Received Date : 13-Feb-2015
Revised Date : 28-Apr-2015
Accepted Date : 06-May-2015
Article type : Original Article

Not all are free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa

S. Geisen^{1, 2, *}, I. Laros³, A. Vizcaíno⁴, M. Bonkowski², G.A. de Groot³

¹ Department of Terrestrial Ecology, Netherlands Institute for Ecology (NIOO-KNAW), Wageningen, The Netherlands

² Department of Terrestrial Ecology, Institute of Zoology, University of Cologne, Germany

³ ALTERRA – Wageningen UR, P.O. Box 47, 6700 AA, Wageningen, The Netherlands

⁴ AllGenetics, Ed. de Servicios Centrales de Investigación, Campus de Elviña s/n, E-15071 A Coruña, Spain

* For correspondence. E-mail s.geisen@nioo.knaw.nl; Tel. (+31) 317473613.

Short title: Protist parasites associated to soil metazoa

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.13238

This article is protected by copyright. All rights reserved.

Keywords: Soil protists; Soil Metazoan Parasites; Potential parasites; Apicomplexa; 454

Metabarcoding; High-throughput sequencing;

Abstract

Protists, the most diverse eukaryotes, are largely considered to be free-living bacterivores, but vast numbers of taxa are known to parasitize plants or animals. High-throughput sequencing (HTS) approaches now commonly replace cultivation-based approaches in studying soil protists, but insights into common biases associated to this method are limited to aquatic taxa and samples.

We created a mock community of common free-living soil protists (amoebae, flagellates, ciliates), extracted DNA and amplified it in the presence of metazoan DNA using 454 HTS. We aimed at evaluating whether HTS quantitatively reveals true relative abundances of soil protists and to investigate whether the expected protist community structure is altered by the co-amplification of metazoan-associated protist taxa. Indeed, HTS revealed fundamentally different protist communities from those expected. Ciliate sequences were highly overrepresented, while those of most amoebae and flagellates were underrepresented or totally absent. These results underpin the biases introduced by HTS that prevent reliable quantitative estimations of free-living protist communities. Furthermore, we detected a wide range of non-added protist taxa likely introduced along with metazoan DNA, which altered the protist community structure. Among those, 20 taxa most closely resembled parasitic, often pathogenic taxa. Therewith, we provide the first HTS data in support of classical observational studies that showed that potential protist parasites are hosted by soil metazoa.

Taken together, profound differences in amplification success between protist taxa and an inevitable co-extraction of protist taxa parasitizing soil metazoa obscure the true diversity of free-living soil protist communities.

Introduction

Soils are among the most complex habitats on earth. Consequently, soils host a huge diversity of functionally different organisms that in turn play pivotal roles in soil functioning. Protists are by far the most diverse and (together with fungi) the most abundant soil eukaryotes (Ekelund & Rønn 1994; Finlay *et al.* 2000; Foissner 1987). Traditionally, protists have been termed protozoa and included as part of metazoa in ecological studies (Foissner 1999a, b, 2008). However, their extreme diversity is illustrated by thousands of protist taxa spreading paraphyletically over the entire eukaryotic tree of life (Adl *et al.* 2005; Adl *et al.* 2012; Cavalier-Smith 2003). Protists hold a key functional position in soil food webs as the major grazers of bacteria due to short turnover-rates and high abundances of up to 100,000 individuals per gram of soil (Clarholm 1981; Darbyshire 1994; Ekelund & Rønn 1994; Geisen *et al.* 2014a).

Despite their functional importance in soils, protists are arguably the least studied group of soil organisms due to several obstacles in studying them (Caron *et al.* 2008; Foissner 1999b). Especially the smaller, more abundant and often transparent flagellates and naked amoebae remain masked by soil particles, while only few groups with larger cells, and more or less constant, distinctive morphological characters that are readily observable, such as shown by ciliates and testate amoebae can directly be observed in suspended soils (Foissner 1999b). Therefore, cultivation-based approaches (e.g. Darbyshire 1974, Geisen *et al.* 2014a) are still indispensable to estimate the abundance and diversity of the numerically dominant protists. Due to profound differences in cultivability, these techniques strongly underestimate the true diversity of soil protists (Caron *et al.* 1989; Ekelund & Rønn 1994; Pawlowski *et al.* 2014). Furthermore, morphological identification of protist taxa is extremely time-consuming and correct identification, especially of flagellates and naked amoebae, remains an expert task (Foissner 1999b; Smirnov *et al.* 2008).

Molecular techniques gave hope to circumvent most of these obstacles (Caron *et al.* 2008). High-throughput sequencing (HTS) approaches have largely replaced morphological methods and become the gold standard to study protist community structures (Adl *et al.* 2014; Bik *et al.* 2012; Stoeck *et al.* 2014). Most common HTS approaches apply a metabarcoding strategy to investigate protists, using specific primers that target highly variable barcoding regions in order to disentangle multiple taxa in a single DNA extract. For most protist groups, the most commonly targeted DNA region is the 18S rDNA as (i) it contains both highly conserved regions that help primer annealing and variable regions that allow detailed taxonomic classification; (ii) it amplifies easily due to high DNA copy-numbers, and (iii) it covers most of the protist diversity in existing online databases, and therefore enables accurate sequence annotations to described species (Guillou *et al.* 2013; Pawlowski *et al.* 2012; Yilmaz *et al.* 2013).

Unfortunately, it is impossible to design specific primers exclusively targeting protists, since protists are taxonomically intermingled with plants, fungi and animals. Accordingly, two different approaches have been applied: most studies employed “general” eukaryotic primers aiming at targeting the majority of protists, especially in aquatic environments (Amaral-Zettler *et al.* 2009; Fonseca *et al.* 2014; Zhan *et al.* 2013). Due to the high load of co-targeted fungi and metazoa (Baldwin *et al.* 2013; Bates *et al.* 2013; Lentendu *et al.* 2014), this approach only recently became feasible for application on soils, thanks to the increase of sequence output from HTS. Alternatively, primers targeting specific protist groups (e.g. ciliates or Cercozoa) largely avoid the amplification of non-target organisms and allow higher sequencing depths and taxonomic resolution of targeted clades (Lentendu *et al.* 2014; Stoeck *et al.* 2010).

Some key problems remain in protist studies: (i) no commonly accepted species concept is applicable for protists (Boenigk *et al.* 2012; Coleman 2002; Schlegel & Meisterfeld 2003); (ii) existing databases suffer from wrongly annotated sequences, and the published protist sequences provide, at best, a fragmentary overview of the expected protist diversity (Guillou *et al.* 2013; Taib *et al.* 2013); (iii) most

Accepted Article

HTS studies, especially metabarcoding approaches, suffer from a wide variety of PCR induced biases such as selectively over- or under-amplification of specific target groups due to unavoidable primer-selectivity and lengths differences in the targeted amplification region, by the formation of chimeras in the amplification step and even the type of polymerase and numbers of PCR cycles impact the outcome (Adl *et al.* 2014; Ahn *et al.* 2012; Behnke *et al.* 2011; Berney *et al.* 2004; Jeon *et al.* 2008; Stoeck *et al.* 2006; von Wintzingerode *et al.* 1997); (iv) the stringent filtering and clustering steps needed in HTS studies to remove high-loads of false positive sequences introduced in the sequencing step also eliminate a substantial part of the true diversity in the investigated samples, which is especially a problem at lower sequencing depths such as in studies applying 454 pyrosequencing (Behnke *et al.* 2011; Kunin *et al.* 2010; Quince *et al.* 2011).

Nevertheless, HTS metabarcoding approaches provide unprecedented possibilities to study soil protists as they allow deciphering the hidden diversity of often uncultivable taxa (Epstein & López-García 2008). HTS studies have revealed the presence of high numbers of protist clades such as Apicomplexa that have never been cultivated and are known to be obligatory parasites, both in aquatic systems (Leander *et al.* 2006; Stoeck *et al.* 2007) and more recently soils (Bates *et al.* 2013; Geisen *et al.* 2015; Ramirez *et al.* 2014). From morphological studies it is known that diverse parasitic protist clades commonly infect soil metazoa and could therefore be the source of the parasites found in HTS studies. Among them are gregarine apicomplexans that infect earthworms (Edwards & Bohlen 1996), microsporidia parasitize collembola and nematodes (Rusek 1998; Troemel *et al.* 2008), whereas oomycetes act as plant pathogens (Altizer *et al.* 2003; Latijnhouwers *et al.* 2003). Therefore, major efforts are still needed to identify whether parasites can actually contribute to the community of “free-living” soil protists to eventually being able to disentangle truly free-living from potentially parasitic endosymbionts taxa.

In a HTS metabarcoding approach, using a variety of mock communities of diverse, but well-characterized soil protists and soil metazoa, we investigated the biases introduced by differential PCR amplification between added taxa. Further we hypothesized that metazoan-associated protists

would be recovered resulting in an increased diversity and change of the expected protist community composition.

Materials and Methods

Mock community creation

Protist cultures established in previous cultivation efforts (Geisen *et al.* 2014a; Geisen *et al.* 2014b; Geisen *et al.* 2014d) were stored in 50 ml cultivation flasks (Sarstedt, Germany) containing 0.15 % wheat grass (WG) medium (Geisen *et al.* 2014d) at 15 °C, and used to establish a mixed protist community. We chose a diverse range of very common soil protists covering a wide phylogenetic diversity of eight orders in seven classes (Table 1), including the ciliate *Colpoda steinii*, the flagellates *Cercomonas plasmodialis* and *Neobodo designis*, and taxonomically diverse amoebae: *Allovahlkampfia* sp., *Saccamoeba* sp., *Vannella* sp., *Acanthamoeba castellanii*, and *Flamella* sp.

Monoclonal protist cultures were individually enumerated using a Neubauer improved counting chamber (Roth, Germany) under an inverted Nikon Diaphot phase contrast microscope (Nikon, Japan) at 400x magnification. We decided not to add equal amounts of the respective protist taxa but a community more closely resembling those found in soil environments (Domonell *et al.* 2013; Geisen *et al.* 2014a). Therefore we added 200,000 protists in a 2 ml volume of culture medium composed of 1.0 % ciliates, 38.5 % flagellates and 60.5 % amoebae (Table 1).

Genomic DNA was then isolated directly from this protist mixture using the guanidine isothiocyanate protocol (Maniatis *et al.* 1982) with slight modifications; briefly, supernatant of WG medium was carefully discarded after centrifugation at 15,000 g for 90 s. The pellet containing all solid material largely composed of protist cell remnants was washed twice with sterile WG medium, which was subsequently replaced by 100 µl guanidine isothiocyanate. Subsequent steps were performed according to the cited protocol (Maniatis *et al.* 1982).

The resulting purified protist DNA was mixed with DNA extracted from five different groups of soil metazoa: microarthropods, enchytraeids, earthworms, and nematodes. Mixed-species communities of microarthropods, enchytraeids and nematodes were extracted in spring 2011, by means of Tullgren and wet funnel extractions, from soil samples taken at a permanent agricultural field near Lusignan (ACBB Lusignan long-term observatory, France). Earthworm specimens came from these same samples, as well as from soil samples taken in spring 2011 in three permanent grasslands near Ede, the Netherlands. Nematodes, enchytraeids and earthworms were washed and left to defecate in tap water. Microarthropods were washed in 96% ethanol. All specimens were then stored in 96% ethanol prior to DNA extraction, which was performed using commercial kits (DNAeasy Blood&Tissue Kit (Qiagen) for microarthropods and earthworms; PureLink DNA Mini Kit (Life Technologies) for nematodes; REALPURE Microspin Kit (Real) for enchytraeids) following the distributor's protocols. DNA of these metazoan groups was mixed with protist DNA in three different mock communities (MC; pools 1-3) containing different relative abundances of the respective groups; pool 1 contained an equal mix of all groups (i.e. 16.7 % protist), pool 2 an enrichment of earthworm DNA (72 % earthworm and 5 % protist DNA) and pool 3 dominated by nematodes and enchytraeids, with protist DNA representing 16 % of total DNA added. Exact relative abundances per group per MC are shown in Fig. S1. Metazoan DNA in changing ratios were added to increase complexity and to identify how non-protist DNA that is commonly co-extracted and co-amplified from soils affects amplification success for known protist taxa and to get an idea of how unattended extraction of high amounts of DNA from e.g. earthworms and the simultaneous reduction sequences assigned to protists can distort the protist community composition (Lentendu *et al.* 2014).

DNA amplification and sequencing

The mixed DNA was subject to HTS metabarcoding using a so-called general eukaryotic primer set to investigate to which degree individual protist species can be recovered and whether relative sequence abundances matched the true abundance of taxa added. For that, the 5' end of the SSU

Accepted Article

rDNA was amplified using forward (Fw) primer EUK20f (5'—TGC CAG TAG TCA TAT GCT TGT – 3') with the reverse (Rev) primer EUK302r+3 (5' – ACC AGA CTT GYC CTC CAA T - 3'), both modified from Euringer & Lueders (2008), to target a 500-700 bp long region of the 18S rDNA. This part of the 18S rDNA contains the highly variable regions V1-V3, enabling high taxonomic discrimination between taxa (Hugerth *et al.* 2014).

Three different 5-bp long multiplex identifiers (MID) were added to the forward and reverse primer sequences (ATCTC, ATCGC, and TACTC respectively for pools 1-3) to allow demultiplexing resulting reads after sequencing. Polymerase chain reactions (PCRs) were performed in a final volume of 25 µl, containing 12.5 µl of Supreme NZYtaq Green PCR Master Mix (NZY Tech, Lisbon, Portugal), 1.25 µl of each primer (10 µM), 2.5 µl of template DNA, and 7.25 µl of ddH₂O.

The following PCR conditions were applied: initial denaturation at 95 °C for 5 min was followed by 35 cycles of denaturation at 95 °C for 30 s, annealing at 48 °C for 60 s and elongation at 72 °C for 2 min with a final extension for 5 min at 72 °C. PCR products were purified using the PCRExtract Mini kit (5 Prime GmbH, Hilden, Germany), quantified using Qubit (Life Technologies) and then pooled in equal concentrations. Library preparation and Roche GS FLX 454 pyrosequencing was performed by Genoscreen (Lille, France) following the manufacturers' guidelines. A total of 57832 sequence reads were generated on 1/16 of a FLX titanium gasket.

Bioinformatic analyses

The metabarcoding pipeline Usearch v7.0.1001 was used for the bioinformatic analyses. Raw sff output from the sequencer was converted to fasta data and phred-like quality values, and subsequently combined into one fastq file using the Python script *faqual2fastq*. The resulting reads were sorted to samples by their primer sequence and labelled with their multiplex identifier (MID) using the *fastq_strip_barcode_relabel2* Python script, removing both sequences from the reads in the process. A stringency of 2 mismatching bases for the primer and no mismatching bases in the MID was used. Since sequencing was done both with the forward and the reverse primers, sorting

and labelling was done both for the forward and the reverse primer, resulting in two files that were processed simultaneously for the steps below.

Using the *fastq_filter* command, reads were cut to different lengths (450, 400, 350, 300 and 250 bp) to evaluate potentially better taxon assignment at longer read lengths with reduced sequence numbers due to increasing sequence errors with extending sequence length. Reads of low quality (maximum expected error of 0.5) at the individual read-cut-offs were discarded and longer reads truncated to the cut-off size. After testing with several fragment lengths (see results) we continued with a threshold size of 250 bases.

Reads were made unique and labelled with the quantity they occurred in using the *derep_fulllength* command, singletons were removed and the remaining unique sequences sorted according to their quantity using the *sortbysize* command. From these abundance-sorted sequences, Operational Taxonomic Units (OTUs) were generated using the *cluster_otus* command with a 97 % similarity threshold. Subsequently, all trimmed sequences (including singletons) were mapped against this set of OTUs using the *usearch_global* command with a 97 % identity threshold. Matrices containing read abundance per OTU per MID were generated using the *uc2otutab* Python script.

OTUs were pre-assigned to a rough taxonomic classification taking the top hit of a Basic Local Alignment Search Tool (BLAST; algorithm v 2.2.23; Altschul *et al.* 1990) search against the NCBI database using an e-value cut-off of $1e^{-5}$, an identity cut-off of 90 %, and a coverage cut-off of 80 % of the query sequence covered in the alignments. NCBI annotations rather than the indisputably more quality controlled SILVA (Yilmaz *et al.* 2013) or PR2 databases (Guillou *et al.* 2013) were used, as the NCBI database is by far largest. All assigned eukaryotic 18S rDNA sequences were then manually verified by BLASTn searches against the NCBI GenBank nt database for correct taxonomic assignments. For that, the best 50 hits were analysed and OTUs conservatively classified if resulting hits showed consistent taxonomic patterns. Correct assignments were further performed using alignment-based phylogenetic analyses of different eukaryotic supergroups, detailed below.

The sequence data generated in this study has been submitted to the ENA databases under accession number PRJEB9172.

Alignments and phylogenetic analyses

As we obtained higher and more diverse OTU numbers than we included in the original protist DNA pools (Table 1) we added sequences of the 5 - 10 best BLASTn hits that were non-identical to each other, including at least one formally described species. We manually aligned those sequences in separate alignments for individual supergroups in Seaview 4 (Gouy *et al.* 2010). The quality of sequences obtained was again controlled for potentially missed chimeras by visual screening of the alignment in search for contradictory sequence signatures (as described in Berney *et al.* (2004)) but none were found.

We performed maximum likelihood phylogenetic analyses on all protist supergroups to identify the taxonomic diversity of the OTUs obtained. Focusing on amoebozoan sequences as the most diverse group added in the original protist MC, more detailed phylogenetic analyses were performed. All eleven amoebozoan OTUs obtained from Fw and Rev primers were aligned together with 56 published 18S rDNA sequences representing close BLASTn hits of all OTUs. First, we performed maximum likelihood analyses using RAxML v. 8.0.2 (Stamatakis 2014) on 1,232 unambiguously aligned positions excluding variable regions as described in Geisen *et al.* (2014b). The GTR+ γ +I model of evolution was applied, as proposed by jModeltest v. 2.1.3 under the Akaike Information Criterion (Darriba *et al.* 2012), the γ approximated by 25 categories running 1000 non-parametric bootstrap pseudoreplicates. Bayesian phylogenetic analyses were performed using Mr Bayes v. 3.2.2 with GTR+ γ +I model of evolution and 8 categories (Ronquist & Huelsenbeck 2003). Two runs of four simultaneous Markov chains were performed for 2,000,000 generations (with the default heating parameters) and sampled every 100 generations; convergence (average deviation of split frequencies < 0.01 of the two runs) was reached after 550,000 generations. The remaining 14,500 trees were used to build a consensus tree.

Results

Overview of sequence reads obtained

In total, 29,266 high-quality sequences were obtained applying a stringent cut-off after 250 bp. Read quality dropped sharply with increasing sequence lengths; 22,413 / 5,295 / 1,353 and 209 sequences remained when cut after 300 / 350 / 400 and 450 bp, respectively. Among those, almost twofold reverse reads than forward reads were obtained (Fig. S2). This difference increased towards the higher cut-off lengths, with reverse reads outnumbering forward reads by 2.7×, 4.6× and 9.8× at sequence cut-offs lengths of 300, 350 and 400 bp, respectively. Sequencing depth differed between pools and was highest in pool 1 (Fw: 4406, Rev: 9215) and pool 2 (Fw: 4414, Rev: 9257) and lowest in pool 3 (Fw: 725, Rev: 1569).

Taxonomic resolution depended on the read lengths, with longer reads enabling higher taxonomic classification (data not shown). However, since deep-taxonomic classification was not the aim of this study, we focused on the more abundant reads (applying a cut-off at 250 bp).

OTU analyses of added protist taxa

All protists were recovered in the sequence data except for the heterolobosean amoeba

Allovahlkampfia sp. Among the added Amoebozoa, *Saccamoeba* sp., *Vannella* sp. and *Flamella* sp. were recovered in all pools from both sequencing directions, while *Acanthamoeba castellanii* was only retrieved from RevOTUs. Both flagellate taxa *Cercomonas plasmodialis* and *Neobodo designis* as well as the ciliate *Colpoda steinii* were recovered in all pools from both sequencing directions.

Relative sequence abundance of individual taxa differed strongly from expected relative abundances based on the added DNA (Fig. 1). Especially *C. steinii*, *Flamella* sp. and *Saccamoeba* sp. were

overrepresented by factors of 13.1, 6.3 and 2.2, respectively. In contrast, *Vannella* sp. (23 % reduction) and *A. castellanii* (54 % reduction) were underrepresented, the latter being entirely absent in FwOTUs. The two flagellates, *C. plasmodialis* (58 % reduction) and *N. designis* (93 % reduction), were also highly underrepresented.

Differences between mock communities

The addition of soil metazoan DNA influenced the efficiency of protist sequence recovery. Clear differences in both the abundance and composition of protist sequences were found between the three mock communities (Fig. S3). Only 8.5 % of all forward sequences and 6.7 % of all reverse sequences resembled protists in pool 1, containing 16.7 % of protist-specific DNA. Similarly, pool 3 yielded lower protist sequences as expected from the DNA content in the mix (16.0 %) with 6.3 % and 5.4 % of all sequences obtained from forward and reverse primers (FwOTUs and RevOTUs, respectively). Interestingly, protist-specific relative sequence abundance in pool 2 was higher than expected from the DNA added (5 %) with 9.4 % and 6.1 % of all sequences assigned into FwOTUs and RevOTUs, respectively (Fig. S3).

Protist community structures changed in line with the differences observed in the mock communities. At a higher taxonomic level, only Ciliophora increased in pool 1 and pool 3, while Amoebozoa, Heterolobosea, Euglenozoa, and Rhizaria were lower than expected (Fig. S3). Similarly, Heterolobosea and Euglenozoa were underrepresented in pool 2, but the relative abundance of Amoebozoa and Rhizaria was as expected (Fig. S3).

All protist clades recovered contained a range of taxonomically diverse taxa (Fig. 2). Amoebozoan sequences could be assigned into 5 distinct FwOTUs and 13 RevOTUs, while only 4 amoebozoan taxa were added in the original pool (Fig. 2, Table 1). A targeted phylogenetic analyses of retrieved and published amoebozoan sequences showed that the OTUs branched in diverse lineages across the supergroup Amoebozoa (Fig. 3). In addition to the expected four OTUs, six placed in four different genera inside the class Tubulinea, one with so-far uncultivated taxa in the class Discosea and three resembled taxa of the class Variosea.

Instead of detecting one OTU in each of the following clades, we retrieved a total of 11 different rhizarian FwOTUs and 8 RevOTUs, as well as 15 ciliate FwOTUs and 13 RevOTUs (Fig. 2). The short sequencing length prevented high-resolution taxonomic assignment in these groups. Two different euglenozoan OTUs were recovered, both resembling *Neobodo designis*, but with nucleotide composition differing by 8.9 % and resembling distinct published sequences of different cultures of *N. designis*. While *Allovahlkampfia* sp. was not found, a sequence perfectly matching the heterolobosean *Vahlkampfia inornata* was recovered. In addition to these added protist clades, other groups were found, among them non-ciliate Alveolates (Apicomplexa and *Colpodella*), stramenopiles and chlorophytes (Fig. 2, Fig. S3).

As expected, the majority of OTUs and sequences were assigned as metazoa, but also sequences of fungi and plants were retrieved (Table S1). A summary of all OTUs obtained from Fw and Rev sequencing direction are shown in Table S2 (Fw) and S3 (Rev).

Potential parasitic taxa

Manual BLASTn searches, conducted to get a better understanding of potential functions, revealed 20 OTUs that most closely resembled potential parasites of taxonomically diverse organisms, including protists (Apicomplexa, Cercozoa, Ciliophora and oomycete Stramenopiles), fungi, and insects (Table 2). Pool 2, enriched in earthworms, hosted the highest relative amount of parasites followed by pool 1 and pool 3 (2.0 %, 1.7 % and 1.3 % of all sequences, respectively). A relationship between earthworms and parasites was further supported by a clear correlation between relative sequence abundances of earthworms and all potential parasites (Fig. 4). In terms of relative abundance, fungi dominated, followed by apicomplexans and ciliates.

Protist OTUs assigned as potential parasites often most closely resembled uncultivated species from environmental surveys (Lesaulnier *et al.* 2008) or from studies targeting animal parasites (Clopton 2009; Valles & Pereira 2003). Four protist OTUs among the supergroup Stramenopiles closely resembled the apicomplexan genera *Gregarina* and *Cryptosporidium* (Table 2). Also 4 OTUs of potential ciliate parasites were identified (Table 2). One OTU resembled the predominantly plant pathogenic oomycete genus *Phytophthora*.

Discussion

Studies applying HTS DNA metabarcoding to describe soil protist communities are rapidly increasing due to plunging sequencing costs. More detailed knowledge on biases of differential amplification and the relative contribution of potential protist parasites accompanying soil metazoa is urgently needed to fully understand the community structure and eventually functions of soil protists.

Sequencing biases change the real community structure of soil protists

Accepted Article

Only few studies on aquatic ciliates have combined morphological and DNA metabarcoding approaches on protists (Bachy *et al.* 2013; Medinger *et al.* 2010; Stoeck *et al.* 2014). Similarly, attempts to evaluate differences between expected and observed protist communities by sequencing an artificially assembled protist community with known compositions have only been conducted for specific aquatic protist clades, such as diatoms and foraminifera (Kermarrec *et al.* 2013; Weber & Pawlowski 2013). All these studies concordantly revealed that morphological quantification and relative sequence abundance of protists differ profoundly mainly due to PCR-induced biases, even within closely related aquatic protist taxa.

Our data clearly support these results, and show that they also apply to studies on protists in soils. Primer-based HTS introduces several biases and therefore cannot be considered to reflect a quantitative estimate of soil protist communities. For example, mismatches in the primer region are known to reduce PCR amplification and might be the reason for the absence of *Allovahlkampfia* sp. and the decrease of *C. plasmodialis* especially in FwOTUs due to the mismatches in the Fw primer. However, the fact that both primers fully matched the sequence of *A. castellanii* and this taxon was absent in RevOTUs demonstrates that even highly abundant protist taxa may not be retrieved, despite exact primer matching of sequences of targeted taxa. The length of the targeted barcoding region also affects amplification success, providing an explanation for the reduction of *Allovahlkampfia* sp., *A. castellanii*, and *Neobodo designis* that share the longest amplicon region of organisms among the mock community in this study. Enhanced amplification due to short 18S rDNA sequences explains the overrepresentation of ciliates in this and environmental HTS studies (Baldwin *et al.* 2013; Bates *et al.* 2013). Amplification rates are also increased by the presence of multiple nuclei and high genomic 18S rDNA copies (Prokopowich *et al.* 2003), both the case for ciliates (Gong *et al.* 2013; Medinger *et al.* 2010) and *Flamella* sp. (Kudryavtsev *et al.* 2009).

These data clearly demonstrate the dependency of PCR-amplification rates on the specific “architecture” of the 18S rDNA of individual protist taxa and provides experimental evidence that

many amoebae and flagellates are underrepresented or completely absent in DNA based HTS studies due to negative discrimination during PCR amplification (Berney *et al.* 2004; Epstein & López-García 2008). Primer-free metatranscriptomic approaches will, with future ongoing decreases in sequencing costs, likely replace these metabarcoding approaches, which have recently been shown to more realistically depict soil protist communities (Geisen *et al.* 2015; Urich *et al.* 2008)

Some additional general methodological problems were uncovered in this study. Interestingly, amplification products obtained from the reverse primer doubled the sequences of the forward primer (Fig. S2). While differences in amplification success between various primer pairs have been documented (Jeon *et al.* 2008; Stoeck *et al.* 2006), we are not aware of studies showing differences between sequencing direction within the same primer pair. We suspect these differences to occur in the final sequencing step of the HTS, because previous steps such as successful PCR depend on an equal amplification of the entire amplicon. A superior amplification of Rev primer-targets due to more suitable sequencing conditions and wider specificity range to target PCR products might explain the observed differences between forward and reverse sequencing products. Further, automated OTU assignment methods that are routinely adopted in HTS studies are problematic and would have resulted in major differences in taxon assignments in this study due to errors and gaps in available databases (Guillou *et al.* 2013; Taib *et al.* 2013). Only evaluation of each OTU individually allowed us to reliably assign OTUs to taxonomy. Thus, profound taxonomic knowledge is required for detailed analyses of protist sequencing studies (Cristescu 2014; Kvist 2013). However, the tedious approach taken here is not feasible for increased sequencing output underpinning the importance of cultivation-based approaches to fill gaps in reference databases. So-called ultrahigh sequencing platforms such as Illumina's MiSeq and HiSeq already provide orders of magnitude more sequences at similar or even lower costs than the outrunning 454 HTS technique (Caporaso *et al.* 2012). Further developments in increases of sequence lengths and error-reductions will eventually allow higher taxonomic resolutions and with the highly increased sequence output allow simultaneous deep-investigations of high sample numbers. This promising outlook will, without doubts, stimulate

research on (parasitic) protists in various habitats and hosts (Bik *et al.* 2012) eventually leading to entirely novel insights in this little studied field.

As we have not aimed at identifying and deciphering individual closely related species, we applied a highly stringed sequence cut-off at 250 bp to avoid increasing sequence error rates (Quince *et al.* 2011) and an OTU clustering of 97 %. These short sequences and conservative OTU clustering, however, undoubtedly lumped together different protist taxa as some even share almost full identity of 18S rDNA sequences (Boenigk *et al.* 2012; Geisen *et al.* 2014c). In line with general practice, we excluded singletons to reduce errors, which likely resulted in filtering out more true protist taxa (Zhan *et al.* 2013), suggesting that we most likely identified only a subset of the protist diversity hidden in our mixed DNA pools. Our primers targeted the first part of the 18S rDNA including the highly variable V2 region (Hadziavdic *et al.* 2014), that has often been used in molecular studies targeting protists in soils (Euringer & Lueders 2008; Lesaulnier *et al.* 2008). The V4 as the recommended barcoding region for protists might have provided higher taxonomic resolution (Pawlowski *et al.* 2012), but we did not aim at high taxonomic resolution as we included taxonomically distinct protist taxa.

A wide range of non-added, often potentially parasitic taxa associated with soil metazoa

We are the first to use a HTS metabarcoding approach to directly reveal the presence of potentially parasitic, diverse eukaryotic protists associated with soil metazoa. The suitability of HTS metabarcoding targeting eukaryotic endoparasites has successfully been proven in ticks, where specific primers for the apicomplexan parasites *Babesia* and *Theileria* were applied (Bonnet *et al.* 2014). Earthworm-associated bacteria have recently been targeted, revealing some earthworm-specific groups such as symbiotic *Verminephrobacter* sp. and parasitic *Serratia* sp. (Pass *et al.* in press).

The increased number of protist sequences obtained from the earthworm enriched pool 2 compared with pool 1 and pool 3 (Fig. 4) and the resulting close correlation of earthworm and sequences of potential parasites indicates that earthworms are major reservoirs for protist taxa. Most of the diversity and abundance of parasitic protists, found in both aquatic (Moreira & López-García 2003; Stoeck *et al.* 2007) and soil surveys (Bates *et al.* 2013; Ramirez *et al.* 2014), might therefore have originated from metazoan endosymbionts. This is of special importance when DNA extraction contains parts of an earthworm that will then strongly influence the taxonomic composition of “soil” protist communities. This scenario is not unlikely, as metazoan sequences, especially from earthworms, are common and abundant in HTS when general eukaryotic primers are used (Baldwin *et al.* 2013; Lentendu *et al.* 2014). Additional protist taxa were likely added by other metazoa, as protist communities differed between pools. Many of these protists most likely represent obligatory parasitic taxa in Apicomplexa, Cercozoa, Ciliophora and oomycete Stramenopiles, but also potentially parasitic fungi and animals were identified (Table 2). A range of parasites such as gregarine protists, nematodes and mites are known to infect earthworms (Edwards & Bohlen 1996), while microsporidian pathogens infect collembola (Rusek 1998) and nematodes (Troemel *et al.* 2008). Despite the fact that a range of protist parasites, such as species of the genera *Cryptosporidium* and *Giardia*, also found in this study, are hazardous both to a range of animals and humans (Salyer *et al.* 2012; Savioli *et al.* 2006), the identity and diversity of most of these organisms remains unknown. Even the protist diversity in the supposedly well-studied human intestine just starts to be disentangled (Parfrey *et al.* 2014). Taken into account that parasites can have profound impacts on host behaviour and fitness (Lefèvre *et al.* 2012) with the potential to control host communities (Moreira & López-García 2003) and change food web structures (Cirtwill & Stouffer 2015), the ecological significance of parasites of soil metazoa needs more attention.

Accepted Article

It has to be mentioned that sequence data cannot provide ultimate proof about functions of individual taxa, and taxa identified as potential parasites might have positive or negative effects on the potential host. Within this study we are confident that the taxa identified as potential parasites are parasitic or at least directly associated to metazoan hosts due to the controlled setup. Commonly performed environmental surveys using HTS, however, only provide an overview of OTUs present in a sample without revealing their microhabitats and detailed combinations of cultivation based with functional studies are needed to determine the true identity and functions of these OTUs (Geisen *et al.* 2015).

In addition, other non-added protists were recovered that most closely resembled non-parasitic taxa in the distinct eukaryotic supergroups Amoebozoa, Archaeplastida, Excavata and SAR. We suspect them to represent either (facultative) parasites or amphizoic organisms colonizing soil organisms. Among them were colpodellids, the closest relatives to parasitic apicomplexans, which are suggested to be free-living predators of other protists (Leander & Keeling 2003; Simpson & Patterson 1996). Recently, however, potentially parasitic colpodellids were documented (Yuan *et al.* 2012), but it remains unclear if the colpodellids found here were parasites or amphizoic organisms that infected metazoan or protist cells. Potentially amphizoic protists have been shown parasitizing fish (Dyková *et al.* 2010; Dyková *et al.* 1998; Dyková *et al.* 1999) and human (Martinez & Visvesvara 1997; Schuster 2002; Visvesvara *et al.* 2007), but have also been retrieved from earthworm guts (Dixon 1975; Edwards & Bohlen 1996; Satchell 1983). As earthworm DNA was isolated after defecation and storage in water from parts of the skin, the surface of which was washed but not sterilized prior to DNA extraction, the exact location of these taxa remains elusive.

Conclusions and outlook

The differences observed in our study between the expected and recovered community composition of protists reveals that the enormous diversity of paraphyletic protist taxa renders both absolute and relative quantification of soil protists impossible. Further, our results reveal that HTS data needs careful interpretation in determining protist community compositions. Ciliates are generally overrepresented in sequence numbers, while many amoeba and flagellate taxa are underrepresented or missing from sequence data. Especially very complex systems such as soils introduce other biases due to the presence of non-targeted taxa such as metazoa, but also fungi and plants. DNA from these taxa could alter the community of “free-living” protist taxa and introduce a range of other, potentially parasitic taxa. We here found a plethora of sequences assigned to protists other than those added that infiltrated the community estimates along with DNA extracted from common soil metazoa. Many of these taxa most closely resembled obligate parasites. These findings have major consequences for past and future soil HTS approaches that target soil protists. Soil protist taxa can clearly be separated into free-living taxa and those tightly associated with soil metazoa as potential parasites, but the ecological importance of especially the latter remains entirely unknown. Future investigations should aim at deciphering metazoan species-specific parasites and how much parasites contribute to the actual diversity of “soil” protists, while functional studies should aim at deciphering the importance of parasites in structuring soil food webs and to act as potential human pathogens.

Acknowledgements

This work was supported by the European Commission within the FP7 project EcoFINDERS (FP7-264465). The data presented here were gathered as part of a larger series of experiments dedicated to the development of methodology for high-throughput molecular characterization of soil biodiversity. The authors would like to sincerely thank the colleagues involved in this task, especially

Marc Bueé, Dalila Costa, Bryan Griffiths, Francis Martin, Rüdiger Schmelz and Dorothy Stone. Further support came from an ERC grant (ERC-Adv 260-55290 (SPECIALS)) of Wim van der Putten.

All authors declare no conflict of interests.

References

- Adl SM, Habura A, Eglit Y (2014) Amplification primers of SSU rDNA for soil protists. *Soil Biol. Biochem.* **69**, 328–342.
- Adl SM, Simpson AGB, Farmer MA, *et al.* (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* **52**, 399–451.
- Adl SM, Simpson AGB, Lane CE, *et al.* (2012) The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59**, 429–514.
- Ahn J-H, Kim B-Y, Song J, Weon H-Y (2012) Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J. Microbiology* **50**, 1071–1074.
- Altizer S, Harvell D, Friedle E (2003) Rapid evolutionary dynamics and disease threats to biodiversity. *Trends Ecol. Evol.* **18**, 589–596.
- 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.
- Bachy C, Dolan JR, Lopez-Garcia P, Deschamps P, Moreira D (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *ISME J.* **7**, 244–255.
- Baldwin DS, Colloff MJ, Rees GN, *et al.* (2013) Impacts of inundation and drought on eukaryote biodiversity in semi-arid floodplain soils. *Mol. Ecol.* **22**, 1746–1758.
- Bates ST, Clemente JC, Flores GE, *et al.* (2013) Global biogeography of highly diverse protistan communities in soil. *ISME J.* **7**, 652–659.
- Behnke A, Engel M, Christen R, *et al.* (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ. Microbiol.* **13**, 340–349.
- Berney C, Fahrni J, Pawlowski J (2004) How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol.* **2**, 13.
- Bik HM, Porazinska DL, Creer S, *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* **27**, 233–243.
- Boenigk J, Ereshefsky M, Hoef-Emden K, Mallet J, Bass D (2012) Concepts in protistology: species definitions and boundaries. *Eur. J. Protistol.* **48**, 96–102.
- Bonnet S, Michelet L, Moutailler S, *et al.* (2014) Identification of parasitic communities within European ticks using next-generation sequencing. *PLoS Negl. Trop. Dis.* **8**, e2753.
- 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.
- Caron DA, Davis PG, Sieburth JM (1989) Factors responsible for the differences in cultural estimates and direct microscopical counts of populations of bacterivorous nanoflagellates. *Microb.*

- Ecol.* **18**, 89-104.
- Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB (2008) Protists are microbes too: a perspective. *ISME J.* **3**, 4-12.
- Cavalier-Smith T (2003) Protist phylogeny and the high-level classification of Protozoa. *Eur. J. Protistol.* **39**, 338-348.
- Cirtwill AR, Stouffer DB (2015) Concomitant predation on parasites is highly variable but constrains the ways in which parasites contribute to food web structure. *J. Anim. Ecol.* **84**, 734-744.
- Clarholm M (1981) Protozoan grazing of bacteria in soil—impact and importance. *Microb. Ecol.* **7**, 343-350.
- Clopton RE (2009) Phylogenetic relationships, evolution, and systematic revision of the septate gregarines (Apicomplexa: Eugregarinorida: Septatorina). *Comp. Parasitol.* **76**, 167-190.
- Coleman AW (2002) Microbial eukaryote species. *Science* **297**, 337.
- Cristescu ME (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends Ecol. Evol.* **29**, 566-571.
- Darbyshire J (1994) *Soil Protozoa* CAB International, Wallingford.
- Darbyshire JF, Whitley RE, Graebes MP, Inkson RHE (1974) A rapid micromethod for estimating bacterial and protozoan populations in soil. *Rev. Ecol. Biol. Sol* **11**, 465-475.
- Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772-772.
- Dixon RF (1975) The astomatous ciliates of British earthworms. *J. Biol. Educ.* **9**, 29-39.
- Domonell A, Brabender M, Nitsche F, Bonkowski M, Arndt H (2013) Community structure of cultivable protists in different grassland and forest soils of Thuringia. *Pedobiologia* **56**, 1-7.
- Dyková I, Kostka M, Pecková H (2010) *Grellamoeba robusta* gen. n., sp. n., a possible member of the family Acramoebidae Smirnov, Nassonova et Cavalier-Smith, 2008. *Eur. J. Protistol.* **46**, 77-85.
- Dyková I, Lom J, Machačková B (1998) *Cochliopodium minus*, isolated from organs of perch *Perca fluviatilis*. *Dis Aquat Org* **34**, 205-210.
- Dyková I, Lom J, Schroeder-Diedrich JM, Booton GC, Byers TJ (1999) *Acanthamoeba* strains isolated from organs of freshwater fishes. *J. Parasitol.* **85**, 1106-1113.
- Edwards CA, Bohlen PJ (1996) *Biology and ecology of earthworms* Springer, New York.
- 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.
- Epstein S, López-García P (2008) “Missing” protists: a molecular prospective. *Biodivers. Conserv.* **17**, 261-276.
- Euringer K, Lueders T (2008) An optimised PCR/T-RFLP fingerprinting approach for the investigation of protistan communities in groundwater environments. *J. Microbiol. Meth.* **75**, 262-268.
- Finlay BJ, Black HJJ, Brown S, et al. (2000) Estimating the growth potential of the soil protozoan community. *Protist* **151**, 69-80.
- Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Prog. Protistol.* **2**, 69-212.
- Foissner W (1999a) Protist diversity: estimates of the near-imponderable. *Protist* **150**, 363-368.
- Foissner W (1999b) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agr. Ecosyst. Environ.* **74**, 95-112.
- Foissner W (2008) Protist diversity and distribution: some basic considerations. *Biodivers. Conserv.* **17**, 235-242.
- Fonseca VG, Carvalho GR, Nichols B, et al. (2014) Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. *Global Ecol. Biogeogr.* **23**, 1293-1302.
- Geisen S, Bandow C, Römbke J, Bonkowski M (2014a) Soil water availability strongly alters the community composition of soil protists. *Pedobiologia* **57**, 205-213.
- Geisen S, Fiore-Donno AM, Walochnik J, Bonkowski M (2014b) *Acanthamoeba* everywhere: high diversity of *Acanthamoeba* in soils. *Parasitol. Res.* **113**, 3151-3158.
- Geisen S, Kudryavtsev A, Bonkowski M, Smirnov A (2014c) Discrepancy between species borders at

- morphological and molecular levels in the genus *Cochliopodium* (Amoebozoa, Himatismenida), with the description of *Cochliopodium plurinucleolum* n. sp. *Protist* **165**, 364-383.
- Geisen S, Tveit AT, Clark IM, *et al.* (2015) Metatranscriptomic census of active protists in soils. *ISME J.*
- Geisen S, Weinert J, Kudryavtsev A, *et al.* (2014d) Two new species of the genus *Stenamoeba* (Discosea, Longamoebia): cytoplasmic MTOC is present in one more amoebae lineage. *Eur. J. Protistol.* **50**, 153-165.
- Gong J, Dong J, Liu X, Massana R (2013) Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* **164**, 369-379.
- Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221-224.
- 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.
- Hadziavdic K, Lekang K, Lanzén A, *et al.* (2014) Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One* **9**, e87624.
- Hugerth LW, Muller EEL, Hu YOO, *et al.* (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One* **9**, e95567.
- Jeon S, Bunge J, Leslin C, *et al.* (2008) Environmental rRNA inventories miss over half of protistan diversity. *BMC Microbiol. Biol.* **8**, 222.
- Kermarrec L, Franc A, Rimet F, *et al.* (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Mol. Ecol. Res.* **13**, 607-619.
- Kudryavtsev AA, Wylezich C, Schlegel M, Walochnik J, Michel R (2009) Ultrastructure, SSU rRNA gene sequences and phylogenetic relationships of *Flamella* Schaeffer, 1926 (Amoebozoa), with description of three new species. *Protist* **160**, 21-40.
- Kunin V, Engelbrektsen A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12**, 118-123.
- Kvist S (2013) Barcoding in the dark?: a critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. *Mol. Phylogenet. Evol.* **69**, 39-45.
- Latijnhouwers M, de Wit PJGM, Govers F (2003) Oomycetes and fungi: similar weaponry to attack plants. *Trends Microbiol.* **11**, 462-469.
- Leander BS, Keeling PJ (2003) Morphostasis in alveolate evolution. *Trends Ecol. Evol.* **18**, 395-402.
- Leander BS, Lloyd SAJ, Marshall W, Landers SC (2006) Phylogeny of marine gregarines (Apicomplexa) — *Pterospira*, *Lithocystis* and *Lankesteria* — and the origin(s) of coelomic parasitism. *Protist* **157**, 45-60.
- Lefèvre T, Chiang A, Kelavkar M, *et al.* (2012) Behavioural resistance against a protozoan parasite in the monarch butterfly. *J. Anim. Ecol.* **81**, 70-79.
- 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.
- Lesaulnier C, Papamichail D, McCorkle S, *et al.* (2008) Elevated atmospheric CO₂ affects soil microbial diversity associated with trembling aspen. *Environ. Microbiol.* **10**, 926-941.
- Maniatis T, Fritsch EF, Sambrook J (1982) *Molecular Cloning, A Laboratory Manual* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Martinez AJ, Visvesvara GS (1997) Free-living, amphizoic and opportunistic Amebas. *Brain Pathology* **7**, 583-598.
- Medinger R, Nolte V, Pandey RV, *et al.* (2010) Diversity in a hidden world: potential and limitation of next generation sequencing for surveys of molecular diversity of eukaryotic microorganisms.

- Mol. Ecol.* **19**, 32-40.
- Moreira D, López-García P (2003) Are hydrothermal vents oases for parasitic protists? *Trends Parasitol.* **19**, 556-558.
- Parfrey LW, Walters WA, Lauber CL, *et al.* (2014) Defining the diversity of microbial eukaryotic communities in the mammalian gut within the context of environmental eukaryotic diversity. *Front. Microbiol.* **5**.
- Pass DA, Morgan AJ, Read DS, *et al.* (in press) The effect of anthropogenic arsenic contamination on the earthworm microbiome. *Environ. Microbiol.*
- Pawlowski J, Audic S, Adl S, *et al.* (2012) CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol.* **10**, e1001419.
- Pawlowski J, Lejzerowicz F, Esling P (2014) Next-generation environmental diversity surveys of Foraminifera: preparing the future. *Biol. Bull.* **227**, 93-106.
- Prokopowich CD, Gregory TR, Crease TJ (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46**, 48-50.
- Quince C, Lanzén A, Davenport R, Turnbaugh P (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**, 38.
- Ramirez KS, Leff JW, Barberán A, *et al.* (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proc. Biol. Sci.* **281**.
- Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.
- Rusek J (1998) Biodiversity of Collembola and their functional role in the ecosystem. *Biodivers. Conserv.* **7**, 1207-1219.
- Salyer SJ, Gillespie TR, Rwego IB, Chapman CA, Goldberg TL (2012) Epidemiology and molecular relationships of *Cryptosporidium* spp. in people, primates, and livestock from western Uganda. *PLoS Negl. Trop. Dis.* **6**, e1597.
- Satchell JE (1983) Earthworm microbiology. In: *Earthworm Ecology* (ed. Satchell JE), pp. 351-364. Springer Netherlands.
- Savioli L, Smith H, Thompson A (2006) *Giardia* and *Cryptosporidium* join the neglected diseases initiative'. *Trends Parasitol.* **22**, 203-208.
- Schlegel M, Meisterfeld R (2003) The species problem in protozoa revisited. *Eur. J. Protistol.* **39**, 349-355.
- Schuster FL (2002) Cultivation of pathogenic and opportunistic free-living amebas. *Clin. Microbiol. Rev.* **15**, 342-354.
- Simpson AGB, Patterson DJ (1996) Ultrastructure and identification of the predatory flagellate *Colpodella pugnax* Cienkowski (Apicomplexa) with a description of *Colpodella turpis* n. sp. and a review of the genus. *Syst. Parasitol.* **33**, 187-198.
- Smirnov AV, Nasonova ES, Cavalier-Smith T (2008) Correct identification of species makes the amoebozoan rRNA tree congruent with morphology for the order Leptomyxida Page 1987; with description of *Acramoeba dendroidea* n. g., n. sp., originally misidentified as '*Gephyramoeba* sp.'. *Eur. J. Protistol.* **44**, 35-44.
- Stamatakis A (2014) RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313.
- 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.
- Stoeck T, Breiner H-W, Filker S, *et al.* (2014) A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology. *Environ. Microbiol.* **16**, 430-444.
- Stoeck T, Hayward B, Taylor GT, Varela R, Epstein SS (2006) A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist* **157**, 31-43.
- Stoeck T, Kasper J, Bunge J, *et al.* (2007) Protistan diversity in the Arctic: a case of paleoclimate shaping modern biodiversity? *PLoS One* **2**, e728.
- Taib N, Mangot J-F, Domaizon I, Bronner G, Debroas D (2013) Phylogenetic affiliation of SSU rRNA

- genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. *PLoS One* **8**, e58950.
- Troemel ER, Félix M-A, Whiteman NK, Barrière A, Ausubel FM (2008) Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. *PLoS Biol.* **6**, e309.
- Urich T, Lanzén A, Qi J, *et al.* (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**, e2527.
- Valles SM, Pereira RM (2003) Use of ribosomal DNA sequence data to characterize and detect a neogregarine pathogen of *Solenopsis invicta* (Hymenoptera: Formicidae). *J. Invertebr. Pathol.* **84**, 114-118.
- Visvesvara GS, Moura H, Schuster FL (2007) Pathogenic and opportunistic free living amoebae: *Acanthamoeba* spp., *Balamuthia mandrillaris*, *Naegleria fowleri*, and *Sappinia diploidea*. *FEMS Immunol. Med. Mic.* **50**, 1-26.
- von Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.* **21**, 213-229.
- Weber AAT, Pawlowski J (2013) Can abundance of protists be inferred from sequence data: a case study of Foraminifera. *PLoS One* **8**, e56739.
- Yilmaz P, Parfrey LW, Yarza P, *et al.* (2013) The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res.*
- Yuan CL, Keeling PJ, Krause PJ, *et al.* (2012) *Colpodella* spp.-like parasite infection in woman, China. *Emerg. Infect. Dis* **18**, 125-127.
- Zhan A, Hulák M, Sylvester F, *et al.* (2013) High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods Ecol. Evol.* **4**, 558-565.

Data Accessibility

Sequence data generated in this study were deposited under ENA accession number *PRJEB9172*. The amoebozoan alignment, fasta files with sequences for the OTUs and bioinformatic scripts were uploaded to Dryad (doi:10.5061/dryad.3m680).

Author Contributions

SG and AdG designed the study. AV generated the sequence data, which were analysed by IL and SG.

SG, MB and AdG wrote the manuscript, aided by all authors.

Table legends

Table 1: Protist taxa included in this study with their morphological and taxonomic identity; # = number of individuals; base pair (bp) differences with Fw primer EUK20F and Rev primer EUK302r+3.

Table 2: OTUs of potentially parasitic taxa with their taxonomic affinity, highest maximum identity (MI) with described taxa (underlined: uncultivated sequence in case of higher MI than described species), Accession numbers (#) of best BLASTn hit, identity of best BLASTn hit, reported host of best BLASTn hit and sequence # as well as relative abundances (rel. abu.) obtained.

Figure legends

Fig. 1: Relative sequence abundances of protist taxa included in the study and their higher taxonomic affinities from both sequencing directions (forward = Fw; reverse = Rev) in three mock communities.

Fig. 2: Expected (left bar) and obtained protist OTUs from all pools individually (Fw1-Fw3, Rev1-Rev3) and summarized from each sequencing direction (Fw all, Rev all); numbers above columns indicate total OTU number of the respective dataset; Fw: forward sequencing primer; Rev: reverse sequencing primer; dotted bars: added protist OTUs; filled bars: additionally recovered protist OTUs; (a): protist taxon added; (n): non-added, newly recovered protist taxa; note: reduced OTU numbers in Fw3 and Rev3 are in line with lower sequencing depth of pool 3.

Fig. 3: Phylogenetic analysis on Amoebozoa including OTUs recovered in this study and sequences associated with those OTUs. In total, 67 sequences with 1,232 unambiguously aligned positions were used; the tree is unrooted; values higher than 50 for maximum likelihood analyses (left) and 0.50 for Bayesian analyses (right) are shown. Black circles indicate full support; **bold**: OTUs and respective amoebozoan genus added that were recovered in OTUs; **underlined bold**: OTUs and respective amoebozoan genera that were not included in the original pools; note: of the four added amoebozoan genera, *Acanthamoeba* was only recovered by one RevOTU.

Fig. 4: Earthworm specific OTUs (1st y-axis) in relative abundance of DNA added (stars) and from OTUs obtained (squares: Fw primer; diamonds: Rev primer) and relative sequence abundance of OTUs of potential parasites (2nd y-axis; crosses: Fw primer; triangles: Rev primer).

Supplementary legends

Table S1: Number (#), percentage (%) of OTUs and % of sequences assigned to non-protist organisms.

Table S2: Detailed OTU table of all 18S rDNA sequences obtained (Seq #) from all three mock

communities with the forward primer (Fw) and their manual identified taxonomic classification; taxonomic classification assigned only to the deepest reliable level; Bit scores and % identity (% Id) with best match only to reliably assigned taxa without taking unknown or uncultivated sequences into account; Pot. Parasitic = Potential parasitic taxon.

Table S3: Detailed OTU table of all 18S rDNA sequences obtained (Seq #) from all three mock communities with the reverse primer (Rev) and their manual identified taxonomic classification; taxonomic classification assigned only to the deepest reliable level; Bit scores and % identity (% Id) with best match only to reliably assigned taxa without taking unknown or uncultivated sequences into account; Pot. Parasitic = Potential parasitic taxon.

Fig. S1: Relative abundances of DNA isolated from communities of six taxonomic groups.

Fig. S2: Differences in obtained sequences from forward (Fw; grey squares) and reverse (Rev; black diamonds) sequencing direction both strongly dropping in numbers with increasing OTU lengths. Regression lines are based on logistic regression in SPSS (Rev: $R^2 = 0.984$; Fw: $R^2 = 0.999$).

Fig. S3: Relative sequence abundance (%) of protists expected and recovered from forward (Fw) and reverse (Rev) sequencing directions in the three mock communities (pools 1-3).