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

Ruibo Cai, Ehsan Kayal, Catharina Alves-De-Souza, Estelle Bigeard, Erwan Corre, Christian Jeanthon, Dominique Marie, Betina M Porcel, Raffaele Siano, Jeremy Szymczak, et al.

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.

Competing Interests: The authors declare any competing financial and/or non-financial interests in relation to the work described

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, parasites of their dinoflagellate congeners. We used genetic and phenotypic characters, applied to 119 individuals sampled locally, to construct practical guidelines for species delineation that could be applied in DNA/RNA based diversity analyses. Based on internal transcribed spacer (ITS) regions, ITS2 compensatory base changes (CBC) and genome k-mer comparisons, we unambiguously defined 8 cryptic species between closely related ribotypes, that differed by less than 97% sequence identity in their SSU rDNA. We then followed the genetic traces of these parasitic species during a three-year survey of summer dinoflagellate blooms. We noticed that these cryptic species of *Amoebophrya* co-occurred and shared the same ecological niche. We also observed a maximal fitness for parasites having low to medium host range, reflecting an elevated cost for infecting a larger host range. This study suggests that a complete taxonomic revision of these parasitic dinoflagellates is long overdue to understand their diversity and ecological role in the marine plankton.

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

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 delineation 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 in marine protists, partially due to the low incidence of observed sexual recombination , 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 delineation 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 delineation is pervasive for parasitic lineages almost exclusively composed of environmental sequences, such as the Marine ALVeolate lineages (MALVs) (10)(11). MALV represents one of the most hyperdiverse lineages (> 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 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 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 delineation 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 descriptions are 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 newly-defined cryptic species of *Amoebophrya* 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 of *Amoebophrya* species, their ecological niches, and population fitness in the field.

Materials and Methods

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 the toxic blooms of the dinoflagellate *A. minutum* that occur during late spring-early summer (May to July) over eight 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) filtered 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 (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 (21,23,9).

Strains and single-cells isolation

Our strategy was to isolate representative phototrophic dinoflagellates from the Rance and the Penzé estuaries, as well as other estuarine systems nearby together with their local syndinid parasites. More details regarding isolation are described in the Supplementary Information. 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 over 3 months. The medium was supplemented with 5% (v/v) soil extract followed by a final filtration (0.22 µm size pore) under sterile conditions. Cultures were grown at 21°C under continuous

light at 100 µEinstein m^2 s^{-1} in vented flasks and a photoperiod of 12h. To maintain parasitic strains, infected hosts were regularly transferred (every 3-7 days) into a healthy host culture in 15 ml culture tubes using a 1/10 dilution rate.

For single-cells, hosts infected by *Amoebophrya*-like parasites in late-stages of infection were detected from freshly collected field samples (less than 3 hours) 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 (24)), then sorted individually by micropipeting, and washed three times into filter sterilized (< 0.22 µm) freshly prepared medium. Hosts were identified according to their morphology, and individual cells (single-cells) were transferred into cryovials with a minimum of medium (3-5 µL), flash-frozen, and stored at -80°C. DNA extraction and purification were performed both on pelleted strains and single-cells using the MasterPure kit (Epicentre).

Genome sequencing

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 (or clades) 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 in order 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, Courtaboeuf, France) 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 details regarding sequencing and genome assembly are described in the Supplementary Methods.

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 to capture the ribosomal operon and the genes of interest. A gene was discarded from the putative single-copy gene list either if i) it was detected in multiple copies using a reciprocal BLAST approach, or ii) had no hit. Genomic reads were then mapped to each of the best hits using Bowtie2 (25). 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 operons per genome. Doing so, we used an average of 21 genes per strain (minimum 7; maximum 55).

Full-length ITS2 sequences were directly annotated using Hidden Markov Models (HMMs) (26) as implemented in the ITS2 database (27) or by alignment to such annotated sequences. Secondary structures were predicted by homology modeling using a relevant template (e.g., (26,27) or by RNA structure using energy minimization and constraint folding (28,29). The predicted secondary structures for the ITS2 resembled the common core structure for eukaryotes (four helices, helix III the longest, helix IV short and divergent (30,25). Helices I, III and IV were highly variable, while Helix IV was lost in some ribotypes (RIB1, 4, and 6) (the inner circle from which the four helices emanate was forced, i.e., no base pairs were allowed between areas separating helices). The phylogenetic analysis of the ITS2 dataset followed the procedures outlined in (33). Specifically, a global multiple sequence-structure alignment was automatically generated in 4SALE v1.7 (32,33), whereby ITS2 sequences and their respective secondary structures were simultaneously aligned using a 12×12 ITS2 sequence-structure specific scoring-matrix (36). 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) or an ITS2 sequence-structure specific general time-reversible (GTR) substitution model, both implemented in ProfDistS v0.9.9 (37). Using the ITS2 sequence and secondary structure simultaneously (encoded by a 12-letter alphabet, (36)), a maximum parsimony tree (MP) was

reconstructed by PAUP (38) based on default settings. A sequence-structure maximum likelihood tree (ML) was calculated using the "phangorn" package (39) in R (40). The R script is available from the 4SALE homepage at http://4sale.bioapps.biozentrum.uni-wuerzburg.de (cf. (36)). Bootstrap support for the sequence-structure trees was estimated based on 100 replicates. A compensatory base change (CBC) table was transferred from 4SALE (35).

Genome comparison using SIMKA *k*-mer analysis

We used adapters and low complexity (i.e., Shannon index < 1.5) filtered DNA-seq reads as input to estimate the *k*-mer distribution of the various genomes with SIMKA ($k = 21$ bp; minimum read size ≥90 bp) (41). 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 (41)) that give more weight to the double presence of *k*-mers (i.e., Kulczynski, Ochiai, and Chord/Hellinger distances) (42). Statistical support 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 the clustering method, considering the above-estimated number of ribotypes as the desired number of clusters.

Cell morphology

In its initial description, Cachon (16) 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 (43). 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 (44)). Free-living stages of these parasites are very small (less than 3-5 µm in diameter), they have nearly the same morphology and very few discriminating characters. Additionally, the dinospore size changes very quickly after their release, with further cell division

happening over a period of a few days. This is the reason why we used the flow cytometry, which allows the rapid analysis of large populations, rather than time-consuming techniques like electronic microscopy, which focused on few cells only. We estimated some of the morphological cell signatures of the cultured strains by flow cytometry using the side scatter (SSC) and the forward scatter (FSC) parameters, as well as the natural green autofluorescence of *Amoebophrya* spp. dinospores when excited by light at 405 nm wavelength (45). 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 (46), 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 through cross-infecting experiments using locally-occurring dinoflagellate strains belonging to three different genera and nine different genetic clade/species, all isolated from the Rance, Penzé and nearby estuaries during the same period of time that the parasitic strains were isolated (Table S1, Fig. S2). Freshly produced dinospores were collected by filtration through 5-µm pore-sized cellulose acetate filters (Minisart, Sartorius, Germany).100 µl aliquots of this filtrate were then inoculated into 1 ml of several exponentially growing dinoflagellate strains 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). All cross-infections were processed 3 to 5 times at different dates.

Environmental metabarcoding survey

We obtained environmental rDNA metabarcode sequences of 48 samples collected in the Penzé estuary during late spring-early summer for three consecutive years (2010-2012). The DNA extraction method was based on a phenol-chloroform protocol (47); the universal TAReuk454FWD1

(5'-CCAGCASCYGCGGTAATTCC-3') and the modified reverse BioMarKs (5'-ACTTTCGTTCTTGATYRATGA-3') primers (48) 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 (49) implemented by the PR2 reference database (50) modified to take into account the *Amoebophrya* species boundary thresholds detected here.

Statistical analyses

All the statistical analyses described below were performed in R software 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 basic '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 (51). 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 (52) on standardized variables (between 0 and 1) based on their minimum and maximum values (53). 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 (54) was first performed to determine the niche position and niche breadth of *Amoebophrya* ribotypes using the function *niche* in the ‘ade4’ package (55). 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 descriptors [i.e., water temperature, salinity, precipitation, tide coefficient, NO_3, PO_4 and $Si(OH)_4$] were included in two separate matrixes (N = 48). Before analysis, relative read abundances were Hellinger transformed (56) whereas the environmental descriptors were standardized to values between 0 and 1 (53). 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 (59). The niche overlap was then estimated by the comparison of the realized niches (i.e., kernel densities) through the calculation of the *D* metric (60) for each pair of *Amoebophrya* ribotypes using the *ecospat.niche.overlap* function from the ‘ecospat’ package (61). 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-infections 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 (N = 18), based on i) their maximal normalized relative read abundance and ii) persistence in the system (e.g., the number of consecutive days in which the non-normalized relative contribution of the ribotype to the total number of reads

was higher than 10%). We then determined if these two fitness indicators were different between groups of *Amoebophrya* ribotypes representing different host ranges (based on the maximal number of infected host species in the cross-infection experiment for each ribotype). This was assessed by performing Kruskal-Wallis tests using the *kruskal.test* function in the basic 'stats' package following [log (x+1)] transformation. In the cases where the Kruskal-Wallis test was significant, the Dunn test was performed as a post-hoc analysis with the *dunnTest* function in the 'FSA' package.

Results and discussion

Ribotypes as cryptic species

We amplified and sequenced part of the ITS1-5.8S-ITS2 region from *Amoebophrya*-like 76 strains in culture and 43 environmental "single-cell" samples (Table S1). The alignment based on the secondary structure of the ITS2 region clustered individuals into eight main ribotypes (RIBs 1-8, Fig.1A-C). These ribotypes displayed low sequence intra-variability (<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 96-100% pairwise sequence identities, but only 93-94% 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 the V4 region and one in the 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 (62). According to Müller et al. (63), CBCs found in the ITS2 region of the rDNA of two seemingly-related specimens correlate (with a probability of 0.93) to the biological species concept (interbreeding populations generating fertile offspring and reproductively isolated from others) of

species (64), whereas the absence of CBC might suggest 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 practical alternative for indicating the potential for discriminating 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 (i.e., the 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 range within ribotypes that clustered into two main groups with no significant intra-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 a genome size disparity likely prevents any sexual reproduction between these two groups. We additionally estimated the number of ribosomal operons per genome ranged 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). Using the DNA-seq reads acquired for 67 individuals (17 of which were environmental ”single-cell” samples), we observed that strains sharing the same ribotype are part of the same cluster estimated by their *k*-mer distribution (Table S2) with high bootstrap support (>90%; Fig. 1A, E). The results of the *k*-mer analysis suggest a low gene flow, if any, between ribotypes. Together our results are consistent with placing each ribotype into a separate cryptic species, awaiting for more formal description.

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 side scatter (SSC) and the forward scatter (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 within a 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 *Amoebophrya* ribotypes during the survey. For that, we made a strong effort in having strains (for both the parasites and their hosts) that co-occurred in the same (or similar environments) and isolated at the same period of the year. As a result, representatives of the three most abundant phototrophic dinoflagellate genera (53 local strains and 9 species/genetic clades) have been isolated and cross-infected in the laboratory with 36 *Amoebophrya* strains (Fig. 2F). All *Amoebophrya* strains can infect the same strain of *Scrippsiella acuminata* STR1, an autotrophic dinoflagellate species ubiquitous in both localities and used as the main host in cultures. Ribotypes 1, 3, 6 and 7 only infected a single dinoflagellate species (i.e., *Scrippsiella acuminata* STR1), while others infected several species in the same *Scrippsiella* genus (RIB5) or even another genera (RIB2 and RIB4 infecting both *Scrippsiella* and *Heterocapsa*; Fig. 2F). 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. The host spectrum is often considered as more permissive in culture experiments compared to the natural environment (70), while higher genomic diversity exists and potentially extends or reduces the host range from that observed in the laboratory. By identifying individualized infected host cells in environmental samples using microscopy, 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, the most closely related ribotypes RIBs 3 and 8 (based on their rRNA sequences), which are considered as different cryptic species based on CBC and *k*-mer analysis, also differed by their host range. As these two ribotypes could not be discriminated on their V4 region is also indicating that rRNA may not be variable enough to address real diversity of *Amoebophrya* lineage.

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 the new species boundaries to environmental data

As a case study, we applied the newly defined *Amoebophrya* cryptic species boundaries to a metabarcoding survey performed during dinoflagellate blooms in the Penzé estuary at late-spring/early-summer time over three consecutive years (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, 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. The niche analysis based on the outlying mean index (OMI) pointed out a substantial interannual variability (Fig. 3B) mainly correlated 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 that took into consideration the niche overlap between parasites and other dinoflagellate 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 (based on the number of hosts detected in the cross-infection experiments and single-cell infections) 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. For the maximal normalized abundance, we did not observe significant difference between ribotypes with low or medium number of hosts (i.e., 1 and 3 hosts, respectively; Dunn test, $p = 0.59$), whereas the maximal normalized head abundances were significantly lower for ribotypes with high number of hosts (i.e., 4–5 hosts) when compared to ribotypes with both low and medium number of hosts (Dunn test, $p < 0.05$) (Fig. 3E). Similar results were obtained when comparing the persistence of the ribotypes in the system, with no differences between ribotypes with low and medium host numbers (Dunn test, $p = 0.13$). However, only ribotypes with medium number of hosts showed higher persistence in the system when compared with ribotypes with hight host numbers (Dunn test, $p < 0.05$). 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, where lower fitness is leveraged by the more generalistic parasitic species/strains.

Conclusions

Here, we provide molecular evidence for the presence of at least eight *Amoebophrya* ribotypes in the Penzé estuary, with genome *k*-mer comparisons and CBC analyses, supporting their classification into individual 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 delineation in the Amoebophryidae clade and that nucleotide differences in the SSU rDNA gene sequence might be enough to delineate putative cryptic species. These results advocate for the use of unique sequences (i.e., 100% threshold sequences similarity) rather than grouping them into OTUs during barcoding studies when using this genetic marker. Considering the diversity of MALV-II lineage in marine waters, a full reassessment of their taxonomy is needed to understand their biogeography and ecology. Applying this novel species definition over a three-year monitoring survey in the Penzé estuary, we observed that most of these cryptic species co-occurred during dinoflagellate blooms, likely competing for similar ecological niches and host resources. We also reported an inverse pattern between population fitness and host range, where the maximal fitness values were observed for the *Amoebophrya* ribotypes having low or intermediate number of hosts, highlighting an elevated cost for infecting a larger host range.

Acknowledgements

We warmly thank Karen Lebret for her help with metabarcoding sequencing, Koester Julie A. for help with English revision of the text, and Ramon Massana for providing unpublished MALASPINA sequences used in Figure S1. 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). This work was promoted in the frame of the GDR Phycotox.

Author Contributions

LG conceived this study. LG, CAdS, EB, CJ, DM, RF participated to 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, & CAdS wrote the paper. All authors edited and approved the final version of this paper.

Data availability

Raw data are available upon request or using the following link: http://application.sb-roscoff.fr/project/hapar. All strains have been deposited at the Roscoff Culture Collection (RCC, http://roscoff-culture-collection.org/).

References

1. Sherr BF, Sherr EB, Caron DA, Vaulot D, Worden AZ. Oceanic Protists. Oceanography. 2007;20(2):130–4.
2. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. Protists are microbes too: A perspective. ISME J. 2009;3(1):4–12.
3. Ruhl MW, Wolf M, Jenkins TM. Compensatory base changes illuminate morphologically difficult taxonomy. Mol Phylogenet Evol; 2010;54(2):664–9.
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(6):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. 2015;348(6237):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(142).
7. Caron DA, Hu SK. Are we overestimating protistan diversity in nature? Trends Microbiol.; 2019;27(3):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(2):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(1841).
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(6820):603–7.
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;10:3349–65.
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(3):734–7.

13. Pernice MC, Logares R, Guillou L, Massana R. General Patterns of diversity in major marine microeukaryote lineages. PLoS One. 2013;8(2):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(6):469–74.

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):1–8.

16. Cachon J. Contribution à l'étude des péridiniens parasites. Cytologie, cycles évolutifs. Ann des Sci Nat Zool Paris. 1964;(12ème série):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;528:520–8.

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;8:267–78.

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):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(7600):465–70.

21. Chambouvet A, Morin P, Marie D, Guillou L. Control of toxic marine dinoflagellate blooms by serial parasitic killers. Science. 2008;322(5905):1254–7.

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. 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 *Alexandrium*

minutum. Mol Ecol. 2014;23:549–60.

24. Coats DW, Bockstahler KR. Occurrence of the parasitic dinoflagellate *Amoebophrya ceratii* in Chesapeake Bay populations of *Gymnodinium sanguineum*. J Eukaryot Microbiol. 1994;41(6):586–93.

25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

26. 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(1–2):50–7.

27. Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F. ITS2 database V: Twice as much. Mol Biol Evol. 2015;32(11):3030–2.

28. 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;11:1616–23.

29. 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(SUPPL. 1):377–80.

30. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA. J Mol Biol. 1999;288(5):911–40.

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

32. 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;2:361–4.

33. Schultz J, Wolf M. ITS2 sequence-structure analysis in phylogenetics: A how-to manual for molecular systematics. Mol Phylogenet Evol. 2009;52(2):520–3.

34. Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M. 4SALE — A tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformatics. 2006;7:498.

35. 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.

36. 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(2):145–9.

37. Wolf M, Ruderisch B, Dandekar T, Schultz J, Müller T. ProfDistS: (profile-) distance based phylogeny on sequence - Structure alignments. Bioinformatics. 2008;24(20):2401–2.

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

39. Schliep KP. phangorn: Phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.

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

41. 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.

42. Legendre P, Legendre LF. Numerical ecology. vol 24. Elsevier; 2012. 852 p.

43. Figueroa RI, Garcés E, Massana R, Camp J. Description, Host-specificity, and strain selectivity of the dinoflagellate darasite *Parvilucifera sinerae* sp. nov. (Perkinsozoa). Protist. 2008;159(August):563–78.

44. Lepelletier F, Karpov S a., Alacid E, Le Panse S, Bigeard E, Garcés E, et al. *Dinomyces arenysensis* gen. et sp. nov. (Rhizophydiales, Dinomycetaceae fam. nov.), a chytrid infecting marine dinoflagellates. Protist; 2014;165(2):230–44.

45. Coats DW, Bockstahler KR, Berg GM, Sniezek JH. Dinoflagellate infections of *Favella panamensis* from two North American estuaries. Mar Biol. 1994;119:105–13.

46. Marie D, Partensky F, Simon N, Guillou L, Vaulot D. Flow cytometry analysis of marine picoplankton. In: Diamond RA, DeMaggio S, editors. Living Colors: protocols in cytometry and cell sorting. Springer Verlag; 2000. p. 421–55.

47. 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(7):2942–51.

48. 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(1):fiw200.

49. 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(23):7537–41.

50. 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(November 2012):597–604.

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

52. Oksanen J, Kindt R, Legendre P, O'Hara B, Henry M, Stevens H. The vegan package. Community Ecol Packag. 2007;10:631–7.

53. 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(4):698–714.

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

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

56. Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. Oecologia. 2001;129:271–80.

57. 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;21:481–97.

58. 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.

59. Duong T. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. J Stat Softw. 2007;21(7):10.18637/jss.v021.i07.

60. Schoener TW. Nonsynchronous spatial overlap of lizards in patchy habitats. Ecology. 1970;51:408–18.

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

62. Gutell RR, Larsen N, Woese CR. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol Mol Biol Rev. 1994;58(1):10–26.

63. Müller T, Philippi N, Dandekar T, Schultz RG, Wolf M. Distinguishing species. RNA. 2007;13:1469–72.

64. Mayr E. The growth of biological thought diversity, evolution, and inheritance. Harvard University Press. 1982;896 pp.

65. 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(2):193–207.

66. 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(8):1531–43.

67. Annenkova NV, Hansen G, Moestrup Ø, Rengefors K. Recent radiation in a marine and freshwater dinoflagellate species flock. ISME J. 2015;9(8):1821–34.

68. 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(5):612–35.

69. 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(1):104–7.

70. Poulin R, Keeney DB. Host specificity under molecular and experimental scrutiny. Trends Parasitol. 2008;24(1):24–8.

71. Mai JC, Coleman AW. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. J Mol Evol. 1997;44:258–271.

Figure 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 from (A) using a GTR substitution model. Bootstrap values >50 derived from NJ, maximum parcimony (MP)- and maximum likelihood (ML) analyses are mapped to above, below, and to the right of the nodes, respectively. (C) An example of ITS2 secondary structure from the *Amoebophyra* RIB2 clade. Helices are numbered from I to IV according to Mai and Coleman (71). (D) Matrix of compensatory base changes (CBCs) between the eight *Amoebophyra* 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 seven (RIBs 1-7) out of the eight *Amoebophrya* ribotypes isolated 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 cross infection experiments where 36 strains of *Amoebophrya* were exposed to 54 host strains belonging to 9 dinoflagellate 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 the *envfit* function from the 'vegan' package) 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 late spring-early summer dinoflagellate blooms.**

(A) Relative abundance (in % of total reads) of *Amoebophrya* ribotypes in the Penzé Estuary (late spring-early summer 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 descriptors (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) Boxplots showing the relationship between the host range (number of host infected by each ribotype detected in the cross-infection experiments) and the field population fitness, defined by the normalized maximal abundance of ribotypes (E) and their permanence in days in the ecosystem (F). Horizontal lines indicate the median for the different descriptors. The red brackets indicate the significant differences between clusters pointed out by the Dunn test. (* $p < 0.05$).