Microb Ecol (2017) 74:923–936
DOI 10.1007/s00248-017-0996-9

METHODS

An 18S rRNA Workflow for Characterizing Protists in Sewage, with a Focus on Zoonotic Trichomonads

Julia M. Maritz[1] · Krysta H. Rogers[2] · Tara M. Rock[1] · Nicole Liu[3] · Susan Joseph[1] · Kirkwood M. Land[3] · Jane M. Carlton[1]

Received: 14 April 2017 / Accepted: 12 May 2017 / Published online: 24 May 2017

Abstract Microbial eukaryotes (protists) are important components of terrestrial and aquatic environments, as well as animal and human microbiomes. Their relationships with metazoa range from mutualistic to parasitic and zoonotic (i.e., transmissible between humans and animals). Despite their ecological importance, our knowledge of protists in urban environments lags behind that of bacteria, largely due to a lack of experimentally validated high-throughput protocols that produce accurate estimates of protist diversity while minimizing non-protist DNA representation. We optimized protocols for detecting zoonotic protists in raw sewage samples, with a focus on trichomonad taxa. First, we investigated the utility of two commonly used variable regions of the 18S rRNA marker gene, V4 and V9, by amplifying and Sanger sequencing 23 different eukaryotic species, including 16 protist species such as *Cryptosporidium parvum*, *Giardia intestinalis*, *Toxoplasma gondii*, and species of trichomonad. Next, we optimized wet-lab methods for sample processing and Illumina sequencing of both regions from raw sewage collected from a private apartment building in New York City. Our results show that both regions are effective at identifying several zoonotic protists that may be present in sewage. A combination of small extractions (1 mL volumes) performed on the same day as sample collection, and the incorporation of a vertebrate blocking primer, is ideal to detect protist taxa of interest and combat the effects of metazoan DNA. We expect that the robust, standardized methods presented in our workflow will be applicable to investigations of protists in other environmental samples, and will help facilitate large-scale investigations of protistan diversity.

Keywords Sewage · Protist · Zoonoses · Trichomonad · 18S rRNA amplicon sequencing · Environmental sequencing

Electronic supplementary material The online version of this article (doi:10.1007/s00248-017-0996-9) contains supplementary material, which is available to authorized users.

✉ Jane M. Carlton
jane.carlton@nyu.edu

[1] Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA

[2] Wildlife Investigations Laboratory, California Department of Fish and Wildlife, Rancho Cordova, CA 95670, USA

[3] Department of Biological Sciences, University of the Pacific, Stockton, CA 95211, USA

Introduction

Microbes are the most abundant and diverse organisms in the biosphere, detected in almost every ecosystem, *e.g.*, in and on humans [1], pets [2], the built environment [3], soil [4], and the ocean [5]. Microbial communities are composite populations of thousands of microorganisms whose collective presence and relative abundance reflect conditions of the surrounding environment. During the last decade, high-throughput sequencing technologies have revolutionized our understanding of these complex systems and their implications for human health. For example, sewage contains microorganisms from human and animal waste as well as from groundwater, soil, and other environments [6]. Sewage is also a major reservoir for human and animal pathogens that are known to vary based on the host source of the waste, exposure to which may pose severe threats to environmental and public health [7]. Additionally, recent studies have shown that sewage accurately reflects the microbial composition of human stool, and it may be possible to identify host-specific microbes from sewage that could serve as indicators of fecal pollution sources [8–10].

The majority of these studies, however, have primarily aimed to characterize the prokaryotic diversity of these communities. In contrast, the protistan component of these ecosystems remains relatively unexplored, largely due to a lack of standard marker genes and reference databases. Protists are important components of terrestrial and aquatic environments, where they are integral constituents of trophic chains and nutrient cycles [11]. This includes human-made ecosystems such as wastewater treatment facilities, where they play roles in the purification process [12]. Human and animal microbiomes are also home to various protist species whose relationships with their hosts vary from parasitic to mutualistic [13]. Zoonotic (i.e., transmissible between humans, domesticated animals, and wildlife) protists such as species of *Giardia*, *Blastocystis*, *Cryptosporidium*, and a variety of trichomonads, *e.g.*, *Tritrichomonas fetus*, are common parasites of humans, livestock, other domestic animals and wildlife, contributing to significant host morbidity and mortality [14]. Exposure to, and risks from, these diseases are compounded in urban environments where contact between host and reservoir species is increased. Despite their ecological and economic importance, little is known about the diversity, incidence, or emergence of zoonotic parasites—or protists in general—in urban environments.

Broad surveys of eukaryotes in sewage have been explored on a very limited basis but their composition has been shown to reflect contributions from various animal, human, and environmental sources [15]. Microbial surveys of raw sewage should reflect community patterns and present an ideal system to monitor zoonotic parasites. For example, recent studies suggest that trichomonads (anaerobic, flagellated protists belonging to the large and diverse groups Trichomonadea and Tritrichomonadea of phylum Parabasalia [16]) are crossing host boundaries [17]. The recent isolation of new species of avian trichomonads responsible for epidemic outbreaks in California with high genetic similarity to human trichomonads highlights the possibility of zoonotic transfer from humans to birds and/or vice versa [18]. With the co-habitation of birds and humans in many urban and suburban areas, bird feces containing these parasites could contaminate human water sources and present a public health threat. The ability to distinguish between and monitor these parasites in sewage may provide insight regarding their distribution and potential transmission routes.

Understanding the prevalence and distribution of trichomonads and other zoonotic protists in sewage samples requires accurate methods for detection and identification. Current methods for high-throughput eukaryotic diversity studies rely on sequencing variable regions of the small ribosomal subunit (18S rRNA gene); however, no single region is universally accepted for environmental marker gene studies [19]. Several previous *in silico* and high-throughput studies have been conducted which suggest the V4 and V9 regions are the most variable, and best suited for microbial eukaryote studies [20–22]. While the utility of the V4 and V9 regions has been discussed in previous literature [23–26], no studies have conducted a direct comparison between primer sets using Illumina technology, nor have many studies investigated their resolution at deeper taxonomic levels, such as between closely related species or strains of zoonotic taxa. Additionally, parameters of sample collection and processing, DNA extraction, and sequencing protocols need to be evaluated, which are particularly important in studies of eukaryotes due to the potential masking effects of host DNA. Some studies have measured the impact of masking effects on the recovery of bacterial communities from stool samples, but such potential biases have yet to be assessed for protists or sewage samples [27].

Here, we describe an optimized workflow for the detection and analysis of protists in sewage samples, with a focus on zoonotic and trichomonad taxa, based on high-throughput amplicon sequencing of existing 18S rRNA markers. First, using Sanger sequencing and *in silico* testing methods, we compared the abilities of two regions (V4 and V9) to distinguish between a variety of human and animal-infectious protist taxa likely to be present in sewage, including *Cryptosporidium parvum*, *Giardia intestinalis*, *Toxoplasma gondii*, and several species of trichomonad. We then developed optimized protocols for sample processing and Illumina sequencing of these regions for raw sewage collected from a private apartment building in New York City. Our study also provides effective methods for high-throughput library construction and deep sequencing for generating high-quality sequencing data. These data provide a standard protocol for the detection of zoonotic protists in sewage, and pave the way forward for further investigation in sewage and other environmental samples.

Methods

Protist and Vertebrate Samples

Genomic DNA was eluted from *C. parvum*, *T. gondii*, *Blastocystis hominis*, *G. intestinalis*, rat, chicken, dog, and horse, following instructions specified by each provider. Genomic DNA was extracted directly from frozen stabilates of *Monotrichomonas carabina*, *Ditrichomonas honigbergii*, *Trichomitus batrachorum*, *Monocercomonas colubrorum*, all *Trichomonas gallinae* stabilates from the American Type Culture Collection (https://www.atcc.org), and both *Tetratrichomonas gallinarum* stabilates, using DNAZol and following the manufacturer's instructions. Other protist species were obtained as genomic DNA, including *Entamoeba invadens* and *Entamoeba histolytica* (from Dr. Daniel Eichinger, New York University School of Medicine); *Dientamoeba fragilis* (from Dr. Graham Clark,

London School of Hygiene and Tropical Medicine); and *Trichomonas tenax* (from Dr. Andrew Brittingham, Des Moines University, Iowa). The scientific name, strain, geographical location, and reference of all known organisms used in this study are shown in Online Resource 1.

For trichomonad species, wild caught birds (including Band-tailed pigeons, Eurasian collared doves, Mourning doves, Ring-necked pheasants, White-winged doves) were sampled between June 2014 and January 2015 in the state of California, USA. Sterile cotton-tipped applicator swabs moistened with sterile saline were used to collect oral swabs from each bird. After collection, swabs were used to inoculate InPouch TF (BioMed Diagnosticts), a transport and culture device designed for the detection and growth of *T. foetus* parasites that has also been shown to work for avian isolates [28]. Inoculated InPouches were incubated anaerobically at 37 °C and examined by microscopy for the presence of protozoa. Positive InPouches, or those containing motile trichomonads as determined by microscopy, were then subcultured into Peptone Yeast Extract Maltose (TYM) medium. Axenic cultures were obtained using subsequent culturing over several weeks in TYM as well as treatment with anti-fungal drugs as described previously in [18, 29]. Isolates still growing after 1 week of antibiotic treatment were then subjected to several rounds of serial dilution to generate parasites for DNA extraction. Genomic DNA was extracted from clonal isolates using a DNAEasy Blood & Tissue extraction kit (QIAGEN, catalog #69504).

Sewage Sample Collection and DNA Extraction

Two 50 mL samples of raw sewage were collected from the aerated feed tank in the basement of a private apartment building in New York City in July and September of 2014, and transported in secure containers to New York University. Sewage samples were handled under Bio Safety Level 2 conditions in a laminar flow hood, with the handler wearing personal protective clothing and goggles. Fresh DNA extractions were performed using 11 1 mL aliquots per sample on the day of sample collection with the PowerSoil DNA Isolation kit (MOBIO, catalog #12888) following manufacturer's instructions. DNA from 10 of the 11 aliquots was concentrated post-extraction using a SpeedVac (Savant) and subsequently pooled to represent DNA from 10 mL of sewage. This resulted in four samples per time point, two representing 1 mL of sewage and two pooled 10 mL samples. DNA extractions were performed on an additional 10 mL of sewage from each of the July samples after they were stored for 1 week at −20 °C as previously described. No extractions were performed on frozen sewage from the September samples. DNA concentration was quantified using the Qubit dsDNA HS Assay kit (Invitrogen, catalog #Q32851).

18S rRNA Amplification for Sanger Sequencing

Primers TAReuk454FWD1 (5′-CCAGCASCYGCGGTAATTCC-3′), TAReukREV3 (5′-ACTTTCGTTCTTGATYRA-3′) [24] were used to target the V4 region of the 18S SSU rRNA gene, and universal primers 1391f (5′-GTACACACCGCCCGTC-3′ [30]) and EukBr (5′-TGATCCTTCTGCAGGTTCACCTAC-3′ [31]) were used to target the V9 variable region. Fragments were amplified from total genomic DNA using either Phusion High-Fidelity DNA Polymerase (NEB, catalog #M0530S) or Phusion High-Fidelity PCR Master Mix (Thermo Scientific, catalog # F-531S) in a 50 μL reaction volume, with 1–5 μL input DNA (depending on concentration), PCR grade water, and a final primer concentration of 0.5 mM. Amplification conditions for the V4 region were 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 49 °C for 30 s, 72 °C for 30 s, and a final step of 72 °C for 10 min. Reaction conditions for the V9 region were 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 62 °C for 30 s, 72 °C for 30 s, and a final step of 72 °C for 10 min. PCR products were visualized on a 1% agarose gel, successful amplifications were purified using a 1.8X ratio of Agencourt AMPure XP beads (Beckman Coulter, catalog #A63880) and sent to Genewiz for Sanger sequencing. For templates that did not amplify initially, 1.5 μl of DMSO per reaction was added and the PCR was repeated.

18S rRNA Amplification for Illumina Sequencing

Environmental DNA extracted from sewage samples was prepared for Illumina sequencing targeting the V4 region, and also the V9 region with and without the addition of the mammal blocking primer. The V4 region was amplified with Illumina primer constructs containing the TAReuk454FWD1 and TAReukREV3 primers; no blocking primer is available for this region. Library synthesis and amplification were performed in triplicate using Phusion High-Fidelity PCR Master Mix (Thermo Scientific, catalog #F-531S), a 20 μL reaction volume, and a two-step PCR amplification strategy as described in [24]: 98 °C for 30 s, 10 cycles of 98 °C for 10 s, 53 °C for 30 s, 72 °C for 30 s; and then 25 cycles of 98 °C for 10 s, 48 °C for 30 s, 72 °C for 30 s, and ending at 72 °C for 10 min.

The V9 fragment of the 18S rRNA gene was amplified using Illumina primer constructs containing the universal primers 1391f-EukBr [23], and the mammal blocking primer designed as described in [32]. Library synthesis and amplification using 5 μL of input DNA was done in triplicate following the Earth Microbiome protocol [33, 34]. All V9 region primers, including the blocking primer, and protocols for amplification and sequencing are available on the EMP website (http://www.earthmicrobiome.org/emp-standard-protocols/18s/).

We included several negative controls, including extraction blanks for DNA purification experiments, and no-template blanks as negative controls in PCR reactions. No bands were visible on agarose gels after amplification, and no DNA was able to be quantified using the Qubit dsDNA HS Assay kit, and thus these control samples were excluded from downstream Illumina sequencing. After amplification, triplicate PCR reactions were pooled and purified with a 1.8X ratio of AMPure XP beads. The size distribution of purified libraries was determined using the 2100 Bioanalyzer or 2200 TapeStation (Agilent Technologies), and quantified via qPCR using the Library Quantification Kit—Illumina/LightCycler® 480 (KAPA Biosystems, Roche® LightCycler 480).

Quantified libraries were individually normalized to 4 nM based on qPCR and Bioanalyzer values and equal amounts of each 4 nM dilution were pooled. Eight samples were multiplexed per Illumina sequencing run. MiSeq preparation and sequencing was performed based on the manufacturer's and Earth Microbiome protocols [35] using the following parameters. Pools for the V4 region were sequenced at a final concentration of 12 pM with a 10% PhiX control spike-in using an Illumina MiSeq 500 cycle V3 kit and 2 × 300 run configuration. Pools for the V9 region were sequenced at a final concentration of 10 pM with a 6% PhiX control spike-in using an Illumina MiSeq 300 cycle V2 reagent kit with a 2 × 100 run configuration.

Analysis of Sanger Sequences

Sanger sequences for the V4 and V9 regions were processed in Geneious 7.1.7 [36]. Sequences were trimmed of poor quality areas using an error probability limit of 0.05, assembled using the de novo option and any remaining primer sequences were removed from the consensus sequence. We used the 18S rRNA SSU sequences in GenBank for *D. fragilis Bi/PA* (U37461), *G. intestinalis* Portland-1 (M54878), and *E. invadens* (AF149905) in all analyses, since these samples did not produce high-quality sequences for both regions.

Multiple sequence alignments were created using the MUSCLE [37] alignment option within Geneious using default parameters, and are available in FASTA format for the V4 data in Online Resource 2 and in Online Resource 3 for the V9 data. MrBayes 3.2.2 [38, 39] was used for phylogenetic analysis of trichomonad sequences and run using 10^6 generations, a sampling frequency of 500, and the GTR substitution model with gamma-distributed rate variation and a proportion of invariable sites. The resulting trees were visualized using FigTree (v1.4.2) [40].

All Sanger sequences from the V4 and V9 regions used in this study were de-replicated and clustered into OTUs at both 97 and 98% identity cutoff values using USEARCH v8.0.1 [41]. GenBank sequences were used where high-quality Sanger sequences were not obtained. Taxonomy was assigned to representative OTU sequences using BLAST [42] within QIIME against the full QIIME compatible SILVA database [43] version 111 (http://www.arb-silva.de/download/archive/qiime/) and an in-house curated version (see below) with an e-value of 1e-15.

Data Analysis of Illumina Sequences

Processing of Illumina sequencing reads from the sewage data was performed at the same time for both regions. Paired-end reads were joined within the QIIME 1.8.0 pipeline [44] using fastq-join [45] with a minimum overlap of 10 bp and allowing a 15% error rate in the overlapping area. Joined reads were de-multiplexed and quality filtered (split_libraries_fastq.py, -q 19 –r 5 –p 0.70), and any reverse primers detected were truncated (truncate_reverse_primer.py). Due to the poor quality of read 2 for the V4 region (average quality value, Q30, of 30.9%), read 1 was trimmed to 250 bp using Trimmomatic [46], then de-multiplexed using the parameters above. De-multiplexed reads for both regions were subject to *de novo* chimera checking, removal of singletons, and clustered into OTUs at 98% identity following the UPARSE pipeline (USEARCH v8.0.1, [47]). Taxonomy was assigned to representative OTU sequences using BLAST within QIIME, first against our curated SILVA database (see below), and subsequently with the QIIME formatted SILVA 111 database clustered at 99% identity. OTUs with no significant hits (<90% identity) after both rounds of taxonomic assignment were labeled as "Unidentified." Both datasets were filtered to remove non-18S sequences (bacterial and archaeal OTUs) and low abundance OTUs making up <0.0005% of reads in the total dataset as recommended for Illumina sequencing data [48].

Eukaryotic OTU tables were rarefied to 600,000 sequences per sample with 10 repetitions prior to alpha diversity analysis. Alpha diversity analyses (Phylogenetic Diversity metric, Shannon Index) were carried out using the QIIME pipeline on the 10 rarefied OTU tables. For metrics requiring a phylogenetic tree, representative OTU sequences were aligned to the curated database using PyNAST [49] at an identity threshold of 70%, a minimum length of 70 bp, and filtered to remove gaps present in >98% of sequences, followed by construction of phylogenetic trees using the default settings of FastTree [50]. Results were plotted using the ggplot2 [51] R package and statistically compared using either the Wilcox Rank Sum or Kruskal-Wallis tests in R. Univariate tests for differentially abundant taxa with respect to extraction volume and primers used was performed on sum-normalized phylum level taxonomic summaries of non-rarefied eukaryotic OTU tables using LEfSe [52] with default settings (Alpha = 0.05, LDA > 2, All-against-all or strict comparison). Samples were grouped into classes and analyzed based on extraction volume, primers used, or both depending on the amplicon region.

Eukaryotic OTU tables were further filtered to represent only protist taxa by excluding all metazoan, fungi, and plant sequences, and low abundance protist OTUs were removed as described above. The resulting OTU tables were rarefied stepwise from 60,000 to 600,000 sequences with 10 repetitions per level and rarefaction curves of the Phylogenetic Diversity metric were generated for each region.

SILVA Database Curation

The SILVA version 111 QIIME formatted database was downloaded (ftp://ftp.microbio.me/pub/QIIME_nonstandard_referencedb/Silva_111.tgz) and the 60,584 eukaryotic sequences from the full, unaligned files without ambiguous bases (Silva_111_full_unaligned_no_ambig.fasta) were filtered to remove all "unidentified," "uncultured," "clone," and "environmental sample" sequences. The remaining eukaryotic sequences were further curated with a focus on protist taxa. This included removing and or re-annotating mislabeled sequences and adding taxonomic placeholders representing super-groups, subphyla, *etc*. Sequences from protists of interest missing from the original database were downloaded from GenBank and added to the curated version. A list of the GenBank accession numbers and taxonomic strings of the sequences added can be found in Online Resource 4. A template alignment for this database was created by aligning it to the eukaryotic representative sequence file clustered at 99% identity (99_Silva_111_rep_set_euk_aligned.fasta), provided in the original QIIME formatted database, using PyNAST and an identity threshold of 60%.

Data Accessibility

Sequence data have been deposited in National Center for Biotechnology Information public databases. Sanger sequences have GenBank accession numbers KU939249, KU939251-KU939275, KU939280-KU939294, KU939300-KU939316, KU939319, KU939320, KU939326, KU939328-KU939352, KU939357-KU939371, and KU939377-KU939394. Raw Illumina sequencing data is archived in the Short Read Archive (SRA) under BioProject PRJNA315104. Mapping files for each 18S region, including SRA IDs, the curated SILVA database, and the eukaryotic abundance filtered OTU tables used for diversity analyses in this study can be downloaded from Figshare [http://dx.doi.org/10.6084/m9.figshare.3114850].

Results

We used two approaches for our method optimization (illustrated in Fig. 1). In part one, we tested the abilities of two variable regions of the 18S rRNA gene, V4 and V9, to detect and distinguish between a variety of human and animal infectious protist taxa likely to be present in sewage using Sanger sequencing. We also used these sequences as a mock community to evaluate parameters for protist community analysis. In part two, we developed experimental protocols for processing and sequencing of raw sewage samples, and further evaluated the use of V4 and V9 regions for protist diversity studies and techniques.

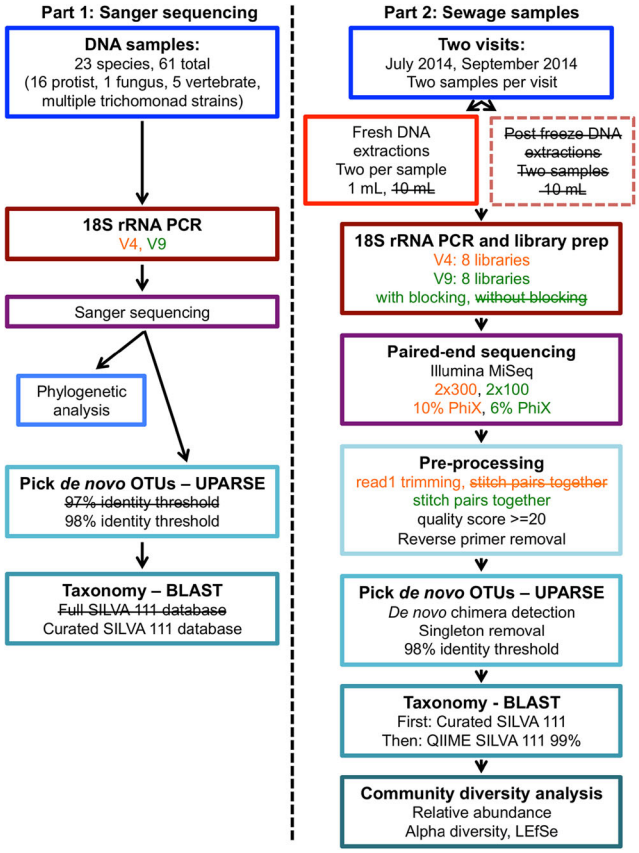

Fig. 1 Workflow employed for testing and optimization of protocols to detect microbial eukaryotes in raw sewage. Part 1 was used to evaluate two variable regions of the 18S rRNA gene, V4 and V9, and determine best practices for zoonotic protists. Part 2 was used to optimize the methods in Part 1 and develop experimental methods for Illumina sequencing of raw sewage samples. Methods that were tested but determined suboptimal are indicated in *boxes with dotted lines* and/or *crossed out text*

Evaluation of Published 18S rRNA Primers and Reference Databases

We first tested the ability of two of the most variable regions of the 18S rRNA gene, V4 and V9, which are widely used in studies of eukaryotic microbial biodiversity, to distinguish between a variety of protist species and distinguish protist DNA from vertebrate DNA likely to be present in sewage. Primer pairs TAReuk454FWD1 and TAReukREV3, and 1391f and EukBr were used to amplify the 18S rRNA gene of 23 different eukaryotic species including 16 protists, 1 fungus, and 5

vertebrates. We also included multiple isolates of several species of trichomonads, totaling 61 DNA samples in all (Online Resource 1). Amplicons ranging from 270 to 387 bp in length for the V4 region and 96–134 bp for the V9 region were subsequently sent for Sanger sequencing. High-quality sequences were obtained for all samples using V4 primers except *G. intestinalis*. High-quality sequences were obtained for all but two samples for the V9 region (*D. fragilis* and *E. invadens*).

V4 region amplicon sequences had a pairwise identity (PI) of 69.7% (Online Resource 2). All four mammalian species had a PI >99% relative to one other, but were distinguishable from the chicken DNA sample. Phylogenetic analysis of the trichomonad V4 sequences is shown in Fig. 2a, where they formed five different groups of identical sequences. Two of the five are composed exclusively of *Trichomonas vaginalis* sequences (Fig. 2a**,** green), and two others are almost entirely *T. vaginalis*-like sequences (Fig. 2a**,** brown and yellow). The fifth group shows the most diversity, with a mixed representation of *T. vaginalis*, *T. tenax*, *T. gallinae*, and *T. gallinarum* sequences (Fig. 2a, orange). All five of these clusters are >97% identical to each other (the standard cutoff used to differentiate OTUs in prokaryotes). V9 region amplicon sequences had an average pairwise identity of 79.4% (Online Resource 3) and a similar pattern as the V4 between the vertebrate species. Phylogenetic analysis of the trichomonad V9 sequences (Fig. 2b) showed less variation within the trichomonad clade compared to the V4 data. Four trichomonad species, *T. vaginalis* (all 24 strains), *T. vaginalis*-like (all 11 strains), *T. tenax*, *T. gallinae* (all four strains), and one strain of *T. gallinarum* (TP-79), had identical sequences for this region. The other strain of *T. gallinarum* (Leverett) was distinguishable from the other trichomonads with a PI of 96% to the trichomonad group.

Next, we used our Sanger sequences as input for a mock community analysis to evaluate the phylogenetic resolution and taxonomic accuracy of these primers in a microbial community setting. OTU clustering of the 63 Sanger sequences from the V4 region, representing 23 different species, returned 19 OTUs at 97% identity and 20 OTUs at 98% identity. At 97% identity, the 42 sequences from the 4 different trichomonad species clustered into 2 OTUs, 1 representing 41/42 sequences and the other representing *T. gallinarum* Leverett (Online Resource 5). The four mammalian species clustered into one OTU. At 98% identity, the trichomonad sequences clustered into three distinct OTUs, one representing 23/24 *T. vaginalis* sequences, the second representing all *T. gallinae* (4 sequences) and *T. vaginalis*-like (11 sequences), *T. tenax*, one *T. gallinarum* (TP-79), and one *T. vaginalis* (C1:NIH) (Fig. 2a, Online Resource 5). The third trichomonad OTU represents *T. gallinarum* Leverett. The 62 V9 Sanger sequences were clustered into 18 OTUs at both the 97% and 98% identity cut off values. The sequence composition of these OTUs was identical using both methods and was in accord with the phylogenetic trees. The pattern of sequence clustering was similar to that of the V4 at 97% identity, with all trichomonad sequences clustered together in one OTU with the exception of *T. gallinarum* Leverett, and the four species of mammals represented by one OTU (Fig. 2b, Online Resource 6).

We then assigned taxonomy to these OTUs using the full SILVA 111 QIIME formatted database and compared the taxonomy of the recovered hit to that of the known query. Primers for the V4 region showed higher taxonomic accuracy than the V9, matching the query ID for 14/20 OTUs at the species level and 15/20 at the genus level. The V9 correctly identified 8/18 OTUs at the species level and 13/18 at the genus level. The unrecovered OTUs were identified as either other species in the same genus or closely related taxa (Online Resource 5 and Online Resource 6). Reference sequences for three taxa (*M. carabina*, *E. histolytica*, *E. invadens*) were not present in the database used for identification.

We curated the existing SILVA 111 QIIME reference database by adding sequences for these missing taxa and others of interest (see Materials and Methods and Online Resource 4). This resulted in a total of 46,094 eukaryotic sequences with curated taxonomy. We then re-assigned taxonomy to the OTUs obtained from our Sanger sequences. After database curation, all 20 V4 OTUs were correctly identified at the genus level, and all except *B. hominis* were correctly identified at the species level (Online Resource 5). Of the V9 OTUs, 11 out of 18 were correctly recovered to the species level, including *M. carabina* and *E. invadens* and 16 out of 18 to the genus level (Online Resource 6). The OTU representing the four mammal sequences was identified as a closely related species, but not one of the four member taxa.

Development of Experimental Protocols for Raw Sewage

We collected samples of raw sewage from a private apartment complex in New York City in order to develop sample processing protocols and further test the utility of the V4 and V9 18S rRNA regions for protist diversity studies. This green apartment building has its own wastewater and rainwater recycling system where treated water is reused for toilets, laundry facilities, and garden irrigation. This “closed” system represents an ideal site for methods testing because the sewage there is mostly from human residents, with little environmental input, which allows for evaluation of the background of metazoan DNA and provides a control for protists that are human based that may be obscured by other taxa in more open systems. Two 50 mL samples were collected at two time points, July and September 2014, and DNA extracted (Fig. 1).

First, we determined optimal preprocessing conditions for sewage samples considering the constraints of same-day extraction, technician fatigue for processing multiple samples,

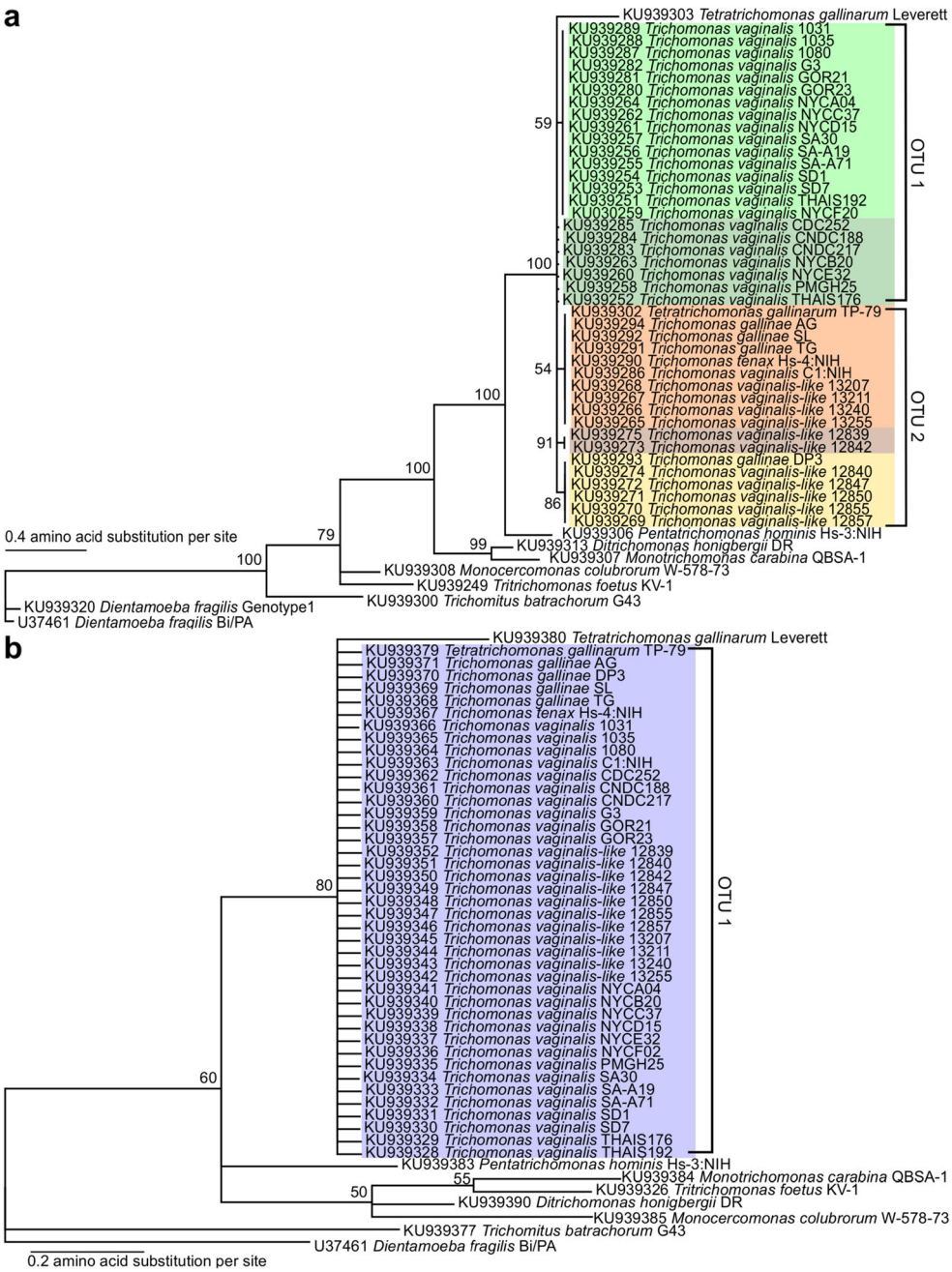

Fig. 2 Phylogenetic tree of trichomonad Sanger sequences from **a** V4 region and **b** V9 region. Numbers at nodes correspond to Bayesian posterior probabilities. GenBank accession numbers are shown at the left for each taxon. Colored taxa are identical to each other and brackets indicate OTU membership and corresponding ID number from the supplementary material after clustering at 98% identity

and potential bias that can be introduced as a result of suboptimal collection and storage conditions. We compared the yield of DNA extracted on the day of collection from 10 mL of sewage to that extracted from 1 mL, for both July samples. DNA extracted from 10 mL of sewage gave higher yields than from 1 mL (Table 1). We also tested stability of DNA extracted from sewage that was processed immediately, compared to an additional 10 mL DNA extraction from each July sample that had been frozen for a week at −20 °C. The amount of DNA extracted from the frozen sample was dramatically lower than from the fresh sample, with an average loss of 20.2 ng/uL (Table 1). Based on these results, we confined the rest of our study to extractions performed on the day of sewage collection.

Next, we compared the effects of extraction volume and use of a vertebrate blocking primer (V9 region only) on the diversity of sewage. We collected an additional two samples in September and extracted DNA from 1 and 10 mL of sewage each on the day of collection. Illumina sequencing libraries were generated from total DNA from fresh extractions for the V4 region (8 libraries total), and for the V9 region with and without the blocking primer (16 libraries total), and run on a MiSeq to generate paired-end (PE) reads. Sequencing of the V4 libraries resulted in 18,269,491 raw PE reads. However,

Table 1 DNA yield from July sewage samples with and without storage at −20 °C for 1 week

Sample ID	Day extracted	Input volume	DNA yield (ng/μL)
S005_1_mL	Day of collection	1 mL	2.9
S005_10_mL	Day of collection	10 mL	35.0
S005_10_mL_frozen	After 7 days at −20 °C	10 mL	5.81
S006_1_mL	Day of collection	1 mL	3.56
S006_10_mL	Day of collection	10 mL	21.0
S006_10_mL_frozen	After 7 days at −20 °C	10 mL	9.8

the quality of the sequencing reads was low; read 1 quality scores decreased significantly in the last 50 bp, and read 2 had an average Q30 of 30.9% (a successful V3 2 × 300 run should have a Q30 >70%). As a result, we were unable to join the majority of read pairs (Table 2). To obtain the highest quality and deepest sequence coverage for downstream analysis, we compared the amount of usable reads returned post de-multiplexing and quality filtering for joined read pairs, read 1 only, and read 1 trimmed to remove the last 50 poor quality bases. Un-joined read 1, trimmed to 250 bp, returned the greatest amount of high-quality (average Q score >19) reads for downstream analysis (Table 3). All further V4 analyses were conducted on this dataset.

De-multiplexed reads were next clustered into OTUs at a 98% identity threshold and filtered to remove low abundance OTUs. This resulted in a total of 6,464,532 eukaryotic sequences and 824 OTUs after data processing and filtering for the V4 region (Table 3). For the V9 region, a total of 29,400,737 PE reads were generated from 2 MiSeq runs. The quality of these reads was much higher compared to the V4 sequences and we were able to join the majority of read pairs resulting in 18,664,054 eukaryotic sequences and 1444 OTUs after filtering (Table 3).

Analysis of Illumina Sequence Data from Raw Sewage Samples

We used QIIME to calculate alpha diversity, a measure of the mean species diversity, for the data from the V4 and V9 libraries. The alpha diversity of the V4 region was always lower than that of the V9 region, regardless of the initial volume of sewage (Fig. 3a). Alpha diversity of both 18S regions was lower for 1 mL samples than 10 mL samples (Fig. 3a), although the only significant pairwise comparison was the V9 phylogenetic diversity metric (Wilcox test, $p < 0.001$). For samples with the blocking primer (V9 region only), the higher extraction volume continued to show higher alpha diversity than the 1 mL extractions; however, the addition of the blocking primer reduced the phylogenetic diversity of those samples (Fig. 3b). Differences in extraction volume and primer set had small effects on alpha diversity independently, but together produced significant differences for the V9 region (Kruskal-Wallis test, $p < 0.01$).

Table 2 Reads returned after joining and/or de-multiplexing the V4 sequencing data

Analysis step	Joined	Read 1 only	Read 1 trimmed
Raw sequences or pairs	18,269,491	18,269,491	18,269,491
Joined pairs of reads	53,526	NA	NA
De-multiplexed reads	42,257	6,906,307	7,027,716

We calculated the relative sequence abundance of taxonomic assignments for all samples and 18S regions (Fig. 4). Ciliates were the dominant group (30–70% of all taxa) in sewage irrespective of extraction volume, blocking primer used, or region amplified. In particular, species of Oligohymenophorea and Phyllopharyngea, free-living bacteriovorous and common freshwater protist inhabitants of sewage were most abundant [12, 15, 53, 54]. Other highly abundant taxa included bacteriovorous flagellates belonging to the Chrysophyceae, which are typically found in fresh water and soil (1–40%). Free-living (*D. honigbergii*) and gut associated (*Pentatrichomonas hominis*, *D. fragilis*) trichomonads, along with a species originally isolated from avian sources (*Trichomonas* sp.) were detected by both regions at low abundances (≤1–5%) in sewage. We also detected low amounts of fungi, and protist species (*Entamoeba* and *Blastocystis*) that are characteristic of the mammalian gut [55]. Higher abundances of these protists were observed in the V9 samples amplified using the blocking primer than in either the V9 samples amplified without it or the V4 region. Other protist taxa of interest (*Cryptosporidium* and *Toxoplasma*) were present at very low levels (<1%) and we did not detect *Giardia* in any of our Illumina sequencing data. The V4 region showed more variable distributions of taxa between the conditions and an increase in uncultured environmental taxa included with "other protists"; in contrast, the V9 had a higher proportion of unidentified OTUs.

The amount of metazoan DNA detected by both regions ranged from <1 to 25% and consisted mostly of invertebrate taxa. The most abundant metazoan groups identified were rotifers, nematodes, and annelids, and much lower levels (<1% in all samples and regions) of human and other mammal DNA were detected. In data amplified with the V4 region, the highest amount of metazoan DNA was present in the 10 mL extraction

Table 3 Summary of Illumina sequence data for each 18S region

Analysis step	V4 region (read 1)	V9 region
Raw sequences or pairs	18,269,491	29,400,737
Joined pairs of reads	NA	22,864,953
De-multiplexed reads	7,027,716	21,395,270
Clustered reads (OTUs)	6,477,688 (2270)	21,205,088 (8827)
Eukaryotic reads (OTUs)	6,477,688 (2270)	18,794,581 (7195)
Filtered eukaryotic reads (OTUs)	6,464,532 (824)	18,664,054 (1444)
Average sequences per sample	808,066	1,196,152

volumes (6–25%; Fig. 4a). High levels of metazoan DNA were also found in the V9 samples amplified without the blocking primer, with the highest amounts present in the 10 mL sample size (13–18%; Fig. 4b). The V4 region showed a much larger range of metazoan DNA abundances and identified a wider variety of metazoan taxa than the V9 region. The lowest overall amounts of metazoan DNA was present in the 1 mL volumes amplified with the blocking primer (<1–2%; Fig. 4b).

Linear discriminant analysis (LDA) effect size (LEfSe) analysis was used to identify biomarkers (enriched or differentially abundant taxa) associated with sample groupings (Fig. 5). Samples from a 10 mL extraction volume were broadly enriched for metazoan and fungal (V9 only) phyla. For the V4 region, these included Rotifera and Platyhelminthes (Fig. 5a), and Rotifera, Cinardia, Choanomonada, and a variety of fungi for the V9 data (Fig. 5b). No clades were consistently present in all 1 mL samples amplified using V4 region primers. Samples from a 1 mL extraction volume amplified with V9 region primers were enriched for protist phyla and those amplified using the blocking primer were enriched for protists of interest, such as Parabasalia (trichomonads), Archamoebae (*Entamoeba*), Opalinata (*Blastocystis*), and Apicomplexa (Fig. 5b). Phylogenetic diversity-based rarefaction curves of protist-only OTU tables approach an asymptote around 350,000 protist sequences for the V4 region and around 400,000 for the V9 region (Online Resource 7). However, after removing host and other non-target sequences, some samples from a 10 mL extraction volume did not contain enough sequences to adequately capture the sample's protist diversity.

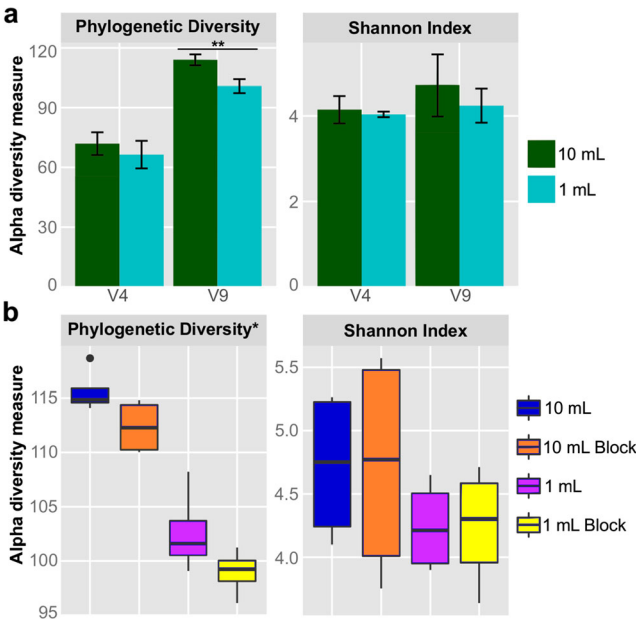

Fig. 3 Alpha diversity analysis for sewage 18S communities using the phylogenetic diversity metric and Shannon Index. All analyses were calculated from replicate OTU tables sampled at a depth of 600,000 sequences. **a** Alpha diversity of the different extraction volumes for the V4 and V9 regions. Values shown represent the measurements from all 1 and 10 mL (fresh samples only) extraction volumes by region. *Error bars* represent ± one standard deviation; *double asterisks* indicates *p* value <0.001. **b** Alpha diversity of sewage samples based on extraction volume and blocking primer use for the V9 region. *Asterisk* indicates *p* value <0.01

Discussion

In this study, we present a method for detecting microbial eukaryotes in raw sewage, while combating the possible effects of metazoan DNA. To our knowledge, this is the first study to optimize methods for investigation of zoonotic protists in raw sewage samples. We first tested the ability of primers from two 18S rRNA variable regions, the V4 and V9, to amplify a representative sample of known animal-infecting protists such *C. parvum*, *G. intestinalis*, *T. gondii*, and several species of trichomonads. Sanger sequencing demonstrated that both amplicons could be used to distinguish between a variety of protist taxa that may be present in sewage samples. The V4 region showed more sequence variability and higher taxonomic accuracy than the V9 region, which is consistent with results of other studies and could suggest that it is more variable at finer scales of taxonomic resolution [19, 24]. High levels of sequence similarity (above the 97% level commonly used to differentiate OTUs in bacteria) were observed in both regions, particularly between mammalian and trichomonad taxa. Due to this similarity, neither region was able to produce OTUs resolving all individual species in our mock analysis.

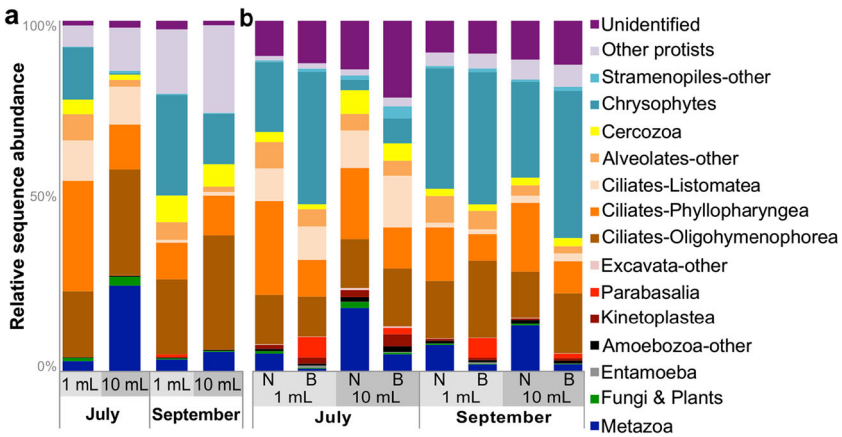

Fig. 4 Relative sequence abundances of eukaryotic taxa across experimental conditions for each region. Each *bar* represents the average of two replicates per condition per time point. All data shown were from fresh extractions. **a** V4 region results from the 1 and 10 mL extractions. **b** V9 region results from the 1 and 10 mL extractions (with (B) and without (N) the blocking primer)

As predicted, the V4 region was better for discriminating between *Trichomonas* and *Tritrichomonas* genera than the V9 region. Although the V4 region could not be used to further differentiate isolates among the group of *Trichomonas* species in the mock analysis, this region could be useful to rapidly distinguish trichomonads that infect non-human mammals, such as dogs and cats, from those that infect avian species. The close grouping of avian *T. vaginalis*-like and human *T. vaginalis* isolates confirms a possible zoonotic potential for these parasites [17]. Monitoring their presence in sewage and other environmental matrices could prove valuable to public health.

We further evaluated the use of the V4 and V9 regions for protist diversity studies by Illumina sequencing raw sewage samples. The V9 region produced a larger number of OTUs at every stage of analysis than the V4 region, even though these represent highly conserved three-domain primers that also amplified low numbers of Eubacteria and Archaea. This conflicts with the data obtained from our analysis of Sanger sequences. Fewer lineages (Phylogenetic Diversity) and lower Shannon Index values were also observed for the V4 data compared to V9 data. This suggests that the primers for the V9 region may have a broader recognition spectrum than the V4 [19, 23]. This may, however, also be a product of the

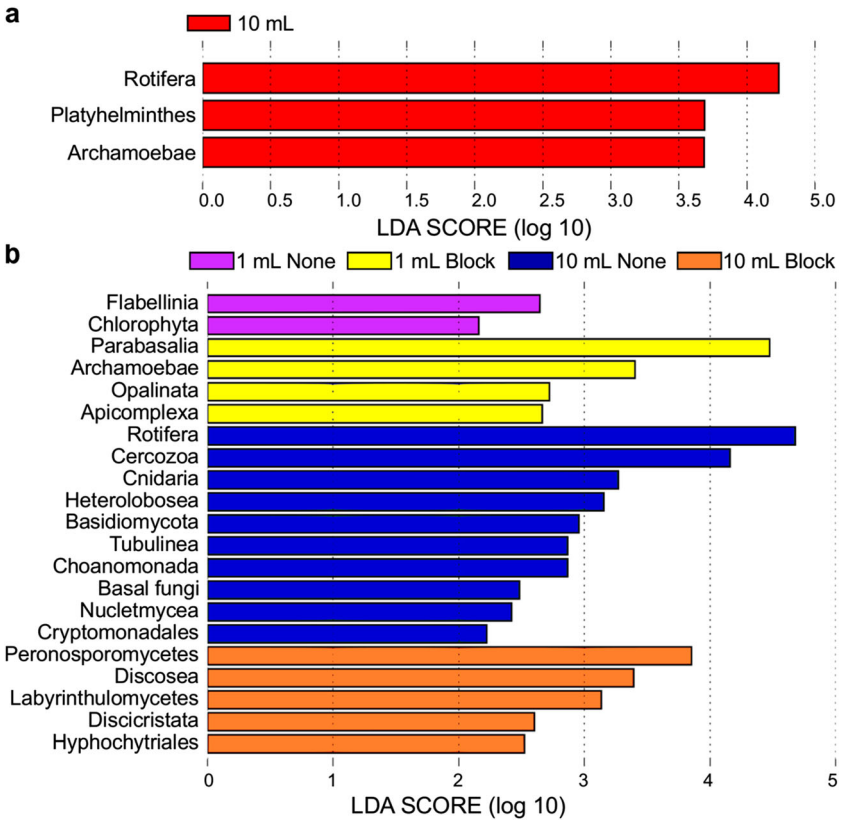

Fig. 5 LEfSe analysis of 18S communities to determine biomarker taxa across sample groupings for each region. All data shown were from fresh extractions. LDA scores for taxa identified as differentially abundant between **a** 10 and 1 mL for the V4 region and **b** 1 mL, 1 mL blocking, 10 mL, and 10 mL blocking samples for the V9 region. Taxa are ranked according to their effect size (LDA score) and associated with the group with the highest median

significantly lower quality and quantity of sequencing reads produced for the V4 region. Previous studies also observed high sequencing error rates in the V4 region due to significant length polymorphism and secondary structure [24, 56, 57]. Additionally, this dataset only used read 1 from the V4 Illumina data as we were unable to join enough read pairs for use in downstream analysis. Performance of the V4 region may improve if full-length sequences can be obtained.

The V4 and V9 regions produced broadly similar protist taxonomic profiles across sewage samples. Each primer set also showed differences in the distribution and overall diversity of taxa detected. In particular, the V9 probe detected much higher levels of Chrysophyte, and unidentified sequences, while the V4 region detected more Cercozoan and uncultured environmental sequences. The V9 region also detected higher levels of protist taxa of interest, particularly trichomonads, than the V4 region. In each case, most of the variation between regions was due to the presence of taxa not recovered by the other primer set. This is likely due to a combination of factors including bias in different primers sets that cause preferential amplification of particular protist taxa, the degree of variability present in each SSU region between particular taxa, and unbalanced representation of sequences in existing databases. Overall, our results confirm previous studies that suggest the V9 region provides a more comprehensive overview across all eukaryotes, and the V4 can better discriminate between some closely related taxa.

Previous studies showed that employing multiple primer sets increases the number of protist species detected in environmental samples [23, 24, 58]. The results of our study lead us to concur and recommend the use of both the V4 and V9 markers to provide a more accurate picture of eukaryotic diversity and better taxonomic resolution of zoonotic protists than either one on its own. Additionally, using both markers facilitates direct comparison with other datasets, as many studies use at least one of these primer sets, and provides a better level of standardization for future studies.

We conducted several experiments to determine optimal methods for processing and sequencing of sewage samples. Our study found that freezing sewage samples dramatically decreased the DNA yield, even from large volumes (Table 1), highlighting the importance of extracting DNA on the day of collection. Comparison of extraction volume and use of a vertebrate blocking primer (V9 region only) in sewage showed that samples with larger extraction volumes have higher overall levels of diversity. These samples, however, were also enriched for metazoan taxa and required deeper sequencing depths to adequately capture the protist diversity present. Surprisingly, the majority of metazoan DNA detected in this study represented invertebrate or nematode taxa, not humans or other mammals, and was successfully reduced by the blocking primer, even though it is designed for vertebrate species. This suggests that the V4 and V9 primers may be inherently biased against recovering vertebrate 18S DNA, and the recovery of vertebrate DNA in raw sewage is not as much of a concern as might be expected. Although samples with 1 mL extraction volumes had reduced eukaryotic diversity, these samples showed higher overall abundances of protist taxa. When combined with the blocking primer, the 1 mL samples were enriched for trichomonads and other taxa of interest, compared to the 10 mL samples. Thus, for our protist-focused research, a combination of small 1 mL extraction volume and incorporating the blocking primer proved most ideal for detecting protist taxa of interest and combating the effects of metazoan DNA.

To expand the applicability of our method, we developed a reliable and optimized protocol for amplicon PCR, library construction, and sequencing of the 18S V4 and the V9 regions on the Illumina MiSeq. Our protocol takes advantage of several experimental techniques including bead-based PCR cleanup and size-based quantification of individual samples to produce high-quality sequencing libraries. This, combined with dilution and pooling of individual libraries for sequencing, resulted in more even sequencing coverage across samples, and reduced the amount of data lost in downstream processing. We also observed that a lower loading concentration, adjusted based on the final concentration of the library pool, increased the overall data output and made it possible to reduce the amount of PhiX control needed. We used as little as ~6% (V9 region, V2 kit) PhiX to produce reads of comparable or higher quality to normal Illumina runs. This strategy maximizes the amount and quality of data generated with less space dedicated to sequencing PhiX and increases the depth of coverage.

Sequencing coverage is an important factor in microbiome studies. Previous research showed that broad sampling with shallow coverage (as few as 100 sequences per sample) is sufficient to adequately capture the diversity present and reveal broad ecological patterns in prokaryotic communities [33, 35, 59]. However, studies of microbial eukaryotes are prone to contamination from non-target taxa such as host or food DNA and in some cases primers that also amplify non-18S rRNA targets. In general, deeper sequencing of these communities is required to provide sufficient coverage after removing non-target OTUs and to capture rare taxa or more subtle ecological effects. For this reason, we employed a deep-sequencing strategy, multiplexing only eight samples per Illumina run (≥800,000 average sequences per sample, Table 3). This strategy more than adequately captured the protist diversity present in sewage as rarefaction curves leveled off around 350,000–400,000 protist sequences. In future runs, we have the flexibility for some cost-cutting modifications of the protocol, such as multiplexing more samples, although we recommend generating a minimum of 200,000 raw sequences per sewage sample to account for host and other non-target sequences. Many of these steps can be adapted for use with automated robotic technology to use in large-scale studies.

OTU clustering was carried out in this study using a sequence similarity cutoff of 98%, as some previous studies suggest that 97% is too conservative for estimating diversity in microbial eukaryotes [24]. Our mock analysis of Sanger sequencing data supported this idea. More species-level OTUs were captured when clustering was performed at 98% similarity for the V4 region. The taxonomic assignment of these reads, however, is still only trustworthy at the genus level. We maximized this accuracy by using a curated SILVA database and adopting a two-step process to assign taxonomy to sewage OTUs. We chose this approach because several target taxa were missing from current databases. After database curation, we correctly identified all V4 OTUs and 16 out of 18 V9 OTUs to the genus level in our mock analysis. Our curated reference database only contains eukaryotic sequences, and when used alone, particularly for the V9 region, returns a large amount of unidentified OTUs due to its three-domain primers.

To prevent non-eukaryotic OTUs from artificially inflating this number, all unidentified OTUs were compared against the full SILVA database and any non-18S rRNA sequences were removed from the dataset. Some of the taxonomic levels added to our curated database, for example Hacrobia, do not reflect the current views of eukaryotic taxonomy from [16] or [60]; however, they were added in an effort to reduce the variability of taxonomic information present between major clades in the database and make summarizing taxonomic information (summarize_taxa.py) in QIIME easier. Even after use of these customized databases, unfortunately, significant amounts of unidentified OTUs remain. This further highlights the need for comprehensive eukaryotic microbe reference databases, and the vast number of eukaryotic microbes that remain to be described.

Our study aimed to provide a workflow for the detection and analysis of protists in sewage samples, with a focus on zoonotic and trichomonad taxa, based on high-throughput amplicon sequencing of existing 18S rRNA markers. As such, the method provides a snapshot of microbial eukaryotic diversity (living or dead) in sewage; we are unable to measure active vs. dormant microbes, nor assess what could be considered "residual" or "transitory" species. Additional complementary analyses will be needed to determine whether the detected species are metabolically active, and to examine the likely sources of these species. The 18S rRNA regions used in this study were limited in their ability to provide fine-scale taxonomic resolutions between species or strains of zoonotic taxa, particularly between closely related trichomonads; however, they can be used to track potential large-scale trends that may influence the distribution of zoonotic microbes in urban environments. Although detection of these biomarker taxa do not provide quantitative estimates of abundance, our method provides a jumping off point for future targeted study design and hypothesis testing to confirm the source, viability, and distribution of these pathogenic species.

Acknowledgements We thank the personnel of the NYU Center for Genomics and Systems Biology GenCore for their sequencing services, and Drs. Laura Wegener Parfrey and Thierry Heger for sharing their Illumina V4 primer constructs and amplification protocol. We also thank members of the Carlton lab, Dr. Steven Sullivan and Dr. Holly Bik, for their reading and discussion of the manuscript as well as the staff of the private housing complex wastewater treatment facility. J.M.M. is supported by the MacCracken Program in the Graduate School of Arts and Science at New York University and a Department of Biology Fleur Strand Fellowship. This study was funded by a New York University Grand Challenge project "Microbes, Sewage, Health and Disease: Mapping the New York City Metagenome" and a grant from the Alfred P. Sloan Foundation to J.M.C. K.M.L. was supported by the Department of Biological Sciences at the University of the Pacific.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Approval This study does not contain any studies with human participants performed by any of the authors. All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted. Bird sampling procedures were approved under the U.S. Fish and Wildlife Service Scientific Collecting Permit (MB03368B) issued to the Wildlife Investigations Laboratory, California Department of Fish and Wildlife.

References

1. Mitreva M (2012) Structure, function and diversity of the healthy human microbiome. Nature 486(7402):207–214. doi:10.1038/nature11234
2. Hand D, Wallis C, Colyer A, Penn CW (2013) Pyrosequencing the canine faecal microbiota: breadth and depth of biodiversity. PLoS One 8(1):e53115. doi:10.1371/journal.pone.0053115
3. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, Brown GZ, Green JL, Bohannan BJ (2014) Bacterial communities on classroom surfaces vary with human contact. Microbiome 2(1):7. doi:10.1186/2049-2618-2-7
4. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME journal 1(4):283–290. doi:10.1038/ismej.2007.53
5. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans c, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science 348(6237):1261359. doi:10.1126/science.1261359

6. Shanks OC, Newton RJ, Kelty CA, Huse SM, Sogin ML, McLellan SL (2013) Comparison of the microbial community structures of untreated wastewaters from different geographic locales. Appl. Environ. Microbiol. 79(9):2906–2913. doi:10.1128/AEM.03448-12
7. Newton RJ, Bootsma MJ, Morrison HG, Sogin ML, McLellan SL (2013) A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. Microb. Ecol. 65(4):1011–1023. doi:10.1007/s00248-013-0200-9
8. Cao Y, Van De Werfhorst LC, Dubinsky EA, Badgley BD, Sadowsky MJ, Andersen GL, Griffith JF, Holden PA (2013) Evaluation of molecular community analysis methods for discerning fecal sources and human waste. Water Res. 47(18):6862–6872. doi:10.1016/j.watres.2013.02.061
9. McLellan SL, Newton RJ, Vandewalle JL, Shanks OC, Huse SM, Eren AM, Sogin ML (2013) Sewage reflects the distribution of human faecal Lachnospiraceae. Environ. Microbiol. 15(8):2213–2227. doi:10.1111/1462-2920.12092
10. Newton RJ, McLellan SL, Dila DK, Vineis JH, Morrison HG, Eren AM, Sogin ML (2015) Sewage reflects the microbiomes of human populations. MBio 6(2):e02574. doi:10.1128/mBio.02574-14
11. Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, Knight R, Fierer N (2013) Global biogeography of highly diverse protistan communities in soil. The ISME journal 7(3):652–659. doi:10.1038/ismej.2012.147
12. Dubber D, Gray NF (2011) The influence of fundamental design parameters on ciliates community structure in Irish activated sludge systems. Eur. J. Protistol. 47(4):274–286. doi:10.1016/j.ejop.2011.05.001
13. Parfrey LW, Walters WA, Knight R (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. Front. Microbiol. 2:153. doi:10.3389/fmicb.2011.00153
14. Fletcher SM, Stark D, Harkness J, Ellis J (2012) Enteric protozoa in the developed world: a public health perspective. Clin. Microbiol. Rev. 25(3):420–449. doi:10.1128/CMR.05038-11
15. Korajkic A, Parfrey LW, McMinn BR, Baeza YV, VanTeuren W, Knight R, Shanks OC (2015) Changes in bacterial and eukaryotic communities during sewage decomposition in Mississippi river water. Water Res. 69:30–39. doi:10.1016/j.watres.2014.11.003
16. Adl SM, Simpson AG, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, Le Gall L, Lynn DH, McManus H, Mitchell EA, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick RS, Schoch CL, Smirnov A, Spiegel FW (2012) The revised classification of eukaryotes. J. Eukaryot. Microbiol. 59(5):429–493. doi:10.1111/j.1550-7408.2012.00644.x
17. Maritz JM, Land KM, Carlton JM, Hirt RP (2014) What is the importance of zoonotic trichomonads for human health? Trends Parasitol. 30(7):333–341. doi:10.1016/j.pt.2014.05.005
18. Girard YA, Rogers KH, Gerhold R, Land KM, Lenaghan SC, Woods LW, Haberkern N, Hopper M, Cann JD, Johnson CK (2014) Trichomonas stableri n. sp., an agent of trichomonosis in Pacific coast band-tailed pigeons (*Patagioenas fasciata monilis*). Int J Parasitol Parasites Wildl 3(1):32–40. doi:10.1016/j.ijppaw.2013.12.002
19. Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, Amaral-Zettler L, Guillou L (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. PLoS One 6(4):e18169. doi:10.1371/journal.pone.0018169
20. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C (2014) Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. PLoS One 9(2):e87624. doi:10.1371/journal.pone.0087624
21. Hu SK, Liu Z, Lie AA, Countway PD, Kim DY, Jones AC, Gast RJ, Cary SC, Sherr EB, Sherr BF, Caron DA (2015) Estimating protistan diversity using high-throughput sequencing. J. Eukaryot. Microbiol. 62(5):688–693. doi:10.1111/jeu.12217
22. Hugerth LW, Muller EE, Hu YO, Lebrun LA, Roume H, Lundin D, Wilmes P, Andersson AF (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS One 9(4):e95567. doi:10.1371/journal.pone.0095567
23. 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(7):e6372. doi:10.1371/journal.pone.0006372
24. Stoeck T, Bass D, Nebel M, Christen R, Jones MD, Breiner HW, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol. Ecol. 19(Suppl 1):21–31. doi:10.1111/j.1365-294X.2009.04480.x
25. Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A, Orsi W, Edgcomb VP (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biol. 7:72. doi:10.1186/1741-7007-7-72
26. Wang Y, Tian RM, Gao ZM, Bougouffa S, Qian PY (2014) Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. PLoS One 9(3):e90053. doi:10.1371/journal.pone.0090053
27. Flores R, Shi J, Yu G, Ma B, Ravel J, Goedert JJ, Sinha R (2015) Collection media and delayed freezing effects on microbial composition of human stool. Microbiome 3:33. doi:10.1186/s40168-015-0092-7
28. Bunbury N, Bell D, Jones C, Greenwood A, Hunter P (2005) Comparison of the InPouch TF culture system and wet-mount microscopy for diagnosis of Trichomonas gallinae infections in the pink pigeon *Columba mayeri*. J. Clin. Microbiol. 43(2):1005–1006. doi:10.1128/JCM.43.2.1005-1006.2005
29. Gerhold RW, Yabsley MJ, Smith AJ, Ostergaard E, Mannan W, Cann JD, Fischer JR (2008) Molecular characterization of the Trichomonas gallinae morphologic complex in the United States. J. Parasitol. 94(6):1335–1341. doi:10.1645/GE-1585.1
30. Lane DJ (1991) 16S/23S sequencing. In: Stackebrant E, Goodfellow M (eds) Nucleic acid technologies in bacterial systematic. Wiley, New York, pp. 115–175
31. Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene 71(2):491–499
32. Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples—a case study on prey DNA in Antarctic krill stomachs. Front. Zool. 5:12. doi:10.1186/1742-9994-5-12
33. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. 108(Suppl 1):4516–4522. doi:10.1073/pnas.1000080107
34. Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. BMC Biol. 12:69. doi:10.1186/s12915-014-0069-1
35. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal 6(8):1621–1624. doi:10.1038/ismej.2012.8
36. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649. doi:10.1093/bioinformatics/bts199

37. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792–1797. doi:10.1093/nar/gkh340
38. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755
39. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572–1574
40. Rambaut A (2014) FigTree v1.4.2. http://tree.bio.ed.ac.uk/software/figtree/.
41. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461. doi:10.1093/bioinformatics/btq461
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2
43. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41(Database issue):D590–D596. doi:10.1093/nar/gks1219
44. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7(5):335–336. doi:10.1038/nmeth.f.303
45. Aronesty E (2011) ea-utils: "Command-line tools for processing biological sequence data". http://code.google.com/p/ea-utils.
46. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. doi:10.1093/bioinformatics/btu170
47. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods 10(10):996–998. doi:10.1038/nmeth.2604
48. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods 10(1):57–59. doi:10.1038/nmeth.2276
49. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26(2):266–267. doi:10.1093/bioinformatics/btp636
50. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5(3): e9490. doi:10.1371/journal.pone.0009490
51. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York
52. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C (2011) Metagenomic biomarker discovery and explanation. Genome Biol. 12(6):R60. doi:10.1186/gb-2011-12-6-r60
53. Hu B, Qi R, An W, Yang M (2012) Responses of protists with different feeding habits to the changes of activated sludge conditions: a study based on biomass data. J. Environ. Sci. (China) 24(12):2127–2132
54. Moreno AM, Matz C, Kjelleberg S, Manefield M (2010) Identification of ciliate grazers of autotrophic bacteria in ammonia-oxidizing activated sludge by RNA stable isotope probing. Appl. Environ. Microbiol. 76(7):2203–2211. doi:10.1128/AEM.02777-09
55. Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, Kodira C, Mohiuddin M, Brunelle J, Driscoll M, Fierer N, Gilbert JA, Knight R (2014) Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. Front. Microbiol. 5:298. doi:10.3389/fmicb.2014.00298
56. Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environ. Microbiol. 11(5):1292–1302. doi:10.1111/j.1462-2920.2008.01857.x
57. Mahe F, Mayor J, Bunge J, Chi J, Siemensmeyer T, Stoeck T, Wahl B, Paprotka T, Filker S, Dunthorn M (2015) Comparing high-throughput platforms for sequencing the V4 region of SSU-rDNA in environmental microbial eukaryotic diversity surveys. J. Eukaryot. Microbiol. 62(3):338–345. doi:10.1111/jeu.12187
58. 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(1):31–43. doi:10.1016/j.protis.2005.10.004
59. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat Meth 7 (10): 813–819 doi:http://www.nature.com/nmeth/journal/v7/n10/abs/nmeth.1499.html - supplementary-information
60. Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ, Katz LA (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst. Biol. 59(5):518–533. doi:10.1093/sysbio/syq037