Received: 11 January 2021 | Revised: 31 May 2021 | Accepted: 6 July 2021

DOI: 10.1111/1755-0998.13465

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

pr2-primers: An 18S rRNA primer database for protists

Daniel Vaulot[1,2] | Stefan Geisen[3,4,5] | Frédéric Mahé[6,7] | David Bass[8,9]

[1]UMR 7144, ECOMAP, Station Biologique de Roscoff, CNRS, Sorbonne Université, Roscoff, France

[2]Asian School of the Environment, Nanyang Technological University, Singapore, Singapore

[3]Department of Terrestrial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands

[4]Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands

[5]Nanjing Agricultural University, Nanjing, China

[6]CIRAD, UMR PHIM, Montpellier, France

[7]PHIM, CIRAD, INRAE, Institut Agro, Univ Montpellier, Montpellier, France

[8]Centre for Environment, Fisheries and Aquaculture Science, Weymouth, UK

[9]Department of Life Sciences, The Natural History Museum, London, UK

Correspondence
Daniel Vaulot, UMR 7144, ECOMAP, Sorbonne Université, CNRS, Station Biologique de Roscoff, 29680 Roscoff, France.
Email: vaulot@gmail.com

Abstract

Metabarcoding of microbial eukaryotes (collectively known as protists) has developed tremendously in the last decade, almost solely relying on the 18S rRNA gene. As microbial eukaryotes are extremely diverse, many primers and primer pairs have been developed. To cover a relevant and representative fraction of the protist community in a given study system, an informed primer choice is necessary, as no primer pair can target all protists equally well. As such, a smart primer choice is very difficult even for experts and there are very few online resources available to list existing primers. We built a database listing 285 primers and 83 unique primer pairs that have been used for eukaryotic 18S rRNA gene metabarcoding. *In silico* performance of primer pairs was tested against two sequence databases: PR^2 version 4.12.0 for eukaryotes and a subset of SILVA version 132 for bacteria and archaea. We developed an R-based web application enabling browsing of the database, visualization of the taxonomic distribution of the amplified sequences with the number of mismatches, and testing any user-defined primer or primer set (https://app.pr2-primers.org). Taxonomic specificity of primer pairs, amplicon size and location of mismatches can also be determined. We identified universal primer sets that matched the largest number of sequences and analysed the specificity of some primer sets designed to target certain groups. This tool enables guided primer choices that will help a wide range of researchers to include protists as part of their investigations.

KEYWORDS
18S rRNA, database, metabrcoding, PCR, primers, protists, R, shiny

1 | INTRODUCTION

Microbes are key players in all Earth ecosystems. Among them are protists that encompass all unicellular or unicellular-colonial eukaryotes, excluding some fungi. Protists perform a range of functions from photosynthesis to organic matter degradation. Although some eukaryotic groups such as unicellular algae (e.g., phytoplankton) have a long tradition of being studied as key players in marine primary production, the importance of protists in other processes and other environments has only been recently recognized, for example their role in nutrient cycling in soils or as symbionts and phagotrophs in marine waters (Geisen, Mitchell, et al., 2018; Worden et al., 2015). This late recognition stems in part from the inherent difficulties of visually identifying them and growing them in culture. In recent years, the development of metabarcoding has provided new tools to study protist diversity and ecology.

Metabarcoding is defined (Taberlet et al., 2012) as the use of a specific marker gene to analyse the composition of natural communities in a specific environment (water, soil, animal gut, faeces, etc). After DNA extraction, the gene is amplified using a pair of primers targeting one specific region, samples are labelled with tag sequences and the resulting DNA is sequenced using a high throughput technology, mostly Illumina currently. This approach was initially developed for bacteria (Sogin et al., 2006) and expanded later for

protists (Amaral-Zettler et al., 2009; Stoeck et al., 2009). The gene most commonly used is the small subunit ribosomal RNA gene (SSU rRNA: 16S rRNA for archaea and bacteria, 18S rRNA for eukaryotes). The SSU rRNA gene is composed of conserved regions that can be used to design general primers and variable regions (V) that can be used to assign taxonomy and design specific probes. In bacteria, the regions targeted are very often V3/V4 or V4/V5 (Parada et al., 2016), although other regions have been suggested as providing better resolution (e.g., Bukin et al., 2019). For eukaryotes, two variable regions of the 18S rRNA gene have mostly been targeted, the V4 and V9 regions: the V4 region is located in the second quarter of the 18S rRNA gene and the V9 region at the end of the 18S rRNA gene, near the internally transcribed spacer (ITS) region. Initially, the V9 region was favoured because of the limitation in sequence size (Amaral-Zettler et al., 2009; Stoeck et al., 2009): for example, initially Illumina sequences were restricted to 2 × 75 bp. However, with the development of the Illumina MiSeq (up to 2 × 300 bp), the V4 region is now preferred, in particular because it is longer, more variable, and better covered in reference databases (Pawlowski et al., 2012). Other eukaryotic genes, and in particular the mitochondrial cytochrome oxidase 1 gene (COI or *cox1)*, have been used for Metazoa (Valentini et al., 2009) but their use is debated in particular because of the lack of universal primers (Andújar et al., 2018; Deagle et al., 2014) and the absence of this gene in lineages that have lost the mitochondrial genome (e.g., Yahalomi et al., 2020). For protists, the 18S rRNA gene appears to be most appropriate as a general marker (Pawlowski et al., 2012), although other genes such as *rbcL* (large subunit of the RUBISCO) have been used for targeting photosynthetic organisms (e.g., Pujari et al., 2019).

Primer selection is critical to obtain an accurate taxonomic profiling of protist communities. Each primer (forward and reverse) must amplify the target community with minimal biases. The region amplified must be long enough to differentiate between closely related taxa by including enough variable positions. Preferably, it should also be short enough to be fully sequenced by the chosen technology, although longer amplicons can be also be partially sequenced. With Illumina sequencing being now the preferred technology, amplicon size must be ideally (although this is not absolutely necessary, see Lambert et al., 2019; Needham & Fuhrman, 2016) about 50 bp smaller than the sum of the forward and reverse sequences (called R1 and R2) to allow enough overlap to reconstruct the complete amplicon: for example, the Illumina MiSeq 2x300 bp chemistry can sequence amplicons of up to 550 bp. A large diversity of primer and primer sets targeting the 18S rRNA gene have been developed over the years, although a only a small number of these dominate in protist metabarcoding studies. Few resources are available that list eukaryotic 18S primers and primer pairs, provide information on their taxonomic specificity, and allow testing of new primer pairs. Most existing primer databases do not focus on protists. For example, the primer database linked to the Barcode of Life Data System project (https://boldsystems.org/index.php/Public_Primer_PrimerSearch) focuses on metazoans, and Probebase (http://probebase.csb.univie.ac.at/node/8, Greuter et al., 2016) focuses on bacteria. A few programming tools have been developed to test primer set specificity, for example ECOPCR (Ficetola et al., 2010), a PYTHON program, or R libraries such as PRIMERMINER (Elbrecht & Leese, 2017). The phylogenic program ARB offers a function to design and test probes and primers (Ludwig, 2004). Unfortunately, these tools need to be installed in a specific computing environment and require some background programming skills. Many existing online tools such as PROBEMATCH (https://rdp.cme.msu.edu/probematch/search.jsp) only allow testing primer sets against bacteria, archaea and fungi. Silva TESTPRIME (https://www.arb-silva.de/search/testprime) is the only tool that covers protists. It provides very detailed feedback on the taxonomy of amplified sequences, and the location of mismatches. Such detailed information comes at the expense of speed, with a typical test needing a few minutes to run. Moreover, the taxonomic annotation of the Silva database for protists is not optimal at this time, particularly for environmental sequences which are often only assigned at the class level or above (for example "Chrysophyceae;uncultured;eukaryotic picoplankton environmental sample").

To fill this gap and to provide protist researchers with a usable tool, we constructed a database of primers and primer sets used for eukaryotic 18S rRNA metabarcoding. These primer sets were tested *In silico* against the PR^2 database (Guillou et al., 2013) that contains more than 180 000 18S rRNA sequences with expert taxonomical annotation and a subset of the Silva database for archaea and bacteria. We developed an R-based web application that allows exploration of the database, to visualize precomputed *In silico* amplification results according to taxonomy (% of amplification, size of amplicons and location of mismatches), and to test any user-defined primer set.

2 | MATERIALS AND METHODS

18S rRNA gene primers (Table S1) and primer sets (Table S2) used in metabarcoding studies were collected from the literature. Primer sequences and primer sets (knowing that several primer sets may share at least one primer) were stored in a *MySQL* database. Primer sets were tested by performing *In silico* amplification of eukaryotic sequences stored in the PR^2 reference database (Guillou et al., 2013) version 4.12.0 (https://github.com/pr2database/pr2database/releases/tag/v4.12.0). We also used a small subset of the Silva database version 132 provided by the mothur website (https://mothur.org/wiki/silva_reference_files) containing 8517 bacteria and 147 archaea sequences to test whether these two groups were amplified. Database sequences with ambiguities were discarded (any nucleotide that is not A, C, G or T). Sequences with length shorter than 1350 bp were not considered except for the V4 region, for which this threshold was lowered to 1200 bp, since most sequences in PR^2 contain the V4 region. In contrast, this limit was extended to 1650 for the V9 region and since many 18S rRNA do not cover the full V9 region, we only kept sequences that contained the canonical sequence GGATC[AT] which is located at the end of the V9 region, just before the start of the internally transcribed spacer 1 (ITS1). An R

(R Development Core Team, 2013) script using the BIOSTRINGS package (Pagès et al., 2020) was used to compute the number of mismatches to the forward and reverse primers, allowing for a maximum of two mismatches for each primer using the function matchPattern with the following parameters: max.mismatch=2, min.mismatch=0, with.indels=FALSE, fixed=FALSE, algorithm="auto". We computed the position of mismatches using the mismatch function with parameter fixed=FALSE. A faster version of the script is also available that does not compute mismatch position using the vectorized form of the matchPattern function (vmatchPattern). The latter function is used in the SHINY application (see below) allowing users to test their own primer or primer sets. The data were tabulated using the DPLYR package and plotted using the GGPLOT2package (Wickham, 2016). An R SHINY application to interact with the database was developed using the following R packages: SHINY, SHINYFEEDBACK and SHINYCSSLOADERS (Sali & Attali, 2020).

3 | RESULTS AND DISCUSSION

3.1 | Database of primers and primer sets

We were able to recover a total of 108 general eukaryotic primers and 177 primers specific to some taxonomic groups from the literature (Table 1 and Table S1, https://app.pr2-primers.org). Some of these primers were designed early on when researchers began to amplify and sequence the 18S rRNA gene (e.g., Medlin et al., 1988). More recently, researchers have been designing primers specific to some taxonomic groups, mostly targeting phylum level (e.g., S19F and S15rF for Foraminifera, Morard et al., 2011) or class level (e.g., primer PRYM03+3 for Prymnesiophyceae, Egge et al., 2013). Some primers were also designed to block specific taxa (e.g., 18SV1V2Block against the coral *Pocillopora damicornis*, Clerissi et al., 2018) to be used in combination with more general primers (18SV1V2F in this case), or to avoid amplification of some groups (e.g., EUK581-F and EUK1134-R which do not amplify Metazoa; Carnegie et al., 2003). These have been modified and adapted for high throughput sequencing of the eukaryotic microbiome of eukaryotic organisms (e.g., corals, oysters) to avoid amplification of host genes (Bass & del Campo, 2020).

We identified a total of 83 unique primer sets (pairs) that have been used in metabarcoding studies (Table S2). Not all primers have been used for metabarcoding, in particular those that amplify the whole 18S rRNA gene, such as EukA and EukB (Medlin et al., 1988). Most metabarcoding primer sets do not target specific groups. The localization of the broadly-targeted primer sets over the 18S rRNA gene is quite diverse, but the vast majority target the V4 region (Table 2 and Figure 1). In contrast, the number of primer sets targeting the other favoured metabarcoding region V9 is much lower. Most of the primer sets targeting a specific taxonomic group are located in the V4 region, and none are in the V9 region (Table 2). In terms of usage, the V4 region is much more popular (about 80% of published studies in marine systems; Lopes Dos Santos et al., 2022), the three most commonly used primer sets being no. 8 (TAReuk454FWD1 and TAReukREV3; Stoeck et al., 2010), no. 17 (E572F and E1009R; Comeau et al., 2011) and no. 16 (TAReuk454FWD1 and V4 18S Next. Rev; Piredda et al., 2017), while for the V9 region the most popular sets are no. 27 (1391F and EukB; Stoeck et al., 2010) and no. 28 (1380F and 1510R; Amaral-Zettler et al., 2009).

TABLE 1 Summary of primers listed in the pr2-primers database. General primers target all eukaryotes and specific primers only certain taxonomic groups

Direction	General primers	Specific primers
fwd	55	89
rev	53	88
Total	108	177

3.2 | Testing primer sets by *In silico* matching

3.2.1 | General primer sets

We used the PR^2 database (Guillou et al., 2013) which currently contains about 180 000 18S rRNA sequences with detailed taxonomic annotations to test all primer sets from the pr2-primers database. We also determined, using a set of more than 8500 sequences representative of diverse archaeal and bacterial groups, whether these primers amplified bacteria or archaea. We only used long sequences (see Section 2) and allowed for a maximum of two mismatches on both forward and reverse primers, that is, a maximum of four mismatches. For general primers, amplification success varied from 32 to more than 97% (Table S3, Figure 2 and Figure S1). In general, the reverse primer had a tendency to have more mismatches than the forward primer (Table S3). Primer sets targeting regions other than V4 or V9 did not perform as well in general (Figure S1), although the best overall performance was for no. 76 targeting the V7 region (F-1183 and R-1443, 97.1% of sequences amplified, Lundgreen et al., 2019). If we focus on the V4 and V9 regions (Figure 2), the best performing primer sets overall were number 6 (616*f and 1132r, 96.5%; Hugerth et al., 2014) and number 29 (1389F and 1510R, 79.8%; Amaral-Zettler et al., 2009). Interestingly, the original study describing this primer set also used another forward primer (1380F, primer set no. 28) on the same samples and recommended using both forward primers together; although this advice which was not followed in subsequent studies (but see Lie et al., 2014). The lower percentage observed for the V9 primers should be interpreted with caution: many 18S reference sequences do not extend to the end of the V9 region and therefore will miss the signature of the reverse primer. To minimize this problem, we retained for the analysis of V9 primer sets only sequences that contain the canonical signature GGATC[AT] located at the 3' end of the V9 region. Despite performing well when allowing for four mismatches, some of these primer sets have at least one mismatch to PR^2 sequences: for example, primer set no. 108 (545F and 1119R, Kataoka et al., 2017) amplifies only 7.9% of the sequences with zero mismatch. Another important consideration is the

TABLE 2 Regions of the 18S rRNA gene targeted by the primer sets from the pr2-primers database

Gene region	General primer sets	Specific primer sets
37F		1
37F–41F		2
V1–V2	1	1
V1–V3		1
V2		3
V2–V3	1	3
V3		1
V3–V4		2
V4	32	15
V4–V5	1	
V5		3
V5–V7	1	
V5–V9		2
V6		1
V6–V8	1	
V7	2	
V7–V8		1
V7–V9	1	1
V8–V9	2	
V9	4	

size of the amplicon. Since most metabarcoding studies currently use Illumina sequencing technology, the maximum possible size to allow some overlap between the two R1 and R2 reads is about 550 bp (assuming that one uses the 2 x 300 bp sequencing kits), although smaller amplicons are preferable to allow more overlap. A sizeable fraction of the primer sets produce amplicons close to or larger than 600 bp (Figure 2). The post sequencing analysis strategy in this case would be to only use one of the reads (R1 is in general less noisy) without trying to assemble R1 and R2 (Lambert et al., 2019) or to assemble the nonoverlapping R1 and R2 reads with an intercalated N base (Needham & Fuhrman, 2016).

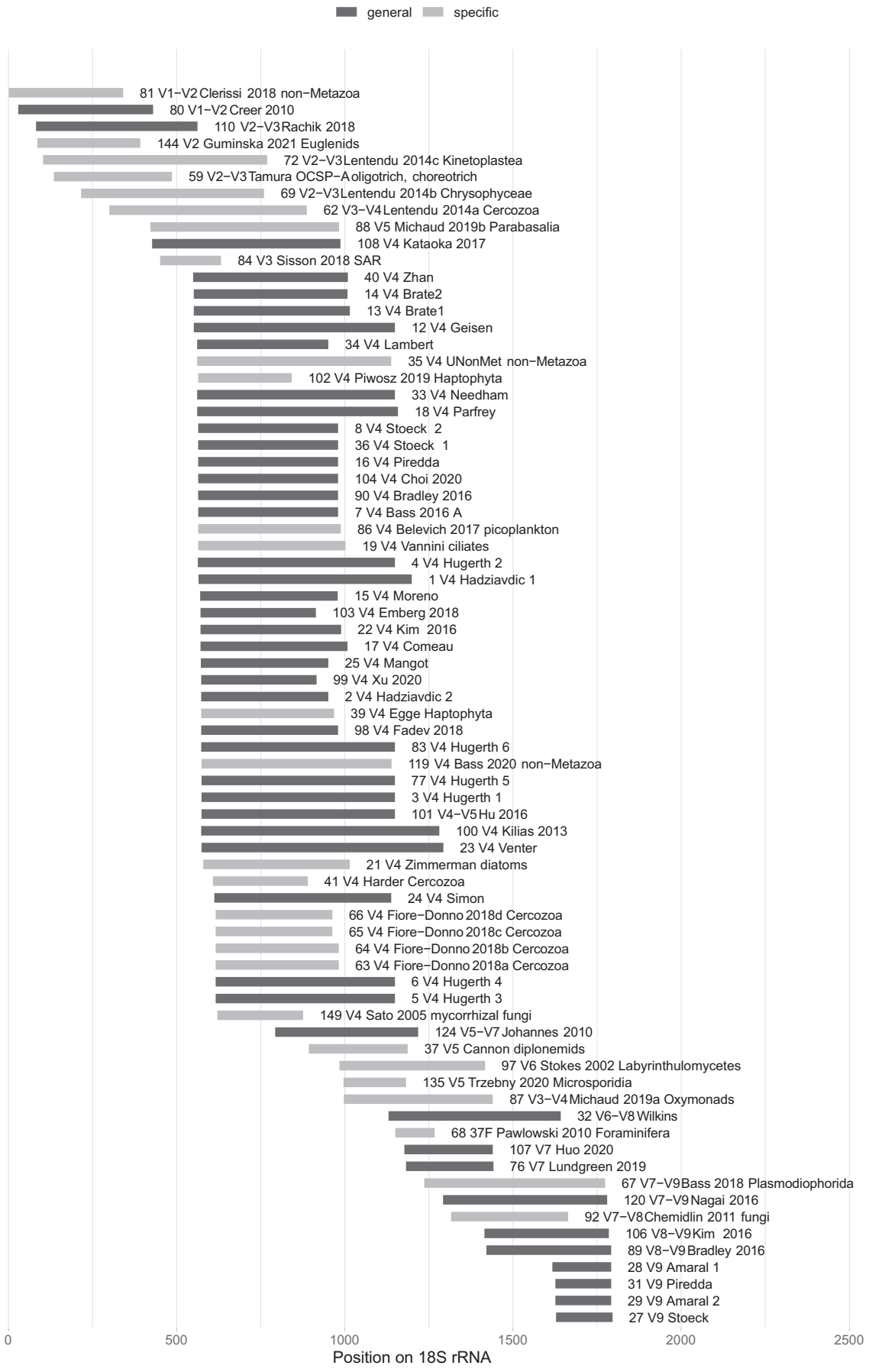
Another important consideration is whether amplification is similar across the whole eukaryotic taxonomic range. Taking as an example the most frequently used primer set targeting V4 (no. 8, Figure 3a) and looking at the amplification efficiency at the supergroup level, a significant fraction of Excavata and to a smaller extent of Rhizaria present at least five mismatches to this primer set (Figure 3a top-left). Amplification is even more unlikely for sequences presenting mismatches with the forward primer because the mismatches are located at the 3' end of the primer (Figure 3a top-right) which is the most unfavourable situation (mismatches at the 5' end are better tolerated). The average size of the amplicon also varies depending on the taxonomic group (Figure 3b bottom). For example, Excavata have on average longer amplicons, in particular because of the presence of introns (Torres-Machorro et al., 2010). Amplicon size is then beyond the current range of Illumina sequencing. This may also induce negative bias during PCR amplification (Geisen et al., 2015). For other groups such as Opisthokonta, although the average size is compatible with Illumina sequencing, there is a large number of outlier sequences with long amplicons. This will mean that taxa corresponding to these sequences (mostly Arthropoda) will be missed from surveys conducted with this primer set, although of course this is less critical when protists are targeted. The situation with the V9 primer set no. 27 (Figure 3) is somewhat similar, although there is less length variation between the different supergroups. However, for some groups, in particular Ascomycota and Bangiophyceae, there is a number of outliers that will be missed by Illumina sequencing. Again, these groups are less relevant when focusing on protists. When looking at all the general primer sets (Figure S2), some sets such as nos. 2, 25, and 110 appear to have more taxonomic biases than others. Overall, Excavata constitute the supergroup that is most often discriminated against.

Most primer sets will not amplify archaea and bacteria, except primer sets such as number 33 (515F and Univ 926R Needham & Fuhrman, 2016) that were specifically designed to amplify both bacteria and eukaryotes (Figures S3 and S4). However, some primer sets assumed to be specific to eukaryotes such as no. 4 (563f and 1132r, Hugerth et al., 2014) amplifies quite well archaea and bacteria. Interestingly, set no. 12 (3NDf and 1132rmod, Geisen, Snoek, et al., 2018) amplify only eukaryotes and archaea, but not bacteria. In most cases we tested, the reverse primer was most discriminating against archaea and bacteria.

3.2.2 | Specific primer sets

In order to access a deeper diversity within a given taxonomic group primer sets have been developed with specific targets (Tables S1 and S2). Target levels are most often at the division (e.g., Haptophyta) and class levels (e.g., Chrysophyceae), although some sets are targeting supergroups (e.g., SAR no. 84). Some primer sets have even more specific targets. One example is primer number 65 targeting Cercozoa (S616F Cerco and S947R Cerco, Fiore-Donno et al., 2018) that contains at least five mismatches to all other divisions (Figure S5) and amplifies all cercozoan groups. Primer number 38 targeting Chlorophyta (ChloroF and ChloroR, Moro et al., 2009) contains at least five mismatches to all other divisions (Figure S5). However, it is does not amplify all Chlorophyta as it misses picoplanktonic green algae such as Mamiellophyceae or Chloropicophyceae (Figure S6). In contrast, several primer sets claimed to be specific of a given group

FIGURE 1 Position of the amplified region when using different primer sets listed in the pr2-primers database along the 18S RNA gene relative to the sequence of the yeast *Saccharomyces cerevisiae* (FU970071). The labels correspond to the primer set id, the 18S region amplified, its identification name and the specific group it eventually targets. Bar shading indicates whether the primer is general (black) or specific (grey) of a taxonomic group


general
specific
81 V1−V2 Clerissi 2018 non−Metazoa
80 V1−V2 Creer 2010
110 V2−V3 Rachik 2018
144 V2 Guminska 2021 Euglenids
72 V2−V3 Lentendu 2014c Kinetoplastea
59 V2−V3 Tamura OCSP−A oligotrich, choreotrich
69 V2−V3 Lentendu 2014b Chrysophyceae
62 V3−V4 Lentendu 2014a Cercozoa
88 V5 Michaud 2019b Parabasalia
108 V4 Kataoka 2017
84 V3 Sisson 2018 SAR
40 V4 Zhan
14 V4 Brate2
13 V4 Brate1
12 V4 Geisen
34 V4 Lambert
35 V4 UNonMet non−Metazoa
102 V4 Piwosz 2019 Haptophyta
33 V4 Needham
18 V4 Parfrey
8 V4 Stoeck 2
36 V4 Stoeck 1
16 V4 Piredda
104 V4 Choi 2020
90 V4 Bradley 2016
7 V4 Bass 2016 A
86 V4 Belevich 2017 picoplankton
19 V4 Vannini ciliates
4 V4 Hugerth 2
1 V4 Hadziavdic 1
15 V4 Moreno
103 V4 Emberg 2018
22 V4 Kim 2016
17 V4 Comeau
25 V4 Mangot
99 V4 Xu 2020
2 V4 Hadziavdic 2
39 V4 Egge Haptophyta
98 V4 Fadev 2018
83 V4 Hugerth 6
119 V4 Bass 2020 non−Metazoa
77 V4 Hugerth 5
3 V4 Hugerth 1
101 V4−V5 Hu 2016
100 V4 Kilias 2013
23 V4 Venter
21 V4 Zimmerman diatoms
41 V4 Harder Cercozoa
24 V4 Simon
66 V4 Fiore−Donno 2018d Cercozoa
65 V4 Fiore−Donno 2018c Cercozoa
64 V4 Fiore−Donno 2018b Cercozoa
63 V4 Fiore−Donno 2018a Cercozoa
6 V4 Hugerth 4
5 V4 Hugerth 3
149 V4 Sato 2005 mycorrhizal fungi
124 V5−V7 Johannes 2010
37 V5 Cannon diplonemids
97 V6 Stokes 2002 Labyrinthulomycetes
135 V5 Trzebny 2020 Microsporidia
87 V3−V4 Michaud 2019a Oxymonads
32 V6−V8 Wilkins
68 37F Pawlowski 2010 Foraminifera
107 V7 Huo 2020
76 V7 Lundgreen 2019
67 V7−V9 Bass 2018 Plasmodiophorida
120 V7−V9 Nagai 2016
92 V7−V8 Chemidlin 2011 fungi
106 V8−V9 Kim 2016
89 V8−V9 Bradley 2016
28 V9 Amaral 1
31 V9 Piredda
29 V9 Amaral 2
27 V9 Stoeck
0
500
1000
1500
2000
2500
Position on 18S rRNA

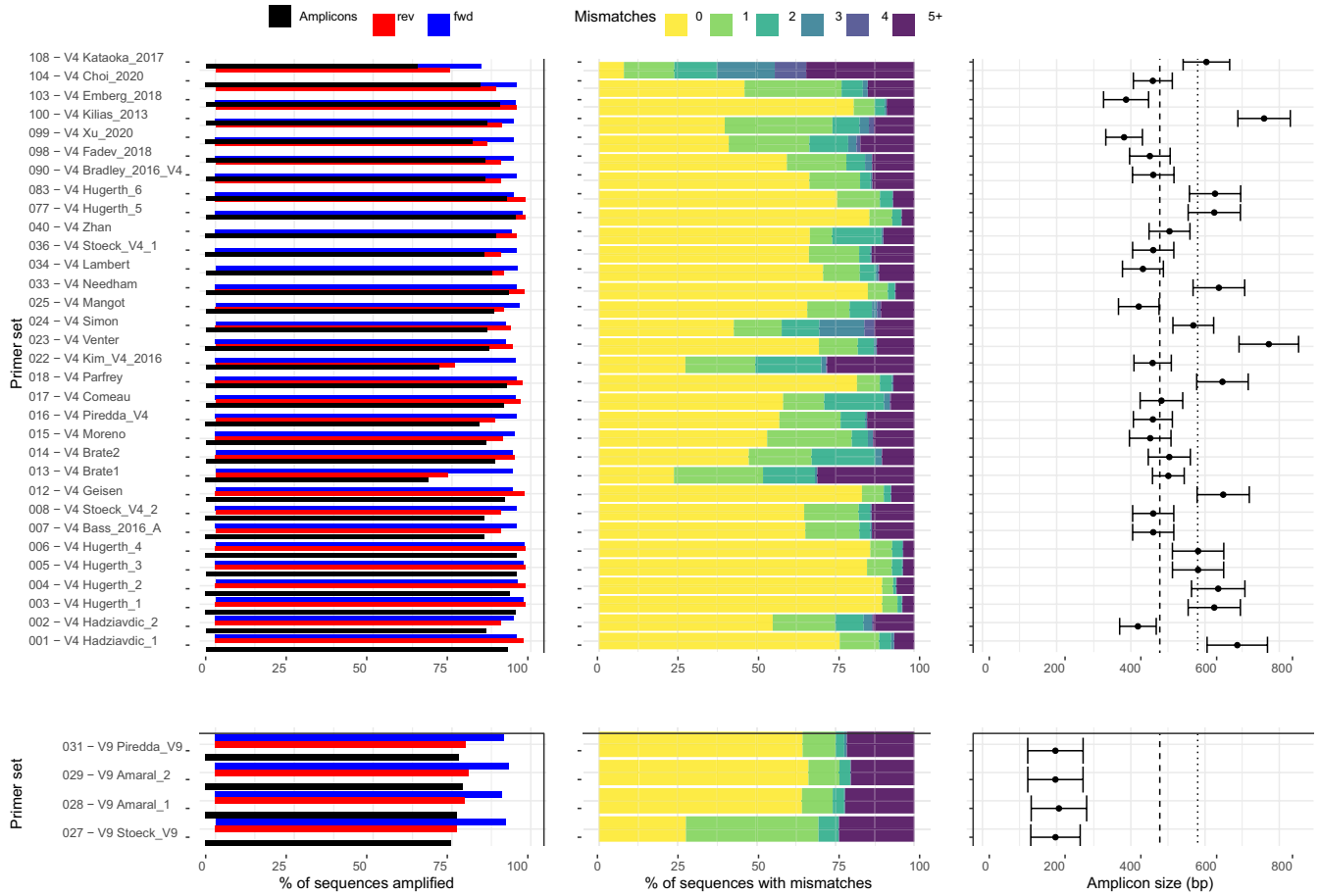


FIGURE 2 Evaluation of general primer sets (Table S2) targeting the V4 (top) and V9 (bottom) regions of the 18S rRNA gene against the PR^2 reference database (version 4.12.0). Left panel. Percentage of reference sequences with at most two mismatches to either forward and reverse primer or to both primers, corresponding to the percentage of sequences amplified by the primer set. Central panel. Number of mismatches for each primer set. Right panel. Amplicon sizes targeted by different primer pairs. The vertical lines correspond to the lengths that can be covered by the most commonly used Illumina sequencers (dashed line: 2 × 250 base pairs [bp]; dotted line: 2 × 300 bp). Error bars represent the standard deviation. See Figure S1 for the complete set of primer sets

are in fact quite general. For example set no. 87 which targets oxymonads (Oxy 18S-F and Oxy 18S-R Michaud et al., 2020) amplifies many other groups (Figures S1 and S5). In this case, this is not critical since oxymonads only occur in termite guts and such primers will only be used in this specific context. Primer set no. 21 (D512for and D978rev, Zimmermann et al., 2011) which was designed to target diatoms would amplify actually most of the Ochrophyta classes but also some green algae (Figure S6).

3.3 | R Shiny application

We have developed a website based on an R SHINY application (https://app.pr2-primers.org) that allows users to visualize and download the pr2-primers database, explore at different taxonomic levels the results of *In silico* amplification against the PR^2 and Silva databases for the primer sets from the pr2-primers database and test their own primer sets. The application is composed of seven panels. The first panel (Figure 4a) provides information on the database as well as a link to report issues or new primers. The second and third panels (Figure 4b) provide an interface to the primer and primer set tables, respectively, with the options of downloading the tables and revealing/hiding specific columns. The fourth and fifth panels are used to display the results of precomputed *In silico* amplification of primer sets from the database. The fourth panel (Figure 4c) shows a synthesis of the results (similar to Figure 2) for all primer sets. The fifth panel (Figure 5a) is a tool to explore amplification properties of a given primer set within a taxonomic level from kingdom to class levels. The right-hand section of this panel shows general amplification characteristics, the location of the mismatches, the number of mismatches for each group and the distribution of the amplicon sizes. Finally, the sixth and seventh panels (Figure 5b) allow users to run an *In silico* amplification with their own primers/probes (panel 6) and primer sets (panel 7) against PR^2 and Silva seed databases. Users can fix the maximum number of mismatches (up to two for each primer). For the sake of speed, only

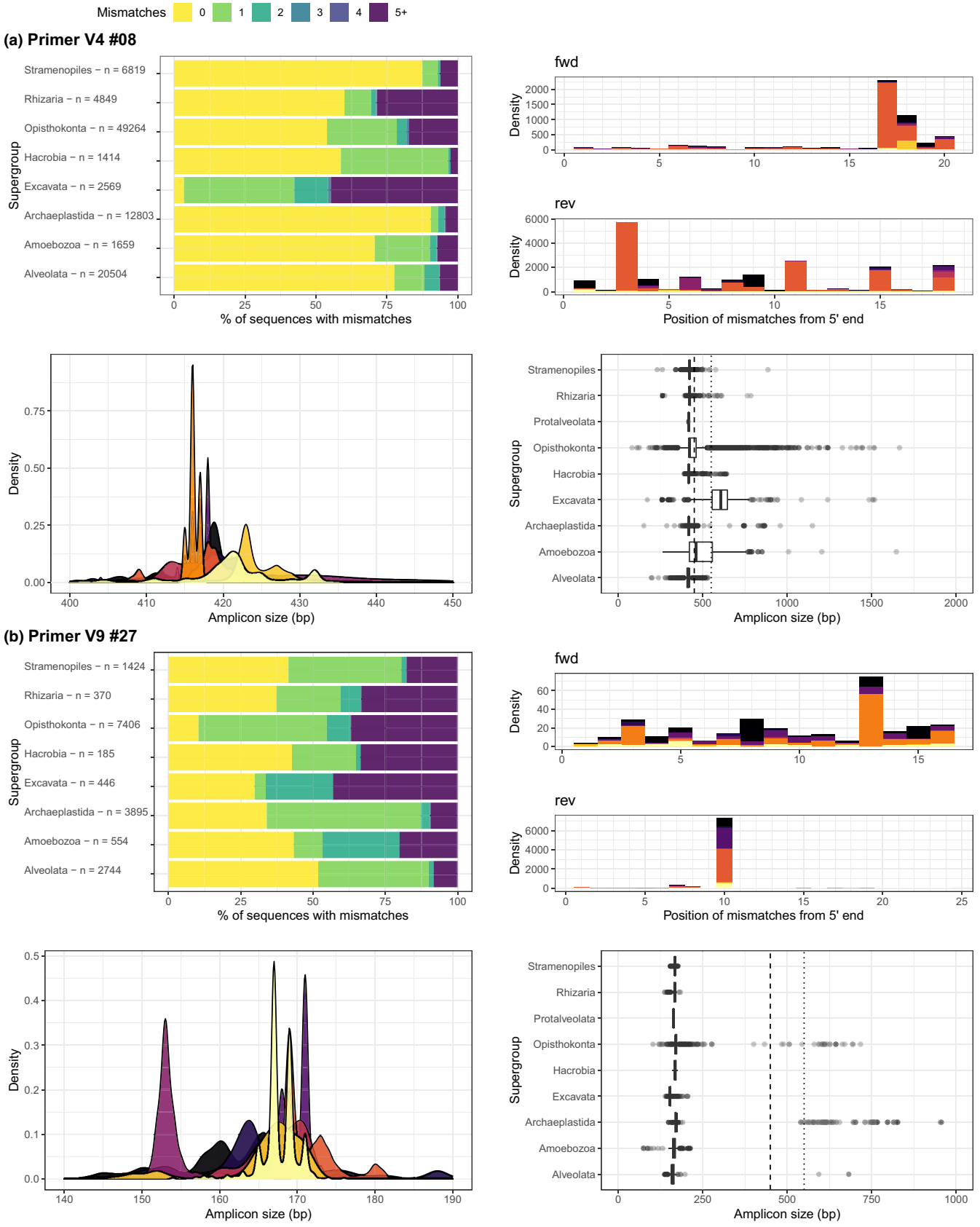


FIGURE 3 Example of analysis for two primer sets amplifying two regions of the 18S rRNA gene: V4 (primer set no. 8, (a) and V9 (primer set no. 27, (b) Top left. Percentage of sequences with a given number of mismatches. Top right. Position of the mismatches for different taxonomic supergroups on the forward and reverse primer, counted from the 5' end. Bottom left. Distribution of amplicon size for different supergroups. Bottom right. Box plots of amplicon size. Colours correspond to taxonomy (division). Hacrobia combine haptophytes, cryptophytes and centrohelids

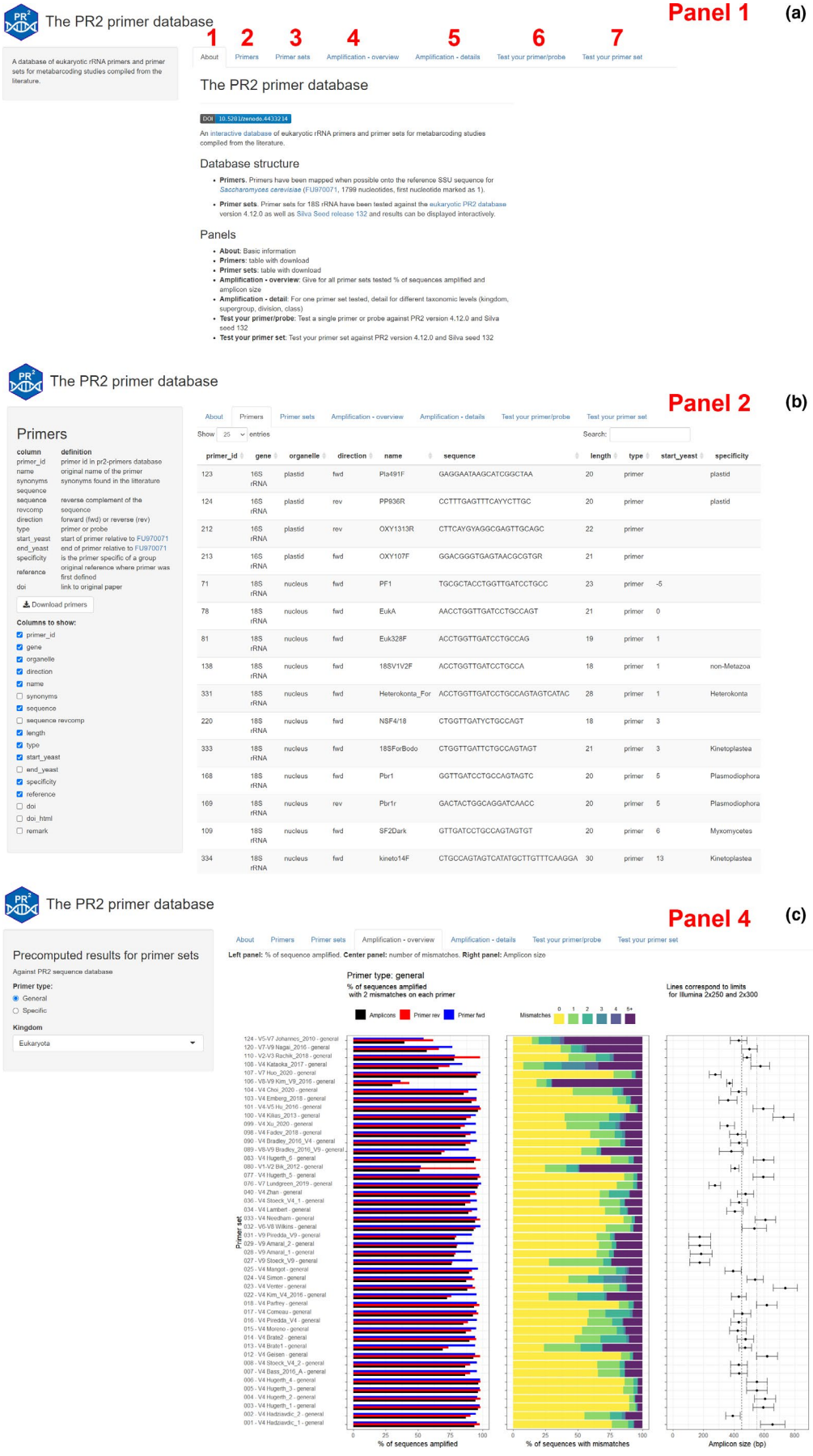


FIGURE 4 Interface to the pr2-primers database. (a) First panel introducing the database. Numbers in red correspond to the different panels. (b) Second panel displaying the list of primers. The third panel is analogous, but for primer sets. (c) Fourth panel showing *In silico* amplification results for all precomputed primer sets

FIGURE 5 Shiny interface to the pr2-primers database. (a) Fifth panel introducing showing *In silico* detailed amplification results for a given precomputed primer set. Taxonomy can be explored in detail. (b) Seventh panel displaying *In silico* amplification results for user-provided primer sets. The sixth panel is identical but for a single primer or probe

the number of mismatches is provided, not their position. Global statistics on the amplification are provided, which can be explored at different taxonomic levels. The R SHINY application has been incorporated into a Docker container available at https://hub.docker.com/repository/docker/vaulot/pr2-primers.

4 | CONCLUSION

The combination of the pr2-primers database with the PR^2 sequence database provides a very useful resource for protist metabarcoding. It will help researchers to select the most suitable primer pairs for both broadly-targeted surveys and studies focusing on target taxonomic groups, and to test and validate *In silico* novel primers. We emphasize that primer pairs must also be tested on reference culture material and natural samples, as actual amplification may differ from *In silico* results. Hopefully this database will grow with time as novel primer pairs are developed and tested on samples from a range of environments. This will contribute to better design and comparability of microbiome analyses, inventories of protist diversity across environments, and increase our understanding of this functionally diverse and important group of organisms.

5 | COMPETING INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGEMENTS

We thank the ABIMS platform of the FR2424 (CNRS, Sorbonne Université) for bioinformatics re-sources. The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Daniel Vaulot and Stefan Gelsen conceived the study. Daniel Vaulot, David Bass and Frédéric Mahé scanned the literature for existing primers and primer sets. Daniel Vaulot developed the database, the analysis scripts and the R shiny application. Daniel Vaulot wrote the first draft of the manuscript and all coauthors edited and approved the final version.

DATA AVAILABILITY STATEMENT

No new data were created or analysed in this study. All scripts, including those for the Shiny application, have been made available at https://github.com/pr2database/pr2-primers (https://doi.org/10.5281/zenodo.4849528). The database is available at https://app.pr2-primers.org.

ORCID

Daniel Vaulot https://orcid.org/0000-0002-0717-5685
Stefan Geisen https://orcid.org/0000-0003-0734-727X
Frédéric Mahé https://orcid.org/0000-0002-2808-0984
David Bass https://orcid.org/0000-0002-9883-7823

REFERENCES

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One*, *4*, e6372. https://doi.org/10.1371/journal.pone.0006372

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the Metazoa. *Molecular Ecology*, *27*, 3968–3975. https://doi.org/10.1111/mec.14844

Bass, D., & del Campo, J. (2020). Microeukaryotes in animal and plant microbiomes: Ecologies of disease? *European Journal of Protistology*, *76*, 125719. https://doi.org/10.1016/j.ejop.2020.125719

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16s rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *6*, 190007. https://doi.org/10.1038/sdata.2019.7

Carnegie, R. B., Meyer, G. R., Blackbourn, J., Cochennec-Laureau, N., Berthe, F. C., & Bower, S. M. (2003). Molecular detection of the oyster parasite *Mikrocytos mackini*, and a preliminary phylogenetic analysis. *Diseases of Aquatic Organisms*, *54*, 219–227. https://doi.org/10.3354/dao054219

Clerissi, C., Brunet, S., Vidal-Dupiol, J., Adjeroud, M., Lepage, P., Guillou, L., Escoubas, J. M., & Toulza, E. (2018). Protists within corals: The hidden diversity. *Frontiers in Microbiology*, *9*, 2043. https://doi.org/10.3389/fmicb.2018.02043

Comeau, A. M., Li, W. K., Tremblay, J. É., Carmack, E. C., & Lovejoy, C. (2011). Arctic ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS One*, *6*, e27492. https://doi.org/10.1371/journal.pone.0027492

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*, 20140562. https://doi.org/10.1098/rsbl.2014.0562

Egge, E., Bittner, L., Andersen, T., Audic, S., de Vargas, C., & Edvardsen, B. (2013). 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: A test for marine haptophytes. *PLoS One*, *8*, e74371. https://doi.org/10.1371/journal.pone.0074371

Elbrecht, V., & Leese, F. (2017). PrimerMiner: An r package for development and *in silico* validation of DNA metabarcoding primers. *Methods in Ecology and Evolution*, *8*, 622–626. https://doi.org/10.1111/2041-210X.12687

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An *In silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 434. https://doi.org/10.1186/1471-2164-11-434

Fiore-Donno, A. M., Rixen, C., Rippin, M., Glaser, K., Samolov, E., Karsten, U., Becker, B., & Bonkowski, M. (2018). New barcoded primers for efficient retrieval of cercozoan sequences in high-throughput environmental diversity surveys, with emphasis on worldwide biological soil crusts. *Molecular Ecology Resources*, *18*, 229–239. https://doi.org/10.1111/1755-0998.12729

Geisen, S., Laros, I., Vizcaíno, A., Bonkowski, M., & De Groot, G. A. (2015). Not all are free-living: High-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa. *Molecular Ecology*, *24*, 4556–4569. https://doi.org/10.1111/mec.13238

Geisen, S., Mitchell, E. A., Adl, S., Bonkowski, M., Dunthorn, M., Ekelund, F., Fernández, L. D., Jousset, A., Krashevska, V., Singer, D., Spiegel, F. W., Walochnik, J., & Lara, E. (2018). Soil protists: A fertile frontier in soil biology research. *FEMS Microbiology Reviews*, *42*, 293–323. https://doi.org/10.1093/femsre/fuy006

Geisen, S., Snoek, L. B., ten Hooven, F. C., Duyts, H., Kostenko, O., Bloem, J., Martens, H., Quist, C. W., Helder, J. A., & van der Putten, W. H.

(2018). Integrating quantitative morphological and qualitative molecular methods to analyse soil nematode community responses to plant range expansion. *Methods in Ecology and Evolution*, *9*, 1366–1378. https://doi.org/10.1111/2041-210X.12999

Greuter, D., Loy, A., Horn, M., & Rattei, T. (2016). ProbeBase-an online resource for rRNA-targeted oligonucleotide probes and primers: New features 2016. *Nucleic Acids Research*, *44*, D586–D589. https://doi.org/10.1093/nar/gkv1232

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (2013). The Protist Ribosomal Reference database (PR^2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*, D597–D604. https://doi.org/10.1093/nar/gks1160

Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., Wilmes, P., & Andersson, A. F. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One*, *9*, e95567. https://doi.org/10.1371/journal.pone.0095567

Kataoka, T., Yamaguchi, H., Sato, M., Watanabe, T., Taniuchi, Y., Kuwata, A., & Kawachi, M. (2017). Seasonal and geographical distribution of near-surface small photosynthetic eukaryotes in the western North Pacific determined by pyrosequencing of 18S rDNA. *FEMS Microbiology Ecology*, *93*, fiw229. https://doi.org/10.1093/femsec/fiw229

Lambert, S., Tragin, M., Lozano, J. C., Ghiglione, J. F., Vaulot, D., Bouget, F. Y., & Galand, P. E. (2019). Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME Journal*, *13*, 388–401. https://doi.org/10.1038/s41396-018-0281-z

Lie, A. A., Liu, Z., Hu, S. K., Jones, A. C., Kim, D. Y., Countway, P. D., Amaral-Zettler, L. A., Cary, S. C., Sherr, E. B., Sherr, B. F., Gast, R. J., & Caron, D. A. (2014). Investigating microbial eukaryotic diversity from a global census: Insights from a comparison of pyrotag and full-length sequences of 18S rRNA genes. *Applied and Environmental Microbiology*, *80*, 4363–4373. https://doi.org/10.1128/AEM.00057-14

Lopes Dos Santos, A., Ribeiro Gérikas, C., Ong, D., Garczarek, L., Shi, X. L., Nodder, S., Vaulot, D., & Gutierrez-Rodriguez, A. (2022). 3.5. Phytoplankton diversity and ecology through the lens of high throughput sequencing technologies. In L. Clementson, R. S. Eriksen, & A. Willis (Eds.), *Advances in Phytoplankton Ecology. Applications of Emerging Technologies* (pp. 1–53). Elsevier.

Ludwig, W. (2004). ARB: A software environment for sequence data. *Nucleic Acids Research*, *32*, 1363–1371. https://doi.org/10.1093/nar/gkh293

Lundgreen, R. B., Jaspers, C., Traving, S. J., Ayala, D. J., Lombard, F., Grossart, H. P., Nielsen, T. G., Munk, P., & Riemann, L. (2019). Eukaryotic and cyanobacterial communities associated with marine snow particles in the oligotrophic Sargasso Sea. *Scientific Reports*, *9*, 8891. https://doi.org/10.1038/s41598-019-45146-7

Medlin, L., Elwood, H. J., Stickel, S., & Sogin, M. L. (1988). The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene*, *71*(2), 491–499. https://doi.org/10.1016/0378-1119(88)90066-2

Michaud, C., Hervé, V., Dupont, S., Dubreuil, G., Bézier, A. M., Meunier, J., Brune, A., & Dedeine, F. (2020). Efficient but occasionally imperfect vertical transmission of gut mutualistic protists in a wood-feeding termite. *Molecular Ecology*, *29*, 308–324. https://doi.org/10.1111/mec.15322

Morard, R., Quillévéré, F., Douady, C. J., de Vargas, C., de Garidel-Thoron, T., & Escarguel, G. (2011). Worldwide genotyping in the planktonic foraminifer *Globoconella inflata*: Implications for life history and paleoceanography. *PLoS One*, *6*, e26665. https://doi.org/10.1371/journal.pone.0026665

Moro, C. V., Crouzet, O., Rasconi, S., Thouvenot, A., Coffe, G., Batisson, I., & Bohatier, J. (2009). New design strategy for development of specific primer sets for PCR-based detection of Chlorophyceae and Bacillariophyceae in environmental samples. *Applied and Environmental Microbiology*, *75*, 5729–5733. https://doi.org/10.1128/AEM.00509-09

Needham, D. M., & Fuhrman, J. A. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology*, *1*, 16005. https://doi.org/10.1038/nmicrobiol.2016.5

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2020). *Biostrings: Efficient manipulation of biological strings*. https://bioconductor.org/packages/release/bioc/html/Biostrings.html

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*, 1403–1414. https://doi.org/10.1111/1462-2920.13023

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., … de Vargas, C. (2012). CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, *10*, e1001419. https://doi.org/10.1371/journal.pbio.1001419

Piredda, R., Tomasino, M. P., D'Erchia, A. M., Manzari, C., Pesole, G., Montresor, M., Kooistra, W. H., Sarno, D., & Zingone, A. (2017). Diversity and temporal patterns of planktonic pro-tist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiology Ecology*, *93*, fiw200. https://doi.org/10.1093/femsec/fiw200

Pujari, L., Wu, C., Kan, J., Li, N., Wang, X., Zhang, G., Shang, X., Wang, M., Zhou, C., & Sun, J. (2019). Diversity and spatial distribution of chromophytic phytoplankton in the Bay of Bengal revealed by RuBisCO Genes (*rbcL*). *Frontiers in Microbiology*, *10*, 1–17. https://doi.org/10.3389/fmicb.2019.01501

R Development Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, *1*, 409. https://doi.org/10.1007/978-3-540-74686-7

Sali, A., & Attali, D. (2020). *Shinycssloaders: Add loading animations to a 'shiny' output while it's recalculating*. https://CRAN.R-project.org/package=shinycssloaders

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 12115–12120. https://doi.org/10.1073/pnas.0605127103

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, *19*, 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M. J., Chistoserdov, A., Orsi, W., & Edgcomb, V. P. (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology*, *7*, 72. https://doi.org/10.1186/1741-7007-7-72

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Torres-Machorro, A. L., Hernández, R., Cevallos, A. M., & López-Villaseñor, I. (2010). Ribosomal RNA genes in eukaryotic microorganisms: Witnesses of phylogeny? *FEMS Microbiology Reviews*, *34*, 59–86. https://doi.org/10.1111/j.1574-6976.2009.00196.x

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*, 110–117. https://doi.org/10.1016/j.tree.2008.09.011

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer International Publishing.

Worden, A. Z., Follows, M. J., Giovannoni, S. J., Wilken, S., Zimmerman, A. E., & Keeling, P. J. (2015). Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, *347*, 1257594. https://doi.org/10.1126/science.1257594

Yahalomi, D., Atkinson, S. D., Neuhof, M., Sally Chang, E., Philippe, H., Cartwright, P., Bartholomew, J. L., & Huchon, D. (2020). A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 5358–5363. https://doi.org/10.1073/pnas.1909907117

Zimmermann, J., Jahn, R., & Gemeinholzer, B. (2011). Barcoding diatoms: Evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity and Evolution*, *11*, 173–192. https://doi.org/10.1007/s13127-011-0050-6

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Vaulot, D., Geisen, S., Mahé, F., & Bass, D. (2022). pr2-primers: An 18S rRNA primer database for protists. *Molecular Ecology Resources*, 22, 168–179. https://doi.org/10.1111/1755-0998.13465