



Systematic Design of 18S rRNA Gene Primers for Determining Eukaryotic Diversity in Microbial Consortia

Luisa W. Hugerth¹, Emilie E. L. Muller², Yue O. O. Hu¹, Laura A. M. Lebrun², Hugo Roume^{2,3a}, Daniel Lundin^{1,3b}, Paul Wilmes^{2*}, Anders F. Andersson^{1*}

1 KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Stockholm, Sweden, **2** Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract

High-throughput sequencing of ribosomal RNA gene (rDNA) amplicons has opened up the door to large-scale comparative studies of microbial community structures. The short reads currently produced by massively parallel sequencing technologies make the choice of sequencing region crucial for accurate phylogenetic assignments. While for 16S rDNA, relevant regions have been well described, no truly systematic design of 18S rDNA primers aimed at resolving eukaryotic diversity has yet been reported. Here we used 31,862 18S rDNA sequences to design a set of broad-taxonomic range degenerate PCR primers. We simulated the phylogenetic information that each candidate primer pair would retrieve using paired- or single-end reads of various lengths, representing different sequencing technologies. Primer pairs targeting the V4 region performed best, allowing discrimination with paired-end reads as short as 150 bp (with 75% accuracy at genus level). The conditions for PCR amplification were optimised for one of these primer pairs and this was used to amplify 18S rDNA sequences from isolates as well as from a range of environmental samples which were then Illumina sequenced and analysed, revealing good concordance between expected and observed results. In summary, the reported primer sets will allow minimally biased assessment of eukaryotic diversity in different microbial ecosystems.

Citation: Hugerth LW, Muller EEL, Hu YOO, Lebrun LAM, Roume H, et al. (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

Editor: Christian R. Voolstra, King Abdullah University of Science and Technology, Saudi Arabia

Received: December 6, 2013; **Accepted:** March 28, 2014; **Published:** April 22, 2014

Copyright: © 2014 Hugerth et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the Swedish Research Councils VR [grant 2011-5689], FORMAS [grant 2009-1174] and EC BONUS project BLUEPRINT partially funded by FORMAS through grants to A.F.A, the Luxemburg National Research fund [grants ATTRACT/A09/03 and CORE/11/BM/1186762 to P.W., PHD-MARP-04 to H.R. and PRD-2011-1/SR to E.M.] and the Chinese Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: paul.wilmes@uni.lu (PW); anders.andersson@scilifelab.se (AA)

^{3a} Current address: University of Ghent Laboratory of Microbial Ecology and Technology, Ghent, Belgium

^{3b} Current address: Ecology and Evolution in Microbial model Systems, Linnaeus University, Kalmar, Sweden

Introduction

The accurate identification of diversity is a key challenge in microbial ecology research. Molecular biology techniques, most notably polymerase chain reaction (PCR)-based amplification and sequencing of the resulting small-subunit ribosomal DNA amplicons, have revolutionized our view of microbial diversity by unveiling a tremendous diversity of bacteria, archaea and eukarya in different environments [1–4]. Recent developments in high-throughput sequencing make deep sequencing of hundreds of samples achievable at affordable costs. This is in turn transforming microbial ecology into a quantitative research field in which, for example, models of spatio-temporal patterns of microbial diversity can be built and tested [5–11], and links between microbial community composition and host genotype [12] as well as phenotype [13–16] can be established. So far, massively parallel sequencing technology has mostly been applied for addressing bacterial and archaeal diversity. However, recently, 454 pyrosequencing has been conducted on 18S rDNA amplicons and revealed an unprecedented diversity of eukaryotes in a range of environments [17–26].

Current high-throughput sequencing platforms can generate billions of short reads, often below 200 bp. Accurate identification

of microbial taxa with this limited amount of information demands a diligent choice of PCR primers, so as to simultaneously prioritize the sequencing of highly informative regions and avoid biases caused by unevenly amplifying different taxa. The size of the amplicon is also a limiting factor for most high-throughput technologies. Considerable efforts have been put into designing optimal primers for bacterial and archaeal identification [6;27–34]. However, despite the efforts of Amaral-Zettler *et al*, Nolte *et al* and Stoeck *et al* [17,36,18], no equally broad and systematic design of eukaryotic primers has yet been conducted by leveraging the richness of sequences in current databases. In fact, several recent studies perform high-throughput sequencing while using primers never fully assessed for this, such as the ones in the pioneering work of Pace and Sogin [1,35] or only assessed for DGGE [37–39]. Both these cases highlight the need for broad taxonomy-range primer pairs for the generation of information-rich amplicons.

Here, in an effort to develop 18S rDNA primer sets compatible with current high-throughput sequencing technologies and which should generate minimally biased, phylogenetically discriminatory sequences, we have used all non-redundant full-length 18S rDNA sequences from the SILVA database [40] to identify optimal primer pairs for partial 18S sequencing. We did this by first

identifying a set of degenerate primers with high coverage among reference sequences and evaluating the taxonomic distribution of their matches. We then combined these into a set of candidate primer pairs for which we assessed *in silico* the phylogenetic information that would be generated with each primer pair when using different massively parallel sequencing technologies. One of the primer pairs that performed best *in silico* was optimised for PCR and subsequently used to amplify the V4-V6 region of the 18S rDNA sequences from a range of environments that were subsequently sequenced by Illumina technology.

The developed primer pairs are optimally suited for identifying eukaryotic taxa through 18S rDNA sequencing using read lengths typically obtained using massive parallel sequencing, which can be as short as pairs of 150 bp paired-end reads or 400 bp single-end reads, while still resulting in phylogenetically discriminatory data.

Materials and Methods

Degenerate Primer Design

The “SSU Ref NR” aligned sequences were downloaded from release 111 of the SILVA database (www.arb-silva.de; [40]), and eukaryotic sequences were extracted from these based on their annotation. After first trimming the alignment to only include positions represented by 90% of the sequences (while maintaining information on where nucleotides had been removed), the program DegePrime (<https://github.com/EnvGen/DEGEPRIME>; Hugerth *et al.* submitted) was used to generate 18 bases long degenerate primers at every alignment position, with maximum degeneracy 12. From these, the number of sequences from each taxon matched by the primers was counted. To generate the modified primers 574* and 616*, sequences corresponding to the taxa that were missed by primers 574 and 616, respectively, were extracted from the SILVA database and DegePrime was run on them. By comparing the resulting oligonucleotides, we identified the degeneracies that should be added to each primer and, in the case of position 574, where the 18th nucleotide would be N, we opted to remove this base.

In Silico Simulations

All reads were simulated from the “SSU Ref NR” database. Error free reads were simulated starting at each primer sequence(s), and extending for 150, 250 or 400 bases, with read length including primer length. When counting unique sequences, in case the paired reads for a primer pair did not span the entire amplicon, they were concatenated and, in case the distance between a primer pair was shorter than their combined read length, the overlapping ends were merged.

For the BLAST-based analyses, reads from 1000 unique sequences were selected at random, including only sequences of defined phylogeny matched by all candidate primers. Stand-alone BLASTN was used for matching reads to sequences in the original database. Hits were kept if they matched at least 95% of the read length with 95% or 99% identity, as stated in the results section, and with an *e*-value below 10E-5. In the case of paired-ended reads, the hit with the lowest combined *e*-value was kept. *E*-values were combined through the mean of their log *e*-values, which corresponds roughly to averaging their bit-scores. Taxonomic accuracy was assessed only in the cases where the selected hit was annotated at least to family level. Distances between full-length query and hit sequences were calculated on Mothur [41], disregarding terminal gaps from the original aligned SILVA file and treating each string of gaps as a single gap. All statistical analyses and graphs were produced in R, with the additional library Vioplots.

DNA Extraction, Amplification and Sequencing

No animals were killed for the purpose of this study. Moose were shot by licensed hunters during Swedish hunting season in 2012. Rumen content was sampled post-mortem, for which no ethical permission is required under Swedish law. The human faeces sample was first used in Forberg's *et al* study [42] and all permissions therein apply.

DNA fractions were extracted from wastewater and moose rumen (*Alces alces*) according to the protocol in Roume *et al* [43] and from soil and marine sediment using the PowerSoil kit (MO BIO Laboratories Inc, Carlsbad CA, USA). One marine sample was extracted using the PowerWater kit (MO BIO Laboratories Inc, Carlsbad CA, USA), the other as described by Riemann *et al* [44], and further purified by ethanol precipitation. The faecal sample was extracted with DNeasy Blood & Tissue kit (Qiagen, Venlo, Netherlands). Optimization of PCR conditions for selective amplification of 18S rDNA was initially carried out with DNA extracted from a pure culture of *Saccharomyces cerevisiae* for eukaryotes as positive control, *Escherichia coli* for bacteria and *Halobacterium* sp. NRC-1 for archaea as negative controls. The best temperature for primer annealing was chosen based on the intensity of the PCR product of the expected size, but also to minimise non-specific products. All primers were ordered dry-frozen from MWG (Ebersberg, Germany). The reaction mixture for each primer pair consisted of 25 μ L of Kapa HiFi Mastermix (Kapa Biosystems, Woburn MA, USA), 2.5 μ L of each primer (10 μ M) and 2.5–7.5 ng of DNA template, depending on sample purity and the proportion of eukaryotic DNA in it. PCR was performed on a Mastercycler Pro S (Eppendorf, Hamburg, Germany). Cycling conditions are 95°C for 5', 98°C for 1', 20–25 cycles of 98°C for 20", annealing temperature for 20" and 72°C for 12" followed by a final elongation step of 72°C for 1'. Annealing temperatures for each primer are presented in table 1. Gel electrophoreses (1% agarose in TAE buffer 1 \times) were carried out to check the size and quality of PCR products.

To prepare libraries for Illumina sequencing, primers 574*f and 1132r were prolonged by a handle, yielding the primer pair 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCGGTAATTCCAGCTCYV-3' and 5'-AGACGTGTGCTCTTCCGATCTACTTAAAGRVATTGACGG-3'. Samples were amplified as described above, with an annealing temperature of 51°C, and cleaned as described by Lundin *et al* [45] with 15% PEG 6000, reducing the sample volume to 23 μ L. To this, 25 μ L of Kapa HiFi Mastermix were added and 1 μ L of each of the primers 5'-AATGATACGGC-GACCACCGAGATCTACAC-X₈-ACACTCTTTCCCTACACGACG-3' and 5'-CAAGCAGAAGACGGCATACGAGAT-X₈-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3', where in each case X₈ represents an 8-bp DNA barcode, so that each sample has a unique combination of forward and reverse. Reaction conditions for this second PCR are 95°C for 1' and ten cycles of 98°C for 10", 62°C for 30" and 72°C for 15", followed by a final amplification step of 1' at 72°C. The product of this reaction was cleaned once more using the same method [45] and sequenced in SciLifeLab/NGI (Solna, Sweden) on a MiSeq (Illumina Inc, San Diego, CA, USA). For more detailed and updated protocols, see <https://github.com/EnvGen/LabProtocols>.

Analysis of Sequencing Data

All sequences generated in this study and at any part of the analysis are available upon request. Raw sequences were trimmed for quality using FastX (http://hannonlab.cshl.edu/fastx_toolkit/links.html), trimming bases with a Phred score below 30 and

Table 1. Oligomers evaluated for their value as 18S rDNA primers.

Primer ID	Yeast 5' position	Sequences matched	Sequence	GC range (%)	Tm range (°C)	Reference
391f	391	26191	YGGAGARGGAGCHTGA	50–67	48.0–54.9	This study
550f	550	29782	GGRCMAGBCTGGTGCCAG	61–78	52.6–59.4	
563f	563	29188	GCCAGCAVYCGCGTAAY	56–72	50.3–57.2	
574f	574	27099	CGGTAAYTCCAGCTCYAV	44–61	45.8–52.6	
574*f	574	29271	CGGTAAYTCCAGCTCYV	47–65	44.6–51.9	
616f	616	27631	TTAAAVGYTCGTAGTYG	28–44	38.9–45.8	
616*f	616	28072	TTAAARVGYTCGTAGTYG	28–50	38.9–48.0	
897f	897	27836	AGAGGTGRAATTCTHRGA	33–44	41.2–45.8	
1132f	1132	29866	AYTTRAGDAATTGACGG	28–44	38.9–45.8	
1132r	1150	29866	CCGTCAATTHCTTYAART	28–44	38.9–45.8	
1182f	1182	29514	AATTYGACTCAACDCRGG	39–56	43.5–50.3	
1266f	1266	29500	RGTGGTGCATGGCCGYTB	56–72	50.3–57.2	
1423f	1423	28365	AACAGGTCHGWRTGCCC	50–61	48.0–52.6	
1423r	1441	28365	GGGCATYWCDGACCTGTT	50–61	48.0–52.6	
1612r	1630	26791	ACAAAKGGCAGGACDYA	44–61	45.8–52.6	
1626r	1644	27697	GACRGGMGGTGTGBACAA	48–67	50–54.9	
1380F	1625	27667	CCCTGCCHTTGTACACAC	53–58	51.1–53.2	Amaral-Zettler, 2009 (17)
1389F	1634	27279	TTGTACACACCGCCC	60	44.7	
1510R	1787	1699	CCTTCYGCAGTTACCTAC	55.60	53.8–55.9	
1391F	1770	27674	GTACACACCGCCGTC	69	51.1	Stoeck, 2010 (18)
EukB	3'	4	TGATCCTCTGCAGGTTACCTAC	50	57.4	
Fwd1	564	27796	CCAGCASCYCGGTAATTCC	60–65	55.9–57.9	
Rev3	981	22175	ACTTTCGTTCTTGATYRA	28–39	38.9–43.5	
fw	366	10744	ATTAGGTTTCGATCCGGAGAGG	50	61.4	Nolte, 2010 (36)
rv	586	27376	CTGGAATTACCGCGSTGCTG	57.1	58.3	

Oligomers evaluated for their value as 18S rDNA primers. Primer ID, where stated, refers to the ID used in the original paper. Primer position is based on *S. cerevisiae*, GenBank accession number Z75578. Primers that were too far downstream to be found in this gene are indicated as "3'". The lowercase letters 'f' and 'r' refer to whether the primer was evaluated for use as forward or reverse. Tm was calculated using Melting Temperature (Tm) Calculation (http://www.biophp.org/minitools/melting_temperature/demo.php).

doi:10.1371/journal.pone.0095567.t001

deleting forward and reverse reads in case any of them was shorter than 150 bp. For clustering, forward reads were trimmed to 230 bp and reverse reads to 180 bp; reads shorter than this were excluded. They were concatenated and clustered using Usearch [46] at decreasing similarity levels of 100%, 99%, 98% and 97%. After clustering, the representative sequence for each cluster was separated into its forward and reverse component. Each read, or representative sequence, as described in the main body of text, were classified using SINA v1.1.13 [40]. The longest consensus taxonomy between forward and reverse reads was taken for the amplicon. Graphs and statistical analyses were produced in R, using libraries MASS, Vegan and Gplots. The taxonomic distribution of reads in environmental samples were produced in Krona (<http://sourceforge.net/projects/krona>). For more detailed and updated protocols, see <https://github.com/EnvGen/Tutorials>.

Results

Identification of Broad Taxonomic Range PCR Primers

In order to identify highly conserved regions in the 18S rDNA, as well as candidate PCR primers from these, we used the program DegePrime, which, for every position of a multiple sequence alignment, outputs an oligomer of defined length and maximum degeneracy that matches an as large number of sequences as possible. Figure 1 shows the results for 31,862 non-redundant 18S rDNA sequences downloaded from the SILVA database. The potential for matching of the best primer for each position largely mirrors the entropy at the same position ($R^2 = 0.89$).

Fifteen oligomers were selected which are located in conserved regions while being flanked by variable regions. These matched 82.2–92.6% of the total pool of 31,862 full-length eukaryotic SILVA sequences (Table 1). The majority of the selected primers match all major eukaryotic lineages represented in SILVA (Fig. 2; Table S1). However, there were notable exceptions, such as the lack of coverage of some primers to the phyla Nematoda, Microsporidia or Discoba, or the Centrohelida class. To partially recover the coverage of Discoba by the candidate primer 616,

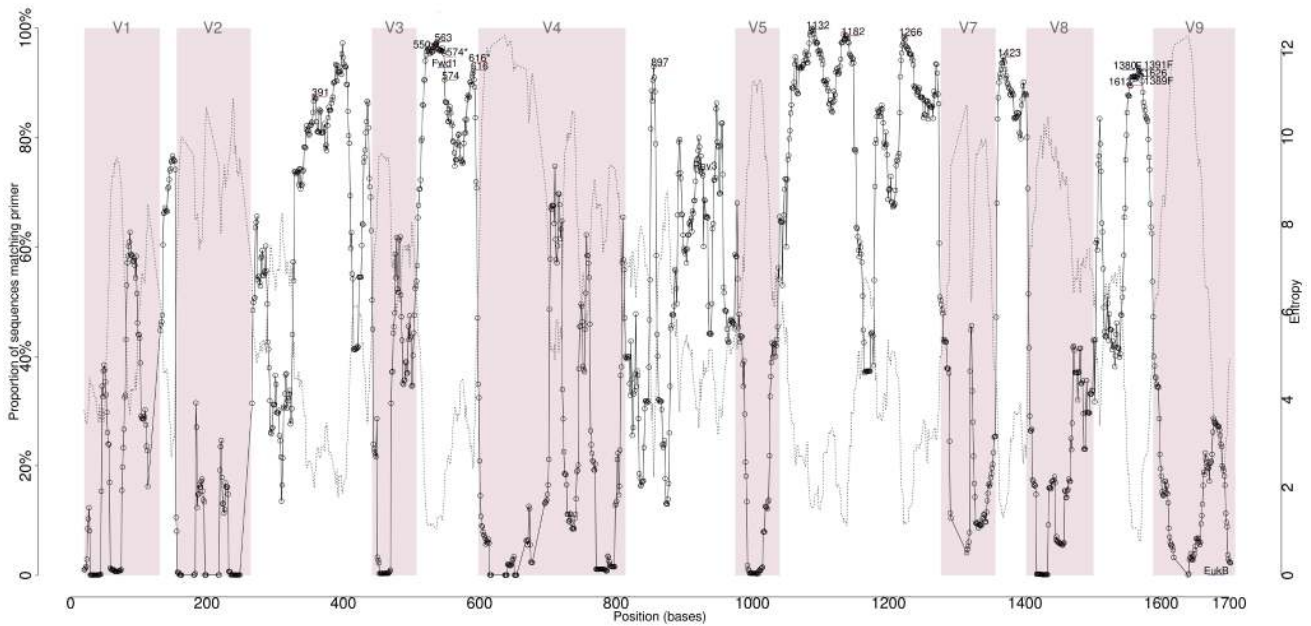


Figure 1. Position and coverage of candidate primers. Eighteen bp oligomers with 12 degrees of degeneracy were designed to match as many of the sequences as possible at each position of an alignment of 31,862 full-length unique eukaryotic 18S rDNA sequences using the DegePrime program. The proportion of the sequences matched by the best oligomer found for each position is depicted in black, with a line connecting adjacent points. The entropy of each position is depicted by a dotted grey line. The position numbering refers to the *Saccharomyces cerevisiae* strain FM-sc-08 18S ribosomal RNA gene, NCBI accession number Z75578. Dark red horizontal bars represent the oligomers chosen as candidate primers in this study. Primers which were later altered are marked in lighter red. Primers found in the literature are depicted in dark blue. doi:10.1371/journal.pone.0095567.g001

another degree of degeneracy was added to the sixth base, and this alternative version was called 616*. In the case of candidate primer 574, coverage of the phyla Rodophyceae, Metamonada and Nematoda was recovered with a 17 bp alternative in the same position. Commonly used 18S annealing primers from the literature were included, and these retain the name used in the original work [17,18]. They generally display good taxonomic coverage, although Rev3 and EukB were excluded from the analysis since they target the distal 3' end of the 18S rDNA, which has a much sparser representation in the SILVA database (Table S1). Primers selected at this stage were combined into seven different primer pair combinations (primer pair 563–1132, 574–1132, 616–1132, 897–1423, 1132–1423, 1132–1391F and 1266–1391F), covering variable regions 4, 6, 7 and 8 and amplifying regions between 167 and 569 bp in length. These pairs were considered throughout the rest of the analysis. When used for the simulation of single-end reads, the letters ‘f’ and ‘r’, for “forward” and “reverse”, respectively, are added to the primer number to indicate the sense of the read.

Assessing the Information Content of Amplified Fragments

In addition to matching a taxonomically broad range of organisms, a good primer pair should amplify sequences that contain as much phylogenetic information as possible. To quantify phylogenetic resolution, we performed three simulations on eukaryotic SILVA sequences which involved the *in silico* generation of amplicons and associated sequencing reads. For the simulations, we used 400 bp single reads reflecting the output of single-end sequencing strategies using 454 or IonTorrent technologies, and 2×150 bp and 2×250 bp paired reads reflecting the output of paired-end sequencing strategies using Illumina technology.

Sequence reads were extracted directly downstream of each primer. Since, typically, different measures of error correction and/or OTU clustering are performed after sequencing, substantially lowering the effect of sequencing errors, we only considered error-free sequences for the subsequent analyses.

The first assessment of the approach is based on the idea that, ideally, two 18S rDNA sequences that differ when comparing their full-length sequences should also differ in the sequencing reads obtained. To quantify this, we assessed how many unique artificial read sequences the non-redundant eukaryotic SILVA database would generate using the different primer pairs, read lengths and types (singlets/pairs). Figure 3 shows the percentage of unique amplicon sequences compared to the number of unique full-length sequences. Numbers differ substantially between primer pairs and read lengths, with the final number of unique reads ranging from 51.3% to 78.3% of the original number of unique sequences. Interestingly, 400 bp single-end reads have similar performance to paired reads of 150 bp or longer. Thus, this figure highlights how the choice of region can play a more significant role than the choice of sequencing technology and read length in resolving eukaryotic phylogenetic diversity by 18S rDNA amplification and high-throughput sequencing.

Estimating Taxonomic Assignment Accuracy

In microbial diversity studies, obtained sequences (either individually or as representatives of operational taxonomic units) are typically compared to a database, and the taxonomic information of the best match is used to describe the sequence, provided that the match fulfils some criteria on similarity. In the second test, we simulated this situation by sub-sampling 1,000 random sequences from a subset of the non-redundant SILVA database containing only sequences matching all selected primers

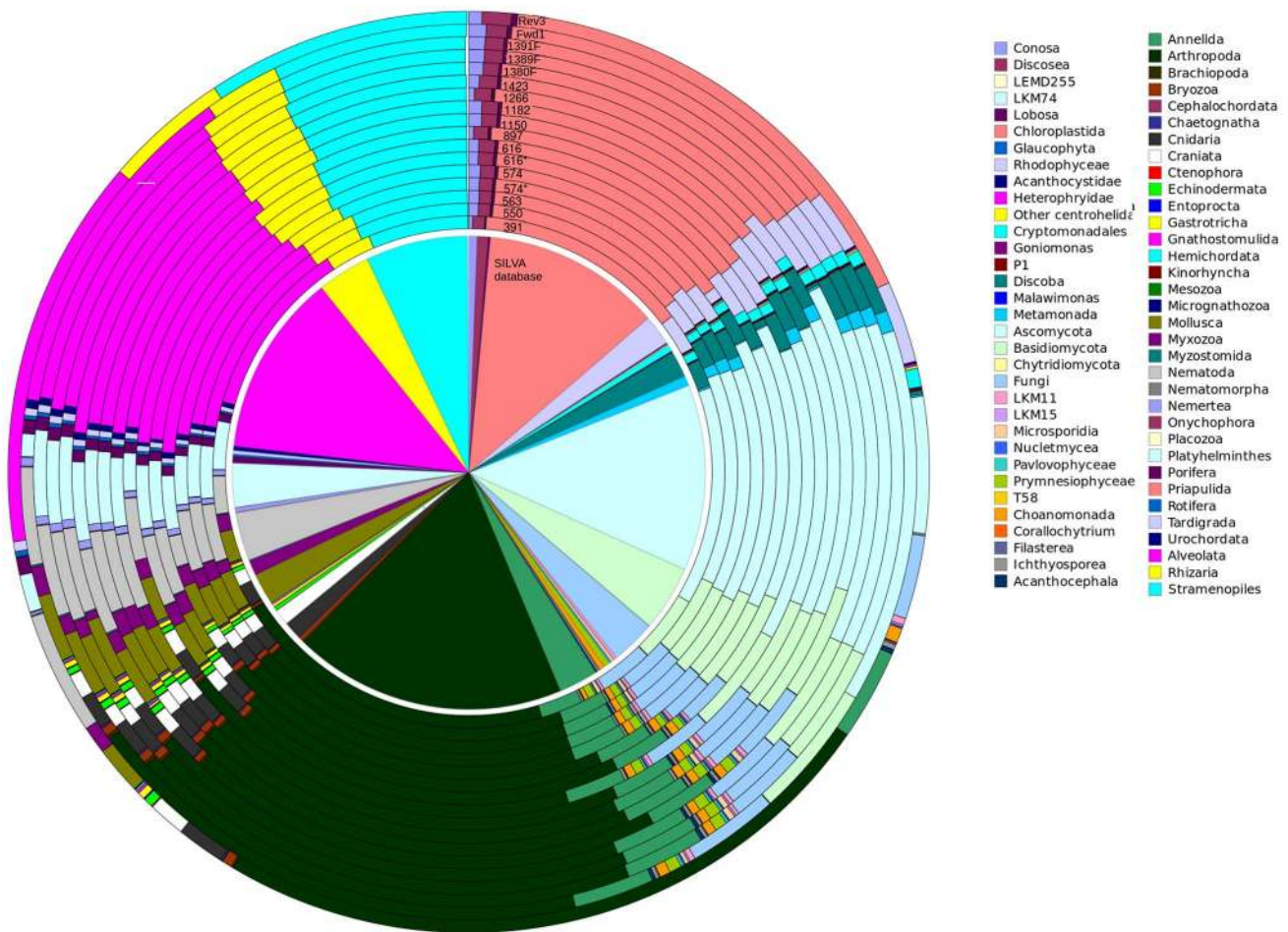


Figure 2. Taxonomic distribution of sequences matching candidate primers. The central circle represents the taxonomic distribution of the SILVA eukaryotic database. Each outer ring corresponds to the taxonomic distribution of sequences matching each primer candidate. Primers are marked in the figure and each colour corresponds to a kingdom or phylum as shown in the legend. doi:10.1371/journal.pone.0095567.g002

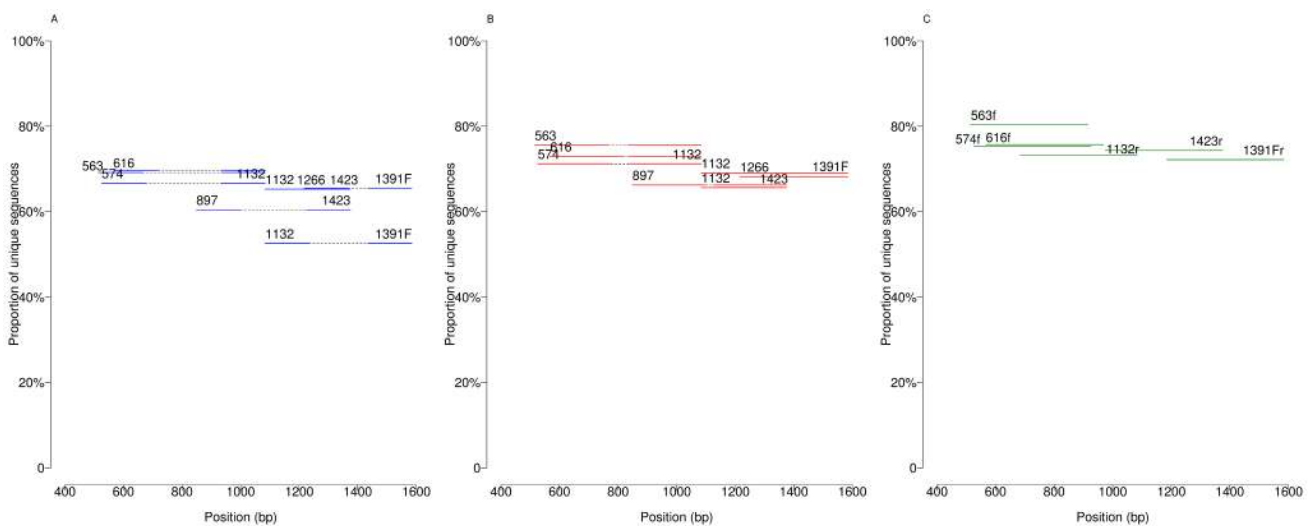


Figure 3. Ratio between number of unique amplicon sequences and full-length sequences. The ratio between the number of unique amplicon sequences and unique near full-length sequences (starting at primer 391 and ending at 1786), for different primer pairs and read lengths/types. (A) Paired-end 150 bp reads (B) Paired-end 250 bp reads (C) Single-end 400 bp reads. Paired-end reads are connected by a black dashed line. doi:10.1371/journal.pone.0095567.g003

and with fully defined phylogeny. Error-free reads were extracted for all selected primer pairs and read lengths. The resulting sequences were BLAST searched against the entire non-redundant eukaryotic SILVA database ($n = 31,862$) using a threshold of 99% identity. While for the single reads we selected the highest scoring sequence that fulfilled the requirements, for the paired reads the sequence with the lowest mean of the log e-values was selected. In all cases, self-hits were disregarded. The taxonomic annotation of the query sequence was compared with that of the selected hit (Fig. 4). The accuracy of the taxonomic annotation varied significantly between primer pairs at all read lengths. While accuracy could be as low as 5.5% and 33.7% at species and genus levels, respectively, the primer pairs covering region V4 had about 30% accuracy at species level and about 75% at genus level for all read lengths. At the level immediately above genus (either family, tribe or other), at least 92% accuracy is achieved by these primer pairs at all read lengths tested. Accuracy falls somewhat at a lower percentage cutoff (Figure S1), and drops further when shorter read lengths are used (not shown).

Estimating Phylogenetic Assignment Accuracy

Since taxonomy is not always linked to evolutionary relationships, we also assessed the similarity of the selected best matches to the query sequences independently of taxonomy. Hence, in the third test, we used all top-scoring query-hit pairs from above and calculated the sequence distances between the full-length sequences of the queries and those of matches. Figure 5 shows the distribution of these distances for the different primer pairs and read lengths/types. With paired-ended reads of 250 bp or a single 400 bp reads, all primers have 75% of their BLAST matches within less than the 3% distance, commonly used for bacterial OTUs. The primers that allow retrieval of the V4 region have, with these read lengths, 80% of their hits at less than 1% distance and over 97% below 3% distance.

DNA Amplification and Sequencing

As primers pairs 563-, 574-, and 616–1132 performed consistently better than any of the other pairs throughout *in silico* tests, they were selected for experimental validation. 574 and 616 were modified to increase their phylogenetic matching (see above and Table S1). Forward primers 563, 574* and 616* were thus combined with reverse primer 1132. After optimization of

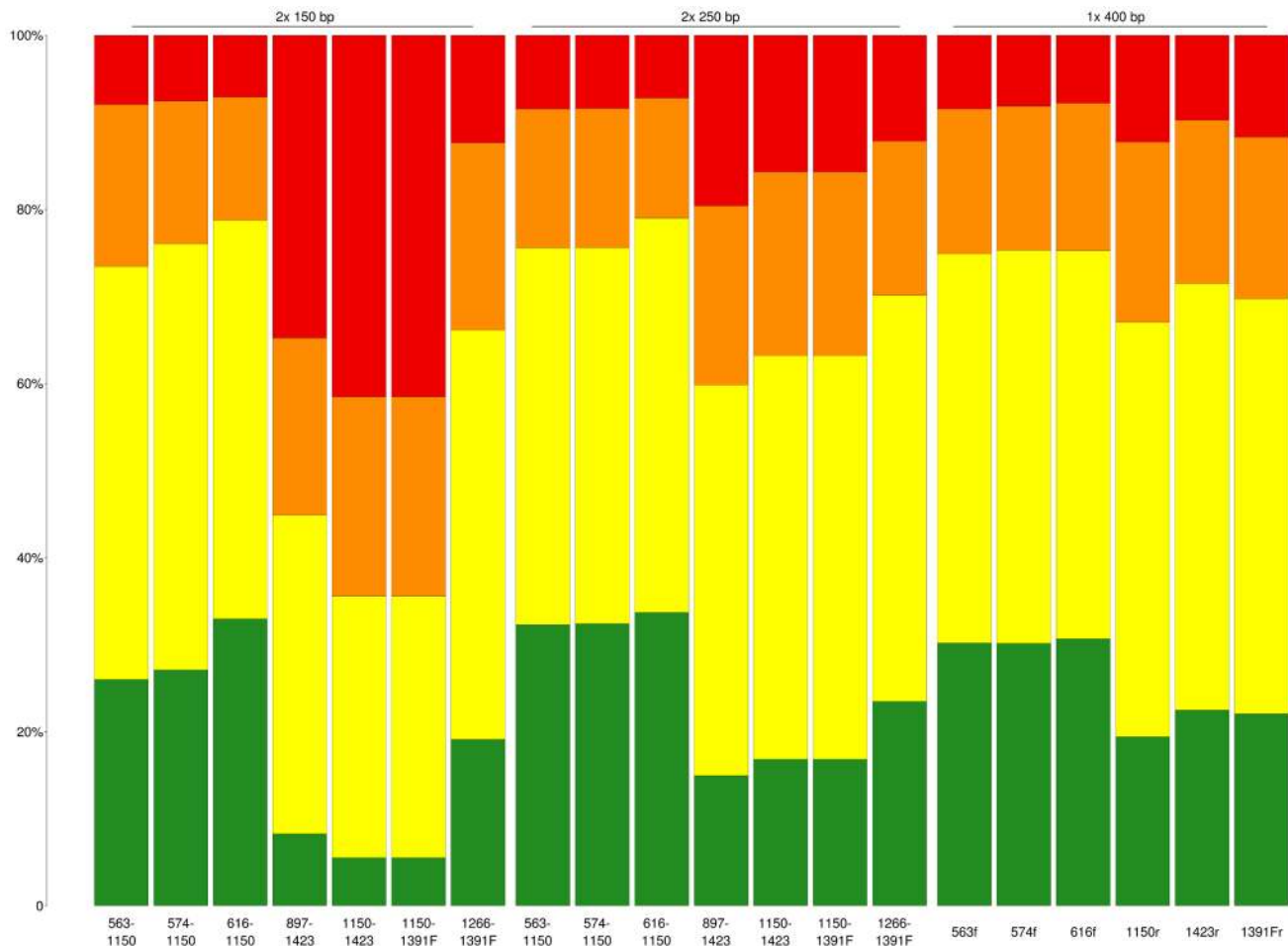


Figure 4. Specificity of taxonomic annotations at different taxonomic levels. Specificity of taxonomic annotations at different taxonomic levels, for the different primer pairs and read lengths/types, when requiring 99% identity to the selected match. Only instances where the selected hit sequence was annotated down to family level are shown, which were on average 78% of the cases. Matches to the correct species are depicted in green, and to the right genus in yellow. Matches to the level annotated immediately above genus are marked in orange. All other matches are considered missassignments and depicted in red.
doi:10.1371/journal.pone.0095567.g004

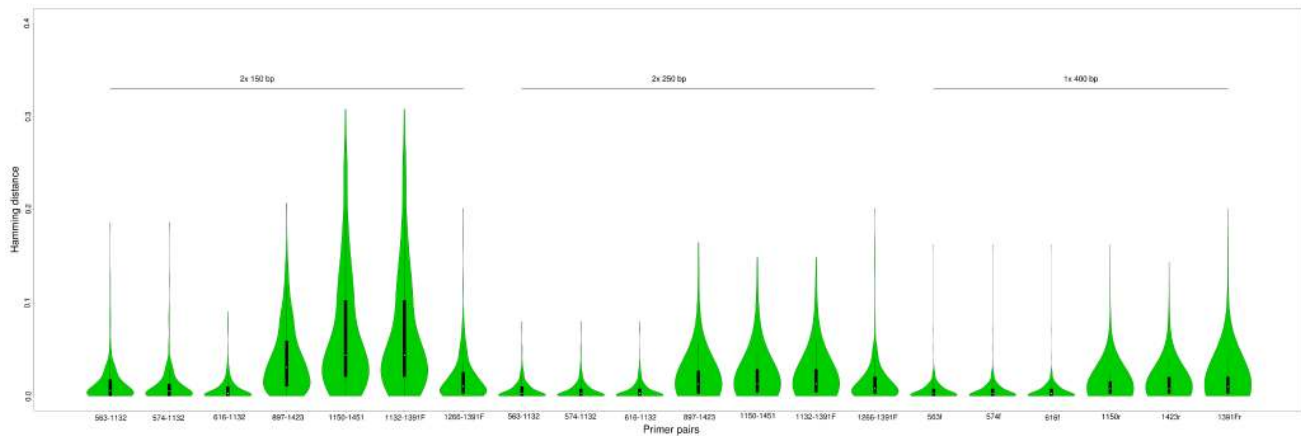


Figure 5. Hamming distance between blast queries and hits. A violin plot of the Hamming distance between the full-length sequence of BLAST queries and hits at a 99% identity cut-off. Inside each violin the boxplot is also depicted. doi:10.1371/journal.pone.0095567.g005

amplification conditions, the three pairs gave strong single bands of expected sizes for the positive control (*Saccharomyces cerevisiae*), and no amplification of the bacterial and archaeal controls. However, tested complex environmental samples sometimes yielded a second amplification product 50–100 bp smaller than the expected size. This band might represent true biological diversity, as the V4 region of the 18S gene is known to present variations in length [47]. However, the pair 574*–1132, which presented a single sharp band, was chosen for further tests.

DNA from three fungal isolates and environmental samples was amplified with primer pair 574*–1132 and sequenced with Illumina MiSeq. After quality trimming, reads were clustered at 97% similarity using Usearch and were classified using the SINA classifier [48]. Starting with 90,000 to 95,000 reads per sample, all isolates produced between 4070 and 5075 clusters, 65% of which singletons. In each fungal isolate case one cluster contained over 90% of remaining sequences and was correctly classified down to genus level. The representative sequence of each of these clusters could be matched exactly to the corresponding species in the SILVA database.

The collection of environmental samples included human faeces, moose rumen content, marine water, lake sediment, wastewater activated sludge and soil. No prokaryotic DNA sequences were observed in these libraries. A first glance shows that host-associated samples are, as expected, much less diverse than samples from open environments. For bacteria, soil and sediment are believed to be among the most diverse environments, orders of magnitude more diverse than for example seawater [49]. Interestingly this was not observed for the eukaryotic communities analysed here (Fig. 6A, Table 2). As might be expected, samples from similar environments have a somewhat similar community composition, with the predominant difference between communities being the separation between free-living and host-associated (Fig. 6B).

Marine 1 emerges as one of the most diverse samples (Fig. 6A, Table 2). This sample was collected close to shore in a closed bay, unlike Marine 2, which was collected in open sea (both central Baltic Sea). While Marine 2 is dominated by the SAR group (Stramenopiles, Alveolata and Rhizaria), and had 53% of its sequences matching to Dinophyceae, Marine 1 presents a more diverse community, including significant amounts of Chloroplastida and Opisthokonta (Fig. 7A, File S1). Also remarkable is the 7% of sequences in Marine 1 assigned to relatively unknown

phylum Katablepharidae [50]. Most previous studies of eukaryotic life in soil have focused on fungi. Our observations are in good agreement with these, such as a large dominance of Basidiomycota and Ascomycota, with much smaller fractions of other phyla such as Chytridiomycota [51]. However, we observe a larger proportion of the deep branching fungi LKM11 and LKM15 in our soil sample (Fig. 7B). The larger proportion of unclassified reads in the lake sediment as compared to the soil sample suggests unexplored microbial diversity, although the total alpha-diversity seems similar (Fig. 6A, File S1). By far the largest group in both wastewater samples was the subclass Peritrichia (Fig. 7C, File S1), which is known to play an important role in wastewater community dynamics [52]. The second largest group observed, the Cercozoa, has been observed in wastewater [53,54] and is also found in marine and fresh water environment samples. Interestingly, Metazoan predators such as Rhabditida and Rotifers were also found in these samples, indicating a complex, multi-layered ecosystem.

In the host-associated communities, little or no host DNA was amplified. For the moose rumen samples, a significant proportion (4 and 53% respectively) of the 18S sequences were derived from feed (Chloroplastida). One of the ruminants also contained substantial amounts of fungal sequences (5% of the sequences). The remainder was dominated by SAR, of which the rumen ciliate *Entodinium* [55] made up a large proportion (Fig. 7D, File S1). In the human gut, interestingly 85% of sequences were classified as Parabasalids (File S1), for which both parasitic and commensal members are known.

Discussion

We defined a set of primer pairs suitable for massively parallel sequencing of the eukaryotic 18S rDNA which covers different variable regions of the gene. Primer pairs 563–1132, 574–1132 and 616–1132 cover regions V4 and V5; 897–1423, V5 and V7; 1132–1423 and 1266–1423, solely V7 and 1132–1391F and 1266–1391F, V7 and V8. Our analyses suggest that regions V4 and V5 are the most information-rich. By using the paired-read information obtained with, for example, Illumina or IonTorrent sequencing, one has the possibility to combine these two regions (combining forward primers 563, 574* or 616* with reverse primer 1132), and the sequences generated are as discriminating as the longer reads achievable with 454 pyrosequencing (Fig. 4 and 5). As both Illumina and IonTorrent sequencing have a much

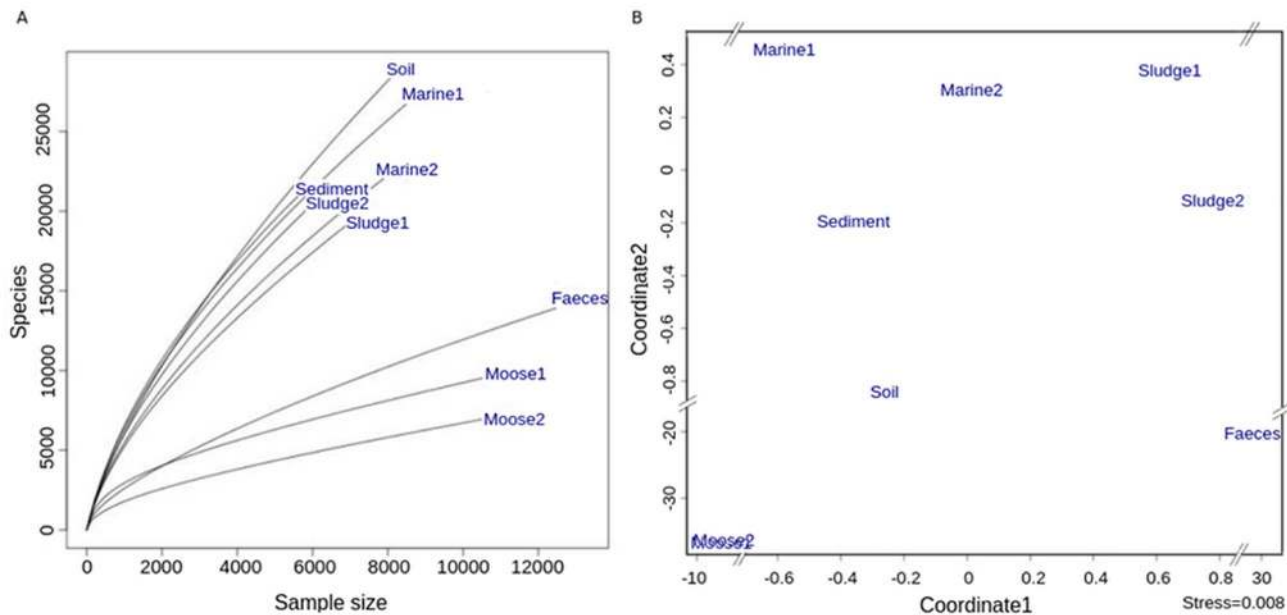


Figure 6. Alpha- and beta-diversity of environmental samples. (A) Rarefaction curves for OTUs at 97% similarity for environmental samples. (B) NMDA plot of the Bray-Curtis distance between 97%-similarity OTU profiles of the same samples. doi:10.1371/journal.pone.0095567.g006

lower cost per read, this allows for deeper sequencing efforts without loss in data quality.

While primer pair 616–1132 performed consistently well in all tests, forward primers 616 and even 616* have the disadvantage of missing some important groups of organisms (Table S1). While the added degeneracy of 616*, as compared to 616, partially rescued its capacity of binding to *Discoba*, both forms still fail to amplify most fungi, Microsporidia and Metamonada (including *Giardia* sp.). In contrast, the modified primer 574* fully recovers binding to Metamonada, but still fails to fully match to Acanthocephala and Microsporidia. Primer 563 is one of the few of the designed primers that doesn't miss any major group of organisms in our detailed analysis (Table S1). It can be suitably combined with primer 1132, which also binds to all major eukaryotic phyla.

It is important to notice that, while primer-binding is a strong limiting factor to the amplification of DNA tags from environmental samples, it is by no means the only one. The presence, structure and composition of a cell wall plays an important role in the efficiency of cell lysis and DNA extraction. Humic and fulvic

acids, as well as pigments and cations, inhibit PCR in a template-specific manner, where the length of the amplicon and the melting temperature of the primer play important parts [56]. The V4 region we propose here can present variations in size of over 100 bp [47]. Furthermore, when dealing with degenerate primers, the annealing temperature is not the same for every variant of it, and therefore to every template present. The length of the amplicon can also affect the efficiency of cluster detection in solid-phase sequencing-by-synthesis, with shorter DNA amplicons producing sharper clusters and better results.

Due to the richness of the data obtained by combining variable regions with paired-end reads, the primer combinations and BLAST parameters used produced median distances between the full-length sequences of queries and hits well below the 3% commonly adopted for binning bacteria into operational taxonomic units at the species level [57] (Fig. 5). However, the taxonomic assignment of reads is still generally only trustworthy at the genus level or higher in the taxonomy (Fig. 4), with 79% accuracy at the genus level for the best primer pair using

Table 2. Estimators of alpha-diversity for the environmental samples sequenced.

Sample	Shannon	Pielou	Chao1
Marine1	4.128	0.6150	1109.4
Marine2	3.605	0.5690	716.0
Soil	3.959	0.5760	1272.3
Sediment	3.674	0.5740	819.2
Sludge1	2.645	0.4370	628.5
Sludge2	2.984	0.4760	692.7
Moose1	1.963	0.3270	270.1
Moose2	0.712	0.1440	189.1
Faeces	1.315	0.2950	144.3

doi:10.1371/journal.pone.0095567.t002

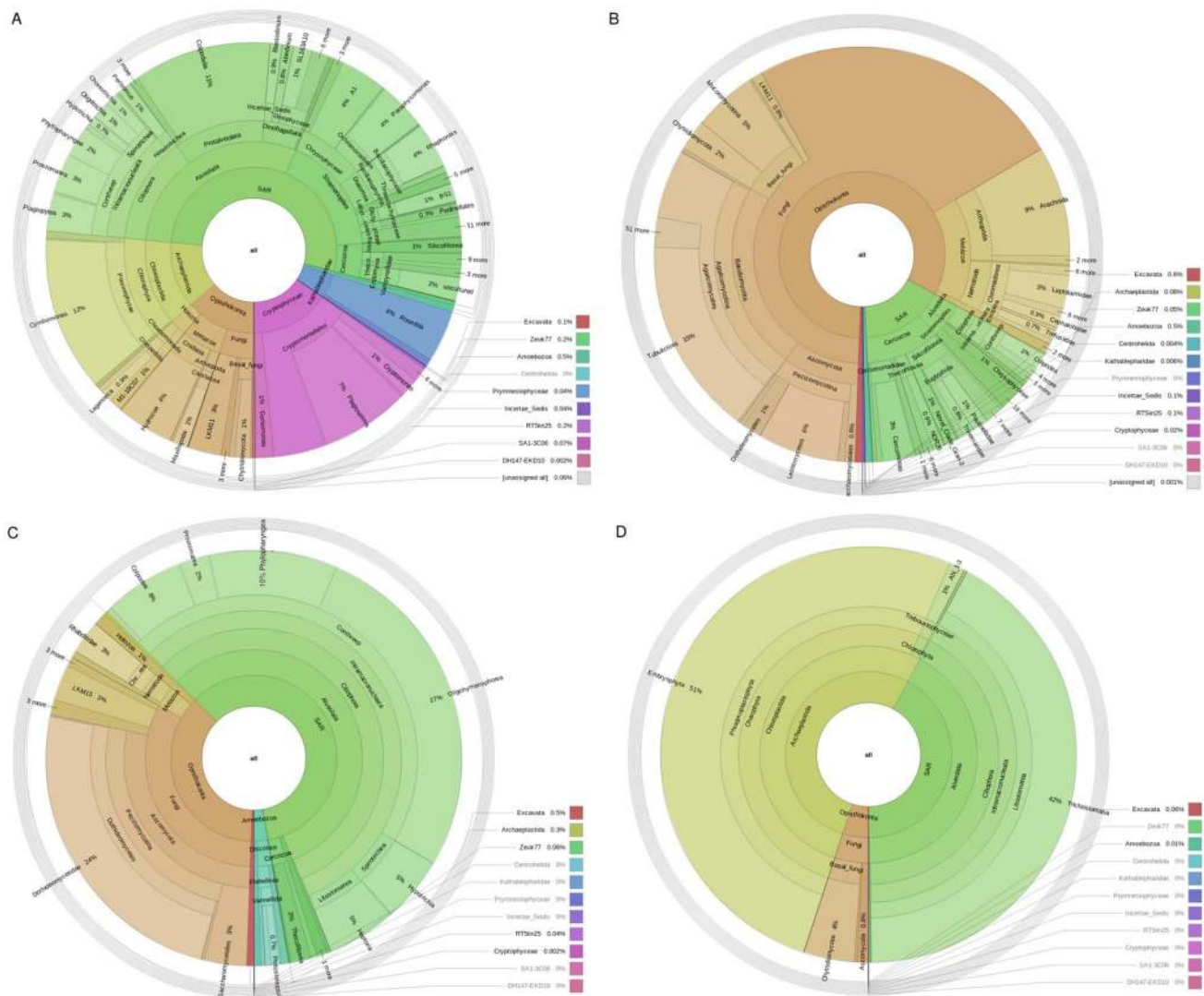


Figure 7. Taxonomic classification of selected environmental samples. (A) Marine water 1 (B) Soil (C) Wastewater sludge 2 (D) Moose rumen 1. An interactive HTML version of these plots and of the other environmental samples at deeper taxonomic resolution can be found as File S1. doi:10.1371/journal.pone.0095567.g007

2×250 bp reads. It is further important to note that the sample taken as query for the BLAST analyses contained only species of known phylogeny, to allow for their comparison. On a sample from a poorly characterized environment (as our environmental sequencing showed), most sequences will probably lack a database relative within close enough sequence distance to infer taxonomy at the genus level.

The 79% accuracy in genus-level taxonomic assignment observed is considerably below what was seen in a similar analysis conducted on the bacterial 16S rDNA, where 94% of genus assignments were correct when reading only 59 bp from variable region V6 [58]. This discrepancy is probably due both to intrinsic characteristics of eukaryotic organisms and to how their taxonomy has been defined [17]. Since eukaryotes are more morphologically distinct than bacteria, they tend to be more finely classified. Organisms with genome differences small enough to be considered to belong to the same species in the bacterial world can be assigned to different genera in the eukaryotic world [59]. This most likely contributed to the relatively poor performance of classification of

the yeast isolates. All isolates belong to the Saccharomycetales order, which includes 1050 sequences from 31 different families in the SILVA 111 database. This abundance of near matches can affect the performance of the classifier, which outputs a conservative class-level classification. For fine-scale phylogenetic analysis of fungi it may be necessary to use more fast evolving sequence regions such as the internal transcribed spacer [60].

The very large variation in gene copy number limits the use of 18S rDNA to a semiquantitative approach [61]. The relative proportion of an OTU can be compared between different samples, but no conclusion can be made on its absolute abundance based only on sequencing read number. One way of side-stepping this issue is to extract RNA from samples and prepare cDNA libraries. This approach reduces the phylogenetic copy number variation, privileging instead the amplification of cDNA fragments from highly physiologically active cells, with many ribosomes [62].

Despite these concerns, the sequencing results presented here show that a meaningful qualitative picture of the eukaryotic microorganisms of complex environments can be obtained by

deep sequencing of the V4 region of the 18S rDNA gene. The relationships between OTU defined by the V4 region and morphospecies is currently under investigation by the CBOL Protist Working Group [63]. This working group will also investigate the suitability and necessity of sequencing lineage-specific gene tags for protistan identification.

Microbial eukaryotes are a fundamental part of all ecosystems, but are often overlooked in microbial research, due to both the relative difficulty in studying them (when compared to bacteria) and historical biases [64]. Here, we present PCR primer pairs that can help bridge this gap, by allowing minimally biased and information-dense 18S rDNA amplicon sequencing.

Supporting Information

Figure S1 Specificity of taxonomic annotations with 95% identity cut-off. Specificity of taxonomic annotations at different taxonomic levels, for the different primer pairs and read lengths/types, when requiring 95% identity to the selected match. Only instances where the selected hit sequence was annotated down to family level are shown. Matches to the correct species are depicted in green, and to the right genus in yellow. Matches to the level annotated immediately above genus are marked in orange. All other matches are considered missassignments and depicted in red. (TIF)

Table S1 Proportion of sequences matched by each primer in each taxon. Amount of sequences from each kingdom or phylum

References

- Pace NR, Stahl DA, Lane DJ, Olsen GJ (1985). Analyzing Natural Microbial Populations by rRNA Sequences. *ASM News*, 51, 4–12.
- Massana R, Murray AE, Preston CM, DeLong EF (1997). Vertical Distribution and Phylogenetic Characterization of Marine Planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol*, 63 (1), 50–6.
- López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D (2001). Unexpected Diversity of Small Eukaryotes in Deep-sea Antarctic Plankton. *Nature*, 409, 603–7.
- Diez B, Pedrós-Alió C (2001). Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-subunit rRNA Gene Cloning and Sequencing. *Appl Environ Microbiol*, 67 (7), 2932–41.
- Chakravorty S, Helb D, Burday M, Connel N, Allan D (2007). A Detailed Analysis of 16S Ribosomal RNA Gene Segments for the Diagnosis of Pathogenic Bacteria. *J Microbiol Methods*, 69 (2), 330–9.
- Wang Y, Qian PY (2009). Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies. *PLoS ONE*, 4 (10), e7401.
- Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009). Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*, 326 (5960), 1694–7.
- Logue JB, Langenheder S, Andersson AF, Bertilsson S, Drakare S, et al (2012). Freshwater Bacterioplankton Richness in Oligotrophic Lakes Depends on Nutrient Availability Rather Than on Species-area Relationships. *ISME J*, 6 (6), 1127–36.
- Herlemann DP, Labrenz M, Jurgens K, Bertilsson S, Waniek JJ, et al. (2011). Transitions in Bacterial Communities Along the 2000 Km Salinity Gradient of the Baltic Sea. *ISME J*, 5 (10), 1571–9.
- Mshvildadze M, Neu J, Shuster J, Theriaque D, Li N, et al. (2011). Intestinal Microbial Ecology in Premature Infants Assessed with Non-Culture-Based Techniques. *Early Hum Dev*, 87 (7), 511–3.
- Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, et al (2012). Defining Seasonal Marine Microbial Community Dynamics. *ISME J*, 6 (2), 298–308.
- Quince C, Lunn EE (2013). The Impact of Crohn's Disease Genes on Healthy Human Gut Microbiota: a Pilot Study. *Gut*, 62 (62), 952–4.
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al (2009). A Core Gut Microbiome in Obese and Lean Twins. *Nature*, 457 7228, 480–4.
- Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, et al (2010). A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary with Inflammatory Bowel Disease Phenotypes. *Gastroenterology*, 129 (6), 1844–54.
- Ichinohe T, Pang IK, Kumamoto Y, Peaper DR, Ho JH, et al (2011). Microbiota Regulates Immune Defense Against Respiratory Tract Influenza A Virus Infection. *PNAS*, 108 (13), 5354–9.
- Abrahamsson TR, Jakobsson HE, Andersson AF, Björkstén B, Engstrand L, et al. (2012) Low Diversity of the Gut Microbiota in Infants with Atopic Eczema. *J Allergy Clin Immunol*, 129 (2), 434–40.
- Amaral-Zettler LA, McCliment EA, Ducklow W, 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.
- Stoeck T, Bass D, Nebel M, Christen R, Jones MD, et al (2010) Multiple Marker Environmental DNA Sequencing Reveals a Highly Complex Eukaryotic Community in Marine Anoxic Water. *Mol Ecol*, 19 (Sup. 1), 21–31.
- Monchy S, Sancier G, Jobard M, Rasconi S, Gerphagnon M, et al (2011) Exploring and Quantifying Fungal Diversity in Freshwater Lake Ecosystems Using rDNA Cloning/sequencing and SSU Tag Pyrosequencing. *Environ Microbiol*, 13 (6), 1433–53.
- Orsi W, Edgcomb V, Jeon S, Leslin C, Bunge J, et al. (2011) Protistan Microbial Observatory in the Cariaco Basin, Caribbean. II. Habitat Specialization. *ISME J*, 5 (8), 1357–73.
- Bik HM, Halanich KM, Sharma J, Thomas WK (2012). Dramatic Shifts in Benthic Microbial Eukaryote Communities Following the Deepwater Horizon Oil Spill. *PLoS ONE*, 7 (6), e38550.
- Bricheux G, Morin L, Le Moal G, Coffe G, Balestrino D, et al (2013). Pyrosequencing Assessment of Prokaryotic and Eukaryotic Diversity in Biofilm Communities from a French River. *Microbiology Open*, doi:10.1002/mbo3.80.
- Findley K, Oh J, Yang J, Conlan S, Deming C, et al (2013) Topographic Diversity of Fungal and Bacterial Communities in Human Skin. *Nature*, 498 (7454), 367–70.
- Auinger BM, Pfandl K, Boenigk J (2008) Improved methodology for identification of protists and microalgae from plankton samples preserved in Lugol's iodine solution: combining microscopic analysis with single-cell PCR. *Appl Environ Microbiol*, 74 (8): 2505–10.
- Brâte J, Logares R, Berney C, Rec DK, Klaveness D, et al (2010) Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *ISME J*, 4 (9): 1144–53.
- Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK (2010) Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. *ISME J*, 4 (9): 1053–9.
- Baker GC, Smith JJ, Cowan DA (2003) Review and Re-analysis of Domain-specific 16S Primers. *J Microbiol Methods*, 55 (3), 541–55.
- Youssef N, Sheik CS, Krumholz LR, Najjar FZ, Roe BA, et al. (2009) Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-generated Fragments in 16S rRNA Gene-based Environmental Surveys. *Appl Environ Microbiol*, 75 (16), 5227–6.
- Gantner S, Andersson AF, Alonso-Sáez L, Bertilsson S (2010) Novel Primers for 16S rRNA-based Archaeal Community Analysis in Environmental Samples. *J Microbiol Methods*, 84 (1), 12–8.

30. Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, et al (2010) Design of 16S rRNA Gene Primers for 454 Pyrosequencing of the Human Foregut Microbiome. *World J Gastroenterol*, 16 (33), 4135–44.31.
31. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of Multimillion-sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-end Illumina Reads. *Appl Environ Microbiol*, 77 (11), 3846–52.32.
32. Kumar PS, Brooker MR, Dowd SE, Carmelento T (2011) Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. *PLoS ONE*, 6 (6), e20956.
33. Soergel DA, Dey N, Knight R, Brenner SE (2012) Selection of Primers for Optimal Taxonomic Classification of Environmental 16S rRNA Gene Sequences. *ISME J*, 6 (7), 1440–4.
34. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, et al (2013) Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and Next-generation Sequencing-based Diversity Studies. *Nucleic Acids Res*, 41 (1), e1.
35. Sogin ML, Gunderson JH (1987) Structural Diversity of Eukaryotic Small Subunit Ribosomal RNAs. Evolutionary Implications. *Ann. NY Acad. Sci.* 503, 125–39.
36. Nolte V, Pandey RV, Jost S, Medinger R, Ottenwalder B, et al (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecology*, 19 (14): 2908–15.
37. Cameron KA, Hodson AJ, Osborn AM (2012) Structure and Diversity of Bacterial, Eukaryotic and Archaeal Communities in Glacial Cryoconite Holes from the Arctic and the Antarctic. *FEMS Microbiol Ecol*, 32 (2), 254–67.
38. Aguilera A, Souza-Egipsy V, Gonzales-Toril E, Rendueles O, Amils R (2010). Eukaryotic Microbial Diversity of Phototrophic Microbial Mats in Two Icelandic Geothermal Hot Springs. *International Microbiology*, 13, 21–32.
39. Pandey PK, Siddharth J, Verma P, Bavdekar A, Patole MS, et al. (2012) Molecular Typing of Fecal Eukaryotic Microbiota of Human Infants and Their Respective Mothers. *J Biosci*, 37 (2), 221–6.
40. Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, et al (2007) SILVA: a Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB. *Nuc Acids Res*, 35 (21), 7188–96.
41. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hatmann M, et al (2009) Introducing Mothur: Open-source, Platform-independent Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol*, 75 (23), 7537–41.
42. Forsberg AM, Kjellstrom L, Agreus L, Nixon AA, Nyhlin H, et al (2012) Prevalence of colonic neoplasia and advanced lesions in the normal population: a prospective population-based colonoscopy study. *Scand J Gastroenterol*, 47 (2): 184–90.
43. Roume H, Muller EE, Cordes T, Renaut J, Hiller K, et al. (2013) A Biomolecular Isolation Framework for Ecosystems Biology. *ISME J*, 7 (1), 110–21.
44. Riemann L, Steward GF, Azam F (2000) Dynamics of Bacterial Community Composition and Activity During a Mesocosm Diatom Bloom. *Appl Environ Microbiol*, 66 (2), 578–87.
45. Lundin S, Stranneheim H, Petterson E, Klevebring D, Lundeberg J (2010). Increased Throughput by Parallelization of Library Preparation for Massive Sequencing. *PLoS ONE*, 5 (4), e10029.
46. Edgar RC (2010) Search and Clustering Orders of Magnitude Faster Than BLAST. *Bioinformatics*, 26 (19), 2460–1.
47. Wuyts J, De Rijk P, Van de Peer Y, Pison G, Rousseeuw P, et al. (2000) Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nuc Ac Res* 28 (23): 4698–708.
48. Pruesse E, Peplies J, Glockner FO (2012) SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, doi: 10.1093/bioinformatics/bts252.
49. Torsvik V, ovreas L, Thingstad TF (2002) Prokaryotic diversity – magnitude, dynamics and controlling factors. *Science*, 296, 1064–6.
50. Okamoto N, Inouye I (2005) The Katablepharids Are a Distant Sister Group of the Cryptophyta. *Protist*, 156 (2), 163–79.
51. McGuire KL, Allison SD, Fierer N, Treseder KK (2013) Ectomycorrhizal-dominated Boreal and Tropical Forest Have Distinct Fungal Communities, but Analogous Spatial Patterns Across Soil Horizons. *PLoS ONE*, 8 (7), e68278.
52. Weber SD, Ludwig W, Schleifer K–H, Fried J (1997) Microbial Composition and Structure of Granular Sewage Biofilms. *Applied Environmental Microbiology*, 73 (19), 6233–40.
53. Paez-Rubio T, Viau E, Romero-Hernandez S, Peccia J (2005). Source Bioresol and rRNA Gene-Based Identification of Microorganisms Aerosolized at a Flood Irrigation Wastewater Reuse Site. *Appl Environ Microbiol*, 71 (2), 804–10.
54. Evans TN, Seviour RJ (2012) Estimating Biodiversity of Fungi in Activated Sludge Communities Using Culture-Independent Methods. *Microb Ecol*, 63 (4), 773–86.
55. Dehority BA (1974) Rumen ciliate fauna of Alaskan Moose (*Alces americana*), Musk-Ox (*Ovibos moschatus*) and Dall Mountain Sheep (*Ovis dalli*). *The Journal of Protozoology*, 21 (1), 26–32.
56. Opel KL, Chung D, McCord BR (2009) A Study of PCR Inhibition Mechanisms Using Real Time PCR. *J Forensic Sci*, 55 (1): 25–33.
57. Schloss PD, Handelsman J (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Appl Environ Microbiol*, 71, 1501–6.
58. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, et al. (2008) Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. *PLoS ONE*, 3 (7), e2836.
59. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, et al. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, 311 (5765), 1283–7.
60. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *PNAS*, doi:10.1073/pnas.1117018109.
61. Gong J, Dong J, Liu X, Massana R (2013) Extremely High Copy Numbers and Polymorphisms of the rDNA Operon Estimated from Single Cell Analysis of Oligotrich and Peritrich Ciliates. *Protist* 164 (3): 369–79.
62. Not F, del Campo J, Balague V, de Vargas C, Massana R (2009) New Insights into the Diversity of Marine Picoeukaryotes. *Plos One*, doi:10.1371/journal.pone.0007143.
63. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, et al (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *Plos Biol*. doi:10.1371/journal.pbio.1001419.
64. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB (2009) Protists Are Microbes Too: a Perspective. *ISME J*, 3 (1): 4–12.