A user guide to environmental protistology: primers, metabarcoding, sequencing, and analyses

Stefan Geisen^{1,2,3*,#}, Daniel Vaultot^{4,5,#}, Frédéric Mahé^{6,7}, Enrique Lara⁸, Colombar de Vargas^{4,9}, David Bass^{10,11}

¹Department of Terrestrial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), PO Box 50, 6700 AB Wageningen, The Netherlands

²Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands

³Nanjing Agricultural University, Nanjing, 210095, P.R. China

⁴Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680 Roscoff, France

⁵Asian School of the Environment, Nanyang Technological University, Singapore

⁶CIRAD, UMR BGPI, F-34398, Montpellier, France

⁷BGPI, Univ Montpellier, CIRAD, IRD, Montpellier SupAgro, Montpellier, France

⁸Real Jardín Botánico, Consejo Superior de Investigaciones Científicas CSIC, Plaza de Murillo 2, 28014 Madrid, Spain

⁹Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* GOSEE, 3 rue Michel-Ange, 75016 Paris, France

¹⁰Centre for Environment, Fisheries and Aquaculture Science, Barrack Road, The Nothe, Weymouth, Dorset DT4 8UB, UK

¹¹Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

*corresponding author: phone +31 317 473 580; email s.geisen@nioo.knaw.nl

authors contributed equally

Abstract

Protists – all eukaryotes besides fungi, animals, and plants - represent a major part of the taxonomic and functional diversity of eukaryotic life on the planet and drive many ecosystem processes. However, knowledge of protist communities and their diversity lags behind that of most other groups of organisms, largely due to methodological constraints. While protist communities differ markedly between habitats and biomes, they can be studied in very similar ways. Here we provide a guide to current molecular approaches used for studying protist diversity, with a particular focus on amplicon-

based high-throughput sequencing (metabarcoding). We highlight that the choice of suitable primers artificially alters community profiles observed in metabarcoding studies. While there are no true ‘universal’ primers to target all protist taxa as a whole, we identify some primer combinations with a wide taxonomic coverage and provide detailed information on their properties. Although environmental protistan ecological research will probably shift towards PCR-free metagenomics or/and transcriptomic approaches in a near future, metabarcoding will remain the method of choice for in-depth community analyses and taxon inventories in biodiversity surveys and ecological studies, due its great cost-efficiency, sensitivity, and throughput. In this paper we provide a guide for scientists from a broad range of disciplines to implement protists in their ecological analyses.

Keywords: Microbial ecology, protists, high-throughput sequencing, environmental microbiology, primer bias, 18S rRNA gene, metabarcoding

Introduction

The past three decades have seen massive developments in our understanding of microbial diversity in most of the Earth ecosystems. Molecular approaches, particularly the advent of high-throughput sequencing (HTS), have enabled cultivation-independent diversity analyses of microorganisms. PCR-based amplicon HTS (here defined as metabarcoding) data for bacteria and archaea have rapidly accumulated throughout this period, revealing a vast and mostly unknown taxonomic diversity (Gans, Wolinsky, & Dunbar, 2005; Leininger et al., 2006; Louca, Mazel, Doebeli, & Parfrey, 2019; Parks et al., 2017; Sogin et al., 2006). While sequencing errors were shown to have artefactually inflated the inferred diversity in many early studies (Huse, Huber, Morrison, Sogin, & Welch, 2007), increased quality control and refined techniques have subsequently confirmed that the global diversity of prokaryotes likely exceeds a million bacterial species-level lineages (Louca et al., 2019; Parks et al., 2017; Thompson et al., 2017), with many previously unknown clades emerging. Similarly, fungal diversity has been studied extensively using HTS and taxon diversity has been estimated to exceed 100,000 (Buée et al., 2009; Tedersoo et al., 2014), if not an order of magnitude greater (Bass &

Richards, 2011; Hawksworth & Lucking, 2017). Protist metabarcoding is lagging behind that of prokaryotes and fungi (Geisen et al., 2017), yet suggesting global species-level diversity estimates in the millions (de Vargas et al., 2015; Mahé et al., 2017).

Protists are of key importance for ecosystem functioning and are major drivers in diverse nutrient cycling pathways. Autotrophs (algae) are the counterparts of land plants as main carbon fixers in aquatic environments (Worden et al., 2015). Heterotrophic protists catalyse nutrient cycling in aquatic and terrestrial environments as selective consumers of bacteria and fungi, being critical drivers of the so-called ‘microbial loop’ (Azam et al., 1983; Bonkowski, 2004). Parasitic protists influence community dynamics of larger eukaryotic hosts, including plants and animals (S. Geisen et al., 2018; Mahé et al., 2017; Worden et al., 2015). Through these abiotic and biotic processes, protists make nutrients available to either smaller or larger members of food webs and enhance eco-system dynamics (Bonkowski, 2004) and/or connectivity (Lima-Mendez et al., 2015). Nevertheless, the diversity of protists is often excluded from microbiome and food-web studies (Lundberg et al., 2012; Mendes et al., 2011). Therefore, we are missing information of an important biotic component that regulates ecosystem functioning.

The protist biodiversity knowledge gap at least partly originates from methodological constraints. Standardized PCR-based approaches with a relatively limited number of recommended primer pairs are applied to study bacteria (J. Gregory Caporaso et al., 2010; J. G. Caporaso et al., 2011; Cole et al., 2014; Knight et al., 2018; Walters et al., 2016) and fungi (Ihrmark et al., 2012; Lindahl Björn et al., 2013; Nilsson et al., 2019). However, a consensus on comparable molecular approaches using a standardized PCR-based approach to study protists does not exist. Several primer sets are regularly used (Adl et al., 2019; Adl, Habura, & Eglit, 2014; Hadziavdic et al., 2014), although their benefits and shortcomings are generally poorly known. Ideally, a PCR protocol would amplify all members of the target group with equal likelihood, avoid excessive amplification of non-target lineages, and provide amplicons with high taxonomic resolution. The target sequence should also be well represented in public databases, and

provide enough phylogenetic signal for eco-evolutionary analyses. However, a single such ideal PCR system cannot be developed. In reality, each primer set has some phylogenetic bias, and the use of many different sets therefore limits comparability across studies (Fouhy, Clooney, Stanton, Claesson, & Cotter, 2016; Ramirez et al., 2018; Stoeck et al., 2010). In fact, primer pairs have often been designed to answer specific questions in protist community profiling (see below), based on the experience or preference of individual researchers. A comparative synthesis of existing ‘protist primers’ for non-experts is still lacking. We therefore review some of the most frequently used primer pairs, evaluate their pros and cons, and provide suggestions for optimal primer choice in different environmental settings. We then introduce guidelines to consider in HTS-based analyses of protist biodiversity. Lastly, we provide an overview of future opportunities and challenges introduced by PCR free HTS approaches.

18S rRNA gene: the universal marker to assess protist biodiversity

The 16S (small subunit) ribosomal RNA gene (Woese, 1987) has been used as a genetic marker for prokaryotic (bacteria and archaea) diversity since the earliest days of amplicon-based diversity studies (Giovannoni & Cary, 1993; Pace, 1997). Its eukaryotic counterpart (18S) was later adopted for studies of micro-eukaryotes (Lopez-Garcia, Rodriguez-Valera, Pedros-Alio, & Moreira, 2001; Moon-van der Staay, De Wachter, & Vaulot, 2001; Moreira & López-García, 2002; van Hannen et al., 1999). The 16S/18S rRNA gene encodes the RNA molecule that forms part of the small subunit of the ribosome. The ribosome is a macro-molecule essential to all organisms to translate mRNA into proteins, its function thus being conserved across the tree of life. While other markers are sometimes preferable for finer-resolution, often group-specific, biodiversity assessment (Fig. 2 in Jan Pawlowski et al. (2012)), the 18S rRNA is the best suited gene for exploring the diversity of protists as a whole. The 18S rRNA gene contains nine relatively variable regions flanked by more conserved regions that together span a range of evolutionary rates, enabling phylogenetic comparisons of both distant and closely related taxa (Neefs, Van de Peer, De Rijk, Chapelle, & De Wachter, 1993). Primers can be designed in more or less

conserved regions to target a wide diversity of taxa. However, there are important considerations to be made when choosing or designing primer sets to amplify 18S rRNA gene regions. Despite being most comprehensive, NCBI's GenBank, is not recommendable to classify 18S rRNA gene sequences due to large amount of errors that include wrongly annotated sequences. At least two general databases are dedicated to the 18S rRNA gene, Silva (Pruesse et al., 2007) and PR² (Guillou et al., 2013), while more specialized databases have been designed for specific taxonomic groups, e.g. foraminifera (Morard et al., 2015) or dinoflagellates (Mordret et al., 2018).

A comparative overview of 18S rRNA gene primers with large taxonomic spectrum

HTS protist metabarcoding studies have employed dozens of different primers in various combinations, targeting a range of variable 18S rRNA gene regions. Among the most widely targeted regions in early protist-focused HTS metabarcoding studies was the V1-3 region (Euringer & Lueders, 2008), a region with relatively high taxonomic resolution (Hadziavdic et al., 2014; Tanabe Akifumi et al., 2015). However, this region is decreasingly being used to target protists due to insufficient reference sequence coverage of the 5' end of the 18S rRNA gene, its excessive length for most current HTS methods, and unsuitable regions for primer annealing (Hugerth et al., 2014). Instead, the V4 and V9 regions are most commonly used nowadays; they are the focus of this review, as summarised in Table 1.

The V9 region was the most frequently used metabarcoding target in early high throughput studies as its relatively short amplicon length (< 200 bp) was better suited to the sequence length limitations of early HTS technologies. This short length makes the V9 region a suitable target also for contemporary ultra-deep metabarcoding technologies, making it a most cost-effective approach to study protist community diversity, while allowing stringent quality control. The downsides of the V9 region is

associated with its short length (Fig. 1), providing limited phylogenetic information. As such, the 400-500 bp V4 region fragment is increasingly used (Table 1). Furthermore, the V4 region is closest of all HTS-suitable 18S rRNA gene regions to that of the full-length gene and allows the high phylogenetic/taxonomic resolution often to species- or genus-level (Dunthorn, Klier, Bunge, & Stoeck, 2012; Hu Sarah et al., 2015).

The V4 region also enables the most accurate taxonomic placement of unassigned HTS amplicon sequences (Mahé et al., 2017). For these reasons it has been suggested as the main universal protist barcode region (Jan Pawlowski et al., 2012). However, some disadvantages are associated with the V4 region such as its length, which is too long for some short-read optimized sequencing methodologies such as Illumina HiSeq and NovaSeq (Fig. 1, Fig. 2). V4-based metabarcoding surveys also suffer from PCR/sequencing biases due to amplicon length variability (Fig. 1) and variable secondary structures.

Challenges of 18S rRNA gene targeted diversity analyses

The 18S rRNA gene is more complex than the prokaryotic 16S rRNA gene. It is on average three hundred base pairs longer (appr. 1,800 bp) and displays profound evolutionary rate variation between different taxa. This renders the design of truly universal primers impossible (Adl et al., 2014; Hadziavdic et al., 2014). Primers described as ‘universal’ are better thought of as ‘broadly-targeted’. Furthermore, even the most broadly-targeted primer pair will not amplify all protist lineages equally well, such as higher taxonomic ranks (e.g. eukaryotic supergroup; Fig. 3), but differential amplification exists also at lower taxonomic ranks such as at the class level (Fig. 4). Some highly divergent lineages will not amplify at all, as their 18S rRNA gene sequences can differ in multiple nucleotide positions even in the most conserved primer sites. This is the case for many parasitic protists (Bass, Stentiford, Littlewood, & Hartikainen, 2015; Hartikainen et al., 2014), but also diverse free-living lineages (Bass et al., 2016; Bass,

Yabuki, Santini, Romac, & Berney, 2012; Fiore-Donno, Weinert, Wubet, & Bonkowski, 2016); at high taxonomic rank this is most obvious for Discoba (within Excavata; Fig. 2). For those groups, alternative approaches are needed to study their communities, such as applying group-specific primers. Even when primers are a perfect sequence match to the template, differences in length and secondary or tertiary structures of the amplified region, due to insertion-deletions (indels) or the presence of introns, which leads to biases, as shorter and structurally simpler fragments are preferentially amplified over longer ones (Figs. 2 and 3). This can influence the sequence representation of protists at all levels, even at supergroups. For instance, Stramenopiles and Alveolates have relatively short V4 amplicon lengths leading to preferential PCR-amplification, while the longer V4 regions of Discoba (including Excavata) and Amoebozoa (Figs. 2 and 3) amplify less efficiently (Geisen, Laros, Vizcaíno, Bonkowski, & de Groot, 2015).

The ribosomal gene operon is present in multiple copies, and copy numbers vary greatly between taxonomic groups by up to three orders of magnitude (Gong, Dong, Liu, & Massana, 2013; Zhu, Massana, Not, Marie, & Vaulot, 2005). There is evidence in some protist groups that the number of 18S rRNA gene reads (suggesting gene copies) positively correlates with relative cell abundance (Giner et al., 2016), biomass (Pitsch et al., 2019) or biovolume (de Vargas et al., 2015), but these correlations do not hold across divergent protist groups (Pitsch et al., 2019). Variable copy numbers between lineages can profoundly alter interpretation of relative abundance or biomass of taxa within protist communities in metabarcoding studies. Inferred lineage diversity can also be biased by intragenomic variation between rRNA gene copies. The degree of intragenomic variation of rRNA gene copies is generally low (0-1%), although in some groups it can reach several percent in some ciliates (Gong et al., 2013), Myxomycetes (García-Martín, Zamora, & Lado, 2019), and Foraminifera (Weber & Pawlowski, 2014). Both of these factors can vary greatly between closely related taxa (Andre et al., 2014; García-Martín et al., 2019; Gong et al., 2013; Nassonova, Smirnov, Fahrni, & Pawlowski, 2010; Weber & Pawlowski, 2014). This can artificially increase diversity estimates as single taxa can then be

treated as different species or, if intraspecific sequence diversity is high, as different taxonomic groups (Caron & Hu, 2019).

The phylogenetic resolution a barcoding region provides is key for HTS studies. However, because taxonomic marker gene evolution occurs at different rates relative to phenotypic evolution across protist groups, there is no fixed or consensus relationship between marker genes dissimilarity and taxonomy (Boenigk, Ereshefsky, Hoef-Emden, Mallet, & Bass, 2012). Therefore, fixed percentage differences between lineages cannot be used to infer taxonomic distances between groups, and often within them. Different solutions to this can be approached for individual groups, such as targeting a combination of diverse variable regions or using barcoding regions other than the 18S rRNA gene. This, however, is often not feasible in larger-scaled ecological studies due to cost-limitations, and a lower resolution has to be accepted. Another approach is the use of selective targeted bioinformatics analysis for specific groups of interest, for example mapping short reads onto robust phylogenetic trees constructed with longer sequences.

Lastly, a metabarcoding approach targeting protists will depict a community of protists that differs from the protist community present *in situ* or *in vivo*, because of the above-mentioned taxon-specific PCR biases and copy number differences. As such, the primer pairs chosen alter the inferred composition of protist communities in a primer-pair specific manner, which is illustrated in studies applying primers targeting both the V4 and V9 regions (Pagenkopp Lohan, Fleischer, Carney, Holzer, & Ruiz, 2016; Stoeck et al., 2010; Tragin, Zingone, & Vaultot, 2018). For example, V9 datasets generally contain relatively more sequences that can only be assigned at the domain level (eukaryotes) compared to V4 assignments (Pagenkopp Lohan et al., 2016).

Group-specific primer sets

Lineages that do not amplify well with broadly-targeted primers or when non-target sequences are expected to dominate the results can be targeted in metabarcoding approaches with group-specific primers. These target a range of different 18S rRNA gene regions (Supplementary Table 1). For

instance, Cercozoa have specifically been targeted in soils to avoid amplification of fungi (Fiore-Donno Anna et al., 2017; Harder et al., 2016; Lentendu et al., 2014). However even within Cercozoa, amplification of some clades requires even more specific primers, such as the divergent coprophilic *Helkesimastix-Guttulinopsis-Rosculus* clade (Bass et al., 2016). Supplementary Table 1 provides a summary of group-specific primer sets that have so far been used in 18S V4 region rRNA gene protistan metabarcoding efforts. Note that this table only includes primers that have been used and proven successful in HTS metabarcoding efforts. Many other primer sets have been proposed (Adl et al., 2019; Adl et al., 2014) that might be applicable but have not yet been tested in metabarcoding approaches.

Genes other than the 18S rRNA can be targeted to increase the taxonomic resolution of specific protist lineages (Jan Pawlowski et al., 2012). For instance, ITS sequences have been used to assess the distribution of uncultivated heterotrophic marine protists (Rodríguez-Martínez, Rocap, Logares, Romac, & Massana, 2011), often providing species-level resolution (Stern et al., 2012). The first subunit of the mitochondrial cytochrome oxidase as the universal barcoding region in animals (Hebert, Ratnasingham, & deWaard, 2003) also allows high-taxonomic discrimination of many protist taxa (Heger et al., 2011; Singer et al., 2018). The binding sites of these primers are so conserved that the same protocols as for animals have been used for several divergent Amoebozoa (Kosakyan et al., 2012; Nassonova et al., 2010). For specific groups of photosynthetic protists such as diatoms the *rbcl* gene can be used (Rimet et al., 2019).

An alternative to group-specific primers that target specific groups of protists is the use of primer combinations that amplify broadly but exclude certain groups of protists and other eukaryotes. As such, fungal sequences can be avoided in soils (Fiore-Donno et al., 2016; Geisen, 2016; Lentendu et al., 2014), and in holobionts for the exclusion of plant or metazoan host-associated sequences.

Assessing host-associated protist diversity

The highly parallel throughput of currently available sequencing techniques means that protist diversity can be effectively sampled and analysed even in samples dominated by non-protist taxa.

However, when targeting host-associated protists, co-amplification of host DNA can overwhelm the protist signal. In these cases, group-specific primer sets can be used to focus on a particular group. This strategy is typically used for high-resolution diversity analyses of fungi or Peronosporomycetes (formerly Oomycetes) inside plant roots (Ramirez et al., 2019; Sapkota & Nicolaisen, 2015). However, the disadvantage of this approach is that multiple groups that might be of ecological importance as symbionts cannot be simultaneously targeted. To fully assess the eukaryotic microbiome ('eukaryome' (Javier del Campo, Bass, & Keeling, 2019)) in animal hosts, broadly-targeted eukaryotic 18S rRNA gene V4 region primers that avoid amplification of most animal sequences are available (Bower et al., 2004; Javier del Campo et al., 2019). Alternatively, blocking primers can also be applied that reduce amplification of selected, non-target DNA and thereby increase the amount of target DNA (Tan & Liu, 2018; Vestheim & Jarman, 2008). To the best of our knowledge, plant-blocking primers have not yet been developed to enable amplification of protists within or on plant tissue.

Downstream sequence analyses

Downstream analyses of raw sequence reads are needed process the data and to ensure high data quality. The key steps are common to analyses of all metabarcoding datasets: merging paired-end sequence reads, and quality control steps such as removing low quality, short-read and chimeric sequences (J. Gregory Caporaso et al., 2010). For the subsequent clustering step of processed reads into sequence types representing taxonomic units, an increasingly wide choice of methods is available that significantly impacts measures of diversity and subsequent interpretation. This step aims to eliminate erroneous sequences created during PCR and sequencing, along with sequences originating from intragenomic and intraspecific polymorphisms (Cédric Berney, Fahrni, & Pawlowski, 2004; Richards & Bass, 2005) that can artificially inflate diversity estimates.

Operational Taxonomic Unit (OTU) clustering commonly groups sequences by a user-defined percentage similarity – typically 97 to 99% - using software such as mothur, qiime, usearch, vsearch, etc. (Edgar, 2010; Rognes, Flouri, Nichols, Quince, & Mahé, 2016), using more flexible high taxonomic

resolution approaches such as the clustering method SWARM (Mahe, Rognes, Quince, de Vargas, & Dunthorn, 2014, 2015) or the denoising method implemented in dada2 to produce amplicon sequence variants (ASVs). This last approach is now widely used for bacterial communities across the globe (Callahan, McMurdie, & Holmes, 2017; Delgado-Baquerizo et al., 2018; Thompson et al., 2017), and is increasingly used to analyse protist diversity (Chénard et al., 2019). It is possible that the application of ASVs might lead to an over-estimate of protistan diversity as intragenomic gene copy variants or and artefactual sequences may be interpreted as separate taxonomic units (Caron & Hu, 2019; Xiong et al., 2019).

Several methods exist for taxonomic annotation of OTUs/swarms/ASVs: assignment by the Uclust consensus taxonomy assigner, blast, GGSearch (Pearson, 2014), SortMeRNA, RTAX, Kraken or naïve Bayes classification (RDP Classifier (Wang, Garrity, Tiedje, & Cole, 2007)). Two main curated reference databases are available for annotation of 18S rRNA gene reads: SILVA (Pruesse et al., 2007) and PR² (Guillou et al., 2013) (<https://github.com/pr2database/pr2database>). SILVA incorporates all existing 18S rRNA sequences and relies on automatic annotation, while PR² focuses on the coherence of a taxonomy framework, for example using resources such as Algaebase (Guiry & Guiry, 2008) for photosynthetic protists, and on taxonomic re-annotations performed by supervised pipelines such as EukRef (Boscaro et al., 2018; Javier del Campo et al., 2018). Taxonomic annotation of the same dataset to different databases can result in discrepancies (Dupont, Griffiths, Bell, & Bass, 2016). A community based and expert driven, universal taxonomic framework for protist is underway (UniEuk (C. Berney et al., 2017)) and aims at providing a unified reference taxonomy for protists that will be implemented through the European Bioinformatics Institute (EBI).

Beyond currently applied metabarcoding approaches

Some new sequencing technologies can generate far longer reads than previously possible (Clarke et al., 2009). Amplicons of thousands of base pairs can be generated with platforms such as Pacific Biosciences and Oxford Nanopore. The potential of these long-read platforms to increase phylogenetic

inference and obtain species-level resolution has been shown for fungi and bacteria (Benítez-Páez, Portune, & Sanz, 2016; Tedersoo, Tooming-Klunderud, & Anslan, 2017), and recently for protists (Jamy et al., 2019). Direct RNA sequencing without a cDNA intermediate is also now possible and could allow untargeted protist and general microbiome community profiling (Graham et al., 2019; Marinov, 2017).

In comparison to sequencing platforms focusing on increased read lengths, other platforms (especially Illumina) focus on increased sequencing depth. These are particularly suitable for PCR-free “omics” strategies, including whole community profiling via metagenomics and metatranscriptomics (Carradec et al., 2018; Prosser, 2015). When PCR amplification to increase product yield is avoided, -omics methods circumvent PCR/global-amplification biases that distort diversity analyses and provide community profiles of all three domains of life plus viruses, as well as functional gene information (Flues, Bass, & Bonkowski, 2017; Karsenti et al., 2011; Pesant et al., 2015; Turner et al., 2013; Urich et al., 2008). The systematic collection of multi-omics data (metagenomic, metatranscriptomic, single-cell omics) in aquatic (Carradec et al., 2018; Cuvelier et al., 2010; Seeleuthner et al., 2018) and soil (Geisen, Tveit, et al., 2015; Jacquioud et al., 2016; Turner et al., 2013) systems are unveiling a more true picture of diversity and community composition of protists. A drawback of omics approaches is the high cost per sample since a much deeper sequencing of 10-100 million reads per sample is required. Coverage of any particular target group is also limited, as metagenomic datasets may be dominated by bacterial sequences or host sequences from animal/plant samples (Tedersoo et al., 2015; Urich et al., 2008). The key opportunity for protists afforded by -omics approaches, i.e. the elucidation of patterns of gene expression and interactions, is so far constrained by the deficiency of reference databases compared to those available for prokaryotes (Caron et al., 2017). As this information gap is increasingly filled, however, functional -omics information will help us to better understand ecosystem processes beyond taxonomy-based inferences of protist functions that are currently largely limited to nutrient uptake modes (Stefan Geisen et al., 2018). The first studies linking protist taxa to biological functions using -omics data have revealed the immense potential of these approaches (Hu et al., 2018; Ottesen et al., 2013). An alternative -omics approach that focuses more on taxonomic diversity

analyses and as such reducing sequencing costs compared with full metagenomic and metatranscriptomic analyses is mitochondrial enrichment (mitogenomics) (Andújar et al., 2015; Liu et al., 2016) – an approach yet to be implemented for studying any microbial group including protists.

Contemporary best practices for protist community analyses

Despite some drawbacks, metabarcoding of the 18S rRNA gene V4 and V9 region represents by far the most cost-efficient way to assess protist environmental diversity across large spatio-temporal scales (i.e. large number of samples). We emphasise that no true eukaryote-wide universal primer pair without biases can be designed for protists due to both the paraphyletic nature of protists and the extreme phylogenetic diversity they encompass. A cumulative overview and information of primer sets used to date that should give an overview for researchers interested in studying protists is shown in Table 1 that together with the other figures provide an overview of advantages and disadvantages. These show, for instance, that the by far most commonly used primer pair 8 (Stoeck et al., 2010) to date, has clear disadvantages and only matches about 67% of all protists (Table 1), especially prevalent in the most common soil protists: Amoebozoa and Rhizaria. Many other primer pairs that can be applied using the highest throughput Illumina sequencing approaches face similar biases and perform only slightly better (Table 1). Therefore, an unbiased comparison between environments and using data obtained so far is difficult and we cannot suggest an ideal primer pair for exhaustive protist community profiling across systems.

Ideally, long-read sequencing using PacBio or Nanopore sequencing is needed to cover most protistan diversity in a sample using primer pair 6 as it covers about 93% of all protists (Table 1). An alternative is accepting shorter read lengths and as such lower phylogenetic resolution using partial single-ended amplicon analyses (Pauvert et al., 2019) with primer pair 6. Pairing sequence reads reduces the possibilities of primers to select from as their amplicon sizes are longer than the most commonly used Illumina platforms can amplify. This means, a lower diversity of protists can be targeted using 2x300bp, 2x250bp or 2x150bp sequencing with the best primers amplifying 88% (primer pair 6), 75% (34) or 63%

(29) of protistan diversity, respectively. If whole microbiome analyses including protists, bacteria, archaea and fungi are envisioned PCR-free metagenomics and metatranscriptomics are suggested, but their costs currently still commonly are preventing large-scaled analyses. A more cost-efficient alternative to get a cumulative microbiome overview is using primer pairs such as 33 that amplify both prokaryotes and eukaryotes.

We summarize the primer selection for a given study system in a user-friendly diagram in Fig. 5. This overview together with additional information provided in this manuscript should help researchers, particularly ecologists, to choose a primer pair that best suits the research question, system studied and sequence platform available. This should facilitate analyses of protistan diversity within mainstream ecological and microbiome studies, and also applied research such as biomonitoring and bioindication (Stefan Geisen et al., 2018; J. Pawlowski, Lejzerowicz, Apotheloz-Perret-Gentil, Visco, & Esling, 2016) . Following this guide, protist community analyses will also become more comparable between studies leading to a greater capacity for, and more meaningful, comparisons of protist communities across eco-systems, similar to that possible for bacteria (Knight et al., 2018) and fungi (Nilsson et al., 2019). Our intention is that this guide develops over time, therefore we intend to regularly update the available metabarcoding primers to study protists (<https://github.com/pr2database/pr2-primers>). We ask the community for their active input in keeping this work up-to-date, in order to fully establish protist community profiling as a standard tool in future microbiome studies (<https://github.com/pr2database/pr2-primers/issues>).

Methodology

18S rRNA gene primer sets used in metabarcoding studies were collected from literature. Primer sequences and primer sets (knowing that several primer sets may share at least one primer) were stored in a custom MySQL database (full list available at <https://github.com/pr2database/pr2-primers/wiki/18S-rRNA-primer-sets> and <https://github.com/pr2database/pr2-primers/wiki/18S-rRNA-primers>). Primer sets targeting the V4 and V9 region of the 18S rRNA gene were selected to determine in silico amplification of sequences stored in version 4.12.0 of the PR² database (Guillou et al. 2013, <https://github.com/pr2database/pr2database/releases/tag/v4.12.0>). Sequences with ambiguities were discarded. For V4 and V9, sequences with length shorter than 1200 and 1650 bp, respectively, were not considered. Moreover, for V9, since many 18S rRNA do not cover the full V9 region, we only kept sequences that contains the canonical sequence GGATC[AT] which is located at the end of the V9 region, just before the start of the internal transcribed spacer 1. A R script was used to compute the % of sequences matching the forward, reverse and both primers using the Biostrings package function vmatchPattern() with the following parameters: max.mismatch=0, min.mismatch=0, with.indels=FALSE, fixed=FALSE, algorithm="auto". The data were tabulated using the dplyr package and plotted using the ggplot2 package. All scripts are available at <https://github.com/pr2database/pr2-primers>.

Acknowledgements

This study was funded by the Gordon and Betty Moore Foundation through grant GBMF5257 (UniEuk) and the International Society of Protistologists. SG was supported by an NWO-VENI grant (016.Veni.181.078) from the Netherlands Organisation for Scientific Research. This work was also supported by NERC grants NE/H009426/1 (DB, CB) and NE/H000887/1 (DB), and by funding from the UK Department of Environment, Food and Rural Affairs (Defra) under contract FC1214 (to DB). EL was supported by a grant "Atracción de Talento Investigador" by the Community of Madrid ([2017-](#)

[T1/AMB-5210](#)). CdV was supported by the French Government “Investissements d’Avenir” program OCEANOMICS (ANR-11-BTBR- 0008).

We thank Linda Amaral-Zettler, Sophie Arnaud-Haond, Sandra Baldauf, Sergio Balzano, Cedric Berney Jens Boenigk, Jane Carlton, Simon Creer, Didier Debroas, Micah Dunthorn, Javier del Campo, Isabelle Domaizon, Virginia Edgcomb, Bente Edwardsen, Noah Fierer, Laure Guillou, Laura Katz, Patrick Keeling, Rob Knight, Franck Lejzerowicz, Purificacion Lopez-Garcia, Connie Lovejoy, Ramon Massana, Sebastian Metz, Edward Mitchell, Angela Oliverio, Xavier Pochon, Chris Seppey, David Singer, Alexey Smirnov, Thorsten Stoeck, Alexandra Worden, Adriana Zingone for initial discussions on primer choices for eukaryotes.

References

- Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., . . . Zhang, Q. (2019). Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology*, 66(1), 4-119. doi:10.1111/jeu.12691
- Adl, S. M., Habura, A., & Eglit, Y. (2014). Amplification primers of SSU rDNA for soil protists. *Soil Biol. Biochem.*, 69, 328–342. doi:<http://dx.doi.org/10.1016/j.soilbio.2013.10.024>
- Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One*, 4(7), e6372.
- Andre, A., Quillevère, F., Morard, R., Ujiie, Y., Escarguel, G., de Vargas, C., . . . Douady, C. J. (2014). SSU rDNA divergence in planktonic foraminifera: molecular taxonomy and biogeographic implications. *PLoS One*, 9(8), e104641. doi:10.1371/journal.pone.0104641
- Andújar, C., Arribas, P., Ruzicka, F., Crampton-Platt, A., Timmermans, M. J. T. N., & Vogler, A. P. (2015). Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Mol. Ecol.*, 24(14), 3603-3617. doi:10.1111/mec.13195
- Azam, F., Fenchel, T., Field, J., Gray, J., Meyer-Reil, L., & Thingstad, F. (1983). The ecological role of water-column microbes in the sea. *Marine Ecology Progress Series* 10, 257-263.
- Balzano, S., Corre, E., Decelle, J., Sierra, R., Wincker, P., Da Silva, C., . . . Not, F. (2015). Transcriptome analyses to investigate symbiotic relationships between marine protists. *Frontiers in Microbiology*, 6(98). doi:10.3389/fmicb.2015.00098
- Bass, D., & Richards, T. A. (2011). Three reasons to re-evaluate fungal diversity ‘on Earth and in the ocean’. *Fungal Biology Reviews*, 25(4), 159-164. doi:<https://doi.org/10.1016/j.fbr.2011.10.003>
- Bass, D., Silberman, J. D., Brown, M. W., Tice, A. K., Jousset, A., Geisen, S., & Hartikainen, H. (2016). Coprophilic amoebae and flagellates, including *Guttulinopsis*, *Rosculus* and *Helkesimastix*, characterise a divergent and diverse rhizarian radiation and contribute to a large diversity of

- faecal-associated protists. *Environmental Microbiology*, 18(5), 1604–1619. doi:10.1111/1462-2920.13235
- Bass, D., Stentiford, G. D., Littlewood, D. T. J., & Hartikainen, H. (2015). Diverse applications of environmental DNA methods in parasitology. *Trends in Parasitology*, 31(10), 499–513. doi:<http://dx.doi.org/10.1016/j.pt.2015.06.013>
- Bass, D., Yabuki, A., Santini, S., Romac, S., & Berney, C. (2012). *Reticulamoeba* is a long-branched granofilosean (Cercozoa) that is missing from sequence databases. *PLoS One*, 7(12), e49090. doi:10.1371/journal.pone.0049090
- Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*, 5(1), 4. doi:10.1186/s13742-016-0111-z
- Berney, C., Ciuprina, A., Bender, S., Brodie, J., Edgcomb, V., Kim, E., . . . de Vargas, C. (2017). UniEuk: Time to Speak a Common Language in Protistology! *Journal of Eukaryotic Microbiology*, 64(3), 407–411. doi:10.1111/jeu.12414
- Berney, C., Fahrni, J., & Pawlowski, J. (2004). How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol.*, 2(1), 13.
- Boenigk, J., Ereshefsky, M., Hoef-Emden, K., Mallet, J., & Bass, D. (2012). Concepts in protistology: species definitions and boundaries. *Eur. J. Protistol.*, 48(2), 96–102. doi:<http://dx.doi.org/10.1016/j.ejop.2011.11.004>
- Bonkowski, M. (2004). Protozoa and plant growth: the microbial loop in soil revisited. *New Phytologist*, 162(3), 617–631. doi:10.1111/j.1469-8137.2004.01066.x
- Boscaro, V., Santoferrara, L. F., Zhang, Q., Gentekaki, E., Syberg-Olsen, M. J., Del Campo, J., & Keeling, P. J. (2018). EukRef-Ciliophora: a manually curated, phylogeny-based database of small subunit rRNA gene sequences of ciliates. *Environ Microbiol*, 20(6), 2218–2230. doi:10.1111/1462-2920.14264
- Bower, S. M., Carnegie, R. B., Goh, B., Jones, S. R. M., Lowe, G. J., & Mak, M. W. S. (2004). Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. *Journal of Eukaryotic Microbiology*, 51(3), 325–332. doi:10.1111/j.1550-7408.2004.tb00574.x
- Brâte, J., Logares, R., Berney, C., Ree, D. K., Klaveness, D., Jakobsen, K. S., & Shalchian-Tabrizi, K. (2010). Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *The ISME Journal*, 4, 1144. doi:10.1038/ismej.2010.39
- <https://www.nature.com/articles/ismej201039#supplementary-information>
- Buée, M., Reich, M., Murat, C., Morin, E., Nilsson, R. H., Uroz, S., & Martin, F. (2009). 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytol.*, 184(2), 449–456. doi:10.1111/j.1469-8137.2009.03003.x
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, 11(12), 2639–2643. doi:10.1038/ismej.2017.119
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335. doi:10.1038/nmeth.f.303
- <https://www.nature.com/articles/nmeth.f.303#supplementary-information>
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., . . . Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA*, 108(Supplement 1), 4516–4522. doi:DOI 10.1073/pnas.1000080107
- Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., . . . Worden, A. Z. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat Rev Micro*, 15(1), 6–20. doi:10.1038/nrmicro.2016.160

- <http://www.nature.com/nrmicro/journal/v15/n1/abs/nrmicro.2016.160.html#supplementary-information>
- Caron, D. A., & Hu, S. K. (2019). Are we overestimating protistan diversity in nature? *Trends in Microbiology*, 27(3), 197-205. doi:<https://doi.org/10.1016/j.tim.2018.10.009>
- Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., . . . Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1), 373. doi:10.1038/s41467-017-02342-1
- Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4, 265. doi:10.1038/nnano.2009.12
- <https://www.nature.com/articles/nnano.2009.12#supplementary-information>
- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., . . . Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1), D633-D642. doi:10.1093/nar/gkt1244
- Comeau, A. M., Li, W. K., Tremblay, J. E., Carmack, E. C., & Lovejoy, C. (2011). Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS One*, 6(11), e27492. doi:10.1371/journal.pone.0027492
- Cuvelier, M. L., Allen, A. E., Monier, A., McCrow, J. P., Messié, M., Tringe, S. G., . . . Worden, A. Z. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences*, 107(33), 14679. doi:10.1073/pnas.1001665107
- de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., . . . Karsenti, E. (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1261605. doi:10.1126/science.1261605
- del Campo, J., Bass, D., & Keeling, P. (2019). The eukaryome: diversity and role of micro-eukaryotic organisms associated with animal hosts. *Functional Ecology*, in press.
- del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L. F., Nenarokov, S., Massana, R., . . . Wegener Parfrey, L. (2018). EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLOS Biology*, 16(9), e2005849. doi:10.1371/journal.pbio.2005849
- del Campo, J., Pons, M. J., Herranz, M., Wakeman, K. C., del Valle, J., Vermeij, M. J. A., . . . Keeling, P. J. (2019). Validation of a universal set of primers to study animal-associated microeukaryotic communities. *Environmental Microbiology*, 0(0). doi:10.1111/1462-2920.14733
- Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., . . . Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science*, 359(6373), 320-325. doi:10.1126/science.aap9516
- Dunthorn, M., Klier, J., Bunge, J., & Stoeck, T. (2012). Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *Journal of Eukaryotic Microbiology*, 59(2), 185-187. doi:10.1111/j.1550-7408.2011.00602.x
- Dupont, A. O., Griffiths, R. I., Bell, T., & Bass, D. (2016). Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. *Environ Microbiol*, 18(6), 2010-2024. doi:10.1111/1462-2920.13220
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460-2461. doi:10.1093/bioinformatics/btq461
- Euringer, K., & Lueders, T. (2008). An optimised PCR/T-RFLP fingerprinting approach for the investigation of protistan communities in groundwater environments. *Journal of Microbiological Methods*, 75(2), 262-268. doi:<http://dx.doi.org/10.1016/j.mimet.2008.06.012>
- Fiore-Donno, A. M., Weinert, J., Wubet, T., & Bonkowski, M. (2016). Metacommunity analysis of amoeboid protists in grassland soils. *Scientific Reports*, 6, 19068.
- Fiore-Donno Anna, M., Rixen, C., Rippin, M., Glaser, K., Samolov, E., Karsten, U., . . . Bonkowski, M. (2017). 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(2), 229-239. doi:10.1111/1755-0998.12729
- Flues, S., Bass, D., & Bonkowski, M. (2017). Grazing of leaf-associated Cercomonads (Protists:
Rhizaria: Cercozoa) structures bacterial community composition and function. *Environmental
Microbiology*, 19(8), 3297-3309. doi:10.1111/1462-2920.13824
- Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J., & Cotter, P. D. (2016). 16S rRNA gene
sequencing of mock microbial populations- impact of DNA extraction method, primer choice
and sequencing platform. *BMC Microbiology*, 16(1), 123. doi:10.1186/s12866-016-0738-z
- Gans, J., Wolinsky, M., & Dunbar, J. (2005). Computational improvements reveal great bacterial
diversity and high metal toxicity in soil. *Science*, 309(5739), 1387-1390.
doi:10.1126/science.1112665
- García-Martín, J. M., Zamora, J. C., & Lado, C. (2019). Evidence of intra-individual SSU polymorphisms
in dark-spored Myxomycetes (Amoebozoa). *Protist*, 170(5), 125681.
doi:<https://doi.org/10.1016/j.protis.2019.125681>
- Geisen, S. (2016). Thorough high-throughput sequencing analyses unravels huge diversities of soil
parasitic protists. *Environmental Microbiology*, 18(6), 1669-1672. doi:10.1111/1462-
2920.13309
- 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. *Mol Ecol*, 24(17), 4556-4569. doi:DOI: 10.1111/mec.13238
- Geisen, S., Mitchell, E. A. D., Adl, S., Bonkowski, M., Dunthorn, M., Ekelund, F., . . . Lara, E. (2018). Soil
protists: a fertile frontier in soil biology research. *FEMS Microbiol Rev*, 42(3), 293-323.
doi:10.1093/femsre/fuy006
- Geisen, S., Mitchell, E. A. D., Adl, S., Bonkowski, M., Dunthorn, M., Ekelund, F., . . . Lara, E. (2018). Soil
protists: a fertile frontier in soil biology research. *FEMS Microbiology Reviews*, 42(3), 293-
323. doi:10.1093/femsre/fuy006
- Geisen, S., Mitchell, E. A. D., Wilkinson, D. M., Adl, S., Bonkowski, M., Brown, M. W., . . . Lara, E.
(2017). Soil protistology rebooted: 30 fundamental questions to start with. *Soil Biology and
Biochemistry*, 111(Supplement C), 94-103. doi:<https://doi.org/10.1016/j.soilbio.2017.04.001>
- Geisen, S., Tveit, A. T., Clark, I. M., Richter, A., Svenning, M. M., Bonkowski, M., & Urich, T. (2015).
Metatranscriptomic census of active protists in soils. *ISME J*, 9(10), 2178-2190.
doi:10.1038/ismej.2015.30
- Giner, C. R., Forn, I., Romac, S., Logares, R., de Vargas, C., & Massana, R. (2016). Environmental
sequencing provides reasonable estimates of the relative abundance of specific
picoeukaryotes. *Applied and Environmental Microbiology*, 82(15), 4757.
doi:10.1128/AEM.00560-16
- Giovannoni, S., & Cary, S. C. (1993). Probing marine systems with ribosomal RNAs. *Oceanography*,
6(3), 95-104.
- Gong, J., Dong, J., Liu, X., & Massana, R. (2013). Extremely high copy numbers and polymorphisms of
the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates.
Protist, 164(3), 369-379. doi:<http://dx.doi.org/10.1016/j.protis.2012.11.006>
- Graham, N., Gillespie, R. G., Kennedy, S. R., Lim, J. Y., Krehenwinkel, H., Pomerantz, A., . . .
Shoobridge, J. D. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables
portable and simple biodiversity assessments with high phylogenetic resolution across broad
taxonomic scale. *GigaScience*, giz006. doi:10.1093/gigascience/giz006
- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., . . . Christen, R. (2013). The Protist
Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA
sequences with curated taxonomy. *Nucleic Acids Research*, 41(D1), D597-D604.
doi:10.1093/nar/gks1160
- Guiry, M. D., & Guiry, G. (2008). AlgaeBase. *AlgaeBase*.

- Hadziavdic, K., Lekang, K., Lanzén, A., Jonassen, I., Thompson, E. M., & Troedsson, C. (2014). Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One*, 9(2), e87624. doi:10.1371/journal.pone.0087624
- Harder, C. B., Ronn, R., Brejnrod, A., Bass, D., Al-Soud, W. A., & Ekelund, F. (2016). Local diversity of heathland Cercozoa explored by in-depth sequencing. *ISME J*, 10(10), 2488-2497. doi:10.1038/ismej.2016.31
- Hartikainen, H., Ashford, O. S., Berney, C., Okamura, B., Feist, S. W., Baker-Austin, C., . . . Bass, D. (2014). Lineage-specific molecular probing reveals novel diversity and ecological partitioning of haplosporidians. *The Isme Journal*, 8, 177. doi:10.1038/ismej.2013.136
- <https://www.nature.com/articles/ismej2013136#supplementary-information>
- Hawksworth, D. L., & Lucking, R. (2017). Fungal Diversity Revisited: 2.2 to 3.8 Million Species. *Microbiol Spectr*, 5(4). doi:10.1128/microbiolspec.FUNK-0052-2016
- Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological sciences*, 270 Suppl 1(Suppl 1), S96-S99. doi:10.1098/rsbl.2003.0025
- Heger, T. J., Pawlowski, J., Lara, E., Leander, B. S., Todorov, M., Golemansky, V., & Mitchell, E. A. D. (2011). Comparing potential COI and SSU rDNA barcodes for assessing the diversity and phylogenetic relationships of cyphoderiid testate amoebae (Rhizaria: Euglyphida). *Protist*, 162(1), 131-141. doi:10.1016/j.protis.2010.05.002
- Hu Sarah, K., Liu, Z., Lie Alle, A. Y., Countway Peter, D., Kim Diane, Y., Jones Adriane, C., . . . Caron David, A. (2015). Estimating protistan diversity using high-throughput sequencing. *Journal of Eukaryotic Microbiology*, 62(5), 688-693. doi:10.1111/jeu.12217
- Hu, S. K., Liu, Z., Alexander, H., Campbell, V., Connell, P. E., Dyhrman, S. T., . . . Caron, D. A. (2018). Shifting metabolic priorities among key protistan taxa within and below the euphotic zone. *Environmental Microbiology*, 20(8), 2865-2879. doi:10.1111/1462-2920.14259
- Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., . . . Andersson, A. F. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One*, 9(4), e95567. doi:10.1371/journal.pone.0095567
- Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7), R143. doi:10.1186/gb-2007-8-7-r143
- Ihrmark, K., Bödeker, I. T. M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., . . . Lindahl, B. D. (2012). New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol. Ecol.*, 82(3), 666-677. doi:10.1111/j.1574-6941.2012.01437.x
- Jacquioud, S., Stenbaek, J., Santos, S. S., Winding, A., Sorensen, S. J., & Prieme, A. (2016). Metagenomes provide valuable comparative information on soil microeukaryotes. *Res Microbiol*, 167(5), 436-450. doi:10.1016/j.resmic.2016.03.003
- Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., . . . Burki, F. (2019). Long metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *bioRxiv*, 627828. doi:10.1101/627828
- Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., . . . the Tara Oceans, C. (2011). A holistic approach to marine eco-systems biology. *PLOS Biology*, 9(10), e1001177. doi:10.1371/journal.pbio.1001177
- Knight, R., Vrbanc, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., . . . Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7), 410-422. doi:10.1038/s41579-018-0029-9
- Kosakyan, A., Heger, T. J., Leander, B. S., Todorov, M., Mitchell, E. A. D., & Lara, E. (2012). COI barcoding of nebelid testate amoebae (Amoebozoa: Arcellinida): extensive cryptic diversity and redefinition of the Hyalospheniidae Schultze. *Protist*, 163(3), 415-434. doi:10.1016/j.protis.2011.10.003

- Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaultot, D., Bouget, F.-Y., & Galand, P. E. (2019). Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, 13(2), 388-401. doi:10.1038/s41396-018-0281-z
- Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G. W., . . . Schleper, C. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, 442(7104), 806-809. doi:http://www.nature.com/nature/journal/v442/n7104/supinfo/nature04983_S1.html
- Lentendu, G., Wubet, T., Chatzinotas, A., Wilhelm, C., Buscot, F., & Schlegel, M. (2014). Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach. *Mol. Ecol.*, 23(13), 3341-3355. doi:10.1111/mec.12819
- Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., . . . Raes, J. (2015). Determinants of community structure in the global plankton interactome. *Science*, 348(6237), 1262073. doi:10.1126/science.1262073
- Lindahl Björn, D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjoller, R., . . . Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist*, 199(1), 288-299. doi:10.1111/nph.12243
- Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., . . . Zhou, X. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, 16(2), 470-479. doi:10.1111/1755-0998.12472
- Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C., & Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, 409(6820), 603-607. doi:10.1038/35054537
- Louca, S., Mazel, F., Doebeli, M., & Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biology*, 17(2), e3000106. doi:10.1371/journal.pbio.3000106
- Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., . . . Dangl, J. L. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, 488(7409), 86-90. doi:<http://www.nature.com/nature/journal/v488/n7409/abs/nature11237.html#supplementary-information>
- Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., . . . Dunthorn, M. (2017). Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution*, 1, 0091. doi:10.1038/s41559-017-0091-<http://www.nature.com/articles/s41559-017-0091#supplementary-information>
- Mahe, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2, e593. doi:10.7717/peerj.593
- Mahe, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3, e1420. doi:10.7717/peerj.1420
- Mangot, J.-F., Domaizon, I., Taib, N., Marouni, N., Duffaud, E., Bronner, G., & Debroas, D. (2013). Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environmental Microbiology*, 15(6), 1745-1758. doi:10.1111/1462-2920.12065
- Marinov, G. K. (2017). On the design and prospects of direct RNA sequencing. *Brief Funct Genomics*, 16(6), 326-335. doi:10.1093/bfpg/elw043
- Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J. H., . . . Bakker, P. A. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, 332(6033), 1097-1100.
- Moon-van der Staay, S. Y., De Wachter, R., & Vaultot, D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, 409(6820), 607-610.
- Morard, R., Darling, K. F., Mahé, F., Audic, S., Ujiié, Y., Weiner, A. K. M., . . . de Vargas, C. (2015). PFR2: a curated database of planktonic foraminifera 18S ribosomal DNA as a resource for studies of plankton ecology, biogeography and evolution. *Molecular Ecology Resources*, 15(6), 1472-1485. doi:10.1111/1755-0998.12410

- Mordret, S., Piredda, R., Vaultot, D., Montresor, M., Kooistra, W. H. C. F., & Sarno, D. (2018). dinoref: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. *Molecular Ecology Resources*, 18(5), 974-987. doi:10.1111/1755-0998.12781
- Moreira, D., & López-García, P. (2002). The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.*, 10(1), 31-38. doi:[http://dx.doi.org/10.1016/S0966-842X\(01\)02257-0](http://dx.doi.org/10.1016/S0966-842X(01)02257-0)
- Moreno, Y., Moreno-Mesonero, L., Amorós, I., Pérez, R., Morillo, J. A., & Alonso, J. L. (2018). Multiple identification of most important waterborne protozoa in surface water used for irrigation purposes by 18S rRNA amplicon-based metagenomics. *International Journal of Hygiene and Environmental Health*, 221(1), 102-111. doi:<https://doi.org/10.1016/j.ijheh.2017.10.008>
- Nassonova, E., Smirnov, A., Fahrni, J., & Pawlowski, J. (2010). Barcoding amoebae: comparison of SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae. *Protist*, 161(1), 102-115.
- Needham, D. M., & Fuhrman, J. A. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology*, 1(4), 16005. doi:10.1038/nmicrobiol.2016.5
- Neefs, J.-M., Van de Peer, Y., De Rijk, P., Chapelle, S., & De Wachter, R. (1993). Compilation of small ribosomal subunit RNA structures. *Nucleic Acids Research*, 21(13), 3025-3049. doi:10.1093/nar/21.13.3025
- Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat Rev Microbiol*, 17(2), 95-109. doi:10.1038/s41579-018-0116-y
- Ottesen, E. A., Young, C. R., Eppley, J. M., Ryan, J. P., Chavez, F. P., Scholin, C. A., & DeLong, E. F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences*, 110(6), E488. doi:10.1073/pnas.1222099110
- Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313), 734-740.
- Pagenkopp Lohan, K. M., Fleischer, R. C., Carney, K. J., Holzer, K. K., & Ruiz, G. M. (2016). Amplicon-based pyrosequencing reveals high diversity of protistan parasites in ships' ballast water: Implications for biogeography and infectious diseases. *Microbial Ecology*, 71(3), 530-542. doi:10.1007/s00248-015-0684-6
- Parfrey, L. W., Walters, W. A., Lauber, C. L., Clemente, J. C., Berg-Lyons, D., Telling, C., . . . Knight, R. (2014). Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Frontiers in Microbiology*, 5, 298. doi:10.3389/fmicb.2014.00298
- Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., . . . Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2(11), 1533-1542. doi:10.1038/s41564-017-0012-7
- Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., . . . Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, 41, 23-33. doi:<https://doi.org/10.1016/j.funeco.2019.03.005>
- Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., . . . de Vargas, C. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biology*, 10(11), e1001419. doi:10.1371/journal.pbio.1001419
- Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, 55, 12-25. doi:<https://doi.org/10.1016/j.ejop.2016.02.003>
- Pearson, W. R. (2014). BLAST and FASTA Similarity Searching for Multiple Sequence Alignment. In D. J. Russell (Ed.), *Multiple Sequence Alignment Methods* (pp. 75-101). Totowa, NJ: Humana Press.

- 715 Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., . . . Wincker, P. (2015).
716 Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2,
717 150023. doi:10.1038/sdata.2015.23
- 718 Piredda, R., Tomasino, M. P., D'Erchia, A. M., Manzari, C., Pesole, G., Montresor, M., . . . Zingone, A.
719 (2017). Diversity and temporal patterns of planktonic protist assemblages at a
720 Mediterranean Long Term Ecological Research site. *FEMS Microbiol Ecol*, 93(1).
721 doi:10.1093/femsec/fiw200
- 722 Pitsch, G., Bruni, E. P., Forster, D., Qu, Z., Sonntag, B., Stoeck, T., & Posch, T. (2019). Seasonality of
723 planktonic freshwater ciliates: are analyses based on V9 regions of the 18S rRNA gene
724 correlated with morphospecies counts? *Frontiers in Microbiology*, 10(248).
725 doi:10.3389/fmicb.2019.00248
- 726 Prosser, J. I. (2015). Dispersing misconceptions and identifying opportunities for the use of 'omics' in
727 soil microbial ecology. *Nature Reviews Microbiology*, 13(7), 439-446.
728 doi:10.1038/nrmicro3468
- 729 Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA:
730 a comprehensive online resource for quality checked and aligned ribosomal RNA sequence
731 data compatible with ARB. *Nucleic Acids Research*, 35(21), 7188-7196.
732 doi:10.1093/nar/gkm864
- 733 Ramirez, K. S., Knight, C. G., de Hollander, M., Brearley, F. Q., Constantinides, B., Cotton, A., . . . de
734 Vries, F. T. (2018). Detecting macroecological patterns in bacterial communities across
735 independent studies of global soils. *Nature Microbiology*, 3(2), 189-196. doi:10.1038/s41564-
736 017-0062-x
- 737 Ramirez, K. S., Snoek, L. B., Koorem, K., Geisen, S., Bloem, L. J., ten Hooven, F., . . . van der Putten, W.
738 H. (2019). Range-expansion effects on the belowground plant microbiome. *Nature Ecology &
739 Evolution*, 3(4), 604-611. doi:10.1038/s41559-019-0828-z
- 740 Richards, T. A., & Bass, D. (2005). Molecular screening of free-living microbial eukaryotes: diversity
741 and distribution using a meta-analysis. *Current Opinion in Microbiology*, 8(3), 240-252.
742 doi:<https://doi.org/10.1016/j.mib.2005.04.010>
- 743 Rodríguez-Martínez, R., Roca, G., Logares, R., Romac, S., & Massana, R. (2011). Low evolutionary
744 diversification in a widespread and abundant uncultured protist (MAST-4). *Molecular Biology
745 and Evolution*, 29(5), 1393-1406. doi:10.1093/molbev/msr303
- 746 Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source
747 tool for metagenomics. *PeerJ Preprints*, 4, e2409v2401. doi:10.7287/peerj.preprints.2409v1
- 748 Sapkota, R., & Nicolaisen, M. (2015). An improved high throughput sequencing method for studying
749 oomycete communities. *J. Microbiol. Meth.*, 110(0), 33-39.
750 doi:<http://dx.doi.org/10.1016/j.mimet.2015.01.013>
- 751 Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., . . . Tara Oceans, C.
752 (2018). Single-cell genomics of multiple uncultured stramenopiles reveals underestimated
753 functional diversity across oceans. *Nature Communications*, 9(1), 310. doi:10.1038/s41467-
754 017-02235-3
- 755 Simon, M., Jardillier, L., Deschamps, P., Moreira, D., Restoux, G., Bertolino, P., & López-García, P.
756 (2015). Complex communities of small protists and unexpected occurrence of typical marine
757 lineages in shallow freshwater systems. *Environmental Microbiology*, 17(10), 3610-3627.
758 doi:10.1111/1462-2920.12591
- 759 Singer, D., Kosakyan, A., Seppely, C. V. W., Pilonel, A., Fernández, L. D., Fontaneto, D., . . . Lara, E.
760 (2018). Environmental filtering and phylogenetic clustering correlate with the distribution
761 patterns of cryptic protist species. *Ecology*, 99(4), 904-914. doi:10.1002/ecy.2161
- 762 Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., . . . Herndl, G. J.
763 (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc.
764 Natl. Acad. Sci. USA*, 103(32), 12115-12120. doi:10.1073/pnas.0605127103

- 765 Stern, R. F., Andersen, R. A., Jameson, I., Küpper, F. C., Coffroth, M.-A., Vaulot, D., . . . Keeling, P. J.
766 (2012). Evaluating the ribosomal internal transcribed spacer (ITS) as a candidate
767 dinoflagellate barcode marker. *PLoS One*, 7(8), e42780. doi:10.1371/journal.pone.0042780
- 768 Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D., Breiner, H. W., & Richards, T. A. (2010).
769 Multiple marker parallel tag environmental DNA sequencing reveals a highly complex
770 eukaryotic community in marine anoxic water. *Mol. Ecol.*, 19(Supplement s1), 21-31.
771 doi:10.1111/j.1365-294X.2009.04480.x
- 772 Tan, S., & Liu, H. (2018). Unravel the hidden protistan diversity: application of blocking primers to
773 suppress PCR amplification of metazoan DNA. *Applied Microbiology and Biotechnology*,
774 102(1), 389-401. doi:10.1007/s00253-017-8565-1
- 775 Tanabe Akifumi, S., Nagai, S., Hida, K., Yasuike, M., Fujiwara, A., Nakamura, Y., . . . Katakura, S. (2015).
776 Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel
777 sequencing-based monitoring of the planktonic eukaryote community. *Molecular Ecology*
778 *Resources*, 16(2), 402-414. doi:10.1111/1755-0998.12459
- 779 Tedersoo, L., Anslan, S., Bahram, M., Pölme, S., Riit, T., Liiv, I., . . . Abarenkov, K. (2015). Shotgun
780 metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in
781 metabarcoding analyses of fungi. *MycKeys*, 10, 1-43.
- 782 Tedersoo, L., Bahram, M., Pölme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., . . . Abarenkov, K.
783 (2014). Fungal biogeography. Global diversity and geography of soil fungi. *Science*,
784 346(6213), 1256688. doi:10.1126/science.1256688
- 785 Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2017). PacBio metabarcoding of Fungi and other
786 eukaryotes: errors, biases and perspectives. *New Phytologist*, 217(3), 1370-1385.
787 doi:10.1111/nph.14776
- 788 Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., . . . The Earth
789 Microbiome Project, C. (2017). A communal catalogue reveals Earth's multiscale microbial
790 diversity. *Nature*, 551, 457-463. doi:10.1038/nature24621
- 791 <https://www.nature.com/articles/nature24621#supplementary-information>
- 792 Tragin, M., Zingone, A., & Vaulot, D. (2018). Comparison of coastal phytoplankton composition
793 estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic
794 groups and especially Chlorophyta. *Environ Microbiol*, 20(2), 506-520. doi:10.1111/1462-
795 2920.13952
- 796 Turner, T. R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., . . . Poole, P. S.
797 (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere
798 microbiome of plants. *ISME J.*, 7, 2248-2258.
- 799 Urich, T., Lanzén, A., Qi, J., Huson, D. H., Schleper, C., & Schuster, S. C. (2008). Simultaneous
800 assessment of soil microbial community structure and function through analysis of the meta-
801 transcriptome. *PLoS One*, 3(6), e2527. doi:10.1371/journal.pone.0002527
- 802 van Hatten, E. J., Zwart, G., van Agterveld, M. P., Gons, H. J., Ebert, J., & Laanbroek, H. J. (1999).
803 Changes in bacterial and eukaryotic community structure after mass lysis of filamentous
804 cyanobacteria associated with viruses. *Applied and Environmental Microbiology*, 65(2), 795-
805 801.
- 806 Venter, P. C., Nitsche, F., Domonell, A., Heger, P., & Arndt, H. (2017). The protistan microbiome of
807 grassland soil: diversity in the mesoscale. *Protist*, 168(5), 546-564.
808 doi:<https://doi.org/10.1016/j.protis.2017.03.005>
- 809 Vestheim, H., & Jarman, S. (2008). Blocking primers to enhance PCR amplification of rare sequences
810 in mixed samples – a case study on prey DNA in Antarctic krill stomachs. *Front. Zool.*, 5(1), 1-
811 11. doi:10.1186/1742-9994-5-12
- 812 Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., . . . Knight, R.
813 (2016). Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed
814 Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems*, 1(1).
815 doi:10.1128/mSystems.00009-15

- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), 5261. doi:10.1128/AEM.00062-07
- Weber, A. A.-T., & Pawlowski, J. (2014). Wide occurrence of SSU rDNA intragenomic polymorphism in foraminifera and its implications for molecular species identification. *Protist*, 165(5), 645-661. doi:<http://dx.doi.org/10.1016/j.protis.2014.07.006>
- Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev*, 51(2), 221-271.
- Worden, A. Z., Follows, M. J., Giovannoni, S. J., Wilken, S., Zimmerman, A. E., & Keeling, P. J. (2015). Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science*, 347(6223), 1257594. doi:10.1126/science.1257594
- Xiong, W., Li, R., Guo, S., Karlsson, I., Jiao, Z., Xun, W., . . . Geisen, S. (2019). Microbial amendments alter protist communities within the soil microbiome. *Soil Biology and Biochemistry*, 135, 379-382. doi:<https://doi.org/10.1016/j.soilbio.2019.05.025>
- Zhan, A., Hulák, M., Sylvester, F., Huang, X., Adebayo, A. A., Abbott, C. L., . . . MacIsaac, H. J. (2013). High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods Ecol. Evol.*, 4(6), 558-565. doi:10.1111/2041-210x.12037
- Zhu, F., Massana, R., Not, F., Marie, D., & Vaulot, D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.*, 52(1), 79-92. doi:10.1016/j.femsec.2004.10.006
- 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 & Evolution*, 11(3), 173. doi:10.1007/s13127-011-0050-6

Data Accessibility

All data used in this work are included in this manuscript. All scripts used are available at <https://github.com/pr2database/pr2-primers>. All data and primer information will continuously be updated at <https://github.com/pr2database/pr2-primers/issues>.

Author Contribution

SG and DB designed the study; SG, DV, FM and EL drafted the tables with help from all authors; DV and SG created the figures; SG and DB wrote the first draft of the manuscript, supplemented with comments from all authors.

849Tables

850Table 1. Primer sets targeting the V4 and the V9 region that were used in metabarcoding high-throughput sequencing efforts to study protists; id: primer pair
851identifier; #: forward or reverse primer name; Fwd: forward primer; Rev: reverse primer; fwd and rev start/end: positions on *Saccharomyces cerevisiae*
852sequence FU970071; note: id33 also targets prokaryotes; id35 selects against metazoan sequences.

region	id	fwd name	fwd seq	fwd start	fwd end	rev name	rev seq	rev start	rev end	length	Reference
V4	1	F-566	CAGCAGCCGCGGTAATTCC	565	583	R-1200	CCCGTGTGAGTCAAATTA AGC	1178	1199	635	(Hadziavdic et al., 2014)
V4	2	A-528F	GCGGTAATTCCAGCTCCAA	573	591	R-952	TTGGCAAATGCTTTCGC	935	951	379	(Hadziavdic et al., 2014)
V4	3	574*f	CGGTAAYTCCAGCTCYV	574	590	1132r	CCGTCAATTHCTTYAART	1132	1149	576	(Hugerth et al., 2014)
V4	4	563f	GCCAGCAVCYGGGTAAY	563	580	1132r	CCGTCAATTHCTTYAART	1132	1149	587	(Hugerth et al., 2014)
V4	5	616f	TTAAAVGYTCGTAGTYG	616	633	1132r	CCGTCAATTHCTTYAART	1132	1149	534	(Hugerth et al., 2014)
V4	6	616*f	TTAAARVGYTCGTAGTYG	616	633	1132r	CCGTCAATTHCTTYAART	1132	1149	534	(Hugerth et al., 2014)
V4	7	V4_1f	CCAGCASCYGCGGTAATWCC	564	583	TAREukREV3	ACTTTCGTTCTTGATYRA	963	980	417	(Bass et al., 2016)
V4	8	TAREuk45 4FWD1	CCAGCASCYGCGGTAATTCC	564	583	TAREukREV3	ACTTTCGTTCTTGATYRA	963	980	417	(Stoeck et al., 2010)
V4	13	3NDf	GGCAAGTCTGGTGCCAG	551	567	V4_euk_R1	GACTACGACGGTATCTRAT CRTCTTCG	989	1015	465	(Bråte et al., 2010)
V4	14	3NDf	GGCAAGTCTGGTGCCAG	551	567	V4_euk_R2	ACGGTATCTRATCRTCTTC G	989	1008	458	(Bråte et al., 2010)
V4	15	EUKAF	GCCGCGGTAATTCCAGCTC	570	588	EUKAR	CYTTCGYCTTGATTRA	963	979	410	(Moreno et al., 2018)
V4	16	TAREuk45 4FWD1	CCAGCASCYGCGGTAATTCC	564	583	V4 18S Next.Rev	ACTTTCGTTCTTGATYRATG A	960	980	417	(Piredda et al., 2017)
V4	17	E572F	CYGCGGTAATTCCAGCTC	571	588	E1009R	AYGGTATCTRATCRTCTTY G	989	1008	438	(Comeau, Li, Tremblay, Carmack, & Lovejoy, 2011)
V4	18	515F	GTGCCAGCMGCCGCGGTAA	561	579	1119r	GGTGCCCTTCCGTCA	1144	1158	598	(Parfrey et al., 2014)

V4	23	590F	CGGTAATTCCAGCTCCAATAG C	574	595	1300R	CACCAACTAAGAACGGCC ATGC	1272	1293	720	(Venter, Nitsche, Domonell, Heger, & Arndt, 2017)
V4	24	EK-565F- NGS	GCAGTTAAAAAGCTCGTAGT	612	631	EUK1134-R	TTTAAGTTTCAGCCTTGCG	1120	1138	527	(Simon et al., 2015)
V4	25	NSF563	CGCGGTAATTCCAGCTCCA	572	590	NSR951	TTGGYRAATGCTTTCGC	935	951	380	(Mangot et al., 2013)
V9	27	1391F	GTACACACCGCCCGTC	1628	1643	EukB	TGATCCTTCTGCAGGTTCA CCTAC	1773	1796	169	(Stoeck et al., 2010)
V9	28	1380F	CCCTGCCHTTTGTACACAC	1617	1635	1510R	CCTTCYGCAGGTTACCTA C	1773	1792	176	(Amaral-Zettler, McCliment, Ducklow Huse, 2009)
V9	29	1389F	TTGTACACACCGCCC	1626	1640	1510R	CCTTCYGCAGGTTACCTA C	1773	1792	167	(Amaral-Zettler et al., 2009)
V9	31	1388F	TTGTACACACCGCCCGTCGC	1626	1645	1510R	CCTTCYGCAGGTTACCTA C	1773	1792	167	(Piredda et al., 2017)
V4	33	515F Univ	GTGYCAGCMGCCGCGGTAA	561	579	926R	CCGYCAATTYMTTTRAGTT T	1130	1149	589	(Needham & Fuhrman, 2016)
V4	34	515FY	GTGYCAGCMGCCGCGGTAA	561	578	NSR951	TTGGYRAATGCTTTCGC	935	951	391	(Lambert et al., 2015)
V4	36	TAReuk45 4FWD1	CCAGCASCYGCGGTAATTCC	564	583	V4RB	ACTTTCGTTCTTGATYRR	963	980	417	(Balzano et al., 2015)
V4	40	Uni18SF	AGGGCAAKYCTGGTGCCAGC	549	568	Uni18SR	GRCGGTATCTRATCGYCTT	991	1009	461	(Zhan et al., 2013)

853

854

Figure legends

856

857Fig. 1. Left panel. Percentage of reference sequences with a perfect match to both forward and reverse
858primers individually and the entire (A) V4 and (B) V9 18S rRNA gene region; right panels show amplicon
859sizes targeted by different primer pairs including lengths that can be covered by the most commonly
860used Illumina sequencers (dotted line: 2x300 base pairs; dashed line: 2x250 base pairs); error bars
861represent the standard deviation. PR² version 4.11.1 was used the reference 18S rRNA database. Note
862that this figure provides an overview of all currently used primer sets to target protists; details for each
863of the primers is given in Table 1.

864

865Fig. 2. Amplicon lengths differences of higher taxonomic level protist lineages exemplified for each of
866a broadly targeted V4 (A) and V9 (B) primer set (Table 1). Lengths differences are prevalent between
867protist groups especially in the V4 region leading to differential amplification efficiency between the
868groups. Note that Hacrobia represents the sum of Haptophytes, Cryptophytes and Centrohelids. n
869represents the number of taxa for a given supergroup present in PR².

870

871Fig. 3. Coverage of some of the most broadly targeting primer pairs (Fig. 1, Table 1) as identified with
872perfect matches of both primer pairs to the target sequences for the main protist lineages. This shows
873that (A) primers do not equally amplify higher taxonomic level lineages of protists and (B) amplicon
874lengths differ between supergroups and depending on primer sets. Note that Hacrobia represents the
875sum of Haptophytes, Cryptophytes and Centrohelids. For details on all primer sets see Suppl. Fig. 1.

876

877

878Fig. 4. Coverage differences at higher taxonomic resolution (which corresponds, approximatively, to
879Class level) as identified with perfect matches of both forward and reverse primers illustrated for
880primer sets 16 and 21, showing that amplification success differences between protist groups (see Fig.
8812) can also be present at lower taxonomic levels. The “universal” primer set 16 does not amplify
882Haptophyta *in silico*, but in fact partially amplified them in natural communities *in vivo* likely because
883the mismatch is located in the middle of the primer (Piredda et al., 2017). Primer set 21 which is
884described as specific of diatoms (Zimmermann, Jahn, & Gemeinholzer, 2011) also amplifies other
885Ochrophyta as well as some green algae.

886

887Fig. 5. Decision-making chart to guide molecular approaches for protist community analyses. The ideal
888primer choice depends on the available sequencing platforms and the study question (protist- focused

889or microbiome- that target all prokaryotes and eukaryotes simultaneously). The longer the sequencing
890read, the broader theoretical coverage of protistan diversity. See Table 1 and main Figs. for detailed
891information on the respective primers.

892

893Supporting Information

894Supplementary Table 1. List of protist group-specific primers so far used in high-throughput sequencing
895based metabarcoding studies

896

Supergroups	Targeted group	primer name	sense	primer sequence 5'-3'	Marker	region in the	average ampli	Sequencing technology	reference paper	Comments
Alveolata	Tintinnids and related	152+ 528-	forward reverse	TTA CAT GGA TAA CCG TGG TAA TTC CCC GGC CCG TTA TTT CTT GT	18S rRNA	V2-V3	308 bp	MISeq	Santoferrara, L., et al. (2018) Journal of Plankton Research, 40, 209-221	
	Symbiodinium spp	SYM_VAR_5.8S2 SYM_VAR_REV	forward reverse	GAATTGCAGAACTCCGTGAACC CGGGTTCWCTTGTGTGACTTCATGC	ITS 2		250 bp	MISeq	Hume, B.C.C., et al. (2018) Peerj, 6, 22.	
	Dinophyta (=Dinoflagellates)	D1R 305R	forward reverse	ACCCGCTGAATTAAAGCATA TTTAAYTCTCTTTCYCAAAGTCC	28S rRNA	D1-D2	365 bp	MISeq	Smith, K.F. et al (2017) New Zealand Journal of Marine and Freshwater Research, 51, 555-576	Original forward primer from a former study, here applied for the first time to metabarcoding
Rhizaria	Cercozoa (general)	Cerc479F Cerc750R	forward reverse	TGTTGCAGTTAAAAAGCTCGT TGAATACTAGCACCCCAAC	18S rRNA	V4	250-300 bp	Pyrosequencing	Harder, C.B., et al. (2016) ISME J, 10, 2488-97.	
		Cer2F Cer1R	forward reverse	ATTCTCGCCCTATCAGCT ATACTAGCACCCCAACT	18S rRNA	V3-V4	600 bp	Pyrosequencing	Lentendu, G. et al. (2014) Molecular Ecology, 23, 3341-3355	
		S616F_Cerco	forward	TTAAAAAGCTCGTAGTTG						Semi nested protocol: the first two primers (S616F_Cerco and S616F_Eocer) are to be mixed in a proportion 80% to 20% respectively, in combination with the third (S963R_Cerco). The obtained products are to be reamplified using the first two in combination with the fourth (S947R_Cerco)
		S616F_Eocer	forward	TTAAAAAGCGCGTAGTTG	18S rRNA	V4	350 bp	Illumina MiSeq	Fiore-Donno, A.M. et al. (2018) Mol Ecol Resour, 18, 229-239.	
		S963R_Cerco S947R_Cerco	reverse reverse	CAACTTTCGTTCTTGATTTAA AAGAAGACATCCTTGGTG						
	Plasmodiophorida	1301f 1801r	forward reverse	GATTGAAGCTCTTTCTTGATCACTTC ACGGAAACCTTGTTCAGACTTC	18S rRNA	V7-V9	500 bp	Pyrosequencing	Bass, D., et al. (2018) Frontiers in Microbiology, 9	
	Foraminifera	s14F1 s15.3	forward reverse	AAGGGCACCAAGAACGC CCTATCACATAATCATGAAAG	18S rRNA	telix 37 and 37	150 bp	Illumina HiSeq	Pawlowski, J., et al., (2014) Molecular Ecology Methods 14, 1129-40	
	Chrysophyceae/Synurophyceae	Chryso_240 Chryso_651	forward reverse	GGAAACCAATGCGGGGCAAC CTATTTTGTCACAGTAAATGACGAG	18S rRNA	V2-V3	430 bp	Pyrosequencing	Lentendu, G., et al. (2014) Molecular Ecology, 23, 3341-55	Primers first developed for Sanger sequencing, here used for HTS. Semi-nested approach with a generic primer; specific primers are to be used in the second PCR
	Bacillariophyceae (diatoms)	Diat_rbcL_708F R3	forward reverse	AGGTGAAGTTAAAGTTTACTACTDAA CCTTCTAATTTACCAACAACGT	bcl (plastidial)		312 bp	Ion Torrent	Chonova, T. et al. (2019) Frontiers in Microbiology, 10	Primers first designed for Sanger sequencing in a previous study
		D512for D978rev	forward reverse	ATTCAGCTCCAATAGCG GACTACGATGGTATCTAATC	18S rRNA	V4	390-410 bp	Pyrosequencing	Zimmermann, J. et al. (2015) Molecular Ecology Resources, 15, 526-542.	Primer pair first developed for Sanger sequencing in a previous study
	Peronosporomycetes (=oomycetes)	ITS100 ITS300	forward forward	GGAAGGATCATTACCACA AGTATGYTGATCAGTG	ITS	ITS1 and ITS2 ITS2	650 bp 350 bp	Illumina MiSeq	Riit, T. et al., Mycokeys, 119-120.	To be used with generic primer ITS4
Discoba	Kinetoplastea	Kineto_80 Kineto_651	forward reverse	CATCAGAGCYAATCTGCCGC TTGGTCGCRCTTYYTTAGTCACAG	18S rRNA	V2-V3	700 bp	Pyrosequencing	Lentendu, G., et al. (2014) Molecular Ecology, 23, 3341-55	Primers first developed for Sanger sequencing, here used for HTS. Semi-nested approach with a generic primer; specific primers are to be used in the second PCR
Metamonadea	Diplomonadida (Fornicata)	DimA DimB	forward reverse	RGGGACRGGTGAAATAGGATG CAAATTGAGCCGCAGACTCC	18S rRNA	V4-V5	280 bp	HiSeq 2500	Cannon, M.V. et al. (2018) Microbiome, 6, 195	shorter 18S rRNA than most eukaryotes
Amoebozoa	Acanthamoeba	SRAca28 SFAca22	forward forward	CCAATTACAAGACTCTTRTCGAG CGGYGAGACTGCGGATGG	18S rRNA	V2	470-510 bp	Pyrosequencing	Fiore-Donno, A.M. (2016) Scientific Reports, 6, 19068.	Semi-nested protocol: first PCR with general primer S1 (Fiore Donno et al., 2008) and SRAca28 second PCR with S1 and SFAca22
	Dark-spored Myxomycetes	SR19Dark SF2Dark	forward forward	GTCCTCTAATTGTTACTCGAD GTTGATCCTGCCAGTAGTGT	18S rRNA	V2	550-590 bp	Pyrosequencing	Fiore-Donno, A.M. (2016) Scientific Reports, 6, 19068.	Semi-nested protocol: first PCR with general primer S1 (Fiore Donno et al., 2008) and SR19Dark second PCR with S1 and SF2Dark
Haptista	Haptophyta	Lhapto8 Lhapto20R_bis	forward reverse	CCATCTCATCCTGCGTGTCTCCGAC TCAGACTCCTGGTCCGTGTTTCT	28S rRNA	D1-D2	350 bp	Pyrosequencing	Bittner, L., et al. (2013) Molecular Ecology 22, 87-101	
		S28Flong PRYM01+7	forward reverse	GCGGTAATTCAGCTCCAA GATCAGTGAACATCCCTGG	18S rRNA	V4	450 bp	Pyrosequencing	Egge, E., et al. (2013) Plos one, 8, e74371

Suppl. Fig. 1. Coverage of all primer pairs used so far in high-throughput sequencing studies as identified with perfect matches of both primer pairs to the target sequences for the main protist lineages. This shows that (A) primers do not equally amplify higher taxonomic level lineages of protists and (B) amplicon lengths differ between supergroups and depending on primer sets. Note that Hacrobia represents the sum of Haptophytes, Cryptophytes and Centrohelids.

903

904

905

906