Measurements of soil protist richness and community composition are influenced by primer pair, annealing temperature, and bioinformatics choices

Applied and Environmental Microbiology
Mau, Rebecca; Hayer, Michaela; Purcell, Alicia; Geisen, S.A.; Hungate, Bruce A. et al
https://doi.org/10.1128/aem.00800-24

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Environmental Microbiology | Full-Length Text

Measurements of soil protist richness and community composition are influenced by primer pair, annealing temperature, and bioinformatics choices

Rebecca L. Mau,[1,2] Michaela Hayer,[1] Alicia M. Purcell,[3] Stefan Geisen,[4] Bruce A. Hungate,[1,2] Egbert Schwartz[1,2]

AUTHOR AFFILIATIONS See affiliation list on p. 14.

ABSTRACT Protists are a diverse and understudied group of microbial eukaryotic organisms especially in terrestrial environments. Advances in molecular methods are increasing our understanding of the distribution and functions of these creatures; however, there is a vast array of choices researchers make including barcoding genes, primer pairs, PCR settings, and bioinformatic options that can impact the outcome of protist community surveys. Here, we tested four commonly used primer pairs targeting the V4 and V9 regions of the 18S rRNA gene using different PCR annealing temperatures and processed the sequences with different bioinformatic parameters in 10 diverse soils to evaluate how primer pair, amplification parameters, and bioinformatic choices influence the composition and richness of protist and non-protist taxa using Illumina sequencing. Our results showed that annealing temperature influenced sequencing depth and protist taxon richness for most primer pairs, and that merging forward and reverse sequencing reads for the V4 primer pairs dramatically reduced the number of sequences and taxon richness of protists. The data sets of primers that targeted the same 18S rRNA gene region (e.g., V4 or V9) had similar protist community compositions; however, data sets from primers targeting the V4 18S rRNA gene region detected a greater number of protist taxa compared to those prepared with primers targeting the V9 18S rRNA region. There was limited overlap of protist taxa between data sets targeting the two different gene regions (80/549 taxa). Together, we show that laboratory and bioinformatic choices can substantially affect the results and conclusions about protist diversity and community composition using metabarcoding.

IMPORTANCE Ecosystem functioning is driven by the activity and interactions of the microbial community, in both aquatic and terrestrial environments. Protists are a group of highly diverse, mostly unicellular microbes whose identity and roles in terrestrial ecosystem ecology have been largely ignored until recently. This study highlights the importance of choices researchers make, such as primer pair, on the results and conclusions about protist diversity and community composition in soils. In order to better understand the roles protist taxa play in terrestrial ecosystems, biases in methodological and analytical choices should be understood and acknowledged.

KEYWORDS protist, microeukaryote, soil, 18S rRNA, small subunit rRNA

Protists are eukaryotic organisms that are not taxonomically classified as plants, animals, or fungi, which lumps together an extremely diverse and polyphyletic group of mostly unicellular organisms, such as flagellates, ciliates, amoebae, and eukaryotic algae (1, 2). Protists facilitate many important processes in terrestrial and aquatic ecosystems, from photosynthetic autotrophs that make up the base of aquatic food webs (3, 4) to heterotrophic bacterivores and fungivores that mediate nutrient

Editor Jeremy D. Semrau, University of Michigan, Ann Arbor, Michigan, USA

Address correspondence to Rebecca L. Mau, Rebecca.Mau@nau.edu.

The authors declare no conflict of interest.

See the funding table on p. 15.

Received 24 April 2024
Accepted 5 June 2024
Published 26 June 2024

cycling directly through regulating the size and composition of bacterial and fungal communities (5, 6) and indirectly by increasing available nutrients to plants and microbes through excretion (7, 8). Despite this, studies focused on protists in soil ecology continues to lag behind other microbial groups, such as bacteria and fungi (9).

One reason protist research may be lagging is that traditional techniques developed for studying these organisms are tedious, involving extracting protists from soil and then counting and identifying individuals under a microscope (10, 11). Critiques of these methods include the inefficient extraction of organisms, lack of morphological differentiating features among taxa, as well as bias in the groups that are amendable to growing under lab conditions (12, 13). Recent technological advances are increasing the number of studies using DNA-based amplicon sequencing to measure the composition and activities of protist communities in soil. These techniques increase our taxonomic resolution of protist organisms in a sample and decrease the time required for processing individual samples.

However, because the advent of molecular studies of protists is relatively recent, a consensus on how to characterize protist communities through nucleic acid-based analyses has not yet emerged (14). For instance, presently, there is no "universal" primer pair for identifying general protist community composition through amplicon sequencing even though most studies only use one primer pair to characterize protist composition. Many studies target the 18S rRNA gene for amplicon sequencing; however, multiple 18S rRNA regions have been proposed [e.g., V4, V9; (15, 16)] and numerous primer combinations within these different regions have been suggested (17). The Protist Ribosomal Reference (PR^2) primer database [app.pr2-primers.org; (18)] is an excellent interactive database that provides simulated taxonomic results for primers curated from the literature for general and group-specific assays (410 primers as of 8/7/2023). And while this is a great *in silico* tool, comparing primers experimentally will directly test their utility in resolving differences in soil protist communities. Despite this, few studies have reported on how primer pair influences protist diversity and composition (19, 20) and these have mostly focused on marine or freshwater environments.

In addition to primer choice, annealing temperature during preparation of the sequencing libraries as well as bioinformatic parameters applied to sequencing reads can play a significant role in the interpretation of richness, diversity, and composition (14, 21). The higher the annealing temperature, the more specific the binding will be to the target DNA and fewer non-specific target sequences will be amplified. However, annealing conditions can become too stringent such that target sequences could also be omitted from the library. Few researchers spend the effort to rigorously test the effect of annealing temperature on community metrics most likely due to time and money constraints, but it has been shown to be important (22). Similarly, bioinformatic parameter choices when processing Illumina next-generation sequencing data such as sequence read trimming length and merging the forward and reverse reads could have important consequences for determining protist community and composition (23). The Illumina quality score of a sequence read, which is the probability that the base call at a position along the sequencing read is correct, tends to decrease toward the end of sequencing reads. One can choose to truncate these reads to minimize low-quality base calls, but it may come at a cost of shorter reads which contain less genetic information. Advances in denoising algorithms, such as DADA2 or Deblur (24, 25), are helping to correct these sequencing errors and help to retain more genetic information. After trimming the sequence reads, or not, the forward and reverse sequencing reads are typically merged to create a single, long sequence which is used to identify taxa. However, some target gene regions are variable in length and, because of this, forward and reverse reads cannot be successfully merged together and are discarded for some sequences because there is not enough overlap between the two reads. This is true for the ITS2 region used to identify fungi (21), as well as the V4 region of the 18S rRNA gene used here to target protists.

In this study, we asked the following questions: (i) Does annealing temperature during sequence library preparation influence sequencing depth, taxon richness, or composition of protist and non-protist organisms; (ii) Does sequence trim length and/or forward and reverse read merging during the DADA2 denoising bioinformatic step influence sequencing depth, taxon richness, or composition of protist organisms; (iii) How does primer pair influence the composition, phylogenetic diversity, and relative abundance of protist taxa? We used 10 soils from temperate, boreal, tropical, arctic, and Antarctic ecosystems to address these questions.

MATERIALS AND METHODS

Soil sites and collection

Samples collected from 10 different ecosystems were used to compare the sequencing results generated with different 18S rRNA eukaryotic primers and bioinformatic parameters. In 2014, soil (0–10cm) was collected from the grassy interspaces of a mixed conifer forest, ponderosa pine forest, pinyon-juniper woodland, and high desert grassland, all part of the C. Hart Merriam Elevation gradient in northern Arizona, USA. Soil from a second grassland site was collected from the Hopland Research and Extension Center in northern California in 2021. Arctic soil from the Arctic LTER site at the Toolik Lake Field Station in Alaska and boreal soil from the SPRUCE (Spruce and Peatland Responses Under Changing Environments) experimental site in Minnesota were collected in 2017, as well as tropical soil from the Sabana Field Research Station in the Luquillo Experimental Forest in Puerto Rico. Soil and moss samples from Antarctica were collected near Palmer Station at the retreating Marr Ice Piedmont Glacier and on Litchfield Island, West Antarctic Peninsula, in 2019. All samples were stored at −80°C until DNA extraction.

DNA extraction and Illumina MiSeq sequencing

DNA was extracted from all samples (n = 3) using a Qiagen PowerSoil Pro (Qiagen LLC, Germantown, MD) kit and following the manufacturer's protocol. Briefly, approximately 0.5 g wet weight soil was extracted and re-suspended in 100 µL. DNA was quantified with a Qubit (Thermo Fisher Scientific, Hampton, NH) using the HS quantification kit and stored at −20°C until sequencing prep. Sample DNA concentrations used for sequence preparation were between 3 and 10 ng/µL.

Samples were prepared for sequencing using a two-step PCR method where the first PCR amplified the target sequence with primers modified with "universal tails (UT)" (UT-Forward: 5´ - ACCCAACTGAATGGAGC - 3´, UT-Reverse: 5´ - ACGCACTTGACTTGTCTT C - 3´), and the second PCR used the UT sequences to attach a unique 8 bp barcode in addition to the Illumina flow cell adapter tails (P5/P7). The initial PCRs were set up in 15 µL reactions, which contained 1× Phusion Green Hot Start II Master Mix (Thermo Fisher Scientific, Hampton, NH), 2 mM $MgCl_2$, 2 µL of template DNA, and 0.5 µM of one of four primer pairs, modified with the UT sequences: 1380F/1510R [(V9; (26), 1391F/EukBr (V9; (19, 26), 616*F/1132R (V4; (16), and TAReuk454FWD1/TAReukRev3 (V4; (19); Table 1). The PCR cycling conditions were as follows: 3 min at 98°C, followed by 35 cycles of 30 s at 98°C, 30 s of annealing (various temperatures, see Table 1), and a 30-s extension at 72°C, with a 2-min final extension cycle. Three different annealing temperatures were tested for the V9 primer pairs: 57°C, 61°C, and 69°C. Fifty-five degrees Celsius and 60°C were used as annealing temperatures for the V4 616*F/1132R primer pair, and 64.5°C and 69°C were used for primer pair TAReuk454FWD1/REV3. The second PCR was a 25 µL reaction containing 1× Phusion Green Hot Start II Master Mix, 2 µL of amplicons from the initial PCR, and 0.1 µM primers. The thermocycling conditions were 2 min at 98°C, followed by 10 cycles of 30 s at 98°C, 30 s at 60°C, and 30 s at 72°C, with a final extension of 5 min at 72°C. PCR products were purified using AMPure magnetic beads (Beckman Coulter, Brea, CA) and quantified using a BioTek plate reader and the Quant-iT DNA quantification

TABLE 1 Primer sequences and annealing temperatures used in sequencing library preparation

Primer name	Sequence (5′-3′)	Annealing temperature	Reference
1380F	CCCTGCCHTTTGTACACAC	57°C, 61°C, 69°C	(26)
1510R	CCTTCYGCAGGTTCACCTAC		
1391F	GTACACACCGCCCGTC	57°C, 61°C, 69°C	(19, 26)
EukBr	TGATCCTTCTGCAGGTTCACCTAC		
616*F	TTAAARVGYTCGTAGTYG	55°C, 60°C	(16)
1132R	CCGTCAATTHCTTYAART		
TAReukFWD1	CCAGCASCYGCGGTAATTCC	64.5°C, 69°C	(19)
TAReukREV3	ACTTTCGTTCTTGATYRA		

kit (Thermo Fisher Scientific, Hampton, NH). Samples from the two V9 primer pairs and three annealing temperatures were pooled together at equimolar concentrations and sequenced on an Illumina MiSeq instrument (Illumina, San Diego, CA) using a v.2 2 × 150 cycle kit. Samples produced with the two V4 primer pairs and two annealing temperatures were pooled together at equimolar concentrations and sequenced on an Illumina MiSeq instrument using a v.3 2 × 300 cycle kit.

Bioinformatics

Sequences were processed using the QIIME2 bioinformatics platform (27). Demultiplexed sequencing reads were imported into QIIME2, and primers were trimmed off using the cutadapt plugin and the "trim-paired" command (28). Sequence trim lengths were determined by looking at the quality scores of the V4 and V9 rRNA gene region sequencing data sets separately and selecting lengths where the quality scores started to decline. Sequences were trimmed to 90 bp or 120 bp for the V9 primers and reads were merged during the denoising step using the DADA2 plugin (25). For the V4 primers, reads were trimmed to 250 bp or 280 bp. The sequences were then either merged together with the DADA2 plugin using the "denoise-paired" command in QIIME2 or only the forward sequencing read was used for subsequent analyses ("denoise-single" command). Reads were then assigned taxonomy using the "feature-classifier classify-sklearn" command (29) and the PR^2 reference database [14.4.0; (30); with the percent identity parameter set to 0.9 to ensure high confidence in taxonomic assignments. Amplicon sequence variants (ASVs) that were assigned to the same taxonomy were then merged together using the QIIME2 "collapse" command at the species level (L9), resulting in 549 unique taxonomic assignments ranging from species level to kingdom (i.e., protist taxa).

Statistics

The influence of annealing temperature and the DADA2 denoising bioinformatic parameters on the composition of the microbial community was tested using the "adonis" permutational multivariate analysis of variation (PERMANOVA) plugin from QIIME2. For each primer pair, a dissimilarity matrix was created using the "jaccard" method, which calculated community differences based on the presence/absence of taxa. Using these matrices, the "adonis" command was run using the formula "DADA2*annealing_temperature" to test for significant differences of these two factors and their interaction. This analysis was run using data from the total sequenced community and from data where non-protistan taxa were filtered out. Non-protistan taxa were considered to be sequences that were assigned with 90% confidence to Archaea, Bacteria, Eukaryota;__, Streptophyta, Opisthokonta;_, Fungi, Mesomycetozoa, Metazoa, Eukaryota:plas, or Unassigned. Taxa associated with significant factors ("indicator species") for each primer pair were identified using the "multipatt" command in the "indicspecies" 1.7.12 R package (31).

To test the effect of annealing temperature and DADA2 denoising parameters on protist taxon richness, samples were rarefied to two different depths (678 and 1,454

sequences) on the data set after non-protist taxa were filtered out as described above. These rarefaction depths were chosen from species accumulation curves to maximize the richness in the data set while retaining the greatest number of samples for analyses (Fig. S1). Protist taxon richness was then calculated in QIIME2 using the "diversity alpha" command and the "observed_features" metric parameter. Analyses of variance (ANOVAs) and Tukey's *post hoc* analyses were performed in R to test for significant differences between primer pair, annealing temperature, sequence trim length, and merging or single read DADA2 parameters and their interactions on protist richness.

Once the optimized annealing temperature and DADA2 denoising parameter combination was identified for each primer pair, as determined by the combination that resulted in the greatest protist taxa richness and least number of non-protist and/or greatest number of protist indicator species, differences in the relative abundances of protist and non-protist taxa between the four primer pairs were tested with ANOVAs in R. Principle coordinate analyses were conducted on protist taxa communities for the four primers using Bray-Curtis and Jaccard dissimilarity matrices, and visualizations were created in QIIME2 using the "emperor" plugin (32). For phylogenetic analyses, a phylogenetic tree was created in QIIME2 using the "align-to-tree-mafft-fasttree" pipeline (33, 34) for primers targeting the same 18S rRNA gene region (e.g., V4 or V9). Sequences were first clustered into operational taxonomic units (OTUs) at 97% identity using the "VSEARCH" plugin in QIIME2 (35). OTUs were chosen over ASVs as a more conservative approach to estimate protist taxon richness (36). A visualization of the phylogenetic diversity detected with each primer pair was created with the Interactive Tree of Life v6 [iTOL; itol.embl.de; (37)]. Phylogenetic diversity differences between the primers were tested using an unweighted UniFrac (38) distance matrix and the "adonis" plugin in QIIME2 (39).

RESULTS

Influence of annealing temperature and bioinformatic parameters on sequencing depth

Higher annealing temperatures decreased the total number of sequences detected for all primer pairs (Table 2). The 616*F/1132R primer pair at 55°C resulted in approximately 3.2 million sequences, nearly three times more than the other primer pair and annealing temperature combinations (all approximately 1 million sequences). The number of total sequences that were retained after the DADA2 denoising bioinformatic steps was not largely different between the 90 bp and 120 bp trimming parameters for the 1380F/1510R primer pair, though increasing annealing temperature increased the percent of initial sequences retained in the data set from approximately 81% to 87.5%. A similar trend with annealing temperature was seen for primer pair 1391F/EukBr with increasing temperature increasing the percent of initial sequences retained; however, there was an effect of trimming length for this primer pair where the percent of retained sequences was greater for the 90 bp trimming parameter compared to the 120 bp parameter. A greater proportion of initial sequences was retained for both V4 primer pairs when the forward and reverse sequencing reads were not merged together and only the forward read was used. For primer pair 616*F/1132R, the merged reads retained a lower percentage of sequences (30.8%) compared to the single read sequences (60.3%). The trend was similar for TAReukFWD1/REV3: 47.9% of the initial sequences were retained when the forward and reverse reads were merged and 61% were retained when only the forward read was used.

The percent of sequences that passed all filtering criteria (e.g., trim length, merging, chimeras) and were taxonomically assigned to protist organisms ranged from 0.5% to 32.4% across all primer pair, annealing temperature, and bioinformatic parameter combinations (Table 2). The percent of sequences that passed all filtering criteria that were assigned to protist taxa were not drastically different between the 90 bp and 120 bp trimming parameters or annealing temperature for primer pair 1380F1510R; however, increasing the trim length from 90 bp to 120 bp considerably increased the

TABLE 2 The influence of annealing temperature and bioinformatic parameters on the number of total sequences and sequences assigned to protist taxa that were retained in data sets generated with four different primer sets[a]

	DADA2_250				DADA2_250_Read1				DADA2_280				DADA2_280_Read1			
	616*F/1132R		TAReukFWD1/REV3		616*F/1132R		TAReukFWD1/REV3		616*F/1132R		TAReukFWD1/REV3		616*F/1132R		TAReukFWD1/REV3	
	55°C	60°C	64.5°C	69°C	55°C	60°C	64.5°C	69°C	55°C	60°C	64.5°C	69°C	55°C	60°C	64.5°C	69°C
Initial number sequences	3200481	1321969	981294	945805	3200481	1321969	981294	945805	3200481	1321969	981294	945805	3200481	1321969	981294	945805
Sequences passed filter (%)	55.1	41.2	61.1	61.4	71.1	69.9	67.2	67.4	47.7	35.9	54.7	54.8	66.6	65.8	64.9	65.1
Merged sequences (%)	39.4	27.3	52.4	56.5	–	–	–	–	37.5	31.4	47.4	50.8	–	–	–	–
Non-chimeric sequences (%)	34.7	25.8	47.0	53.4	60.0	65.5	61.3	63.7	32.9	29.6	42.9	48.1	54.4	61.3	58.1	61.0
Sequences assigned protist taxa (%)	0.5	1.5	31.7	30.8	4.8	14.0	28.3	26.6	0.6	1.8	32.4	31.1	5.4	15.0	28.6	27.5
Number of protist taxa	39	21	327	315	301	315	361	322	49	43	325	312	295	316	365	331

	DADA2_120						DADA2_90					
	1380F/1510R			1391F/EukBr			1380F/1510R			1391F/EukBr		
	57°C	61°C	69°C	57°C	61°C	69°C	57°C	61°C	69°C	57°C	61°C	69°C
Initial number sequences	1135109	928535	778679	1122867	1032829	792780	1135109	928535	778679	1122867	1032829	792780
Sequences passed filter (%)	93.6	95.4	96.1	56.6	65.3	91.2	96.2	96.7	96.8	93.0	95.4	96.8
Merged sequences (%)	86.7	89.5	90.4	51.2	59.9	85.9	88.1	90.0	90.1	81.6	84.3	89.0
Non-chimeric sequences (%)	80.7	85.1	87.8	49.2	57.1	83.9	81.4	85.0	87.3	78.4	80.8	86.9
Sequences assigned protist taxa (%)	18.6	20.2	19.2	16.4	14.7	16.2	18.5	20.3	19.2	10.2	10.4	15.8
Number of protist taxa	217	219	203	180	189	206	206	215	200	179	189	206

[a]The percentages of sequences that passed filter, merged, and were non-chimeric are relative to the initial number of sequences, while the percentage of sequences assigned protist taxa are relative to the number of sequences that passed filter, merged, and were non-chimeric. Dashes indicate there are no data for this parameter.

percent of sequences assigned to protist taxa for primer pair 1391F/EukBr from ~10.3% at the two lower temperatures to ~15.6%. Merging the forward and reverse reads for primer pair 616*F/1132R dramatically decreased the number of sequences assigned to protist taxa to <2% regardless of annealing temperature. Single read sequences that were assigned to protists for the 616*F/1132R primer pair remained relatively low (<5.5% at 55°C and <15% at 60°C). A similar pattern was observed for the TAReukFWD1/REV3 primer pair, but the absolute values were not as small with an average of 31.5% of sequences assigned to protist taxa for the merged reads and 27.8% for the forward sequencing reads only.

Effect of annealing temperature and DADA2 denoising parameters on community composition and protist taxon richness

Both annealing temperature and DADA2 denoising parameters had a significant effect on the composition of all taxa sequenced and the protist community composition for primer pairs targeting the V4 18S rRNA region (616*F/1132R and TAReukFWD1/REV3; Table 3; Fig. S3). Both annealing temperature and DADA2 denoising parameters influenced the composition of all taxa sequenced for primer pair 1381F/EukBr, while only annealing temperature affected the composition of protist taxa (Table 3). Primer pair 1380F/1510R was the most stable in terms of community composition where annealing temperature was the only significant factor influencing the community composition of all taxa sequenced (Table 3).

Annealing temperature was significantly associated with 119 indicator species for primer pair 616*F/1132R. Of the 109 indicator species in data generated with an annealing temperature of 55°C, 97 were bacteria, 5 were archaea, 6 were fungi, and 1 was a plant mitochondrion, suggesting that this low annealing temperature is not optimal for targeting protist taxa. There were 10 indicator species associated with annealing temperature of 60°C, five being protist taxa. Sequencing results generated with single read DADA2 parameters trimmed to 250 bp and 280 bp had 291 and 296 indicator species associated with them, respectively, while the paired 280 bp and 250 bp trimmed reads had 52 and 6 indicator species, respectively. Eighty-eight percent of the indicator species for the paired reads were non-protist taxa (e.g., fungi), while the majority of the indicator species for the single read DADA2 parameters (~61%) were protist taxa, suggesting that merging reads together is not a successful strategy when targeting protist taxa using this primer pair.

Indicator species analysis found 73 taxa associated with annealing temperature for primer pair TAReukFWD1/REV3. For annealing temperature of 64.5°C, 33 of the 57 species were protist species (58%), whereas only 3 of the 16 indicator species were protist organisms for annealing temperature of 69°C (19%). There were 123 indicator species associated with DADA2 denoising parameters. The single read DADA2 parameters had almost twice the number of indicator species associated with them (58 taxa) compared to the paired reads (30.5 taxa). However, there was no obvious differentiation between the single read DADA2 parameter trimmed to 250 bp or 280 bp, suggesting that these two trim lengths are not an important factor influencing the composition of microbial taxa.

Indicator species analysis showed that 27 taxa were significantly associated with the DADA2 90 bp trimming parameter for primer pair 1391F/EukBr; none of which were

TABLE 3 Factors influencing the description of taxon composition of total taxa sequenced and protist taxa only[a]

	1380F/1510R		1391F/EukBr		616*F/1132R		TAREukFWD1/REV3	
	All taxa (*n* = 179)	Target taxa (*n* = 177)	All taxa (*n* = 179)	Target taxa (*n* = 179)	All taxa (*n* = 239)	Target taxa (*n* = 218)	All taxa (*n* = 239)	Target taxa (*n* = 239)
Annealing temp	**2.81** (**0.002**)	1.09 (0.314)	**4.34** (**0.001**)	**2.24** (**0.003**)	**10.21** (**0.001**)	**1.71** (**0.016**)	**5.17** (**0.001**)	**2.66** (**0.001**)
DADA2 parameters	0.15 (1.0)	0.16 (1.0)	**4.36** (**0.001**)	0.28 (1.0)	**10.20** (**0.001**)	**10.04** (**0.001**)	**2.45** (**0.001**)	**2.55** (**0.001**)
Temp x DADA2	0.03 (1.0)	0.009 (1.0)	0.33 (0.99)	0.02 (1.0)	0.59 (1.0)	0.78 (0.96)	0.21 (1.0)	0.16 (1.0)

[a]PERMANOVA *F*-statistics (*P*-value in parentheses) from Jaccard's distance matrices are presented. Bold text indicates a significant value at $P < 0.05$.

protist taxa, in contrast to the DADA2 120 bp trimmed data set. The highest annealing temperature (69°C) was associated with the most protist indicator taxa (14), whereas annealing temperatures of 57°C and 61°C were both only associated with two protist taxa. Therefore, amplicons prepared with an annealing temperature of 69°C and the paired reads trimmed to 120 bp during the DADA2 denoising bioinformatic step resulted in data with the highest number of protist indicator species.

For primer pair 1380F/1510R, indicator species analysis showed 26 indicator species were associated with the annealing temperature used in library construction. Nine of 10 taxa associated with annealing temperature of 57°C only were bacteria. There were four indicator species at annealing temperature of 61°C, one fungus and three protists, and two indicator species at 69°C. The remaining 10 taxa were associated with both 57°C and 61°C temperatures; four of them were protist taxa. We chose the data set from the 61°C annealing temperature in subsequent analyses because bacterial indicator species were prevalent at the 57°C temperature and there was not a significant difference in protist taxon richness between the two lowest annealing temperatures. Since the DADA2 denoising parameters were not significantly different, we used data from the 120 bp trimming parameters in the subsequent analysis of comparing community composition between the different primer sets to be consistent with the other V9 primer pair (1391F/EukBr) that did show a significant difference in results generated with these different trimming parameters.

The taxon richness of protist taxa increased with increasing sequencing depth, with an average taxon richness of 36.7 (± 3.0) for data rarefied to 678 sequences and 41.1 (± 4.0) for data rarefied to 1,454 (Fig. 1). Primer pair was significant at both rarefaction depths, with the TAReukFWD1/REV3 primer pair consistently resulting in the greatest protist richness, about 2.3 times greater than primer pair 616*F/1132R and 1.3 times greater than primer pair 1391F/EukBr. The interaction of primer and annealing temperature on protist taxon richness was also significant at both rarefaction depths, with increasing annealing temperature increasing protist taxon richness for primers 1391F/EukBr and 616*F/1132R, decreasing richness for primer pair TAReukFWD1/REV3, and having no effect on primer pair 1380F/1510R (Fig. 1). There was also a significant interaction between primer and merging or single read DADA2 parameter, where merging forward and reverse reads resulted in a dramatic decrease in protist taxon richness for primer pair 616*F/1132R and no difference for primer pair TAReukFWD1/REV3.

For subsequent analyses comparing the influence of primer pair on community composition, we chose one dataset from each primer pair with what we considered optimized annealing temperatures and denoising parameters (trim length and single read or merged reads). This decision was based on results from the indicator species analysis that resulted in the greatest number of protist taxa and least number of non-protist taxa as well as results from the protist taxon richness analysis. The annealing temperature and denoising parameter combinations for each primer pair were as follows: 61°C, paired 120 bp for primer pair 1380F/1510R; 69°C, paired 120 bp for primer pair 1391F/EukBr; 60°C, single read 280 bp for 616*F/1132R primer pair; 64.5°C, single read 280 bp for primer pair TAReukFWD1/REV3.

Effect of primer pair selection on taxon composition of protist and non-protist taxa

Primer pair had a significant effect on measurements of the relative abundance of protist and non-protist taxa (Fig. 2). TAReukFWD1/REV3 detected the highest average relative abundance of protists (31.7% ± 4%) and primer pair 616*F/1132R the lowest (18.1% ± 4.7%; $F = 2.9$, $P = 0.04$). The 616*F/1132R primers amplified a significantly larger proportion of bacteria and archaea than the other three primer pairs (0.38 ± 0.05; $F = 47.4$, $P < 0.0001$). The relative abundance of fungi was almost two times greater in data sets from the primers targeting the V9 region of the 18S rRNA gene (1380F/1510R and 1391F/EukBr) than the V4 primers (616*F/1132R and TAReukFWD1/REV3; $F = 11.6$, $P < 0.0001$). In

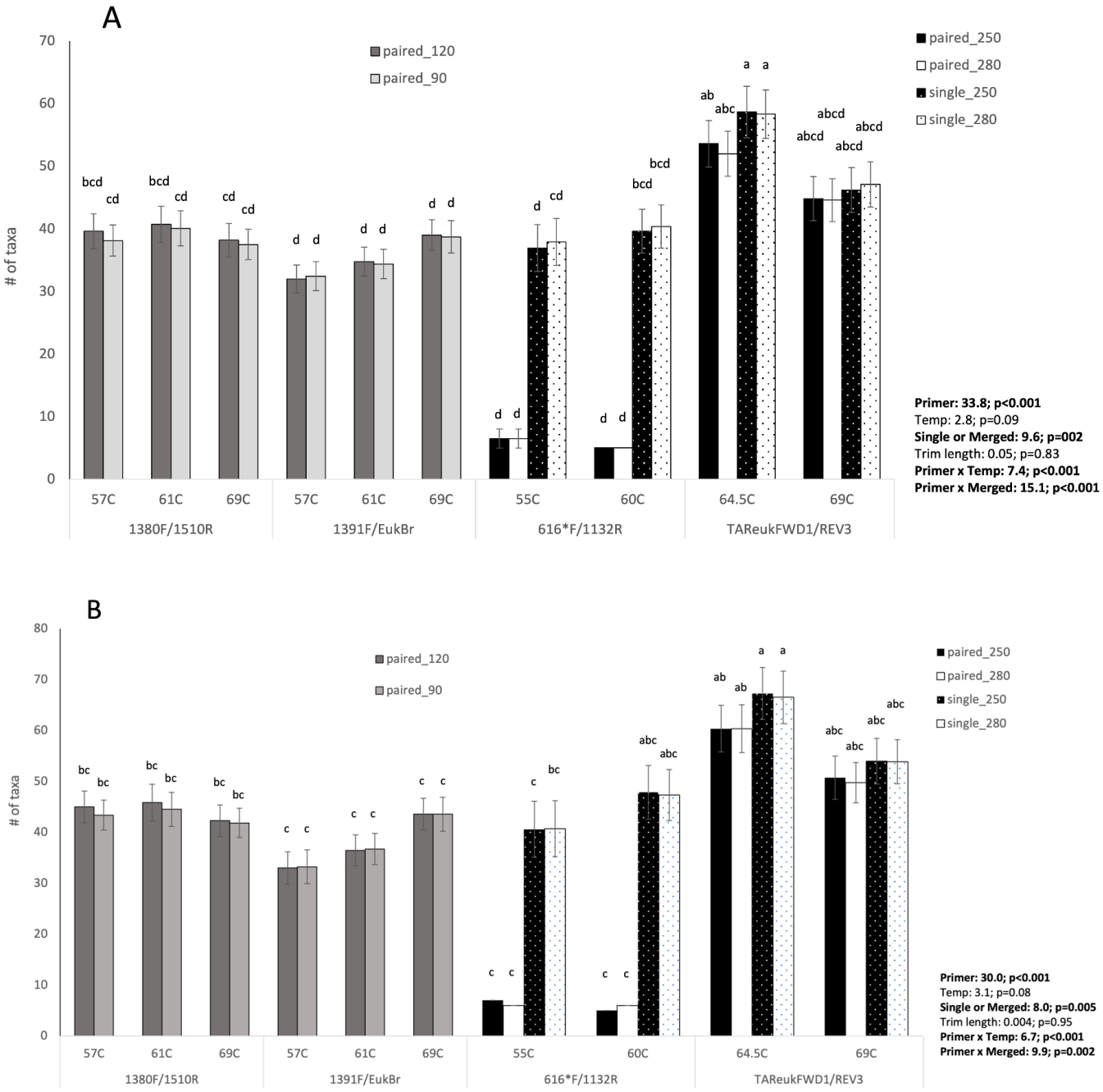
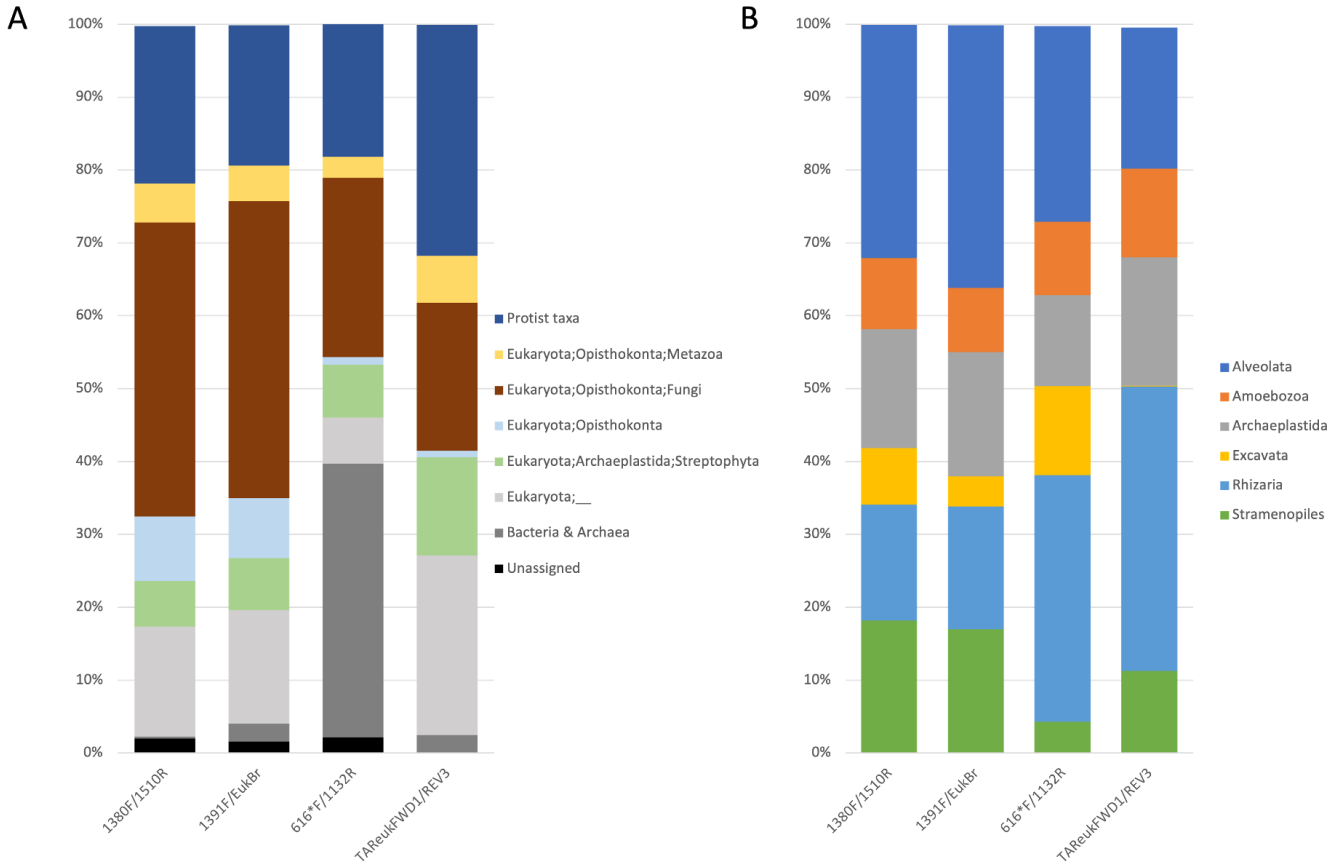


FIG 1 Protist taxa richness described by data sets rarefied at 678 (A) and 1,454 (B) sequences. The left panel shows richness detected within sequences produced with the V9 18S rRNA primers at three different annealing temperatures and two DADA2 trimming parameters, and the panel on the right shows richness of protists detected with the V4 18S rRNA primers for two annealing temperatures and four DADA2 trimming and merging parameters. *F*-statistic and *P*-values from ANOVAs comparing richness at each rarefication depth are reported next to each panel with bold text indicating significance at $P < 0.05$.

addition, the V9 primers detected a greater proportion of metazoan sequences than the V4 primers (~0.8 vs 0.095; $F = 3.1$, $P = 0.03$).

The composition of protist taxa was significantly influenced by the identity of the primer pair used to generate the sequencing libraries (Fig. 3), and this variation was greater between primer pairs than replicate soil samples. The relative abundance of Stramenopiles was greater in data sets produced with primers targeting the V9 18S rRNA region compared to the V4 rRNA region (Fig. 2; $F = 5.1$, $P = 0.003$). In contrast, the primers targeting the V4 18S rRNA region detected a greater relative abundance of Rhizaria than



C

Taxa	F-value	p-value
Bacteria & Archaea	**47.4**	**<0.001**
Archaeplastida	**12.7**	**<0.001**
Streptophyta	1.2	0.3
Eukaryota	**13**	**<0.001**
Opisthokonta	**31.1**	**<0.001**
Fungi	**11.6**	**<0.001**
Metazoa	1.4	0.233
Protist Taxa	**2.9**	**0.04**
Unassigned	**3.1**	**0.03**

D

Taxa	F-value	p-value
Alveolata	2.1	0.1
Amoebozoa	0.5	0.68
Archaeplastida	0.3	0.82
Excavata	**12.1**	**<0.001**
Rhizaria	**10.6**	**<0.001**
Stramenopiles	**5.1**	**0.003**

FIG 2 Average relative abundance of all taxa sequenced (A) and of protist taxa only (B) in sequencing data sets generated using four different primer pairs. The data presented here are from the optimized annealing temperature and DADA2 bioinformatic parameters data sets for each primer pair (n = 30): 61°C, paired 120 bp for primer pair 1380F/1510R; 69°C, paired 120 bp for primer pair 1391F/EukBr; 60°C, single read 280 bp for 616*F/1132R primer pair; 64.5°C, single read 280 bp for primer pair TAReukFWD1/REV3. Tables present ANOVA results comparing the relative abundances of taxa of all sequences (C) and protist taxa only (D) between the four primers, with taxa in bold indicating significant differences at $P < 0.05$.

the primers targeting the V9 18S rRNA region (Fig. 2; $F = 10.6$, $P < 0.001$). There was also a significant effect of primer pair identity on the abundance of Excavata (Fig. 2; $F = 12.1$, $P < 0.001$) in sequencing data sets, with primer pair 616*F/1132R detecting the greatest relative abundance and TAReukFWD1/REV3 detecting the lowest relative abundance of this group.

Overall, the V4 primers detected a greater number of protist OTUs than the V9 primers (2,173 vs 1,006; Fig. 4). The TAReukFWD1/REV3 primers detected the greatest

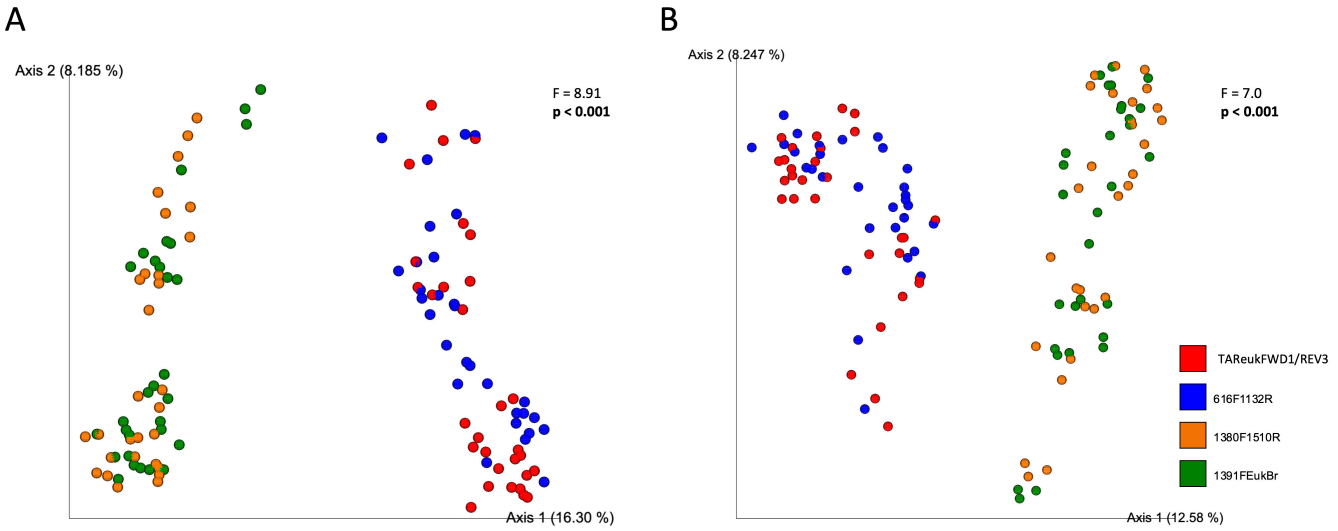


FIG 3 Principal coordinates analysis from Jaccard (A) and Bray-Curtis (B) dissimilarity matrices of protist taxa in sequences produced with four different primer pairs. Only the sequencing data generated with optimized annealing temperatures and DADA2 parameters of each primer pair are represented (61°C, paired 120 bp for primer pair 1380F/1510R; 69°C, paired 120 bp for primer pair 1391F/EukBr; 60°C, single read 280 bp for 616*F/1132R primer pair; 64.5°C, single read 280 bp for primer pair TAReukFWD1/REV3). *F*-statistics and *P*-values testing the effect of primer on protist composition are presented in each panel.

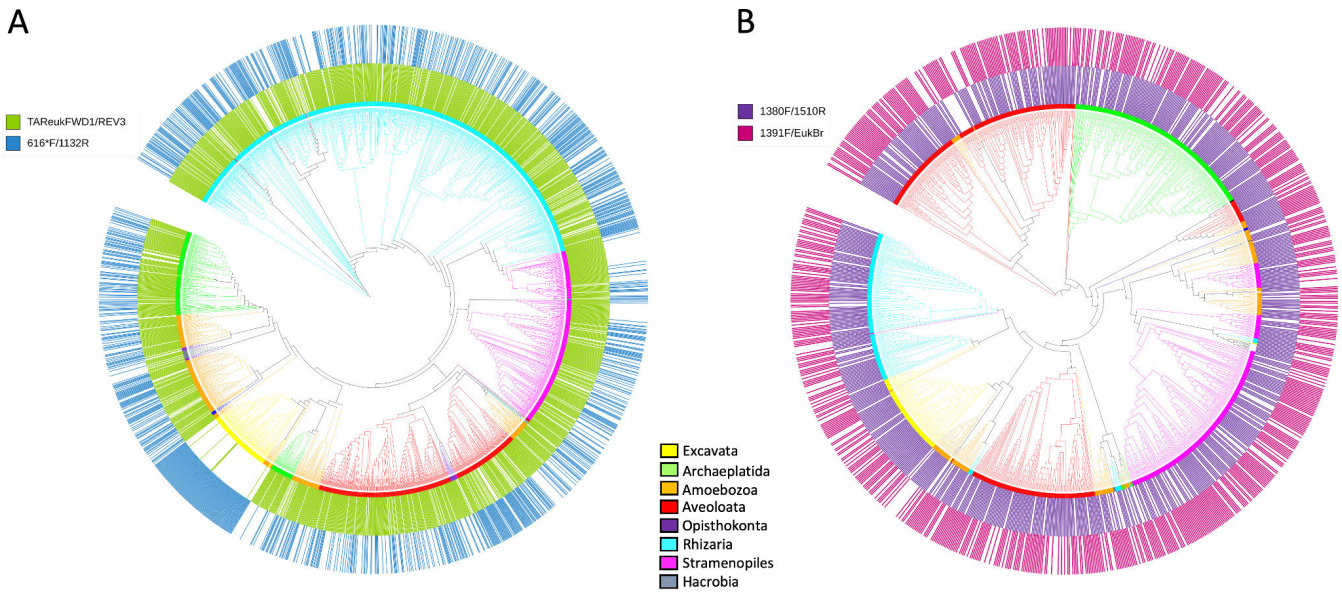
number of OTUs (1,799) followed by primer pair 616*F/1132R with 1,099 OTUs, 1380F/1510R with 849 OTUs, and finally 1391F/EukBr with 749 OTUs (Fig. 4). The TAReukFWD1/REV3 primers identified the most OTUs in all supergroups except Excavata, where the 616*F/1132R primers detected a much greater number (Fig. 4). The phylogenetic diversity described by the amplicons generated with the two different sets of V9 primers was not significantly different from each other ($F = 0.9$, $P = 0.58$; Fig. S3B), whereas there was a significant difference in measures of phylogenetic diversity described by the sequences generated with the different V4 primer combinations ($F = 4.8$, $P < 0.001$; Fig. S3A).

There were 549 unique taxonomic assignments in our data set, 419 of which had a taxonomic assignment to at least the genus level with 90% confidence (Table S1). The TAReukFWD1/REV3 primer pair detected the greatest number of taxa (365; 66.5% of all taxa) compared to the 616*F/1132R (316; 57.6% of taxa), 1380F/1510R (219; 39.9% of taxa), and 1391F/EukBr (206; 37.5% of taxa) primer pairs (Table S1). There were only 80 taxa that were shared between all four primer sets (14.6%; Fig. S4). Data sets generated with primer pairs targeting the V9 18S rRNA region shared 64 taxa (11.7%), while primers targeting the V4 region resulted in data sets that shared 160 taxa (29.1%). Primer pair TAReukFWD1/REV3 described the most unique taxa (90) of the primer pairs tested here and the 1391F/EukBr primer pair the least (20).

DISCUSSION

Three objectives of this study were to investigate how annealing temperature during library preparation, sequence trimming and merging parameters during the denoising bioinformatics step, and primer pair choice influenced sequencing data sets focused on protist community composition and richness in soil environments. Here, we show that primers targeting the V4 and V9 region of the 18S rRNA gene resulted in different inferences about the richness and composition of protist taxa in 10 globally distributed soils. In addition, the influence of annealing temperature and bioinformatic choices were specific to different primer pairs.

Annealing temperature during library preparation had a significant effect on the composition of protist and non-protist taxa for all primer pairs. The lower annealing temperatures tested here for primer pairs 1380F/1510R, 1391F/EukBr, and 616*F/1132R



C

	V9 primers (1006 OTUs)		V4 primers (2173 OTUs)	
	1380F/1510R	1391F/EukBr	616*F/1132R	TAReukFWD1/REV3
Alveolata	253	235	181	**322**
Amoebozoa	103	80	156	**230**
Apusozoa	2	2	2	**6**
Archaeplastida	140	122	81	**180**
Excavata	67	48	**122**	7
Hacrobia	1	3	5	**16**
Opisthokonta;Choanoflagellida	0	0	9	**18**
Protalveolata	**1**	0	0	**1**
Rhizaria	121	99	428	**726**
Stramenopiles	161	160	115	**293**
Total OTUs	849	749	1099	1799

FIG 4 Phylogenetic distribution of OTUs clustered at 97% identity in sequencing data sets produced with primers targeting the V4 (A) and V9 (B) 18S rRNA regions. Clade branches and the innermost ring is colored by the identity of protist supergroups. The outer rings represent the presence of an OTU with colors indicating if the OTU is detected by a specific primer pair (blue, 616*F/1132R; green, TAReukFWD1/REV3; purple, 1380F/1510R; pink, 1391F/EukBr). Panel C presents number of OTUs for each supergroup for each primer pair, with the greatest number of OTUs presented in bold text.

had a higher abundance on non-protistan taxa (e.g., bacteria, fungi) than higher annealing temperatures. This can have important consequences for interpreting and comparing sequencing data sets even when the same primer sets are used. For example, a search of the literature revealed that annealing temperatures used to prepare sequencing libraries using primer pair 1380F/1510R ranged from 50°C (40) to 68°C (41) and annealing temperatures for primer pair TAReukFWD1/REV3 ranged from 47°C (42, 43) to 60°C (44). Our results suggest that even though these different studies used the same primers to target protist communities, they cannot be directly compared because protist composition is significantly influenced by annealing temperature.

Bioinformatic parameters can also play a large role in the interpretation of sequencing data sets (21, 45). We observed the largest effects of the bioinformatics approach

between the merged and non-merged sequence data sets for the V4 18S rRNA primers where merging the sequencing reads resulted in a decrease in sequences. One reason that forward and reverse reads may not be successfully merged depends on the sequencing quality scores, which tend to be lower for the reverse reads and toward the end of either read in long sequencing kits (e.g., 2 × 300 bp). And this is indeed the case here; however, there was no significant difference in the quality scores at 260 bp or 280 bp between the two V4 18S rRNA primers, meaning that the significant differences between merging the 616*F/1132R and TAReukFWD1/REV3 is most likely not from low quality scores preventing merging. A second reason that forward and reverse reads may not be successfully merged is that the amplicon is larger than the sequencing platform allows for and there is not enough overlap between the forward and reverse reads to confidently merge them together to create a single, long read. This is most likely the case here as the average amplicon length for the 616*F/1132R primers is right at the edge of what a MiSeq 2 × 300 bp kit can sequence (18). The average amplicon length for the TAReukFWD1/REV3 is an average of 118 bp smaller, allowing for a higher proportion of the forward and reverse reads to be successfully merged with only minimal loss of taxa for this primer pair (18). Here, we showed that only using the forward read from an Illumina sequencing data set resulted in greater sequencing depth, and thus taxon richness, and was particularly pronounced for data sets generated with primer pair 616*F/1132R. Although more genetic information, and potentially more specific taxonomic classification, can be captured when merging forward and reverse reads together to create a longer sequence read, results will be biased toward organisms with shorter (<600 bp) amplicon lengths. This limitation can be overcome by using a different sequencing platform, such as Oxford Nanopore Technologies or Pacific Biosciences (Pacbio), which allows for longer sequencing reads. However, these platforms tend to be more expensive per basepair.

Amplicons produced with the TAReukFWD1/REV3 primers and an annealing temperature of 64.5°C without merging the forward and reverse reads together consistently resulted in the highest protist richness and detected the greatest number of OTUs among all primer sets investigated in this study. This is in contrast with a number of aquatic protist surveys comparing data sets targeting V4 and V9 18S rRNA gene regions, where V9 primers consistently detected more OTUs than V4 primer pairs (20, 46–48). Our TAReukFWD1/REV3 data set also detected the largest relative abundance of protist taxa (~32%) in our optimized sequencing library, meaning that the number of non-protistan sequences (e.g., derived from bacteria or fungi) that would be filtered out of a data set intended to describe protist communities is minimized with the use of this primer pair. While the amount of data that are assigned to non-protist organisms, or in other words is "filtered out," is not often reported in studies surveying protist communities, we would argue that it is an important metric when choosing a primer pair. This is because when using primers that result in a higher proportion of reads that are not "filtered out," more samples can be multiplexed on a single Illumina sequencing run for a desired sequencing depth, potentially reducing sequencing costs for researchers. Of the 549 protist taxa that were assigned in our whole data set, the TAReukFWD1/REV3 primer pair identified 365 (66.5%), representing the most in each supergroup with the exception of Excavata. In general, this is the best primer pair of the ones tested in this study for protist richness surveys in soil environments with the caveat that this primer pair did not perform well detecting protists in the Excavata supergroup.

The Excavata supergroup is made up of six phyla divided into two groups, the Discoba and Metamonada kingdoms (1). Many groups in the Discoba kingdom are known bacterial predators, such as Heterolobosea (1), and as such, they may play a vital role in soil food webs. The 616*F/1132R and 1380F/1510R primer pairs detected higher relative abundances of Excavata compared to the 1391F/EukBr and TAReukFWD1/REV3 primers. Similarly, bias against Excavata in data sets targeting the V4 compared to the V9 18S rRNA gene region has been reported in aquatic systems as well (20, 46). It is unclear how functionally important this supergroup is in soils, as they made up the smallest

average relative abundance of the protist community in the 10 soils studied here, though this was to be expected given the number of primer mismatches and generally longer amplicon lengths of this group (18). Three global assessments of protist taxa did not describe taxa from this group in their studies (4, 49, 50), while one other global survey of 52 soils did report them (51). While more studies are needed to assess the functional significance of this group in terrestrial environments, it is clear that many "universal" protist primer pairs should not be used to advance these investigations.

The two primer pairs targeting the V9 region of the 18S rRNA gene performed equally well when annealing temperature was optimized for the 1391F/EukBr primer pair. The richness and composition of protist taxa described by these two primers were similar, with no statistical difference in the composition or phylogeny. These V9 primers did target a larger proportion of fungi than the V4 primer pairs, so increasing sample sequencing depth in soils with high abundance of fungi may be required. Of the four primer pairs tested here using an Illumina MiSeq sequencing platform, the data set prepared with the TAReukFWD1/REV3 primer pair where the forward and reverse reads were not merged during the denoising step resulted in the greatest soil protist richness. There were no stark differences in the protist community composition or richness between the two primer pairs targeting the V9 18S rRNA region, though the data set prepared with the 1391F/EukBr primer pair was sensitive to annealing temperature while the 1380F/1510R primer pair was the most robust against differences in annealing temperature. Using a combination of the TAReukFWD1/REV3 and 1380F/1510R primer pairs, 87% of the protist taxa detected in this study would be included in the analysis (479/549 taxa).

In summary, we recommend more attention be given to annealing temperature during sequencing library preparation if maximizing protist sequences is the goal of the study. We also conclude that using a combination of primer pairs targeting the V4 and V9 regions of the 18S rRNA gene will maximize the richness and phylogenetic diversity of soil protist taxa surveys, a result that is consistent with protist surveys in aquatic systems (20, 47). In contrast to surveys of aquatic protist communities though, here, we find that data sets using primers targeting the V4 18S rRNA gene region (TAReukFWD1/TAReukREV3) result in the highest richness of protists in soil ecosystems.

ACKNOWLEDGMENTS

We would like to thank three anonymous reviewers for their constructive criticism to help improve the manuscript.

This project was funded by the Department of Energy grant IDs DE-AC52-07NA27344, DE-SC0020172, and DE-SC0023126.

AUTHOR AFFILIATIONS

[1]Center for Ecosystem Science and Society (Ecoss), Northern Arizona University, Flagstaff, Arizona, USA
[2]Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona, USA
[3]Department of Biological Sciences, Texas Tech University, Lubbock, Texas, USA
[4]Laboratory of Nematology, Wageningen University & Research, Wageningen, the Netherlands

AUTHOR ORCIDs

Rebecca L. Mau http://orcid.org/0009-0006-7833-8863
Bruce A. Hungate http://orcid.org/0000-0002-7337-1887

FUNDING

Funder	Grant(s)	Author(s)
U.S. Department of Energy (DOE)	DE-AC52-07NA27344	Bruce A. Hungate Egbert Schwartz
U.S. Department of Energy (DOE)	DE-SC0020172	Bruce A. Hungate Egbert Schwartz
U.S. Department of Energy (DOE)	DE-SC0023126	Bruce A. Hungate Egbert Schwartz

AUTHOR CONTRIBUTIONS

Rebecca L. Mau, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft | Michaela Hayer, Conceptualization, Writing – review and editing | Alicia M. Purcell, Investigation, Writing – review and editing | Stefan Geisen, Methodology, Writing – review and editing | Bruce A. Hungate, Funding acquisition, Writing – review and editing | Egbert Schwartz, Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review and editing

DATA AVAILABILITY

Raw sequencing data have been uploaded to NCBI's Sequence Read Archive (SRA) under BioProject number PRJNA1013338.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (AEM00800-24-s0001.docx). Figures S1 to S4; Table S1.

REFERENCES

1. Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. J Eukaryot Microbiol 66:4–119. https://doi.org/10.1111/jeu.12691
2. Geisen S, Lara E, Mitchell E. 2023. Contemporary issues, current best practice and ways forward in soil protist ecology. Mol Ecol Resour 23:1477–1487. https://doi.org/10.1111/1755-0998.13819
3. Gaedke U. 2009. Trophic dynamics in aquatic ecosystems, p 499–504. In Encyclopedia of inland waters. Elsevier Inc.
4. Xiong W, Jousset A, Li R, Delgado-Baquerizo M, Bahram M, Logares R, Wilden B, de Groot GA, Amacker N, Kowalchuk GA, Shen Q, Geisen S. 2021. A global overview of the trophic structure within microbiomes across ecosystems. Environ Int 151:106438. https://doi.org/10.1016/j.envint.2021.106438
5. Glücksman E, Bell T, Griffiths RI, Bass D. 2010. Closely related protist strains have different grazing impacts on natural bacterial communities. Environ Microbiol 12:3105–3113. https://doi.org/10.1111/j.1462-2920.2010.02283.x
6. Hünninghaus M, Koller R, Kramer S, Marhan S, Kandeler E, Bonkowski M. 2017. Changes in bacterial community composition and soil respiration indicate rapid successions of protist grazers during mineralization of maize crop residues. Pedobiologia 62:1–8. https://doi.org/10.1016/j.pedobi.2017.03.002
7. Clarholm M. 1985. Interactions of bacteria, protozoa and plants leading to mineralization of soil nitrogen. Soil Biol Biochem 17:181–187. https://doi.org/10.1016/0038-0717(85)90113-0
8. Clarholm M. 1989. Effects of plant-bacterial-amoebal interactions on plant uptake of nitrogen under field conditions. Biol Fert Soils 8:373–378. https://doi.org/10.1007/BF00263171
9. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. 2009. Protists are microbes too: a perspective. ISME J 3:4–12. https://doi.org/10.1038/ismej.2008.101
10. Coleman DC, Blair JM, Elliot ET, Wall DH. 1999. Soil Invertebrates, p 349–377. In Robertson GP (ed), Standard soil methods for long-term ecological research. Oxford University Press.
11. Adl SM, Coleman DC. 2005. Dynamics of soil protozoa using a direct count method. Biol Fertil Soils 42:168–171. https://doi.org/10.1007/s00374-005-0009-x
12. Elliott ET, Coleman DC. 1977. Soil protozoan dynamics in a shortgrass prairie. Soil Biol Biochem 9:113–118. https://doi.org/10.1016/0038-0717(77)90046-3
13. Foissner W. 1999. Soil protozoa as Bioindicators: Pros and cons, methods, diversity, representative examples. Agriculture, Ecosystems & Environment 74:95–112. https://doi.org/10.1016/S0167-8809(99)00032-8
14. Geisen S, Vaulot D, Mahé F, Lara E, de Vargas C, Bass D. 2019. A user guide to environmental protistology: primers, metabarcoding, sequencing, and analyses. bioRxiv. https://doi.org/10.1101/850610
15. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C. 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. PLoS One 9:e87624. https://doi.org/10.1371/journal.pone.0087624
16. Hugerth LW, Muller EEL, Hu YOO, Lebrun LAM, Roume H, Lundin D, Wilmes P, Andersson AF. 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
17. Adl SM, Habura A, Eglit Y. 2014. Amplification primers of SSU rDNA for soil protists. Soil Biol Biochem 69:328–342. https://doi.org/10.1016/j.soilbio.2013.10.024

18. Vaulot D, Geisen S, Mahé F, Bass D. 2022. Pr2-primers: an 18S rRNA primer database for protists. Mol Ecol Resour 22:168–179. https://doi.org/10.1111/1755-0998.13465
19. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, Richards TA. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19 Suppl 1:21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x
20. Choi J, Park JS. 2020. Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform. Sci Rep 10:6519. https://doi.org/10.1038/s41598-020-63561-z
21. Pauvert C, Buée M, Laval V, Edel-Hermann V, Fauchery L, Gautier A, Lesur I, Vallance J, Vacher C. 2019. Bioinformatics matters: the accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. Fungal Ecol 41:23–33. https://doi.org/10.1016/j.funeco.2019.03.005
22. Schmidt PA, Bálint M, Greshake B, Bandow C, Römbke J, Schmitt I. 2013. Illumina metabarcoding of a soil fungal community. Soil Biol Biochem 65:128–132. https://doi.org/10.1016/j.soilbio.2013.05.014
23. Mohsen A, Park J, Chen YA, Kawashima H, Mizuguchi K. 2019. Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks. BMC Bioinformatics 20:581. https://doi.org/10.1186/s12859-019-3187-5
24. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191-16. https://doi.org/10.1128/mSystems.00191-16
25. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869
26. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. 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
27. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9
28. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j 17:10. https://doi.org/10.14806/ej.17.1.200
29. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90. https://doi.org/10.1186/s40168-018-0470-z
30. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, et al. 2013. The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41:D597–D604. https://doi.org/10.1093/nar/gks1160
31. De Cáceres M, Legendre P, Moretti M. 2010. Improving indicator species analysis by combining groups of sites. Oikos 119:1674–1684. https://doi.org/10.1111/j.1600-0706.2010.18334.x
32. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16. https://doi.org/10.1186/2047-217X-2-16
33. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
34. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490. https://doi.org/10.1371/journal.pone.0009490
35. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584. https://doi.org/10.7717/peerj.2584
36. Caron DA, Hu SK. 2019. Are we overestimating protistan diversity in nature? Trends Microbiol 27:197–205. https://doi.org/10.1016/j.tim.2018.10.009
37. Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301
38. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. 2011. UniFrac: an effective distance metric for microbial community comparison. ISME J 5:169–172. https://doi.org/10.1038/ismej.2010.133
39. Anderson MJ. 2001. A new method for non - parametric multivariate analysis of variance. Austral Ecology 26:32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x
40. Liu M, Xue Y, Yang J. 2019. Rare plankton subcommunities are far more affected by DNA extraction kits than abundant plankton. Front Microbiol 10:454. https://doi.org/10.3389/fmicb.2019.00454
41. Ma L, Xie Y, Han Z, Giesy JP, Zhang X. 2017. Responses of earthworms and microbial communities in their guts to triclosan. Chemosphere 168:1194–1202. https://doi.org/10.1016/j.chemosphere.2016.10.079
42. Bystrianský L, Hujslová M, Hršelová H, Řezáčová V, Němcová L, Šimsová J, Gryndlerová H, Kofroňová O, Benada O, Gryndler M. 2019. Observations on two microbial life strategies in soil: planktonic and biofilm-forming microorganisms are separable. Soil Biol Biochem 136:107535. https://doi.org/10.1016/j.soilbio.2019.107535
43. Mahé F, Mayor J, Bunge J, Chi J, Siemensmeyer T, Stoeck T, Wahl B, Paprotka T, Filker S, Dunthorn M. 2015. Comparing high-throughput platforms for sequencing the V4 region of SSU-rDNA in environmental microbial eukaryotic diversity surveys. J Eukaryot Microbiol 62:338–345. https://doi.org/10.1111/jeu.12187
44. Schulz G, Schneider D, Brinkmann N, Edy N, Daniel R, Polle A, Scheu S, Krashevska V. 2019. Changes in trophic groups of protists with conversion of rainforest into rubber and oil palm plantations. Front Microbiol 10:240. https://doi.org/10.3389/fmicb.2019.00240
45. O'Sullivan DM, Doyle RM, Temisak S, Redshaw N, Whale AS, Logan G, Huang J, Fischer N, Amos GCA, Preston MD, et al. 2021. An inter-laboratory study to investigate the impact of the bioinformatics component on microbiome analysis using mock communities. Sci Rep 11:10590. https://doi.org/10.1038/s41598-021-89881-2
46. Okamoto N, Keeling PJ, Leander BS, Tai V. 2022. Microbial communities in sandy beaches from the three domains of life differ by microhabitat and Intertidal location. Mol Ecol 31:3210–3227. https://doi.org/10.1111/mec.16453
47. Maritz JM, Rogers KH, Rock TM, Liu N, Joseph S, Land KM, Carlton JM. 2017. An 18S rRNA workflow for characterizing protists in sewage, with a focus on zoonotic trichomonads. Microb Ecol 74:923–936. https://doi.org/10.1007/s00248-017-0996-9
48. Tragin M, Lopes dos Santos A, Christen R, Vaulot D. 2016. Diversity and ecology of green microalgae in marine systems: an overview based on 18S rRNA gene sequences. pip 3:141–154. https://doi.org/10.1127/pip/2016/0059
49. Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, Knight R, Fierer N. 2013. Global biogeography of highly diverse protistan communities in soil. ISME J 7:652–659. https://doi.org/10.1038/ismej.2012.147
50. Oliverio AM, Geisen S, Delgado-Baquerizo M, Maestre FT, Turner BL, Fierer N. 2020. The global-scale distributions of soil protists and their contributions to belowground systems. Sci Adv 6:eaax8787. https://doi.org/10.1126/sciadv.aax8787
51. Ramirez KS, Leff JW, Barberán A, Bates ST, Betley J, Crowther TW, Kelly EF, Oldfield EE, Shaw EA, Steenbock C, Bradford MA, Wall DH, Fierer N. 2014. Biogeographic patterns in below-ground diversity in New York city's central park are similar to those observed globally. Proc Biol Sci 281:20141988. https://doi.org/10.1098/rspb.2014.1988