

DR. CHRISTIAN WURZBACHER (Orcid ID : 0000-0001-7418-0831)

Article type : Resource Article

Introducing ribosomal tandem repeat barcoding for fungi

Running title: Tandem repeat barcoding for fungi

Christian Wurzbacher^{1,2,3}, Ellen Larsson^{1,3}, Johan Bengtsson-Palme^{4,5}, Silke Van den Wyngaert⁶, Sten Svantesson^{1,3}, Erik Kristiansson⁷, Maiko Kagami^{6,8,9}, R. Henrik Nilsson^{1,3}

Affiliations

1. Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 40530 Göteborg, Sweden.
2. Chair of Urban Water Systems Engineering, Technical University of Munich, Am Coulombwall 3, Garching 85748, Germany
3. Gothenburg Global Biodiversity Centre, Box 461, 405 30 Göteborg, Sweden
4. Wisconsin Institute for Discovery, University of Wisconsin-Madison, 330 North Orchard Street, Madison WI 53715, Wisconsin, USA.
5. Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, 413 46, Göteborg, Sweden

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12944

This article is protected by copyright. All rights reserved.

6. Leibniz-Institute of Freshwater Ecology and Inland Fisheries Berlin, Alte Fischerhütte 2, 16775 Stechlin, Germany
7. Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 412 96 Göteborg, Sweden
8. Department of Environmental Science, Faculty of Science, Toho University, 2-2-1 Miyama, Funabashi, Chiba, Japan
9. Graduate School of Environmental and Information Sciences, Yokohama National University, Tokiwadai 79-7, Hodogayaku, Yokohama, Kanagawa, Japan

Corresponding author: christian@wurzbacher.cc

Abstract

Sequence comparison and analysis of the various ribosomal genetic markers are the dominant molecular methods for identification and description of fungi. However, new fungal environmental fungal lineages known only from DNA data reveal significant gaps in our sampling of the fungal kingdom both in terms of taxonomy and marker coverage in the reference sequence databases. To facilitate the integration of reference data from all of the ribosomal markers, we present three sets of general primers that allow for amplification of the complete ribosomal operon from the ribosomal tandem repeats. The primers cover all ribosomal markers: ETS, SSU, ITS1, 5.8S, ITS2, LSU, and IGS. We coupled these primers successfully with third generation sequencing (PacBio and Nanopore sequencing) to showcase our approach on authentic fungal herbarium specimens (Basidiomycota), aquatic chytrids (Chytridiomycota), and a poorly understood lineage of early diverging fungi (Nephridiophagidae). In particular, we were able to generate high-quality reference data with

Nanopore sequencing in a high-throughput manner, showing that the generation of reference data can be achieved on a regular desktop computer without the involvement of any large-scale sequencing facility. The quality of the Nanopore generated sequences was 99.85 %, which is comparable with the 99.78 % accuracy described for Sanger sequencing. With this work, we hope to stimulate the generation of a new comprehensive standard of ribosomal reference data with the ultimate aim to close the huge gaps in our reference datasets.

Keywords

ribosomal operon, IGS, Sanger, PacBio, Nanopore, third generation sequencing

Introduction

In 1990 it became clear that ribosomes are common to all extant organisms known today (Woese, Kandler, & Wheelis, 1990). The ribosomal genetic markers are located in the ribosomal operon, a multi-copy region featuring genes and variously also spacers and other poorly understood elements (Rosenblad et al., 2016). Clearly defined fragments of these regions and spacers have been identified as suitable markers for various scientific pursuits in different groups of organisms (Hillis & Dixon, 1991; Tedersoo et al., 2015) depending on their individual substitution rates and length. Well-known examples include the SSU (16S/18S), which has been used to explore the phylogeny of prokaryotes and microeukaryotes and the ITS region, which is the formal DNA barcode for molecular identification of fungi (Schoch et al., 2012).

The fungal kingdom is vast, with estimates ranging from 1.5 to 6 million extant species (Taylor et al., 2015; Hawksworth & Lücking, 2017). On the other hand, a modest ~143,000 species are formally recognized (<http://www.speciesfungorum.org/Names/Names.asp>, accessed August 2018),

underlining a considerable knowledge gap. The discrepancy between the known and unknown fungi becomes readily apparent when doing environmental sequencing, where it is not uncommon to find for instance that >10% of all fungal species hypotheses (a molecular based species concept akin to operational taxonomic units, OTUs; Blaxter et al., 2005; Kõljalg et al., 2013) do not fall in any known fungal phylum (Nilsson et al., 2016). This hints at the presence of a large number of unknown branches on the fungal tree of life (Tedersoo, Brahm, Puusepp, Nilsson, & James, 2017). Compounding this problem, not all described fungi have a nucleotide record, which is often but not always related to older species descriptions made before the advent of molecular biology (Hibbett et al., 2016). Consequently, many studies fail to classify more than 15-20% of the fungal sequences to genus level (e.g., Wurzbacher et al., 2016), which severely hampers the interpretation of the results.

While the ITS is a well chosen barcode, it is less suitable for phylogenetic analysis and it is not optimal for all fungal lineages. For instance, the Cryptomycota taxonomy is based on the ribosomal SSU gene as the primary genetic marker (e.g. Lazarus & James, 2015), while Chytridiomycota taxonomists mainly work with ribosomal large subunit data (LSU/23S/26S/28S) (Letcher, Powell, Churchill, & Chambers, 2006). Similarly, studies on Zygomycota often employ the SSU and LSU (White et al., 2006), while work on yeast species is often done using the LSU (Burgaud et al., 2016). Researchers studying fungal species complexes regularly need to consider genetic markers with even higher substitution rates than the ITS (e.g., the IGS region, O'Donnell et al., 2009; Nilsson et al., 2018).

There are thus serious gaps in the reference databases relating to 1) taxonomic coverage and 2) marker coverage. Some groups have ample SSU data; others have a reasonable ITS and LSU coverage; others are known only from ITS or LSU data. We argue that it is crucial to close these two

types of gaps – ideally at the same time – to achieve robust data-driven progress in mycology. Having access to all ribosomal markers at once solves a range of pertinent research questions, such as trying to obtain a robust phylogenetic placement for an ITS sequence (e.g. James et al., 2006), trying to prove that an unknown taxon does indeed belong to the true fungi (Tedersoo et al., 2017), or moving forward in spite of the fact that an ITS sequence produces no BLAST matches at all in the public sequence databases (Heeger et al., 2018). Thus, an urgent goal is to fill the taxonomic and marker-related gaps in our reference sequence databases (e.g. SILVA: Glöckner et al., 2017, UNITE: Kõljalg et al., 2013, and RDP: Cole et al., 2014).

In this study we explore a promising way to close the gaps in the reference databases and simultaneously unite them through the use of emerging long-read sequencing technologies. Here, we aim to generate high-quality *de novo* reference data for the full ribosomal operon and the adjacent intergenic regions, which would unify five or more distinct marker regions: the SSU gene, the ITS region including the 5.8S rRNA gene, the LSU gene, and the IGS (that often but not always contains the 5S gene, too). Fortunately, the eukaryotic ribosomal operon is arranged in tandem repeats in the nuclear genome, which makes its amplification by PCR comparatively straightforward.

In theory at least, DNA sequencing of the full ribosomal operon is perfectly possible. So far, it has not been feasible to sequence such long DNA stretches in a simple, time and cost-efficient way. In principle, Sanger sequencing with maybe 10 internal sequencing primers is a possibility, or alternatively shot-gun sequencing could do it, which however would require prior fragmentation of the long amplicon. Both of these approaches are less than straightforward in that they require substantial time, multiple rounds of sequencing, and significant laboratory expertise.

In contrast, emerging third-generation sequencing technologies – MinION (Oxford Nanopore Technologies; <https://nanoporetech.com/>) and PacBio SMRT sequencing (Pacific Biosciences; <http://www.pacb.com/>) – offer the possibility to sequence long DNA amplicons in a single read, much like Sanger does so well for short amplicons. Both of these technologies are suitable for high-accuracy, long-range sequencing (Singer et al., 2016; Benitez-Paez & Sanz, 2017; Tedersoo, Tooming-Klunderud, & Anslan, 2018b; Karst et al., 2018). PacBio excels where high accuracy is crucial due to its circular consensus sequencing mode. The advantage of Nanopore sequencing is the moderate price and the fast processing time. In addition, there is no need for a sequencing provider when it comes to Nanopore sequencing, because the sequencing can be done through a regular desktop computer. We believe that these features combine to make both technologies invaluable for our envisioned generation of comprehensive reference data.

In this work, we present PCR primers to cover the whole fungal nuclear ribosomal region in either two shorter amplifications of 5 kilo-bases (kb) each or in a single long amplicon of approximately 10 kb. The end product of both approaches is a 10 kb long stretch of nucleotide data that comprises all ribosomal markers, thus forming reference sequence data that satisfy many different research questions at once. Our secondary objective is to provide a cost-efficient and easy-to-use system that can be adopted even in small laboratories with limited budgets to facilitate broad generation of complete ribosomal reference data that may, as a joint effort, eventually help to fill our knowledge gaps in mycology and elsewhere. We targeted three fungal phyla (Chytridiomycota, Basidiomycota, and the enigmatic Nephridiophagidae lineage; Radek et al. 2017) separated by more than a hundred million years of evolution (Tedersoo et al. 2018a) in an attempt to establish the wide taxonomic scope of our primers and approach.

Methods

Tested samples

We tested several samples of various origins to evaluate the primer systems for our respective fields of research with a focus on reference material from herbarium specimens (Supplemental Material S1). For Basidiomycota species within our target class Agaricomycetes, we tested DNA extracted from herbarium specimens deposited at the infrastructure of University of Gothenburg, Herbarium GB (n = 66). DNA extraction from the fungal material was performed with the DNA Plant Mini Kit (Qiagen, Hilden, Germany). We furthermore evaluated the use of the primers for a few early diverging and poorly known environmental fungal lineages. For parasitic uncultured aquatic fungi of the phylum Chytridiomycota we employed micromanipulation as described in Ishida et al. (2015). Briefly, cells or sporangia were identified under a microscope, picked by a micropipette, washed, transferred to a microreaction tube, and lysed. This was followed by a whole genome amplification (illustra Single Cell GenomiPhi V1/V2 DNA amplification kit; GE Healthcare), which provided the DNA template for our ribosomal PCRs (n = 9). Finally, to test the amplification of an extremely distant fungal lineage from an animal host, we worked with DNA extracted from Malpighian tubule tissue from two cockroaches infected with members of the early diverging fungal lineage *Nephridiophaga*, derived from previous work (Radek et al., 2017).

Primer design

The operon PCR was performed with newly designed and adapted primer pairs. The primer pairs were modified from previous primers, namely the universal NS1 primer that offers a broad coverage of many eukaryotic lineages (White, Bruns, Lee, & Taylor 1990) and a Holomycota/Nucleomycea-specific primer derived from the RD78 primer (Wurzbacher et al., 2014). The modified and further developed primers were designed and tested in ARB v. 6 (Ludwig et al., 2004) against the SILVA

reference databases v. 123 (Glöckner et al., 2017) for SSU and LSU. NS1 was shortened by three bases at the 3' end to avoid mismatches to major lineages in the Chytridiomycota and Cryptomycota. Furthermore, to maintain an acceptable melting temperature, the primer was prolonged with two bases at the 5' end from a longer version of NS1 mentioned in Mitchell & Zuccharo (2011), and was named NS1short: CAGTAGTCATATGCTTGTC¹. We further developed the RD78 primer with the Probedesign and Probematch tools integrated in ARB by shifting it towards the LSU 5' by several bases, so that the mismatches to outgroups such as Eumetazoa fall in the 3' end of the primer region. This will facilitate the application of the primer in, e.g., mixed-template samples such as environmentally derived DNA. Similar to RD78, the resulting primer is highly specific to true fungi, and we named it RCA95m (CTATGTTTTAATTAGACAGTCAG) since it matches more than 95% of all fungi in the SILVA LSU database. However, we found exceptions (mismatches) in a few long-branched Kickxellomycota (e.g., *Dimargaris* of the Dimargaritales) and in the close vicinity of the genus *Neurospora* (Sordariales, Ascomycota). A phylum-based *in silico* analysis of all primers used in this study can be found in Supplemental Material S1. This primer pair (NS1short and RCA95m) was used to amplify the greater part of the ribosomal operon (rDNA PCR; Figure 1) of the ribosomal tandem repeat. In order to amplify the missing parts of the ribosomal region (the 5' end of the LSU, IGS, and the ETS region), we simply used the reverse complementary version of each primer: NS1rc (ACAAGCATATGACTACTG) and RCA95rc (CTGACTGTCTAATTTAAACATAG). As a third primer pair we developed a primer pair based on RCA95m that binds to a single region in the LSU: the forward primer Fun-rOP-F (CTGACTGTCTAATTTAAACAT) amplifies in the 3' direction of the LSU, while the reverse primer Fun-rOP-R (TCAGATTCCTTGTCCGTA) amplifies in the 5' direction (Figure 1). Note

1

We noticed that there are a few cases in the reference data where this 5' prolongation „CA“ is replaced by „TC“ in several not necessarily related fungal species. Closer inspection of these cases by BLASTing and aligning to the SILVA SSU reference database (v. 128) showed that this was due to incomplete trimming of the sporadically used primer PNS1 (Hibbett, 1996), which produces severe 5' mismatches to the fungal backbone.

that the last four nucleotides of both Fun-rOP primers are pairing and that these four nucleotides resemble the position overlap in the template (CTGA) at the exact *Escherichia coli* reference position 1770-1773 of the LSU (SILVA LSU reference position). This allows a subsequent end-to-end assembly of the full ribosomal region (LSU-IGS-ETS-SSU-ITS1-5.8S-ITS2-LSU) extracted from the ribosomal tandem repeat. All primer pairs were barcoded following the dual indexing strategy of Illumina sequencing (Part#15044223Rev.B; Illumina, San Diego). That means that we introduced the forward barcode series S500 to the 5' end of each forward primer and the N700 barcode series to the 5' end of the reverse primer, which allows the simultaneous sequencing of more than 100 samples, at least in theory. These barcodes are exchangeable and could also be replaced by e.g., longer Nanopore barcodes. After each barcode we added one or two extra nucleotides as a precaution against nuclease activity. Between barcode and primer nucleotides we added a two-nucleotide wide spacer (see Supplemental Material S1) that has a mandatory mismatch to the fungal kingdom at these two positions. These two mismatches were validated by using ARB and the respective SILVA reference databases (SSU and LSU). We did not test all samples with all amplicons, since our primary focus in this study was the herbarium specimens (Table 1).

Long range PCRs

In general, we applied the PrimeStar GLX polymerase (Takara) for all primer systems. For the IGS stretch (IGS PCR) and the full ribosomal region tandem repeat PCR (TR PCR), we employed only the PrimeStar GLX (Takara). The PCR was performed in 40 μ l reactions with 1.5 μ l enzyme, 12 pmol of each barcoded primer (a unique combination for each sample), 1 mM dNTPs, and 1 μ l of template (with a concentration of approximately 1-40 ng/ μ l). For all primer pairs we ran an initial denaturation of 1 min at 98°C, then 36 cycles at 98 °C for 10 sec, 55 °C for 15 sec, and 68 °C for 2.5 min. We increased the elongation step of the TR PCR from 2.5 to 4 minutes. For the Chytridiomycota samples we exchanged the PrimeStar polymerase for Herculase II (Agilent Technologies) for the

rDNA PCR. The reason for this is that we worked on these samples in a second laboratory, in which the Herculase II was the established polymerase. We ran a two-step protocol for Herculase II. An initial PCR with native (non-barcoded) NS1short/RCA95m primers and 3% BSA (molecular grade, Carl-Roth) as additive was run with the following program: 5 min at 95°C, then 35 cycles at 95 °C for 30 sec, 55 °C for 30 sec, and 68° C for 4 min. The PCR product was then used as template in a second PCR with 10 cycles but otherwise identical conditions, exchanging the native primers with barcoded primers. In general, we recommend the use of a polymerase that is specifically engineered for long amplicons for all three amplifications (e.g., PrimeStar GLX).

Library preparation and sequencing

The PCR products were purified with either 0.8 (v/v) of AMPure beads (Beckmann) or with PCR purification plates (Qiagen) following the respective manufacturer's recommendations. After that, the purified PCR products were quantified using Nanodrop 2000 (Thermo Scientific) and pooled in an approximately equimolar way. This final pool was purified anew with AMPure beads using 0.4 (v/v) of beads and eluted in a 50-100 µl molecular grade water. The concentration of the amplicon pool was quantified with a Qubit instrument (Invitrogen). Approximately 2-4 µg were sent for sequencing with PacBio RSII (Pacific Biosciences) at the Swedish SciLife Lab in Uppsala, Sweden. Another batch of 800 ng was used for Oxford Nanopore library preparation following the manufacturer's protocol and recommendations for D2 sequencing (LSK-208; Oxford Nanopore Technologies; discontinued as of May 2017, with the R9.4 chemistry) or alternatively 1D² sequencing (LSK-308; with the most recent R9.5 chemistry as of May 2017). In brief, both protocols consist of end-repair, adapter ligation, and purification steps that take approximately two hours in the laboratory. Sequencing took place locally on a MinION instrument (Oxford Nanopore Technologies) operated with FLO-107 flowcells. We aimed for more than 2,000 sequences per sample and stopped the sequencing as soon as we achieved this goal, which took 2-8 hours depending on the pool size and amplicon length.

This article is protected by copyright. All rights reserved.

Sequence data processing

The first step after obtaining the Nanopore data is the 2D base calling, which was done with Albacore (v2.4; Oxford Nanopore Technologies). We observed that for a successful calculation of the required sequencing depth, usually 20% of the sequences are retained as good quality 2D reads, which is one of the limitations of the 1D² chemistry. Not all reads are complementary reads and currently the base caller can only base call ~50% as complementary (i.e. paired reads), while the other 50% remains unpaired. Unpaired reads have a higher error rate than paired reads and were therefore discarded in this study. These former as complementary identified sequences subsequently form pairs, resulting in 25% of the initial reads. Finally, ~5% did not pass the quality filtering step, so that as a rule of thumb, a total of 20% of the initial reads remain as high-quality paired reads for generation of the consensus reference sequences.

For the PacBio data, we only worked with the “reads of insert” (ROS) data in the next steps of the data processing. After these initial steps, all sequences from both sequencing platforms are processed in the same way. An initial quality-filtering step (USEARCH v8.1; Edgar, 2010) was performed. The maximum allowed error rate was set to 0.02 for PacBio sequences. After testing this quality filtering for the Nanopore data (which come pre-filtered at an error rate of 0.08), ranging from 0.04-0.08, however, we came to the conclusion that quality filtering had no beneficial effect on the final consensus quality (Supplemental Material S2). We thus removed it from the pipeline. Additionally, we filtered the sequences by length using Biopython (v1.65; Cock et al., 2009) to exclude too short and too long sequences as detected in the histograms. This helped to increase the quality of the subsequent alignment. Then all quality filtered and trimmed sequences were demultiplexed as FASTA files into individual samples according to their combined barcodes (Flexbar, v2.5, Dodt, Roehr, Ahmed, & Dieterich, 2012). Barcodes and adapters were removed in this step. All sequences from each individual sample were subsequently aligned using MAFFT (v7.397, Katoh & Standley, 2013) using the auto-alignment option. The aligned sequences were clustered in mothur

(v1.39, Schloss et al., 2009) using the Opticlust algorithm, and the consensus sequences for each operational taxonomic unit were built using a custom-made Perl script (Consension) available at <http://microbiology.se/software/consension/>. The optimal OTU clustering threshold for Nanopore data was determined to be 0.07 for shorter amplicons (rDNA and IGS PCR) and 0.08 for the long TR PCR (Supplemental Material S3). To counter spurious OTUs we determined a dynamic OTU size cut-off that is provided to Consension. It was calculated as:

$$K = [\text{number of sample reads}] \cdot [\text{error rate}] / [\text{length correction}] \quad (1)$$

The length correction is an integer and defined as amplicon length [kb] divided by 5 kb of the expected amplicon length. K_{\min} (the minimum number of sequences a OTU can hold) was set to 3 for PacBio and 5 for Nanopore sequences. The consensus sequences were finally compared by inspecting the alignment visually in SeaView (v.4.7; Gouy, Guindon, & Gascuel, 2009) and by calculating sequence similarities with local BLAST searches (nucleotide BLAST+, v2.2.28). Visualization of the BLAST-based TR results matching all other sequences (rDNA, IGS, and ITS) was done with BRIG (v 0.95, Alikhan, Petty, Zakour, & Beatson, 2011). In the few cases where we obtained more than 1 OTU after consensus generation, we only used the most abundant OTU for subsequent similarity comparisons. In addition, we evaluated the effect of polishing the Nanopore consensus sequences by mapping the FASTQ files to the consensus sequences with Racon (v.1.3; <https://github.com/isovic/racon>). Finally, we examined the variation across all ribosomal regions on PacBio generated TR sequences from 13 *Inocybe* (Agaricomycetes) species to provide a first look at the explanatory power of full-length rDNA sequencing. For this analysis we first annotated all regions and subsequently extracted them (excluding introns) to evaluate the pairwise similarity of each region by inferring the respective distance matrices.

Results

All primer pairs worked successfully on our target herbarium specimens. The rDNA primers also worked both with samples from the early diverging lineages of Chytridiomycota (whole genome amplified DNA of infected single algal cells) and with cockroach tissue infected with *Nephridiophaga*. This confirmed the fungal specificity and the broad spectrum of the primers, which should, based on *in silico* analysis, cover all fungal phyla with the few within-phyla exceptions mentioned above (see Primer design, Supplemental Material 1). The performance of the rDNA and IGS PCR was robust and worked for > 93% of the herbarium extracts. We noted, however, that the longest amplicon (TR PCR) was amplified in only 50% of the otherwise successfully amplified herbarium specimens (29 of 58 samples), potentially due to DNA integrity issues in older herbarium specimens (see Discussion section and Larsson & Jacobsson, 2004). To demonstrate what could be gained in mycological efforts by sequencing the full ribosomal region (either by combining the two shorter rDNA and the IGS amplicons or by the TR amplicon alone), we analyzed the inter-species variation for 13 species of the genus *Inocybe* (Table 2).

An example of a full comparison between Sanger, PacBio, and Nanopore- generated sequences and all applied primer pairs for one of our herbarium DNA samples can be seen in Figure 2 for specimen GB-0158876 (*Inocybe melanopus* EL263-16).

As expected, PacBio-derived sequences had a high accuracy and matched high-quality Sanger sequences with 100% identity in 23 of 64 cases, while Nanopore sequences achieved this only in 3 of 41 cases. The discrepancy between PacBio and Sanger was in most cases related to mismatches in the distal ends of the ITS sequences, potentially reflecting quality issues of Sanger sequences (Supplemental Material S5). Similarly, Nanopore-derived sequences (1D² chemistry) had on average

only 0.15% mismatches to PacBio sequences, resulting in a consensus accuracy of 99.85%, identical to the median Sanger similarity to PacBio sequences (Table 3). In the alignment view, most of the Nanopore-based mismatches could be identified as indels in homopolymeric regions. The discontinued D2 chemistry in combination with its outdated base calling reached a consensus accuracy of 99.4%.

In a few samples (on average in 12% of the amplicons) we saw two or more distinct consensus sequences, both of which crossed the consensus threshold K . Often these low-abundance alternative consensus sequences differed significantly – by more than 100 bases – in length, and it is likely that these represented genomic variants of the ribosomal regions. In this study we disregarded these sequences, but they may be worth a second look in future studies as they may have implications for the outcome of phylogenetic analyses. We found that a lower clustering threshold (0.08 or below) is important in retaining special cases of these variants separate, e.g., identical sequences with one large intron (Supplemental Material 3).

Discussion

The disjunct distribution of ribosomal marker sequences across distinct databases is not only a mycological problem but one that pertains to most DNA-based studies targeting bacteria, algae, and protists (e.g. De Vargas et al., 2015; Wurzbacher, Nilsson, Rautio, & Peura, 2017). We are currently missing a lot of reference data in the databases we all use on a regular basis. In a time where biodiversity screening by high-throughput sequencing methods is becoming routine, these deficits in taxonomic and marker-related sampling of genetic material are developing into severe problems. Any strategy that may help closing these gaps in the future will be extremely valuable. The genetic markers of the ribosomal operon differ from each other, and across the fungal tree of life, in length

and level of conservation. As a consequence, different markers have been used to address research questions in different parts of the fungal tree of life and at different evolutionary time scales. The incompleteness and fragmentation of extant ribosomal data is clearly problematic, and our sampling of fungal ribosomal DNA sequences should be augmented with full-length reference sequences that span several regions suitable for everything from conservative (high-level) taxonomic classification to intraspecies assignment (see Table 2). The primers we present here extend the currently longest sequenced ribosomal fragments (Karst et al., 2018, Tedersoo et al., 2018b) and enable generation of data in an easy and straightforward way for the whole fungal ribosomal tandem repeat region, solving the problem of non-overlapping sets of ribosomal markers.

The primer pairs we introduce with the present paper worked not only for the Basidiomycota species examined, but also for Chytridiomycota and the distant genus *Nephridiophaga*. Indeed, according to the *in silico* primer design and evaluation, they should be suitable for the greater majority of fungal species, with the chief exception of some long-branching Zygomycota lineages, for which primer adaptations may be required. The primers RCA95m and Fun-rOP-F are fairly fungus-specific and will in many cases preferentially amplify fungi from mixed-DNA samples (e.g. our cockroach host tissue, see also Heeger et al., 2018 for an application of the rDNA amplicon on environmental samples). Users wanting to examine the performance of the primers on as-yet untested fungal taxa without having to do actual PCR runs can evaluate the primers through simulated PCR runs (e.g., ecoPCR; Ficetola et al. 2010) on authentic, long rDNA reads available in, e.g., SILVA.

The decreased success rate that we observed for the TR amplification in comparison to the rDNA and IGS PCR is probably linked to the DNA integrity in the sense that DNA is known to degrade (fragment) over time in herbarium settings (Larsson & Jacobsson, 2004). The TR-PCR approach, in contrast, requires long genomic fragments with intact ribosomal tandem repeats. In the present

Accepted Article

study we used herbarium specimens, which are usually moderately to fairly fragmented depending on age of collection and storage conditions, so the degree of fragmentation may have hindered the amplification of the TR fragment, while the shorter rDNA and IGS PCRs still worked. Although the accuracy of TR fragments (99.63 %) is slightly lower, it is still within the range of Sanger sequencing. A polishing step was not successful in increasing the overall quality (Table 3). The reason could lie in the underlying alignment algorithm, which may not have been optimized for long fragments with high individual error rates. In summary, the concerted sequencing of rDNA and IGS is a robust way to obtain the complete ribosomal operon and its adjacent regions in terms of sequence quality and amplification success (e.g., in cases of difficult DNA extractions and lower template DNA integrity).

Long-read sequencing offers the possibility to generate reference data for fungi, potentially other eukaryotes, and bacteria at high read quality. Given that our PacBio data are almost perfect (99.99% accuracy; Travers, Chin, Rank, Eid, & Turner, 2010), the average Nanopore consensus quality of 99.85% is already as good as the average Sanger quality of 99.78% (Nilsson et al., 2017) or for our data 99.73% (average of TR and rDNA results, Table 3). We argue, therefore, that the use of Nanopore sequencing is justified. Our results are in line with recently applied Nanopore-driven *CO1* barcoding for metazoa (Srivathsan et al., 2018). In particular, Nanopore sequencing offers a cheap method to generate full-length ribosomal data independently of amplicon length and at excellent quality. It does not require additional clean-up steps, as may be necessary for PacBio (See Supplemental Material S4 for a direct comparison between PacBio and Nanopore in terms of produced fragment lengths). We anticipate that Nanopore sequencing will prove to be a valuable tool for small laboratories, culture collections, herbaria, field work, and single cell workflows for the generation of high-throughput reference data, potentially also for mixed environmental samples (Karst et al., 2018; Calus, Ijaz, & Pinto 2018). The sequencing can be done in-house within a couple of days, significantly speeding up the generation of reference data and rendering it suitable in the

context of affordable high-throughput solutions (Srivathsan et al., 2018). Importantly, Nanopore sequencing does not rely on sending DNA for sequencing at large-scale facilities but is, rather, amenable to analysis on a modern desktop computer.

Similar to mock communities in environmental samples (Heeger et al., 2018), we consider it as absolutely crucial to spike in control DNA in Nanopore-driven studies. If no partial or complete reference data are available, we recommend doing complementary ITS Sanger/PacBio sequencing as an internal standard control. The generated data should be deposited to sequence repositories just as carefully and as richly annotated as in the case of Sanger sequences to avoid errors derived from the experimental procedure (cf. Nilsson et al., 2017). The sequencing error rate, determined through the use of known high-quality reference data, should always be included in the submission, included either in the sequence header or as additional data.

In conclusion, we hope that this study lays out the first steps for a new way of generating full-length reference data for fungi. This will enable mycologist to comprehensively fill up the taxonomic and marker-related gaps in the sequence databases in a straightforward and cost-efficient way. We found the adaptation to high-throughput data generation to be surprisingly easy and to require only an initial investment in barcoded primers, as well as good sample and data management. Our approach does place some demands on availability of bioinformatics expertise, testifying to the multidisciplinary nature of contemporary mycology.

Acknowledgements

We would like to thank Renate Radek for her help with providing material of *Nephridiophaga*; Keilor Rojas-Jimenez for whole genome amplifications; and Magnus Alm Rosenblad, Michael M. Monaghan, Elizabeth C. Bourne, and Felix Heeger for joint discussions on the implementation of long-read sequencing. The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure. CW and RHN gratefully acknowledge financial support from Stiftelsen Olle Engkvist Byggmästare, Stiftelsen Lars Hiertas Minne, Kapten Carl Stenholms Donationsfond, and Birgit och Birger Wålhströms Minnesfond. SVdW was supported by a IGB postdoc fellowship and the German Science Foundation (DFG). EL was supported by The Swedish Taxonomy Initiative, ArtDatabanken SLU Uppsala.

References

Alikhan, N. F., Petty, N. K., Zakour, N. L. B., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*(1), 402.

Balzi, E., Di Pietro, A., Goffeau, A., van Heerikhuizen, H., & Klotwijk, J. (1985). The RNA polymerase I initiation site and the external transcribed spacer of the fission yeast *Schizosaccharomyces pombe* ribosomal RNA genes. *Gene*, *39*(2), 165-172.

Benitez-Paez, A., & Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. *GigaScience*, *6*(7), 1-12.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1462), 1935-1943.

Burgaud, G., Coton, M., Jacques, N., Debaets, S., Maciel, N. O., Rosa, C. A., ... & Casaregola, S. (2016). *Yamadazyma barbieri* f.a. sp. nov., an ascomycetous anamorphic yeast isolated from a Mid-Atlantic Ridge hydrothermal site (– 2300 m) and marine coastal waters. *International Journal of Systematic and Evolutionary Microbiology*, *66*(9), 3600-3606.

Calus, S. T., Ijaz, U. Z., & Pinto, A. J. (2018). NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform. *bioRxiv*, 244517.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422-1423.

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.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... & Carmichael, M. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1261605.

Doty, M., Roehr, J. T., Ahmed, R., & Dieterich, C. (2012). FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*, 1(3), 895-905.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460-2461.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., ... & Pompanon, F. (2010). An in silico approach for the evaluation of DNA barcodes. *BMC genomics*, 11(1), 434.

Glöckner, F. O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., ... & Ludwig, W. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology*, 261, 169-176.

Gouy, M., Guindon, S., & Gascuel, O. (2009). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2), 221-224.

Hawksworth, D. L., & Luecking, R. (2017). Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiology Spectrum*, 5(4). doi: 10.1128/microbiolspec.FUNK-0052-2016.

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., ... & Monaghan, M. T. (2018). Long-read DNA metabarcoding of ribosomal rRNA in the analysis of fungi from aquatic environments. *Molecular Ecology Resources*, in press. <https://doi.org/10.1111/1755-0998.12937>

Hibbett, D. S. (1996). Phylogenetic evidence for horizontal transmission of group I introns in the nuclear ribosomal DNA of mushroom-forming fungi. *Molecular Biology and Evolution*, 13(7), 903-917.

Hibbett, D., Abarenkov, K., Kõljalg, U., Öpik, M., Chai, B., Cole, J., ... & Herr, J. R. (2016). Sequence-based classification and identification of Fungi. *Mycologia*, 108(6), 1049-1068.

This article is protected by copyright. All rights reserved.

Hillis, D.M. & Dixon, M.T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology*, 66(4), 411-453.

Ishida, S., Nozaki, D., Grossart, H. P., & Kagami, M. (2015). Novel basal, fungal lineages from freshwater phytoplankton and lake samples. *Environmental microbiology reports*, 7(3), 435-441.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., ... & Lumbsch, H. T. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113), 818.

Karst, S. M., Dueholm, M. S., McIlroy, S. J., Kirkegaard, R. H., Nielsen, P. H., & Albertsen, M. (2018). Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nature Biotechnology*, 36, 190–195.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772-780.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., ... & Douglas, B. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, 22(21), 5271-5277.

Larsson, E. & Jacobsson, S. (2004). Controversy over *Hygrophorus cossus* settled using ITS sequence data from 200 year-old type material. *Mycological Research*, 108(7), 781-786.

Lazarus, K. L., & James, T. Y. (2015). Surveying the biodiversity of the Cryptomycota using a targeted PCR approach. *Fungal Ecology*, 14, 62-70.

O'Donnell, K., Gueidan, C., Sink, S., Johnston, P. R., Crous, P. W., Glenn, A., ... & Van Der Lee, T. (2009). A two-locus DNA sequence database for typing plant and human pathogens within the *Fusarium oxysporum* species complex. *Fungal Genetics and Biology*, 46(12), 936-948.

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.

Monchy, S., Sancier, G., Jobard, M., Rasconi, S., Gerphagnon, M., Chabé, M., ... & Viscogliosi, E. (2011). Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. *Environmental Microbiology*, 13(6), 1433-1453.

Mitchell, J. I., & Zuccaro, A. (2006). Sequences, the environment and fungi. *Mycologist*, 20(2), 62-74.

Nilsson, R. H., Wurzbacher, C., Bahram, M., Coimbra, V. R., Larsson, E., Tedersoo, L., ... & Ryberg, M. K. (2016). Top 50 most wanted fungi. *MycoKeys*, 12, 29.

Nilsson, R. H., Sánchez-García, M., Ryberg, M. K., Abarenkov, K., Wurzbacher, C., & Kristiansson, E. (2017). Read quality-based trimming of the distal ends of public fungal DNA sequences is nowhere near satisfactory. *MycoKeys*, 26, 13.

Nilsson, R. H., Taylor, A. F., 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.

Radek, R., Wurzbacher, C., Gisder, S., Nilsson, R. H., Owerfeldt, A., Genersch, E., ... & Voigt, K. (2017). Morphologic and molecular data help adopting the insect-pathogenic nephridiophagids (Nephridiophagidae) among the early diverging fungal lineages, close to the Chytridiomycota. *MycoKeys*, 25, 31.

Rosenblad, M. A., Martín, M. P., Tedersoo, L., Ryberg, M. K., Larsson, E., Wurzbacher, ... & Nilsson, R. H. (2016). Detection of signal recognition particle (SRP) RNAs in the nuclear ribosomal internal transcribed spacer 1 (ITS1) of three lineages of ectomycorrhizal fungi (Agaricomycetes, Basidiomycota). *MycoKeys*, 13, 21.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Sahl, J. W. (2009). Introducing mothur: open-source, platform-independent, community-supported software

for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23), 7537-7541.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., ... & Miller, A. N. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, 109(16), 6241-6246.

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., ... & Hallam, S. J. (2016). High-resolution phylogenetic microbial community profiling. *The ISME Journal*, 10(8), 2020-2032.

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W. X., Bertrand, D., Ng, A. H. Q., ... & Meier, R. (2018). A MinION-based pipeline for fast and cost effective DNA barcoding. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12890.

Taylor, D. L., Hollingsworth, T. N., McFarland, J. W., Lennon, N. J., Nusbaum, C., & Ruess, R. W. (2014). A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecological Monographs*, 84(1), 3-20.

Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., ... & Bork, P. (2015). Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys*, 10, 1.

Tedersoo, L., Bahram, M., Puusepp, R., Nilsson, R. H., & James, T. Y. (2017). Novel soil-inhabiting clades fill gaps in the fungal tree of life. *Microbiome*, 5(1), 42.

Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., Bahram, M., Döring, M., Schigel, D., ... & Abarenkov, K. (2018a). High-level classification of the Fungi and a tool for evolutionary ecological analyses. *Fungal Diversity*, 90(1), 135-159.

Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2018b). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist*, 217(3), 1370-1385.

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.

White, T. J., Bruns, T., Lee, S. J. W. T., & Taylor, J. L. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications*, 18(1), 315-322.

White, M. M., James, T. Y., O'Donnell, K., Cafaro, M. J., Tanabe, Y., & Sugiyama, J. (2006). Phylogeny of the Zygomycota based on nuclear ribosomal sequence data. *Mycologia*, 98(6), 872-884.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12), 4576-4579.

Wurzbacher, C., Rösel, S., Rychła, A., & Grossart, H. P. (2014). Importance of saprotrophic freshwater fungi for pollen degradation. *PloS one*, 9(4), e94643.

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.

Wurzbacher, C., Nilsson, R. H., Rautio, M., & Peura, S. (2018). Poorly known microbial taxa dominate the microbiome of permafrost thaw ponds. *The ISME Journal*, 11, 1938–1941.

Data Accessibility Statement

The Sanger ITS sequences can be found in GenBank under the accession numbers MH310744-MH310802 (herbarium sequences) and MH356537-MH356545 (Chytridiomycota sequences). The Nanopore and PacBio reads were deposited in ENA under the project number PRJEB26696.

Author Contributions

CW and RHN designed the experiment; JBP, EK, and CW implemented the bioinformatics steps and performed the analyses; MK, SVW, and CW designed the experiment on single cells; EL and SS prepared the herbarium specimens; MK, SVW, CW, and EL performed the molecular work and prepared the sequencing libraries; and CW and RHN wrote the manuscript. All authors read and approved the final manuscript version.

Tables and Figures

Figure Legends

Figure 1. Schematic representation of the fungal ribosomal tandem repeat with two copies of the ribosomal operon and its transcribed and nontranscribed regions (precursor rRNA and IGS, respectively). Please note that the 5S is not always present (Balzi, Pietro, Goffeau, Heerikhuizen, & Klootwijk, 1985). The primers and the positions of the primer binding sites for the three employed amplicons (rDNA: 4-6 kb, IGS: 2-6 kb, and TR: 7-13 kb) are indicated.

Figure 2. Left panel: graphic view on the BLAST+ comparison of the ribosomal regions generated with the three amplicons and the Sanger reference sequence of the ITS region. The TR amplicon generated by PacBio sequencing was used as reference. Similarities are displayed for each amplicon, respectively. Right panel: photograph of in situ basidiomata of herbarium specimen GB-0158876 (*Inocybe melanopus*).

Table 1. Amplification overview for the sample types

technology	herbarium specimens (n = 66)	Chytridiomycota cells (WGA) (n = 9)	<i>Nephridiophaga</i> (host tissue) (n = 2)
PacBio	rDNA, IGS, TR	n.t.	rDNA
Nanopore	TR, rDNA	rDNA	n.t.

n.t. (not tested); rDNA refers to the amplicon produced by NS1short/RCA95m; IGS refers to the amplicon produced by RCA95rc/NS1rc; TR refers to the amplicon produced by Fun-rOP-F/Fun-rOP-T; WGA (whole genome amplification).

Table 2. Sequence variation analysis of the ribosomal tandem repeat regions for 13 *Inocybe* species

ribosomal region	average pairwise identity (%)	minimum pairwise identity (%)	maximum pairwise identity (%)	consensus length (bases)
5'-ETS	54.2	43.2	67.5	500
SSU	98.4	97.5	99.3	1803
ITS	53.3	42.6	66.4	401
5.8S	98.2	95.6	100	159
LSU	95.8	93.6	97.9	3386
3'-ETS†	49.6	29.8	68.3	439
5S	92.9	79.5	100	121
IGS	24.5	3.5	35.1	1870
all regions	63.3	54.1	73.8	8720

† annotation of the 3'-ETS region is uncertain and may falsely include a second conservative region of unknown identity.

Table 3. Similarities of sequences compared to Sanger or PacBio sequences

	PacBio data (vs. Sanger ITS)		Nanopore data (vs. Sanger ITS)		Nanopore data (vs. PB amplicons)			difference (Nanopore vs. PacBio)		D2-chemistry (vs. Sanger)
	rDNA	TR	rDNA	TR	rDNA	TR	TR _{Polished}	rDNA	TR	rDNA
average	99.68	99.78	99.20	99.52	99.85	99.63	99.63	0.15	0.37	99.4
median	99.86	99.85	99.37	99.59	99.85	99.71	99.71	0.15	0.29	99.5
st.dev.	0.45	0.19	0.60	0.38	0.05	0.20	0.20	0.05	0.20	0.40
min.	98.31	99.37	98.09	98.5	99.7	99.16	99.19	0.30	0.84	98.73
max.	100	100	100	100	99.91	99.83	99.85	0.09	0.17	99.85
n	45	19	17	24	17	18	18	n.a.	n.a.	9

rDNA refers to the amplicon produced by NS1short/RCA95m; TR refers to the amplicon produced by Fun-rOP-F/Fun-rOP-T; TR_{Polished} refers to the TR data, which were processed with Racon v 1.3.; st.dev. (standard deviation); min. (minimum); max. (maximum).

