Metabarcoding Approaches for Soil Eukaryotes, Protists, and Microfauna

Microbial Environmental Genomics (MEG)
Lentendu, Guillaume; Lara, Enrique; Geisen, Stefan
https://doi.org/10.1007/978-1-0716-2871-3_1

Chapter 1

Metabarcoding Approaches for Soil Eukaryotes, Protists, and Microfauna

Guillaume Lentendu, Enrique Lara, and Stefan Geisen

Abstract

There have been major developments in the molecular characterization of soil protist and micrometazoan diversity, leading to a better understanding of these minute soil eukaryotes. Like in all newly developing research fields, several approaches are currently used in parallel to study these organisms. Here, we synthesize these various approaches and propose a best practice manual that should help researchers to efficiently target soil eukaryotic diversity as a whole. We cover the whole working pipeline, ranging from sampling to nucleic acids extraction to bioinformatic processing and sequence identification. Synchronous approaches to molecularly survey microbial-sized eukaryotes and other soil biodiversity groups are needed in order to provide a cumulative knowledge of soil biodiversity, as here shown for the soil eukaryome. This will be crucial in understanding the important ecosystem functions provided by soil biodiversity.

Key words Metabarcoding, eDNA, Soil, Protists, Nematodes, Microfauna, Arthropodes, Bioinformatics, ASV

1 Introduction

Soils are the most diverse systems on the planet, and this diversity is predominantly microbial. Estimated species numbers of these microbial bacteria, archaea, fungi, and protists are in the tens of millions [1, 2]. Soil microbes drive major ecosystem functions, such as controlling the cycling of carbon and other elements through anabolic and catabolic activities. Plants also would hardly grow without their mutualistic microbes and those that facilitate plant nutrient uptake. While we have gained major insights on species diversity and biogeography of bacteria and fungi [3–6], this knowledge is still limited for other microbial eukaryotes. This bias has often been attributed to methodological constraints that have limited the possibilities to reliably study nonfungal soil eukaryotes (i.e., protists, nematodes, and microarthropods), the so-called soil eukaryome. With methodological issues increasingly being solved,

Francis Martin and Stephane Uroz (eds.), *Microbial Environmental Genomics (MEG)*,
Methods in Molecular Biology, vol. 2605, https://doi.org/10.1007/978-1-0716-2871-3_1,

we have obtained new information on this eukaryome. For example, we uncovered the unexpected richness of species and operational taxonomic unit (OTU) of soil protists. This includes studies revealing the presence of millions of protist species in tropical regions, including a previously unknown diversity of animal parasites [7] and studies that showed that the global distribution of soil protists [8–10]. In addition, these experimental approaches revealed that protists are the most responsive microbiome community to anthropogenic changes [11] or are predictors of plant health [12, 13]. These latter examples emphasized the need to include protists in soil microbiome analyses. Here, we describe easy-to-use approaches to survey protists for the nonexperts.

Protists have long been ignored in soil microbiome studies due to their paraphyletic diversity [14]. This feature precluded the design of a targeted PCR approach to survey environmental protist diversity only. In-depth description of the entire protistan communities was therefore not possible by using sequencing approaches based on cloning and Sanger technology due to the cosequencing of fungi. Protistan taxon-specific approaches were often used at that time (as also highlighted in the previous version of this chapter [15]). Now, however, most high-throughput studies commonly use primer pairs that target the whole domain Eukarya—and therefore the entire protistan diversity—as broadly as possible [16, 17]. This was allowed by the increase in sequencing depth by orders of magnitude in the last 15 years. We propose here this approach in order to uncover the broadest possible range of soil eukaryotes.

However, soil biodiversity is not only represented by unicellular organisms. Many animal species inhabit soils, with nematodes being the most abundant and diverse group. The field of soil nematology has developed over the last century, with recent studies showing the immense abundance and deterministic community-level global biogeography of these nematodes [18, 19]. Similar to protists, nematodes include many functional groups including bacterivores, fungivores, omnivores, predators, and, most notoriously, plant feeders. The latter have been of major interest as damage caused by these root-penetrating organisms is estimated in the billion euros for the crop and vegetable productions globally. Due to this functional diversity and well-established morphology-based methods, nematodes have gained considerable interest in basic and applied researches. Yet, deep taxonomic insights into nematode communities require profound expertise and time and therefore are expensive [20]. High-throughput sequencing approaches have thus long been proposed [21, 22] and are now gaining momentum [20, 23]. The same principles and constraints hold for soil microarthropods which are still even more studied based on morphological features, but molecular tools are emerging. Metabarcoding of terrestrial microarthropods has been validated against mixture of

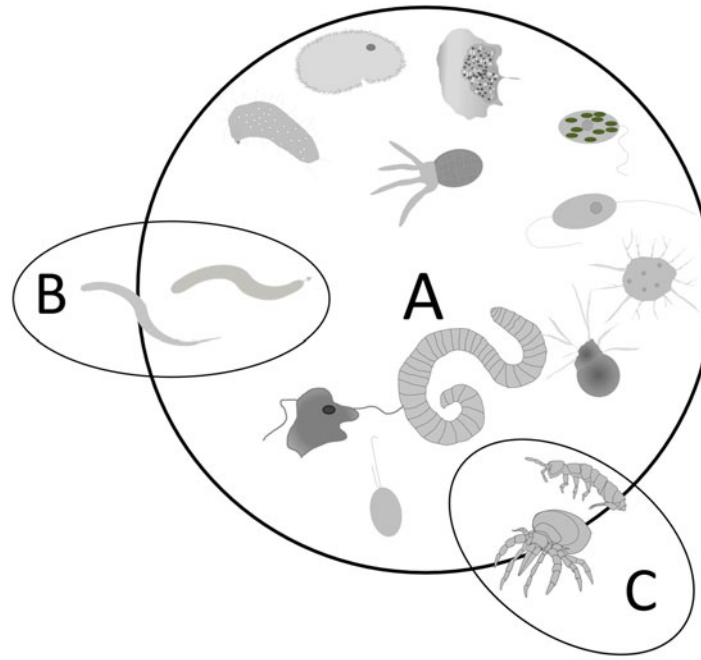

Fig. 1 Schematic representation of soil eukaryotic metabarcoding. (**a**) Metabarcoding of all eukaryotes after direct DNA extraction using a eukaryote-wide ("universal") primer pair. (**b**) Metabarcoding of nematodes after nematode isolation as a representative example of group-focused approaches using the same eukaryote-wide primer. (**c**) Metabarcoding of microarthropods after extracellular DNA extraction as a representative of group-focused approaches using a specific primer pair. The taxonomic resolution increases for the given organism groups, here shown for nematodes and microarthropods, while the taxonomic coverage is more restricted than with the universal barcoding approach

morphologically identified organisms [24, 25] to then be more widely applied to soil environmental DNA for biodiversity surveys [2, 26, 27]. We here use nematodes and microarthropods as example groups for targeted molecular sequencing. We also highlight that similar approaches are envisionable to study other soil animal groups such as rotifers [28] or specific protist groups (e.g., *chrysophytes, kinetoplastids, microalgae, Cercozoa* [29], and *ciliates* [30]).

We here focus on metabarcoding-based soil eukaryome analyses including DNA extraction, amplicon preparation, sequencing, and taxonomic identification (Figs. 1 and 2). We acknowledge that increasing read output with sequencing platforms will slowly shift the field of eukaryome studies toward PCR-free metagenomic and metatranscriptomic approaches [31, 32]. As the number of characterized sequences of eukaryotic genes and genomes is still limited in the omics datasets, metabarcoding approaches will remain the method of choice to study soil eukaryotic biodiversity in multiple samples for the next several years.

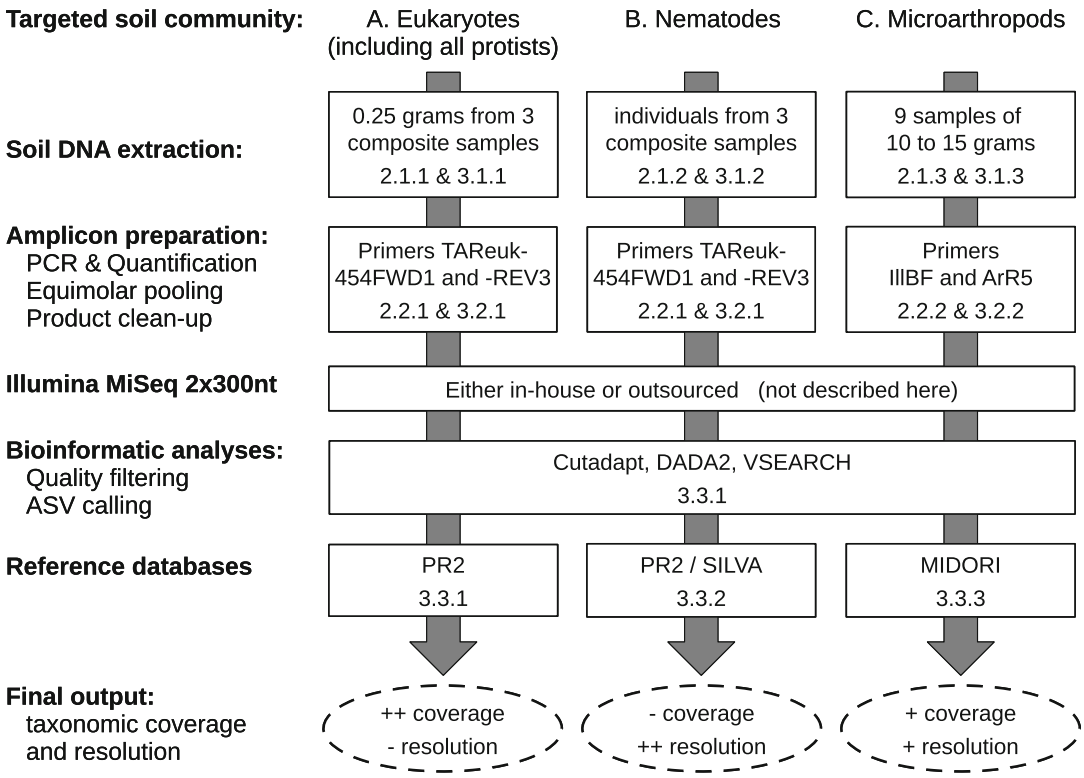

Fig. 2 Schematic overview of metabarcoding strategy targeting the entire eukaryome (**a**) or individual groups in more focused analyses such as nematodes (**b**) and microarthropods (**c**). Distinct sample preparation and wet laboratory processing are needed for the three targets. Sequencing and bioinformatic processing are identical for all three targets, except for the sequence reference databases which need to be adapted to the sequenced target and clade. General eukaryotic primers (**a**) allow for broad coverage of all eukaryotes while having the same time a limited taxonomic resolution (hardly discriminate species); general eukaryotic primers when used with nematode extracts (**b**) allow for specific amplification and sequencing of this clade and enable good species resolution; arthropod primers (**c**) have intermediate taxonomic coverage and resolution

2 Materials

2.1 DNA Extraction

2.1.1 Soil DNA Extraction for the Entire Eukaryome

1. Soil core sampler (ø = 6 cm; length depending on the desired sampling depth, generally 5 cm).
2. Plastic bags and/or PVC rings (1 per sample; 5 cm high, ø = 6 cm) for carrying soil cores.
3. Hammer.
4. Sharp knife.
5. Cooling box.
6. Water, brush, and 70% v/v ethanol to clean sampler.
7. 5 and 2 mm mesh size sieves.
8. 70% v/v ethanol or bleach and distillate water to clean sieves.

9. DNA extraction kit for soil: DNeasy PowerSoil (Qiagen) or NucleoSpin Soil (MACHEREY-NAGEL).
10. Spectrophotometer for DNA concentration measurement (e.g., NanoDrop, Thermo Fisher Scientific).
11. Optional: DNA preservation solution (e.g., LifeGuard, RNA-later, DMSO-EDTA-salt solution).

2.1.2 Nematode Isolation From Soils with Subsequent DNA Extraction

1. Use the same sampling material as in Subheading 2.1.1 up to **item 6**.
2. Oostenbrink elutriator or any established elutriator to extract nematode individuals.
3. Same DNA extraction kit as in Subheading 2.1.1 or a tissue/blood specific DNeasy or NucleoSpin kit.

2.1.3 Soil DNA Extraction for Microarthropods

1. Use the same material as in Subheading 2.1.1 except **items 7** and **8**.
2. Plastic bottle of 100 ml or 50 ml centrifuge tubes.
3. Silica gel in 3 g bags.
4. Phosphate buffer (Na2HPO4; 0.12 M; pH ~ 8).

2.2 Amplicon Preparation for Metabarcoding

2.2.1 All Eukaryotes with Special Focus on Protists and Nematodes

1. DNA template with concentration standardized to ~5 to 10 ng/μL.
2. PCR Hot-Start Taq polymerase and buffer (GoTaq® G2 Hot-Start Taq Polymerase, Promega).
3. PCR equipment (Thermocycler, 96-well plates).
4. PCR product quantification device (Picogreen or Qubit).
5. PCR cleanup kit (Wizard® Plus SV Minipreps DNA Purification System, Promega).
6. A set of 24 forward and reverse primers TAReuk454FWD1 (5'-CCA GCA SCY GCG GTA ATT CC-3') and TAReuk-REV3 (5'-ACT TTC GTT CTT GAT YRA-3') [33], preceded with a 5'-end spacer of 2 to 4 N and a one of the 24 8 bp barcode sequence (*see* **Note 1**).

2.2.2 Microarthropods

1. Same as **items 1** to **5** of Subheading 2.2.1.
2. A set of 24 forward and reverse primers IllBF (5'-CCN GAY ATR GCN TTY CCN CG-3') [34], and ArR5 (5'-GTR ATN GCN CCN GCN ARN AC-3') [35], preceded with a 5'-end spacer of 2 to 4 N and a one of the 24 8 bp barcode sequence.

2.3 Software for Bioinformatic Processing

2.3.1 In a Bash Operating System (Linux or macOS)

1. cutadapt (https://cutadapt.readthedocs.io).
2. vsearch (https://github.com/torognes/vsearch).
3. GNU parallel (https://www.gnu.org/software/parallel/).
4. biom-format (https://biom-format.org/index.html).

2.3.2 In R v3.4 and Beyond (https:/cran.r-project.org/)

1. plyr (https://cran.r-project.org/web/packages/plyr/index.html).
2. dada2 (https://benjjneb.github.io/dada2/index.html).
3. seqinr (https://cran.r-project.org/web/packages/seqinr/index.html).
4. digest (https://cran.r-project.org/web/packages/digest/index.html).

2.3.3 Example Data and Scripts

https://github.com/lentendu/V4_SSU_ASV_bioinformatic_pipeline/.

3 Methods

3.1 Soil DNA Extraction

3.1.1 Soil Total DNA Extraction for the Entire Eukaryome

1. Using a split soil corer, take a total of 9 samples (ca 250 g each), evenly distributed along the outer edge of a circular plot or evenly scattered inside a square plot with a 1 to 10 m radius/side length. Plot design has to be consistent all over the study or habitat type investigated. Soil core depth depends on habitat type, soil horizon, and targeted community.
2. Seal the samples individually in plastic bags and transported to the laboratory. Keep samples in a cool box during transport and store at 4 °C.
3. Homogenize 3 adjacent samples of the same plot by sieving them at 5 and 2 mm in order to produce a total of three sieved composite samples per plot (i.e., replicate samples). Only use one fifth (~ 50 g) of the original samples to keep material for nematodes and arthropods extraction. If the samples have to be carried during more than 2 days to reach the laboratory, sieve on site or at a base camp and store aliquots of 0.5 to 1 g of soil in at least two 2 mL tubes (fill until half the volume of the tube) and add 1 mL of DNA preservative solution in each tube. Then, maintain tubes at 4 °C whenever possible during transportation. In the laboratory, either store aliquots with preservative medium at −20 °C or prepare aliquots of 10 grams per composite sieved soil sample in 15 mL tubes and store at −20 °C.

4. One DNA extraction is performed for each composite sample using 0.3 g of soil material following the soil DNA extraction kit instructions. DNA preservation solution, if any, is removed by centrifugation (remove supernatant after 1 min at 2500 × g) prior to DNA extraction. When using a salt-based solution for preservation, the samples need two rounds of washing with PCR grade water prior to DNA extraction to remove any trace of preservative solution, which may otherwise interfere with the DNA extraction protocol.
5. Obtained DNA extracts of the same plot may or may not be pooled prior to PCR amplification in order to limit the number of samples considered for amplicon sequencing (Subheading 3.2.1). This pooling step depends on the need to integrate the spatial heterogeneity/variability within the sampled plot.

3.1.2 DNA Extraction of Nematodes

1. Follow **Steps 1** to **2** of Subheading 3.1.1.
2. Using half of each soil sample (ca. 100–150 g), pool soils of three adjacent samples to create three composite samples per plot.
3. Extract nematodes (ca. 5–10 k individuals) using one of several nematode elutriators (e.g., Oostenbrink, Seinhorst) according to the specific instructions.
4. Concentrate extracted nematodes in the first tube of the DNA extraction kit and store at −20 °C.
5. Proceed with nematodes DNA extraction following kit instructions.
6. Consider pooling DNA extracts of the same plot; *see* Subheading 3.1.1, **step 5**.

3.1.3 Soil Extracellular DNA Extraction for Microarthropods

1. Follow **steps 1** to **2** of Subheading 3.1.1.
2. Place 15 to 30 g of each soil sample into a plastic bottle with silica gel and store at room temperature. Silica gel may need to be replaced after few days for wet samples.
3. Remove silica gel and add phosphate buffer into plastic bottle (1/1 v/v). Shake horizontally for 20 min.
4. Retrieve two times 2 mL of supernatant into a 2 mL tube and centrifuge at 10000 g for 1 min. Use 500 μL of supernatant to start standard soil DNA extraction kit protocol at the first DNA binding step (e.g., start at **step 6** in the NucleoSpin Soil DNA extraction kit). If the DNA is not enough concentrate, use until 2 mL of supernatant to bind DNA.
5. Consider pooling DNA extracts of samples from the same plot; *see* Subheading 3.1.1, **step 5**.

3.2 Amplicon Preparation

Perform all steps in sterile conditions to prevent contamination.

3.2.1 Metabarcoding of the Soil Eukaryome, Including Protists and Nematodes

1. Use primers TAReuk454FWD1 and TAReukREV3 [33] to target a 390–420 bp long region of the 18S rDNA gene of a wide range of eukaryotes including protists, fungi, and fauna (*see* **Note 2**). Primers may include an heterogeneity spacer at their 5'-end (*see* **Note 3**).
2. Carry out PCR reactions in 96-well plates in 20 μL volume consisting of 1.2 μL of each primer (5 μM), 4 μL of buffer, 2 μL of 25 mM $MgCl_2$, 0.6 μL of 10 mM DNTPs, 0.1 μL of 5 U/μL Taq polymerase, 9.9 μL of PCR grade H_2O, and 1 μL of 5 to 10 ng/μL of template DNA (*see* **Note 4**). Only use unique combination of forward and reverse barcoded primers.
3. Apply the following PCR setup: initial denaturation at 95 °C for 5 min, followed by 35 cycles of denaturation at 94 °C for 30s, annealing at 47 °C for 45 s and elongation at 72 °C for 1 min with a final extension for 10 min at 72 °C (*see* **Note 5**). Use forward and reverse primers containing exclusive barcodes for individual samples (*see* **Note 6**).
4. PCRs (**items 1** to **3**) have to be replicated 2 to 3 times. The duplicate or triplicate PCR products are pooled together in order to reduce PCR amplification bias and obtain enough PCR product.
5. Quantify PCR products using a fluorometric quantification device, such as Picogreen™ or Qubit™ (Invitrogen).
6. Pool the PCR products of multiple samples in equimolar concentrations. Pool only PCR products with unique barcoded primers combinations, for a maximum of 24 samples. Each library ideally contains a PCR positive or mock community, a PCR negative (replace DNA template by PCR grade water) and a DNA extraction negative (soil replaced by PCR grade water for DNA extraction). In case of >21 samples, allocate the amplicons to multiple libraries (*see* **Note 6**).
7. Purify library DNA using membrane-based purification kit.
8. Send purified libraries for library preparation and Illumina MiSeq 2 × 300 bp sequencing using the company's standard protocol (or *see* **Note 7**).

3.2.2 Metabarcoding of Microarthropods

1. Use primers IllBF [34] and ArR5 [35, 36] to target a 315 bp long region of the mitochondrial COI gene of most arthropods (*see* **Note 8**).
2. Use same PCR reaction mixes as in Subheading 3.2.1, **step 2**.
3. Apply the following PCR setup: initial denaturation at 95 °C for 5 min, followed by 40 cycles of denaturation at 95 °C for

30s, annealing at 47 °C for 45 s and elongation at 72 °C for 45 s, with a final extension phase for 10 min at 72 °C (*see* **Note 5**). Use forward and reverse primers containing exclusive barcodes for individual samples in each library (*see* **Note 6**).

4. Continue with the same **steps 4** to **8** as for Subheading 3.2.1. The sequencing can be performed on an Illumina MiSeq 2 × 300 bp or Illumina MiSeq 2 × 250 bp platform, as the amplified fragment is shorter than the one for whole eukaryotes.

3.3 Bioinformatic Analyses

Below, we describe pipelines for the bioinformatic analysis of the sequencing output, largely making use of the Cutadapt [37] and VSEARCH software [38] as well as the DADA2 R package [39]. An example of the exact list of scripting command lines as applied to sequencing data for the general eukaryotes primer set (as in Subheading 3.2.1) is provided on the companion website (https://github.com/lentendu/V4_SSU_ASV_bioinformatic_pipeline/).

3.3.1 Metabarcoding of All Eukaryotes with a Focus on Protists

Two cases can arise depending on the library preparation strategy:

(a) The standard Illumina protocol with one or two PCR steps ensures that the forward sequencing adapter is attached to the forward biological primer and conversely. This procedure produces reads holding exclusively the forward primer in the R1 library and exclusively the reverse primer in the R2 library (latter refer as "case a"). If raw reads are already demultiplexed, start at **item 4**; otherwise, start at **item 2**.

(b) In the ligation protocol, the forward and reverse sequencing adapters are equally attached to the forward and the reverse biological primers. This produces reads holding both the forward and the reverse primer at the 3'-end in the R1 as well as in the R2 libraries (latter refer as "case b"). Processing starts at **item 1**.

1. Reads with different orientations (i.e., reads with either the forward or the reverse primer at the 5'-end) have to be first separated inside each R1 and R2 library. The forward and the reverse primers are searched toward the 5'-end inside both the R1 and R2 libraries, using the -g, -G, --no-indels, and --trimmed-only options of Cutadapt. The list of primer sequences in fasta format is provided to the g/G options, with the sequence identifiers being the primer names which are used to label the output fastq filenames. Multiple mismatches (until 30%) can be allowed on the primer sequence as this is only a presorting of reads to facilitate the demultiplexing. The option --action=none prevent removing any

nucleotide from the reads. There should be only minimal loss of reads at that step.

2. Raw reads are demultiplexed using Cutadapt. The reads of each library and each orientation are searched for the barcode sequences at their 5'-end using the -g, -G, --no-indels, and --trimmed-only options. The list of barcode sequences in fasta format is provided to the g/G options, with the sequence identifiers being the sample names, which are used to label the output fastq filenames. It is recommended to use noninternal adapters to improve the barcode detection by adding the "X" symbol at the beginning of the barcode sequences (*see* **Note 9**). Depending on the length of the barcode sequences and the combinations of barcode used, until 2 mismatches (max 25%) can be allowed. This step has to be repeated for the reads in the other orientation for case b. The outputs are one (case a) or two (case b) pairs of raw read files for each combination of forward and reverse barcode.
3. The expected combination of forward and reverse barcodes is assigned to their respective sample's names.
4. The primers are stripped from the 5'-end using Cutadapt option -g and -G. Multiple mismatches can be allowed on the primer sequence as the quality of this fragment is not necessarily linked to the quality of the biological sequence (till 25%).
5. The quality and number of reads in each sample are assessed using the VSEARCH option –fastq_stat and --fastq_eestats2. In order to proceed further with amplicon sequence variant (ASV) calling using error model correction [39], the reads of each paired libraries and of each orientation need to be trimmed at a fixed length (*see* **Note 10**). The maximum expected error (maxEE) can be used to select an appropriate length, while keeping enough nucleotide in each R1 and R2 to allow for pair-end assembly (*see* **Note 11**).
6. The reads of each library and each orientation are truncated to the same length and filtered with the same maxEE threshold using the VSEARCH options --fastq_filter, --fastq_trunclen, and --fastq_maxee. As the filter does not accept paired fastq files, the command has to be run independently for R1 and R2 libraries. Only the reads passing the filter in both directions will be subsequently selected by first joining the sequence identifiers passing the filter in both libraries and then using the --fastx_getseq option of VSEARCH to select reads in each paired files.
7. The reads are dereplicated in each sample, library, and orientation separately using the DADA2 R package command derepFastq.

8. The error rate is inferred using all samples of the same Illumina run for each library and each orientation separately using the DADA2 command learnErr (*see* **Note 12**).
9. The sequence variants are assessed over all samples of the same run, same library, and same orientation using the DADA2 command dada, with the pool option set to TRUE.
10. Pairs of reads from R1 and R2 libraries are merged for each orientation using the DADA2 command mergePairs. A minimum overlap of 10 nt is required and the maximum number of mismatches can be adjusted, with a good starting point at 10% of the minimum overlap.
11. The count table is generated with the DADA2 command makeSequenceTable. For case b (ligation-based library), the paired reads originally with the reverse primer at the 5'-end of R1 library are reverse-complemented using the SeqinR R package command c2s [40].
12. Sequence count tables are merged throughout runs and orientations using the DADA2 command mergeSequenceTables.
13. Chimera are detected and removed using the DADA2 command removeBimeraDenovo, with the pool option set to TRUE.
14. The count table is exported to a TAB-separated values file. ASV sequences are labeled with their SHA1 hash using the digest R package command sha1 [41] and exported to a fasta file.
15. The taxonomy of the ASV is assigned by comparing to sequences of the Protist Ribosomal Reference database (***PR2;*** [42]) using the global pairwise alignment (--usearch_global option) of VSEARCH. Pairwise identity is computed without taking into account terminal gaps (option "--iddef 2"). All best hits with a minimum similarity of 60% are conserved (options "--id 0.6 --top_hits_only --maxaccepts 0").
16. A consensus taxonomy is generated for ASV assigned to multiple best hits with divergent taxonomy using a 60% threshold for each taxonomic rank.
17. ASV count table and taxonomic information are assembled into one table and format into a BIOM table for sparse and long-term storage [43].

3.3.2 Metabarcoding of Nematodes

1. Use the same **steps 1** to **17** of Subheading 3.3.1. The **SILVA SSU** reference database [44] can be used as an alternative at **item 15** instead of PR2.

3.3.3 Metabarcoding of Microarthropods

1. Use the same **steps 1** to **17** of Subheading 3.3.1, but use the **MIDORI Reference 2 COI** ("CO1," [45]) reference database at **item 15** instead of PR2.

4 Notes

1. Barcodes of at least 8 nt length have to be selected to have an almost equal share of each four bases at each position when pooled together and with at least 4 bp pairwise distance (without gap allowed). Sets of 26 forward and 26 reverse barcodes were designed and successfully tested in previous studies [46]. 12-nt long barcodes allow to use up to 2167 different barcodes with at least 8 bp pairwise distance, which need to be considered when more than 21 samples have to be pooled in a single library [47]. Barcodes attached to the forward and the reverse primers of a unique barcode pair can be identical.
2. The most common alternative primer pair used to study eukaryotic diversity in environmental samples is 1389F (5'-TTG TAC ACA CCG CCC-3') and 1510R (5'-CCT TCY GCA GGT TCA CCT AC-3'). These primers target a 120 to 150 bp long region of the V9 SSU rDNA gene [48]. This shorter fragment is sequenced on other Illumina platforms producing shorter reads (e.g., Miseq 2 × 150 bp, HiSeq, or NovaSeq) and has generally a similar taxonomic resolution than the proposed primer pair [16, 17]. A disadvantage of this approach is that many publicly available 18S rDNA sequences stop before the V9 region. On the other hand, HiSeq and NovaSeq have greater sequencing depth, and flanking regions are more conserved in eukaryotes, which imply relatively less amplification biases. Other primers may be used to amplified the full-length SSU rRNA gene or the rRNA operon (18S, ITS, 28S), to be sequenced on third-generation high-throughput sequencing platform like Nanopore or PacBio [16, 49]. Every primer pair will have its own specific set of amplification biases, some groups being better/less amplified with certain primers [16, 17].
3. The addition of a 0 to 4 base long heterogeneity spacer at the 5'-end of the primer is a valid approach to increase the base heterogeneity at 5'-end of the sequences which increase the quality of the Illumina sequencing and reduce the need for PhiX addition to sequencing libraries [50].
4. When more volume of PCR product is needed in the following steps, PCR reactions might be prepared in 50 μL final volume, in which case all volumes have to be multiplied by a factor of 2.5.
5. PCR conditions might vary depending on the polymerase and thermocycler used and should be tested using a gradient PCR before application. An altered annealing temperature will, however, provide a different picture of the resulting community as higher annealing temperatures will benefit those targets that

optimally bind primers, while lower temperature will also amplify less-specific targets. The approach taken depends on the experimental question but, in order to compare between different studies, adopting an identical protocol is recommended.

6. If more libraries are to be created (*see* Subheading 3.2.1, **step 6**), the same primers (i.e., with the same barcodes) can be reused in different libraries, thereby reducing primer costs. In this case, the distribution of samples over libraries should be considered already before the PCR steps, so that samples can randomly be allocated to a library.
7. For ligation-based library preparation with Illumina TruSeq kit, avoid the T4 DNA polymerase blunt-ending and the post-ligation PCR as these steps are not necessary for amplicon libraries, and they cause tag jumps [51].
8. Alternatives primers targeting the mitochondrial 16S rRNA gene (Chiar16SF: 5'-TAR TYC AAC ATC GRG GTC-3', Chiar16SR: 5'-CYG TRC DAA GGT AGC ATA-3', ~ 350 bp long) [52] can be used here.
9. The X symbol at 5'-end of the barcode sequences allows for partial match at the 5'-end only. This is relevant when primers are produced without HPLC purification as they could have 1 to 2 nucleotides missing at their 5'-end, which can cause mis-tagging due to reduced length of the barcode sequences. Using a heterogeneity spacer in front of the barcode sequence is a good solution to mitigate this issue (e.g., 2 to 4 N's at 5'-end); this produces also higher overall reads quality (*see* **Note 3**).
10. The quality of the reads is generally dropping toward the 3'-end, a tendency that is generally more pronounced in the R2 libraries.
11. To ensure proper alignment, pair-end assembly should be done on at least ten nucleotides, so that the cumulative length of R1 and R2 reads after length truncation have to be at least the maximal length of the biological sequence plus ten. The --fastq_eestats2 option in VSEARCH is in this sense ideal to simulate the number of reads retained when using different length and maxEE cutoffs. The maxEE filter can be tested over a range from 0.5 to 4. As rule of thumb, length truncation and maxEE filtration should not remove more than 20% of reads. If more reads are dropped, there might be quality issues with the sequencing run.
12. The more data are provided to the model, the better the model prediction will be. Also, the better ASVs that are rare in a sample but abundant in others will be detected. It is possible to infer the error rate for each sample individually, which allows

for parallel computation and a large reduction of computation time. Single sample error rate inference tends to remove rare ASV, which may or may not be desirable depending on scientific questions. These sequences are often present in low abundance, but spread over multiple samples. Further guidelines are available on the tool website (https://benjjneb.github.io/dada2/).

References

1. FAO, ITPS, CBD, et al (2020) State of knowledge of soil biodiversity - status, challenges and potentialities: report 2020. FAO, Rome, Italy
2. Geisen S, Briones MJI, Gan H et al (2019) A methodological framework to embrace soil biodiversity. Soil Biol Biochem 136:107536
3. Bahram M, Hildebrand F, Forslund SK et al (2018) Structure and function of the global topsoil microbiome. Nature 560:233
4. Delgado-Baquerizo M, Oliverio AM, Brewer TE et al (2018) A global atlas of the dominant bacteria found in soil. Science 359:320–325
5. Tedersoo L, Bahram M, Põlme S et al (2014) Global diversity and geography of soil fungi. Science 346:1256688
6. Thompson LR, Sanders JG, McDonald D et al (2017) A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457
7. Mahé F, de Vargas C, Bass D et al (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. Nat Ecol Evol 1:0091
8. Oliverio AM, Geisen S, Delgado-Baquerizo M et al (2020) The global-scale distributions of soil protists and their contributions to belowground systems. Sci Adv 6:eaax8787
9. Singer D, Seppey CVW, Lentendu G et al (2021) Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. Environ Int 146:106262
10. Xiong W, Jousset A, Li R et al (2021) A global overview of the trophic structure within microbiomes across ecosystems. Environ Int 151: 106438
11. Zhao Z-B, He J-Z, Geisen S et al (2019) Protist communities are more sensitive to nitrogen fertilization than other microorganisms in diverse agricultural soils. Microbiome 7:33
12. Guo S, Xiong W, Hang X et al (2021) Protists as main indicators and determinants of plant performance. Microbiome 9:64
13. Xiong W, Song Y, Yang K et al (2020) Rhizosphere protists are key determinants of plant health. Microbiome 8:27
14. Adl SM, Bass D, Lane CE et al (2019) Revisions to the classification, nomenclature, and diversity of eukaryotes. J Eukaryot Microbiol 66:4–119
15. de Groot AG, Laros I, Geisen S (2016) Molecular identification of soil Eukaryotes and focused approaches targeting Protist and Faunal Groups using high-throughput Metabarcoding. In: Martin F, Uroz S (eds) Microbial environmental genomics (MEG). Springer, New York, pp 125–140
16. Geisen S, Vaulot D, Mahé F et al (2019) A user guide to environmental protistology: primers, metabarcoding, sequencing, and analyses. bioRxiv:850610
17. Vaulot D, Geisen S, Mahé F et al (2022) pr2-primers: an 18S rRNA primer database for protists. Mol Ecol Resour 22:168–179
18. Luan L, Jiang Y, Cheng M et al (2020) Organism body size structures the soil microbial and nematode community assembly at a continental and global scale. Nat Commun 11:6406
19. van den Hoogen J, Geisen S, Routh D et al (2019) Soil nematode abundance and functional group composition at a global scale. Nature 572:194–198
20. Geisen S, Snoek LB, ten Hooven FC et al (2018) Integrating quantitative morphological and qualitative molecular methods to analyse soil nematode community responses to plant range expansion. Methods Ecol Evol 9:1366–1378
21. Chen XY, Daniell TJ, Neilson R et al (2010) A comparison of molecular methods for monitoring soil nematodes and their use as biological indicators. Eur J Soil Biol 46:319–324
22. Porazinska DL, Giblin-Davis RM, Faller L et al (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Mol Ecol Resour 9:1439–1450
23. Wilschut RA, Geisen S, Martens H et al (2019) Latitudinal variation in soil nematode communities under climate warming-related range-

expanding and native plants. Glob Change Biol 25:2714–2726

24. Ji Y, Ashton L, Pedley SM et al (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett 16:1245–1257
25. Yu DW, Ji Y, Emerson BC et al (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623
26. Oliverio AM, Gan H, Wickings K et al (2018) A DNA metabarcoding approach to characterize soil arthropod communities. Soil Biol Biochem 125:37–43
27. George PBL, Lallias D, Creer S et al (2019) Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems. Nat Commun 10: 1107
28. Fontaneto D, Eckert EM, Anicic N et al (2019) We are ready for faunistic surveys of bdelloid rotifers through DNA barcoding: the example of Sphagnum bogs of the Swiss Jura Mountains. Limnetica 38:213–225
29. Lentendu G, Wubet T, Chatzinotas A et al (2014) Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach. Mol Ecol 23:3341–3355
30. Forster D, Lentendu G, Filker S et al (2019) Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. Environ Microbiol 21: 4109–4124
31. Geisen S, Tveit AT, Clark IM et al (2015) Metatranscriptomic census of active protists in soils. ISME J 9:2178–2190
32. Thompson AR, Geisen S, Adams BJ (2020) Shotgun metagenomics reveal a diverse assemblage of protists in a model Antarctic soil ecosystem. Environ Microbiol 22:4620–4632
33. Stoeck T, Bass D, Nebel M et al (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19:21–31
34. Hajibabaei M, Spall JL, Shokralla S et al (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecol 12:28
35. Gibson J, Shokralla S, Porter TM et al (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc Natl Acad Sci 111:8007–8012
36. Hajibabaei M, Porter TM, Wright M et al (2019) COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. PLoS One 14:e0220953
37. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12
38. Rognes T, Flouri T, Nichols B et al (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584
39. Callahan BJ, McMurdie PJ, Rosen MJ et al (2016) DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583
40. Charif D, Lobry JR (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE et al (eds) Structural approaches to sequence evolution: molecules, networks, populations. Springer, Berlin, Heidelberg, pp 207–232
41. Eddelbuettel D (2020) digest: Create Compact Hash Digests of R Objects, 0.6.27, digest
42. Guillou L, Bachar D, Audic S 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
43. McDonald D, Clemente JC, Kuczynski J et al (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience 1:7
44. Quast C, Pruesse E, Yilmaz P et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596
45. Leray M, Knowlton N, Machida RJ (2022) MIDORI2: A collection of quality controlled preformatted and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. Environmental DNA 4(4):894–907. 10.1002/edn3.303
46. Cordier T, Esling P, Lejzerowicz F et al (2017) Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. Environ Sci Technol 51:9118–9126
47. Caporaso JG, Lauber CL, Walters WA et al (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624
48. Amaral-Zettler LA, McCliment EA, Ducklow HW et al (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–e6372

49. Jamy M, Foster R, Barbera P et al (2020) Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. Mol Ecol Resour 20:429–443
50. Jensen EA, Berryman DE, Murphy ER et al (2019) Heterogeneity spacers in 16S rDNA primers improve analysis of mouse gut microbiomes via greater nucleotide diversity. BioTechniques 67:55–62
51. Carøe C, Bohmann K (2020) Tagsteady: a metabarcoding library preparation protocol to avoid false assignment of sequences to samples. Mol Ecol Resour 20:1620–1631
52. Marquina D, Andersson AF, Ronquist F (2019) New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. Mol Ecol Resour 19: 90–104