsiella* species of similar size. Journal of phycology.

Kim, S., Park, M. G., KIM, K. Y., KIM, C. H., Yih, W., Park, J. S., & Coats, D. W. (2008). Genetic diversity of parasitic dinoflagellates in the genus *Amoebophrya* and its relationship to parasite biology and biogeography. Journal of eukaryotic microbiology, 55(1), 1-8.

Kim, Y. O., & Han. (2000). Seasonal relationships between cyst germination and vegetative population of *Scrippsiella trochoidea* (Dinophyceae). Marine Ecology Progress Series, 204: 111–118

Knowlton, N. (1993). Sibling species in the sea. Annual review of ecology and systematics, 24(1), 189-216.

Kohli, G. S., John, U., Figueroa, R. I., Rhodes, L. L., Harwood, D. T., Groth, M., ... & Murray, S. A. (2015). Polyketide synthesis genes associated with toxin production in two species of *Gambierdiscus* (Dinophyceae). BMC genomics, 16(1), 410.

Kremp, A., & Parrow, M. W. (2006). Evidence for asexual resting cysts in the life cycle of the marine peridinoid dinoflagellate, *Scrippsiella hangoei*. Journal of Phycology, 42(2), 400-409.

Kuvardina, O. N., Leander, B. S., Aleshin, V. V., Myl'nikov, A. P., Keeling, P. J., Simdyanov, T. G. (2002). The phylogeny of colpodellids (Alveolata) using small subunit rRNA gene sequences suggests they are the free-living sister group to apicomplexans. Journal of Eukaryotic Microbiology, 49(6), 498-504.

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.

Leander, B. S., & Keeling, P. J. (2004). Early evolutionary history of dinoflagellates and apicomplexans (Alveolata) as inferred from hsp90 and actin phylogenies. Journal of Phycology, 40(2), 341-350.

Lee, S. Y., Jeong, H. J., You, J. H. & Kim, S. J. (2018). Morphological and genetic characterization and the nationwide distribution of the phototrophic dinoflagellate *Scrippsiella lachrymose* in the Korean waters. Algae 33:21-35.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.

Lewis, J. (1991). Cyst-theca relationships in *Scrippsiella* (Dinophyceae) and related orthoperidinioid genera. Botanica marina, 34(2), 91-106.

Lidie, K. B., & Van Dolah, F. M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. Journal of Eukaryotic Microbiology, 54(5), 427-435.

Liew, Y. J., Li, Y., Baumgarten, S., Voolstra, C. R. & Aranda, M. (2017). Condition-specific RNA editing in the coral symbiont *Symbiodinium microadriaticum*. PLoS Genet. 13, e1006619.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., ... & Bittner, L. (2015). Determinants of community structure in the global plankton interactome. Science, 348(6237), 1262073.

Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., . . . Morse, D. (2015). The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. Science, 350, 691–694.

Lin, S., Zhang, H., Zhuang, Y., Tran, B., & Gill, J. (2010). Spliced leader–based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. Proceedings of the National Academy of Sciences, 107(46), 20033-20038.

Liu, H., Stephens, T. G., González-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P., ... & Miller, D. J. (2018). *Symbiodinium* genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. Communications biology, 1(1), 95.

López-García, P., López-López, A., Moreira, D., & Rodríguez-Valera, F. (2001). Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. FEMS Microbiology Ecology, 36(2-3), 193-202.

M., Murray, S. A. (2015). Polyketide synthesis genes associated with toxin production in two species of *Gambierdiscus* (Dinophyceae). BMC Genomics, 16, 410.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. PeerJ, 2, e593.

Mai, J. C., & Coleman, A. W. (1997). The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. Journal of Molecular Evolution, 44(3), 258-271.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., ... & Dolan, J. R. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environmental microbiology, 17(10), 4035-4049.

Mayden, R. L. (1997). A hierarchy of species concepts: the denouement in the saga of the species problem. In Species: The Units of Biodiversity (Claridge, M.F., Dawah, H.A. & Wilson, M.R., editors), 381-424. Chapman and Hall, London.).

Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., ... & Not, F. (2018). Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Molecular ecology, 27(10), 2365-2380.

Miller, J. J., Delwiche, C. F., & Coats, D. W. (2012). Ultrastructure of Amoebophrya sp. and its changes during the course of infection. Protist, 163(5), 720-745.

Montagnes D. J. S., Chambouvet A., Guillou L., Fenton A. (2008). Can microzooplankton and parasite pressure be responsible for the demise of toxic dinoflagellate blooms? Aquatic Microbial Ecology. 53:201-210.

Moon-van der Staay, S. Y., De Wachter, R., & Vaulot, D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature, 409(6820), 607.

Moore, R. B., Oborník, M., Janouškovec, J., Chrudimský, T., Vancová, M., Green, D. H., ... & Šlapeta, J. (2008). A photosynthetic alveolate closely related to apicomplexan parasites. Nature, 451(7181), 959.

Müller, T., Philippi, N., Dandekar, T., Schultz, J., & Wolf, M. (2007). Distinguishing species. Rna, 13(9), 1469-1472.

Murray, S. A., Suggett, D. J., Doblin, M. A., Kohli, G. S., Seymour, J. R., Fabris, M., & Ralph, P. J. (2016). Unravelling the functional genetics of dinoflagellates: a review of approaches and opportunities. Perspect. Phycol, 3(1), 37-52.

Murray, S., Jørgensen, M. F., Ho, S. Y., Patterson, D. J., & Jermiin, L. S. (2005). Improving the analysis of dinoflagellate phylogeny based on rDNA. Protist, 156(3), 269-286.

Norén, F., Moestrup, Ø., & Rehnstam-Holm, A. S. (1999). *Parvilucifera infectans* Norén et Moestrup gen. et sp. nov.(Perkinsozoa phylum nov.): a parasitic flagellate capable of killing toxic microalgae. European journal of protistology, 35(3), 233-254.

Olli, K., & Anderson, D. M. (2002). High encystment success of the dinoflagellate *Scrippsiella* cf. *lachrymosa* in culture experiments. Journal of Phycology, 38(1), 145-156.

Orr, R. J. S., St€uken, A., Murray, S. A., & Jakobsen, K. S. (2013). Evolutionary acquisition and loss of saxitoxin biosynthesis in dinoflagellates: The second "core" gene, sxtG. Applied and Environment Microbiology, 79, 2128–2136.

Orr, R. J., 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.

Parfrey, L.W., Grant, J., Tekle, Y.I., Lasek-Nesselquist, E., Morrison, H.G., Sogin, M.L., Patterson, D.J., Katz, L.A. (2010). Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst. Biol. 59, 518-533.

Park, M. G., Kim, S., Shin, E. Y., Yih, W., & Coats, D. W. (2013). Parasitism of harmful dinoflagellates in Korean coastal waters. Harmful algae, 30, S62-S74.

Pernice, M. C., Logares, R., Guillou, L., & Massana, R. (2013). General patterns of diversity in major marine microeukaryote lineages. PLoS One, 8(2), e57170.

Persson, A., Godhe, A., & Karlson, B. (2000). Dinoflagellate cysts in recent sediments from the west coast of Sweden. Botanica marina, 43(1), 69-79.

Persson, A., Smith, B. C., Wikfors, G. H., & Alix, J. H. (2013). Differences in swimming pattern between life cycle stages of the toxic dinoflagellate *Alexandrium fundyense*. Harmful Algae, 21, 36-43.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). Ecological niches and geographic distributions (MPB-49) (Vol. 56). Princeton University Press.

Pfiester, L. A., Anderson, D. M. (1987). Dinoflagellate reproduction. In:Taylor, F.J.R. (Ed.), The Biology of Dinoflagellates. BlackwellScience, Oxford, pp. 611–648.

Piganeau, G., Eyre-Walker, A., Grimsley, N., & Moreau, H. (2011). How and why DNA barcodes underestimate the diversity of microbial eukaryotes. PloS one, 6(2), e16342.

Reyes-Prieto, A., Moustafa, A., & Bhattacharya, D. (2008). Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. Current Biology, 18(13), 956-962.

Roy, S., & Morse, D. (2012). A full suite of histone and histone modifying genes are transcribed in the dinoflagellate *Lingulodinium*. PLoS One, 7(4), e34340.

Saldarriaga, J. F., Cavalier-Smith, T., Menden-Deuer, S., & Keeling, P. J. (2004). Molecular data and the evolutionary history of dinoflagellates. European journal of protistology, 40(1), 85-111.

Saldarriaga, J. F., McEwan, M. L., Fast, N. M., Taylor, F. J. R., & Keeling, P. J. (2003). Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. International Journal of Systematic and Evolutionary Microbiology, 53(1), 355-365.

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.

Shalchian-Tabrizi, K., Minge, M. A., Cavalier-Smith, T., Nedreklepp, J. M., Klaveness, D., & Jakobsen, K. S. (2006). Combined heat shock protein 90 and ribosomal RNA sequence phylogeny supports multiple replacements of dinoflagellate plastids. Journal of Eukaryotic Microbiology, 53(3), 217-224.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., ... & Hamada, M. (2013). Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. Current biology, 23(15), 1399-1408.

Siano, R., Alves-de-Souza, C., Foulon, E., Bendif, E. M., Simon, N., Guillou, L., & Not, F. (2011). Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. Biogeosciences, 8(2), 267-278.

Sibbald, S. J., & Archibald, J. M. (2017). More protist genomes needed. Nature Ecology and Evolution, 1(5), 145.

Skovgaard, A., Meneses, I., & Angélico, M. M. (2009). Identifying the lethal fish egg parasite *Ichthyodinium chabelardi* as a member of Marine Alveolate Group I. Environmental microbiology, 11(8), 2030-2041.

Soehner, S., Zinssmeister, C., Kirsch, M., & Gottschling, M. (2012). Who am I—and if so, how many? Species diversity of calcareous dinophytes (Thoracosphaeraceae, Peridiniales) in the Mediterranean Sea. Organisms Diversity & Evolution, 12(4), 339-348.

Soyer-Gobillard, M. O. & Geraud, M. L. (1992). Nucleolus behaviour during the cell cycle of a primitive dinoflagellate eukaryote, *Prorocentrum micans* Ehr., seen by light microscopy and electron microscopy. J Cell Sci, 102(3), 475-485.

Stephens, T. G., Ragan, M. A., Bhattacharya, D., & Chan, C. X. (2018). Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. Scientific reports, 8(1), 17175.

Stosch, H. V. (1972). La signification cytologique de la «cyclose nucléaire» dans le cycle de vie des Dinoflagellés. Bulletin de la Société Botanique de France, 119(sup1), 201-211.

Strassert, J. F., Karnkowska, A., Hehenberger, E., del Campo, J., Kolisko, M., Okamoto, N., ... & Hallam, S. J. (2018). Single cell genomics of uncultured marine alveolates shows paraphyly of basal dinoflagellates. The ISME journal, 12(1), 304.

Takayama, H. (1985) Apical grooves of unarmoured dinoflagellates. Bulletin of Plankton Society of Japan, 32(2), 129-140.

Tas, S., & Yilmaz, I. N. (2015). Potentially harmful microalgae and algal blooms in a eutrophic estuary in Turkey. Mediterranean Marine Science, 16(2), 432-443.

Taylor, F. J. R. (1987). General group characteristics, special features of interest, short history of dinoflagellate study. In: Taylor FJR, editor. The Biology of dinoflagellates. Oxford.: Blackwell Scientific Publications. 798.

Taylor, F. J. R., Hoppenrath, M., & Saldarriaga, J. F. (2008). Dinoflagellate diversity and distribution. Biodiversity and conservation, 17(2), 407-418.

Tikhonenkov, D. V., Janouškovec, J., Mylnikov, A. P., Mikhailov, K. V., Simdyanov, T. G., Aleoshin, V. V., & Keeling, P. J. (2014). Description of *Colponema vietnamica* sp. n. and *Acavomonas peruviana* n. gen. n. sp., two new alveolate phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and their contributions to reconstructing the ancestral state of alveolates and eukaryotes. PLoS One, 9(4), e95467.

Tillmann, U., Salas, R., Gottschling, M., Krock, B., O'Driscoll, D., & Elbrächter, M. (2012). *Amphidoma languida* sp. nov. (Dinophyceae) reveals a close relationship between *Amphidoma* and *Azadinium*. Protist, 163(5), 701-719.

Velo-Suárez, L., Brosnahan, M. L., Anderson, D. M., & McGillicuddy Jr, D. J. (2013). A quantitative assessment of the role of the parasite *Amoebophrya* in the termination of *Alexandrium fundyense* blooms within a small coastal embayment. PLoS One, 8(12), e81150.

Villarino, M. L., Figueiras, F. G., Jones, K. J., Álvarez-Salgado, X. A., Richard, J., & Edwards, A. (1995). Evidence of in situ diel vertical migration of a red-tide microplankton species in Ria de Vigo (NW Spain). Marine biology, 123(3), 607-617.

von Dassow, P., & Montresor, M. (2010). Unveiling the mysteries of phytoplankton life cycles: patterns and opportunities behind complexity. Journal of Plankton Research, 33(1), 3-12.

von Stosch, H. V. (1973). Observations on vegetative reproduction and sexual life cycles of two freshwater dinoflagellates, *Gymondinium pseudopalustre* Schiller and *Woloszynskia apiculata* sp. nov. British Phycological Journal, 8(2), 105-134.

Waller, R. F., & Jackson, C. J. (2009). Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. Bioessays, 31(2), 237-245.

Wang, S., Tang, D., He, F., Fukuyo, Y., & Azanza, R. V. (2008). Occurrences of harmful algal blooms (HABs) associated with ocean environments in the South China Sea. Hydrobiologia, 596(1), 79-93.

Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate genome evolution. Annual review of microbiology, 65, 369-387.

Yamaguchi, A., Kawamura, H., & Horiguchi, T. (2006). A further phylogenetic study of the heterotrophic dinoflagellate genus, *Protoperidinium* (Dinophyceae) based on small and large subunit ribosomal RNA gene sequences. Phycological Research, 54(4), 317-329.

Zhang, H., Bhattacharya, D., & Lin, S. (2007). A three-gene dinoflagellate phylogeny suggests monophyly of prorocentrales and a basal position for *Amphidinium* and *Heterocapsa*. Journal of Molecular Evolution, 65(4), 463-474.

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. Proceedings of the National Academy of Sciences, 104(11), 4618-4623.

Zhang, Z., Green, B. R., & Cavalier-Smith, T. (1999). Single gene circles in dinoflagellate chloroplast genomes. Nature, 400(6740), 155.

Zhao, Y., Yi, Z., Warren, A., & Song, W. B. (2018). Species delimitation for the molecular taxonomy and ecology of the widely distributed microbial eukaryote genus *Euplotes* (Alveolata, Ciliophora). Proceedings of the Royal Society B: Biological Sciences, 285(1871), 20172159.

Annexes

Rapid protein evolution and invasive intronic elements in two marine protistan parasites

Sarah Farhat[1], Phuong Le[1], Ehsan Kayal[1], Benjamin Noel[1], Estelle Bigeard, Erwan Corre, Florian Maumus, Isabelle Florent, Adriana Alberti, Jean-Marc Aury, Tristan Barbeyron, Ruibo Cai, Corinne Da Silva, Benjamin Istace, Karine Labadie, Dominique Marie, Jonathan Mercier, Tsinda Rukwavu, Thierry Tonon, Catharina Alves-de-Souza, Pierre Rouzé, Yves Van de Peer, Patrick Wincker, Stephane Rombauts, Betina M. Porcel[2,*], Laure Guillou[2,*]

1 co 1er author

2 co-last author

* Corresponding author

Sarah Farhat, sfarhat@genoscope.cns.fr, address: Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

Phuong Le, phule@psb.vib-ugent.be, address:

1- Center for Plant Systems Biology, VIB, Ghent, Belgium,

2- Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

Ehsan Kayal, ehsan.kayal@sb-roscoff.fr, address: Sorbonne Université, CNRS, FR2424, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

Benjamin Noel, bnoel@genoscope.cns.fr, address: Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Estelle Bigeard, bigeard@sb-roscoff.fr, address: Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France

Erwan Corre, erwan.corre@sb-roscoff.fr, address: Sorbonne Université, CNRS, FR2424, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

Florian Maumus, florian.maumus@inra.fr, address: URGI, INRA, Université Paris-Saclay, 78026, Versailles, France

Isabelle Florent, isabelle.florent@mnhn.fr, address: Sorbonne Université, Muséum national d'histoire naturelle, CNRS, UMR 7245, 57 rue Cuvier, 75231 Paris Cedex 05, France

Adriana Alberti, aalberti@genoscope.cns.fr, address: Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

Jean-Marc Aury, jmaury@genoscope.cns.fr, adress : Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Tristan Barbeyron, barbeyron@sb-roscoff.fr, address: Sorbonne Université, CNRS, UMR 8227, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

Ruibo Cai, ruibo.cai@sb-roscoff.fr, address: Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France

Corinne Da Silva, dasilva@genoscope.cns.fr, address: Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Benjamin Istace, bistace@genoscope.cns.fr, address: Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Karine Labadie, klabadie@genoscope.cns.fr, address: Genoscope, Institut François Jacob, CEA, 91057 Evry, France

Dominique Marie, marie@sb-roscoff.fr, address: Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France

Jonathan Mercier, address: Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Tsinda Rukwavu, t.rukwavu@gmail.com, address: Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Thierry Tonon, thierry.tonon@york.ac.uk, address: Centre for Novel Agricultural Products, Department of Biology, University of York, Heslington, York, YO10 5DD, UK.

Catharina Alves-de-Souza,cathsouza@gmail.com, address: Algal Resources Collection, MARBIONC, Center for Marine Sciences, University of North Carolina Wilmington, 5600 Marvin K. Moss Lane, Wilmington, NC 28409, USA

Pierre Rouzé, pirou@psb.vib-ugent.be, address:

1- Center for Plant Systems Biology, VIB, Ghent, Belgium,

2- Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

Yves Van de Peer, yves.vandepeer@psb.ugent.be

1- Center for Plant Systems Biology, VIB, Ghent, Belgium,

2- Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

3- Department of Biochemistry, Genetics and Microbiology, Pretoria, South Africa

Patrick Wincker, pwincker@genoscope.cns.fr, address: Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

Stephane Rombauts, strom@psb.vib-ugent.be, address:

1- Center for Plant Systems Biology, VIB, Ghent, Belgium,

2- Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

Betina M. Porcel, betina@genoscope.cns.fr, address: Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

Laure Guillou, lguillou@sb-roscoff.fr, address: Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France

Abstract:

Dinoflagellates are successful marine protists that harbour atypical genomes in terms of size (3-250 Gb), gene organization and gene expression patterns. Here, we investigated the very compact genomes (~115 Mb) of two *Amoebophrya* strains, basal intracellular dinoflagellate parasites. We discovered a strong tendency for gene co-orientation and high levels of synteny between the genomes of both parasites, despite the low interspecific protein sequence similarity. Most strikingly, we recovered predominantly non-canonical introns unique among eukaryotes regarding the broad variety of splicing motifs, suggesting a novel splicing mechanism, with a tendency to spread over their respective genomes, similar to transposable elements. These NCIs expand the range of possible genome organization in eukaryotes, and might be correlated to the speciation of these co-occurring parasites.

One Sentence Summary :

High protein divergence combined with seemingly invasive intronic elements in the genomes of Syndiniales parasites of dinoflagellates.

Main Text:

Dinoflagellates (Alveolata) are relevant single-cell eukaryotes with a wide range of lifestyles. Approximately half of total known dinoflagellates constitute important primary producers found in oceans worldwide, some of which are responsible for toxic blooms while others live in symbiotic relationships, such as the Symbiodiniaceae found in corals [1]. Unlike other alveolates, dinoflagellates display a wide range of genome sizes (~ 3 to 250 Gb; [2]), which are relatively gene-rich [3] and remain nearly permanently packed into condensed liquid-crystalline dinokaryons (20-270 chromosomes). Their genetic material is associated with nucleoproteins originating from phycodnaviruses (DVNPs, [2]) and histone-like proteins derived from bacterial HU-like proteins [4]. Gene expression in dinoflagellates involves trans-splicing of messenger RNAs [5] through the addition of a 5'-end splice leader (DinoSL) sequence [6][7] still identifiable in the genomic sequence in likely retro-transposed transcripts [8]. These trans-spliced messengers harbor unusual GC/GA dinucleotide pairs at their 5' donor splice site [9], and a putatively translational – rather than transcriptional – gene regulation mechanism [10]. The exploration of early-diverging dinoflagellate lineages such as Syndiniales (also known as environmental Marine Alveolates or MALVs [11]), will likely shed light on the emergence of such atypical genomic features. A recent draft genome of the *Amoebophrya* sp. AT5 strain infecting the toxic autotrophic dinoflagellate *Alexandrium catenella* reported an unusual aerobic mitochondrion likely lacking a genome [16]. Known Syndiniales comprise exclusively of parasites infecting marine eukaryotic species that eventually kill their host to complete their life cycle. *Amoebophrya* species are

intracellular parasites of dinoflagellates, as well as radiolarians, ciliates, and other *Amoebophrya* [12][13]. A single infection by *Amoebophrya*-like parasites can lead to the production of hundreds of dinospores, the parasite's infective and flagellated propagules. While the host specificity varies among strains, *Amoebophrya* spp. are generally observed to be highly host-specific and are likely drivers of successive dinoflagellate blooms [11][14][15]. Consequently, a comparative genome analysis approach of *Amoebophrya* strains with different host-range spectrum could provide insights into the evolution of parasitism in dinoflagellates.

Using a combination of Illumina MiSeq paired-end short-read and ONT MinION long-read sequencing technologies, we sequenced and assembled the high-quality genomes of two additional *Amoebophrya* strains (A120 and A25; Table S1), belonging to the MALV-II clade 2 (following the nomenclature by Guillou et al. [11]), and able to infect the same non-toxic autotrophic dinoflagellate *Scrippsiella acuminata* species. Despite a low SSU rDNA sequence divergence (Fig. 1A, Table S2-S3) and temporal co-occurrence of these two parasites, A25 displays a natural host-range restricted to one species while A120 is able to infect a wider range of hosts belonging to at least two dinoflagellates genera. Both genome assemblies (115.5 and 116 Mb, Table S3) were consistent in size with estimates obtained by *k*-mer analysis (113.59 and 118.57 Mb, Fig. S1) and flow cytometry (125.25 ±5.24 and 131.60 ±5.39 Mb) for A120 and A25, respectively (Suppl. Text 1), displaying better assembly metrics than the draft genome of the AT5 strain [16]. We found the A120 assembly is harbored almost twice as many repetitive elements compared to A25 (23.8% versus 13.1%; Fig. S2). The majority of those repeated elements remains unclassified. Additionally, both genomes contain a diversity of autonomous transposable elements corresponding to several retro-element families, including long terminal repeat (LTR) and non-LTR retrotransposons. We predicted 26,441 and 28,091 protein-coding genes, 59.3% and 63.7% with a functional assignment, for A120 and A25, respectively (Table S3,). These relatively small *Amoebophrya* genomes with high gene density values (232 and 248 genes/Mb in A120 and A25, respectively) are reminiscent of the genomes of other parasites such as *Perkinsus marinus* (Table S3, Suppl. Text 2), but remains far smaller than the predicted ∼4.8 Gb genome of the Syndiniales *Hematodinium* sp., parasite of the Norway lobster [17].

Amoebophrya genomes contain far fewer gene families (that include ≥2 genes) than described in *Symbiodinium* spp. (*i.e.*, 25% in *Amoebophrya* vs. 55-65% in *Symbiodinium* spp.), Table S3). In contrast, we observed a stronger tendency for genes co-orientation than in *Symbiodinium* spp. (Fig. S5). We find no correlation between gene co-orientation and their function nor their expression profiles over the full *Amoebophrya* life-cycle (Suppl. Text 5, [19]). We identified a truncated dinoflagellate-specific spliced leader (DinoSL) motif (Fig. S3, Suppl. Text 4) at the 5'-end of at least 37.8% (A120) and 18.5% (A25) of gene transcripts, together with a single complete DinoSL 22-nucleotide (nt) SL-like coding sequence in each genomes (Fig. S4). Despite high levels (96-97%) of SSU rDNA sequence similarities, we only identified 8,118-9,490 orthologous genes between AT5, A120 and A25, representing 36-47%

of the total number of predicted protein genes (Fig. S6). These orthologs shared 48.2-51.2% amino acid sequence identity on average, a level similar to that observed when comparing each *Amoebophrya* strain with *Symbiodinium* spp., *P. marinus* and *P. falciparum* (Fig.1B). These observations suggest a rapid protein sequence divergence in *Amoebophrya*, potentially linked to parasitism, as it often coincides with relaxed functional constraints, leading to higher substitution rates [20]. Beside their genetic distance at the protein level, A25 and A120 genomes exhibited a strong synteny with 64% of orthologous genes (6,908 out of 9,490) clustered into 196 syntenic regions representing 84% (A120) and 80% (A25) of the total number of genes (Fig. 1C). We found rather high levels of synteny of orthologous genes between AT5 and our strains (57% for A120 and 49% for A25, Figs. S7-S8). Strong synteny between A120 and A25, and AT5 to a lesser extent, might suggest a strong evolutionary constraint on gene order in this clade. Conversely, no synteny was observed within the *Symbiodinium* spp. complex, where several species were recently reclassified into separated genera [18]. Alternatively, the combination of elevated high levels of synteny or collinearity, along with a rather high SSU rDNA sequence similarity, are hallmark of a relatively recent speciation event, which is contradicted by the overall low level of protein similarity among *Amoebophrya* species.

More surprisingly, we observed a high proportion (66-67%) of non-canonical introns (NCIs) in both the A25 and A120 genomes (Tables S3-S4, Fig. S9, Suppl. Text 6), unlike AT5 where most introns (99.98%) were predicted to be canonical, i.e., with GT-AG splice sites [16]. When looking at orthologous genes in both *Amoebophrya* A120 and A25 strains, we found that 98.6% of introns at conserved positions displayed canonical splice sites, corresponding to 19.4% and 19.9% of total introns, respectively. Furthermore, we observed a positive correlation between the increased portion of conserved introns and the level of protein similarity between orthologous pairs (Fig. S10-S11),

A deeper investigation revealed that about 30% (A120) and 11% (A25) of NCIs contained direct repeat (DR) motifs of 3-5 nucleotides overlapping the exon/intron boundaries and 8-20 nt inverted repeat (IR) motifs (Fig. 2A, Figs S12-S16, Suppl. Text 6). Comparatively, we observed a similar organization in 15% (A120) and 1% (A25) of canonical introns. IR motifs can produce hairpin structures (Fig. 2A and 2B) allowing the joining of exon boundaries, a mechanism that might be involved in RNA splicing (Fig. 2A). Structurally, the presence of IR and DR along with the absence of internal transposase-encoded genes in repeated elements is reminiscent of non-autonomous terminal inverted repeat (TIR) DNA transposons, where the TIR represents a unique hallmark for each DNA transposon family. Using hidden Markov model (HMM) based profiles obtained from an initial set of IR motifs from both genomes (Fig. S17), we detected 29,850 (68% of NCIs) and 2,039 (20%) repeated introns, hereafter called introners (Fig. S18). We further classified these repeated NCIs into 1,954 and 252 families in the A120 and A25 genomes, respectively (Table S5). Their narrow length distribution and GC percentage range values regrouped them into one (in A25) or two (in A120) populations (Fig. S19, Suppl. Text 6). Clustering analysis grouped these introners into strain-specific clades suggesting strain-specific

amplification (Fig.S18, Suppl. Text 6). We identified several identical pairs of NCIs, in each *Amoebophrya* genome (97 in A120; 64 in A25) suggesting that the spread of introners is ongoing. The remaining NCIs generally have similar DR and IR structure as the introners (24,976 and 28,467 in A120 and A25, respectively), but no relationship between them could be found, and hence remain singletons. Altogether, introners represent 17 and 8% of the genome in A120 and A25, respectively.

Both the unconventional intron splice site [9] and identically repeated intron boundary (IRIB) sequences [21] have been described in dinoflagellates before, but not to such an extent and diversity of repeated motifs, like those described here. Structure-wise, *Amoebophrya* introners are similar to the fixed-lengths DR- and TIR-containing canonical introners described in the genome of the green alga *Micromonas pusilla* and the Stramenopiles *Aureococcus anophagefferens* [22]. Nevertheless, *Amoebophrya* introners are unique with respect to the predominance of non-canonical sites, suggesting the necessity for an alternative to the ubiquitous eukaryotic splicing machinery. Moreover, a broad spectrum of DR motifs and sizes (3-8 nt in length) (Suppl. Text 6) indicates little or no sequence conservation at what we would expect to be the splicing position (based on protein similarity and RNAseq coverage). We identified the near complete eukaryotic canonical machinery expected for processing canonical introns in both *Amoebophrya* genomes, excluding the small nuclear RNA (snRNA) U1 that binds the 5'-donor splice site of introns during splicing (Fig.2B, Figs. S20-S24, Table S6, Suppl. Text 7). Given that snRNAs are highly conserved throughout eukaryotes, *Amoebophrya* snRNA U1 has either diverged significantly to accommodate the novel NCI-borders, or has been completely lost and likely replaced by another rRNA type or a protein equivalent. We were also unable to detect several minor spliceosome subunits (U11, U12, U4atac and U6atac snRNAs), which are involved in the splicing of non-canonical introns [23] in other eukaryotes. We searched for putative DNA transposons in the genomes of the two *Amoebophrya* because non-autonomous TIR-containing DNA transposons are mobilized by transposases encoded by autonomous elements [24]. We identified two putative transposases in A25, but none in A120, ruling out the general transposase-mediated mobilization of *Amoebophrya* introners. These make the *Amoebophrya* introners a novel type of repetitive elements with a still unknown splicing mechanism.

Overall, the *Amoebophrya* genomes presented here display unique characteristics among dinoflagellates and other eukaryotes. The small number of transposable elements, short introns and intergenic regions, and limited number of gene families, all contribute to the high level of compactness and lack of redundancy of their genomes compared to other dinoflagellates. Strong synteny can suggest some evolutionary constraints for the maintenance of gene order through a low rate of chromosomal duplication and rearrangement within *Amoebophrya* species, possibly linked to primary transcriptional mechanistic constraints as described in trypanosomatid protozoa [25]. Given the high level of SSU rDNA sequence similarity between A120 and A25, and AT5 to a lesser degree, recent speciation between A120 and A25, may alternatively have been driven by evolutionary processes that accumulated

protein sequence divergence while maintaining synteny. Part of the answer may lie in the unusual and actively replicating NCIs predominantly located within less conserved genes in the two *Amoebophrya* genomes described here, and their structural peculiarities. In fact, these NCIs share several unique characteristics including extremely variable DR sequences, and elevated levels of IRs sequence identity and a secondary structure for the formation of secondary stem-loops potentially involved in their mobilization. Hypothetically, these secondary structures could contribute to a certain level of stability of transcripts potentially needed during invasion of the parasitic life cycle. The recurrent absence of the snRNA U1 subunit needed for 5'-donor site recognition in both genomes suggests a splicing mechanism common to both *Amoebophrya* strains that could enable the spliceosome complex to recognize and process unusual intron-exon boundaries, maybe through the recruitment of a protein-based subunit. Such mechanism common to both *Amoebophrya* strains must have predated their divergence, enabling the retention and proliferation of introners. The sequencing of additional *Amoebophrya* genomes might shed light on the origin and spread of NCIs, and the potential impact of NCIs on protein evolution. Those and genomes from other basal Syndiniales shall shed light into the mechanisms underlying the contrasting genome organizations observed in dinoflagellates, from highly compact genomes to relax gigantism.

Acknowledgments:

Funding: This research was funded by the ANR—Agence Nationale de la Recherche, Grant ANR-14-CE02-0007—, the CEA and the Région Bretagne (RC doctoral grant ARED, EK postdoctoral grant SAD).

Author contributions:

LG conceived this study. EB, DM, CAdS, RC and LG collected samples. AA and KL acquired sequencing data. SF, BN, JM, BI, TR, CDS, JMA and BP performed genome assemblies and annotation. SF, BN, BP and JMA worked on the gene and genome analyses. PL, SR, PR, SF, BN and BP analyzed introns. PL, SR and FM analyzed introners and repetitive elements. EK, EC, TB, IF, TT, LG, BP, BN, SF worked on functional annotation. SF, EK, BN, BP, FM, SR, PW and LG wrote the manuscript. All authors edited and approved the final version of this paper. We thank Dr. Julie Koester for the English review of the manuscript and Loraine Guéguen for providing the jbrowse data access.

Competing interests: Authors declare no competing interests.

Materials availability: Strains have been deposited at the RCC (http://roscoff-culture-collection.org/, RCC4398 and RCC4383). Genomes and transcriptomes: http://application.sb-roscoff.fr/project/hapar/

Supplementary Materials:

Materials and Methods

Figures S1-S24

Tables S1-S8

References (1-77)

Fig. 1. SSU rDNA sequence identity, gene orthology and synteny in A120 and A25 genomes.

(A) SSU rDNA sequence identity (in percentage) between *Amoebophrya* A25 (in dark), A120 (in pink) and a selection of other alveolates, including *Amoebophrya* AT5. (B) Violin distribution of the percentage of identity of orthologous genes defined by best reciprocal hits (BRHs) between *Amoebophrya* A25 (in dark), A120 (in pink) and a selection of other alveolates, including *Amoebophrya* AT5. Squares represent median values for each distribution. (C) Dot-plot of the synteny observed between the longest scaffolds for each of the *Amoebophrya* A120 (y-axis, 21 scaffolds) and A25 (x-axis, 53 scaffolds) genomes. For each genome, genes are sorted by their rank on the scaffolds. Each dot represents a pair of orthologous genes defined by BRH. Blue lines highlight syntenic regions.

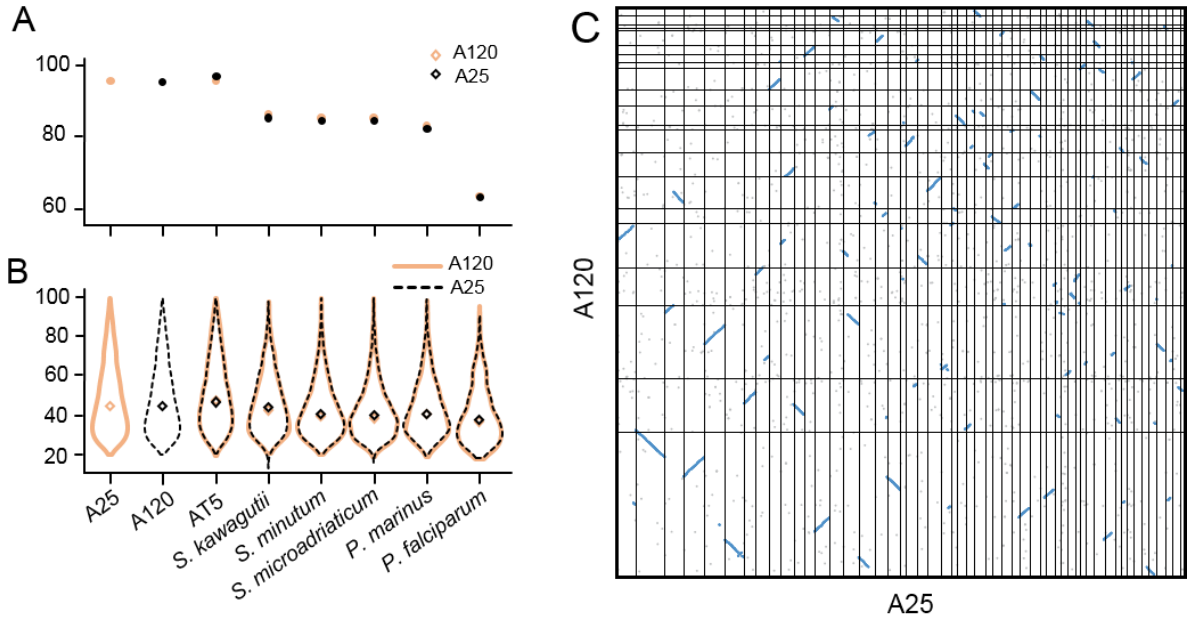
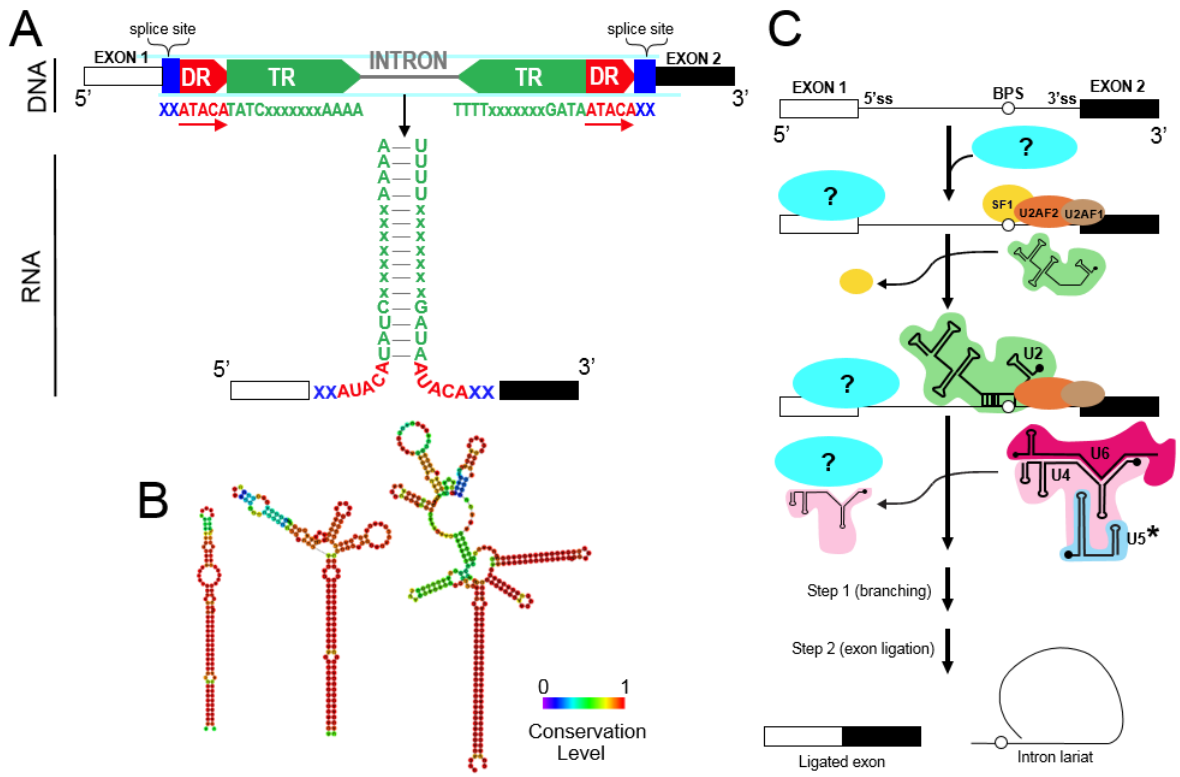


Fig. 2. Predicted hairpin secondary structure of introners and their putative splicing mechanism.

(A) A schematic structure of introners containing direct repeat (DR) and inverted repeat (IR) motifs in the genome (DNA), and (B) the predicted secondary structure (RNA). (C) Homologous proteins putatively involved in the splicing mechanism identified by sequence homology in the A120 and A25 proteomes. In grey with interrogative mark: missing U1 protein in both genomes. *= U5 identified only in A120.



References

1. Taylor FJR, Hoppenrath M, Saldarriaga JF. Dinoflagellate diversity and distribution. *Biodivers Conserv* 2007; 17: 407–418.

2. LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW. Symbiodinium (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. *J Phycol* 2005; 41: 880–886.

3. Hou Y, Lin S. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: Gene content estimation for dinoflagellate genomes. *PLoS One* 2009; 4.

4. Janouškovec J, Gavelis GS, Burki F, Dinh D, Bachvaroff TR, Gornik SG, et al. Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proc Natl Acad Sci* 2017; 114: E171–E180.

5. Bachvaroff TR, Place AR. From stop to start: Tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PLoS One* 2008; 3: e2929.

6. Lidie KB, Van Dolah FM. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. *J Eukaryot Microbiol* 2007; 54: 427–435.

7. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, et al. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci* 2007; 104: 4618–4623.

8. Slamovits CH, Keeling PJ. Widespread recycling of processed cDNAs in dinoflagellates. *Curr Biol* 2008; 18: 550–552.

9. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft assembly of the symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Curr Biol* 2013; 23: 1399–1408.

10. Moustafa A, Evans AN, Kulis DM, Hackett JD, Erdner DL, Anderson DM, et al. Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS One* 2010; 5: e9688.

11. Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, et al. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* 2008; 10: 3349–3365.

12. Cachon J. Contribution à l'étude des péridiniens parasites. Cytologie, cycles évolutifs. *Ann des Sci Nat Zool Paris* 1964; 1–158.

13. Park MG, Yih W, Coats DW. Parasites and phytoplankton, with special emphasis on dinoflagellate infections. *J Eukaryot Microbiol* 2004; 51: 145–155.

14. Montagnes DJS, Chambouvet A, Guillou L, Fenton A. Responsibility of microzooplankton and parasite pressure for the demise of toxic dinoflagellate blooms. *Aquat Microb Ecol* 2008; 53: 211–225.

15. Alves-de-Souza C, David P, Emilie LF, Sébastien M, Cécile R, Behzad M, et al. Significance of plankton community structure and nutrient availability for the parasitic control of dinoflagellate blooms by parasites: a modeling approach. *PLoS One* 2015; e0127623.

16. John U, Lu Y, Wohlrab S, Groth M, Janouškovec J, Kohli GS, et al. An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. *Sci Adv* 2019; 5: eaav1110.

17. Gornik SG, Febrimarsa, Cassin AM, MacRae JI, Ramaprasad A, Rchiad Z, et al. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc Natl Acad Sci* 2015; 112: 5767–5772.

18. Lajeunesse TC, Parkinson JE, Gabrielson PW, Jeong HJ, Reimer JD, Voolstra CR, et al. Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Curr Biol* 2018; 28: 2570–2580.

19. Farhat S, Florent I, Noel B, Kayal E, Da Silva C, Bigeard E, et al. Comparative time-scale gene expression analysis highlights the infection processes of two amoebophrya strains. *Front Microbiol* 2018; 9: 1–19.

20. Johnson KP, Allen JM, Olds BP, Mugisha L, Reed DL, Paige KN, et al. Rates of genomic divergence in humans , chimpanzees and their lice. *Proc R Soc B Biol Sci* 2014; 281.

21. Mendez GS, Delwiche CF, Apt KE, Lippmeier JC. Dinoflagellate gene structure and intron splice sites in a genomic tandem array. *J Eukaryot Microbiol* 2015; 62: 679–687.

22. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. *Nature* 2016; 538: 533–536.

23. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: Splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* 2013; 4: 61–76.

24. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007; 41: 331–368.

25. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al. Gene synteny and evolution of genome architecture in trypanosomatids. *Mol Biochem Parasitol* 2004; 134: 183–191.

Materials and Methods

Origin of strains stock cultures and gDNA extraction

Amoebophrya-like parasites infecting dinoflagellates were detected from surface water samples in the Penzé estuary (North-West of France, English Channel, 48°37'N; 3°56'W, Table S1) thanks to their natural bright green auto-fluorescence using an epifluorescence microscope (BX51, Olympus) equipped with a U-MWB2 cube (450- to 480-nm excitation, 500-nm emission [1]). Infected hosts in late-stage of infection were isolated and incubated in exponentially growing host cultures maintained in an F/2 medium (Marine Water Enrichment Solution, Sigma). The F/2 medium was prepared with filtered and autoclaved natural seawater from the Penzé estuary collected in June one year earlier at 27 salinity and stored in the dark, and complemented with 5% (v/v) local soil extract [2], followed by a final filtration using a 0.22 µm pore-size filter under sterile conditions. We isolated two *Amoebophrya*-like strains (A25 in *Scrippsiella acuminata* ST147 and A120 in *Heterocapsa triquetra* HT150) by sequential (three and six times for A120 and A25, respectively) re-isolation of the parasites in its host. All stock cultures were maintained at 19°C and on a L:D cycle of 12:12 h at 80 µEinstein m2 s−1. Parasite stock cultures were maintained by transferring 300 µL of infected host cultures into 3mL of exponentially growing uninfected hosts every 2-3 days. A120 and A25 were grown in the host *Scrippsiella acuminata* ST147 (prepared in non-ventilated flasks) by transferring the host culture into fresh medium (ratio 1:1) every 2-3 days. A protocol detailing cell harvesting for genomic and transcriptomic analyses can be found at the protocole.io dx.doi.org/10.17504/protocols.io.vrye57w. The genomic DNA was extracted using the kit Nucleospin Plant II following the manufacturer's instructions (Macherey-Nagel, Kit Midi 740771.20 and Kit Mini 740770.250).

Short-read Illumina library preparation and sequencing

DNA was quantified on a Qubit Fluorometer using with the Quant-iT dsDNA Assay Kit (Life Technologies, Carlsbad, California, USA) and its quality checked by electrophoresis in a 0.7% agarose gel. For both strains, an overlapping paired-end (PE) library and a mate-pair library (MP) were prepared for Illumina sequencing. PE overlapping library preparations were carried out from 250 ng of genomic DNA using a semi-automated protocol. Briefly, DNA was sheared with the Covaris E210 instrument (Covaris, Inc., Woburn, Massachusetts, USA) to generate fragments of 150-400 bp in size. End repair, A-tailing and ligation with Illumina compatible adaptors (Bioo Scientific Austin, Texas, USA) were performed using the SPRIWorks Library Preparation System and a SPRI-TE instrument (Beckmann Coulter, Danvers, Massachusetts, USA) according to the manufacturer's protocol. Library fragments 200-400 bp in size were selected and amplified by 12 cycles of PCR with the Pfx Platinum Taq polymerase (ThermoFisher, Waltham, Massachusetts, USA) and Illumina adapter-specific primers. Amplified library fragments were size selected on 3% agarose gel around 300 bp and purified.

For strain A25, a mate-pair (MP) library was prepared according to the initial Illumina protocol (Illumina Mate Pair library kit, Illumina, San Diego, CA) with about 10 µg of genomic DNA subjected to Covaris fragmentation in the first step. For strain A120, the MP library was prepared with the Nextera Mate Pair Sample Preparation Kit (Illumina) using 4 µggenomic DNA that was simultaneously fragmented by enzymatic treatment and tagged with a biotinylated adaptor. The resulting fragmented and tagged (tagmented) DNA was subjected to size selection (8-11 kb) by gel electrophoresis, and circularized overnight by incubation with a ligase. Linear, non-circularized fragments were digested and circularized DNA was fragmented to generate fragments of 300-1000 bp in size with the Covaris E210 system. Biotinylated DNA was immobilized on streptavidin beads, end-repaired, 3'-end adenylated, and ligated with Illumina adapters. DNA fragments were amplified by PCR with Illumina adapter-specific primers and purified.

The quality of all Illumina libraries was evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) and quantified by qPCR with the KAPA Library Quantification Kit (KapaBiosystems Inc., Woburn, MA, USA) on a MxPro instrument (Agilent Technologies). Libraries were sequenced using 101 bp PE reads chemistry on a HiSeq2000 Illumina sequencer.

We cleaned all Illumina PE and MP reads in a four-step process using fastx_clean (http://www.genoscope.cns.fr/fastxtend), an internal software based on the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), by discarding: i) sequencing adapters and low-quality nucleotides (quality value < 20); ii) sequences located between the second unknown nucleotide (N) and the end of the read; iii) reads shorter than 30 nucleotides after trimming; iv) reads and their mates mapping onto run quality control sequences (the PhiX genome).

Long-read Nanopore library preparation and sequencing

Genomic DNA was first size selected (10-50 Kb for both organisms and 20-80 Kb cut-offs for A120 only) using a BluePippin (Sage Science, Beverly, MA, USA) and repaired depending upon the DNA quantity recovered using the NEBNext FFPE Repair Mix (New England Biolabs, Ipswich, MA, USA). Following end-repair and 3'-A-tailing with the NEBNext® Ultra™ II End Repair/dA-Tailing Module (NEB), sequencing adapters provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, UK) were ligated using Blunt/TA Ligase Master Mix (NEB). Each library was then mixed with the Running Buffer with Fuel Mix (ONT) and the Library Loading Bead (Oxford Nanopore Technologies) and loaded on MinION R9.4 SpotON Flow Cells. Two and three libraries were run for the A25 and A120 strains, generating a total yield of 2.5 Gb and 14 Gb, respectively (Table S2).

Read event data were generated by the MinKNOW control software (versions 1.3.25, 1.3.30 or 1.4.3) and base-calling done with the Metrichor software version 2.43.1 or 2.45.3 (1D Basecalling RNN for LSK108 workflow). The data generated (pores metrics, sequencing, and base-calling data) by the MinION softwares were stored and organized using a Hierarchical Data Format. FASTA reads were extracted from MinION Hierarchical Data Format files using poretools [3].

Genome size estimation

We estimated the genome sizes of the two parasite strains using both flow cytometry and k-mer analysis. For the flow cytometry, we extracted nuclei by mixing 50 µL of freshly produced dinospore with 450 µL of 0.25X NIB buffer [4], containing SYBR Green-I at final concentration of 1/5,000. We used 2 µL of a culture of exponential growing *Micromonas pusilla* RCC299 (1C = 20.9 fg) as an internal reference. The mixture was then incubated at least for 30 min in the dark, before being analyzed using a FACS Canto II flow cytometer equipped with a 488 nm laser and the standard filter setup, where signal was triggered by green fluorescence. The ratio between the mean distribution of the dinospores and the RCC299 was used for the evaluation of the DNA content.

For k-mer size estimation, we analyzed the k-mer distribution of Illumina 100 bp paired-end reads in order to derive an independent estimate of the haploid genome size of the parasites. We used Jellyfish [5] with the following parameters: -m 31 -s 2048M –C to generate a 31-mer distribution and the k-mer histogram was uploaded to the GenomeScope website (http://qb.cshl.edu/genomescope/).

Genome assembly

We used both short Illumina and long Nanopore reads to generate genome assemblies for the two *Amoebophrya* strains. First, we obtained a draft Illumina-based assembly from the combination of

Illumina paired-end and mate pair reads using the All-PathsLG [6] program with default parameters. Gaps were closed using GapCloser from the SOAPdenovo package [7]. In order to detect and remove chimeric junctions that are present in Illumina scaffolds, we aligned Nanopore reads on the Illumina assemblies using the Last aligner package [8]. Then, we used NanoSV [9] to detect any mis-mapping in reads that could indicate a chimeric scaffold. Finally, we cut the scaffolds sequence at each breakpoint indicated by NanoSV.

Second, we generated a Nanopore-only draft assembly for each genome. For A25, we used all Nanopore reads (corresponding to an estimated 23x genome coverage) as inputs to the SMARTdenovo assembler (Jue Ruan, Ultra-fast de novo assembler using long noisy reads, 2016, available at https://github.com/ruanjue/smartdenovo) using the –k 17 parameter to increase k-mer size (as advised by the developers on large genome sizes) and –c 1 to generate a consensus. For A120, we selected the longest Nanopore reads corresponding to an estimated 30x (out of 120x) coverage of the genome as input to the SMARTdenovo assembler as previously described [10][11] with the –k 17 and –c 1 parameters. Then, we aligned the Illumina short reads onto the Nanopore assemblies using bwa mem [12] in order to correct non-random mainly homopolymeric Nanopore errors, and gave the resulting alignments as input to Pilon [13] in order to correct the consensus of the Nanopore-only assemblies.

Finally, we decided to preserve the original Illumina scaffolds by organizing them into super-scaffolds based on the Nanopore-only assemblies. We aligned the Illumina scaffolds of each genome onto its respective Nanopore-only assembly using Nucmer [14] and kept only the best match with the delta-filter command. We considered a match only if the alignment covered more than 90% of the Illumina scaffold with at least 85% identity. Thanks to this list of matches, we organized the Illumina scaffolds along the Nanopore assemblies as the final assembly for gene annotation.

Detection of genome repeat sequences for gene prediction

Most of the genome comparison analyses described below were performed with repeat-masked sequences. To do so, we searched several kinds of repeats in parallel using the following tools: the RepeatMasker program version 3.3.0 (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015, http://www.repeatmasker.org) to look for alveolates known repeats and transposable elements included in the RepBase database [15]; the TRF program version 4 [16] for the tandem repeats; the DUST program [17] for low complexity repeats. In parallel, we also performed an *ab initio* detection of repeat patterns with the RepeatScout program [18].

Transcriptome assembly

We filtered the raw transcriptome data from a previous study [19] in order to remove clusters with too much intensity and ribosomal RNA-like reads were excluded using the SortMeRNA program [20]. All reads from each time point were pooled before producing transcriptome assemblies for several life stages of each parasite using oases v. 0.2.08 [21] with a k-mer size of 51. We cleaned the assemblies with dustmasker from the ncbi-blast-2.2.27+ toolkit [17] and trimmed the 5’ and 3’ low-complexity ends. Assembly statistics are shown in Table S7& S8.

Contigs longer than 150 bp and containing more than 75% of unmasked nucleotides from all transcriptomes were kept and used for the gene prediction of each genome separately.

Gene prediction

A first attempt to align the transcriptomes against the assembled genomes revealed an unusually high rate of non-canonical splice sites, rendering the use of classical mappers and *ab initio* gene prediction softwares unfit for annotating the *Amoebophrya* genomes. A customized annotation pipeline was developed to take into account the non-canonical intron whose splice sites have been confirmed by the RNA-seq data. Transcriptomes of the several life stages of the parasites were mapped onto the genome assemblies using the EST2GENOME software. Given that EST2GENOME expects canonical GT/AG splicing sites, we explored the possibility of alternative exon-intron boundaries by aligning the transcripts to the genome assemblies with BLAT (≥ 90% sequence identity and ≥85% aligned query length), keeping only the best match per transcript (Table S5).

We also aligned 456,355 alveolate proteins downloaded from the UniProtKB [22] databank (9/2014) to the genome assemblies using BLAT [23]. Subsequently, we extracted the genomic regions without protein hit and realigned the Uniprot proteins with more permissive parameters using BLAST [24]. Each significant match was then refined using Genewise [25] in order to refine exon/intron boundaries. Given that Genewise settings use canonical spliced sites model, these protein alignments were essentially used to finding open reading frames (ORFs).

Alignments of *Amoebophrya* assembled transcripts and conserved proteins were used as input to Gmove (www.genoscope.cns.fr/gmove), an in-house combiner program, to predict gene models for both A25 and A120 strains. Ab initio gene prediction software was not used for gene model prediction, due to the large amount of non-canonical introns observed in both A25 and A120 *Amoebophrya* genomes. Briefly, putative exons and introns boundaries extracted from the alignments were used to build a simplified graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched ORFs consistent with the protein alignment evidences. Finally, a selection step was made on all candidate genes based on gene structure, where the model with the longest (>100nt) ORF per coding locus was selected. We removed all intron-less genes (ORF < 300nt in size) as well as overlapping spliced genes.

We assessed the quality of our gene prediction approach using the Eukaryota database of the Benchmarking Universal Single-Copy Orthologs (BUSCO v2 Eukaryotic dataset, [26]) and by remapping RNA-seq reads.

Functional annotation

We defined INTERPRO domains for both *Amoebophrya* proteomes using InterProScan [27] and predicted the most probable function for each gene as described elsewhere [19]. In order to ensure reproducibility of our annotation approach, we re-annotated the proteomes of the coral symbiont *Symbiodinium kawagutti*, the malaria parasite *Plasmodium falciparum* and the perkinsozoan *Perkinsus marinus* using the same method. We then scored the completeness of KEGG pathways in each organism by estimating the fraction of predicted enzymatic reactions present in the query organism when compared to the canonical pathways defined by the KEGG database using the KEGG MODULE reconstruction pipeline with default parameters [28]. We checked missing annotations of the major metabolic pathways in our genomes by comparing them to those of *Toxoplasma gondii* obtained from the (Liverpool) Library of Apicomplexan Metabolic Pathways (LAMP; http://www.llamp.net/) and of *P. falciparum* obtained from the Parasite Metabolic Pathways (MPMP; http://mpmp.huji.ac.il/). We validated the identity of candidate genes by the presence of functional domains and sequence alignments with closely related proteins.

Orthologs and synteny in the genomes of A25 and A120

We conducted gene family analysis by comparing the predicted proteomes of both *Amoebophrya* strains with those of 12 other protist species: *S. kawagutii* ([29]; http://web.malab.cn/symka_new/), *S. minutum* ([30]; http://marinegenomics.oist.jp/symb/viewer/info?project_id=21), *S. microadiaticum* ([31]; http://smic.reefgenomics.org/), *P. marinus* (http://protists.ensembl.org/Perkinsus_marinus_atcc_50983/Info/Index), *P. falciparum* 3D7 ([32]; http://plasmodb.org/plasmo/), *T. gondii* ME49 strain ([33]; http://toxodb.org/toxo/), *Trypanosoma brucei* TREU 927 strain ([34]; http://tritrypdb.org/tritrypdb/ release 9.0), *Leishmania major*; http://tritrypdb.org/tritrypdb/), *Theileria equi* ([35]; http://eupathdb.org/), *Chromera velia* CCMP 2878 ([36]; http://eupathdb.org/), *Vitrella brassicaformis* CCMP 3155 ([36]; http://eupathdb.org/) and *Cryptosporidium parvum* ([37]; http://cryptodb.org/cryptodb/). We performed all-against-all BLASTp (e-value = 1e-5; min. alignment length of the shortest protein = 50%) searches using the NCBI Blast+ 2.2.28 package between the 14 proteomes, and clustered the proteins into OrthoGroups (OG) using a Markov cluster (MCL 14-137) algorithm [38].

Orthologs and synteny in the genomes of A25 and A120

We used the Smith-Waterman algorithm ([39]) (BLOSUM62, gapo=10, gape=1) to build pairwise protein alignments for both *Amoebophrya* strains, *S. kawagutii*, *S. minutum*, *S. microadiaticum*, *P. marinus*, and *P. falciparum* 3D7, and retained all alignments with a score >300. From these alignments, we defined orthologous and paralogous genes between *Amoebophrya* and other alveolate species using a Best Reciprocal Hits (BRH) approach. We represented orthologous genes between A25 and A120 as a dot-plot graph according to their locations on the genome assemblies where clusters of genes (composed of at least five genes) were used to define syntenic regions. For a syntenic region to be valid, we defined a maximum distance of fifteen genes between two genes of the same cluster.

Tandem duplication detection

Finally, we detected tandemly duplicated genes by comparing the protein sequences of predicted genes in each genome, where only highly similar pairs were retained (identity percent ≥95% at protein level with a minimum alignment length of 90% of the total). Then, we grouped proteins together according to their similarty values using single a linkage clustering algorithm. For each cluster, two genes were defined as co-localized if they were contiguous by their rank (i.e. genomic location) on the genome, where only one gene without match against the genes in the same cluster was allowed between the pair.

Co-orientation

Finally, we computed the distribution of gene orientation changes using a non-overlapping 10 genes sliding window [31]. We defined co-oriented gene blocks of at least five contiguous genes (based on their rank along the genome sequences) with the same orientation with a maximum of two contiguous genes in an opposite orientation.

Detection of spliced-genes

In order to identify putative trans-spliced genes in *Amoebophrya*, we searched the 3'-end 16 nt of the dinoflagellate spliced leader (dinoSL) sequence [40] in the RNAseq data using a k-mer approach with kfir (www.genoscope.cns.fr/kfir) and a k-mer size equal to 8. The reads containing the dinoSL-like motif were aligned against their respective genome assembly using bwa mem [12]. Only the reads containing the last 5 nt (TCAAG) of the dinoSL were later selected among the soft-clipped part of the

alignments. Trying to define the SL sequence of *Amoebophrya* A120 and A25 strains, we extended up to 13 nt upstream toward the 5'-end soft clipped position in the genome without divergence from the dinoSL consensus sequence (Fig. S3). The first match after the soft-clipped region in the RNA-genome alignment was considered as the putative SL junction . If the two last bases before this position didn't correspond to the 3'-end dinoSL 'AG' dinucleotides, the putative SL junction was shifted upstream while the dinoSL sequence was verified. We used a multiple sequence alignment to define the SL sequence for the *Amoebophrya* A120 and A25 strains.

We then compared the locations on the genome assemblies of these putative SL junctions with gene predictions. A putative SL junction was associated with a gene either if it overlapped the 5' UTR region of the corresponding gene or the first coding exon. The putative SL junctions located in intergenic region were linked to the nearest gene models.

Intron analyses

We obtained RNA-seq validated intron sequences with Hisat2 (--very-sensitive --qc-filter --max-intron length 10000; [41]) and Regtools (junctions extract -a 8 -i 40 -I 10000). Only introns validated with minimum coverage of three RNA-seq sequences at the splice-junctions and a length window of 40-1000 bp were used for further analyses. We used a consensus canonical motif to differentiate canonical introns from non-canonical introns. The non-canonical (GT/AG) introns were compared to each other using BlastN (all-against-all, E value=1e-5; [24]) and clustered using orthoMCL (I=5, [32]). All intronic sequences from each cluster were subsequently aligned with Muscle (v. 3.8.31, -diags) [42]. We used Patscan tool v20110223 [43] to identify conserved palindrome motifs (referred to as inverted repeats, IRs) around the splice sites,. We then regrouped NCIs into families based on their IRs (100% identity in sequence composition and length) and intronic (identity>=30%) sequences using the CD-hit tool. We constructed HMM profiles for each repeated NCI (rNCI or intron) family using hmmbuild (E value=1e-5) from the HMMer v. 3.1b package [44]. To classify the super families of introners, we used hierarchical clustering (hclust, method=euclidean, ward.D) in R (v 3.2.2). We estimated the percent identity and the length of the introners using the 'Needle' sequence aligner from the Emboss v. 6.1.0 package [45] and analyzed the median percent identity and length using the ggplot2 and ggdendro scripts from the R packages.

Spliceosome component

We identified orthologs of the small nuclear ribonucleoproteins (snRNPs) in the proteomes of A120 and A25 using the Markov cluster MCL 14-137 algorithm [38] with queries from *P. falciparum*, *T. gondii* and *H. sapiens* [46][47].

We also searched for the U1, U2, U4, U5 and U6 snRNAs in both *Amoebophrya* genomes and transcriptomes using BLASTn [24] searches with the default parameters with homologues from *P. falciparum*, *S. minutum*, *H. sapiens* and *Saccharomyces cerevisiae* as queries. For each positive match from the transcriptomes, a BLASTn search against the uninfected host RNA-seq sample was performed in order to eliminate transcripts belonging to the host. The remaining predicted snRNA sequences were verified by remapping genomic reads using bwa mem. Finally, the secondary structure of each snRNA was predicted with RNAfold tool from the Vienna RNA package [48].

Conserved introns between orthologous genes

We compared intron position conservation between orthologous genes for *Amoebophrya* A120 and A25 by building homologous protein gene alignments with Muscle v3.7 [42] and filtering out highly variable positions with Gblocks (v0.91b). We tagged the last amino acid of each spliced exon in the alignements, and considered any intron as conserved if it was present at the same location in the two ortholog proteins, in the same phase and in a conserved block in the alignment.

Transposable elements

We annotated repetitive elements in the *Amoebophrya* genomes using the REPET package [49]. We also built libraries of consensus sequences representative of repetitive elements found in A120 and A25 assemblies separately using the TEdenovo pipeline [49], and used these libraries to annotate similar regions in the assemblies using the TEannot pipeline [50].

We searched for putative transposase genes that may mediate the movement of repetitive elements by building a library of conserved protein domains belonging to DNA transposons from the Repbase database [51], and used this library as query to search A25 and A120 assemblies by reverse position-specific (RPS) BLAST searches. We also used detectMITE [52] to identify the putative MITE elements in two genomic scaffolds.

Supplementary Texts

Suppl. Text 1: Genome sizes and genome assemblies

We used a combination of short-read Illumina and long-read Nanopore sequencing technologies in order to assemble the genomes of the two *Amoebophrya* strains A120 and A25. The final assemblies were 115.5 and 116 Mb in size for A120 and A25, respectively, which are rather close to genome sizes estimated by the kmer analysis (Fig. S1) and FACS (Fluorescence-activated cell sorting) analysis (118.6 Mb and 125.25 ±5.24 for A120, and 113.6 Mb and 131.60 ±5.39 Mb for A25, respectively). These two Syndiniales genomes reported here are smaller than those of *Symbiodinium* species, already described as the smallest genomes (~1.5-4.8 pg/cell) in the dinoflagellate's clade [53], but similar to that of *Perkinsus marinus* in size (Table S3). In comparison, the genome assembly of *Amoebophrya* AT5 was 87.7 Mb, for a genome size estimated by flow cytometry of ~120 Mb [54]. The resulting assemblies of A120 and A25 were composed of 50 and 557 scaffolds with an N50 of 9.24 Mb and 1.08 Mb, respectively (Table S3). The longest scaffolds are 16.51 Mb in A120 and 3.01 Mb in A25. Assemblies are complete enough to detect telomeres at the end of few scaffolds. Telomeres display highly conserved motifs over evolution, yet we identified two sets of telomere-like repeat elements in the *Amoebophrya* genomes: the plant-like TTTAGGG/TTTGGGG motifs in A120, observed in Dinophyceae [55] and a novel TTTGGGA motif in A25.

Suppl. Text 2: Number of genes over dinoflagellates

The 26,441 and 28,091 (for A120 and A25, respectively) predicted genes, as well as the less numerous 19,925 genes predicted for the AT5 strain [54], were in similar in numbers than the 23,654 genes described in *P. marinus*. By comparison, *Symbiodinium* species (excluding *S. kawagutii*) harbor a larger number of multi-exonic genes (~43,000-53,000 predicted genes), which are also longer. All *Amoebophrya* and *P. marinus* genomes are more compact (232-273 genes/Mb) than those of other dinoflagellates, where gene density range in the smaller *Symbiodinium* genomes is around 39-69 genes/Mb. However, the differences in predicted gene numbers between *Amoebophrya*, *P. marinus* and *Symbiodinium* cannot be explained by their parasite lifestyle given that *Amoebophrya* AT5 genomes encode the nearly full metabolic pathway of a typical heterotrophic organism [54].

Suppl. Text 3: Reduced organelles in Amoebophrya

As reported for AT5 [54], we found no evidence of a mitochondrial genome nor trace of a vestigial plastid in the *Amoebophrya* A120 and A25 genomes. We observed the concomitant loss of several plastidial metabolic pathways, such as the plastidial alternative non-mevalonate (DOXP) pathway for the synthesis of isoprenoids [56] as well as the fatty acid elongation (FASII) pathway. Similarly, we were unable to identify genes encoding the squalene synthase and the squalene monooxygenase involved in the production of the squalene 2,3-epoxide from isopentenyl pyrophosphate (IPP) pathway retained in apicoplasts [57]. These observations advocate for the total absence of a plast-like organelle in *Amoebophrya*, suggesting at least two independent complete losses of plastids in Alveolates, one leading to the *Cryptosporidium* lineage and the other to Syndiniales before the divergence of *Amoebophrya* and *Hematodinium* [58].

The mitochondrial genome of myzozoans is drastically reduced and encodes only two (*cox1* and *cox3* in *Chromera velia*) to three protein-coding (*cox1*, *cox3* and *cob* in the rest), and fragments of ribosomal RNA (*rns* and *rnl*) genes, the rest of the genes encoding for the components of the oxidative phosphorylation (OXPHOS) pathway having been transferred to the nucleus [59][60]. Despite intensive search in both genomic and transcriptomic resources generated for both *Amoebophrya* strains, we were

unable to identify the two (*cox3* and *cob*) canonical mitochondrial-encoded protein genes, as reported for AT5 [54]. We did find two partial candidate sequences with homology to *cox1* (*cox1a* and *cox1b*) containing two transmembrane helices each, where *cox1a* predicted peptides aligned to the iron-binding site. Those putative ORFs were a part of the nuclear genome assemblies and located on the assembled transcripts of other larger nuclear genes in A120 (GSA120T00019965001 and GSA120T00004436001 for *cox1a*; GSA25T0001315001 for *cox1b*), suggesting that these sequences are likely NUMTs (Nuclear copies of Mitochondrial DNA).

Suppl. Text 4: Trans-Splicing in Amoebophrya

Dinoflagellates and kinetoplastids are thought to share several genomic features [61] including the addition of a splice leader (SL) at the 5' end of gene transcripts, and the presence of polycistronic mRNAs, though a more recent study has challenged the latter assumption [62]. The SL trans-splicing appears to be ubiquitous in dinoflagellates [63], including in Perkinsozoa [64]. Starting from the 22-nucleotides dinoflagellate SL (dinoSL) sequence [65], we identified a 13-nucleotides conserved sequence corresponding to the 3'-end of the dinoSL at the 5'-end of 37.8 and 18.5% of A120 and A25 genes (Fig. S3). At the same time, we identified one putative SL-encoding gene in each *Amoebophrya* genome (Fig. S4), not linked to the spliceosomal gene clusters as described in dinoflagellates [65]. Trans-splicing has been linked to the resolution of putative pre-mature polycistronic pre-mRNAs and mRNA stability in several lineages [66]. For instance, genes organized in directional clusters in kinetoplastid genomes are transcribed into polycistronic mRNAs [67]. However, the absence of evidence for polycistronic mRNAs [62] of unidirectional clusters of genes [68] challenges such role in dinoflagellates.

Suppl. Text 5: Co-orientation of genes

In all *Amoebophrya* genomes, genes were packed into long co-oriented chromosomic regions: 98.5% into 516 blocks in A120; 98.1% into 587 blocks in A25; 83% into 1245 blocks in AT5. The average shift of orientation for a 10-genes window was lower in AT5 (0.93) compared to the other two *Amoebophrya* strains (about 0.15 and 0.17 in A120 and A25, respectively), but remained higher than what has been described in most *Symbiodinium* genomes (2.32 for *S. microadriaticum*, 2.11 for *S. kawagutii*, and 0.64 for *S. minutum*; Fig. S5). This indicates a tendency for clustering unidirectional genes in dinoflagellates. In *Amoebophrya* A25 and A120, these co-oriented genes do not have similar gene expression nor similar functions. One predicted benefit of the gene co-orientation is the reduction of potential conflicts between replication and transcription [69], that may be dictated by an absence of temporal separation between DNA transcription and replication over the whole infection cycle of both *Amoebophrya* strains [19].

Suppl. Text 6: Non-canonical introns in Amoebophrya *genomes*

Focusing on introns supported by RNA-seq (coverage >3), *Amoebophrya* genomes consist of 66,565 and 55,290 introns in A120 and A25 respectively, providing a similar density (1.42 and 1.47 introns per kilobase of coding sequence in A120 and A25, respectively) with what is commonly observed in alveolates and eukaryotes in general [70]. Interestingly, more than 60% of those introns were non-canonical introns (NCIs; Table S4), where the donor-acceptor site differed from the canonical GT-AG pair (Fig. S9). Instead, 29,850 (A120) and 2,039 (A25) of NCIs contained a long (<20 nt) repeated motif at a variable (3-5nt) nucleotide distance from the presumed exon/intron border, forming a complementary sequence between the 5'- and the 3'-end of the same intron (Fig. S12). While the

repeated motifs were different between A120 and A25 NCIs (Fig. S12), they have the potential to form a secondary hairpin structure bringing adjacent exons close to each other (Fig. 2A). We defined introners as belonging to NCI families sharing high sequence similarity (Table S5, Fig. S18) and containing inverted repeat (IR) regions (Fig. 2A), where the IR sequence is likely involved in the formation of a hairpin structure that would bring the donor- and acceptor sites together. We identified 29,850 and 2,039 NCIs members clustered into cluster 1,954 and 252 distinct families in A120 and A25 respectively, with 24,976 and 28,467 singleton IR-containing NCIs without homology to any introner family. For instance, IR family motifs in A120 started with TAT followed by seven less conserved nucleotides and ending with a minimum of three conserved A (Fig. S18). In A25, IR family motifs started with the conserved TTA motif followed by two purines (A or G) and ended with a conserved G (Fig. S18). The introners grouped into two distinctive populations of different length (~130 nt and 260 nt) and GC content in A120 but only one group of homogenous length distribution (~110 nt) and GC content in A25 (Fig. S19). Among the introner families, we identified, 97 (in A120) and 64 (in A25) pairs that shared 100% sequence identity. All of these observations suggest that the proliferation of introners started from a common origin maybe independently in each *Amoebophrya* strain, and that the process is still ongoing, in a fashion arguably similar to transposable elements.

By comparing orthologous genes between the two strains, we identified that only 19.4 and 19.9% of introns in A120 and A25, respectively, occurred at the same exact locations, 98.6% of which displaying canonical splice site. Globally, the number of these conserved introns (introns that are at the same position between the 2 orthologs) increased with the level of sequence similarity between orthologous proteins (Fig. S10). Comparatively, only 24.8 and 32.6% (in A120 and A25) of species-specific intron positions (found in one strain but not in the other) displayed the canonical GT-AT splice. We identified a slight bias in the distribution of introners in favor of genes involved in translation and ribosome (GO:03010:Ribosome term, Fig. S11), indicating a potential link existing between sequence divergence and introners.

We identified a short (3-8 nt) unidirectional duplicated sequence (referred to as direct repeats or DRs) flanking the IRs at the expected splice sites. The DRs varied in length, composition, and position (Fig. S12-S15): the most abundant DRs in A120 consisted of four nucleotides upstream of the 5'-end and within two nucleotides downstream of the 3'-end of the IR motifs (Fig. S15); in A25, the most abundant DRs were overlapping the 5'-end and were one nucleotide downstream of 3'-end of the IR motifs (Fig. S15). Given the large diversity of the DR sequences, no specific insertion phases could be identified, suggesting a likely uncommon insertion mode in *Amoebophrya*. The absence of identifiable NCIs in the AT5 strain should be hampered by the lower quality of the genome assembly [54].

DNA transposons can degenerate into non-autonomous transposable elements (commonly known as miniature inverted-repeat transposable elements or MITEs) that display often short (≥10 bp) DRs (resulting from target site duplications or TSDs) and IRs but lack transposase genes. Instead, MITEs rely on the activity of transposases encoded by cognate full-length autonomous transposons through a cut-and-paste transposition mechanism by recognizing the IR motifs for mobilization. MITEs have been detected in numerous eukaryotes including some plants, fungi, protozoans, metazoans [71][72], and more recently in viruses [40]. We attempted to classify *Amoebophrya* introners using the current classification of MITEs. We found that only a small proportion of introners (31% and 10% for A120 and A25, respectively) could be assigned to putative but unknown MITE families, and no family-specific IR motifs could be detected.

Suppl. Text 7: Spliceosome components

Eukaryotic genes are transcribed as precursor mRNAs (pre-mRNAs) that are maturated into mRNAs by the spliceosome, a multimega-dalton ribonucleoprotein (RNP) complex, via removal of noncoding

sequences (introns and UTR regions) that results in the joining of the coding sequences (exons) together and the addition of 5'- and 3'-end protective regions [73]. Two spliceosome complexes coexist in most eukaryotes: the major spliceosome which catalyzes canonical GT/AG introns and the minor spliceosome which catalyzes AT/AC introns [74], the latter spliceosome occurring at low frequency in diverse eukaryotic taxa [75]. The major spliceosome is composed of five well conserved small nuclear ribonucleoproteins (snRNPs), each including a uridine-rich snRNA, a common set of proteins and a variable number of particle-specific proteins [73]. We identified all known spliceosomal proteins but 6 in *Amoebophrya* A120 and A25 strains (Table S6, Fig. 2B). Missing spliceosomal proteins in A120 and A25 have roles in the U4/U6 complex (snRNP27), in U5 complex (CD2BP2), in the specification of U5 and interactions with RNA (BCAS2, SYF2), in the SR (a protein with domain having long repeats of Serine and arginine amino acid Residues) and hnRNP family (heterogeneous nuclear ribonucleoprotein) (PTBP2 and hnRNP U). hnRNP proteins and SR proteins have an additional putative role in alternative splicing [76]. Absence of such proteins has already been reported in other alveolate parasites, such as hnRNP U in *P. falciparum* [47], CD2BP2 in *T. gondii* [46], and SYF2, a pre-mRNA processing factor, in both.

We additionally identified four out of the five snRNAs excluding U1 (Table S6, Figs. S19-S22), all having conserved secondary structures (Fig. S23). We identified U5 snRNA only in A120. The *Amoebophrya* U2 and U4 snRNA sequences diverge from those of *P. falciparum*, *S. minutum* and *H. sapiens*. A survey of both the transcriptome and the genomes suggest that the genome of A120 encodes at least four copies of U4 while that of A25 encodes at least two, whilst many copies (11 in A25 and 15 in A120) of U6 snRNA were identified in both genome assemblies.

To date, no component of the minor spliceosome has been identified in Alveolata [77]. The apparent absence of key components (U11, U12, U4atac and U6atac snRNAs) of the minor spliceosome along with the very low proportion ($0.4\ 10^{-3}$ in A120 and $0.8\ 10^{-3}$ in A25) of introns with AT/AC splicing site suggest the absence of this complex in the *Amoebophrya* strains A120 and A25.

References

1. Coats DW, Bockstahler KR, Berg GM, Sniezek JH. Dinoflagellate infections of Favella panamensis from two North American estuaries. *Mar Biol* 1994; 119: 105–113.

2. Starr R, Zeikus J. UTEX-The culture collection of algae at the University of Texas at Austin. *J Phycol* 1993; 29: 1–106.

3. Loman NJ, Quinlan AR. Poretools : a toolkit for analyzing nanopore sequence data. *Bioinformatics* 2014; 30: 3399–3401.

4. Marie D, Partensky F, Simon N, Guillou L, Vaulot D. Flow cytometry analysis of marine picoplankton. In: Diamond RA, DeMaggio S (eds). *Living Colors: protocols in cytometry and cell sorting*. 2000. Springer Verlag, pp 421–455.

5. Marçais G, Kingsford C. A fast , lock-free approach for efficient parallel counting of occurrences of k -mers. *Bioinformatics* 2011; 27: 764–770.

6. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS* 2011; 108: 1513–1518.

7. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012; 1: 18.

8. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011; 21: 487–493.

9. Stancu MC, Roosmalen MJ Van, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* 2017; 8: 1326.

10. Istace B, Friedrich A, D'Agata L, Faye S, Payen E, Beluche O, et al. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* 2017; 6: giw018.

11. Schmidt MHW, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De novo assembly of a new Solanum pennellii accession using nanopore sequencing. *Plant Cell* 2017; 29: 2336–2348.

12. Li H, Durbin R. Fast and accurate short read alignment with Burrows – Wheeler transform. *Bioinformatics* 2009; 25: 1754–1760.

13. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon : An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014; 9: e112963.

14. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol* 2004; 5: R12.

15. Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Diversity of retrotransposable elements Repbase update , a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005; 467: 462–467.

16. Benson G. Tandem repeats finder : a program to analyze DNA sequences. *Nucleic Acids Res* 1999; 27: 573–580.

17. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* 2006; 13: 23 Jun.

18. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics* 2005; 21: 351–358.

19. Farhat S, Florent I, Noel B, Kayal E, Da Silva C, Bigeard E, et al. Comparative time-scale gene expression analysis highlights the infection processes of two amoebophrya strains. *Front Microbiol* 2018; 9: 1–19.

20. Kopylova E, Noé L, Touzet H. SortMeRNA : fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012; 28: 3211–3217.

21. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases : robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012; 28: 1086–1092.

22. The Uniprot Consortium. UniProt : a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019; 47: D506–D515.

23. Kent WJ. BLAT — The BLAST -like alignment tool. *Genome Res* 2002; 12: 656–664.

24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; 215: 403–410.

25. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res* 2004; 14: 988–995.

26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. Genome analysis BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015; 31: 3210–3212.

27. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011 : new developments in the family and domain prediction database. *Nucleic Acids Res* 2012; 40: D306–D312.

28. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res* 2019; 47: D590–D595.

29. Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, et al. The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. *Science (80-)* 2015; 350: 691–694.

30. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft assembly of the symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Curr Biol* 2013; 23: 1399–1408.

31. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, et al. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci Rep* 2016; 6: 1–15.

32. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB : a functional genomic database for malaria parasites. *Nucleic Acids Res* 2009; 37: D539–D543.

33. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, et al. ToxoDB : an integrated Toxoplasma gondii database resource. *Nucleic Acids Res* 2008; 36: D553–D556.

34. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB : a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* 2010; 38: D457–D462.

35. Kappmeyer LS, Thiagarajan M, Herndon DR, Ramsay JD, Caler E, Djikeng A, et al. Comparative genomic analysis and phylogenetic position of Theileria equi. *BMC Genomics* 2012; 13: 603.

36. Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* 2015; 4: e06974.

37. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, et al. Complete genome sequence of the Apicomplexan, Cryptosporidium parvum. *Science (80-)* 2004; 304: 441–445.

38. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002; 30: 1575–1584.

39. Codani JJ, Comet JP, Aude JC, Glémet E, Wozniak A, Risler J, et al. Automatic analysis of large-scale pairwise alignments of protein sequences. *Methods Microbiol* 1999; 28: 229–244.

40. Zhang HH, Zhou QZ, Wang PL, Xiong XM, Luchetti A, Raoult D, et al. Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes. *Mob DNA* 2018; 9: 19.

41. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2016; 12: 357–360.

42. Edgar RC. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; 32: 1792–1797.

43. Dsouza M, Larsen N, Overbeek R. Searching for patterns in genomic data. *Genetwork* 1997; 13: 497–498.

44. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 2010; 11: 431.

45. Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. *Trends Genet* 2000; 16: 276–277.

46. Suvorova ES, White MW. Transcript maturation in apicomplexan parasites. *Curr Biol Microbiol* 2014; 20: 82–87.

47. Sorber K, Dimon MT, DeRisi JL. RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions , alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* 2011; 39: 3820–3835.

48. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatshefte ftir Chemie* 1994; 125: 167–188.

49. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in De Novo annotation approaches. *PLoS One* 2011; 6: e16526.

50. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* 2005; 1: e22.

51. Bao W, Kojima KK, Kohany O. Repbase Update , a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015; 6: 11.

52. Ye C, Ji G, Liang C. detectMITE : A novel approach to detect miniature inverted repeat transposable elements in genomes. *Sci Rep* 2016; 6: 19688.

53. LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW. Symbiodinium (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. *J Phycol* 2005; 41: 880–886.

54. John U, Lu Y, Wohlrab S, Groth M, Janouškovec J, Kohli GS, et al. An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. *Sci Adv* 2019; 5: eaav1110.

55. Fulnečková J, Ševčíková T, Fajkus J, Lukešová A, Lukeš M, Čestmír V, et al. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol Evol* 2013; 5: 468–483.

56. Bentlage B, Rogers TS, Bachvaroff TR, Delwiche CF. Complex ancestries of isoprenoid synthesis in dinoflagellates. *J Eukaryot Microbiol* 2016; 63: 123–137.

57. Mcfadden GI, Yeh E. the apicoplast: now you see it, now you don't. *Int J Parasitol* 2017; 47: 137–144.

58. Gornik SG, Febrimarsa, Cassin AM, MacRae JI, Ramaprasad A, Rchiad Z, et al. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc Natl Acad Sci* 2015; 112: 5767–5772.

59. Flegontov P, Michálek J, Janouškovec J, Lai D-H, Jirků M, Hajdušková E, et al. Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. *Mol Biol Evol* 2015; 32: 1115–1131.

60. Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture : Reoccurring themes , but significant differences at the extremes. *PNAS* 2015; 112: 10177–10184.

61. Lukeš J, Leander BS, Keeling PJ. Cascades of convergent evolution : The corresponding evolutionary histories of euglenozoans and dinoflagellates. *PNAS* 2009; 106: 9963–9970.

62. Beauchemin M, Roy S, Daoust P, Dagenais-Bellefeuille S, Bertomeu T, Letourneau L, et al. Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic. *PNAS* 2012; 109: 15793–15798.

63. Lin S. Genomic understanding of dinoflagellates. *Res Microbiol* 2011; 162: 551–569.

64. Zhang H, Campbell DA, Sturm NR, Dungan CF, Lin S. Spliced leader RNAs , mitochondrial gene frameshifts and multi-protein phylogeny expand support for the genus Perkinsus as a unique group of alveolates. *PLoS One* 2011; 6: e19933.

65. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, et al. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci* 2007; 104: 4618–4623.

66. Lasda EL, Blumenthal T. Trans-splicing. *Adv Rev* 2011; 2: 417–434.

67. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. *Science (80-)* 2005; 309: 436–442.

68. Bachvaroff TR, Place AR. From stop to start: Tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PLoS One* 2008; 3: e2929.

69. Wang JD, Berkmen MB, Grossman AD. Genome-wide coorientation of replication and transcription reduces adverse effects on replication in Bacillus subtilis. *PNAS* 2007; 104: 5608–5613.

70. Csuros M, Rogozin IB, Koonin E V. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* 2011; 7: 1–9.

71. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007; 41: 331–368.

72. Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P, et al. Miniature inverted-repeat transposable elements: discovery, distribution, and activity. *Genome* 2013; 56: 475–486.

73. Will CL, Lührmann R. Spliceosome structure and function. In: Atkins JF, Gesteland RF, Cech TR (eds). *Cold Spring Harbor Perspectives in Biology*. 2011. p 3:a003707.

74. Patel AA, Steitz JA. Splicing double: Insights from the second spliceosome. *Mol Cell Biol* 2003; 4: 960–970.

75. Burge C, Padgett R, Sharp P. Evolutionary fates and origins of U12-type introns. *Mol Cell* 1998; 2: 773–785.

76. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. *Nature* 2010; 465: 53–59.

77. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: Splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* 2013; 4: 61–76.

Fig. S1. Distribution of k-mer in A120 and A25

Analysis of k-mer from Illumina 100 bp paired end genomic reads of both *Amoebophrya*, including the genome size estimation. **A**: A25. **B**: A120.

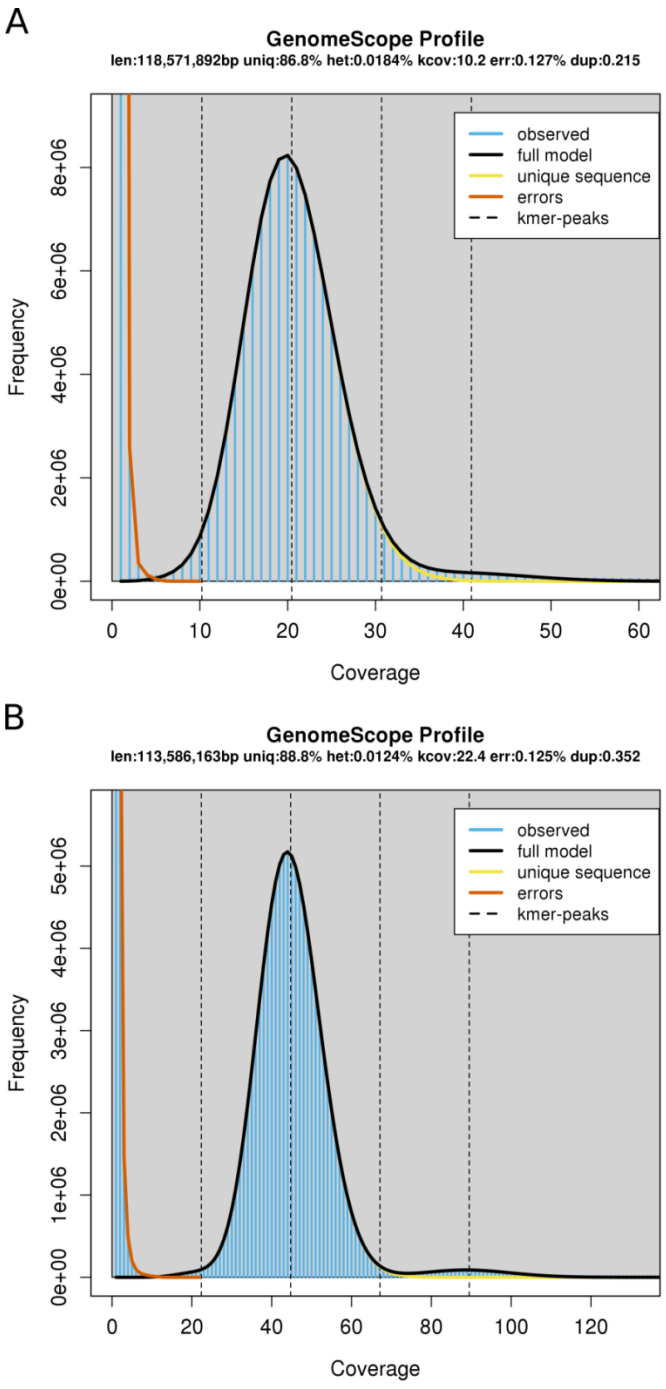


Fig. S2. Classification of repeated elements in *Amoebophrya* genomes (AT5 [54], A120 and A25) using REPET.

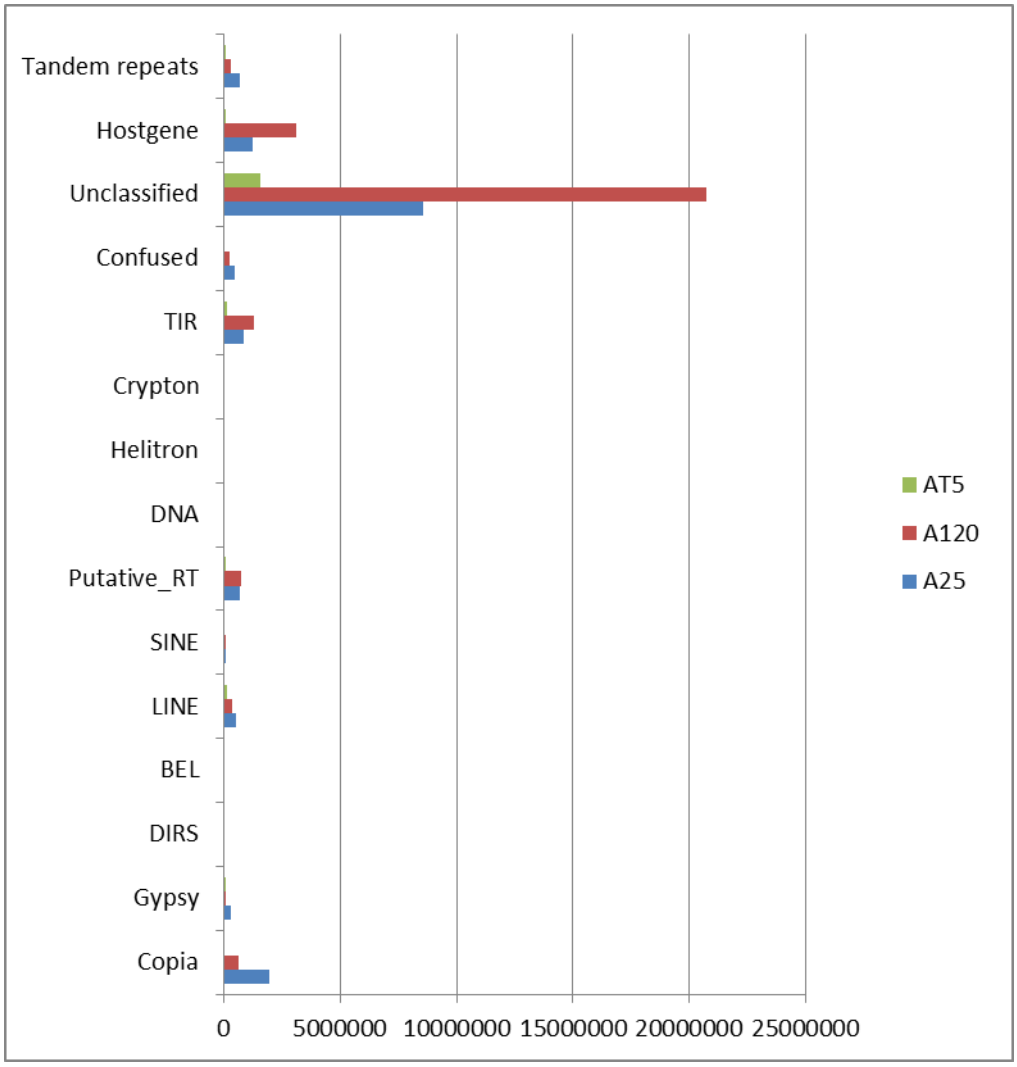


Fig. S3. Conserved motif of the putative splice leader (SL) in A25 and A120

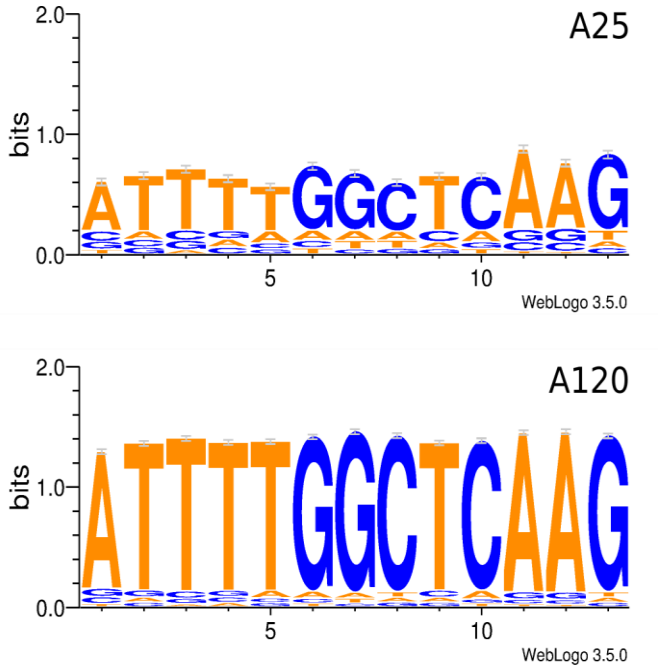


Fig. S4.Alignments of gene encoding the putative spliced leader (SL) gene in A120 and A25.

A120 and A25 SL encoding gene compared to other published sequence from *Amoebophrya* (Amo-XX) and *Karenia* spp. (Kmi and K. brevis) by Zhang et al. 2011.

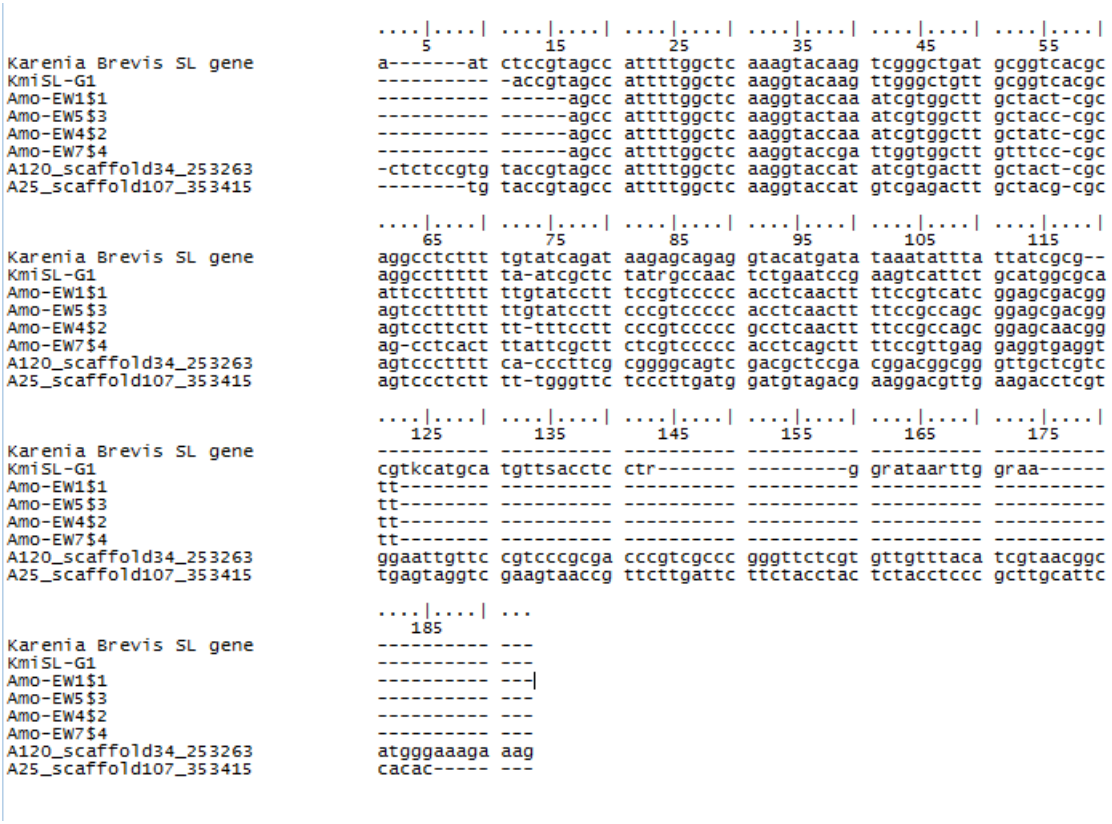


Fig. S5. Gene orientation change rate in *Amoebophrya* genomes.

Number of changes in gene orientation in *Amoebophrya* A120 and A25 compared to *Amoebophrya* AT5 (AT5), *Symbiodinium kawagutii* (Skav), *S. microadriaticum* (Smic) and *S. minutum* (Smin). Gene orientation was computed using a non-overlapping 10 genes sliding window.

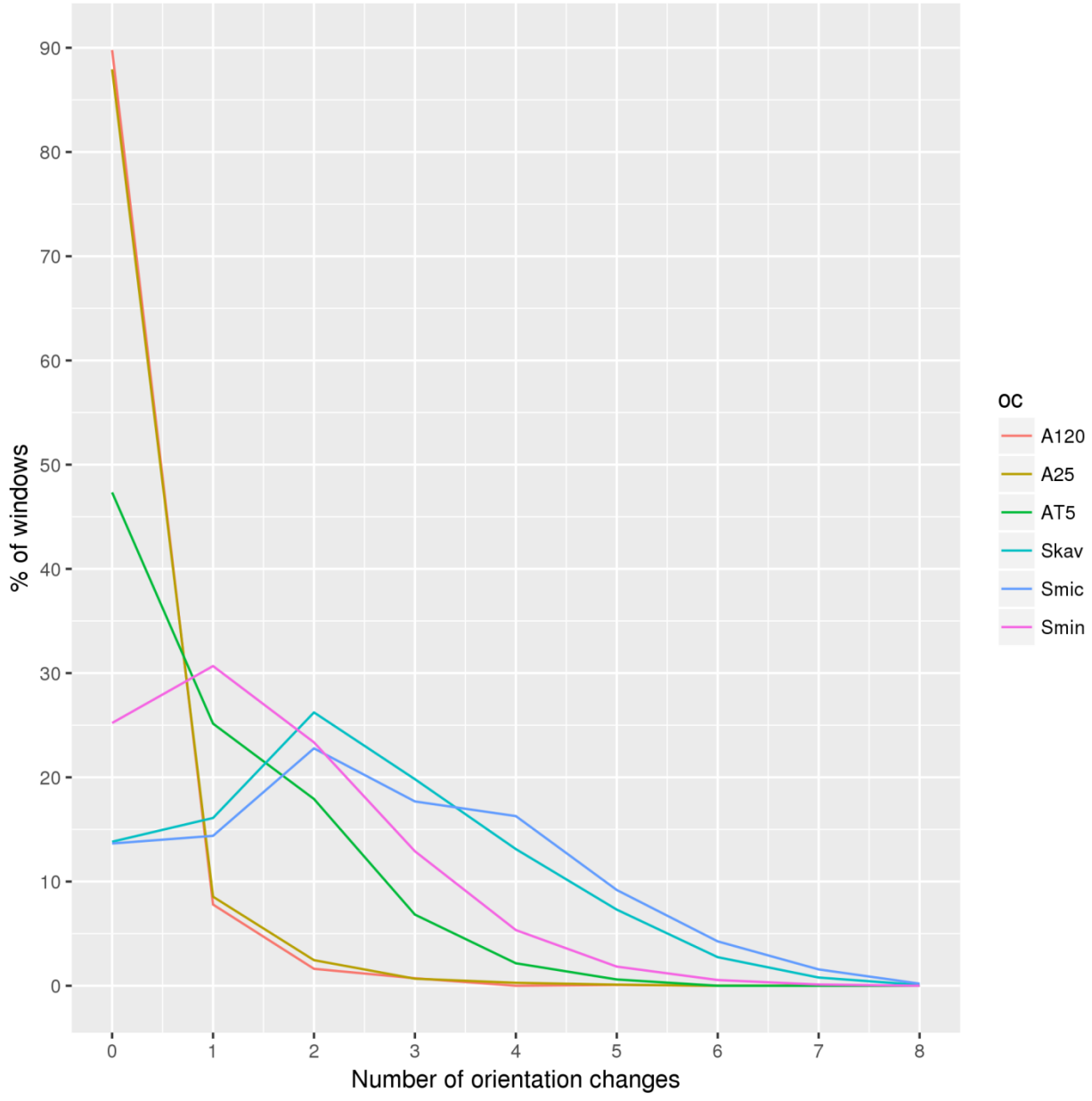


Fig. S6. Distribution of the numbers of orthologous and paralogous genes.

Number of orthologous and paralogous genes defined by Best Reciprocal Hit (BRH) searches between A120 (yellow), A25 (blue) and *Amoebophrya* AT5, *P. falciparum*, *P. marinus*, *S. kawagutii*, *S. microadriaticum*, and *S. minutum* predicted proteomes..

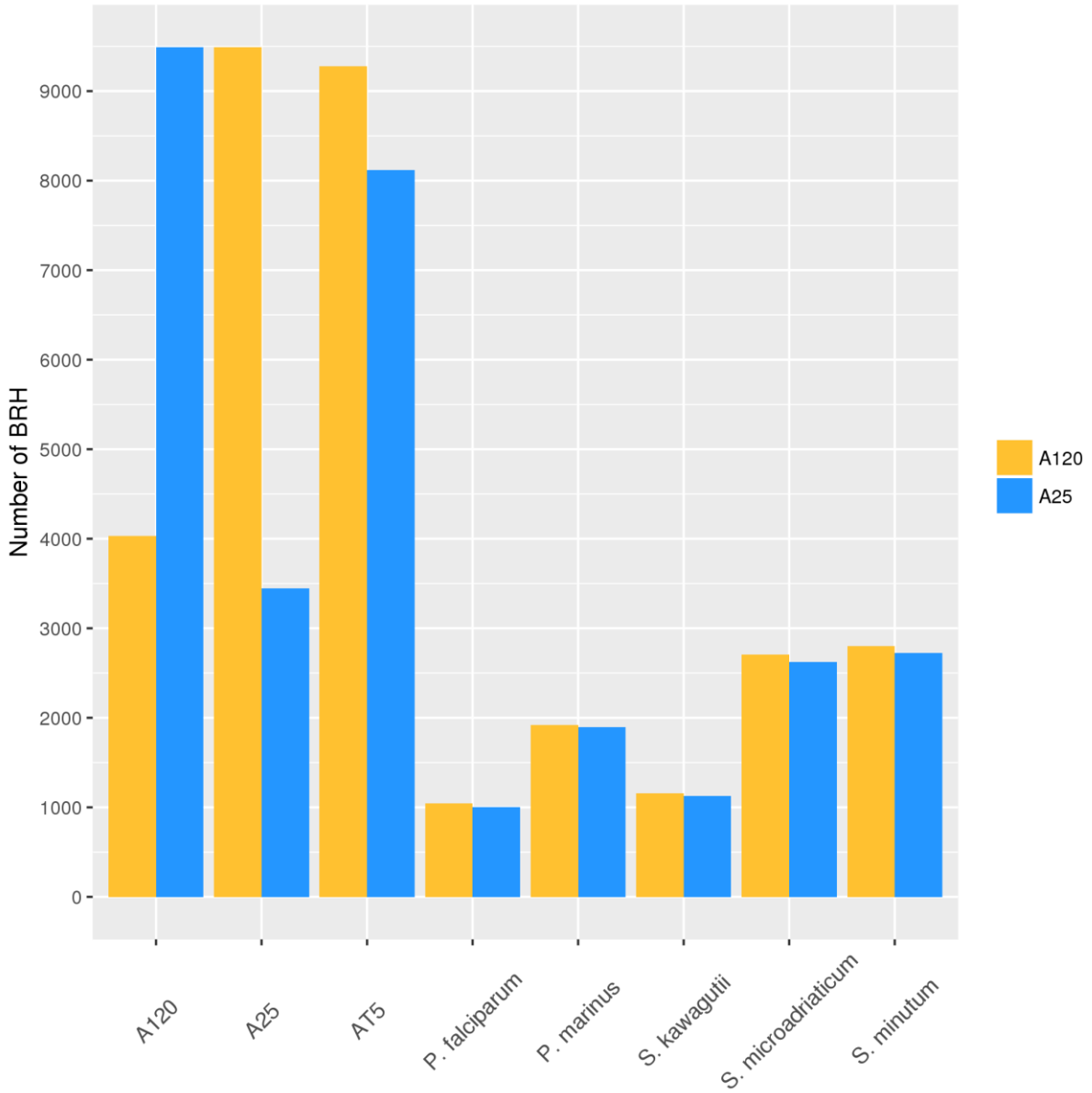


Fig. S7. Synteny dot-plot between *Amoebophrya* A120 and AT5.

Dot-plot of the synteny between the longest scaffolds for each of the *Amoebophrya* AT5 and A120 genomes. The 100 scaffolds (AT5) and the 100 scaffolds (A120) are shown on the x and y axes, respectively. For each genome, genes are sorted by their rank on the scaffolds. Each blue point represents a pair of orthologous genes defined by BRH.

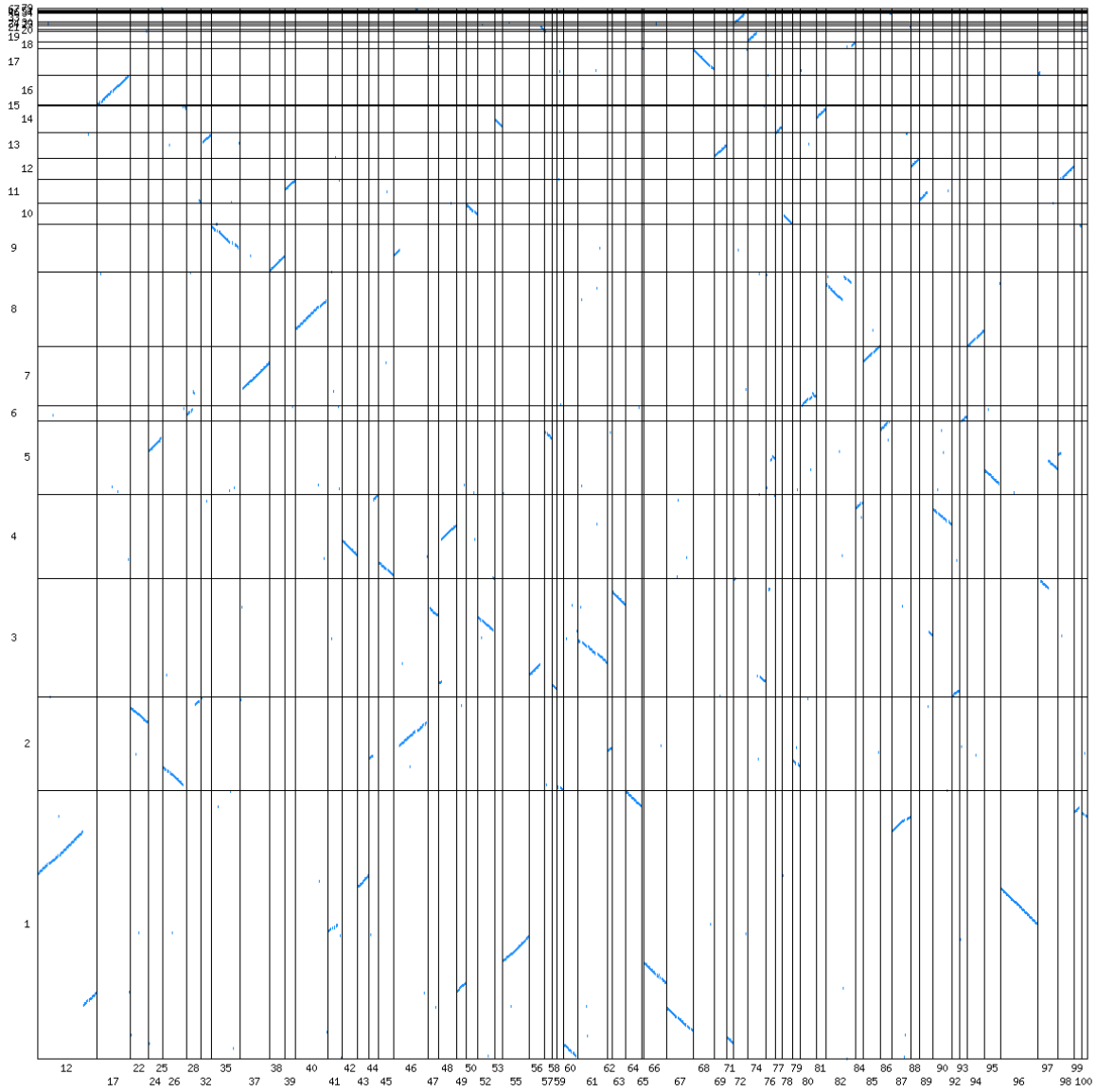


Fig. S8. Synteny dot-plot between *Amoebophrya* A25 and AT5

Dot-plot of the synteny between the longest scaffolds for each of the *Amoebophrya* AT5 and A25 genomes. The 100 scaffolds (AT5) and the 100 scaffolds (A25) are shown on the x and y axes, respectively. For each genome, genes are sorted by their rank on the scaffolds. Each blue point represents a pair of orthologous genes defined by BRH.

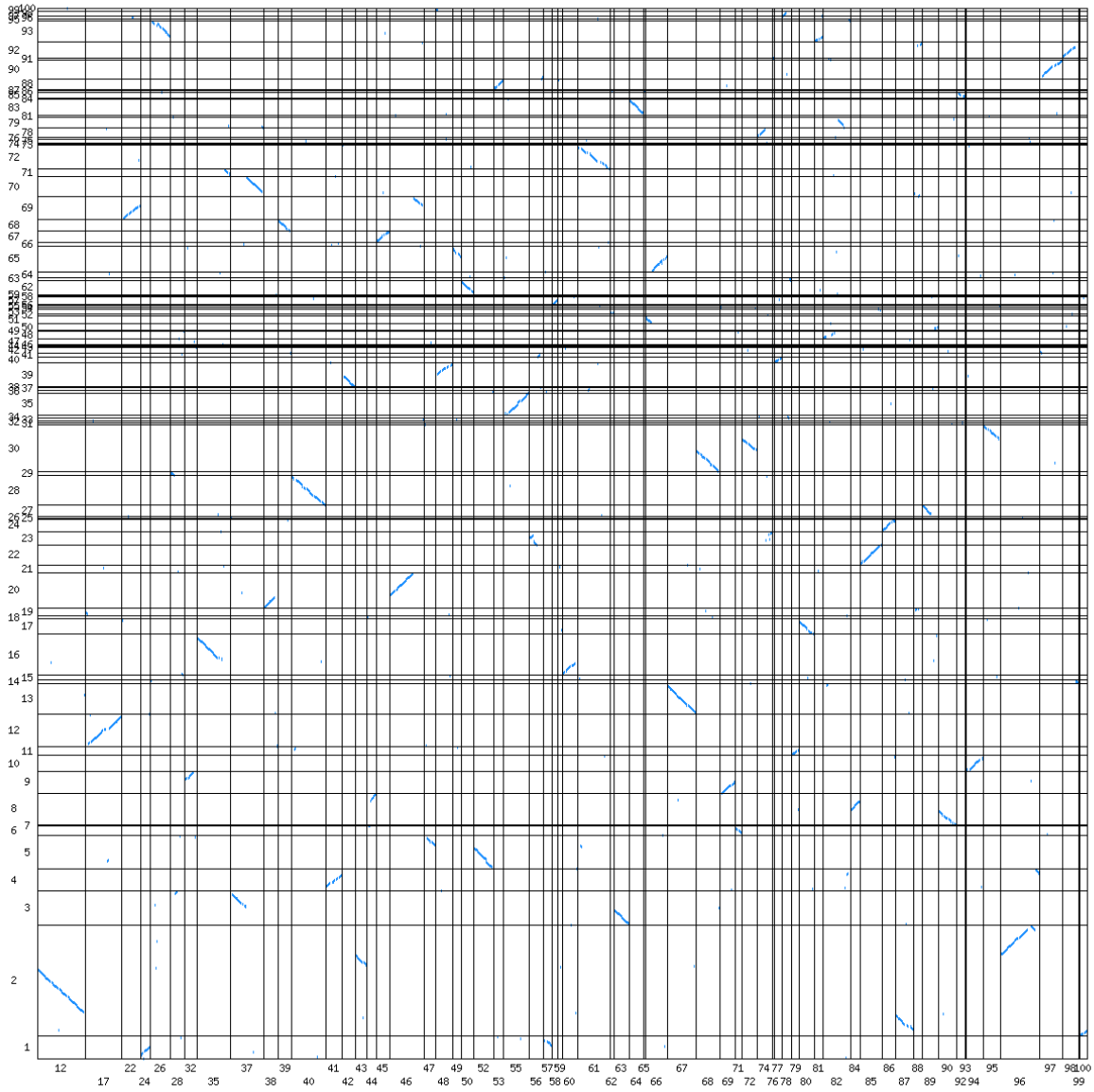


Fig. S9. Intron splicing motifs in A120 (upper) and A25 (down)

Canonical introns: square delimiting the intron, including the canonical donor and acceptor motifs. Shaded area up- and downstream of the intron represent exon sequence.

Non-canonical introns: line above logos indicate intron region with palindromic motifs forming the hairpin (sold line). Splice sites relative to the hairpin-motif are variable (dashed line). Also, exact position of intron border remains unknown (shaded gradient).

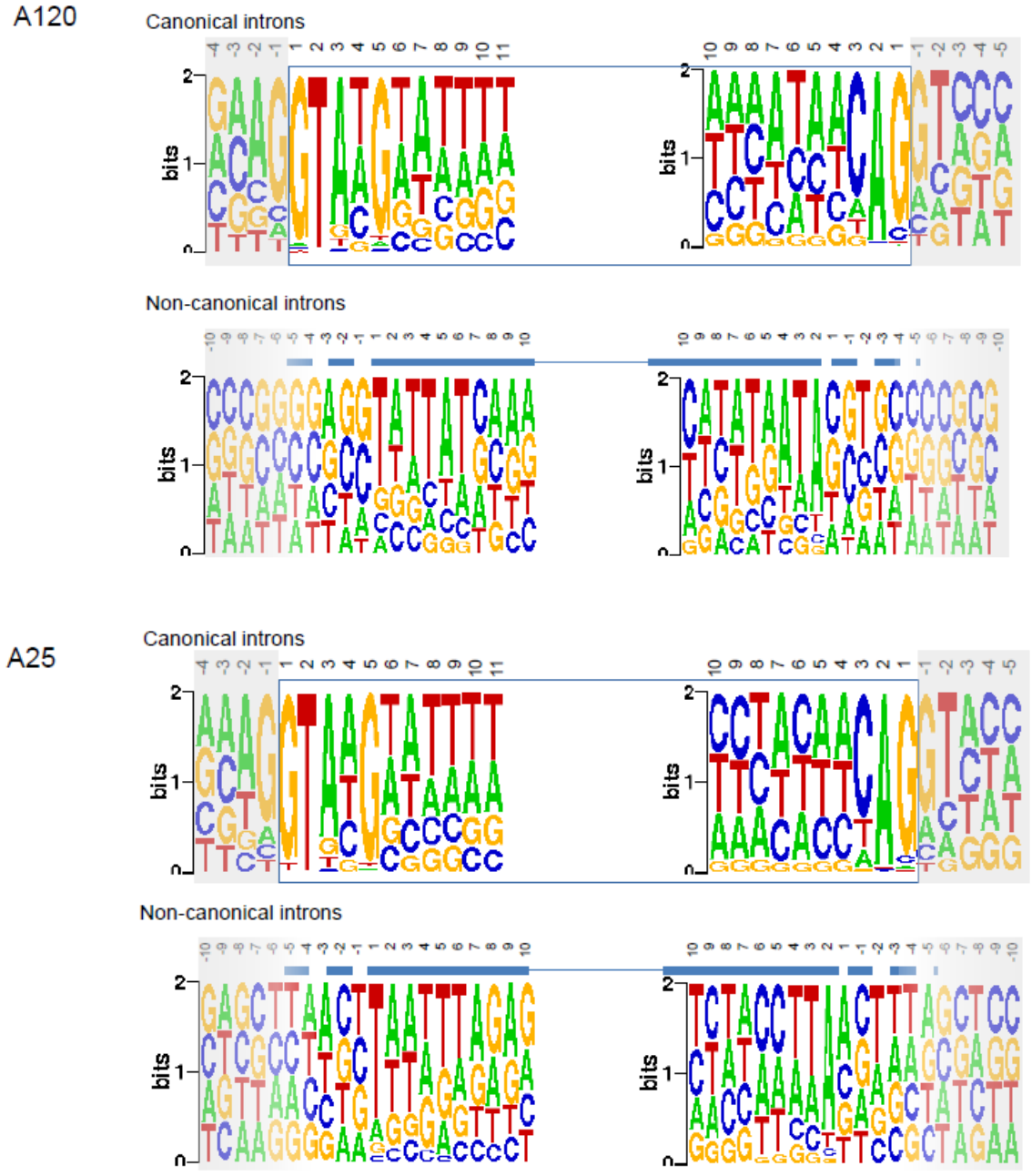


Fig. S10. Distribution of conserved introns.

Violin plot distribution of the ratio of conserved intron based upon the level of amino acid level identity of aligned orthologous genes. Percent identity is shown on the x axis. A diamond represents the average ratio of conserved introns for each violin plot. The minimum alignment length for each orthologous pair was >80%.

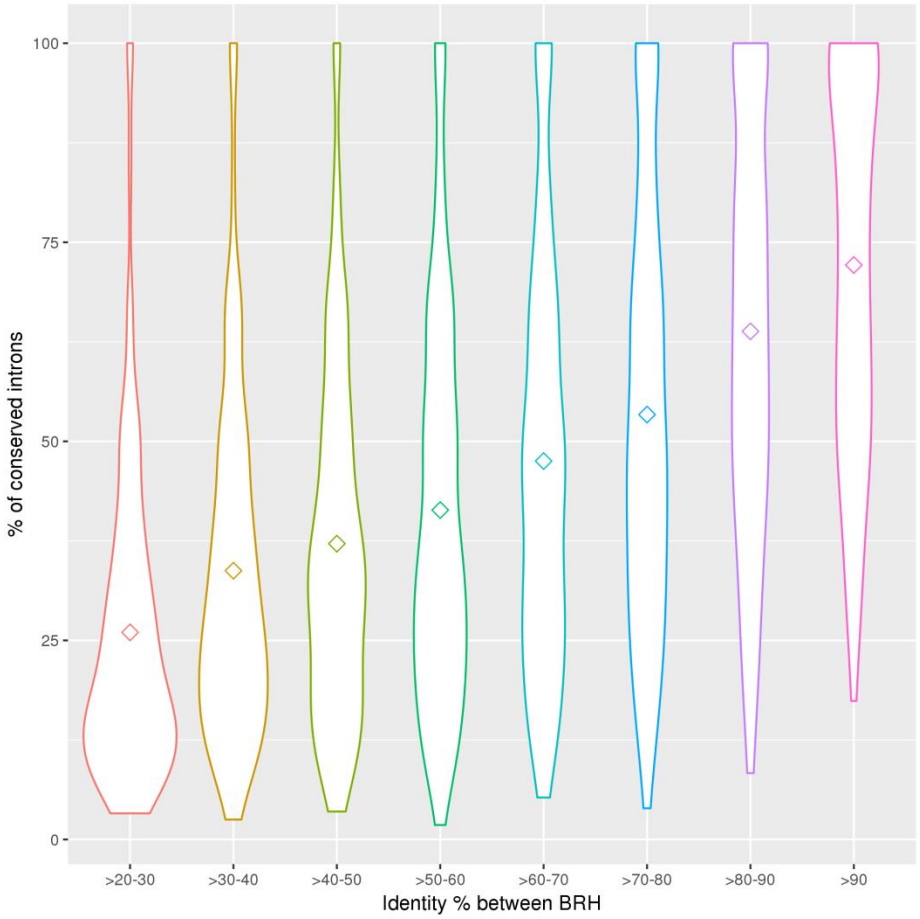
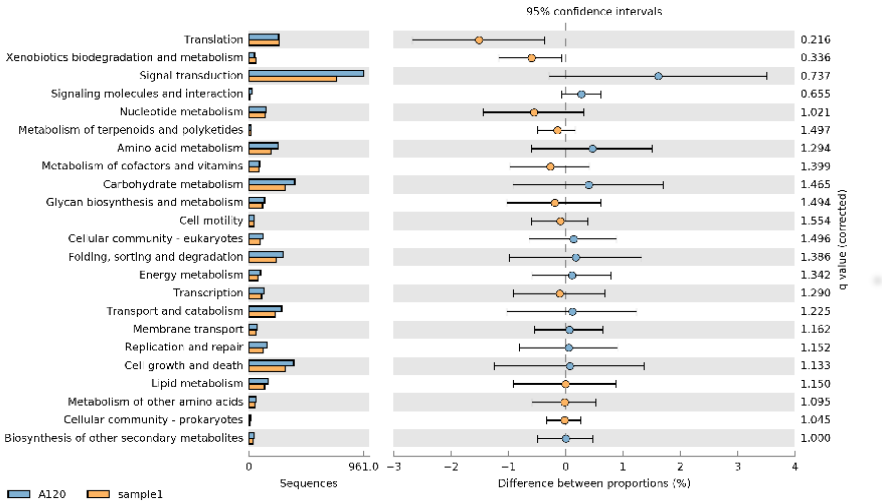


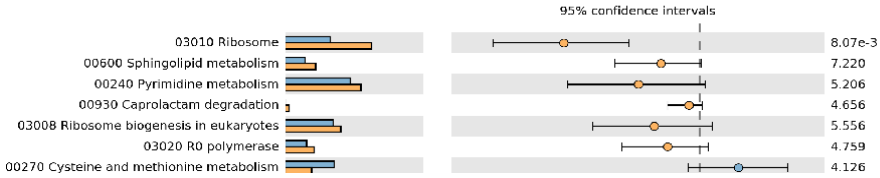
Fig. S11. The difference in the proportion of IEs-containing-genes compared to +/- IEs genes in KEGG assignment in A120 and A25.

Integration of IEs is statistically different in genes involved in the translation (A25), and ribosome (A120 and A25).

A120



A25



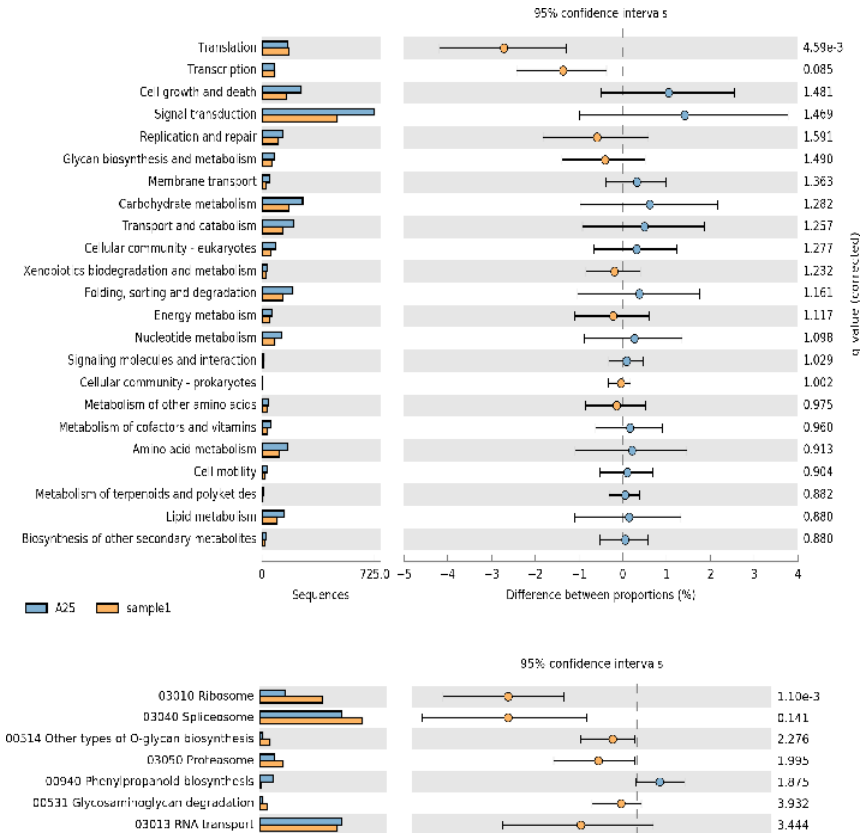


Fig. S12. Example of introner elements (IEs) in Amoebophrya.

Direct repeats of 5 bp are in blue. Inverted repeats (red) are at the introner element ends. Squares are the exon sequence border prediction.

Fig. S13. Distribution of the direct repeats in size range from 3-8 nucleotides in A120.

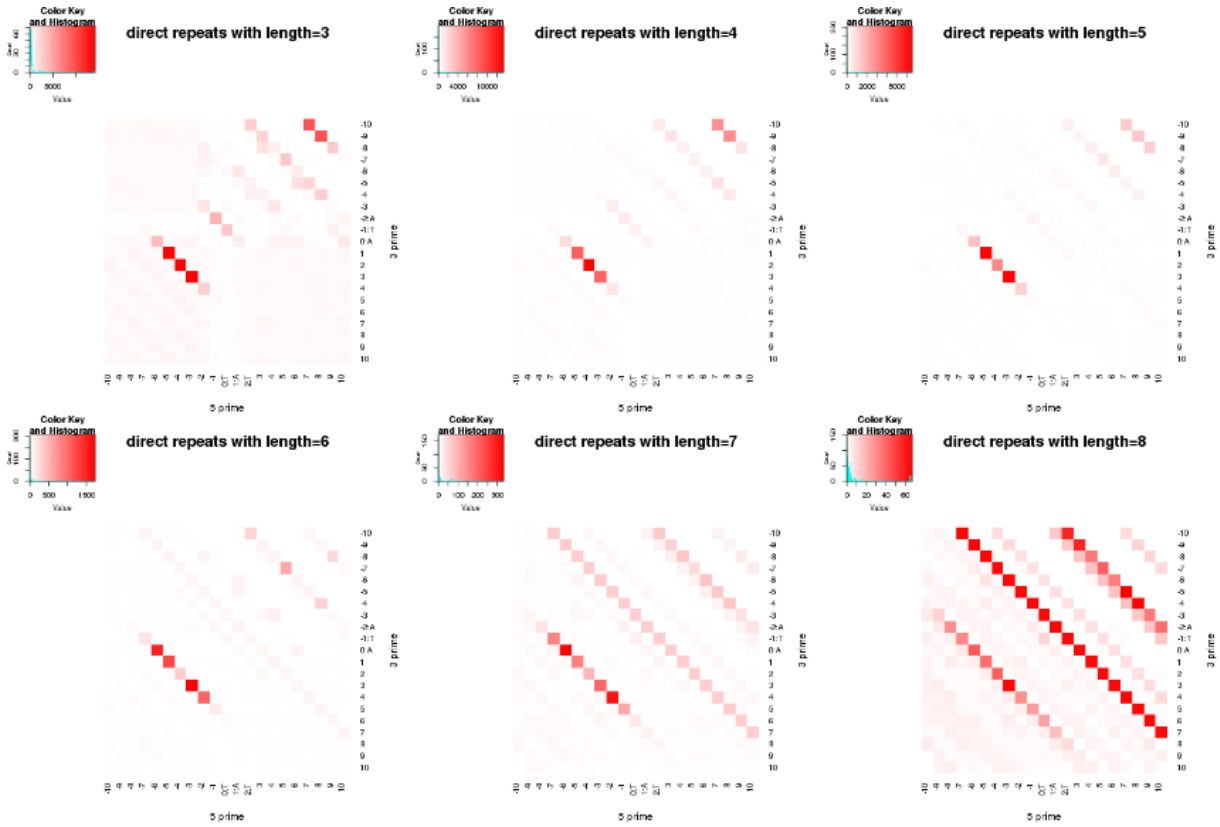


Fig. S14. Distribution the direct repeats in size range from 3-8 nucleotides in A25.

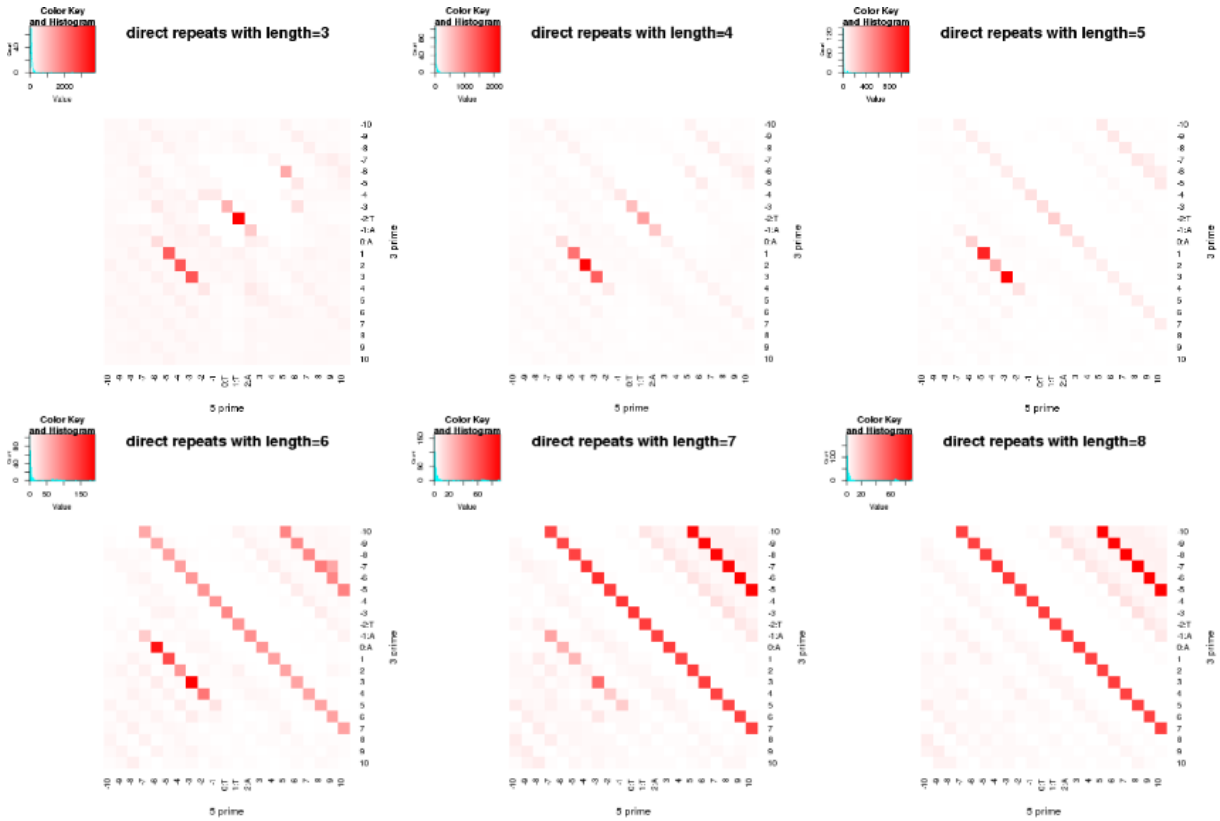


Fig. S15. Composition of direct repeats in introners elements.

The diversity in composition of the three (a, b, c) most abundant of direct repeats in introner elements in A120 (up) and A25 (down).

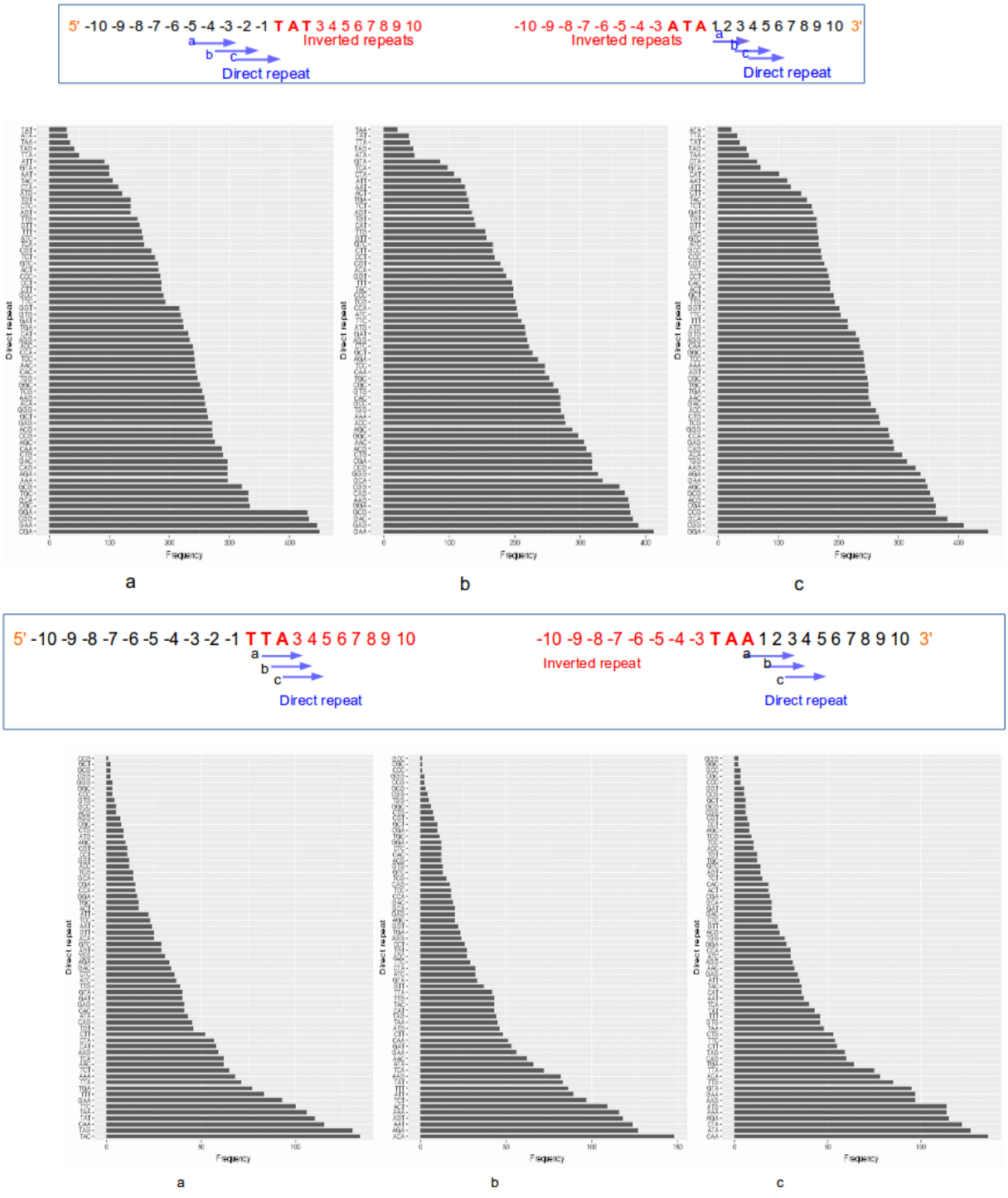


Fig. S16. Terminal inverted repeat locations around the splicing sites in A120 and A25.

The position of inverted repeats via the splice sites in A120 and A25. The inverted repeats of A120 are located at 1-5 the nucleotides upstream and downstream of splice sites. b. The inverted repeats of A25 are located at the 1-6 nucleotides in upstream and downstream of splice sites.

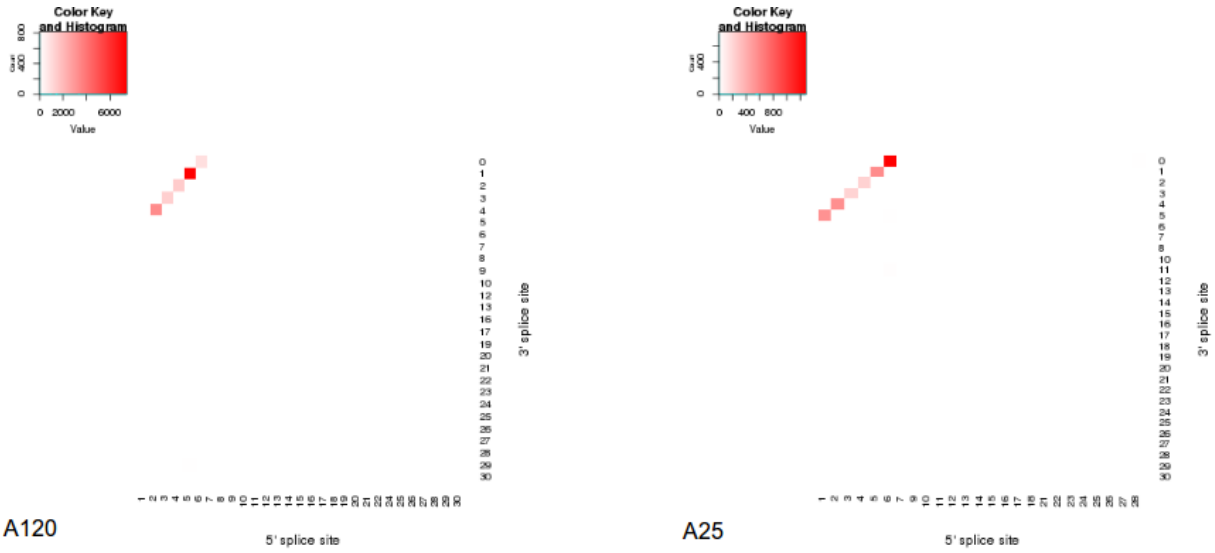


Fig. S17. The flowchart of the search of introner elements.

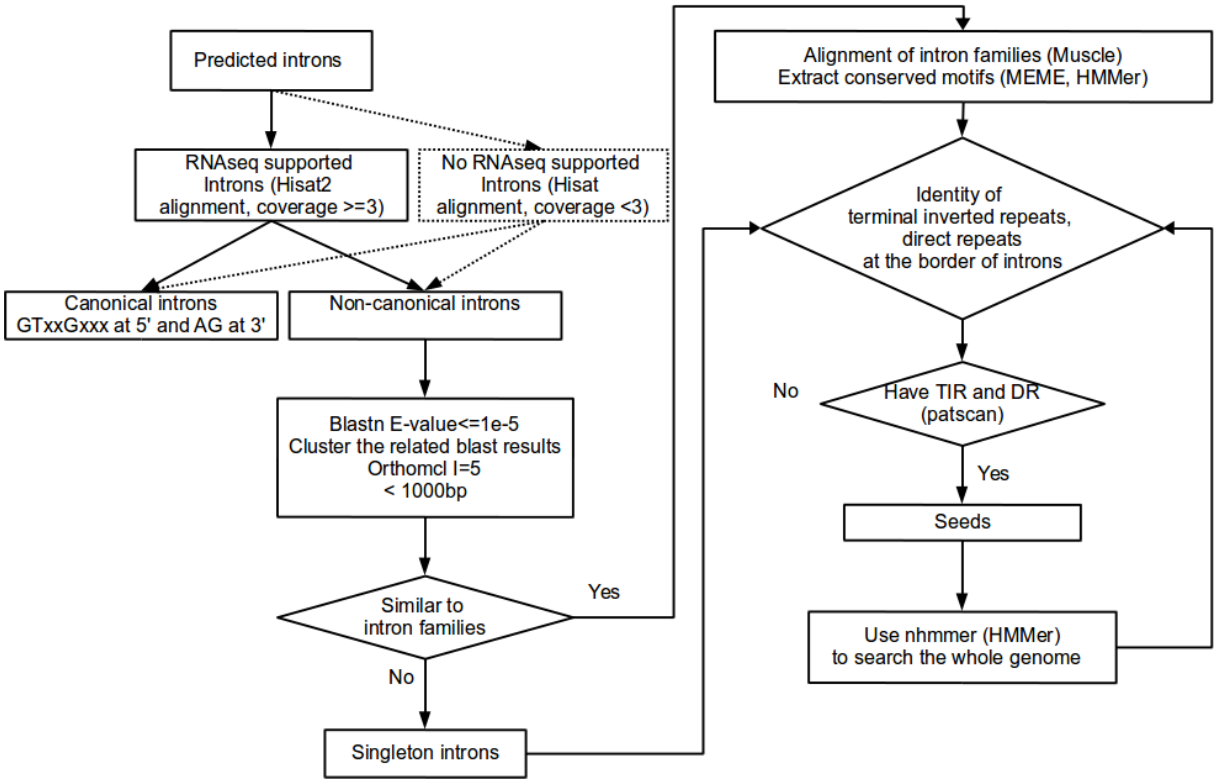


Fig. S18.
Hierarchical clustering analysis (pairwise similarity, OrthoMCL) of all intron families and the inverted repeats in A25 and A120.

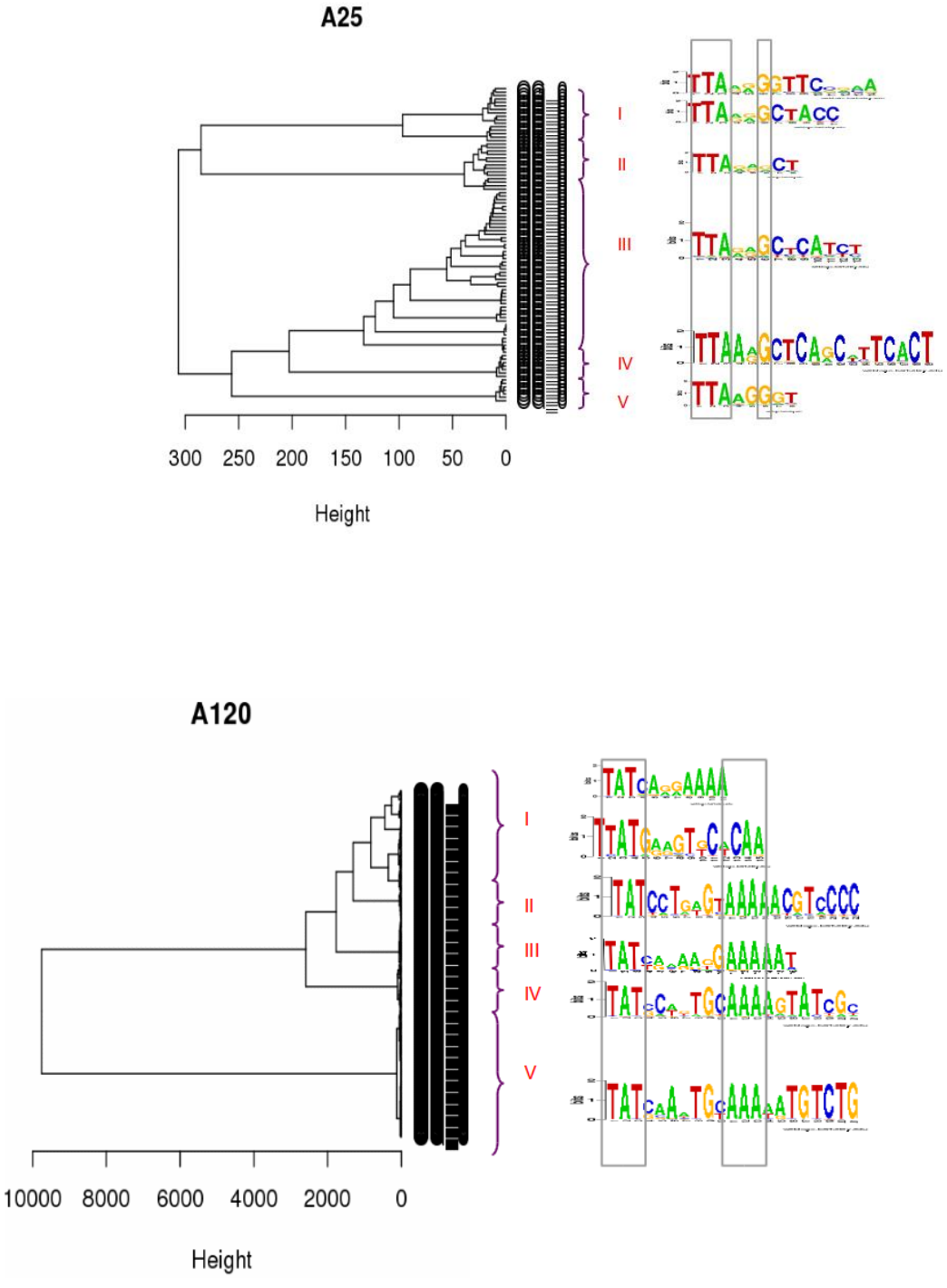


Fig. S19. Distribution of length, GC%, and % identity of introners (IES) in the two Amoebophrya A120 and A25 strains

The percent identity is the median of percent identity in each intron family. All of these introners split into distinct populations having distinctive length (the length is the median length of each intron family, two size peaks in A120 (pink): ~130 nt and 260 nt, one size in A25 (blue): ~110nt, **left**) either different GC content (two populations in A120: 37% and 45%, one population in A25: 44-45%, **middle**), although no correlation between length and GC-content could be made. The percent pairwise identity between individual introner elements were 77-88% in A120, and 74-95% in A25 (**rigth**), indicating that IEs in both genomes evolved recently and probably IEs in A25 were more recent than IEs in A120.

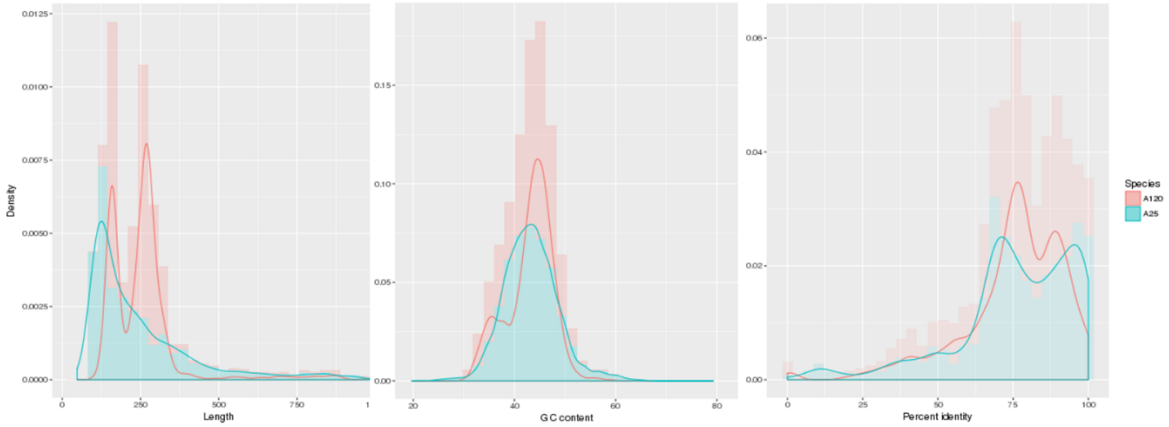


Fig. S20. Multiple alignments of U2 snRNAs

Multiple alignment of U2 snRNAs of *Amoebophrya* A120 and A25 (A120-U2A_T12, A120-U2B_T18, A120-U2C_T24, A120-U2D_T30, A120-U2E_T6 and A25-U2A_Dinospore, A25-U2B_T12, A25-U2C_T24, A25-U2D_T30, A25-U2E_T36, A25-U2F_T42, A25-U2G_T6) compared to *P. falciparum*, *S. minutum* and *H. sapiens* U2 snRNAs (Pf-u2snRNA, Sm-u2snRNA and Hs-u2snRNA)

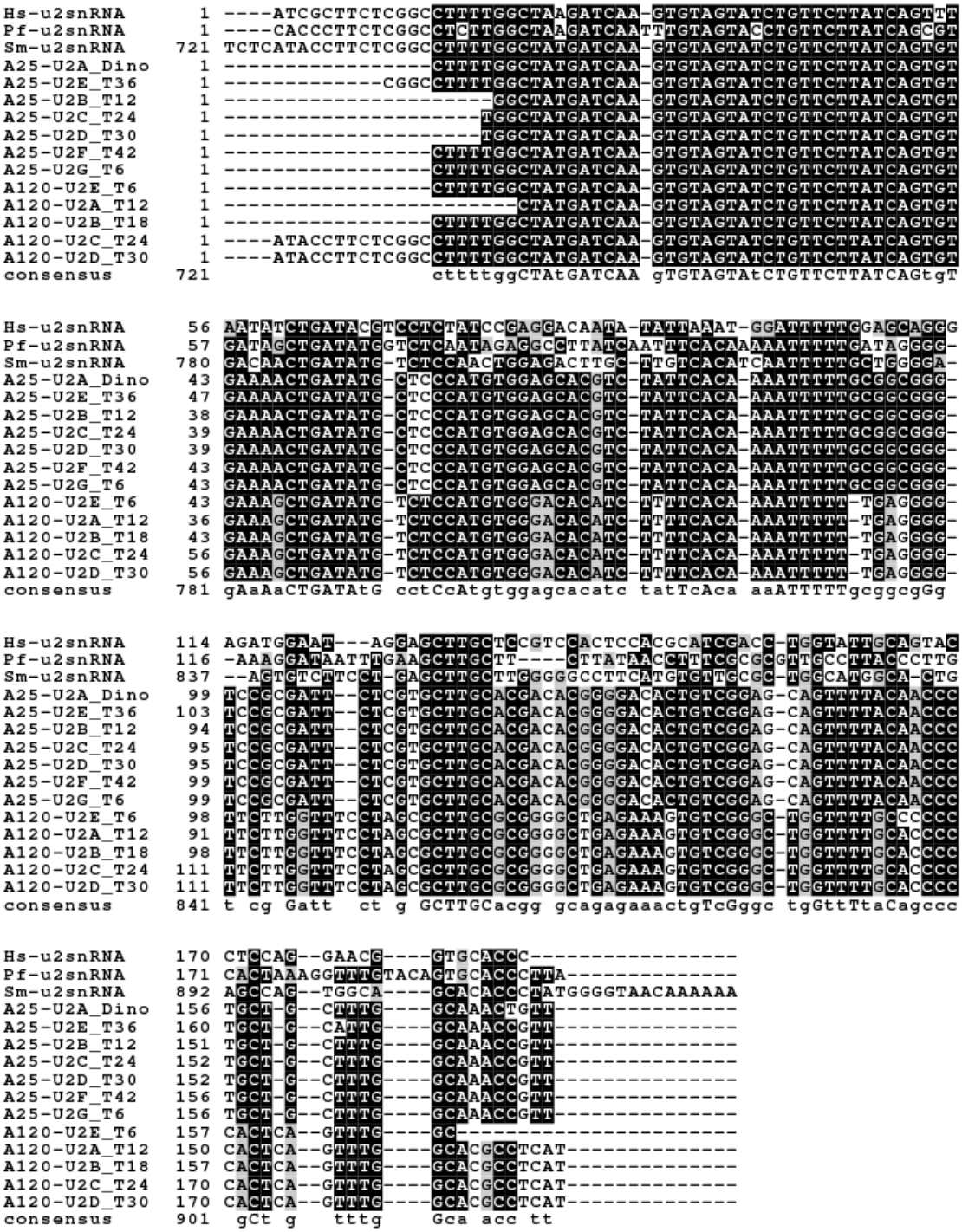


Fig. S21. Multiple alignments of U4 snRNAs

Multiple alignment of U4 snRNAs of *Amoebophrya* A120 and A25 (A120-U4A_Dinospore, A120-U4B_T18, A120-U4C_T30, A120-U4D_T36 and A25-U4_Dinospore) compared to *P. falciparum*, *S. minutum* and *H. sapiens* U4 snRNAs (Pf-u4snRNA, Sm-u4snRNA and Hs-u4snRNA)

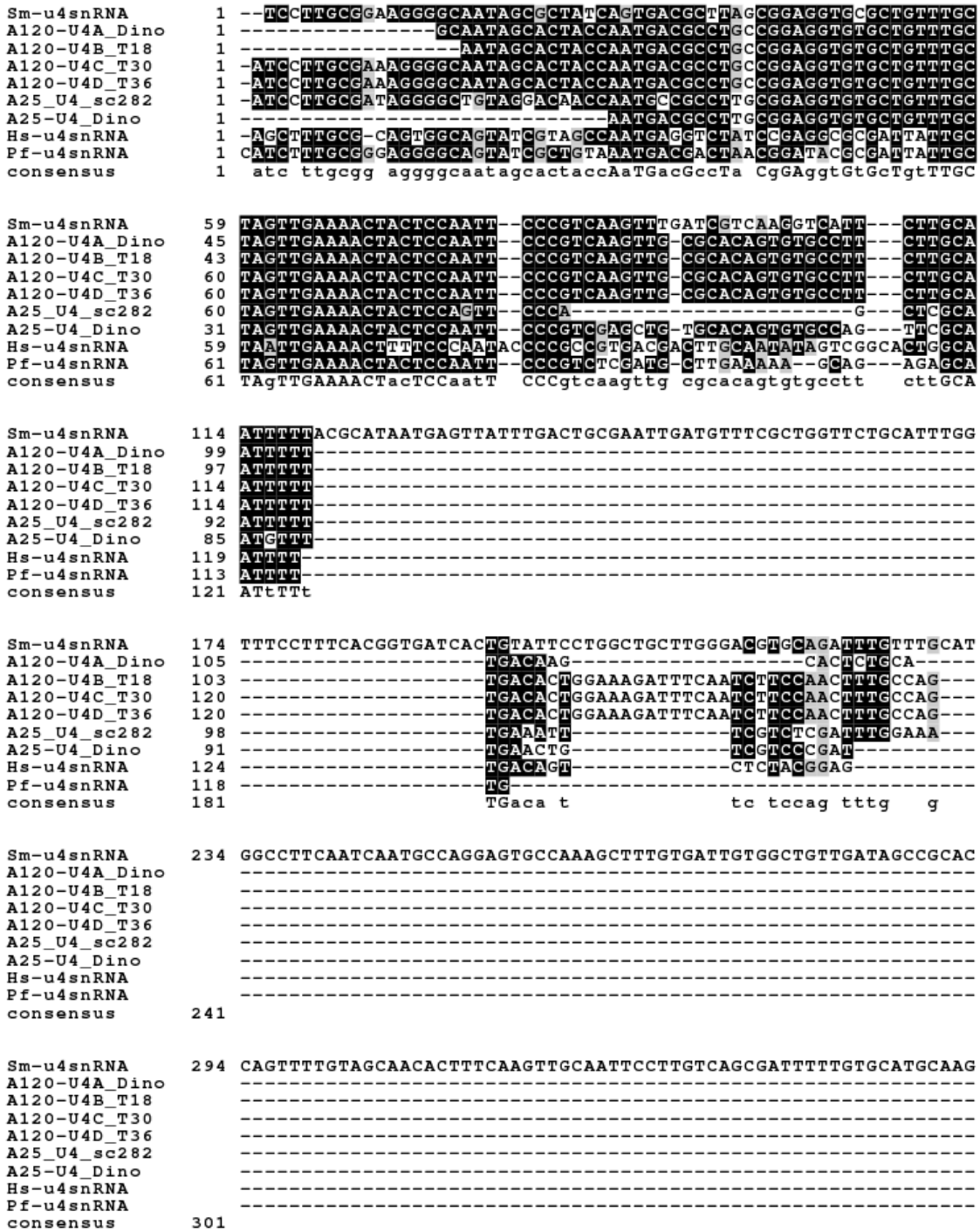


Fig. S22. Multiple alignments of U5 snRNAs

Multiple alignment of U5 snRNAs of Amoebophrya A120 (A120-U5A_T30 and A120-U5B_T36) compared to *P. falciparum*, *S. minutum* and *H. sapiens* U4 snRNAs (Pf-u5snRNA, Sm-u5snRNA and Hs-u5snRNA)

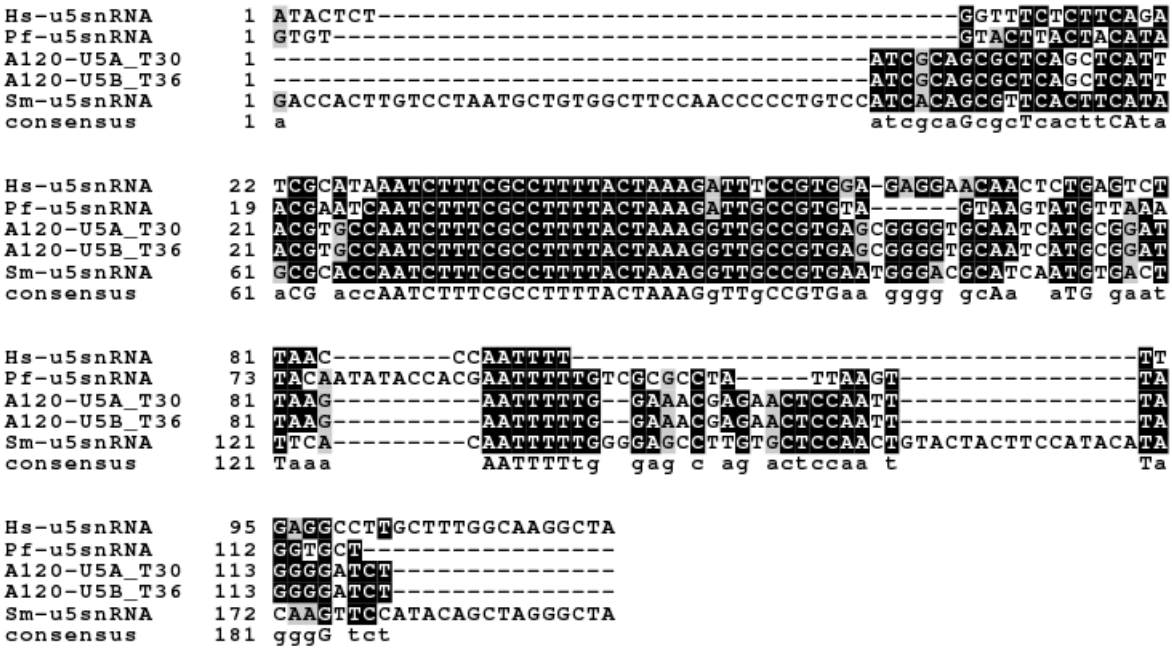


Fig. S23. Multiple alignments of U6 snRNAs

Multiple alignment of U6 snRNAs of Amoebophrya A120 and A25 compared to *P. falciparum*, *S. minutum* and *H. sapiens* U6 snRNAs (Pf-u6snRNA, Sm-u6snRNA and Hs-u6snRNA)

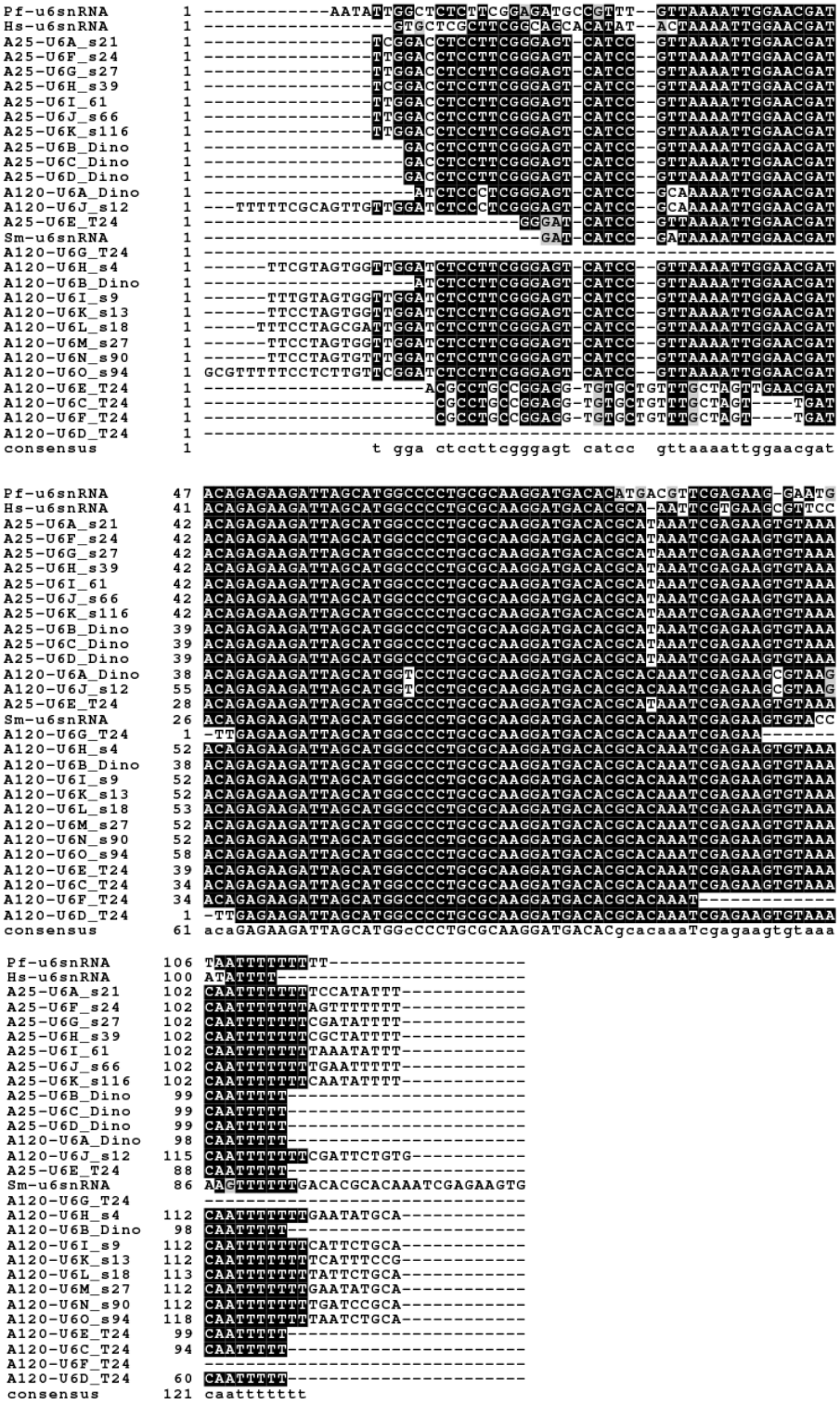


Fig. S24. Secondary structure of *Amoebophrya* snRNA.

U2, U4, U5 and U6 secondary structure of *H. sapiens*, A120 and A25. (Lack of A25 U5 snRNA). A is U2 snRNAs, B is U4 snRNA, C is U5 snRNA and D is U6 snRNA.

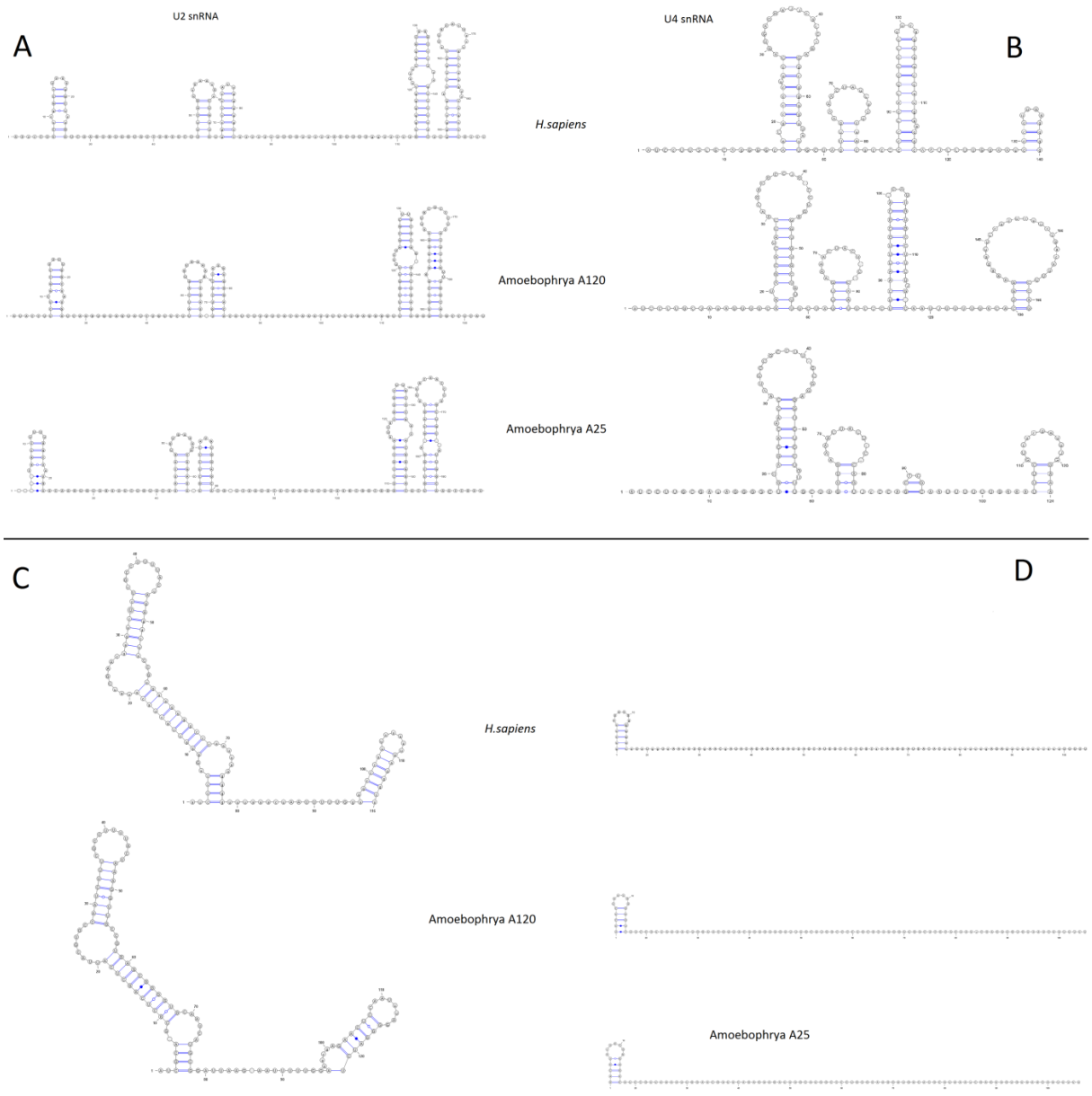


Table S1.

RCC number, date and site of isolation of strains

Strain	**RCC number**	**Date and site of isolation**
Scrippsiella acuminata (ST147)*	RCC1627	2005 from sediment
Heterocapsa triquetra (HT150)	RCC3596	July 6th 2007
A25**	RCC4383	15th of June 2009
A120***	RCC4398	13th of June 2011

* previously known as *S. trochoidea*, Kretschmann et al. (2015)

http://dx.doi.org/10.11646/phytotaxa.220.3.3

** First isolated in *Scrippsiella acuminata*

*** First isolated in *Heterocapsa triquetra*, then maintained in *Scrippsiella acuminata*

Table S2.

Metrics of Nanopore runs for the two *Amoebophrya* strains

	A25	**A120**
Number of runs	2	3
Cumulative size (nt)	2 534 247 679	14 022 482 812
Average size (nt)	15 244	9 344
N50 (nt)	19 456	14 562

Table S3.

Assembly and annotation metrics for *Amoebophrya* A120 and A25, and AT5, *Symbiodinium minutum* (Smin), *S. kawagutii* (Skav), *S. microadriaticum* (Smic) and *P. marinus* (Pmar).

	A25	A120	AT5	Skav	Smin	Smic	Pmar
Assembly							
Number of scaffolds	557	50	2,351	30,040	21,899	9,695	17,897
Cumulative size (Mb)	116	115.5	87.7	935	609	808	87
N50 / L50	1.082Mb / 35	9.243Mb / 5	83.9kb / 298	381kb / 772	125kb / 1448	574kb / 420	158kb / 124
N90 / L90	423kb / 106	1.464Mb / 18	19.6kb / 1,095	109kb / 2,477	31kb / 5103	146kb / 1442	1.2kb / 9,284
Max. size	3.013Mb	16.512Mb	537kb	1.914Mb	811kb	3.145Mb	1.8Mb
%N	2.27	1.41	2.25	3.4	0.9	7.7	0.64
%GC	47.8	51.2	55.92	45.5	43.5	50.5	47.4
Genes							
Number	28,091	26,441	19,925	36,850	41,925	49,109	23,654
Density (genes/Mb)	247.78	232.18	227.2	39.4	68.78	60.8	273.1
Average length (bp)	2,965	3,482	2,782	3,788	11,961	12,898	1,581
Median length (bp)	1,890	2,442	1,803	2,039	7,899	7,255	1,038
Exons							
Number	117,411	121,327	67,639	150,118	985,369	1,072,528	133,410
Av. length (bp)	475	541	578	256	99	109	177
Median length (bp)	235	265	319	81	53	51	112
Longest (bp)	79,744	44,016	14,772	11,064	14,818	13,755	16,293
Average number of exons / gene	4.18	4.59	3.39	4.07	20.96	21.8	5.64
% GC	51.9%	56.3%	54.7%	52.7%	50.8%	56.9%	50.95%
Introns							
Number	81,610	90,882	47,714	113,268	938,355	1,023,342	109,756
% of spliced genes	69.8%	66.9%	71.3%	64.1%	95.4%	98.6%	72.4%
Average length (bp)	345	335	337	893	517	505	124
Median length (bp)	208	247	228	501	297	231	49
Longest (bp)	90,415	35,152	3,556	9,977	88,176	177,825	11,034
% GC	44%	46.5%	49.4%	44.5%	41.8%	47.1%	43.4%
% of introns with GT-AG splice sites	34.02%	30.41%	99.98%	65.38%	48.23%	0,26	99.3%
% of introns with GC|GA-AG splice sites	0.45%	2.95%	0.02%	25.30%	51.77%	73.95%	0.7%
% of introns with other splices sites	65.53%	66.64%	0%	9.32%	0%	0.05%	0%
CDS							
Average coding size (bp)	1,337	1,773	1,962	1,041	1,916	2,375	4,839
Genome coverage of coding bases, % in brackets	32.4%	40.6%	44.6%	4.1%	13.1%	14.4%	26.4%
Gene families							

Number of genes belonging to families, % in brackets	7,074 (25.2)	7,428 (28.1)	ND	20,374 (55.3)	25,809 (61.5)	32,796 (66.8)	18,258 (77.2)
Avg. of genes in a family	3.5	3.6	ND	6.7	5.9	7	ND
Max. of genes in a family	171	157	ND	889	703	831	ND
Annotation							
Number of proteins with at least one significative match	8,360	8,690	4,366	29,720	13,813	5,538	ND
Number of proteins with KO assignation	5,774 (21%)	5,983 (23%)	2,018	14,926 (40%)	10,954 (65%)	3,008 (54%)	ND
Number of proteins with BRITE assignation	5,774	5,856		14,764	10,755	2,960	ND
Number of proteins of with an IPR domains	8,444	9,054	7,404	16,895	13,541	4,059	ND
Number of proteins with UniProt matches (%)	9,101 (32.4)	9,404 (35.6)	ND	ND	ND	ND	ND

Table S4. Number of the different types of introns

Number of validated introns, canonical and non-canonical introns. Detailed known splicing sites motifs are shown. Ratio of others splicing sites are not shown.

	A120	A25
Introns in total	90,882	81,610
Supported RNA-seq introns (coverage>3)	66,565	55,290
Canonical introns (GT-AG)	35%	40%
Non-canonical introns	65%	60%
Introns with GC-AG splice sites	1%	0.2%
Introns with AT-AC splice sites	1.5%	0.9%
Introns with GG-AG splice sites	0.8%	0.04%
Introns with GT-TG splice sites	1.3%	0.5%
Introns with GT-CG splice sites	2.1%	0.3%
Introns with CT-AG splice sites	2.6%	0.3%

Table S5.

Classification into families of non-canonical introns in A120 and A25.

Clustering analyses considering non-canonical introns having RNA-seq to validate the introns junctions (coverage>3) and length ≤ 1k, classification based on the IR sequence and sequence similarity (introners) or MITEs-like elements.

	A120		A25	
	Introners	MITEs	Introners	MITEs
Family number	1,954	1,121	252	34
Family member	29,850	13,748	2,039	249

Table S6.

Putative *Amoebophrya* A120 and A25 snRNP homologs

Amoebophrya snRNPs prediction identified by reciprocal hits analysis and mcl gene family groups with *H. sapiens*, *P. falciparum* and *T. gondii*.

Complex subunit	***H. sapiens* (NP)**	***T. gondii* (TGME49_)**	***P. falciparum* (PF3D7)**	**A120**	**A25**
snRNP core (stability and function of U1, U2, U4 and U5 snRNPs)					
SNRPB	_937859	300280	_1414800	GSA120T00015799001	GSA25T00009626001
SNRPD1	_008869	267350	_1125500	GSA120T00006136001	GSA25T00020775001
SNRPD2	_004588	270830	_0218500	GSA120T00007224001	GSA25T00005714001
SNRPD3	_004166	309740	_0909800	GSA120T00003650001	GSA25T00025008001
SNRPE	_003085	275750	_1350200	GSA120T00006897001	GSA25T00019804001
SNRPF	_003086	213410	_1126900	GSA120T00013291001	GSA25T00025362001
SNRPG	_003087	314790	_0822300	GSA120T00009095001	GSA25T00001657001
U1 snRNP					
U1-70K	_003080	205180	_1367100	GSA120T00005676001	GSA25T00019007001
U1A	_004587	309800	_1306900	GSA120T00025754001	GSA25T00018322001
U1C	_003084	306380	_0812700	GSA120T00018889001	GSA25T00015584001
PRP40 (A/B)	_001026868	306220	_1316500	GSA120T00025896001	GSA25T00014867001
RBM25	_067062	270770	_0610200	GSA120T00003034001 GSA120T00002189001	GSA25T00026804001 GSA25T00021865001
DDX5	_004387	_236650	_1445900	GSA120T00022090001	GSA25T00013401001
CA150	_001035095	316180	_1111200	GSA120T00004838001	GSA25T00009802001
U4/U6 snRNP					
PRP3	_004689	219790	_1309300	GSA120T00002339001	GSA25T00014061001

PRP4	_001231855	243540	_1343900	GSA120T00005896001	GSA25T00019043001
CypH	_006338	_230520 _205700 _285760	_0322000 _0804800 _1115600	GSA120T00006823001 GSA120T00000182001	GSA25T00019816001 GSA25T00027189001 GSA25T00006816001
PRP31	_056444	244100	_0409100	GSA120T00000517001	GSA25T00008446001
Snu113	_004999	236580	_1123900	GSA120T00010172001	GSA25T00018752001
snRNP27	_006848	264010	_0818000	ND	ND
Sad1	_006581	294360	_1317000	GSA120T00014855001	GSA25T00023451001
Snu66	_005137	318140	_0323700	GSA120T00018941001	GSA25T00016665001
Snu23	_653324	275310	_1243100	GSA120T00006991001	GSA25T00019720001
PRP38A	_060531	266030	_1132600	GSA120T00001068001	GSA25T00023879001
PRP38B	_116253	285230	_1407300	GSA120T00014867001 GSA120T00000117001 GSA120T00001396001	GSA25T00022798001 GSA25T00014687001
U2 snRNP					
U2A	_003081	229210	_1369700	GSA120T00010616001	GSA25T00003636001
U2B	_003083	209690	_0935000	GSA120T00009931001	GSA25T00010392001
SF1	_004621	314860	_0623600	GSA120T00017418001 GSA120T00012361001	GSA25T00002046001
SF3A1	_001005409	246500	_1474500	GSA120T00009218001	GSA25T00001539001
SF3A2	_009096	228000	_0619900	GSA120T00015561001	GSA25T00018040001
SF3A3	_006793	221950	_0924700	GSA120T00002021001	GSA25T00013095001
SF3B1	_001005526	205010	_0308900	GSA120T00005385001	GSA25T00012209001
SF3B2	_006833	314740	_1461600	GSA120T00002421001	GSA25T00016286001
SF3B3	_036558	230960	_1234800	GSA120T00022307001	GSA25T00020920001
SF3B4	_005841	224580	_1420000	GSA120T00005800001	GSA25T00026577001

SF3B5	_116147	248250	_1018500	GSA120T00001050001	GSA25T00021783001
SFB125	_031398	ND	ND	GSA120T00022140001	GSA25T00011042001
SFB14	_057131	305010	_1224900	GSA120T00014493001	GSA25T00009327001
U2AF65	_001020374	234520	_1468800	GSA120T00005469001	GSA25T00025555001
U2AF35	_001020374	236910	_1119300	GSA120T00008087001	GSA25T00001003001
PUF60	_001258027	224850	_1224300	GSA120T00007081001	GSA25T00018129001
SPF30	_005862	286440	_0323500	GSA120T00024076001	GSA25T00020066001
SPF45	_116294	214820	_1454000	GSA120T00008338001	GSA25T00001210001
CHERP	_006378	321560	none	GSA120T00012641001	GSA25T00008586001
SR140	_001073884	240710	_1402700	GSA120T00003030001	GSA25T00011628001
PRP43	_001349	_233520 _312280 _263650	_0917600 _1030100	GSA120T00001099001 GSA120T00013085001 GSA120T00018299001	GSA25T00010117001 GSA25T00027084001 GSA25T00000348001
U5 snRNP					
DDX23/PRP28	_004809	298020	_0518500	GSA120T00011443001	GSA25T00013599001
CD2BP2	_006101	ND	_1031600	ND	ND
Snu114	_004238	_205470 _286080	_1003800 _1451100	GSA120T00024060001 GSA120T00010700001	GSA25T00020040001 GSA25T00003575001
Brr2	_054733	_249810 _233390	_1439100 _0422500	GSA120T00004033001 GSA120T00026232001 GSA120T00018887001	GSA25T00014432001 GSA25T00003083001
PRP6	_036601	205220	_1110200	GSA120T00011220001	GSA25T00000514001
PRP8	_006436	231970	_0405400	GSA120T00009175001	GSA25T00001593001
PRP8BP	_004805	310860	_0822800	GSA120T00011213001	GSA25T00000528001
DIB1	_006692	270140	_1231500	GSA120T00017514001	GSA25T00001939001

U6 core (stability and function of U6 snRNP)					
LSM2	_067000	297140	_0520300	GSA120T00011592001	GSA25T00014305001
LSM3	_055278	298970	_0819900	GSA120T00007224001	GSA25T00020303001 GSA25T00005714001
LSM4	_036453	278950	_1107000	GSA120T00002193001	GSA25T00020504001 GSA25T00027880001
LSM5	_036454	247610	_1443300	GSA120T00004015001	GSA25T00014455001
LSM6	_009011	261470	_1325000	GSA120T00001208001 GSA120T00013291001	GSA25T00025362001
LSM7	_057283	286560	_1209200	GSA120T00018185001	GSA25T00010229001
LSM8	_057284	272630	_0829300	GSA120T00011562001	GSA25T00014265001
hPrp19/CDC5 (specification of U5 and U6 interactions with RNA)					
PRPF19	_055317	320210	_0308600	GSA120T00014925001 GSA120T00003237001	GSA25T00022601001 GSA25T00004472001
CRNKL1	_057736	269200	_0403700	GSA120T00014810001	GSA25T00024249001
CDC5L	_001244	275480	_1033600	GSA120T00019196001	GSA25T00003897001
ISY1	_065752	203870	_1472000	GSA120T00020760001	GSA25T00017110001
BCAS2	_005863	243620	_0614400	ND	ND
XAB2	_064581	305240	_1235900	GSA120T00006726001	GSA25T00018347001
PLRG1	_002660	218420	_0302000	GSA120T00013703001	GSA25T00024611001
SYF2	_056299	ND	ND	ND	ND
SNW1	_036377	233190	_0218700	GSA120T00023369001	GSA25T00013244001
BUD31	_003901	246620	_0522800	GSA120T00021456001	GSA25T00002733001
PPIE	_006103	ND	ND	GSA120T00000182001	GSA25T00006816001 GSA 25T00027189001
CCDC12	_001264003	279430	_1451500	GSA120T00023334001	GSA25T00002423001

AQR	_055506	314410	_1352700	GSA120T00000623001	GSA25T00019323001
CWC15	_057487	270740	_0722500	GSA120T00017667001	GSA25T00005543001
PPIL1	_057143	270560	_0528700	GSA120T00023192001	GSA25T00021726001
Non-snRNP factors (second step factors) (RNA release)					
DHX16	_003578	ND	_1030100	GSA120T00013085001 GSA120T00001099001	GSA25T00027084001 GSA25T00000348001
DDX39B	_004631.1	ND	_0209800	GSA120T00015321001	GSA25T00012721001
DDX46	_001287789	ND	_0508700	GSA120T00001622001	GSA25T00024018001
SLU7	_006416	ND	_0610100	GSA120T00016851001	GSA25T00002586001
DHX38 (PRP16)	_054722.2	ND	_1364300	GSA120T00006589001	GSA25T00008585001
CDC40	_056975.1	ND	_1220100	GSA120T00002225001	GSA25T00020557001
PRPF18	_003666	ND	_0922700	GSA120T00006464001	GSA25T00000726001
SR and hnRNP family					
SRSF1	_008855	ND	_0517300	GSA120T00017256001	GSA25T00003447001 GSA25T00003651001
PTBP2	_067013	ND	_0606500	ND	ND
SRSF4	_005617	ND	_1022400	GSA120T00022021001 GSA120T00015640001	GSA25T00008688001
hnRNP A	_006796	264610	_0916700	GSA120T00001992001	GSA25T00013062001
hnRNP D0	_112738	265530	ND	GSA120T00021290001	GSA25T00016342001
hnRNP H	_005511	236540	ND	GSA120T00006941001	GSA25T00019779001
hnRNP M	_112480	262620	_1006800	GSA120T00021834001	GSA25T00022289001
hnRNP U	_114032	290270	ND	ND	ND

Table S7.

Assembly statistics for *Amoebophrya* A120 and A25 transcriptomes at different time of the infection within the host and at the free-living stage host (dinospore).

Time of infection (h)	Number of reads (M)	Assembled reads (%)	Number of contigs	Contigs average length (nt)	N50 contigs (nt)
A120					
Dinospore only	217	93	44,591	2,251	4,313
0h (host only)	183	90	202,829	945	1,581
6	136	90	203,593	909	1,543
12	144	87	210,414	825	1,386
18	143	89	222,810	922	1,556
24	134	89	225,120	890	1,510
30	153	90	225,189	986	1,705
36	166	92	222,213	976	1,693
A25					
Dinospore only	145	95	41,322	2,419	4,801
0 (host only)	144	83	186,115	777	1,269
6	131	81	190,459	807	1,341
12	151	85	198,130	852	1,425
18	131	82	178,304	744	1,170
24	158	86	200,408	843	1,356
30	155	89	228,002	871	1,461
36	157	91	239,274	937	1,594

42	157	91	234,415	995	1,685
44	144	91	228,678	964	1,617

Table S8.

Contigs alignments on different stages of infection for *Amoebophrya* A120 (A120) and A25 (A25). ND corresponds to "not determined" when no measure was done.

Time of infection (h)	**Number of contigs**		**Number of aligned contigs (%)**	
	A120	A25	A120	A25
Dinospore stage	44.591	41,322	41,810 (94%)	37,239 (90%)
0h (host only)	202,829	186,115	82 (0,04%)	0
6	203,593	190,459	11,592 (5,7%)	8,088 (4%)
12	210,414	198,130	14,781 (7%)	7,740 (3,9%)
18	222,810	178,304	29,270 (13%)	10,860 (6%)
24	225,120	200,408	38,252 (17%)	17,489 (8,7%)
30	225,189	228,002	44,571 (19,8%)	31,668 (14%)
36	222,213	239,274	46,519 (21%)	39,368 (16%)
42	ND	234,415	ND	41,240 (17,6%)
44	ND	228,678	ND	37,271 (16,3%)

Summary

Parasitism is a frequent lifestyle in nature and a major source of evolutionary pressure for both hosts and their parasites. Dinoflagellates are successful marine protists found in oceans worldwide, some of which are responsible for toxic blooms while others live in mutualistic relationships with myriad of corals. *Amoebophrya ceratii* species complex (Syndiniales) includes a large number of parasites which have the potential for regulating dinoflagellate blooms. A high sequence diversity has been observed for this group in both cultures and environmental investigations. This thesis was aimed to answer whether the sequence diversity represents the species diversity. Based on a polyphasic approach involving genetic and phenotypic characters applied on 119 closely related individuals, all able to infect the same host species (the bloom-forming dinoflagellate *Scrippsiella trochoidea*), I defined 8 ribotypes which likely correspond to different species. These results advocated for considering unique sequences (i.e., with any nucleotide differences) of 18S-V4 or 18S-V9 (small subunit ribosomal RNA genes) regions for species delimitation rather than grouping them into operational taxonomic units (OTUs). Then I investigated the existence of a set of genes specifically involved in meiosis in two fully sequenced genomes and thereby provided the *in silico* evidence that sexual reproduction may occur in *Amoebophrya*. I observed that these genes over-expressed during the free-living stage of the parasite, providing an interesting track to explore. Overall, this thesis offers new insights into the highly underestimated species diversity in *Amoebophrya* lineage and lays the basis for further study on their biological traits.

(Translation)

Le parasitisme est un style de vie fréquent dans la nature, et une force évolutive majeure pour les hôtes comme pour les parasites. Les dinoflagellés sont des protistes marins très répandus dans tous les océans, certaines espèces étant même responsables d'efflorescences algales toxiques tandis que d'autres vivent en symbioses mutualistes avec de nombreux coraux. Le complexe d'espèces *Amoebophrya ceratii* (Syndiniales) inclue de très nombreux parasites de dinoflagellés capables potentiellement de contrôler les efflorescences de ces dinoflagellés. Une très grande diversité a été observée au sein de ce groupe, soit en culture soit dans l'environnement. Ce travail de thèse a pour but d'étudier si la diversité de ces séquences correspond à la diversité au niveau spécifique. Sur la base d'une approche polyphasique faisant intervenir des caractères génétiques et phénotypiques appliqués à 119 individus proches phylogénétiquement, et tous capables d'infecter le même hôte (le dinoflagellé producteur d'efflorescence *Scrippsiella trochoidea*), j'ai défini 8 ribotypes qui correspondent vraisemblablement à des espèces différentes. Ces résultats prônent l'utilisation de séquences uniques (i.e., divergentes par un seul nucléotide) pour délimiter les espèces dans la région V4 ou V9 du 18S (la petite sous-unité du

ribosome) plutôt que de les grouper en unité opérationnelle taxonomique (OTUs). J'ai ensuite recherché l'existence d'une collection de gènes spécifiquement impliqués dans la méiose au sein de deux génomes de référence, et fournit des évidences *in silico* qu'une reproduction sexuée peut avoir lieu chez ce parasite. J'ai observé que ces gènes étaient surexprimés durant la phase libre du parasite, offrant une piste intéressante pour de prochaines études. En conclusion, ce travail de thèse offre de nouvelles perspectives concernant la diversité largement sous-estimée des *Amoebophrya* et posent de nouvelles bases pour l'étude de leurs traits biologiques.

Acknowledgements

I thank Laure Guillou. She provided me with this chance to do a PhD with her.

I thank Fabrice Not. Without his help for time scheduling, this thesis would not be finished in time.

I thank Harold Amen Moundoyi-mboungou, who helped me a lot to adapt to the new life when I came to Roscoff. He helped me with my visa and my phone number, etc. I'm thankful to him.

I thank the university students at UMPC enrolled in the year 2016 and 2017. They shared a lot of fun and activities with me. We had a party on the Ille de Batz, playing games and dancing. That's nice memory. I'm thankful to them.

I thank Ulysse Guyet, who is really a good friend both in life and in work. He helped me to fix my bike and took me to the supermarket. When I asked for help, he always lent me a hand. I'm thankful to him.

I thank Nicolas Henry. He helped me a lot with the R scripts when I got stuck. He is excellent at processing data.

I thank Pierre-Yves Mocaer, who is also a really good friend. During those days when we were AJC members, we had a lot of fun organizing activities together. He also helped with the registration at the UPMC. I'm thankful to him.

I thank Solene Breton. She invited me and the others to her birthday party. That's really nice. She even invited me for a coffee break. Although I rejected, I am really thankful to her.

I thank Martin Gachenot. He taught me how to do fishing in the sea and offered me drinks once.

I thank Laura Rubinat. She is shy but very kind. She once helped me with scripts. I'm thankful.

I thank Martina Strittmatter. When it's my birthday, she made cookies as gifts for me. I'm thankful.

I thank Florence le Gall, Priscillia Gourvil, Christophe Six, Domique Marie, Colomban de Vargas, Christian Jeanthon, Fabienne Rigaut-Jalabert, Morgane Ratin, Nathalie Simon, Ewen Corre, Miguel Méndez Sandín, Anne-claire Baudoux, Laurence Garczarek, and Frederic Partensky. They greet me with nice smiles when we meet on the way. I'm thankful to all of them.

Last but the most important, I thank my families, especially my father and mother. During the past three years, I have been working very hard. I have not gone back to see them since I started my PhD. But they always provide me with the strongest support. I'm thankful to them.

In short, many thanks to everyone who helped me and supported me to finish this PhD! Best wishes to all of you!

Ruibo

Curriculum vitae

Ruibo Cai

E-mail : rcai@sb-roscoff.fr; ruibo.dr@gmail.com

Education

09/2007-07/2011 Henan Normal University, Bachelor of Science, Xinxiang, China

Major: Biotechnology

2013-2016 Beijing Forestry University, Master of Agriculture, Beijing, China

Major: Conservation Genetics

2016-present, UPMC, PhD student, Paris, France

Publication

- *Ruibo Cai, Aaron B.A. Shafer, Alice Laguardia, Zhenzhen Lin, Shuqiang Liu, Defu Hu*[*]*. Recombination and selection in major histocompatibility complex of the endangered forest musk deer (Moschus berezovskii), Scientific Reports, 2015, 5; doi: 10.1038/srep17285.*
- *Xiaoning Sun,# Ruibo Cai,# Xuelin Jin, Aaron B. A. Shafer, Xiaolong Hu, Shuang Yang, Yimeng Li,1 Lei Qi,.Blood transcriptomics of captive forest musk deer (Moschus berezovskii) and possible associations with the immune response to abscesses. Scientific Reports, 2018, 1; doi: 10.1038/s41598-017-18534-0. (# co 1*[st]*-author)*
- *Ruibo Cai, Ehsan Kayal, Catharina Alves-de-Souzac, Estelle Bigeard, Erwan Corre, Christian Jeanthon, Dominique Marie, Betina Porcel, Raffaele Siano, Jeremy Szymczak, Matthias Wolf, Laure Guillou. Cryptic species in the parasitic Amoebophrya species complex revealed by a polyphasic approach. (in review)*
- *Sarah Farhat, Phuong Le, Ehsan Kayal, Benjamin Noel, Estelle Bigeard, Erwan Corre, Florian Maumus, Isabelle Florent, Adriana Alberti, Jean-Marc Aury, Tristan Barbeyron,* ***Ruibo Cai****, Corinne Da Silva, Benjamin Istace, Karine Labadie, Dominique Marie, Jonathan Mercier, Tsinda Rukwavu, Thierry Tonon, Catharina Alves-de-Souza, Pierre Rouzé, Yves Van de Peer, Patrick Wincker, Stephane Rombauts, Betina M. Porce, Laure Guillou. Rapid protein evolution and invasive intronic elements in two marine protistan parasites (submitted)*

Professional Experiences

07/2011-07/2012 Beijing Tianzhu Shidai Biological Technology Co., Ltd., Beijing, China

Promoted lab instruments to universities and institutes

08/2012-10/2012 Lab Assistant in Key Laboratory for Protein and Peptide Pharmaceuticals, National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China.

01/2013-04/2013 Lab Assistant in Institute of Zoology, Chinese Academy of Sciences, Beijing, China

08/2013-10/2013 Assistant Researcher in Professor Hu Defu's Lab, Beijing Forestry University, Beijing, China

Course done

- Discipline-unrelated courses:
  - Scientific career plan (3 days, 2017.10, Concarneau, 21H)
  - French(Roscoff, 2018.5-2018.9, 20H)
- Discipline-related courses:
  - UNIX (Roscoff, Abims, 2017)
    Module 1: Linux Avance, 29th, Nov, 2017, (7H)

    Module 2: Utilisation of Cluster, 4th Dec, 2017 (7H)
  - Population genomics: background and tools (CNR, Napoli, Italy, 2018.4.21-4.27, 40H)
  - Python (Roscoff, Abims, 2018/6.13-6.14, 6.20-6.21)
    Module 1: python 16-18 Jan. (21H)
  - R (Roscoff, Abims, 2018/6.11-6.12)
    Module 1: R initiation, 11th June 2018 (7H)

    Module 2: R advanced, 12th June 2018 (7H)

    Module 3: Python initiation, 13-14th, June 2018 (14H)

    Module 4: Python advanced, 20-21th, June 2018 (14H)
  - Concepts in ecology and NGS, Summer Colloquium 2018 (CSIC, Barcelona, Spain, 2018/7.9-7.13, 40H)
  - Prevention of incidental, biological and chemical risk (Roscoff, 1H)