The ISME Journal (2016) 10, 2488–2497

www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Local diversity of heathland Cercozoa explored by in-depth sequencing

Christoffer Bugge Harder[1,2], Regin Rønn[1], Asker Brejnrod[3], David Bass[4,5], Waleed Abu Al-Soud[3] and Flemming Ekelund[1]

[1]*Section of Terrestrial Ecology, Department of Biology, University of Copenhagen, Copenhagen, Denmark;* [2]*Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, Oslo, Norway;* [3]*Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark;* [4]*Department of Life Sciences, The Natural History Museum, Cromwell Road, London, UK and* [5]*Centre for Environment, Fisheries and Aquaculture Science (Cefas), The Nothe, Weymouth, Dorset, UK*

Cercozoa are abundant free-living soil protozoa and quantitatively important in soil food webs; yet, targeted high-throughput sequencing (HTS) has not yet been applied to this group. Here we describe the development of a targeted assay to explore Cercozoa using HTS, and we apply this assay to measure Cercozoan community response to drought in a Danish climate manipulation experiment (two sites exposed to artificial drought, two unexposed). Based on a comparison of the hypervariable regions of the 18S ribosomal DNA of 193 named Cercozoa, we concluded that the V4 region is the most suitable for group-specific diversity analysis. We then designed a set of highly specific primers (encompassing ~270 bp) for 454 sequencing. The primers captured all major cercozoan groups; and >95% of the obtained sequences were from Cercozoa. From 443 350 high-quality short reads (>300 bp), we recovered 1585 operational taxonomic units defined by >95% V4 sequence similarity. Taxonomic annotation by phylogeny enabled us to assign >95% of our reads to order level and ~85% to genus level despite the presence of a large, hitherto unknown diversity. Over 40% of the annotated sequences were assigned to Glissomonad genera, whereas the most common individually named genus was the euglyphid *Trinema*. Cercozoan diversity was largely resilient to drought, although we observed a community composition shift towards fewer testate amoebae.

The ISME Journal (2016) **10,** 2488–2497; doi:10.1038/ismej.2016.31; published online 8 March 2016

Introduction

Protozoa are essential organisms in soil ecosystems primarily because they have a significant role in the soil food web as bacterial grazers (Ekelund and Rønn, 1994). Moreover, protozoan diversity is a good index of ecosystem function (Griffiths *et al*., 2000); in particular, it has been suggested that the high diversity allows the soil protozoan community to respond to both seasonal and environmental change (Bamforth, 1995). As fundamentally aquatic organisms, one would *a priori* expect protozoa to be negatively affected by drought, however empirical evidence is ambiguous. Whereas Eisenhauer *et al.* (2012) reported that drought decreased protozoan abundance, Schmitt and Glaser (2011) found that drought increased the protozoan diversity. Because of the large unknown protozoan diversity, and the difficulty in identifying many species morphologically, it has been problematic to evaluate such contrasting views.

However, the fast development in high-throughput sequencing techniques (HTS) now offers tools to answer such questions; still, only few such studies have specifically targeted soil protozoa. One reason for this is that heterotrophic soil protozoa include members from at least four kingdoms/supergroups (Baldauf *et al*., 2000). Therefore, they can inherently only be amplified with general eukaryotic primers, which will unavoidably also amplify DNA from other organisms abundant in soil. Thus, sequences from metazoans, fungi and plants often dominate samples, and must be removed prior to further analysis. Further, in case of micro-eukaryotes, we only have a comprehensive reference database for the small subunit ribosomal RNA (18S) gene. The copy number of this gene may vary considerably between different eukaryotic groups (Zhu *et al*., 2005), which complicates quantitative comparison of community composition based on sequence abundance.

Here, we tackle these problems by targeting a specific protozoan group, the Cercozoa; for a phylogenetic overview of this group see Bass *et al.* (2009b).

Correspondence: F Ekelund, Section of Terrestrial Ecology, Department of Biology, University of Copenhagen, Universitetsparken 15, Copenhagen DK-2100, Denmark.
E-mail: fekelund@bio.ku.dk
Received 14 August 2015; revised 27 November 2016; accepted 8 January 2016; published online 8 March 2016

Cercozoa encompass a high morphological diversity. They include large testate amoebae such as *Euglypha* and *Trinema*, naked filose/reticulate amoebae such as vampyrellids, granofiloseans and *Filoreta*, and gliding flagellates such as cercomonads, glissomonads and thaumatomonads (Bass *et al*., 2009a; Hess *et al*., 2012; Berney *et al*., 2013). Owing to the high morphological and physiological diversity, Cercozoa also show high functional and ecological diversity. Hence, changes in the relative abundance of different cercozoan groups could potentially be a valuable indicator of environmental change. For example, paleohydrological studies have shown that testate amoebae are sensitive indicators of water content in peat bogs (Charman and Warner, 1992; Booth, 2001, 2008) and there is evidence that testate amoebae also respond negatively to drought in more arid soils (Lousier, 1974a, b; Wilkinson and Mitchell 2010). Hence, it is likely that testate Cercozoa, such as *Euglypha* and *Trinema*, would decrease in abundance in response to drought.

Morphological methods have long suggested that Cercozoa is one of the dominant groups of free-living eukaryotic microorganisms in temperate soils (Sandon, 1927; Ekelund *et al*., 2001). This has been confirmed by recent HTS-based studies. Bates *et al.* (2013) found that Cercozoa accounted for ~30% of the identifiable protozoan 18S reads in arid or semi-arid soils and ~15% in more humid soils. In a transcriptomic analysis of soil protist activity, Geisen *et al.* (2015) found that 40–60% of all identified protozoan small subunit ribosomal RNAs in forest and grassland soils could be assigned to Cercozoa. Cercozoa are also abundant in marine benthic and interstitial communities; a recent HTS study using general eukaryotic primers found Cercozoa to comprise between 9 and 24% of all assigned eukaryotic operational taxonomic units (OTUs) on the ocean floor (Pawlowski *et al*., 2011).

Several recent papers have named many new cercozoan taxa at species, genus and family level from temperate topsoil (Bass *et al*., 2009b; Howe *et al*., 2009, 2011; Chatelain *et al*., 2013). Sanger sequencing of environmental DNA in Cercozoa (Bass and Cavalier-Smith, 2004) has shown a further undescribed diversity. However, the number of sequences obtained by this method is low compared with HTS methods, and though HTS is now ubiquitous in studies of microbial ecology, to our knowledge no HTS-based study has yet targeted Cercozoa directly. Reasons for this likely include lack of standard protocols with consensus on using a suitable region of the 18S ribosomal RNA gene and lack of good primers. Moreover, existing reference databases with sequences for OTU identification are rather limited. For example, as of November 2015, the Silva database contained only ~300 full-length cercozoan 18S sequences.

Here we present an HTS-based analysis of soil Cercozoa. We first identified the best suited 18S region in Cercozoa; next we designed an appropriate primer to target this region. We then used this primer to target soil Cercozoa directly in an experiment where we tested the hypothesis that testate Cercozoa respond negatively to drought. To annotate sequences, we used a yet unpublished database (David Bass, in preparation) that contains 966 cercozoan 18S sequences of high-quality, which cover the entire group. We hope that we provide theoretical and practical foundations needed to establish a frame for future comparative molecular analyses of cercozoan diversity.

Materials and methods

Study site and soil sampling

Projected climate change for Denmark in the 2100th century indicates drier summers, which are experimentally simulated on the CLIMAITE study site (Larsen *et al*., 2011). The CLIMAITE experimental site is situated in a dry heath-/grassland 50 km NW of Copenhagen, Denmark (55° 53′ N, 11° 58′ E). The mineral fraction of the soil consists of 92% sand, 5.8% silt and 2.2% clay (Nielsen *et al*., 2009). The site is well drained with an organic top layer (O-horizon). The pH_{CaCl2} in the O-horizon is 3.3 increasing to 4.5 in the lower B-horizon. The dominant vegetation consists of the dwarf shrub *Calluna vulgaris* (c. 30% cover) and the perennial grass *Deschampsia flexuosa* (c. 70% cover). The annual mean temperature is 8 °C with a mean precipitation of 613 mm (Danish Meteorological Institute, www.dmi.dk). Since 2005, a complete three-factorial treatment with increased CO_2, temperature and summer drought has been maintained in 12 four-chamber octagons (7 mm in diameter), where the fourth octagon is a control plot with no treatments. Each treatment is replicated seven times. These treatments are intended to mimic the projected climate change for the region in 2075. Drought is induced once or twice a year by automatic rain shelters, which exclude the precipitation continuously for 2–5 weeks until the water content plunges below 5% by volume in the upper 20 cm of the soil. Larsen *et al.* (2011) provide a detailed description of the experiment.

In November 2010, we sampled topsoil (0–8 cm, O-horizon) from two of the six drought plots 1–2 (Dry1) and 3–1 (Dry2) and from two of the control plots 9–1 (Control1) and 11–4 (Control2). To minimise the spatial variation, we pooled and mixed three subsamples from each plot for the subsequent analyses.

Comparisons of hypervariable regions

The entire 18S region typically spans a length of ~2000 bp, whereas HTS methods have not been able to amplify >500–700 bp at most. Therefore, it is necessary to identify good representative parts of the

18S region for HTS analyses. Prime candidates are the eight hypervariable regions labelled V1–V5 and V7–V9 (the V6 hypervariable region found in bacterial 16S is absent from eukaryotic 18S (Howe *et al.*, 2011). As the V4 region is the longest of the eight hypervariable regions, we found it the most attractive for taxonomic annotation. However, we wished to make sure that the diversity in V4 correlated reasonably well with the total 18S diversity as compared with the other hypervariable regions.

An appropriate clustering level-threshold constitutes another special problem in HTS analyses. Thus, it is necessary to choose a percentage-wise separation threshold when clustering the obtained sequences into OTUs. To obtain a robust theoretical foundation for HTS analysis, we first obtained two data sets of named cercozoan 18S Sanger-generated sequences. In June 2015, we obtained one set of 63 species from GenBank. This set contained the whole 18S region including all eight hypervariable regions in their entirety. We used the 63 sequence set to identify the V4 region as the best representative of the entire 18S diversity. The other, larger set consisted of 193 partial sequences that were at least 1500 bp long and all contained the V4 region. We included only named species documented in recent papers (Ekelund *et al.*, 2004; Hoppenrath and Leander, 2006; Lara *et al.*, 2007; Wylezich *et al.*, 2007; Bass *et al.*, 2009a, b; Burki *et al.*, 2010; Chantangsi and Leander, 2010a, b; Heger *et al.*, 2010; Heger *et al.*, 2011; Howe *et al.*, 2011; Yabuki and Ishida, 2011). This sequence set represented all nine cercozoan classes sensu Cavalier-Smith and Chao (2003), and contained no duplicate names or synonyms. We used this set of 193 sequences to evaluate the interspecific distances in the 18S region most suitable for separation of OTUs in the Cercozoa, and to test the effects of different OTU separation thresholds. Names and accessions numbers of the 63 +193 sequences are listed in Supplementary data (tables 1 and 2).

In order to precisely identify the position of the hypervariable regions in the Cercozoa, we use the E-ins-i algorithm in MAFFT (Katoh and Standley, 2013) to align the 63 sequences along with the complete sequence of the fungus AF258606 *Scytalidium hyalinum*, which had the start and end of each V1–V9 regions fully annotated in its documentation on GenBank. Using the command dist.seqs in MOTHUR (Schloss *et al.*, 2009), counting all indels as one event without penalising end gaps, we then calculated all possible uncorrected *P*-distances between the sequences for the whole 18S and for each of the eight hypervariable regions V1–V9, and correlated these *P*-distances for each region with the complete 18S. In this manner, we tested how well the genetic diversity of the respective regions correlated with that of the complete 18S region. All graphics and statistics were done in R (Ihaka and Gentleman, 1996).

Sequence variation between known species

To examine the congruence between already described Cercozoa and the genetic distances in V4, we aligned the data set of 193 sequences with MAFFT using the E-ins-i algorithm and calculated the uncorrected *P*-distances for all 18 527 pairs in MOTHUR. We then clustered the V4 region of these 193 sequences with the furthest neighbour-algorithm (implemented in MOTHUR) for all thresholds between 0 and 10%.

DNA extraction, primer design and initial cloning check

We extracted DNA from 0.5 g of fresh soil within 24 h of soil sampling. We used a genomic mini spin kit for universal DNA isolation (A&A biotechnology, Gdynia, Poland) with a standard protocol (Yu and Mohn, 1999). Based on the 193-sequence alignment, we designed primers that would amplify the majority of named key soil cercozoan genera within Granofilosea, Imbricatea, Cryomonadida, Cercomonadida, Glissomonadida and Euglyphida with no—or in some cases one—mismatch in the primer sequence. We accepted a slight bias against some members within these taxa, and against some genera, for example, *Cyphoderia*, *Platyreta*, *Arachnula* and *Filoreta* (two mismatches) and some bias against Chlorarachniophyta, Phytomyxea and Ascetosporea (notably Haplosporida and Mikrocytida) and other endomyxan lineages, and the genera *Helkesimastix*, *Sainouron*, *Cholamonas* (Cavalier-Smith *et al.*, 2009), Reticulamoeba (Bass *et al.*, 2012), and *Rosculus* and *Guttulinopsis* (Bass *et al.*, submitted) with three or more mismatches in each primer.

From the alignment of all these genera, we used the representative sequence AF411270 *Cercomonas longicauda* as template in Primer3 (Rozen and Skaletsky, 2000) to find two compatible primers: the forward primer Cerc479F (5'TGTTGCAGT TAAAAAGCTCGT-3', Tm = 57.8 °C) and the reverse primer Cerc750R (5'TGAATACTAGCACCCCCAAC-3', Tm = 57.5 °C). To check the specificity of the primers, we performed an initial PCR and cloning of 50 sequences. The PCR master mix (25 µl) consisted of 1 × High Fidelity buffer (Invitrogen, Carlsbad, CA) with $MgCl_2$, 0.25 mM deoxynucleotides mixture, 1 µl 100 × bovine serum albumin, 0.5 IU Phusion Hot Start DNA polymerase (5 units μl^{-1}, Invitrogen 0.4 µM) of each primer, 1 µl DNA template. The PCR incubation conditions consisted of an initial denaturation step of 94 °C for 5 min; 30 cycles of denaturation at 94 °C for 60 s, annealing at 55 °C for 60 s and elongation at 68 °C for 60 s; and finally, an extension step of 72 °C for 7 min. We chose to lower the annealing temperature from the theoretical optimum of the primers to compensate for the mismatches. Cloning was performed using TOPO TA Cloning Kit from Invitrogen, and sequencing of 50 supposedly positive clones from this PCR was done by MACROGEN in Seoul, South Korea.

DNA amplification and GS-FLX Pyrosequencing

The samples were prepared for GS-FLX pyrosequencing in a two-step PCR. We used a Platinum *Taq* DNA High Fidelity polymerase (5 units μl^{-1}, Invitrogen); otherwise the master-mix and the PCR incubation conditions were as above. To eliminate as many primer-dimers as possible, the products were incubated at 70 °C for 4 min and then stored immediately on ice before electrophoresis. We loaded the PCR products on a 1% agarose gel with ethidium bromide, which confirmed the presence of a single band in the desired length of ~250–300 bp with ultraviolet illumination. The bands of PCR products were excised from the agarose gel and purified by the Montage DNA Gel Extraction kit (Millipore, Bedford, MA).

The second PCR amplification was performed with fusion primers consisting of the raw primers above with the B-adaptors and four MID-tag barcodes of 10 bp added upon the forward primer and was amplified using only 15 PCR cycles. Otherwise, PCR incubation conditions, electrophoresis, gel excision and purification were as above. The amplified DNA from the second PCR was quantified with the Qubit dsDNA HS Assay Kit and the Qubit fluorometer (Invitrogen, Life technologies, Carlsbad, CA, USA) and mixed in approximately equal molar concentration (5×10^6 copies μl^{-1}) to ensure an approximately equal representation of sequences on each sample. A GS-FLX Titanium sequencing run was then performed on a 70_75 GS PicoTiterPlate (PTP) using a GS-FLX Titanium pyrosequencing system according to manufacturer instructions (Roche Diagnostics, Basel, Switzerland) at the National High-throughput DNA Sequencing Centre (Copenhagen, Denmark).

Bioinformatic analyses

In several taxonomic groups, including Rhizaria, error rates primarily linked to singletons and homopolymers may cause a considerable overestimation of diversity; especially in the V4 compared with the V9 region. GS-FLX Titanium was particularly susceptible to such errors compared with the GS-FLX standard kit, even when reads are clustered up to a 3% level (Behnke *et al.*, 2011). Hence, to eliminate such errors, we applied a strict quality sorting approach in our analysis; we eliminated singletons and long homopolymers, and chose a conservative 5% clustering threshold.

The titanium run produced 689 988 reads. We analysed it through the Qiime pipeline (Caporaso *et al.*, 2010) and discarded all reads that had a quality score below 25 or had any mismatches in the primer or MID-tag sequences. We also discarded reads with a length outside 200–1000 bps, as the shortest cercozoan among the 193 named species had a V4 length of 218 bp. This left 494 963 reads, which were run through ACACIA (Bragg *et al.*, 2012) to discard homopolymers >6 bp. Chimeras were removed with UCHIME (Edgar *et al.*, 2011). This removed further 22 254 and 12 310 reads, respectively. The rest were clustered at 5% with UCLUST (Edgar, 2010), and 838 post-clustering singletons were subsequently discarded. Representative sequences from the resulting 1745 OTUs were blasted using nblast (Altschul *et al.*, 1990). We removed another 160 OTUs that either had no BLAST hit (two OTUs), were presumed chimera with different BLAST hits of the 5' and 3' end (2), had top hits to non-target organisms (17 fungi, 2 ciliates, 1 heterokont), or had a query coverage of 60% or below (136 OTUs); and thus perhaps were chimeras. All the rest blasted to Cercozoa with a similarity of 80% or more to the most similar hit in GenBank. The final data set consisted of 443 350 sequences, distributed on the plots with 85 305 from Control1, 94 975 from Control2, 116 388 from Dry1 and 146 682 from Dry2. To obtain comparable data for rarefaction curves and statistical tests, we further resampled down to 85 305 sequences per plot. The data (the sff file) and barcode information has been archived on GenBank in the Sequence read Archive under the experiment number SRX1054896.

Taxonomic affiliation of OTUs

In some groups of organisms, an argument for choosing a particular clustering threshold can be made by identifying a 'barcoding gap' (or barcoding window); that is, a gap between non overlapping distributions of taxa. This approach has, for example, been used in several fungal groups (Frøslev *et al.*, 2007; Jargeat *et al.*, 2010; Harder *et al.*, 2013). Unfortunately, our analysis of the 193 sequences assigned with a name shows that no such barcoding gap exists in Cercozoa (see also results and discussion), as the distribution of the interspecific diversity extends continuously all the way down to 0%. Hence, to eliminate as much artefactual diversity as possible without overcompromising phylogenetic resolution, we chose a conservative 5% level for OTU separation. We first blastn-searched the remaining 1143 OTUs locally against a custom cercozoan database (David Bass, in preperation). This enabled us to group them roughly into higher level groups approximating to Order/Class. We then used GenBank to blastn-search for representatives for these higher level groups. The top blastn hits, including as many named or otherwise characterized database sequences as possible, were retrieved and aligned with the OTUs generated in this study using the E-ins-i algorithm in MAFFT (Katoh and Standley, 2013) and phylogenetically analysed using RaxML BlackBox (Stamatakis *et al.*, 2008). We constructed an ML tree in RAxML BlackBox (using the GTRGAMMA) of the HTS sequences within a set of longer 18S reference sequences using the approach described in Dunthorn *et al.* (2014). An approach using only V4 for the whole analysis gave a similar result but with less backbone resolution

because this approach removes informative data from the analysis.

We used the resulting trees to assign OTUs as far as possible to named genera, higher level groups or environmental clades. We used a similar taxonomic framework as the one used in several recent studies of cercomonads, glissomonads, Granofilosea and other rhizopodial forms, as well as Cercozoa in general (Bass *et al.*, 2009a, b; Howe *et al.*, 2009, 2011). We included the % sequence identity in the OTU name to indicate the degree of similarity of the OTU to the best-matched database sequence. We made no attempt to assign any OTUs to species level. This may be possible for 100% complete 18S ribosomal DNA reads, however, in the vast majority of cases it would be misleading to imply such a high resolution. We considered OTUs assigned to the orders Euglyphida, Cryomonadida, and Thecofilosea, and the genera *Trinema, Rhogostoma, Corythion, Cyphoderia, Ovulinata, Euglypha, Trachelocorythion, Assulina, Pseudodifflugia, Tracheleuglypha* and *Ebria* as testate.

Results and discussion

Primer specificity, choice of amplification region and clustering threshold

The V4 regions is the longest of the 18S hypervariable regions (average V4 length = 253.5 bp) and the only one long enough for serious phylogenetic analysis. For this reason we would have considered this region preferable even if it had correlated slightly worse than a much shorter region. However, the sequence variability of the V4 region also turned out to correlate more strongly with the entire 18S of the 63 complete sequences than the other hypervariable regions ($R^2 = 0.89$, Figure 1). The V2 region also correlates well ($R^2 = 0.88$) but is much shorter with 161.8 bp on average. All other hypervariable regions are shorter than 100 bp on average and had R^2 values below 0.7 (Figure 1). This is important as the V9 region has been used extensively in HTS analyses of eukaryotic diversity (Amaral-Zettler *et al.*, 2009; Stoeck *et al.*, 2009; Behnke *et al.*, 2011) and has been suggested as a prime candidate especially for measuring protist lineage richness (Amaral-Zettler *et al.*, 2009). There were three reasons for recommending V9. First, it is comparatively short length (75–150 bp) enabled sequencing with first-generation Illumina/Solexa; second, it had an apparently lower sequencing error rate compared with the longer V4 region (Behnke *et al.*, 2011). Third, it appeared to yield less-biased results across the broad taxonomic groups with general eukaryotic primers (Stoeck *et al.*, 2010).

However, the rapid development in HTS analyses means that the read lengths of, for example, standard Illumina MiSEQ (2 × 250 bp) easily amplifies any hypervariable region. Furthermore, Behnke *et al.* (2011) found that the error rate in V4 could be greatly reduced by rigorous elimination of singletons and conservative clustering approaches, as we did here. Finally, our analysis of the hypervariable regions in Cercozoa shows that V4 correlates much better with

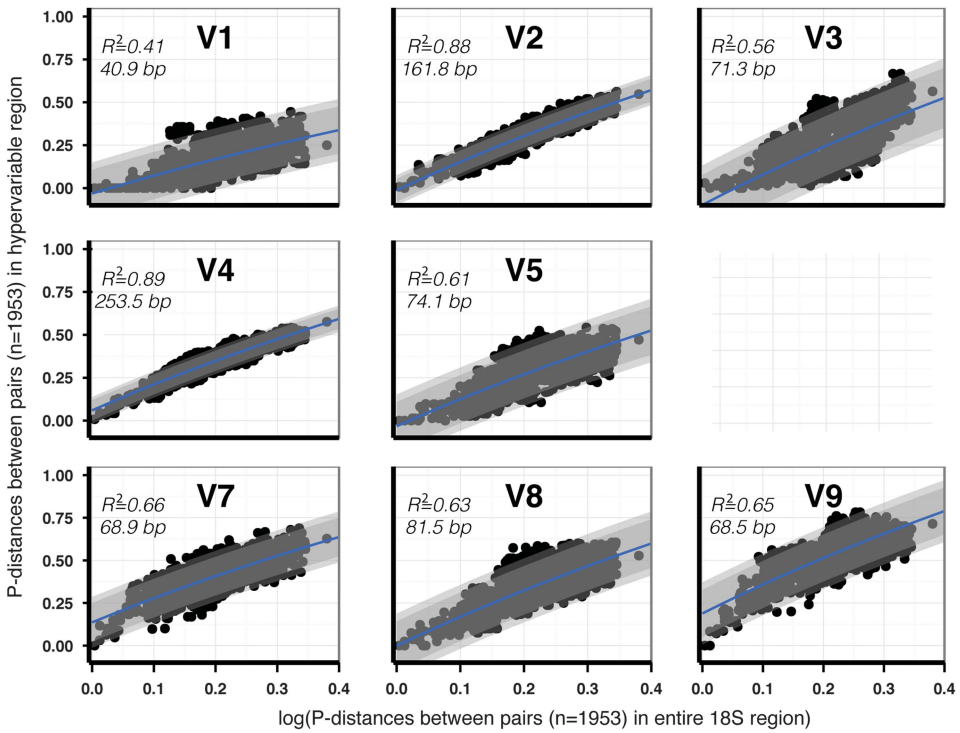

Figure 1 Linear regression of the percentage-wise DNA distance (uncorrected *P*-distance) of the eight hypervariable regions (V1–V9) against the entire 18S sequence from 63 complete cercozoan sequences. Dark grey shading = 99% confidence intervals, light grey shading = 95% confidence intervals.

the entire 18S than V9. This is in line with results found for other protist groups. Thus, the V4 seems more appropriate as barcode region than V9 in specific studies on ciliates (Dunthorn *et al.*, 2012), it has been suggested as a barcode in diatoms (Luddington *et al.*, 2012), and has been argued to be the best 'pre-barcoding' region for protists as a whole (Pawlowski *et al.*, 2012). In dinoflagellates, the V1–V4 regions outperform the other half of the 18S in taxonomic resolution (Ki, 2011). A recent study of oceanic eukaryotes with general V9 primers found that although they obtained a good representative sample across eukaryotic phyla, close to 90% of the protist sequences could not be assigned to genus or species level (de Vargas *et al.*, 2015). All in all, this suggests that, although V9 may be a good choice for studies that target eukaryotes using general primers, V4 (and adjacent regions) is a better choice for 18S-based diversity analyses that specifically target subgroups such as Cercozoa and several other protist groups.

However, we must stress that the V4 region is too conservative to separate several of the sequences assigned a species name in the 193-sequence set that we used. Many 18S-identical cercozoan strains exhibit consistently different phenotypes and ecological preferences, and one single 18S-type may harbour several different ITS1-types (Bass *et al.*, 2007). Hence, application of an OTU separation threshold of close to zero would be needed to get close to identifying species such as *Sandona mutans*, *S. dimutans, S. trimutans, S. tetramutans, S. pentamutans,* or *Bonamia* sp. ex *Ostrea chilensis and B.* sp. ex *Crassostrea ariakensis* in clustering analyses (Figure 2a). Accordingly, V4 contains no barcoding gap (Figure 2b); as the distribution of the pairwise distances between the 193 named known sequences extend continuously all the way below 1%. We would need ITS or another more sensitive marker to analyse diversity at this level. However, at present, the use of ITS for HTS studies in Cercozoa would be impractical because of the lack of good databases. At present, Genbank contains <50 full-length ITS sequences for named cercozoan species, and in an HTS analysis one would only be able to annotate a fraction of the resulting OTUs. For the time being, HTS community analyses of Cercozoa will therefore have to be based on 18S data. To compare such analyses, researchers must agree on a specific region and on a reasonable clustering threshold.

Our analysis shows that the decision about an appropriate clustering threshold in V4 region of Cercozoa can only be pragmatically based. Our clustering threshold of 5% is high. However, when we used the Bayesian-based ribosomal database

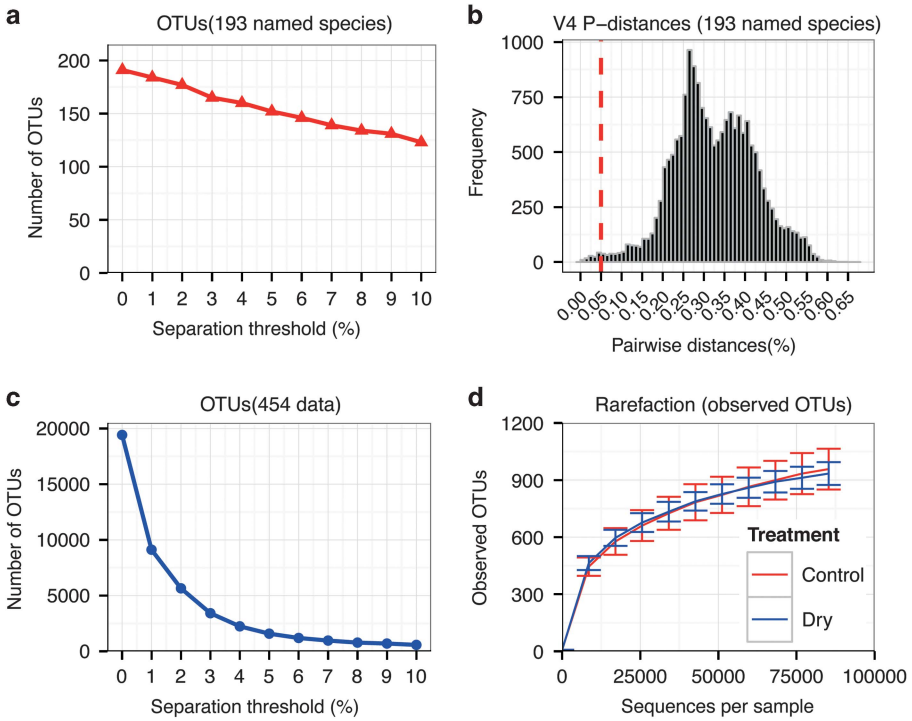

Figure 2 (**a**) Number of OTUs in 193 named cercozoan morpho-species as a function of OTU separation threshold of the V4 region. (**b**) Histogram of the pairwise comparisons of the V4 region from 193 named cercozoan morpho-species. The dashed red line shows our separation threshold of 5%. The bulk of the interspecific diversity is well above dissimilarities of 3 or 5%, but it extends all the way down to zero with no separation window. (**c**) Number of OTUs as a function of OTU separation threshold in the cleaned 454 data set of 443 350 V4 sequences. Choosing a separation threshold of 0.5–1% consistent with the morpho-species concepts would retrieve >9000 OTUs. (**d**) Rarefaction curve of the observed species (OTUs at 5% threshold) with error bars from combining the two drought and control plots, resampled to 85 305 sequences per tag. The samples are reasonably saturated with no difference in raw diversity between the two treatments.

project classifier (Wang *et al*., 2007) and the Silva database (at a 60% bootstrap support) to compare the 1585 OTUs at the 5% level and the 3421 OTUs returned by the 3% level, we found very little difference in the proportion of OTUs that could be assigned to the different taxonomic groups. For example, Silicofilosea comprises 17.1% of the reads at the 5%-threshold compared with 16.3% at the 3%-threshold, and the percentages of testate genera at the 5% and 3%-thresholds are 13.4 and 13.7%, respectively. As the choice of threshold had little apparent effect on our overall conclusions and could not be supported by taxonomy, we ultimately preferred the high value of 5%. This minimises the biases from sequencing artefacts in HTS sequencing (especially on the now defunct GS-FLX Titanium platform) that artificially inflate diversity. The clearly exponential decline in diversity as a function of clustering threshold in Figure 2c suggests that this effect is still substantial at the 3% level.

Still, most of the interspecific genetic variation is well above 5% in the diversity of the 193 known sequences (Figure 2b). The number of retained OTUs only declines from 165 to 152 when the dissimilarity threshold increases from 3 to 5% (Figure 2a). When the dissimilarity threshold increases from 0 to 5%, nearly all the 30–40 known species that lump in this process are congeneric and/or otherwise closely related. Hence, when we take into account both HTS error rates and genetic diversity in known reference species, we conclude that a dissimilarity threshold of 5% in cercozoan V4 sequence data is reasonable for capturing the main taxa and major functional groups.

Annotation, diversity and ecology

It is commonplace that HTS studies find a high unknown diversity of microorganisms in almost all habitats. Hence, we were not surprised to find a high level of unknown cercozoan diversity. However, we find it remarkable that even with our highly conservative data treatment, the number of OTUs on our single biotope still exceeds by a factor of more than four the number of Cercozoa that have been assigned a name based on their morphology. This is high for a single biotope by any standard or OTU separation threshold (Figure 2c). Moreover, ITS1-level diversity of Cercozoa is likely to be many times higher than that at 18S level. Thus, our study shows that an extrapolation of the existing practice of assigning names to Cercozoa harbours an immense potential for naming new hitherto unknown species.

The numbers of OTUs in the control and drought sites were almost identical (Figure 2d). However, our HTS data also allow us to evaluate the relative abundances of different Cercozoan genera on this soil type, as can be seen on the heatmaps in Figure 3. It has been suggested that because of their small size (5–10 μm), Glissomonadida are the quantitatively dominant Cercozoa in soils (Howe *et al*., 2009), and

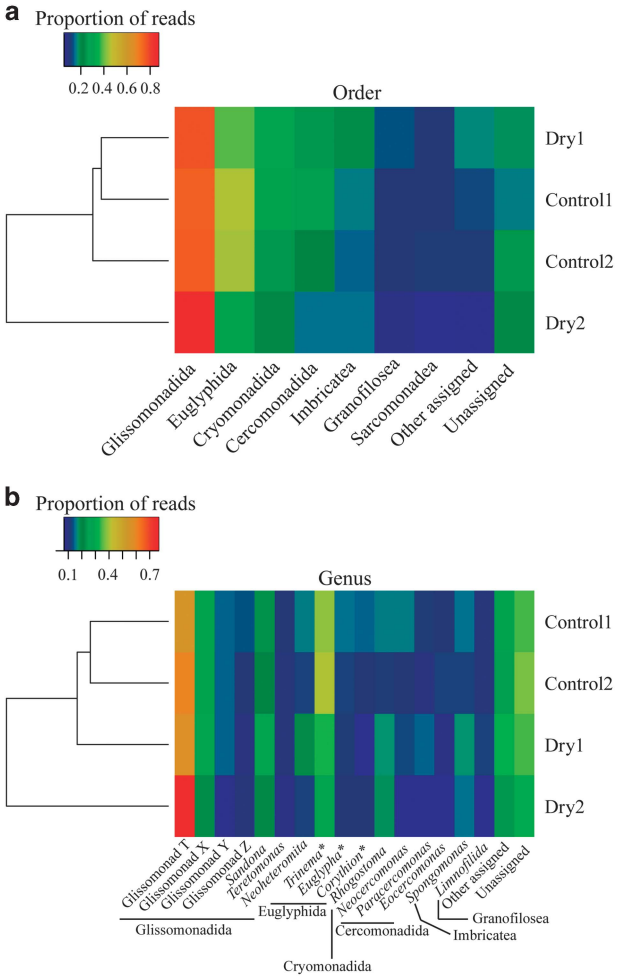

Figure 3 Heatmaps displaying the taxonomic abundance and distribution of the sequences ($n = 85\,305$) between the four sample sites, at (**a**) order and (**b**) genus level. The heatmaps are drawn using the heatmap.2 function of the gplots package, and dendrogram distances are based on Euclidean distance. Taxa constituting <1% of the total taxa have been lumped into the 'other assigned' category. Testate taxa are marked with an asterisk. Glissomonads and Euglyphids are by far the most abundant Cercozoa in all samples. Of all taxa, the genus *Trinema* shows the largest relative change in response to the drought treatment.

indeed, glissomonads are the most abundant group in our analysis (Figure 3). Among the taxa that could be identified confidently, the euglyphid testate amoeba *Trinema* was the most abundant genus (Figure 3b). *Trinema* also presented the proportionally largest changes in abundance of cercozoan genera between the two treatments. Overall, testate amoebae constituted ~19.9 and 17.7% on the two control sites and 15.9 and just below 9% on the drought sites, respectively (Figure 3a). Thus, although our results are not significant, possibly due to the small sample size, they suggest a negative response of testate amoebae to drought in accordance with previous findings (Lousier, 1974a, b; Wilkinson and Mitchell, 2010).

Testate amoebae are generally large (members of *Trinema* usually in the length range of 30–100 μm, or

6–12 times longer than most glissomonads). Hence, even though cell number of glissomonads in the samples exceed that of testate amoebae by a factor of three or four, the same will not be true for their biomass. Since-cell size of protozoan predators is one of the most important morphological factors determining the bacterial community composition (Glucksman *et al*., 2010), our results suggest a quantitative importance of cercozoan testate genera that merits more attention.

Conclusions

We found the V4 hypervariable region to be the best single region in 18S ribosomal DNA for exploring cercozoan diversity by HTS analyses. Further our analysis using this region showed a high diversity at the studied sites. Cercozoan species have traditionally been defined by morphology supported by 18S Sanger sequencing, and if this concept is to be taken to represent the best estimate of the 'true diversity', our analysis of the V4 region from cultured cercozoans species demonstrates that more variable taxonomic markers should be investigated. However, as our conservative treatment of V4 HTS sequence data reveal the existence of a large unknown diversity in just one single biotope, the diversity revealed by less-conservative markers would be quite high indeed.

Our results suggest that the soil protozoan diversity *per se* is largely resilient to the levels of drought expected from the climate change scenarios projected for the Northern temperate latitudes over the 21st century, at least over shorter timeframes. However, they also suggest that among Cercozoa, testate forms are the most sensitive to drought and hence good indicator organisms to detect soil community changes in the early stages of climate change. Finally, our data indicate that a sampling of close to 10^5 sequences is necessary to reach sampling saturation in HTS studies of Cercozoa in soil samples. This should be taken into account in future HTS studies, especially those that wish to use general eukaryotic primers to gain an understanding of eukaryotic subgroups.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank the CLIMAITE project and the generous funding of this by the Villum Kann Rasmussen foundation, Marie Merrild for collecting samples, and Anders Michelsen and Merian Schouw Haugwitz for providing valuable suggestions and information about the study site. We are also grateful for the technical assistance provided by Karin Vestberg and Anette Løth, without which this study would not have seen the light of day, and for the bioinformatic support from Lars H Hansen. RR and FE were supported by the Carlsberg Foundation (2012_01_0677) and by the Danish Council for Independent Research (DFF-4002-00274); DB was supported by NERC Standard Research Grant (NE/H009426/1); CBH and FE were supported by The Danish Council for Strategic Research grant (2104-08-0012, MIRESOWA).

Author contributions

CBH, RR and FE conceived and designed the experiments. CBH and RR carried out the experiments. CBH, RR, AB, DB and WAS analysed results. CBH, RR, DB, AB and FE wrote the paper.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

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.

Baldauf SL, Roger A, Wenk-Siefert I, Doolittle W. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**: 972–977.

Bamforth SS. (1995). Interpreting soil ciliate biodiversity. *Plant Soil* **170**: 159–164.

Bass D, Cavalier-Smith T. (2004). Phylum-specific environmental DNA analysis reveals remarkably high global biodiversity of Cercozoa (Protozoa). *Int J Syst Evol Micr* **54**: 2393–2404.

Bass D, Richards TA, Matthai L, Marsh V, Cavalier-Smith T. (2007). DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evol Biol* **7**: 162.

Bass D, Chao EE, Nikolaev S, Yabuki A, Ishida K, Berney C *et al.* (2009a). Phylogeny of novel naked Filose and Reticulose Cercozoa: Granofilosea cl. n. and Proteomyxidea revised. *Protist* **160**: 75–109.

Bass D, Howe AT, Mylnikov AP, Vickerman K, Chao EE, Edwards Smallbone J *et al.* (2009b). Phylogeny and classification of Cercomonadida (Protozoa, Cercozoa): *Cercomonas*, *Eocercomonas*, *Paracercomonas*, and *Cavernomonas* gen. nov. *Protist* **160**: 483–521.

Bass D, Yabuki A, Santini S, Romac S, Berney C. (2012). *Reticulamoeba* is a long-branched Granofilosean that is missing from sequence databases. *PLoS One* **7**: e49090.

Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, Knight R. (2013). Global biogeography of highly diverse protistan communities in soil. *ISME J* **7**: 652–659.

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T. (2011). Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ Microl* **13**: 340–349.

Berney C, Romac S, Mahe F, Santini S, Siano R, Bass D. (2013). Vampires in the oceans: predatory cercozoan amoebae in marine habitats. *ISME J* **7**: 2387–2399.

Booth RK. (2001). Ecology of testate amoebae (Protozoa) in two Lake Superior coastal wetlands: implications for paleoecology and environmental monitoring. *Wetlands* **21**: 564–576.

Booth RK. (2008). Testate amoebae as proxies for mean annual water-table depth in Sphagnum-dominated peatlands of North America. *J Quaternary Sci* **23**: 43–57.

Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW. (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat Methods* **9**: 425–426.

Burki F, Kudryavtsev A, Matz MV, Aglyamova GV, Bulman S, Fiers M *et al.* (2010). Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. *BMC Evol Biol* **10**: 377.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Cavalier-Smith T, Chao EE-Y. (2003). Phylogeny and classification of phylum Cercozoa (Protozoa). *Protist* **154**: 341–358.

Cavalier-Smith T, Oates B, Lewis R, Chao EE, Bass D. (2009). *Helkesimastix marina* n. sp. (Cercozoa: Sainouroidea superfam. n.) a gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist* **160**: 452–479.

Chantangsi C, Leander BS. (2010a). Ultrastructure, life cycle and molecular phylogenetic position of a novel marine sand-dwelling cercozoan: *Clautriavia biflagellata* n. sp. *Protist* **161**: 133–147.

Chantangsi C, Leander BS. (2010b). An SSU rDNA barcoding approach to the diversity of marine interstitial cercozoans, including descriptions of four novel genera and nine novel species. *Int J Syst Evol Micr* **60**: 1962–1977.

Charman DJ, Warner BG. (1992). Relationship between testate amoebae (Protozoa: Rhizopoda) and microenvironmental parameters on a forested peatland in northeastern Ontario. *Can J Zool* **70**: 2474–2482.

Chatelain AP, Meisterfeld R, Roussel-Delif L, Lara E. (2013). Sphenoderiidae (fam. nov.), a new clade of euglyphid testate amoebae characterized by small, round scales surrounding the aperture. *Protist* **164**: 782–792.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Dunthorn M, Klier J, Bunge J, Stoeck T. (2012). Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J Eukaryot Microbiol* **59**: 185–187.

Dunthorn M, Otto J, Berger SA, Stamatakis A, Mahé F, Romac S *et al.* (2014). Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol Biol Evol* **31**: 993–1009.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Eisenhauer N, Cesarz S, Koller R, Worm K, Reich PB. (2012). Global change belowground: impacts of elevated CO2, nitrogen, and summer drought on soil food webs and biodiversity. *Glob Change Biol* **18**: 435–447.

Ekelund F, Rønn R, Griffiths BS. (2001). Quantitative estimation of flagellate community structure and diversity in soil samples. *Protist* **152**: 301–314.

Ekelund F, Daugbjerg N, Fredslund L. (2004). Phylogeny of *Heteromita*, *Cercomonas* and *Thaumatomonas* based on SSU rDNA sequences, including the description of *Neocercomonas jutlandica* sp. nov., gen. nov. *Eur J Protistol* **40**: 119–135.

Ekelund F, Rønn R. (1994). Notes on protozoa in agricultural soil with emphasis on heterotrophic flagellates and naked amoebae and their ecology. *FEMS Microbiol Rev* **15**: 321–353.

Frøslev TG, Jeppesen TS, Læssøe T, Kjøller R. (2007). Molecular phylogenetics and delimitation of species in *Cortinarius* section Calochroi. (Basidiomycota, Agaricales) in Europe. *Mol Phylogenet Evol* **44**: 217–227.

Geisen S, Tveit AT, Clark IM, Richter A, Svenning MM, Bonkowski M *et al.* (2015). Metatranscriptomic census of active protists in soils. *ISME J* **9**: 2178–2190.

Glucksman 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.

Griffiths BS, Ritz K, Bardgett RD, Cook R, Christensen S, Ekelund F *et al.* (2000). Ecosystem response of pasture soil communities to fumigation-induced microbial diversity reductions: an examination of the biodiversity-ecosystem function relationship. *Oikos* **90**: 279–294.

Harder CB, Læssøe T, Frøslev TG, Ekelund F, Rosendahl S, Kjøller R. (2013). A three-gene phylogeny of the *Mycena pura* complex reveals 11 phylogenetic species and shows ITS to be unreliable for species identification. *Fungal Biol* **117**: 764–775.

Heger TJ, Mitchell EA, Todorov M, Golemansky V, Lara E, Leander BS *et al.* (2010). Molecular phylogeny of euglyphid testate amoebae (Cercozoa: Euglyphida) suggests transitions between marine supralittoral and freshwater/terrestrial environments are infrequent. *Mol Phylogenet Evol* **55**: 113–122.

Heger TJ, Pawlowski J, Lara E, Leander BS, Todorov M, Golemansky V *et al.* (2011). Comparing potential COI and SSU rDNA barcodes for assessing the diversity and phylogenetic relationships of cyphoderiid testate amoebae (Rhizaria: Euglyphida). *Protist* **162**: 131–141.

Hess S, Sausen N, Melkonian M. (2012). Shedding light on vampires: the phylogeny of vampyrellid amoebae revisited. *PLoS One* **7**: e31165.

Hoppenrath M, Leander BS. (2006). Dinoflagellate, Euglenid, or Cercomonad? The ultrastructure and molecular phylogenetic position of *Protaspis grandis* n. sp. *J Eukaryot Microbiol* **53**: 327–342.

Howe AT, Bass D, Scoble JM, Lewis R, Vickerman K, Arndt H *et al.* (2011). Novel cultured protists identify deep-branching environmental DNA clades of cercozoa: New Genera *Tremula*, *Micrometopion*, *Minimassisteria*, *Nudifila Peregrinia*. *Protist* **162**: 332–372.

Howe AT, Bass D, Vickerman K, Chao EE, Cavalier-Smith T. (2009). Phylogeny, taxonomy, and astounding genetic diversity of glissomonadida ord.nov., the dominant gliding zooflagellates in soil (Protozoa:Cercozoa). *Protist* **160**: 159—189.

Ihaka R, Gentleman R. (1996). R: A language for data analysis and graphics. *J Comput Graph Stat* **5**: 299–314.

Jargeat P, Martos F, Carriconde F, Gryta H, Moreau PA, Gardes M. (2010). Phylogenetic species delimitation in ectomycorrhizal fungi and implications for barcoding: the case of the *Tricholoma scalpturatum* complex (Basidiomycota). *Mol ecol* **19**: 5216–5230.

Katoh K, Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Ki J-S. (2011). Hypervariable regions (V1–V9) of the dinoflagellate 18S rRNA using a large dataset for marker considerations. *J Appl Phycol* **24**: 1035–1043.

Lara E, Heger TJ, Mitchell EA, Meisterfeld R, Ekelund F. (2007). SSU rRNA reveals a sequential increase in shell complexity among the euglyphid testate amoebae (Rhizaria: Euglyphida). *Protist* **158**: 229–237.

Larsen KS, Andresen LC, Beier C, Jonasson S, Albert KR, Ambus P *et al.* (2011). Reduced N cycling in response to elevated CO2, warming, and drought in a Danish heathland: Synthesizing results of the CLIMAITE project after two years of treatments. *Glob Change Biol* **17**: 1884–1899.

Lousier JD. (1974a). Response of soil Testacea to soil moisture fluctuations. *Soil Biol Biochem* **6**: 235–239.

Lousier JD. (1974b). Effects of experimental soil moisture fluctuations on turnover rates of Testacea. *Soil Biol Biochem* **6**: 19–26.

Luddington IA, Kaczmarska I, Lovejoy C. (2012). Distance and character-based evaluation of the V4 region of the 18S rRNA gene for the identification of diatoms (Bacillariophyceae). *PLoS One* **7**: e45664.

Nielsen PL, Andresen LC, Michelsen A, Schmidt IK, Kongstad J. (2009). Seasonal variations and effects of nutrient applications on N and P and microbial biomass under two temperate heathland plants. *Appl Soil Ecol* **42**: 279–287.

Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, Amaral-Zettler L *et al.* (2011). Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One* **6**: e18169.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C *et al.* (2012). CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol* **10**: e1001419.

Rozen S, Skaletsky H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365–386.

Sandon H. (1927). *The composition and distribution of the protozoan fauna of the soil*. Oliver and Boyd: Edinburgh, UK.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microl* **75**: 7537–7541.

Schmitt A, Glaser B. (2011). Organic matter dynamics in a temperate forest soil following enhanced drying. *Soil Biol Biochem* **43**: 478–489.

Stamatakis A, Hoover P, Rougemont J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* **57**: 758–771.

Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A *et al.* (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol* **7**: 72.

Stoeck T, Bass D, Nebel M, Christen R, Jones MD, Breiner HW *et al.* (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.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Wilkinson DM, Mitchell EAD. (2010). Testate amoebae and nutrient cycling with particular reference to soils. *Geomicrobiol J* **27**: 520–533.

Wylezich C, Mylnikov AP, Weitere M, Arndt H. (2007). Distribution and phylogenetic relationships of freshwater Thaumatomonads with a description of the new species Thaumatomonas coloniensis n. sp. *J Eukaryot Microbiol* **54**: 347–357.

Yabuki A, Ishida K-I. (2011). Mataza hastifera n. g., n. sp.: a possible new lineage in the Thecofilosea (Cercozoa). *J Eukaryot Microbiol* **58**: 94–102.

Yu ZT, Mohn WW. (1999). Killing two birds with one stone: simultaneous extraction of DNA and RNA from activated sludge biomass. *Can J Microbiol* **45**: 269–272.

Zhu F, Massana R, Not F, Marie D, Vaulot D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *Fems Microbiol Ecol* **52**: 79–92.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)