Received: 3 April 2018 | Revised: 5 July 2018 | Accepted: 28 July 2018

DOI: 10.1111/1755-0998.12937

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments

Felix Heeger[1,2*] | Elizabeth C. Bourne[1,2*] | Christiane Baschien[3] | Andrey Yurkov[3] | Boyke Bunk[3] | Cathrin Spröer[3] | Jörg Overmann[3] | Camila J. Mazzoni[2,4†] | Michael T. Monaghan[1,2†]

[1]Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany

[2]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany

[3]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

[4]Leibniz Institute of Zoo- and Wildlife Research (IZW), Berlin, Germany

Correspondence

Felix Heeger, Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany.
Email: heeger@igb-berlin.de

Funding information

Leibniz Association, Grant/Award Number: SAW-2014-IGB-1

Abstract

DNA metabarcoding is widely used to study prokaryotic and eukaryotic microbial diversity. Technological constraints limit most studies to marker lengths below 600 base pairs (bp). Longer sequencing reads of several thousand bp are now possible with third-generation sequencing. Increased marker lengths provide greater taxonomic resolution and allow for phylogenetic methods of classification, but longer reads may be subject to higher rates of sequencing error and chimera formation. In addition, most bioinformatics tools for DNA metabarcoding were designed for short reads and are therefore unsuitable. Here, we used Pacific Biosciences circular consensus sequencing (CCS) to DNA-metabarcode environmental samples using a ca. 4,500 bp marker that included most of the eukaryote SSU and LSU rRNA genes and the complete ITS region. We developed an analysis pipeline that reduced error rates to levels comparable to short-read platforms. Validation using a mock community indicated that our pipeline detected 98% of chimeras de novo. We recovered 947 OTUs from water and sediment samples from a natural lake, 848 of which could be classified to phylum, 397 to genus and 330 to species. By allowing for the simultaneous use of three databases (Unite, SILVA and RDP LSU), long-read DNA metabarcoding provided better taxonomic resolution than any single marker. We foresee the use of long reads enabling the cross-validation of reference sequences and the synthesis of ribosomal rRNA gene databases. The universal nature of the rRNA operon and our recovery of >100 nonfungal OTUs indicate that long-read DNA metabarcoding holds promise for studies of eukaryotic diversity more broadly.

KEYWORDS

aquatic, chimera formation, fungi, metabarcoding, mock community, PacBio

1 | INTRODUCTION

DNA metabarcoding is widely used in the study of microbial communities (Davison et al., 2015; Hamad et al., 2018; McFrederick & Rehan, 2016; Tedersoo et al., 2014; Wurzbacher et al., 2016; Yu et al., 2012), whereby one or more marker regions in the genome are PCR-amplified and sequenced using a next-generation sequencing (NGS) platform. Reads are quality-filtered, and sequences are clustered according to sequence similarity into putative taxa (Operational Taxonomic Units = OTUs). OTUs are then classified using marker-specific and sometimes taxon-specific databases. DNA metabarcoding has become a commonly used tool because it

[*] These authors contributed equally to this work.

[†] These authors contributed equally to this work.

provides an estimate of biodiversity, including that of taxa that cannot be cultured, and identification relies on relatively stable genetic information rather than often variable and subtle phenotypic characters. Limitations of the method include the fact that marker regions and PCR primers must be selected a priori to detect the taxa of interest and that the variability of the marker region, and how well the taxa are represented within a given reference database, determines the success of the identification of the members of an assemblage (Nilsson et al., 2018).

There is a fundamental trade-off between using a marker that is conserved enough to be amplified across a broad range of taxa, but variable enough to distinguish among closely related species. Marker length also has consequences for how many OTUs can be identified and to what taxonomic resolution (Porras-Alfaro, Liu, Kuske, & Xie, 2014). Shorter markers within a given locus may include less genetic variation than longer markers, reducing the ability to distinguish closely related species (Singer et al., 2016). One consequence of the read length limits on typical NGS sequencing platforms (e.g., 2× 300 bp on an Illumina MiSeq) is that shorter but highly variable regions are often used as DNA metabarcoding markers. While variable regions may increase taxonomic resolution in groups for which reference sequences are available, sequence homology can be difficult or impossible to establish. This precludes phylogeny-based analyses and can result in the complete failure of classifying OTUs at any taxonomic level (Lindahl et al., 2013).

More recent (i.e., third-generation sequencing) technologies provide much longer (several kbp) sequencing reads than current popular second-generation platforms (Goodwin, McPherson, & McCombie, 2016); however, their use in studies of environmental samples remains limited. The few existing studies, using full-length (~1.5 kbp) bacterial 16S (Franzén et al., 2015; Schloss, Jenior, Koumpouras, Westcott, & Highlander, 2016; Singer et al., 2016) and parts of the eukaryotic rRNA operon including ITS (up to 2.6 kbp) (Schlaeppi et al., 2016; Tedersoo, Ave, & Sten, 2017; Walder et al., 2017), have reported increased taxonomic resolution. The Pacific Biosciences (PacBio) RSII platform generates reads of >50 kbp by single-molecule real-time (SMRT) sequencing. Single-pass error rates of 13%–15% (Goodwin et al., 2016) limit their value in DNA metabarcoding because species identification is unreliable at those levels of uncertainty. However, the circular consensus sequencing (CCS) version of SMRT sequencing greatly reduces the error rate. In CCS, double-stranded DNA amplicon molecules are circularized by the ligation of hairpin adapters. The sequencing polymerase passes around the molecule and reads the same insert multiple times (Travers, Chin, Rank, Eid, & Turner, 2010). The repeated reads of the same amplicon molecule, together with the random nature of sequencing error, can then be used to reduce the final error rate to <1% (Goodwin et al., 2016) by generating consensus sequences.

Aside from the higher per-base cost, a primary reason why long-read approaches have not been applied to DNA metabarcoding is the fact that most of the existing bioinformatic tools have been optimized for the analysis of data from short-read technologies (e.g., Illumina). Their performance on PacBio CCS reads has not been systematically evaluated, and it is thus not clear whether they will perform well in all circumstances. Longer sequences have more errors because even high-quality reads with low error rates will accumulate more total errors as a function of length. The types of errors in PacBio reads also differ from those of short-read technologies, with CCS reads tending to have more insertions and deletions, compared to substitutions that are more common in short-read data (Goodwin et al., 2016). Schloss et al. (2016) explored the error profile of CSS reads and steps that can be taken when targeting the 16S for a bacterial mock community and for environmental samples. They found that the error rate of their longest amplicon (V1-V9) was only 0.68% and could be further reduced to 0.027% by preclustering at 99% similarity. Another potential problem of long reads is that chimera formation rates may be increased in longer markers since longer amplicons may suffer premature elongation terminations, leading to more possibilities for the resulting incomplete amplicons to act as primers in the next PCR cycle and thus more chimeras to be formed (see also Laver et al., 2016). Existing algorithms commonly used to detect chimeras are optimized for short reads, and to our knowledge, their performance has not yet been evaluated on long reads.

Fungi are ecologically important eukaryotes. They play diverse roles in the cycling of carbon and nutrients, occupy a range of niches, including as decomposers, parasites and endophytes, and are ubiquitous in terrestrial and aquatic habitats alike (Tedersoo et al., 2014; Wurzbacher et al., 2016). Microbial fungal communities are increasingly studied with DNA metabarcoding (Roy et al., 2017), taking advantage of the increased detection of taxa without the need to culture and the reduced cost of sequencing that has permitted ever higher sequencing coverage. The broad phylogenetic diversity of fungi has the consequence that fungal DNA metabarcoding studies typically use markers that vary depending on the taxonomic group of interest and the resolution desired.

Different regions of the eukaryotic rRNA operon have been widely utilized for barcoding fungi due to its universality and the fact that short stretches have been able to provide reasonable power for fungal identification. Within this region, the most commonly applied barcode is the internal transcribed spacer (ITS; Schoch et al., 2012). This comprises the ITS1, the 5.8S rRNA gene and the ITS2. Depending on the lineage, it varies from 300 to 1,200 bp in length. In fungal DNA metabarcoding, the ITS2 region is widely used to assess fungal diversity in environmental samples (Blaalid et al., 2013; Kõljalg et al., 2013) and allows for good taxonomic identification of many fungal groups (Porras-Alfaro et al., 2014; Tedersoo et al., 2015). However, it has a lower identification success compared to the full-length ITS (Tedersoo et al., 2017). For early diverging fungal lineages, such as those found in many aquatic habitats (Monchy et al., 2011; Rojas-Jimenez, Fonvielle, Ma, & Grossart, 2017; Rojas-Jimenez, Wurzbacher, et al., 2017; Wurzbacher et al., 2016), sequences from the small subunit (SSU) rRNA gene (18S) can provide affiliation of higher taxonomic ranks, but are often not variable enough to distinguish among species (Cole et al., 2014). The large subunit (LSU) rRNA gene (18S) has higher variability than the SSU and is often used for identification of some fungal groups (e.g., Glomeromycota and Chytridiomycota). Databases have been established for all three different markers, for example, UNITE for ITS (Kõljalg et al., 2013), SILVA for SSU (Quast et al., 2013) and RDP for LSU (Cole et al., 2014). Nevertheless, database coverage remains poor for several fungal lineages, for example, Glomeromycota (Ohsowski, Zaitsoff, Öpik, & Hart, 2014), Chytridridiomycota (Frenken et al., 2017) and Cryptomycota, and for species from less well studied habitats such as aquatic, indoor and marine environments.

Here, we describe a pipeline for analysing SMRT CCS of a long (*ca.* 4,500 bp) DNA metabarcode that includes the three major regions of the eukaryotic rRNA operon (SSU, ITS and LSU) in a single sequencing read (Figure 1). We first sequenced cultured isolates and a mock community comprising a broad phylogenetic range to derive rates of sequencing error and chimera formation. We found error rates to be comparable to short-read approaches after filtering with our pipeline and chimera formation rates to be comparable to those found in studies with shorter amplicons. We used these results to develop a new bioinformatics pipeline designed for the analysis of full-length rRNA operon amplicon metabarcoding. We applied this method to field-collected environmental samples from a temperate lake and identified 947 OTUs, 848 of which could be classified to phylum, 486 to family, 397 to genus and 330 to species.

2 | MATERIALS AND METHODS

2.1 | Isolates, mock community and environmental samples

Sixteen fungal isolates (Table 1) were selected for the creation of our mock community. We aimed for phylogenetic diversity (9 Ascomycota, 5 Basidiomycota, 1 Chytridiomycota and 1 Mucoromycota) and relevance for the aquatic environment, but also considered the availability of genomic data and isolate cultures or DNA. Isolates were obtained from the culture collection of the German Collection of Microorganisms and Cell Cultures or had been isolated from freshwater lakes in North-East Germany (strain numbers in Table 1). Cultures were maintained on 2% malt extract agar (MEA, Carl Roth, Karlsruhe, Germany), and mycelia were harvested directly for DNA extraction (see Supporting Information Appendix S1 for further details). Isolates from Rojas-Jimenez, Fonvielle, et al. (2017) were donated as DNA. Following DNA extraction (see below), a mock community was created by pooling the fungal isolate DNA at varying concentrations (Table 1) to account for low DNA concentrations in some isolates, while ensuring sufficient mix of a single mock community for all PCR comparisons. This community was used to test PCR and library preparation protocols that were later applied to environmental samples and to quantify the efficiency of de novo and reference-based chimera detection in our long-read bioinformatics pipeline described below.

Environmental samples were collected from Lake Stechlin, an oligo-mesotrophic lake in North-East Germany (53.143°N 13.027°E) in October 2014. Littoral water samples (30 L total) were collected and pooled from surface water in the shallow zone along three 10-m transects, located within 5 m of the lake shore or reed belt. Pelagic water samples (30 L total) were collected from the deeper zone of the lake by pooling samples taken at multiple depths (0–65 m) at one point, using a Niskin bottle (Hydro-Bios, Kiel, Germany). A subsample (2 L) of each (littoral or pelagic) was filtered through 0.22-µm Sterivex filters (Merck Millipore, Darmstadt, Germany) using a peristaltic pump (GT-EL2 Easy Load II, UGT, Müncheberg, Germany). Excess water was expelled using a sterile syringe and parafilm used to seal the ends. Sediment samples were collected from four locations in each zone (littoral and pelagic) using a PVC sediment corer (63 mm diameter) on a telescopic bar (Uwitec, Mondsee, Austria). The uppermost 2 cm from each sediment core was pooled in the field and divided into 2 ml subsamples for storage. Sterivex filters and sediment subsamples were frozen in liquid nitrogen in the field and returned to the laboratory for long-term storage at −80°C.

2.2 | DNA extraction

Genomic DNA was extracted from fungal isolates and was optimized using three different methods based on initial tests applied to each isolate (Table 1). Briefly, mycelia was harvested directly from MEA plates and homogenized directly in lysis buffer, or lyophilized and homogenized dry, before adding lysis buffer (see Supporting Information Appendix S1 for full details). DNA was checked on a gel and quantified in duplicate using a Qubit HS dsDNA Assay (Invitrogen, Carlsbad, USA). A mock community was created by pooling DNA according to the proportions described in Table 1, after validation that each could be individually amplified by the PCR protocol described below.

Environmental DNA was extracted from water and sediment samples using a modified phenol–chloroform method (after Nercessian, Noyes, Kalyuzhnaya, Lidstrom, & Chistoserdova, 2005; Table 1, see also Supporting Information Appendix S1). In brief, frozen Sterivex cartridges were broken open and sterilized forceps were used to

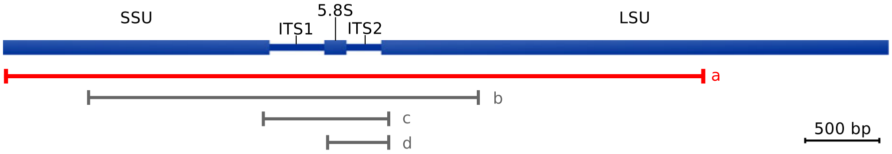

FIGURE 1 Region of the eukaryotic rRNA operon covered by the primer pair used in this study (a) compared to the primer pair SSU515Fngs-TW13 used by Tedersoo et al., 2017 (b), and the widely used (e.g. Schoch 2012) primer pairs ITS5-ITS4 (c) and ITS3-ITS4 (d) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Fungal isolates used and their contribution to the mock community by pooling of DNA at different concentrations

Taxon	Code	Isolate	DNA pooled (ng)	% of mock community
Clavariopsis aquatica	CA	DSM 29862[a]	60	7.6
Chytridiomycota	CHY1	CHY1[b]	60	7.6
Cladosporium sp.	Csp1	KR4[b]	20	2.5
Clonostachys rosea	CR	DSM 29765[c]	60	7.6
Cystobasidium laryngis	CL	CBML 151a[c]	5	0.6
Cladosporium herbarum	CH	KR13[b]	20	2.5
Exobasidium vaccinii	EV	DSM 5498[c]	60	7.6
Leucosporidium scottii	LS	CBML 203[c]	60	7.6
Metschnikowia reukaufii	MR	DSM 29087[c]	60	7.6
Mortierella elongata	ME	CBML 271[c]	60	7.6
Penicillium brevicompactum	PB	KR5[b]	80	10.2
Phanerochaete chrysosporium	PC	DSM 1547[c]	60	7.6
Phoma sp.	Psp1	KR1[b]	3	0.4
Saccharomyces cerevisiae	SC	DSM 70449[c]	60	7.6
Trichoderma reesei	TR	DSM 768[a]	60	7.6
Ustilago maydis	UM	DSM 14603[a]	60	7.6

[a]Extracted using Qiagen DNeasy Plant Mini Kit. [b]Extracted using peqGOLD Tissue DNA Mini Kit. [c]Extracted using MasterPure Yeast DNA Purification Kit.

transfer half of each fragmented filter into a single 2-ml tube, resulting in two subsamples per cartridge. Sediment samples were thawed and aliquoted as two subsamples into two 2-ml tubes, each containing 200 mg. Tubes were preprepared by adding beads (0.1 and 1.0 mm zirconium, and 3 × 2.5 mm glass beads, Biospec, Bartlesville, USA) to 0.2 volume of the tube. DNA was extracted following an initial CTAB cell lysis step, mechanical disruption, a series of phenol–chloroform phase clean-ups, precipitation by PEG/NaCl and a final ethanol wash (see Supporting Information Appendix S1 for full details). Subsamples were pooled to give 100 µl nucleic extract per sample. RNA was removed by the addition of 0.5 µl (5 µg) RNase A (10 mg/ml DNase and protease free, Thermo Fisher Scientific, Waltham, USA) to 80 µl of the pooled sample, incubated at 37°C for 30 min and cleaned using the PowerClean Pro DNA Clean-Up Kit (MoBio Laboratories, Carlsbad, USA). DNA was quantified in triplicate using a Qubit HS dsDNA Assay and gel-checked for quality.

2.3 | PCR and chimera formation tests

Approximately 4,500 bp of the eukaryotic rRNA operon (Figure 1), including SSU, ITS1, 5.8S, ITS2 and partial LSU (D1-D8 region, Wurzbacher, Rösel, Rychła, & Grossart, 2014) regions, was PCR-amplified using the primers NS1_short (5′- CAGTAGTCATATGCTTGTC-3′) and RCA95m with a 5′ mismatch spacer (5′-ACCTATGTTTTAATTAGA-CAGTCAG-3′) (Wurzbacher et al., 2018). Symmetric (reverse complement) 16-mer barcodes were added to the 5′ ends of primers (Supporting Information Table S1) following the PacBio manufacturer's guidelines on multiplexing SMRT sequencing.

We aimed to minimize chimera formation by keeping the number of PCR cycles performed per sample low. Cycle numbers were chosen after amplifying all samples with a variable number of cycles (13–30) and identifying the exponential phase of PCR (Lindahl et al., 2013) according to band visibility on an agarose gel. Based on these results, we used 15–20 cycles to amplify isolates (3–8 ng template DNA), 13–30 cycles for mock community samples (2–20 ng) and 22–26 cycles for environmental samples (10 ng). Barcodes were allocated to the different PCR conditions tested as shown in Supporting Information Tables S2 and S3. All standard PCRs were conducted in 25 µl reactions using 0.5 µl high processivity Herculase II Fusion enzyme (Agilent Technologies, Cedar Creek, USA), 5 µl of 5× PCR buffer, 0.62 µl each primer (10 µM), 0.25 µl dNTPs (250 mM each) and 0.3 µl BSA (20 mg/ml BSA, Thermo Fisher Scientific, Waltham, USA) on a SensoQuest Labcycler (SensoQuest Gmbh, Göttingen, Germany) with 2-min denaturation at 95°C, 13–30 cycles (see above) of 94°C for 30 s, 55°C for 30 s and 70°C for 4 min, and a final elongation at 70°C for 10 min. Multiple PCRs were required for each environmental sample, depending on the sample type (Stechlin water 7–18, Stechlin sediment 24–37 reactions), to ensure sufficient product for library preparation (1 µg purified PCR product). We also included a two-step emulsion PCR (emPCR) of the mock community in order to test whether emPCR could reduce chimera formation rate by the physical isolation of DNA template molecules (Boers, Hays, & Jansen, 2015). The Micellula DNA Emulsion Kit (Roboklon GmbH, Berlin) was used for a first amplification of 25 cycles using 20 ng of mock community DNA, with 2 µl of the cleaned template used in a second 25-cycle emPCR. For a full description, see Supporting Information Appendix S2.

2.4 | Library preparation and sequencing

Replicate PCRs were pooled back to sample level, and products were cleaned with 0.45 × CleanPCR SPRI beads (CleanNa, Waddinxveen, Netherlands), precleaned according to PacBio specifications (C. Koenig, pers. comm.). In brief, this involved removing the beads from the CleanPCR solution, washing the beads three times in molecular grade water, a final TE rinse and suspension in the original solution. This was to ensure the removal of compounds that could interfere with the SMRT cell sequencing enzyme. Amplicons were quantified twice using a Qubit HS dsDNA Assay and quality-checked on an Agilent® 2100 Bioanalyzer System (Agilent Technologies, Santa Clara, USA). Samples were then pooled into libraries (as described in Supporting Information Table S3) before being quality-checked on an Agilent® 2100 Bioanalyzer.

SMRTbell™ template libraries were prepared according to the manufacturer's instructions following the Procedure & Checklist—Amplicon Template Preparation and Sequencing (Pacific Biosciences, Inc., Menlo Park, CA, USA). Briefly, amplicons were end-repaired and ligated overnight to hairpin adapters applying components from the DNA/Polymerase Binding Kit P6 (Pacific Biosciences). We included enough DNA from each sample to obtain the required library concentration (37 ng/μl) for end repair. Reactions were carried out according to the manufacturer´s instructions. Conditions for annealing of sequencing primers and binding of polymerase to purified SMRTbell™ template were assessed with the Calculator in RS Remote (Pacific Biosciences). SMRT sequencing was carried out on the PacBio *RSII* (Pacific Biosciences) taking one 240-min movie.

In total, we ran eight libraries and 27 SMRT cells. Three of the isolates (*Trichoderma reesei*, *Clonostachys rosea* and a species belonging to the phylum Chytridiomycota) were sequenced on one SMRT cell to test the protocol for CCS. The remaining 13 isolates and one of the mock community conditions (30 PCR cycles) were prepared as part of the libraries containing the environmental samples (Supporting Information Table S3), which were each run on three SMRT cells. Mock community samples and the emPCR sample were pooled in equimolar ratio and sequenced using two SMRT cells.

Demultiplexing and extraction of subreads from SMRT cell data were performed applying the RS_ReadsOfInsert.1 protocol included in SMRTPortal 2.3.0 with a minimum of two full passes and a minimum predicted accuracy of 90%. Barcodes were provided as FASTA files, and barcode extraction was performed in a symmetric manner with a minimum barcode score of 23 within the same protocol. Mean amplicon lengths of 3,800–4,500 kbp were confirmed. Demultiplexed reads were downloaded from the SMRT Portal as fastq files for further analysis.

2.5 | Long-read metabarcoding pipeline

We developed an analysis pipeline for PacBio CCS reads using the python workflow engine SNAKEMAKE (version 3.5.5, Köster & Rahmann, 2012). Our pipeline combines steps directly implemented in python with steps that use external tools. The implementation is available on github (https://github.com/f-heeger/long_read_metabarcoding), and parameters used for the external tools can be found in the Supporting Information Methods (Supporting Information Appendix S3).

2.5.1 | Read Processing stage

Reads longer than 6,500 bp were excluded to remove chimeric reads formed during adapter ligation as well as reads containing double inserts due to failed adapter recognition during the CCS generation. Reads shorter than 3,000 bp were removed to exclude incompletely amplified sequences and other artefacts. Remaining reads were then filtered by a maximum mean predicted error rate of 0.004 that was computed from the Phred scores (see qualityFilter rule in Supporting Information Appendix S3 for details). Reads with local areas of low quality were removed if predicted mean error rate was >0.1 in any sliding window of 8 bp. CUTADAPT (version 1.9.1, Martin, 2011) was used to remove forward and reverse amplification primers and discard sequences in which primers could not be detected. Random errors were reduced by preclustering reads from each sample at 99% similarity using the cluster_smallmem command in vsearch (version 2.4.3, Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Reads were sorted by decreasing mean quality prior to clustering to ensure that high-quality reads were used as cluster seeds. vsearch was configured to produce a consensus sequence for each cluster.

2.5.2 | OTU clustering and classification stage

Chimeras were detected and removed with the uchime_denovo command in vsearch. Based on tests using mock community samples (see below), we determined this was a suitable method of chimera detection following the Read Processing stage (above). Only sequences that were classified as nonchimeric were used for further analysis. The rRNA genes (SSU, LSU and 5.8S) and internal transcribed spacers (ITS1 and ITS2) in each read were detected using ITSx (version 1.0.11, Bengtsson-Palme et al., 2013). To generate OTUs, the ITS region (ITS1, 5.8S and ITS2) was clustered using vsearch at 97% similarity. SSU and LSU sequences were then placed into clusters according to how their corresponding ITS was clustered. OTUs were taxonomically classified using the most complete available database for each marker. For the ITS, we used the general FASTA release of the UNITE database (version 7.1, 20.11.2016, only including singletons set as RefS, Kõljalg et al., 2013); for the SSU, we used the truncated SSU release of the SILVA database (version 128, Quast et al., 2013), excluding database sequences with quality scores below 85 or Pintail chimera quality below 50; and for the LSU, we used the RDP LSU data set (version 11.5, Cole et al., 2014). The ITS, SSU and LSU regions of the representative sequence of each OTU were locally aligned to the database using LAMBDA (version 1.9.2, Hauswedell, Singer, & Reinert, 2014). For LSU and SSU, the alignment parameters had to be modified to allow for longer alignments (see Supporting Information Appendix S3). From the alignment results, a classification was determined by filtering the best matches and generating a lowest common ancestor (LCA) from their classifications as follows. For each query sequence, matches were filtered by a maximum e-value (10^{-6}), a minimum identity (80%) and a minimum coverage of the shorter of the query or database sequence (85%). For the SSU and LSU, nonoverlapping matches between each query and database sequence were combined. For each query sequence, a cut-off for the bit score was established representing 95% of the value for the best match, above which all matches for that given sequence were considered. For the SSU and LSU, bit scores were normalized by the minimum length of query and database sequences to account for the varying lengths of database sequences. To determine the LCA from the remaining matches, their classifications were compared at all levels of the taxonomic hierarchy starting at kingdom (highest) and ending at species (lowest) level. For each OTU, the LCA classification was determined by comparing the classifications of matches remaining after the filtering described above at each taxonomic rank. If >90% of matches were to the same taxon, then this was accepted. If <90% were the same, then the OTU remained unclassified at this and all lower ranks.

2.6 | Error rates based on isolate sequences

Isolate sequences were processed using the Read Processing stage of the pipeline (described above) in order to generate error-corrected consensus sequences from preclusters. The consensus sequences of the largest precluster for each isolate were >99% identical to the Sanger sequencing data obtained from the same isolate (not shown), with most differences found in bases that were of low quality in the Sanger sequence data. We therefore used the consensus sequence of the largest cluster for each isolate as a reference for that species in all further analysis. CCS reads from each isolate were then aligned with the respective consensus sequence using blasr (github comit 16b158d, Chaisson & Tesler, 2012) to estimate error rates of CCS reads. Sequences after filtering steps were also compared in order to estimate remaining errors.

2.7 | Evaluating chimera detection

De novo and reference-based chimera classifications were compared as a way of estimating the reliability of de novo chimera calls. The CCS reads from the mock community samples were tested for chimeras with vsearch once in de novo mode (uchime_denovo) and once with a reference-based approach (uchime_ref). For the de novo approach, reads were processed with the Read Processing stage of the pipeline (above) to generate error-corrected sequences from preclusters. Cluster sizes resulting from the preclustering step were used as sequence abundances. For the reference-based approach, a reference file was created from the consensus sequence of the largest cluster for each isolate sample. A random subset of reads (100 sequences, 1.3% of the data) was generated from the mock community sample with the highest chimera rate and the most reads (30 PCR cycles). The subset of reads was aligned to the consensus sequences from the isolate samples and visually inspected for chimeras in GENEIOUS (version 7.1.9, Kearse et al., 2012). These “manual” chimera calls were then used to verify reference-based chimera classifications for these reads. Chimeras identified by the reference-based approach were used to compute the chimera formation rate under different PCR conditions.

2.8 | Mock community classification

We tested classification with the DNA metabarcoding pipeline using the mock community sample with the most reads (30 PCR cycles). In the pipeline, chimeras were classified de novo and OTU classification was performed using the public databases. We manually classified the same OTUs using consensus sequences from our isolate samples as reference. For each read, chimeras were detected with a reference-based approach using vsearch and the classification of the read was determined by mapping reads to the isolate sample sequences with blasr. To better understand the resolution that can be expected from the different regions of the rRNA operon, each region (SSU, ITS1, 5.8S, ITS2 and LSU) was clustered independently. Chimeras were first removed using the reference-based approach with our isolate sequences as references. The different regions in each read were separated with ITSx, dereplicated and clustered at 97%.

2.9 | Environmental community classification

Sequences from the environmental samples from Lake Stechlin were processed with the full rRNA metabarcoding pipeline described above. Chimeras were detected using the de novo approach, which we conclude provides a very good diagnosis of chimeras based on our validation using the mock community to compare de novo and reference-based approaches (see Results). The resulting classifications obtained with SSU, ITS and LSU markers were then compared at each taxonomic level. OTUs with only one read (singletons) were excluded from this comparison.

3 | RESULTS

Sequencing resulted in a total number of 233,176 CCS reads, which were submitted to the NCBI Sequence Read Archive (SRR6825218–SRR6825222). A total of 215,720 of these reads were within the targeted size range of 3,000–6,500 bp (Table 2). After stringent filtering using average and window quality criteria, 69,342 reads remained that contained an identifiable amplification primer sequence (Table 2). Preclustering of isolate samples with the metabarcoding pipeline resulted in one large (>80 reads) precluster for each sample. Besides these big clusters, six samples had

TABLE 2 Number of sequencing reads remaining after each step in the bioinformatics pipeline for each sample type

Analysis step	Isolates	Mock community	Environmental samples	Total
Raw CCS	46,740	60,448	125,988	233,176
Length-filtered	44,595	53,730	117,395	215,720
Average quality-filtered	18,532	16,054	48,778	83,364
Window quality-filtered	16,353	11,263	43,385	71,001
Primer-filtered	16,082	10,891	42,369	69,342

additional very small (<3 reads) clusters. For isolates sequenced on two different SMRT cells, consensus sequences of the large preclusters were identical across cells except for *Saccharomyces cerevisiae* where a T homopolymer in the ITS2 was 6 bases long in one consensus and 7 in the other and *Ustilago maydis* which shows a 1 bp difference in the LSU. Consensus sequences of large clusters were used as reference for further analysis and submitted to gene bank (MH047187–MH047202). The mean sequencing error rate of quality-filtered CCS reads, based on comparison with the consensus sequences of the large clusters (taken to be our reference for each isolate), was 0.2216% (*SD* 0.1621%). Deletions were by far the most common error (0.1756%), with insertions and substitutions much lower (Table 3).

3.1 | Chimera formation and detection

Using reference-based chimera detection in the mock community, chimera formation rate (i.e., sequences classified as chimeras or as unsure) rose from <2% of sequences at 13–18 PCR cycles to 16.3% at 30 cycles (Figure 2, Supporting Information Table S2). The emPCR (25 cycles) resulted in 4.4% of sequences classified as chimeric (Figure 2, Supporting Information Table S2), compared to 14.1% for 25 cycles under standard PCR conditions. Template DNA amounts played no measurable role in chimera formation rate, with 2, 8 and 20 ng of DNA all resulting in <2% chimeric sequences (18 cycles). Manual inspection of 100 randomly chosen isolate sequences classified 16 of these as chimeras. Reference-based detection identified 15 of these as chimeric and one as "suspicious." Of the 84 confirmed as nonchimeric by manual inspection, the reference-based algorithm classified 82 (97.6%) as nonchimeric and 2 as "suspicious." De novo chimera detection (i.e., in the absence of a reference) classified 98.6% of the reads in the sample in the same way as using the reference-based approach.

3.2 | Mock community classification

OTU clustering based on the ITS region variability resulted in 16 nonsingleton OTUs. Twelve OTUs consisted of sequences from one species as well as a few chimeric sequences, one contained sequences from *Cladosporium herbarum* and the other *Cladosporium* sp., and three smaller OTUs were entirely made up of chimeric sequences (Table 4). *Mortierella elongata* and *Cystobasidium laryngis* did not appear in any OTUs. This is because the few reads (<10) we did observe from these species in the mock community raw data were removed during quality filtering.

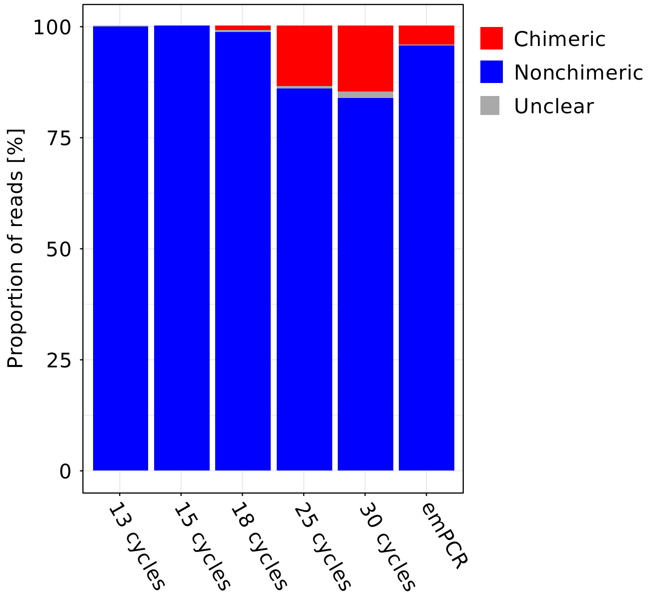

FIGURE 2 Chimera calls by vsearch with reference-based approach for different PCR conditions. Reads were classified as "chimeric" (red), "non-chimeric" (blue) or "unclear" (grey) [Colour figure can be viewed at wileyonlinelibrary.com]

When mock community reads were independently clustered by each of the five rRNA operon regions (SSU, ITS1, 5.8S, ITS2 and LSU) with a 97% similarity threshold, seven species were consistently distinguished, that is, formed a single cluster for all five regions (Figure 3). *Metschnikowia reukaufii* produced multiple clusters for ITS1 and ITS2, as expected based on previous reports of extraordinarily high rRNA operon variation in this genus (Lachance, Bowles, & Starmer, 2002; Sipiczki, Pfliegler, & Holb, 2013). *Clavariopsis aquatica* and *Phoma* sp. were separated by all regions except SSU, which grouped them together. *Trichoderma reesei* and *Clonostachys rosea* were separated by ITS1, ITS2 and LSU but not with SSU and 5.8S genes. *Cladosporium herbarum* and *Cladosporium* sp. were differentiated only with the ITS2, although one of the two clusters was mixed (Figure 3).

OTUs were classified into varying taxonomic ranks by the three different genetic markers (Table 4). The SSU gene provided mainly order- and family-level classifications, the ITS region provided family- to species-level classifications, and the LSU gene provided genus-level classifications in some cases and higher level classifications in others. The *Metschnikowia reukaufii* OTU was classified to different species by ITS (*M. cibodasensis*) and LSU (*M. bicuspidata*). Different genus-level classifications by ITS and LSU for the Chytrid species were the result of different taxonomies used in the UNITE and the

TABLE 3 Mean error rates (with standard deviation in parentheses) in CCS reads computed by mapping to consensus sequences of isolates

Analysis step	Substitutions mean (*SD*)	Insertions mean (*SD*)	Deletions mean (*SD*)	Total mean (*SD*)
Raw CCS	0.0453% (0.1277%)	0.3140% (0.6108%)	0.8650% (1.1960%)	1.2243% (1.5575%)
Filtered	0.0080% (0.0273%)	0.0380% (0.0476%)	0.1756% (0.1550%)	0.2216% (0.1621%)

TABLE 4 Mock community OTU classification with our analytical pipeline. Manual classifications were made by comparison with full-length reference sequences from species known to be present in the mock community. rRNA gene region classifications were made automatically by the pipeline with the LCA approach based on reference sequences in SILVA (SSU), UNITE (ITS) and RDP (LSU) databases. Size indicates the number of reads

		Classification method			
OTU	Size	Manual	SSU	ITS	LSU
11	6	*Clavariopsis aquatica*	Pleosporales (Order)	*Clavariopsis aquatica*	Pleosporales (Order)
6	44	Chytridiomycota	Chytridiomycetes (Class)	Globomyces (Genus)	Rhizophydium (Genus)
4	344	*Cladosporium* sp. + *Cladosporium herbarum*	Cladosporium (Genus)	Cladosporium (Genus)	Davidiella (Genus)
5	140	*Clonostachys rosea*	Hypocreales (Order)	Bionectriaceae (Family)	Hypocreales (Order)
1	4,165	*Metschnikowia reukaufii*	Saccharomycetales (Order)	*Metschnikowia cibodasensis*	*Metschnikowia bicuspidata*
2	1,096	*Leucosporidium scottii*	Basidiomycota (Phylum)	Leucosporidiaceae (Family)	*Leucosporidium* (Genus)
3	719	*Saccharomyces cerevisiae*	Saccharomycetaceae (Family)	Saccharomyces (Genus)	*Saccharomyces* (Genus)
7	37	*Penicillium brevicompactum*	Trichocomaceae (Family)	Penicillium (Genus)	Fungi (Kingdom)
8	34	*Ustilago maydis*	Ustilaginaceae (Family)	Ustilaginaceae (Family)	*Ustilago maydis*
9	20	*Exobasidium vaccinii*	Exobasidiales (Order)	*Exobasidium vaccinii*	Exobasidium (Genus)
10	21	*Phanerochaete chrysosporium*	Agaricomycetes (Class)	*Phanerochaete* sp.	Agaricomycetes (Class)
12	5	*Phoma* sp.	Pleosporales (Order)	Pleosporales Incertae sedis (Family)	Didymellaceae (Family)
13	5	*Trichoderma reesei*	Hypocreaceae (Family)	Trichoderma (Genus)	Hypocreaceae (Family)
14	3	chimeric	Saccharomycetales (Order)	Nectriaceae (Family)	*Metschnikowia bicuspidata*
16	3	chimeric	Saccharomycetales (Order)	*Metschnikowia cibodasensis*	unknown
17	9	chimeric	Saccharomycetales (Order)	*Metschnikowia cibodasensis*	Bionectria (Genus)

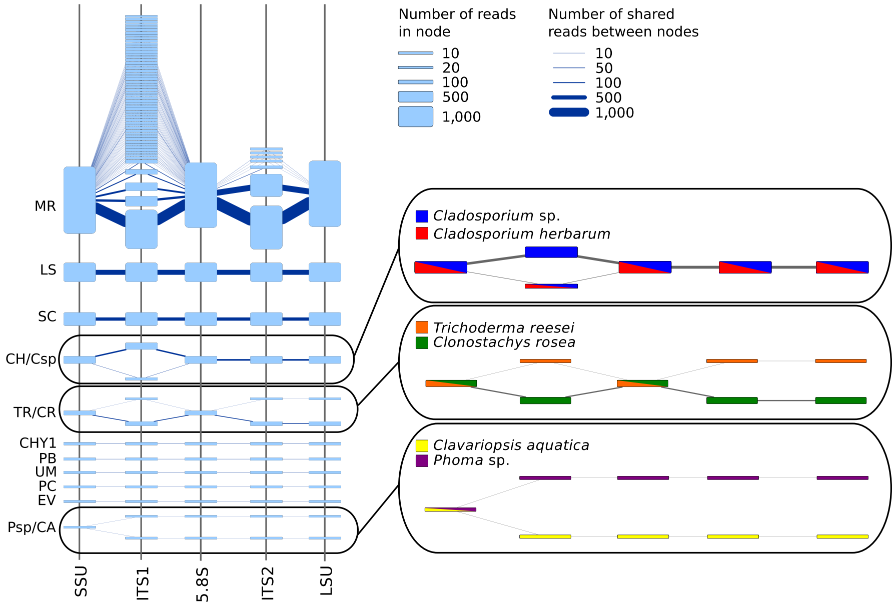

FIGURE 3 Resolution of different regions of the rRNA operon for our mock community. Each node represents a cluster, and each edge between two clusters represents shared reads between the clusters. Node height and edge thickness are proportional to read number. Nodes and edges with less than three reads are not shown. Identification codes are given in Table 1. Components with multiple species are shown in detail on the right. Nodes are coloured by species appearing in them. The graph was initially created with Cytoscape (version 3.2.1, Shannon 2003) and manually adapted for better readability [Colour figure can be viewed at wileyonlinelibrary.com]

RDP databases. The best match in both databases was *Globomyces pollinis-pini*, but the classification at higher ranks differs among the databases. Similar discrepancies caused by differences in database taxonomy also occurred for some of the other species. Other than that, classifications by all three markers were consistent with each other and with the manual classification.

3.3 | Environmental community classification

OTU clustering of the environmental samples produced 947 nonsingleton OTUs (Supporting Information Table S4), of which 799 (84%) were classified as fungi by at least one of the three markers (SSU, ITS and LSU). The SSU database also allowed identification of nonfungal sequences, and 112 OTUs were assigned to Metazoa, 10 to Discicristoidea, 2 to Stramenopiles, 2 to Alveolata and 1 to Chloroplastida. The 200 most abundant fungal OTUs (91% of fungal reads; 61% of total reads) were consistently classified to phylum level by all three markers except for nine cases in which SSU and LSU gave different classifications for the same OTU (Figure 4). There were no conflicts between SSU and ITS, although the SILVA and UNITE databases use different names for the phylum Cryptomycota/Rozellomycota (Figure 4). Classification at the phylum level was most successful with SSU (188 reads, i.e., 94% of the 200 most abundant fungal OTUs). Fewer OTUs were classified to phylum with LSU (126, 63%) and ITS regions (36, 18%). Classification to the species level was most successful with LSU (55, 27.5%) and less successful for ITS (20, 10%) and SSU (13, 6.5%; Figure 4).

Extended to all 947 OTUs, the results were similar. SSU provided the most classifications, especially for higher taxonomic ranks, and *ca.* 20% of these were classified the same using the ITS (Figure 5a) and *ca.* 66% were classified the same by LSU (Figure 5b). ITS classifications matched those of SSU (Figure 5c) and LSU (Figure 5d) at ranks from kingdom to class. At family, genus and species rank, most OTUs that were classified by ITS were not classified by SSU (Figure 5c) and many were classified differently by LSU (Figure 5d). At higher taxonomic rank (kingdom to class), OTUs classified by LSU were classified the same way as by SSU. But more than 50% were either not assigned to any taxon or were classified differently by SSU at lower ranks (order to species; Figure 5e). More than 60% of the OTUs classified by the LSU were not classified by ITS at all. Of those that were, classifications did however agree. At the order to species rank, OTUs classified by both LSU and ITS were rare and differences between the markers were more common (Figure 5f).

4 | DISCUSSION

Long sequencing reads have the potential to provide many benefits for DNA metabarcoding. These include taxonomic assignment of OTUs at lower taxonomic levels (Franzén et al., 2015; Porter & Golding, 2011), the use of homology-based classification and phylogenetic reconstruction (Tedersoo et al., 2017) and higher sequencing quality for standard length DNA barcodes in reference databases (Hebert et al., 2018). Disadvantages of long reads include lower sequence quality (D'Amore et al., 2016; Glenn, 2011), a possible increase in the rate of chimera formation and the fact that most bioinformatics tools are optimized for shorter reads. Here, we produced DNA metabarcodes nearly twice as long as any used to date (*ca.* 4,500 bp), comprising almost the whole eukaryotic rRNA operon (SSU, ITS and partial LSU). We combined circular consensus

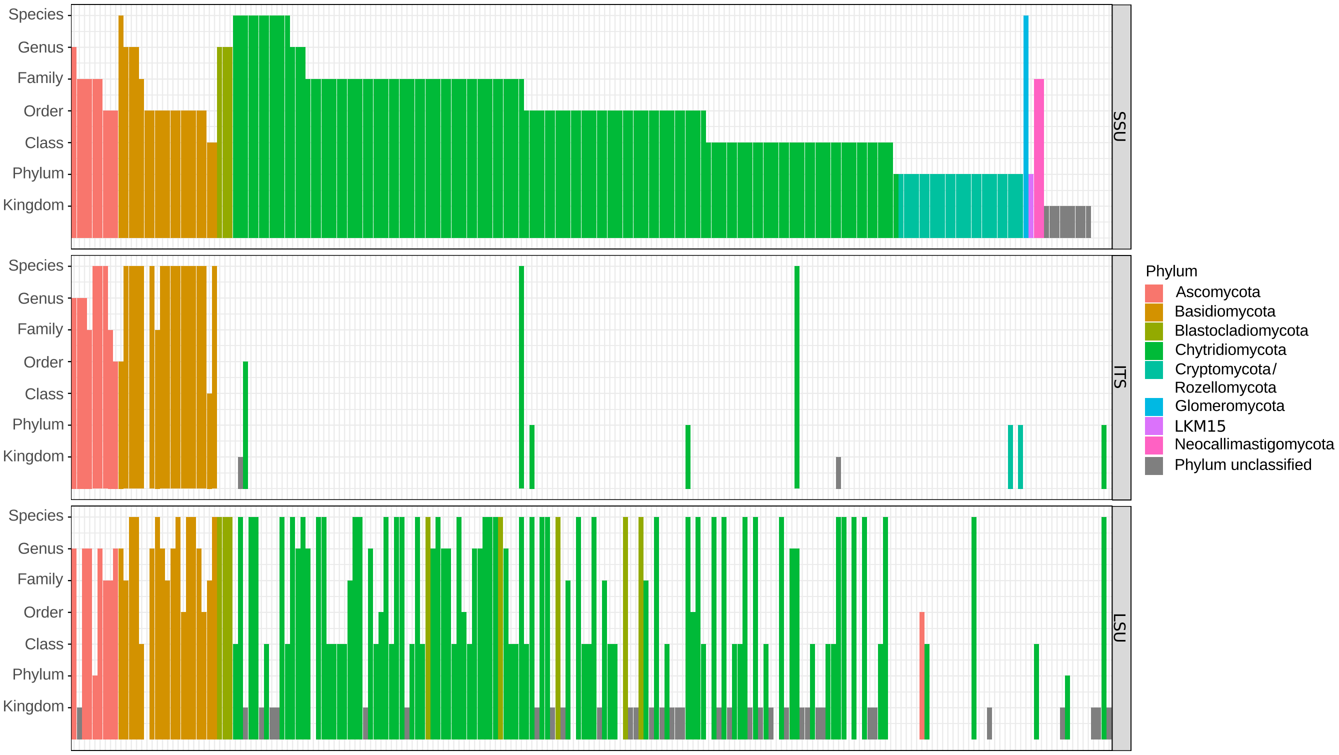

FIGURE 4 Classification specificity of the 200 most abundant fungal OTUs for the three different regions (SSU, ITS and LSU). The three rows give classifications by the three different regions. Each OTUs classification is given by a bar in each row. The height of the bar represents level of classification. Bars are coloured by phylum

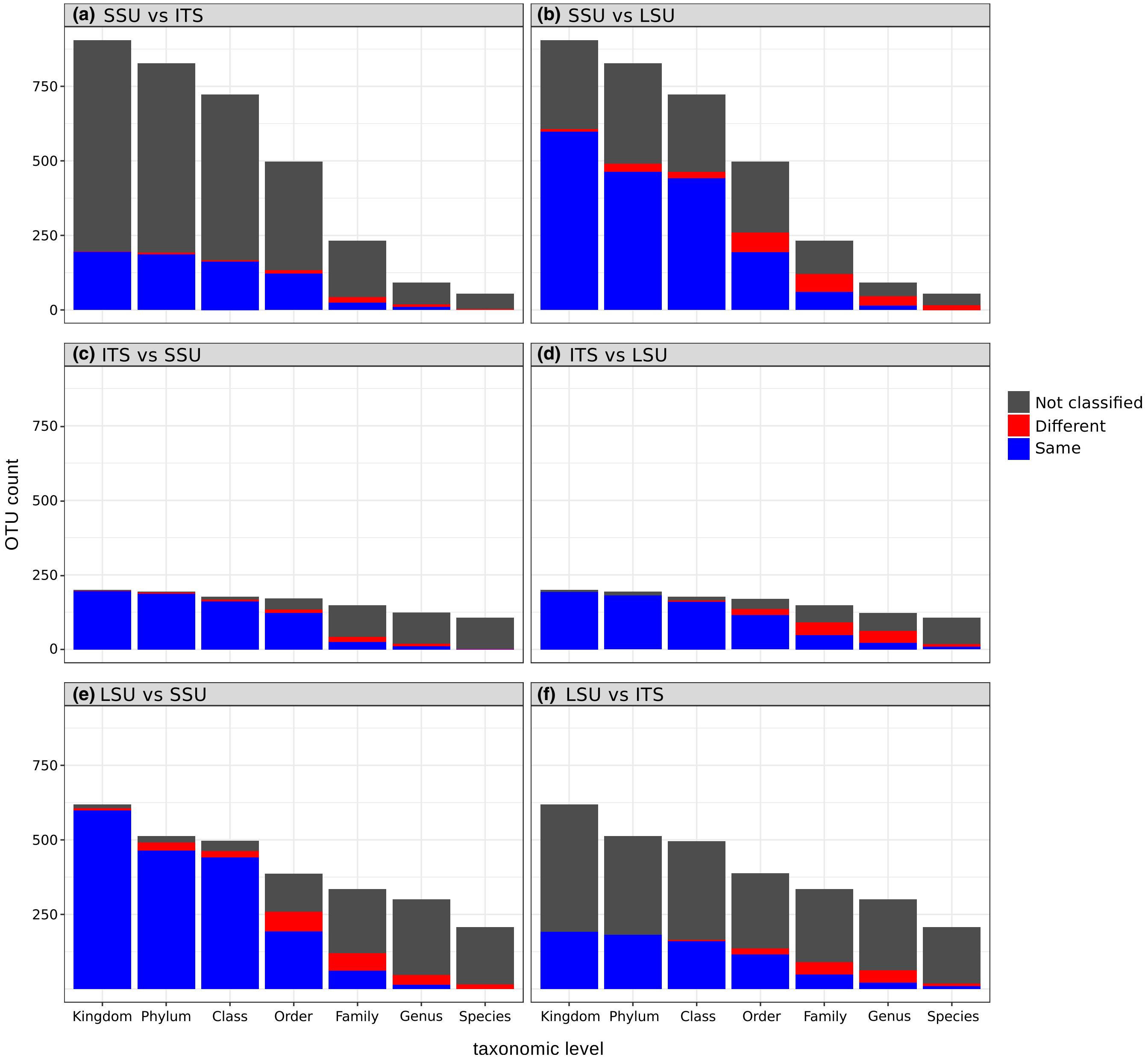

FIGURE 5 Agreement of classifications of all OTUs by the different regions. Each panel represents a comparison between two regions. Each set of stacked bars shows numbers of agreeing (blue), disagreeing (red) and unknown (grey) OTU classifications in the second region of the comparison compared to the first, at each taxonomic level

sequencing with our newly developed bioinformatics pipeline and obtained error rates comparable to short-read Illumina sequencing (D'Amore et al., 2016; Glenn, 2011). The use of multiple markers allowed us to use the ITS region for OTU delineation (clustering) and automated species-level taxonomic classifications for environmental OTUs with both ITS and LSU sequences. Finally, the inclusion of the SSU rRNA gene into the analyses allowed us to classify OTUs that were not represented in ITS and LSU databases, including many fungi that belong to basal lineages and are common in freshwater habitats (Rojas-Jimenez, Fonvielle, et al., 2017; Rojas-Jimenez, Wurzbacher, et al., 2017; Wurzbacher et al., 2016).

4.1 | Challenges of long reads

A significant challenge in using longer reads for DNA metabarcoding of mixed samples is the fact that most bioinformatics tools have been designed for the analysis of short sequences (typically 200–600 bp). Although we obtained very high-quality CCS reads, the higher indel rate and accumulation of errors in long reads require analyses that differ from that of more commonly used sequencing platforms like Illumina. For example, the clustering algorithm applied by swarm (Mahé, Rognes, Quince, Vargas, & Dunthorn, 2015) relies on a low total number of errors per sequence (ideally one error). In long sequences, even with low error rates, the total number of errors is higher, making it unfeasible to use this algorithm. Other widely used clustering tools such as uclust (Edgar, 2010) or vsearch use heuristics to choose starting points for clustering. Reads are first dereplicated and those with the most identical copies are used as cluster starting points, based on the assumption that they are the most likely to be real sequences. Because the comparably high nucleotide deletion rate and the long read length made almost all reads unique, dereplication would not lead to large clusters in our data and would thus not provide useful starting points for clustering.

In the future, it might be beneficial to develop specialized software for clustering and correcting PacBio long-range amplicons. Here, we used heuristic clustering starting with high-quality reads and with a high similarity threshold (99%), and a consecutive consensus calling for correction of random sequencing errors. This also gave us clusters of highly similar sequences, that we could use for chimera detection and OTU clustering instead of the groups of identical reads resulting from dereplication, that are normally used for these steps.

An important consideration of SMRT CCS is the much higher cost per base compared to short-read methods (Goodwin et al., 2016). Here, we used three SMRT cells per environmental sample to generate more than 20,000 reads per sample. This allowed us to apply a very strict quality filtering and still retain a number of reads (>7,000) per sample that was comparable to other short-read metabarcoding studies on fungi (Rojas-Jimenez, Fonvielle, et al., 2017; Rojas-Jimenez, Wurzbacher, et al., 2017; Yang et al., 2016; Zhang, Wang, Zhang, Liu, & Yu, 2016). We also found that the read number per OTU in the mock community (Table 4) did not correlate with the input DNA concentration (Table 1), and for two species, we did not recover any reads. Our main goal in this study was not quantitative representation of the community but the improved identification of novel taxa, but it is clear that additional optimization of the primers would improve species recovery.

One of the problems in any study applying PCR to mixed samples is chimera formation. Our comparison of de novo and reference-based chimera detection found them to produce the same classifications in >98% of cases. This indicates that de novo chimera classification in our long-read pipeline provided a good estimate of chimera formation rate and is suitable for data sets where no complete reference database is available. We can therefore be confident in our results for the environmental samples, even where no reference sequences were available in databases. Interestingly, a manual inspection of conflicting read assignments in the independent clustering of the different regions (data not shown) found a few cases (9 of 6,585 reads in the one mock community sample) of chimeras that could not be detected. Neither reference-based nor de novo approaches detected these chimeras because 3’ and 5’ ends were both from the same species, and only the central section originated from a second species. Most chimera detection software, including vsearch, models chimeras from two origins, that is, different 3' and 5' ends, but not more. These methods would then fail to identify chimeras if the 3' and 5' ends are from the same species and a second species is in the middle, as we observed. Although this was very rare in our data (0.1% of reads investigated), it created small OTUs made up almost entirely of these complex chimeras in our mock community (OTU 14, 16 and 17, see Table 4). As a general rule, chimeras are most likely to be found associated with the most frequent sequences in a PCR sample (Sommer, Courtiol, & Mazzoni, 2013) and this is also true for the complex chimeras we observed here. In fact, all three chimeric OTUs found in our mock community involved the species with the highest read abundance, *Metschnikowia reukaufii*. DECIPHER (Wright, Yilmaz, & Noguera, 2012) is one tool that may detect these chimeras, but requires a complete reference database of possible parent sequences and is therefore unsuitable for use with environmental samples (for which reference sequences are difficult to obtain) and long reads.

We also attempted to minimize chimera formation in the laboratory, by exploring the influence of reduced PCR cycle numbers, emulsion PCR and template concentration. Although we were initially concerned that our *ca.* 4,500 bp amplicon length would lead to higher chimera formation rates during PCR, the mock community sample that was amplified with the highest cycle number (30) formed chimeras at a rate within the range reported by short-read studies (ca. 4%–36%; Ahn, Kim, Song, & Weon, 2012; Qiu et al., 2001). We observed reduced chimera formation with fewer cycles which is also consistent with short-read studies (D'Amore et al., 2016; Lahr & Katz, 2009; Qiu et al., 2001). Unlike other studies (D'Amore et al., 2016; Lahr & Katz, 2009), we did not find a notable influence of DNA template concentration in our samples, possibly because at 18 cycles all reactions were still in the exponential phase, before depletion of reagents (see below). Chimera formation rates in our mock community may underestimate rates in environmental samples because the lower species richness in the mock community may have led to reduced chimera formation (Fonseca et al., 2012). However, the chimera rate detected by de novo chimera detection in our environmental data was <1%, that is, even lower than the de novo detection rate in the less diverse mock community samples. Chimera formation occurs primarily during the saturation phase of a PCR, when a large amount of PCR product has accumulated and the template:primer ratio increases (Judo, Wedel, & Wilson, 1998). Normally, PCR cycle number is used as a proxy for the amount of accumulated product. In our case, for a given cycle number, the amount of accumulated product may differ between the environmental and mock community samples, because although a similar amount of template DNA was used in mock community (8 ng) and environmental (10 ng) samples, the amount of template available for primer binding might be lower in the latter because they also contain non-fungal DNA. Environmental samples often contain PCR inhibitors (Schrader, Schielke, Ellerbroek, & Johne, 2012), which would reduce PCR efficiency. Together, these two effects could delay the saturation phase to a higher cycle number in environmental samples compared to the mock community. This could make the relation between cycle number and chimera formation rate noncomparable between mock community and environmental samples. Optimization of DNA extraction and amplification could make lower PCR cycle numbers feasible and thus further reduce the problem of chimera formation. Our emPCR results also indicate that this might be a promising way of reducing chimera formation when more PCR cycles are required.

4.2 | Classification

Although the ITS region has been proposed as a standard barcode for fungi (Schoch et al., 2012), other regions of the rRNA operon remain popular choices as fungal barcodes (Roy et al., 2017; Stielow et al., 2015; Wurzbacher et al., 2016). Compared to rRNA genes, ITS1 and ITS2 often exhibit higher interspecific variability and thus can provide greater species delineation power (i.e., more OTUs) than SSU and (in most fungal groups) LSU (Schoch et al., 2012). Indeed, we found that isolate species of the same genus (Cladosporium) and even from the same order (Hypocreales) and subdivision (Pezizomycotina) could not be separated by the SSU (Figure 3) and that the use of ITS resulted more often in classification to species level than SSU and LSU in Dikarya (Figure 4). At the same time, the often higher variability of ITS also means that for new species that are not represented in the database it can be more difficult to find comparable sequences and thus to identify them to any level. In these cases, longer sequencing reads that include more conserved regions with a stable evolutionary rate are likely to be helpful in making classifications based on sequence similarity as we did here or by phylogenetic methods (e.g., Tedersoo et al., 2017). Sequences of species from the phylum Chytridiomycota, which are often found in aquatic environments and were highly abundant in our environmental samples, are underrepresented in sequence databases (Frenken et al., 2017). We observed many OTUs from this phylum that could not be classified using our pipeline and parameters with the ITS at all, while the SSU provided at least class or family rank classifications and the LSU often even provided classifications at species rank (Figure 4). In fact, we obtained more species-level classifications with the LSU than with the ITS, but this is most likely due to the fact that the LSU has more reference sequences available for Chytridiomycota than the ITS.

For the classification of the mock community, the different degrees of taxonomic resolution provided by the different markers were clear. The mock community consisted of species that are represented in the reference databases with sequences that were identical or very similar to the sequence that we found. In these cases, ITS was a superior marker region, since its greater variability allowed for higher resolution classification. While almost all classifications were correct, those obtained for ITS went down to at least family rank in all cases, and even to species rank for a third of the OTUs. LSU and SSU both provided far fewer specific classifications. Using the LSU marker, species levels classifications could be obtained for some OTUs, but others were only classified to higher taxonomic ranks (up to kingdom). Using the SSU marker, classification results were obtained between the ranks of order and family. In our environmental samples, the disadvantage of ITS becomes clear. If no closely related reference sequence was available, multiple matches to different taxa with similar alignment scores were found, thus precluding an unambiguous assignment of the OTU to any single taxon. In these cases, SSU and LSU markers provided at least classification at family or class level, while many OTUs stayed completely unclassified with the ITS.

The independent clustering of the different regions (SSU, ITS1, 5.8S, ITS2 and LSU) of the rRNA operon (Figure 2) also showed the higher resolution of ITS1 and ITS2, which were the only regions that separate almost all species from each other. On the other hand, for *Metschnikowia reukaufii*, they formed multiple clusters for one species. This is most likely the result of high variability of rRNA operon copies in *Metschnikowia* (Lachance et al., 2002; Sipiczki et al., 2013) in combination with the short ITS1 and ITS2 sequences (70 bp and 75 bp, respectively) which mean that very few (3) differences already constitute an identity difference of 3%.

4.3 | Classification conflicts and synergies

The conflicts we observed between classifications based on different marker regions and databases provide insights into a number of interesting problems. In some cases, they may either represent uncertainty in classification using at least one of the markers or genuine chimeric reads. In other cases, conflicts may highlight incompatibility among the taxonomies used by the databases or even errors in the databases (see also Nilsson et al., 2006). Many conflicts resulted from differences in naming convention and taxonomic placement in the different databases. One significant challenge is the different rate at which new phylogenetic insights are incorporated into the taxonomies of the databases. For example, multiple OTUs were classified with LSU and the RDP database to the more recently defined orders Rhizophydiales (Letcher, Powell, Churchill, & Chambers, 2006) and Lobulomycetales (Simmons, James, Meyer, & Longcore, 2009), but were classified with SSU and the SILVA database as Chytridiales, the older classification for these new orders. Other conflicts exemplify how minor problems in the databases can lead to major differences in classification. In our environmental data, several high (read) abundance OTUs were classified as Chytridiomycota with SSU but as Blastocladiomycota with LSU. Closer inspection of the LSU alignments indicated that, for many of these OTUs, only the second best hit was to a Blastocladiomycota. The best match was *Rhizophlyctis rosea* which is a Chytridiomycota but has no classification beyond kingdom in the RDP database file we used and was thus ignored for classification. In addition, the second best match which was used for classification was to a sequence from the genus *Catenomyces* which belongs to the phylum Blastocladiomycota according to RDP, but according to SILVA belongs to the phylum of Chytridiomycota. Thus, a minor error in the database file, in combination with inconsistencies in the taxonomy used by different databases, can lead to completely different classifications when using different markers.

These conflicts in classification clearly highlight problems with the databases, but classifications using three different markers from the same molecule, as obtained from the full rRNA operon, can help us to evaluate how confident we can be in our classification. A classification that is supported by three markers, with largely independent databases, can be considered more trustworthy than one that is only supported by one or even shows conflicts when using different markers. In addition, long DNA barcodes could be used to create synergies between the databases and to support short-read studies. For example, if a sequence was classified to the same family by SSU (SILVA) and LSU (RDP), the ITS region could be added to the Unite database (even if it is not classified to the species level) to help future studies that use ITS markers. The possibility to sequence SSU, ITS and LSU at the same time therefore offers the opportunity to contribute to different databases in parallel, with the future potential to generate a new reference data set with nearly full-length rRNA operon sequences.

5 | CONCLUSIONS

We used a DNA metabarcode nearly twice the length of any used to date and created a long-read (*ca*. 4,500 bp) bioinformatics pipeline that results in rates of sequencing error and chimera detection that are comparable to typical short-read analyses. The approach enabled the use of three different rRNA sequence reference databases, thereby providing significant improvements in taxonomic classification over any single marker. While ITS is likely to remain a short-metabarcode region of choice for some time, a clear limitation of ITS is that its high variability, in combination with the incompleteness of databases for early diverging fungal lineages, often leads to classification failing. In these cases, the other rRNA markers are beneficial. In particular, classification based on SSU or LSU was superior in more basal fungal groups. The universal nature of the rRNA operon and our recovery of >100 nonfungal OTUs indicate that the method could (with adapted primers) also be suitable for more general studies of eukaryotic biodiversity.

ACKNOWLEDGEMENTS

We thank Lars Ganzert, Katrin Premke and Robert Taube (IGB) for helping with field sampling; Keilor Rojas and Silke Van den Wyngaert (IGB) for providing isolates; Christian Wurzbacher (Univ. Gothenburg, now TU Munich) for providing primers and for initial interesting discussions on long-read sequencing; and Nicole Heyer and Simone Severitt (DSMZ) for helping with sequencing. Research was partially funded by the Leibniz Association Pakt/SAW project "MycoLink" (SAW-2014-IGB-1). We appreciate the comments of L. Tedersoo and two anonymous reviewers on an earlier version of the manuscript.

AUTHOR CONTRIBUTIONS

The study was conceived and designed by F.H., E.C.B., C.B., A.Y., J.O., C.J.M. and M.T.M.; molecular laboratory work was designed and performed by E.C.B.; the analysis pipeline was designed and implemented and the analysis was carried out by F.H.; sequencing strategy was advised and library preparation and sequencing was performed by B.B. and C.S.; isolates for the mock community were chosen and cultivated by E.C.B, C.B. and A.Y.; the first draft was written by F.H., E.C.B., C.J.M. and M.T.M.; and the final manuscript was contributed by all authors.

DATA ACCESSIBILITY

Raw SMRT CCS read are available at the NCBI Short Read Archive (SRR6825218 – SRR6825222). Consensus sequences of isolate amplicons are available on GenBank (MH047187 – MH047202).

ORCID

Felix Heeger http://orcid.org/0000-0003-3519-2973
Elizabeth C. Bourne http://orcid.org/0000-0002-6369-9133
Andrey Yurkov http://orcid.org/0000-0002-1072-5166
Boyke Bunk http://orcid.org/0000-0002-8420-8161
Jörg Overmann http://orcid.org/0000-0003-3909-7201
Michael T. Monaghan http://orcid.org/0000-0001-6200-2376

REFERENCES

Ahn, J.-H., Kim, B.-Y., Song, J., & Weon, H.-Y. (2012). Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *Journal of Microbiology*, *50*(6), 1071–1074. https://doi.org/10.1007/s12275-012-2642-z

Bengtsson-Palme, J., Martin, R., Martin, H., Sara, B., Zheng, W., Anna, G., … Michael, B. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution*, *4*(10), 914–919. https://doi.org/10.1111/2041-210X.12073

Blaalid, R., Kumar, S., Nilsson, R. H., Abarenkov, K., Kirk, P. M., & Kauserud, H. (2013). ITS1 versus ITS2 as DNA metabarcodes for fungi. *Molecular Ecology Resources*, *13*(2), 218–224. https://doi.org/10.1111/1755-0998.12065

Boers, S. A., Hays, J. P., & Jansen, R. (2015). Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Scientific Reports*, *5*(1), 14181. https://doi.org/10.1038/srep14181

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics*, *13*(1), 238. https://doi.org/10.1186/1471-2105-13-238

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(D1), D633–D642. https://doi.org/10.1093/nar/gkt1244

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., … Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, *17*, 55. https://doi.org/10.1186/s12864-015-2194-9

Davison, J., Moora, M., Öpik, M., Adholeya, A., Ainsaar, L., Bâ, A., … Zobel, M. (2015). Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science*, *349*(6251), 970–973. https://doi.org/10.1126/science.aab1161

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Fonseca, V. G., Nichols, B., Lallias, D., Quince, C., Carvalho, G. R., Power, D. M., & Creer, S. (2012). Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Research*, *40*(9), e66–e66. https://doi.org/10.1093/nar/gks002

Franzén, O., Hu, J., Bao, X., Itzkowitz, S. H., Peter, I., & Bashir, A. (2015). Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome*, *3*, 43. https://doi.org/10.1186/s40168-015-0105-6

Frenken, T., Elisabet, A., Berger, S. A., Bourne, E. C., Mélanie, G., Hans-Peter, G., … Ramsy, A. (2017). Integrating chytrid fungal parasites into plankton ecology: Research gaps and needs. *Environmental Microbiology*, *19*(10), 3802–3822. https://doi.org/10.1111/1462-2920.13827

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. https://doi.org/10.1111/j.1755-0998.2011.03024.x

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Hamad, I., Abdallah, R. A., Ravaux, I., Mokhtari, S., Tissot-Dupont, H., Michelle, C., … Bittar, F. (2018). Metabarcoding analysis of eukaryotic microbiota in the gut of HIV-infected patients. *PLoS One*, *13*(1), e0191913. https://doi.org/10.1371/journal.pone.0191913

Hauswedell, H., Singer, J., & Reinert, K. (2014). Lambda: The local aligner for massive biological data. *Bioinformatics (Oxford, England)*, *30*(17), i349–i355. https://doi.org/10.1093/bioinformatics/btu439

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., … Zakharov, E. V. (2018). A sequel to sanger: Amplicon sequencing that scales. *BMC Genomics*, *19*, 219. https://doi.org/10.1186/s12864-018-4611-3

Judo, M. S., Wedel, A. B., & Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Research*, *26*(7), 1819–1825. https://doi.org/10.1093/nar/26.7.1819

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … 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. https://doi.org/10.1093/bioinformatics/bts199

Kõljalg, U., Henrik, N. R., Kessy, A., Leho, T., Taylor, A. F. S., Bahram, M., … Larsson, K.-H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277. https://doi.org/10.1111/mec.12481

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Lachance, M.-A., Bowles, J. M., & Starmer, W. T. (2002). Metschnikowia santaceciliae, Candida hawaiiana, and Candida kipukae, three new yeast species associated with insects of tropical morning glory. *FEMS Yeast Research*, *3*(1), 97–103. https://doi.org/10.1111/j.1567-1364.2003.tb00144.x

Lahr, D. J. G., & Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, *47*(4), 857–866. https://doi.org/10.2144/000113219

Laver, T. W., Caswell, R. C., Moore, K. A., Poschmann, J., Johnson, M. B., Owens, M. M., … Weedon, M. N. (2016). Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Scientific Reports*, *6*, https://doi.org/10.1038/srep21746

Letcher, P. M., Powell, M. J., Churchill, P. F., & Chambers, J. G. (2006). Ultrastructural and molecular phylogenetic delineation of a new order, the Rhizophydiales (Chytridiomycota). *Mycological Research*, *110*(8), 898–915. https://doi.org/10.1016/j.mycres.2006.06.011

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., … Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers–a user's guide. *The New Phytologist*, *199*(1), 288–299. https://doi.org/10.1111/nph.12243

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420. https://doi.org/10.7717/peerj.1420

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

McFrederick, Q. S., & Rehan, S. M. (2016). Characterization of pollen and bacterial community composition in brood provisions of a small carpenter bee. *Molecular Ecology*, *25*(10), 2302–2311. https://doi.org/10.1111/mec.13608

Monchy, S., Sanciu, G., Jobard, M., Rasconi, S., Gerphagnon, M., Chabé, M., … Sime-Ngando, T. (2011). Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. *Environmental Microbiology*, *13*(6), 1433–1453. https://doi.org/10.1111/j.1462-2920.2011.02444.x

Nercessian, O., Noyes, E., Kalyuzhnaya, M. G., Lidstrom, M. E., & Chistoserdova, L. (2005). Bacterial populations active in metabolism of C1 compounds in the sediment of Lake Washington, a freshwater lake. *Applied and Environmental Microbiology*, *71*(11), 6885–6899. https://doi.org/10.1128/AEM.71.11.6885-6899.2005

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.-H., & Kõljalg, U. (2006). Taxonomic reliability of DNA sequences in public sequence databases: A fungal perspective. *PLoS One*, *1*(1), e59. https://doi.org/10.1371/journal.pone.0000059

Nilsson, R. H., Taylor, A. F. S., Adams, R. I., Baschien, C., Bengtsson-Palme, J., Cangren, P., … Abarenkov, K. (2018). Taxonomic annotation of public fungal ITS sequences from the built environment – a report from an April 10–11, 2017 workshop (Aberdeen, UK). *MycoKeys*, *28*, 65–82. https://doi.org/10.3897/mycokeys.28.20887

Ohsowski, B. M., Zaitsoff, D. P., Öpik, M., & Hart, M. M. (2014). Where the wild things are: Looking for uncultured Glomeromycota. *New Phytologist*, *204*(1), 171–179. https://doi.org/10.1111/nph.12894

Porras-Alfaro, A., Liu, K.-L., Kuske, C. R., & Xie, G. (2014). From genus to phylum: Large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition. *Applied and Environmental Microbiology*, *80*(3), 829–840. https://doi.org/10.1128/AEM.02894-13

Porter, T. M., & Golding, B. G. (2011). Are similarity- or phylogeny-based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? *New Phytologist*, *192*(3), 775–782. https://doi.org/10.1111/j.1469-8137.2011.03838.x

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., & Zhou, J. (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology*, *67*(2), 880–887. https://doi.org/10.1128/AEM.67.2.880-887.2001

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rojas-Jimenez, K., Fonvielle, J. A., Ma, H., & Grossart, H.-P. (2017). Transformation of humic substances by the freshwater Ascomycete *Cladosporium* sp. *Limnology and Oceanography*, *62*(5), 1955–1962. https://doi.org/10.1002/lno.10545

Rojas-Jimenez, K., Wurzbacher, C., Bourne, E. C., Chiuchiolo, A., Priscu, J. C., & Grossart, H.-P. (2017). Early diverging lineages within Cryptomycota and Chytridiomycota dominate the fungal communities in ice-covered lakes of the McMurdo Dry Valleys. *Antarctica. Scientific Reports*, *7*, https://doi.org/10.1038/s41598-017-15598-w

Roy, J., Reichel, R., Brüggemann, N., … M. C. (2017). Succession of arbuscular mycorrhizal fungi along a 52-year agricultural recultivation chronosequence. *FEMS Microbiology Ecology*, *93*(9),fix102. https://doi.org/10.1093/femsec/fix102

Schlaeppi, K., Bender, S. F., Mascher, F., Russo, G., Patrignani, A., Camenzind, T., … van der Heijden, M. G. A. (2016). High-resolution community profiling of arbuscular mycorrhizal fungi. *The New Phytologist*, *212*(3), 780–791. https://doi.org/10.1111/nph.14070

Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L., & Highlander, S. K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ*, *4*, e1869. https://doi.org/10.7717/peerj.1869

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Fungal Barcoding Consortium. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Schrader, C., Schielke, A., Ellerbroek, L., & Johne, R. (2012). PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology*, *113*(5), 1014–1026. https://doi.org/10.1111/j.1365-2672.2012.05384.x

Simmons, D. R., James, T. Y., Meyer, A. F., & Longcore, J. E. (2009). Lobulomycetales, a new order in the Chytridiomycota. *Mycological Research*, *113*(4), 450–460. https://doi.org/10.1016/j.mycres.2008.11.019

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., … Woyke, T. (2016). High-resolution phylogenetic microbial community profiling. *The ISME Journal*, *10*(8), 2020–2032. https://doi.org/10.1038/ismej.2015.249

Sipiczki, M., Pfliegler, W. P., & Holb, I. J. (2013). Metschnikowia species share a pool of diverse rRNA genes differing in regions that determine hairpin-loop structures and evolve by reticulation. *PLoS One*, *8*(6), e67384. https://doi.org/10.1371/journal.pone.0067384

Sommer, S., Courtiol, A., & Mazzoni, C. J. (2013). MHC genotyping of non-model organisms using next-generation sequencing: A new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, *14*(1), 542. https://doi.org/10.1186/1471-2164-14-542

Stielow, J. B., Lévesque, C. A., Seifert, K. A., Meyer, W., Iriny, L., Smits, D., … Robert, V. (2015). One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia : Molecular Phylogeny and Evolution of Fungi*, *35*, 242–263. https://doi.org/10.3767/003158515X689135

Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., … Abarenkov, K. (2015). Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys*, *10*, 1–43. https://doi.org/10.3897/mycokeys.10.4852

Tedersoo, L., Ave, T.-K., & Sten, A. (2017). PacBio metabarcoding of Fungi and other eukaryotes: Errors, biases and perspectives. *New Phytologist*, *217*(3), 1370–1385. https://doi.org/10.1111/nph.14776

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., … Abarenkov, K. (2014). Global diversity and geography of soil fungi. *Science*, *346*(6213), 1256688. https://doi.org/10.1126/science.1256688

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159. https://doi.org/10.1093/nar/gkq543

Walder, F., Schlaeppi, K., Wittwer, R., Held, A. Y., Vogelgsang, S., Heijden, V. D., & A, M. G., (2017). community profiling of fusarium in combination with other plant-associated Fungi in different crop species using SMRT sequencing. *Frontiers in Plant Science*, *8*, 2019. https://doi.org/10.3389/fpls.2017.02019

Wright, E. S., Yilmaz, L. S., & Noguera, D. R. (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Applied and Environmental Microbiology*, *78*(3), 717–725. https://doi.org/10.1128/AEM.06516-11

Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., … Nilsson, R. H. (2018). Introducing ribosomal tandem repeat barcoding for fungi. *bioRxiv*. https://doi.org/10.1101/310540

Wurzbacher, C., Rösel, S., Rychła, A., & Grossart, H.-P. (2014). Importance of saprotrophic freshwater fungi for pollen degradation. *PLoS One*, *9*(4), e94643. https://doi.org/10.1371/journal.pone.0094643

Wurzbacher, C., Warthmann, N., Bourne, E., Attermeyer, K., Allgaier, M., Powell, J. R., … Monaghan, M. T. (2016). High habitat-specificity in fungal communities in oligo-mesotrophic, temperate Lake Stechlin (North-East Germany). *MycoKeys*, *16*, 17–44. https://doi.org/10.3897/mycokeys.16.9646

Yang, C., Schaefer, D. A., Liu, W., Popescu, V. D., Yang, C., Wang, X., … Yu, D. W. (2016). Higher fungal diversity is correlated with lower CO_2 emissions from dead wood in a natural forest. *Scientific Reports*, *6*, 31066. https://doi.org/10.1038/srep31066

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Functional Ecology*, *3*, 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zhang, T., Wang, N.-F., Zhang, Y.-Q., Liu, H.-Y., & Yu, L.-Y. (2016). Diversity and distribution of aquatic fungal communities in the Ny-Ålesund region, Svalbard (High Arctic). *Microbial Ecology*, *71*(3), 543–554. https://doi.org/10.1007/s00248-015-0689-1

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Heeger F, Bourne EC, Baschien C, et al. Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Mol Ecol Resour*. 2018;18:1500–1514. https://doi.org/10.1111/1755-0998.12937