METHODS

Code availability. All codes used here can be found in HTML format (**Supplementary File S1**).

Neotropical soil samples. The sampling design allowed for a broad view of the diversity of soil-inhabiting protists in three lowland Neotropical rainforest types. Samples were taken at La Selva Biological Station, Costa Rica in October 2012 and June 2013; Barro Colorado Island (BCI), Panama October 2012 and June 2013; and Tiputini Biodiversity Station, Ecuador in October 2013 (see **Supplementary File 1** for sample coordinates). The aim was to collect soils from around each field station, excluding areas where humans were not permitted and areas of secondary growth. Fieldwork and government-issued sampling permits were made possible by the Organization for Tropical Studies and the Ministerio del Ambiente y Energía and Sistema Nacional de Areas de Conservación in Costa Rica, the Smithsonian Tropical Research Institute in Panama, and David Romo Vallejo and Consuelo de Romo at the Universidad San Francisco de Quito in Ecuador.

After removal of the loose-leaf layer, stones, visible plant roots and visible animals, five ml of the top-most layer of surface soil were immediately placed into eight ml of LifeGuard Soil Preservation Solution (MO BIO, Carlsbad, CA). DNA was isolated with the PowerSoil DNA Elution Accessory Kit (MO BIO). Like other large-scale sequencing studies, e.g.,[1,2], only DNA was used in this study, not RNA which can likewise remain even in ancient sediments[3,4]. See **Extended Data Figure 8** to see that RNA can be found in cysts, and for a comparison of DNA and RNA abundances; additionally, the low similarity of OTUs among samples within each

forest, even those that were collected near each other (**Extended Data Figure 7**), point to the soils not having the same taxa that would be expected if most of the DNA was from a large seed bank of dead or inactive species that collected overtime. The samples were taken mostly at least 100 meters apart. DNA from every two consecutive samples was combined in equal concentration; many of these samples failed to amplify, and only those that worked are shown here. The combined samples were amplified with a high fidelity polymerase for the V4 region of 18S rRNA following Mahé et al.[5] with universal eukaryotic V4 primers (TAReuk454FWD1 and TAReukRev3)[6]. This primer pair has been used in a wide variety of environments, they have been shown to amplify a wide variety of eukaryotic lineages, and they have not shown a preference for amplifying Apicomplexa[6-10] (see also **Extended Data Figure 3**); however, some protist lineages may have been missed[11,12]. Additionally, the V4 primers were able to amplify all Oomycota in GenBank in an *in silico* analysis.

After PCR cleanup, a second PCR with primers containing sample-specific tags and Illumina MiSeq's sequencing adapters was performed following manufacturer's instructions. Amplicons were gel-size selected for appropriate length, and library quality was assessed using the Bioanalyzer 2100 (Agilent, Santa Clara, CA) and Qubit 2.0 (Life Technologies, Darmstadt, Germany). Illumina MiSeq sequencing was performed at GATC Biotech AG (Konstanz, Germany) with v3 chemistry and software MCS v2.3.0.3 and RTA v1.18.42.0. PhiX DNA was spiked-in to the MiSeq flowcell to increase base diversity. Twenty amplified products (each from a combined two samples) were loaded onto one MiSeq flowcell.

Ten samples from La Selva, Costa Rica, and ten samples from Barro Colorado Island, Panama, were also amplified with primers designed to specifically target the Ciliophora[13], which, like the Apicomplexa, is classified within the Alveolata[14]. After PCR cleanup, a second, nested

PCR was performed with V4 primers, and then sequenced with Roche/454 GS FLX+ with Titanium chemistry and software v2.6. The 20 amplified products were loaded onto 1/4 of a plate.

Marine samples. Reads from the open oceans came from the *Tara*-Oceans consortium (http://oceans.taraexpeditions.org/en)[1]. Starting with *Tara*-Oceans' raw reads (SRA accessions PRJEB6610 and PRJEB7315) allowed us to treat the sequencing data with the same bioinformatic pipeline. Our pipeline's goal was to improve the sensitivity of OTU detection by applying most quality- and frequency-based filtering steps after clustering instead of applying them before, and removal of low abundant OTUs was performed on the combined data and not at the sample level. We also replaced the paired-read merger FLASH[15] with the newer PEAR[16], which is advertised to yield higher quality results. After a careful investigation, it turned out that the large increase in OTU number we observed compared to de Vargas et al.[1] is mainly due to that later modification.

Samples from European near-shore marine sites (see **Supplementary File 1** for coordinates for each sample) came from the BioMarKs consortium (http://www.biomarks.eu)[9,17]. Total nucleic acids from sediment samples were extracted using the RNA Power Soil Total Isolation kit combined with DNA Elution Accessory kit (MO BIO) following manufacture's instructions. To remove contaminating DNA in the RNA extracts, the samples were DNase treated twice with DNase for 25min at 37°C with 2U of Turbo DNA-free kit (Ambion) followings manufacturer's instructions. RNA was reverse-transcribed into first-strand cDNA using Invitrogen's Superscript III (Thermo Fisher Scientific, Waltham, MA) following manufacturer's instructions. The hyper-variable V4 region of 18S rRNA was amplified with

general eukaryotic primers (TAReuk454FWD1 and TAReukRev3)[6], and the hyper-variable V9 region of 18S rRNA was amplified with general eukaryotic primers (1389F and 1510R)[18]. Amplified products were sequenced on an Illumina Genome Analyser IIx sequencer at CEA Genoscope (Évry, France).

Read cleaning and clustering. Fastq files were assembled with PEAR v0.9.8[16] using default parameters and converted to fasta format. Following Mahé et al.[5], assembled paired-end reads were filtered with Cutadapt v1.9[19] and retained if they had both primers and no Ns (or ambiguously called nucleotides). Reads were dereplicated into strictly-identical amplicons with Vsearch v1.6.0[20], then clustered with Swarm v2.1.5[21] using *d*=1 with the fastidious option on. The most abundant amplicon in each OTU was searched for chimeric sequences with Vsearch v1.6.0[20], and their OTUs were removed even if they occurred in multiple samples.

Taxonomic assignment. Taxonomic assignment used Vsearch's global pairwise alignments with the Protist Ribosomal Reference (PR^2) database v203[22] based on GenBank release v203. Using Cutadapt, the PR^2 database was extracted for just the specific regions that were amplified and sequenced to allow for comparisons using a global pairwise alignment. Amplicons were assigned to their best hit, or co-best hits, in the reference database as reported by Vsearch. The different steps of that taxonomic assignment strategy are grouped in a pipeline called Stampa (https://github.com/frederic-mahe/stampa). Low abundant OTUs were removed from the combined dataset only if they included ≤2 reads, and were found in only one sample, and were <99% similar to an accession in PR2. All reads were deposited at GenBank's SRA bioproject # SUB582348.

Stampa plots. To assess globally the results of Stampa, our taxonomic assignment pipeline, we produced "Stampa plots"[23]. Stampa plots are simple distribution plots showing the number of reads per similarity value, where the similarity value is the best match between environmental and reference sequences. Stampa plots are a visual assessment of the taxonomic coverage of our reference database sequences: if most environmental reads or OTUs have high similarity values with references, then the coverage is good. On the contrary, low similarity values indicate a lack of coverage, which can be the sign of amplification-sequencing artifacts (chimeras for instance), the presence of known taxa without reference sequences, the presence of new taxa, or if the coverage deficit is massive, a previously unexplored environment (such as Neotropical rainforest soils). The Stampa website is at: https://github.com/frederic-mahe/stampa.

Phylogenetic placement pipeline. To put the 50,118,536 soil protist reads into a phylogenetic context, they were dereplicated into 10,567,804 strictly identical amplicons and placed onto a comprehensive eukaryotic reference tree. The corresponding multiple sequence alignment (MSA) used to build this tree contained 512 full length sequences from all major eukaryotic clades (see **Supplementary File 2** for GenBank accession numbers) based on the taxon sampling as summarized in a range of pan-eukaryotic phylogenomic studies reviewed in [24,25], with a bias towards lineages known to occur in soils e.g.,[26,27,28]. Only high quality, full- or near-full length 18S rRNA reads were selected, and reads that have previously been observed to form long branches omitted, to reduce phylogenetic artifacts. It is acknowledged that single gene trees are unable to resolve the backbone of the eukaryote phylogeny; however, this was not the intended purpose of this taxon selection or the phylogenetic analyses. The intent was to produce a comprehensive eukaryote sample tree on which to place V4 high-throughput sequencing reads

generated by this study. Phylogenetic placements of query OTU representatives that were taxonomically assigned to the Apicomplexa were conducted using a full-length reference MSA and corresponding phylogeny comprising 190 taxa from all major Alveolate clades (see **Supplementary File 2** for GenBank accession numbers). The taxon sampling was collated from a range of publications relating to eukaryotic diversity in general and Alveolata diversity and phylogeny in particular; e.g.[14,29-35]. It is also acknowledged that single gene trees are unable to resolve the backbone of the Alveolate phylogeny.

For a full explanation of the phylogenetic placement see **Supplementary File 2**. In brief, we used RAxML v8.1.15[36] to infer reference trees from the reference alignments, then used PaPaRa v2.4[37] to align the query sequences to those reference alignments, and finally used the Evolutionary Placement Algorithm (EPA)[38] to place the queries onto the trees. Placement results were visualized as "Heattrees" using Genesis[39]. The trees were inferred with and without taxonomic constraints, although the results were similar.

Phylogenetic placement of Haptophyta. For the 28 Haptophyta OTUs, we fist downloaded 106 GenBank accession based off of the taxon sampling of [40,41], and aligned MAFFT v7 [42]. OTU representatives were then added to this alignment using the MAFFT "add-in" function, and the tree was inferred with RAxML, with substitution model GTRGAMMAI, algorithm New rapid hill-climbing and 100 bootstrap runs.

Oomycete analyses. To tentatively assign functional roles of the Oomycota OTUs, we first compared each individual OTU with related sequences in Genbank using BLAST[43] (Altschul et al., 1997) for a taxonomic affiliation as accurate as possible. In ambiguous cases, we proceeded

to a quick phylogenetic neighbor joining analysis to determine as closely possible its affiliation. Functional affiliation of the OTUs was then determined mostly based on Lara et al.[44].

Statistical analyses. We analyzed frequency count data derived from OTU clustering to estimate the total (i.e., observed + unobserved) OTU richness in the population from which the observed samples were drawn. These data consisted of the number of OTUs with one representative in the sample (the "singletons"), the number with two representatives, three, and so on. We used two sets of statistical procedures. The first was implemented in the software package Breakaway v2.0[45]. This method estimates a family of statistical models, known as "Kemp"-type distributions, from the frequency count data, by fitting ratios of successive frequency counts via nonlinear regression; an optimal selected model is returned, along with error and goodness-of-fit assessments. In the case of the combined (global) OTU data the lowest-order model was selected, which is equivalent to fitting a weighted linear regression to the ratios of successive frequency counts. The second was implemented in the software package CatchAll[46]. This method fits a family of mixed-Poisson models to the frequency data, as well as calculating a suite of nonparametric procedures. It returns an optimal selected model with errors and goodness-of-fit assessments, and comparisons of parametric and nonparametric analyses. CatchAll selected a mixture of three exponentials-mixed Poisson models. In all cases the results were virtually identical, showing almost no diversity extant in the population beyond that observed in the sample. The R package Vegan [47] was used to analyze frequency count data derived from OTU clustering. Different functions of Vegan were called to compute diversity indexes (Shannon, Simpson, and Rényi diversities), to randomly subsample our samples (rrarefy function) and to

estimate and compare species compositions, using Bray-Curtis distance and NMDS ordination (monoMDS function). Figures were made with R[48] and ggplot2[49].

Detecting RNA in cysts. To assess if viable and sequence-able RNA can be detected in protist cysts, we designed fluorescent in situ hybridization (FISH) probes specific to the kinetoplastids with 18S rRNA probes coupled to Cy3 fluorochrome and designed to hybridize to ribosomes: KIN516, a wide-spectrum kinetoplastic probe (ACCAGACTTGTCCTCC); and Bodo1757, a probe designed on the V8 region and specific to the clade *Bodo* (CGAGCAAGTGAAACACTCGCC). Probes were checked for specificity using pure cultures of *Neobodo designis* strain AND31 (AY965872) and *Bodo saltans* strain SCCAP BS364. Cultures were fixed with three volumes of paraformaldehyde 4% (PFA) for two hours. Fixed cell suspensions were collected on a 0.2 µm nitrocellulose ISOPORE filter (Millipore, Billerica, MA, USA) and rinsed with demineralized water. Hybridizations were performed at 45°C for 90 min in a buffer containing 0.9 M NaCl, 20mM Tris HCl, and 0,01% SDS, pH 7.2. Probes were used at a concentration of 50ng/µl. The optimal amount of formamide in the hybridization buffer to reduce non-specific binding of the probe was set by performing hybridizations with increasing formamide concentrations for the different strains investigated. In general, hybridization with 35% formamide did not result in any unspecific binding of the probes to non-target organisms. For unspecific staining of all cells, DAPI (4', 6-diamidino-2-phenylindole) was added at a concentration of 1 µg/ml to the hybridization buffer. After hybridization, filters were washed for 15 min at 48°C in a buffer containing 20 mM Tris-HCl (pH 7.2), 10 mM EDTA, 0.01% SDS and 80 mM NaCl, subsequently rinsed with distilled water and air-dried. Prior to microscopic observation, filters were mounted with Citifluor solution. Probes were applied to a soil taken

from a garden compost in a garden in Switzerland. Soil was mixed and sieved through a 2 mm grid to remove large debris. The soil was then spread as a 1cm layer on a 16 x 23 cm plastic tray, and wetted with autoclaved demineralized water (without flooding). Two layers of lens-cleaning tissues were spread on the surface, and four microscope cover slips measuring 24 per 50 mm each (Menzel Gläser, Braunschweig, Germany) were laid upon the wet tissues. The tray was left in the dark at 18°C for one week. Protists attached to the coverslips were removed by rinsing with Neff's Amoeba Saline buffer and collected into a 10 ml Falcon tube. Cells were directly fixed with PFA on a nitrocellulose filter (0.2µm) and hybridized using the specific designed probes as well as with DAPI as described above.

Measuring DNA and RNA. Ciliates were measured for the DNA and RNA abundances of the 18S rRNA locus: *Colpoda magna*, from the American Type Culture Collection # 50128; *Dileptus* sp., from Carolina Biological Supply Company; *Halteria grandinella*, collected on October 2013 at a pond at the University of Kaiserslautern; *Paramecium tetraurelia* strain D4-2, provided by Martin Simon, Saarland University; and *Tetrahymena thermophila* wildtype B, provided by Josef Loidl, University of Vienna. Clonal populations were cultured with wheat grass powder medium inoculated with the bacterium *Klebsiella minuta*. Standard curve were made with cells from each species were amplified for the 3' end of 18S rRNA with primers 1391-F[50] and Euk-B[51]. PCR products were cleaned with MinElute PCR Purification Kit from Qiagen and ligated to pGEM®-T vector (Promega, Madison, Wisconsin, U.S.A.) and cloned. Plasmids were isolated with the FastPlasmid Mini Kit (5 Prime, Hilden, Germany), and checked with an amplification using vector primers and Sanger sequencing. Plasmids were diluted to a six tenfold standard for the qPCR ranging from 10^{-1}-10^{-6} ηg/µl. For quantitative real-time PCR

(qPCR) of RNA, total RNA was extracted 12 separate times from single cells first washed 3X in Volvic water, then lysed with the Ambion Single Cell Lysis Kit (Thermo Fisher Scientific) with an integrated DNase step. cDNA was generated using the QuantiTect Reverse Transcription Kit (Qiagen, Venlo, Netherlands) with another integrated DNase step. For quantitative real-time PCR (qPCR) of DNA, measurements were taken directly from 12 separate single cells. Amplifications were performed in a final volume of 20 µl, containing 10 µl iQ™ SYBR Green Supermix from Bio-Rad (Hercules, California, U.S.A.), 1.0 µl template DNA (cDNA or a single cell), 10 pg of each primer, and 7.0 µl RNase free water. Cycling conditions were: 95 °C for 2 min; and 40 cycles of 95 ^{0}C for 15s, 57 ^{0}C for 25s, and 72 ^{0}C for 40s. Melt curve data were collected after cycle 2 between 57 ^{0}C and 95 ^{0}C with a temperature increase at a rate of 0.5 ^{0}C/10s. qPCR reactions were performed with iQ Single Color Real-Time PCR Detection System (Bio-Rad), with negative control and the six tenfold dilution in triplicates of standard curves plotting against Ct values. The amplification efficiency E was estimated with the equation $E=(10^{-1/slope})-1$ with the calculated slope of the standard curves. We obtained efficiencies E between 90.1% and 105.5% and R^2 between 0.987 and 1.000. Copy numbers were calculated using the following equation: number of copies = (amount × 6.022×10^{23}) / (length × 1×109 × 660) where amount is the concentration of DNA or cDNA (in ng), and length is the length of the linear fragment plus plasmid, 660 is the average molecular weight of one base pair for double strand DNA and 6.022×10^{23} is the molar constant[52].

References for Methods and Extended Data

1 de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).

2 Tedersoo, L. *et al.* Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).

3 Orsi, W., Biddle, J. F. & Edgcomb, V. Deep sequencing of subseafloor eukaryotic rRNA reveals active fungi across marine subsurface provinces. *PLoS ONE* **8**, e56335 (2013).

4 Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T. & Wilding, T. Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Mol. Ecol. Res.* **14**, 1129-1140 (2014).

5 Mahé, F. *et al.* Comparing high-throughput platforms for sequencing the hyper-variable V4 region in environmental eukaryotic diversity surveys. *J. Eukaryot. Microbiol.* **62**, 338-345 (2015).

6 Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **19**, 21-31 (2010).

7 Filker, S., Gimmler, A., Dunthorn, M., Mahé, F. & Stoeck, T. Deep sequencing uncovers high and substantial novel diversity in protistan plankton of solar saltern ponds. *Extremophiles* **19**, 283-295 (2015).

8 Forster, D. *et al.* Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol.* **13**, 16 (2015).

9 Logares, R. *et al.* Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813-821 (2014).

10 Richards, T. A. *et al.* Molecular diversity and distribution of marine fungi across 130 European environmental samples. *Proc. R. Soc. B* **282**, 20152243 (2015).

11 Lentendu, G. *et al.* Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach. *Molec. Ecol.* **23**, 3341-3355 (2014).

12 Fiore-Donno, A. M., Weinert, J., Wubet, T. & Bonkowski, M. Metacommunity analysis of amoeboid protists in grassland soils. *Scientific reports* **6**, 19068 (2016).

13 Lara, E., Berney, C., Harms, H. & Chatzinotas, A. Cultivation-independent analysis reveals a shift in ciliate 18S rRNA gene diversity in a polycyclic aromatic hydrocarbon-polluted soil. *FEMS Microbiol. Ecol.* **62**, 365-373 (2007).

14 Adl, S. M. *et al.* The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59**, 429-493 (2012).

15 Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).

16 Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620 (2014).

17 Massana, R. *et al.* Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ. Microbiol.* **17**, 4035-4049 (2015).

18 Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistian diversity using massively parallel sequencing of V9 hypervariable region of small-subunit ribosomal RNA genes. *PLoS ONE* **4**, 7 (2009).

19 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).

20 https://github.com/torognes/vsearch.

21 Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, e1761 (2015).

22 Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**, D597-604 (2013).

23 Stampa: sequence taxonomic assignment by massive pairwise alignments (https://github.com/frederic-mahe/stampa, 2016).

24 Burki, F. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect Biol.* **6**, a016147 (2014).

25 del Campo, J. *et al.* The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* **29**, 252-259 (2014).

26 Bates, S. T. *et al.* Global biogeography of highly diverse protistan communities in soil. *ISME J.* **7**, 652-659 (2013).

27 Dupont, A., Griffiths, R. I., Bell, T. & Bass, D. Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. *Environ. Microbiol.* **doi: 10.1111/1462-2920.13220** (2016).

28 Geisen, S. *et al.* Metatranscriptomic census of active protists in soils. *ISME J.* **9**, 2178-2190 (2015).

29 Barta, J. R., Ogedengbe, J. D., Martin, D. S. & Smith, T. G. Phylogenetic position of the adeleorinid coccidia (Myzozoa, Apicomplexa, Coccidia, Eucoccidiorida, Adeleorina) inferred using 18S rDNA sequences. *J. Eukaryot. Microbiol.* **59**, 171-180 (2012).

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

31 Gomez, F., López-Garcia, P., Nowaczyk, A. & Moreira, D. The crustacean parasites *Ellobiopsis* Caullery, 1910 and *Thalassomyces* Niezabitowski, 1913 form a monophyletic divergent clade within the Alveolata. *Syst. Parasitol.* **74**, 65-74 (2009).

32 Rueckert, S. & Leander, B. S. Molecular phylogeny and surface morphology of marine archigregarines (Apicomplexa), *Selenidium* spp., *Filipodium phascolosomae* n. sp., and *Platyproteum* n. g. and comb. from North-Eastern Pacific peanut worms (Sipuncula). *J. Eukaryot. Microbiol.* **56**, 428-439 (2009).

33 Rueckert, S., Chantangsi, C. & Leander, B. S. Molecular systematics of marine gregarines (Apicomplexa) from North-eastern Pacific polychaetes and nemerteans, with descriptions of three novel species: *Lecudina phyllochaetopteri* sp. nov., *Difficilina tubulani* sp. nov. and *Difficilina paranemertis* sp. nov. *Int. J. Syst. Evol. Microbiol.* **60**, 2681-2690 (2010).

34 Slapeta, J. & Linares, M. C. Combined amplicon pyrosequencing assays reveal presence of the apicomplexan "type-N" (cf. *Gemmocystis cylindrus*) and *Chromera velia* on the Great Barrier Reef, Australia. *PLoS One* **8**, e76095 (2013).

35 Wakeman, K. C. & Leander, B. S. Molecular phylogeny of Pacific Archigregarines (Apicomplexa), including descriptions of *Veloxidium leptosynaptae* n. gen., n. sp., from the sea cucumber *Leptosynapta clarki* (Echinodermata), and two new species of *Selenidium*. *J. Eukaryot. Microbiol.* **59**, 232-245 (2012).

36 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

37 Berger, S. A. & Stamatakis, A. Aligning short reads to reference alignments and trees. *Bioinformatics* **27**, 2068-2075 (2011).

38 Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under Maximum Likelihood. *Syst. Biol.* **60**, 291-302 (2011).

39 Genesis. A toolkit for working with phylogenetic data (https://github.com/lczech/genesis, 2016).

40 Egge, E., Eikrem, W. & Edvardsen, B. Deep-branching novel lineages and high diversity of haptophytes in the Skagerrak (Norway) uncovered by 454 pyrosequencing. *J. Eukaryot. Microbiol.* **62**, 121-140 (2015).

41 Shalchian-Tabrizi, K., Reier-Røberg, K., Ree, D. K., Klaveness, D. & Bråte, J. Marine-freshwater colonizations of haptophytes inferred from phylogeny of environmental 18S rDNA sequences. *J. Eukaryot. Microbiol.* **58**, 315-318 (2011).

42 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066 (2002).

43 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402. (1997).

44 Lara, E. & Belbahri, L. SSU rRNA reveals major trends in oomycete evolution. *Fungal Divers.* **49**, 93-100 (2011).

45 Willis, A. & Bunge, J. Estimating diversity via frequency ratios. *Biometrics* **71**, 1042-1049 (2015).

46 Bunge, J. *et al.* Estimating population diversity with CatchAll. *Bioinformatics* **28**, 1045-1047 (2012).

47 Vegan: community ecology package. R package version 2.0-7 (2013).

48 R: a language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).

49 Wickham, H. *gplot2: elegant graphics for data analysis*. (Springer, 2009).

50 Lane, D. J. in *Nucleic acids techniques in bacterial systematics* (eds E. Stackebrandt & M. Goodfellow) 115-147 (John Wiley & Sons, 1991).

51 Medlin, L., Elwood, H. J., Stickel, S. & Sogin, M. L. The characterization of enzymatically amplified eukaryotes 16S-like ribosomal RNA coding regions. *Gene* **71**, 491-500 (1988).

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

53 Creer, S. *et al.* Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.* **19**, 4-20 (2010).

54 Geisen, S., Laros, I., Vizcaino, A., Bonkowski, M. & de Groot, G. A. Not all are free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa. *Mol Ecol* **24**, 4556-4569 (2015).

55 Jordan, R. W., Cros, L. & Young, J. R. A revised classification scheme for living haptophytes. *Micropaleontol.* **50**, 55-79 (2004).

56 Gong, J., Dong, J., Liu, X. & Massana, R. Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* **164**, 369-379 (2013).

57 Klobutcher, L. A. & Herrick, G. Developmental genome reorganization in ciliated protozoa: the transposon link. *Prog. Nucleic Acid Res. Mol. Biol.* **56**, 1-62 (1997).

58 Riley, J. L. & Katz, L. A. Widespread distribution of extensive genome fragmentation in ciliates. *Mol. Biol. Evol.* **18**, 1372-1377 (2001).

59 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).

60 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

61 Ghodsi, M., Liu, B. & Pop, M. DNACLUST: accurate and efficient clustering of phylogenetuc marker genes. *BMC bioinformatics* **12**, 271 (2011).

62 Bittner, L. *et al.* Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Mol. Ecol.* **22**, 87-101 (2013).

63 Bråte, J. *et al.* Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *ISME J* **4**, 1144-1153 (2010).

64 Dunthorn, M. *et al.* Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol. Biol. Evol.* **31**, 993-1009 (2014).

65 Egge, E. S. *et al.* Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. *Molec. Ecol.* **24**, 3026-3042 (2015).

66 Grossmann, L. *et al.* Protistan community analysis: key findings of a large-scale molecular sampling. *ISME J.* **doi:10.1038/ismej.2016.10** (2016).

67 Hu, S. K. *et al.* Estimating protistan diversity using high-throughput sequencing. *J. Eukaryot. Microbiol.* **62**, 688-693 (2015).

68 Monier, A. *et al.* Oceanographic structure drives the assembly processes of microbial eukaryotic communities. *ISME J.* **9**, 990-1002 (2015).

69 Orsi, W., Song, Y. G., Hallam, S. & Edgcomb, V. Effect of oxygen munium zone formation on communites of marine protists. *ISME J.* **6**, 1586-1601 (2012).

70 Caron, D. A. *et al.* Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl. Environ. Microbiol.* **75**, 5797-5808 (2009).

71 Nebel, M., Pfabel, C., Stock, A., Dunthorn, M. & Stoeck, T. Delimiting operational taxonomic units for assessing ciliate environmental diversity of small subunit rRNA gene sequences. *Environ. Microbiol. Reports* **3**, 154-158 (2011).

72 Pernice, M. C., Logares, R., Guillou, L. & Massana, R. General patterns of diversity in major marine microeukaryote lineages. *PLoS ONE* **8**, e57170 (2013).

73 Dunthorn, M., Klier, J., Bunge, J. & Stoeck, T. Comparing the hyper-variable V4 and V9 regions for assesment of ciliate environmental diversity. *J. Eukaryot. Microbiol.* **59**, 185-187 (2012).

74 Hu, S. K. *et al.* Estimating protistan diversity using high-throughput sequencing. *J. Eukaryot. Microbiol.* **62**, 688-693 (2015).

75 Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014).

76 Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14812-14817 (2004).

77 Borisenko, A. V., Lim, B. K., Ivanova, N. V., Hanner, R. H. & Hebert, P. D. N. DNA barcoding in surveys of small mammal communities: a field study in Suriname. *Mol. Ecol. Resour.* **8**, 471-479 (2008).

78 CBOL Plant Working Group *et al.* A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12794-12797 (2009).

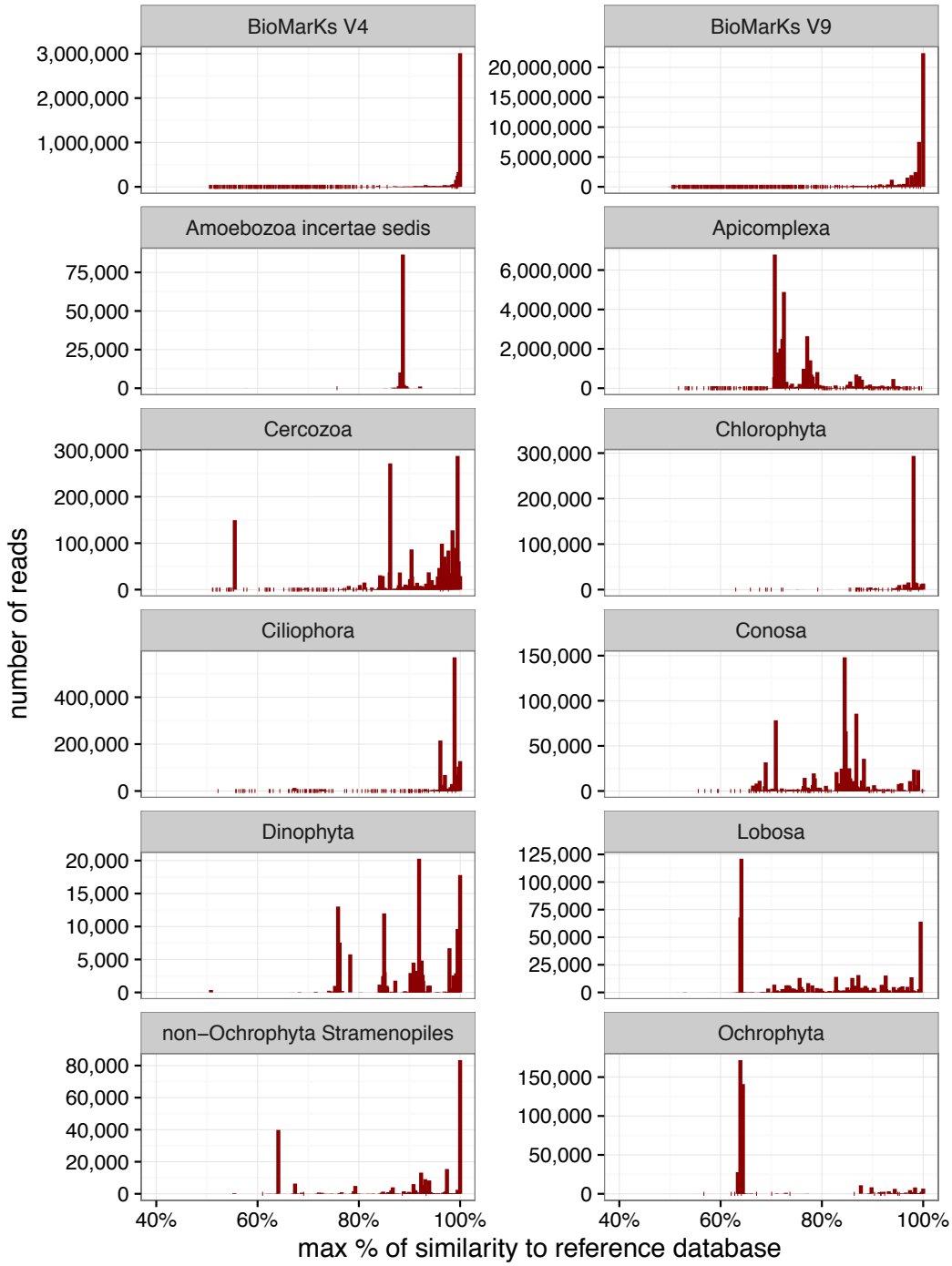
79 Kress, W. J., Wurdack, K., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. Use of DNA barcodes to identify flowering plants. *Proc. Nat. Acad. Sci. U.S.A.* **102**, 8369-8374 (2005).

80 Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6241-6246 (2012).

81 Pawlowski, J. *et al.* CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol.* **10**, e1001419 (2012).

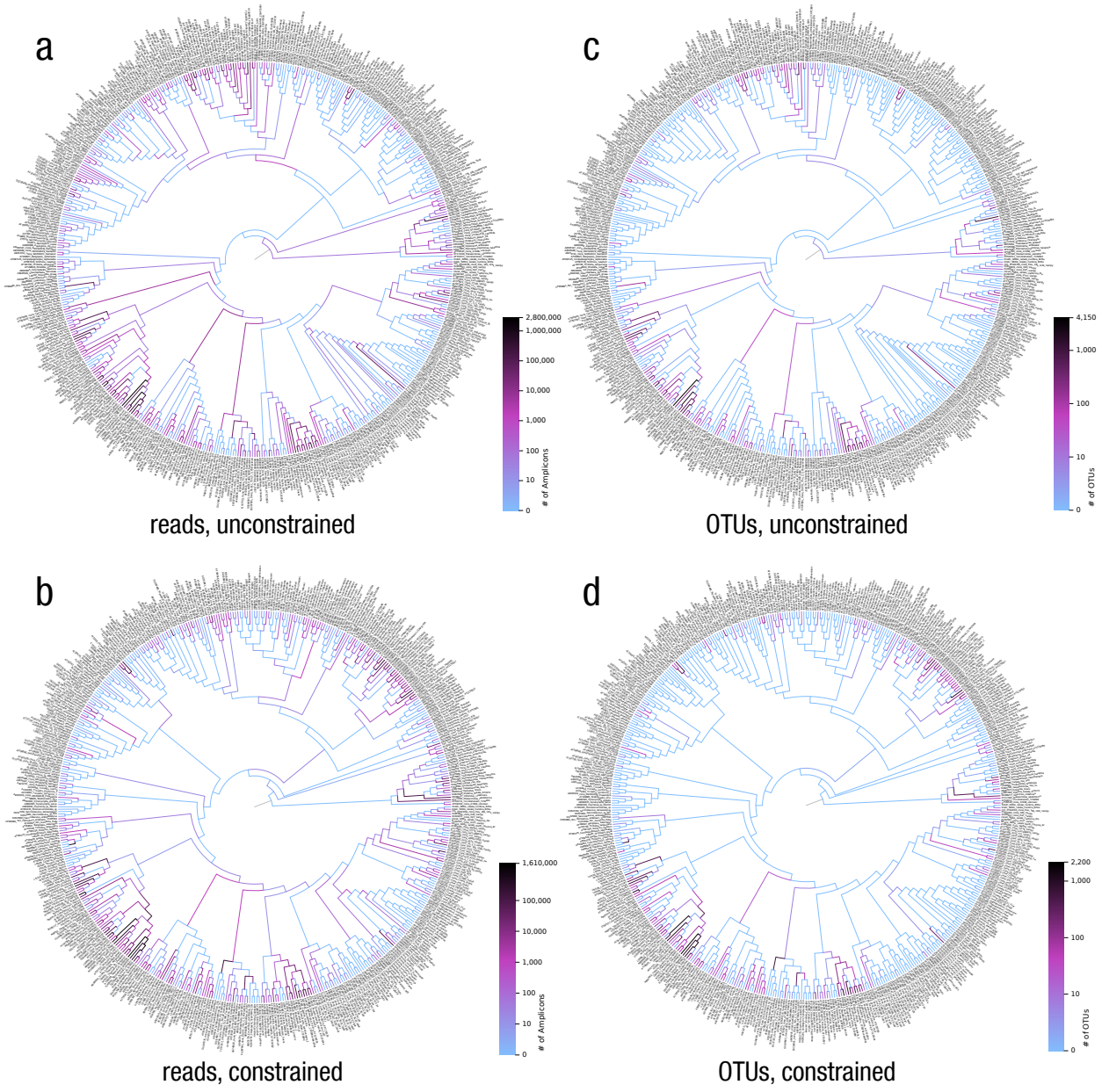
82 Pawlowski, J. & Holzmann, M. A Plea For DNA Barcoding of Foraminifera *J. Foramin. Res.* **44**, 62-67 (2014).

83 Stoeck, T., Przyboś, E. & Dunthorn, M. The D1-D2 region of the large subunit ribosomal DNA as barcode for ciliates. *Molec. Ecol. Res.* **14**, 458-468 (2014).

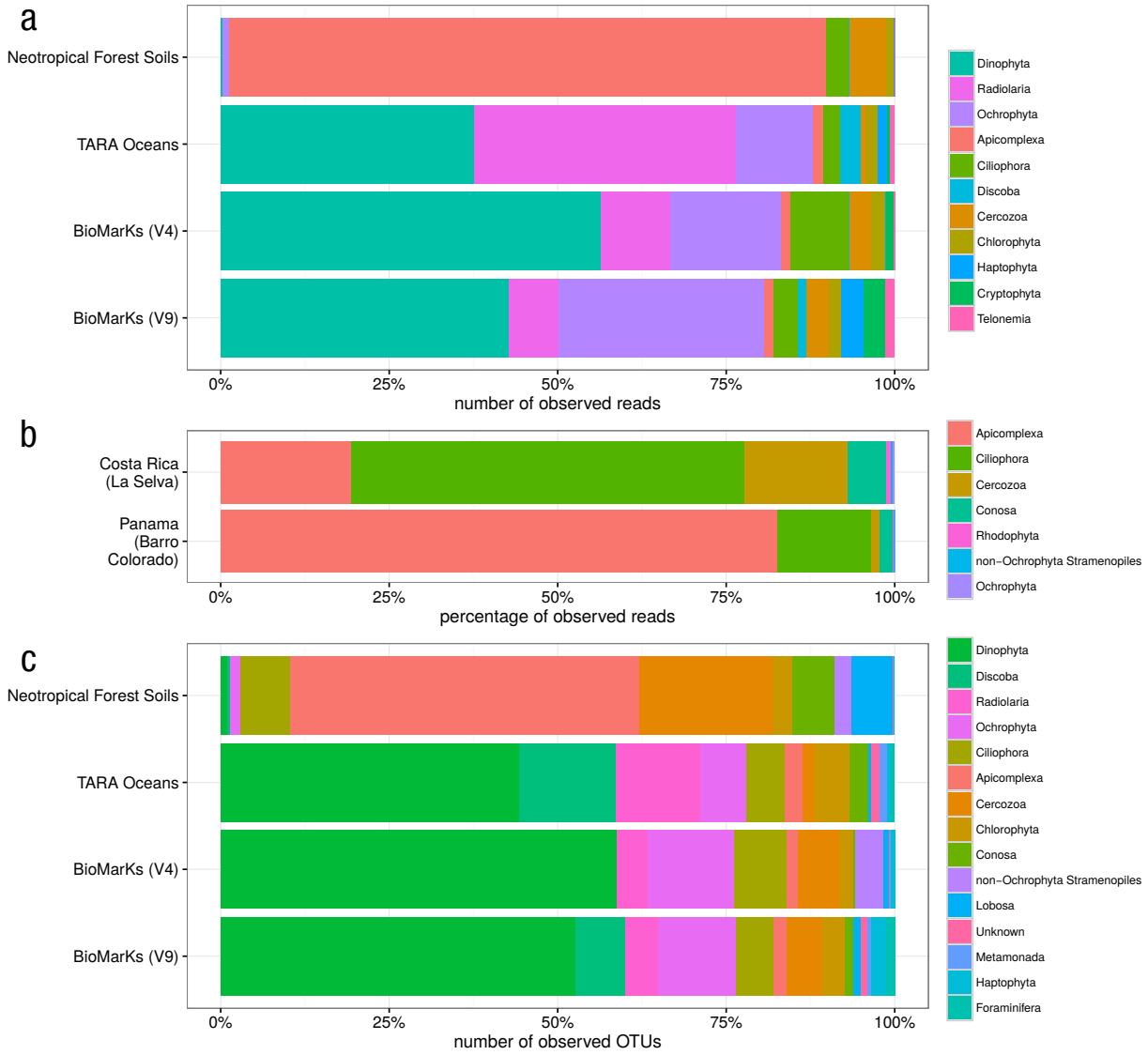


Extended Data Figure 1 | Similarity of protists reads to the taxonomic reference database. Top two panels: In contrast to the V4 protist data from the soils of the three Neotropical rainforests, but similar to the V9 protists reads from our re-analysis of the of the world's open

oceans (**Fig. 1**), the European near-shore marine sites were even more similar to the reference database: only 1.3% of the 6,092,106 protist V4 reads and only 1.9 % of the 49,794,719 protist V9 reads were <80% similar to PR[2]. Other panels: reads from the rainforest soils broken down into the top ten most abundant taxonomic groups.

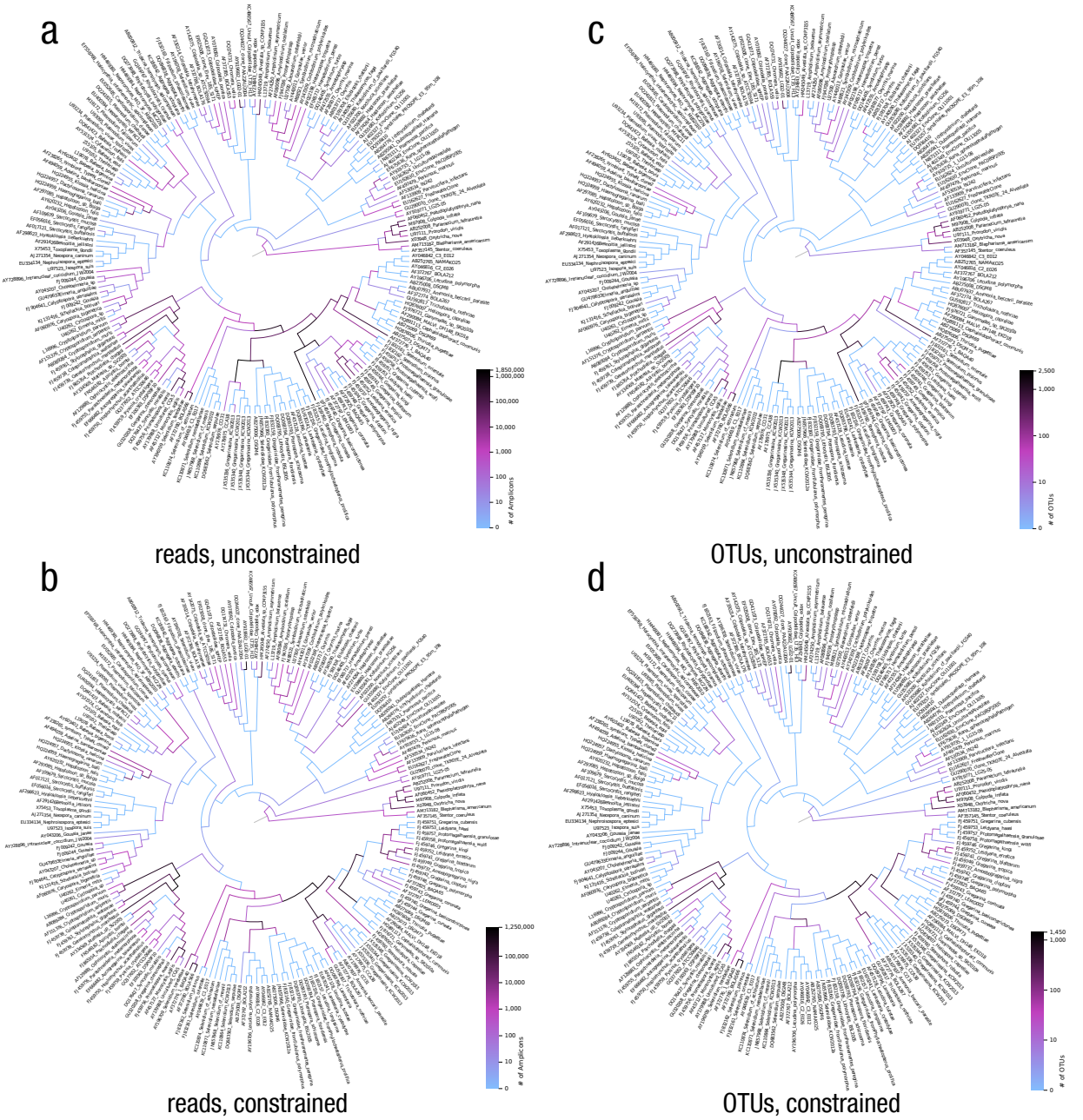


Extended Data Figure 2 | Phylogenetic placement of reads and OTUs onto the eukaryotic tree of life. Reads were dereplicated into strictly identical amplicons. The inferred relationships may differ from those obtained with phylogenomic data. **a,** Amplicons on a taxonomically unconstrained tree (see also **Fig. 2**). **b,** Amplicons on a taxonomically constrained tree. **c,** OTUs on a taxonomically unconstrained tree. **d,** OTUs on a taxonomically constrained tree.



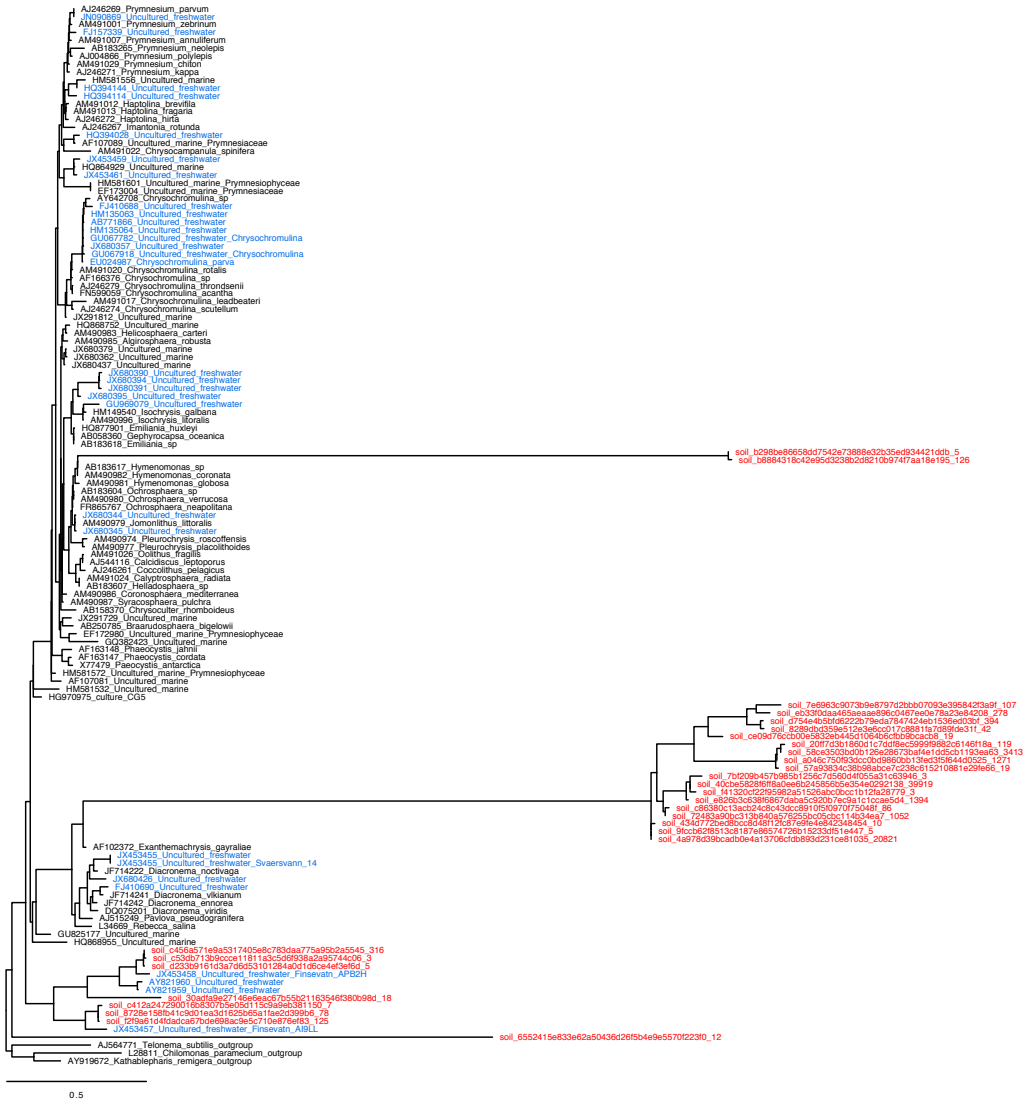
Extended Data Figure 3 | Taxonomic identity of protist reads and in different environments using different molecular markers. a, Taxa shown each represent >5% of the total reads. Apicomplexa abundances amount to an unprecedented 84.4% of reads in the combined rainforest soils. Apicomplexa reads from the other re-analyzed or newly sequenced environments are: 1.5% for V9 reads from the world's open oceans, 1.3% for V4 reads and 1.3% for V9 reads from European near-shore samples. Using metatranscriptomics, the Alveolata composed 57% of reads in an Arctic peatlands, although it is not clear how many of those where just Apicomplexa[28]. In a

study of a some samples from La Selva Biological Station, Cost Rica, protists were not taxonomically identified beyond "Protozoa"[53]; and a world-wide study of soils using Roche/454 of a region including the V4, Apicomplexa accounted for a little more than 50% of the reads in one of two Peru samples, with the other three humid sites having <50%[26]. An increase in Apicomplexa and other parasites were observed when animals were added to a mock community of soil protists[54]. **b,** Taxa shown each represent >0.1% of the total reads. About 47.8% of reads derived from ciliate-specific primers were assigned to the Apicomplexa. **c,** Taxa shown each represent >5% of the total OTUs.



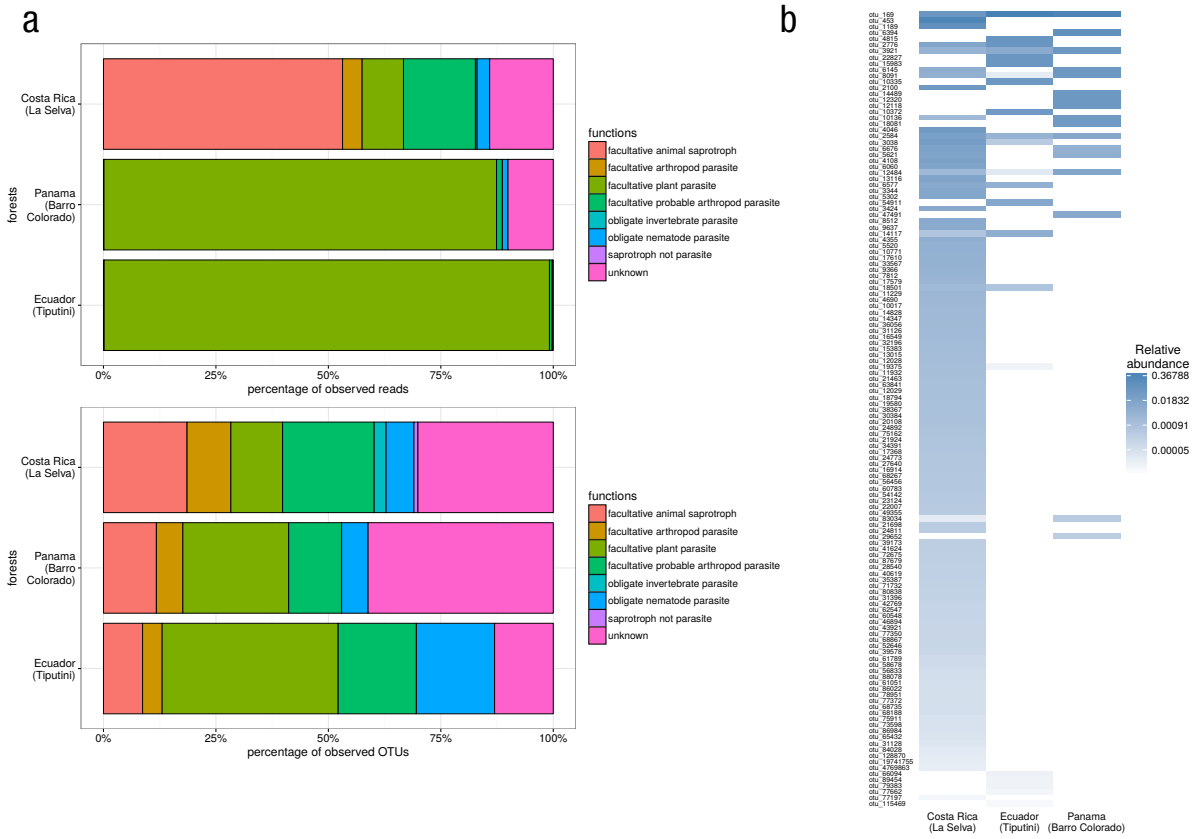
Extended Data Figure 4 | Phylogenetic placement of Apicomplexa OTUs and amplicons found in the three Neotropical rainforests onto taxonomically unconstrained and constrained trees. Most of the OTU placed with the gregarines, which are largely parasites of invertebrates. **a,** OTUs placed on a taxonomically unconstrained tree. **b,** OTUs placed on a taxonomically

constrained tree. **c,** Amplicons placed on a taxonomically unconstrained tree. **d,** Amplicons placed on a taxonomically constrained tree.

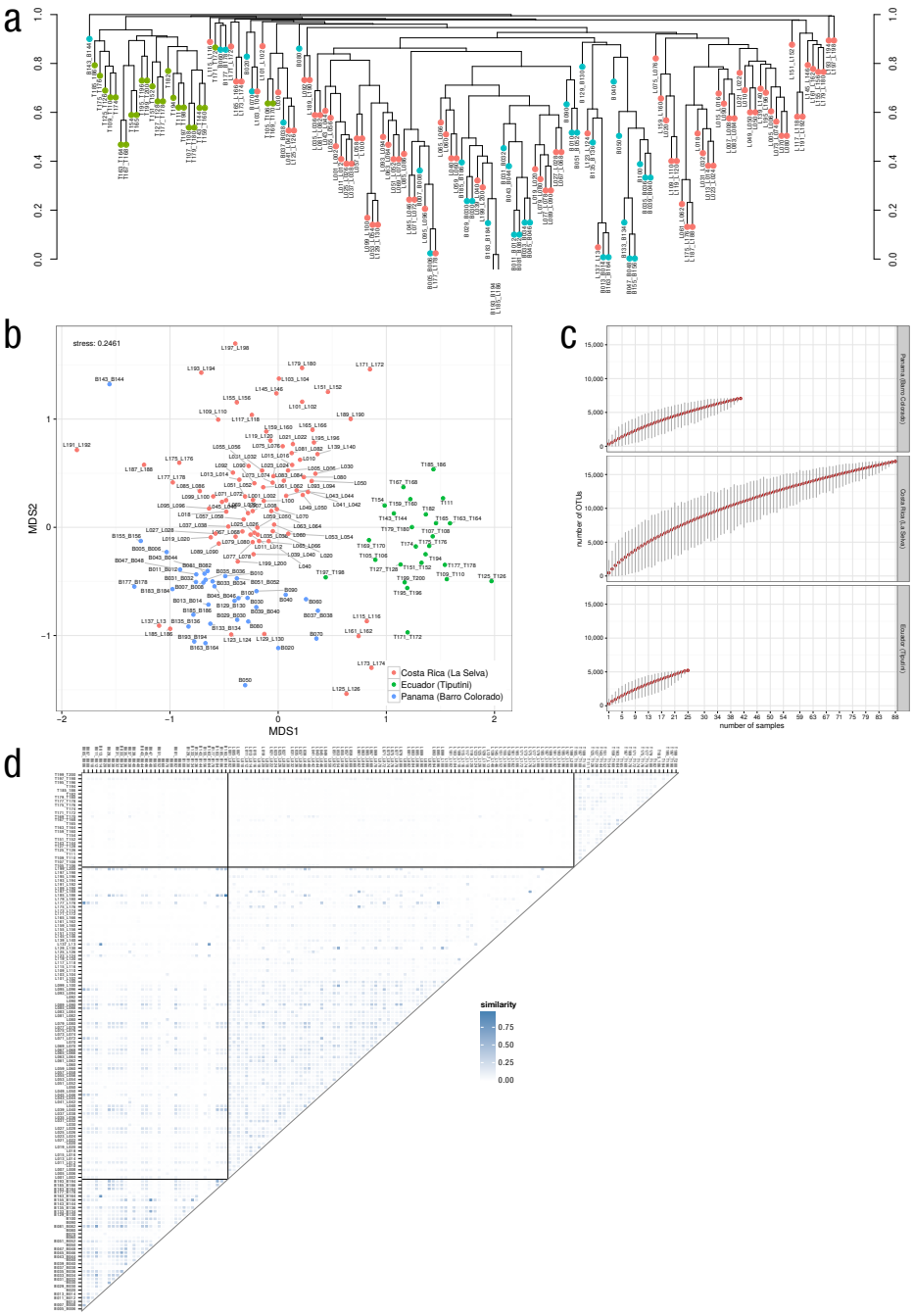


Extended Data Figure 5 | Phylogenetic placement of Haptophyta OTUs found in the three Neotropical forests. The Haptophyta are mostly known from marine environments, where they are primarily photo- or mixo-trophic; e.g., the bloom forming *Emiliania huxleyi* and *Phaeocystis* sp.[55]. However, a few freshwater Haptophyta are likely to have evolved from marine ancestors[41]. Sequencing of the soil samples resulted in 28 Haptophyta OTUs. Phylogenetic placement of the OTU centroids shows that some of them nest within already known freshwater clades, while

others form two new clades and one new branch. Blue= previously known freshwater species or environmental sequences; red= OTUs found in the Neotropical rainforests soils.

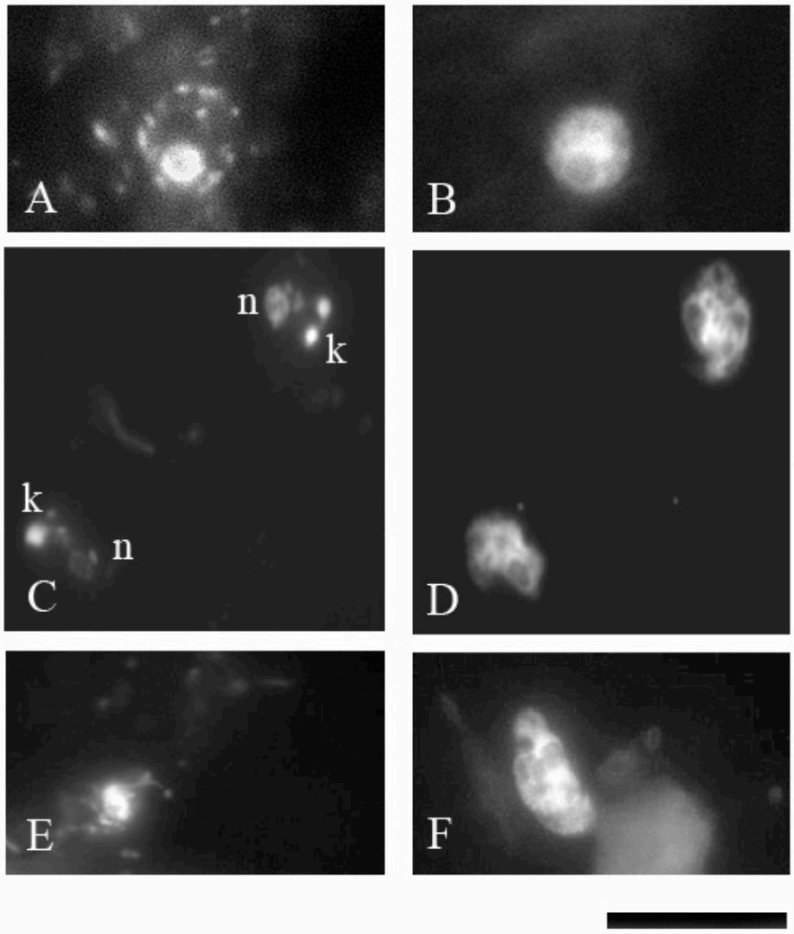


Extended Data Figure 6 | Oomycota in three Neotropical rainforests. a) Putative functional annotation of reads and OTUs. Most OTUs are not plant-specific parasites. **b)** Heatmap of Oomycota OTUs' relative abundance averaged for each forest. Most of the OTUs are not shared between the forests.



Extended Data Figure 7 | OTU composition differences between soil samples from three Neotropical rainforests. Note: samples were combined (see online Methods). **a,** Jaccard similarity index; the lower the similarity value, the higher the OTU turnover between samples. **b,** Non-metric multidimensional scaling; grouping according to taxonomic compositional similarity

(Bray-Curtis similarity). **c,** Jaccard similarity index; the lower the similarity value, the higher the OTU turnover between samples. **d,** OTU accumulation plot within each forest; circles represent average number of OTUs observed when accumulating *n* samples, and vertical bars represent tminimal and maximal observed numbers of OTUs in tested samples combinations. B = Barro Colorado Island, Panama, L = La Selva Biological Station, Costa Rica, and T = Tiputini Biodiversity Station, Ecuador.

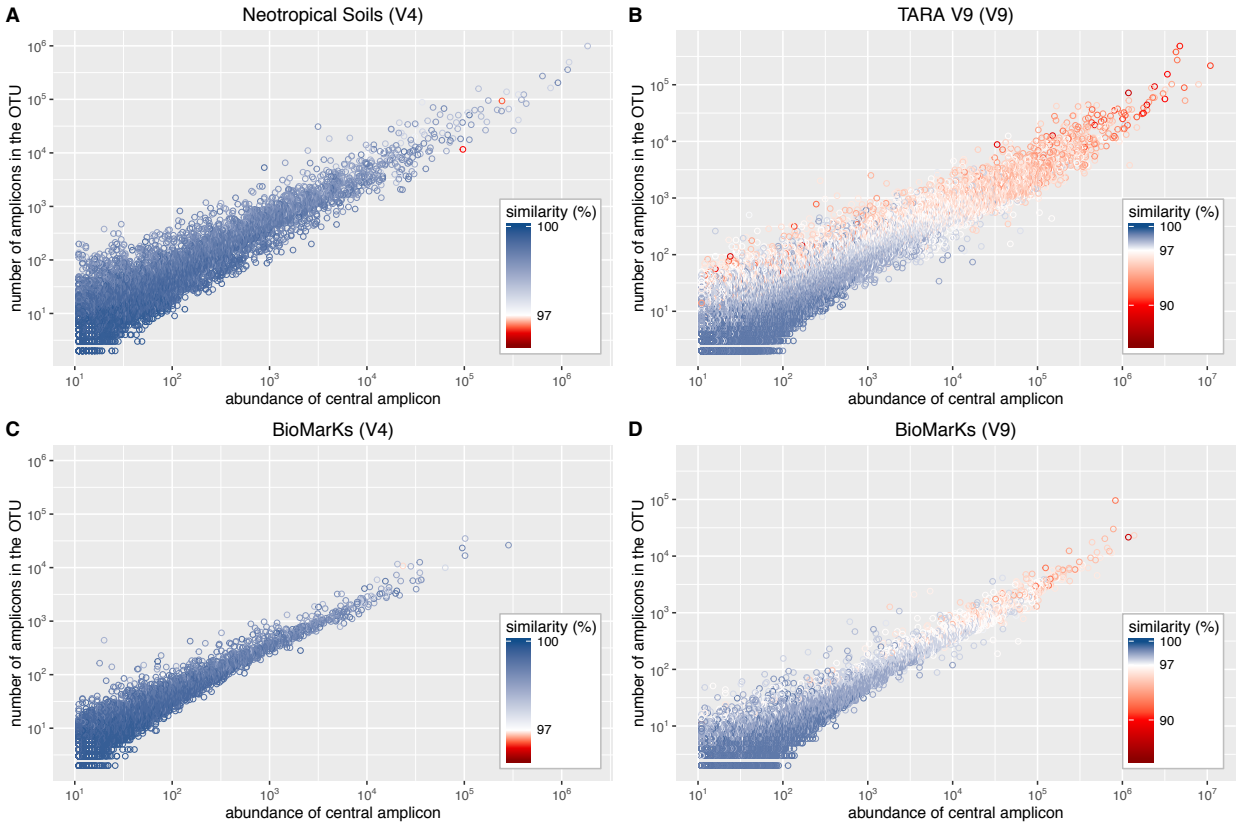


G

Species	mean DNA copy # (stand. dev.)	mean RNA copy # (stand. dev.)
Colpoda magna	45,211 (79,673)	1,182,250 (1,056,064)
Dileptus sp.	249,955 (326,520)	267,567 (206,271)
Halteria grandinella	54,411 (21,609)	166,244 (123728)
Paramecium tetraurelia	61,205 (73,909)	258,725 (784,481)
Tetrahymena thermophila	6,101 (3,765)	23,654 (21,702)

Extended Data Figure 8 | DNA vs. RNA in microbial environmental sequencing studies. a-f) Recent studies have shown that RNA can remain in the environment and be amplified[3,4]. Using FISH, here we show that sequenceable RNA can be found in encysted (i.e. dormant) kinetoplastid cells. **A,C,E)** DAPI-stained cells. **B)** Cyst with probe KIN516. **D)** Cells with KIN516. **F)** Cells with probe Bodo1757. Scale bar = 10μm. **g)**. Some Ciliophora have hundreds of DNA copies of each genes while others have thousands of copies[56-58], which can lead to

spurious abundance estimates. Here we show that RNA copy number differences of 18S rRNA can also lead to spurious abundance estimates. Taken together, these results from FISH and qPCR analyses show that, like DNA, there are problems with using RNA in environmental sequencing studies of protists.



Extended Data Figure 9 | Global clustering thresholds do not apply to all taxa. Most *de novo* OTU construction methods use a global similarity value (e.g.,[59-61]), which results in OTUs with radii or diameters (depending on the method) at most equal to that value. In protists, values around 97% to 99% are often used[6,9,11,17,62-69] but rarely justified[70-74]. Swarm[21,75] uses a local clustering value, resulting in OTUs whose radii that can be larger or smaller than 97% similarity. **a)** V4 amplicons from rainforesr soil samples, **b)** V9 amplicons from *Tara*-Oceans consortium open-ocean samples, and **c)** V4 and **d)** V9 amplicons from marine near-shore samples. Only a small proportion of the OTUs have a 97%-radii.

a

Taxa	# of reads	# of OTUs
animals	5,715,300	4,374
plants	3,874,669	3,089
fungi	44,430,656	17,849
protists	46,652,187	26,860

b

Forest	observed # of OTUs	estimated # of OTUs	standard error
LaSelva	16,935	18,667	50
Barro	7,065	7,935	37
Tiputini	5,217	5,707	26

Extended Data Table 1 | Observed and estimated diversity. a, Number of observed reads and OTUs from eukaryotes found in three Neotropical rainforests. These V4 data likely underestimate the diversity of all taxa equally. In barcoding studies aimed at distinguishing species, the mitochondrial COI is used in animals[76,77], plastid genes and the nuclear ITS are used in plants[78,79], and the nuclear ITS is used in fungi[80]. For the protists, while the V4 is considered an excellent pre-barcode, other group specific DNA regions have been proposed to differentiate species[81-83]. **b,** Number of observed and estimated OTUs of just the protists in each of three forests.