Long- and short-read metabarcoding technologies reveal similar spatio-temporal structures in fungal communities

Brendan Furneaux[1,*], Mohammad Bahram[2,3], Anna Rosling[4], Nourou S. Yorou[5], Martin Ryberg[1]

[1]Program in Systematic Biology, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

[2]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[3]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[4]Program in Evolutionary Biology, Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

[5]Research Unit in Tropical Mycology and Plant-Fungi Interactions, LEB, University of Parakou, Parakou, Benin

[*]Corresponding author, brendan.furneaux@ebc.uu.se

Abstract

Fungi form diverse communities and play essential roles in many terrestrial ecosystems, yet there are methodological challenges in taxonomic and phylogenetic placement of fungi from environmental sequences. To address such challenges we investigated spatio-temporal structure of a fungal community using soil metabarcoding with four different sequencing strategies: short amplicon sequencing of the ITS2 region (300–400 bp) with Illumina MiSeq, Ion Torrent Ion S5, and PacBio RS II, as well as long amplicon sequencing of the full ITS and partial LSU regions (1200–1600 bp) with PacBio RS II. Resulting community structure and diversity depended more on statistical method than sequencing technology. The use of long-amplicon sequencing enables construction of a phylogenetic tree from metabarcoding reads, which facilitates taxonomic identification of sequences. However, long reads present issues for denoising algorithms in diverse communities. We present a solution that splits the reads into shorter homologous regions prior to denoising, and then reconstructs the full denoised reads. In the choice between short and long amplicons, we suggest a hybrid approach using short amplicons for sampling breadth and depth, and long amplicons to characterize the local species pool for improved identification and phylogenetic analyses.

1 Introduction

Fungi are key drivers of nutrient cycling in terrestrial ecosystems. One important guild of fungi form ectomycorrhizas (ECM), a symbiosis between fungi and plants in which fungal hyphae enclose the plant's fine root tips. The fungi provide nutrients and protection from pathogens in exchange for carbon from the plant (Smith & Read, 2010). Approximately 8% of described fungal species are thought to take part in ECM symbiosis (Ainsworth, 2008; Rinaldi et al., 2008). Although only about 2% of land plant species form ECM, these include

ecologically and economically important stand-forming trees belonging to both temperate and boreal groups such as Pinaceae and Fagaceae, and tropical groups such as Dipterocarpaceae, *Uapaca* (Phyllanthaceae) and Fabaceae tr. Amherstieae (Brundrett, 2017).

Although ECM fungi form many well-known mushrooms (e.g., *Amanita*, *Cantharellus*, *Boletus*), some instead produce inconspicuous (e.g., *Tomentella*) or no (e.g., *Cenococcum*) fruit bodies. Even when fruitbodies are large, they are ephemeral, so study of ECM communities is facilitated by looking at vegetative structures (Horton & Bruns, 2001). Unlike many saprotrophic fungi which grow easily in axenic culture, ECM fungi are usually difficult to culture, so DNA barcoding is increasingly used to investigate vegetative structures in the field. The advent of high-throughput sequencing (HTS) has facilitated such studies by providing enough sequencing depth for metabarcoding of bulk environmental samples such as soils (Lindahl et al., 2013).

As additional techniques and methods are developed for HTS, there is an increasing array of choices for researchers investigating fungal communities. Fungal metabarcoding studies using short-read HTS technologies have targeted the ITS1 or ITS2 regions, which provide sufficient resolution to distinguish fungal species in many groups, and which are usually short enough for HTS (Lindahl et al., 2013; Schoch et al., 2012). The resulting sequencing reads are clustered by sequence similarity to form operational taxonomic units (OTUs), which are then used as the units for further community analysis (Lindahl et al., 2013). If taxonomic identification is desired in order to put OTUs in a wider context and associate functional information, it has usually been performed by database searches using BLAST (Altschul et al., 1990; Lindahl et al., 2013). However, this approach comes with some potential weaknesses.

While ITS1 and ITS2 often have suitable variation to distinguish species, they cannot be reliably aligned over the fungal kingdom (Lindahl et al., 2013; Tedersoo, Tooming-Klunderud, et al., 2018). Additionally, the wide range of length variation of these regions may introduce

bias in recovery of different taxa. Further bias is introduced by variation in the 5.8S region which separates the two ITS regions, as well as in the 5' end of LSU, which makes it difficult to design primers that are suitable for all fungi (Tedersoo et al., 2015).

Distance-based clustering conflates intra-species variation and sequencing error, and results are dataset-specific. In contrast, more recent denoising methods such as DADA2 (Callahan et al., 2017), Deblur (Amir et al., 2017), and UNOISE2 (Edgar, 2016b) utilize read quality information to control for sequencing error while preserving intra-species variation. The resulting units are known as amplicon sequence variants (ASVs) or exact sequence variants (ESVs), as they should represent true amplicon sequences from the sample. Unlike cluster-based OTUs, ASVs can capture variation of as little as one base pair, and are less dataset specific (Callahan et al., 2017).

Assignment of taxonomic identities using BLAST requires *a priori* choice of thresholds for different taxonomic ranks. Several algorithms specifically designed for taxonomic assignment have been published which use information about variability within different taxa in the reference database to assign unknown sequences, along with confidence estimates for these assignments (e.g., Edgar, 2016a; Murali et al., 2018a; Wang et al., 2007). In addition, methods have been published which integrate predictions from multiple algorithms to increase the reliability of assignments (Gdanetz et al., 2017; Somervuo et al., 2016).

Recent long-read HTS technologies such as Pacific Biosciences Single Molecule Real Time sequencing (PacBio) enable sequencing longer amplicons which include both the ITS regions and the flanking, more highly conserved SSU and/or LSU regions (Tedersoo, Tooming-Klunderud, et al., 2018). This can potentially improve taxonomic placement of sequences that lack close database matches and allow the alignment of metabarcoding reads for subsequent phylogenetic analysis. Information from phylogenetic trees produced from long-amplicon metabarcoding has the potential to both improve taxonomic assignment and provide al-

ternative measures of community alpha and beta diversity. Because OTU clustering may both "clump" different species into a single OTU, and "split" a single species into multiple OTUs (Ryberg, 2015), diversity measures based on counting species within a community or shared species between two communities may give different results depending on the clustering threshold. In contrast, phylogenetic community distance measures (Wong et al., 2016) are relatively insensitive to species/OTU delimitation, but require a phylogenetic tree. Phylogenetic placement algorithms have been developed to place short amplicon reads onto a reference tree (Berger et al., 2011; Matsen et al., 2010), but are not easy to apply to ITS sequences because they require that the query sequences be aligned to a reference alignment. Additionally, methods exist to place OTUs on a simplified tree based on taxonomic assignments (Tedersoo, Sánchez-Ramírez, et al., 2018). However, long amplicon sequencing allows the inclusion of alignable regions for construction of more fully resolved phylogenetic trees directly from metabarcoding reads. However, long-read technologies are currently more expensive per read compared to short-read sequencing, and so their use entails a trade-off with sequencing depth and/or sample number (Kennedy et al., 2018).

Here we investigated the effects of different sequencing strategies and post-analysis on biological conclusions using measurement of the spatiotemporal turnover rate of the fungal community in an ECM-dominated Soudanian woodland in Benin by metabarcoding of bulk soil, sampled at narrow intervals, over two years. We compare three different sequencing platforms (PacBio RS II, Illumina MiSeq, Ion Torrent Ion S5), long and short amplicons, three different taxonomic assignment algorithms (RDP classifier, SINTAX, IDTAXA) and reference databases (Unite, Warcup, RDP), and two different community distance measures (Bray-Curtis vs. weighted UNIFRAC). We also present new algorithms for dividing the rDNA into regions, combining denoising results from multiple regions, and incorporating phylogenetic information into taxonomic assignments.

2 Materials and Methods

2.1 Sampling

Sampling was conducted at two sites (Ang: N 9.75456° W 2.14064°; Gan: N 9.75678° W 2.31058 °) approximately 30 km apart in the *Forêt Classée de l'Ouémé Supérieur* (Upper Ouémé Forest Reserve) in central Benin. Both sites were located in woodlands dominated by the ECM host tree *Isoberlinia doka* (Caesalpinioideae). At each site, 25 soil samples were collected along a linear transect at intervals of 1 m in May 2015. One third of the sample locations (3 m spacing) were resampled one year later in June 2016. For each sample, any coarse organic debris was removed from the soil surface and a sample of approximately 5cm×5cm×5cm was extracted with a sterilized knife blade. Each sample was sealed in a plastic zipper bag and homogenized by shaking and manually breaking apart soil aggregations. Approximately 250 mg total of soil was collected from two locations in the homogenized soil sample and placed into a separate 2.0 mL microtube containing 750 mL of lysis buffer and lysis beads (Xpedition^{TM} Soil/Fecal DNA miniprep, Zymo Research Corporation, Irvine, California, USA) and lysed in the field using a handheld bead-beater (TerraLyser^{TM}; Zymo Research Corporation).

An additional sample was collected at every sampling location (1-m spacing) in 2016 using LifeGuard^{TM} Soil Preservation Solution (MO BIO, Carlsbad, CA; USA) for preservation, without field lysis. Sequencing results for these samples differed significantly (PERMANOVA with 9999 permutations, $p < 0.0001$, $R^2 = 0.06$) from samples preserved using the Xpedition^{TM} lysis buffer (Figures S1, S2, and S3); as such these samples were excluded from our spatial analyses. However, reads from these samples were included in the full bioinformatics workflow, including ASV calling, OTU clustering, and phylogenetic trees.

2.2 DNA extraction, amplification, and sequencing

After field lysis, DNA was extracted using the Xpedition^{TM} Soil/Fecal Prep kit (see above). Samples preserved using LifeGuard were first centrifuged at 10000 g for 1 minute, after which the supernatant was removed and DNA was extracted from the remaining soil using the Soil/Fecal Prep kit as for the other samples. DNA was quantified using fluorometrically using Quant-iT^{TM} PicoGreen^{TM} dsDNA (Thermo Fisher Scientific, Waltham, MA, USA) fluorescent indicator dye on a Infinite F200 plate spectrofluorometer (Tecan Trading AG, Männedorf, Switzerland) according to the manufacturer's protocol.

Two different fragments of the nuclear rDNA were amplified (Figure S4). The short amplicon (approximately 300 bp) targeted the full ITS2 region as well as parts of the flanking 5.8S and large subunit (LSU) rDNA, using gITS7 (Ihrmark et al., 2012) as the forward primer and a mix of ITS4 (White et al., 1990) and ITS4a (Urbina et al., 2016) as the reverse primer (hereafter, ITS4m). The long amplicon (approximately 1500 bp) targeted the full ITS region including the 5.8S rDNA and approximately 950 bp at the 5' end of the LSU, including the first three variable regions (Figure S4), using ITS1 (White et al., 1990) as the forward primer and LR5 (Vilgalys & Hester, 1990) as the reverse primer. Each PCR run also included a blank sample and a positive control consisting of freshly extracted DNA from a commercially purchased fruitbody of *Agaricus bisporus*.

The gITS7 primers for the short amplicon were indexed for multiplexing (Supplementary File 1). Amplification was performed by polymerase chain reaction (PCR) in 20µl reactions containing 200 µM dNTP mix, 250 µM indexed gITS7 primer, 150µM ITS4m, 2mM $MgCl_2$, 0.1 U *Taq* polymerase (Dream *Taq*, Thermo Fisher Scientific, Waltham, MA, USA) and 3–7 ng purified DNA in Dream *Taq* buffer. The reaction conditions were 10 min at 95°, followed by 35 cycles of 60 s at 95°, 45 s at 56°, and 50 s at 72°, and finally 3 min at 72°. Each reaction was conducted in three technical replicates to reduce the effect of PCR

stochasticity, which were pooled after amplification.

Both primers for the long amplicon were indexed for multiplexing (Supplementary File 2). PCR was performed as for the short amplicons, but with 500 µM of each of the two primers. Reaction conditions were 10 min at 95°, 30 cycles of 45 s at 95°, 45 s at 59°, and 90 s at 72°, and finally 10 min at 72°. Each reaction was performed in three technical replicates as for short amplicons.

Amplicons were purified using SPRI beads (Vesterinen et al., 2016) and quantified fluorometrically as above. An aliquot of 100 ng of DNA from each sample (or the total PCR product if less than 100 ng) was pooled into two libraries each for long and short amplicons. Each library was sequenced using Single Molecule Real Time (SMRT) sequencing on a Pacific Biosciences (PacBio) RS II sequencer at the Uppsala Genome Center (UGC; Uppsala Genome Center, Science for Life Laboratory, Dept. of Immunology, Genetics and Pathology, Uppsala University, BMC, Box 815, SE-752 37 UPPSALA, Sweden). Short amplicon libraries were sequenced on two SMRT cells each, while long amplicon libraries were sequenced on four SMRT cells each.

Additionally, the short amplicon libraries were combined and sequenced using an Ion S5 (Ion Torrent) sequencer using one 520 chip at UGC, and a MiSeq (Illumina Inc.) sequencer using v3 chemistry with a paired-end read length of 300 bp at the SNP&SEQ Technology Platform (Dept. of Medical Sciences, Uppsala University, BMC, Box 1432, SE-751 44 UPPSALA, Sweden). The Illumina library was pooled with samples for another project, with half of the reads from one lane devoted to the current study.

2.3 Bioinformatics

Circular consensus sequence (CCS) basecalls for PacBio sequences were made using `ccs` version 3.4 (Pacific Biosciences, 2016, July 13/2019) using the default settings. The resulting

sequences, as well as the paired-end Illumina sequences, were demultiplexed and sequencing primers were removed using `cutadapt` version 2.8 (Martin, 2011). Sequencing primers were similarly removed from the Ion Torrent sequences, but interference between the tagged gITS7 primers and the Ion XPress tags used in library prep made full demultiplexing of the Ion Torrent sequences impossible, and these reads were thus only analyzed as a pool. For Ion Torrent and PacBio, reads were discarded if they did not have the appropriate primers on both ends. Reads were searched in both directions, and reads where the primers were found in the reverse direction were reverse complemented before further analysis. For Illumina sequences, read pairs were only retained when PCR primers were detected at the 5' ends of both the forward and reverse read. Primers were also searched for and removed on the 3' ends of the reads, in case of readthrough with short amplicons. Read pairs where the primers were found in reverse orientation were kept in separate files, but were retained in their original orientation until after denoising.

2.3.1 Denoising

We attempted to denoise both long and short PacBio amplicons using DADA2 according to the steps outlined in the supplementary information in Callahan et al. (2019). However, only 38 amplicon sequence variants (ASVs) were obtained for the long amplicons, representing 12% of the trimmed reads. We conclude that this poor performance was due to a combination of long read length and low sequencing depth relative to community diversity. The DADA2 algorithm requires that the seed sequence of each ASV be represented by at least two error-free reads (Callahan et al., 2016). If sequencing errors are uniformly distributed, then the probability that a given read will be error-free is $(1-\epsilon)^L$, where ϵ is the sequencing error rate and L is the read length in base pairs. Then the number of reads of a given sequence that would be required to obtain two error-free reads in expectation is $2/(1-\epsilon)^L$. For the combination of long reads (median $L = 1509$ bp after trimming) and moderate error

rate (mean $\epsilon = 0.0066$ based on ccs quality scores) for the long amplicon in this study, the expected number of reads required to achieve two error-free reads is 43,720. Given the high diversity relative to sequencing depth in this study (485 ASVs based on PacBio short amplicons, 104,305 trimmed long amplicon reads), this requirement could not have been met for the long amplicons except by the most abundant sequences. In comparison, the equivalent requirement for the short amplicon ($L = 265$ bp, $\epsilon = 0.0022$) is only 3.6 reads. We therefore developed a new workflow to assemble ASVs from the long amplicons, as follows:

Raw reads were divided into shorter regions by matching to covariance models (CM), which are similar to stochastic hidden markov models (HMM), but account for both nucleotide sequence and RNA secondary structure (Eddy & Durbin, 1994). First, the 5.8S rDNA was located in each read by searching for Rfam model RF0002 (Kalvari et al., 2018) using `cmsearch` from Infernal 1.1.2 (Nawrocki & Eddy, 2013), and all bases before the 5.8S were assigned to ITS1. No attempt was made to remove the approximately 12 bp fragment of the SSU from the 5' end of ITS1 in the long amplicons; it was too short to be reliably detected by a CM or the HMMs employed by ITSx (Bengtsson-Palme et al., 2013). A reference alignment including conserved RNA base pairing between and within the 5.8S and relevant portions of LSU was generated from the fungal 28S RNA seed alignment from the Ribosomal Data Project (RDP) release 11.5 (Cole et al., 2014; Glöckner et al., 2017) by truncating after the LR5 primer site and using the reference line to annotate the variable regions *sensu* Michot et al. (1984) and Raué et al. (1988). A CM was generated from the alignment using `cmbuild` from Infernal. The fragment of each read beginning with the 5.8S rDNA was then aligned to the CM using `cmalign` from Infernal. The annotation line in the CM alignment for each read was then used to split the reads into alternating more-conserved and less-conserved regions as shown in Figure S4, where LSU1-4 represent the conserved regions of LSU flanking the variable D1-3 regions (Michot et al., 1984). For short amplicons, only (partial) 5.8S, ITS2, and (partial) LSU1 were extracted. Code to extract the regions, including annotated seed

alignments and CMs, is available in the new R package `LSUx`.

Each of the extracted regions was independently filtered for length (Supplementary Table S1) and a maximum of three expected errors. Sequences were then dereplicated and denoised into amplicon sequencing variants (ASVs) using DADA2 version 1.12.1 (Callahan et al., 2016; Callahan et al., 2019). The error model for DADA2 denoising was fit using the 5.8S region for long amplicons, and using the entire read for short amplicons. Independent error models were fit for each sequencing run (i.e., long *vs.* short amplicons, different sequencing technologies). For PacBio libraries, DADA2 was run with complete pooling and a band size of 16. For Ion Torrent libraries, pseudo-pooling and a band size of 32 were used, and the homopolymer gap penalty was set to -1, as recommended by the DADA2 FAQ (https://benjjneb.github.io/dada2/faq.html). Chimeras within each region were removed using `removeBimeraDenovoTable` from DADA2.

For each ITS2 ASV from the long amplicon data set, the denoised sequences for the other regions corresponding to the same sequencing reads were concatenated to form a set of full-length reads. For reads which were not assigned a denoised sequence for each region, the raw read for the region was used instead. Because ITS2 is the most variable of the amplified regions (Figure S5), reads with identical ITS2 regions are expected to have highly similar sequences in the other regions, unless the amplicon was chimeric. The concatenated ASVs representing each long read were aligned in R using the `DECIPHER` package (Wright, 2015). Outlier sequences, as determined by mean pairwise distance from the rest of the alignment, were removed from each alignment using the `odseq` package (Jehl et al., 2015), using the default threshold of 0.025. The consensus of the remaining aligned sequences was assigned as the full-length ASV sequence. Full-length ASV sequences with more than three ambiguous bases (i.e., no nucleotide >50% at a given position) were removed. The count and sample distribution of reads assigned to each full-length ASV were calculated in order to form a sample × ASV community matrix. A similar process was used to generate a consensus ITS

(ITS1–5.8S–ITS2) and LSU (LSU1–D1–LSU2–D2–LSU3–D3–LSU4) sequence for each ASV. The process of assigning consensus full-length ASVs was carried out using the new `tzara` package for R.

Because the Illumina dataset consisted of paired-end reads, regions were not extracted prior to denoising. ASVs were instead generated according to a standard workflow for DADA2. Demultiplexed reads were truncated after the first base with quality score <= 10, and then reads with more than 3 expected errors in either read were discarded. Forward and reverse reads were denoised using DADA2 version 1.12.1 (Callahan et al., 2016) using separate error models and pseudo-pooling, and then forward and reverse reads were merged. The ITS2 region was extracted from the ASVs using `LSUx` for comparison to the other technologies.

2.3.2 Taxonomy assignment

Taxonomic annotations of the Ribosomal Data Project's LSU fungal training set (RDP) version 11.5 (Cole et al., 2014) and Warcup ITS training set (Deshpande et al., 2016) were mapped to the taxonomic classification system used in the Unite database version 8 (Nilsson et al., 2019). In particular, the classification for fungi was according to Tedersoo, Sánchez-Ramírez, et al. (2018), and for non-fungal eukaryotes was according to the proposed system of Tedersoo (2017a) as described in (Tedersoo, 2017b). Although the latter system is not formally published, it is consistent with the annotations for non-fungal eukaryotes in the Unite database. Additionally, it is a system with both purportedly monophyletic taxa and a uniform set of taxon ranks, which make it more appropriate for sequence-based taxonomic assignment algorithms than more accepted classification systems such as that of the International Society of Protistologists (Adl et al., 2019), which utilizes hierarchical nameless ranks.

Taxonomic assignment was performed to genus level separately on the ITS region using Unite

and Warcup and on the LSU region using RDP, respectively, as taxonomic references. For each region/reference combination, taxonomy was assigned using three algorithms: the RDP Naïve Bayesian Classifier (RDPC, Wang et al., 2007) as implemented in DADA2; SINTAX (Edgar, 2016a) as implemented in VSEARCH v2.9.1 (Rognes et al., 2016); and IDTAXA (Murali et al., 2018b). Each full-length ASV was thus given up to nine preliminary taxonomic assignments (three references × three algorithms). ASVs from the short-amplicon datasets for which no matching long-amplicon ASV could be reconstructed were taxonomically assigned using Unite and Warcup on the full length of the short amplicon.

Sequences were assigned as ECM based on taxonomic assignments using the FUNGuild database (Nguyen et al., 2016) via the R package `FUNGuildR` (https://github.com/brendanf/FUNGuildR). All taxa which included "Ectomycorrhiza" in the guild assignment at any level of confidence were included.

2.3.3 Clustering

For comparison with clustering-based methods, ASVs were clustered into operational taxonomic units (OTUs) at 97% similarity using VSEARCH v2.9.1 (Rognes et al., 2016).

2.3.4 Alignment and phylogenetic inference

Full length long amplicon ASVs were aligned using DECIPHER (Wright, 2015) with up to 10 iterations of progressive alignment and conserved RNA secondary structure calculation and 10 refinement iterations. This alignment was truncated at a position after the D3 region corresponding to base 907 of the *Saccharomyces cerevisiae* S288C reference sequence for LSU, because several sequences had introns after this position, as also observed in several fungal species by Holst-Jensen et al. (1999).

An ML tree was produced using RAxML version 8.2.12 (Stamatakis, 2014) using the

GTR+GAMMA model and rapid bootstrapping with the MRE_IGN stopping criterion. The tree was rooted outside the kingdom Fungi by using the most abundant ASV which was confidently assigned to a non-fungal kingdom by all 6 applicable taxonomic assignment methods. Assignments based on Warcup were not used at this step because non-Fungi are not included in the dataset. The kingdom Fungi was identified as the minimal clade containing all ASVs which were confidently identified (consensus of at least 6 of 9 assignments) to a fungal phylum. ASVs falling outside this clade were not included in downstream fungal community analysis.

Taxonomic assignments of ASVs from the long amplicon dataset were refined using the phylogenetic tree (Figure S6). A taxon at a particular rank was assigned to a node and all its descendants if that taxon was consistent with the reference-based taxonomic assignments for each of the descendants. A taxon assignment was considered to be consistent if at least one algorithm assigned that taxon at greater than 50% confidence, or if no algorithm successfully classified the sequence at greater than 50% confidence. The result of this process is twofold. First, it gives a taxonomic assignment to ASVs which were previously unassigned if they are nested within a clade which is consistently given an assignment. Second, it clarifies the assignment of ASVs where different algorithms had resulted in different assignments, but only one of these is consistent with the assignments of other ASVs in the same clade. This refinement algorithm is referred to as "PHYLOTAX".

ASVs from the short amplicon datasets were refined using only the final strict consensus step, i.e., an assignment at a given rank was accepted if there was no conflict between the different assignment algorithms at greater than 50% confidence. This refinement method is referred to as "Consensus". Additionally, a hybrid method, was applied to the short amplicon datasets, in which assignments from PHYLOTAX were used for ASVs which could be linked by an identical ITS2 region to a long amplicon, and assignments from Consensus were used for the remaining ASVs.

2.4 Community comparison

The fungal communities recovered by the three sequencing strategies that were successfully demultiplexed (Illumina, PacBio Short, PacBio Long) were compared by PERMANOVA. In order to detect bias at larger taxonomic scales, ASVs were clustered according to the assigned taxonomic class. Only samples where all three strategies yielded at least 100 fungal reads (34 samples), and classes which represented at least 1% of reads in at least one sample (14 classes), were included. PERMANOVA included three terms: an indicator for soil sample, comprising all spatiotemporal effects; amplicon length (long vs. short); and sequencing technology (Illumina MiSeq vs. PacBio RS II). The marginal significance of each term for explaining variation in the Bray-Curtis community dissimilarity matrix was performed using the `adonis2` function in the R package `vegan` (Oksanen et al., 2019), with 9999 permutations. Partial Principal Coordinates Analysis (PPCoA) was applied to the same dissimilarity matrix using the `capscale` function in `vegan` (Oksanen et al., 2019). Spatiotemporal effects were partialled out in order to visualize effects due to sequencing technology and amplicon length.

A similar analysis was also applied to only fungi classified as ECM, clustered at the family level.

2.5 Spatiotemporal analysis

Turnover scale is the distance at which two communities can be considered to be independent samples of the local species pool. Knowledge of turnover scale is import when planning studies of local diversity and its environmental correlates. It varies between different ecosystems and taxonomic groups. Turnover scale is often measured by the range at which a Mantel correlogram indicates significant autocorrelation, or by fitting a function to an empirical distance-decay curve of community dissimilarity vs. distance (Legendre & Legendre, 2012).

Ecological community dissimilarity matrices were calculated using the ASV/OTU based Bray-Curtis metric (both long and short amplicons) and the phylogenetically based weighted UNIFRAC method (only long amplicons) in `phyloseq` version 1.26.0. Each of these distance matrices was used to calculate a Mantel correlogram for distances of 0–12 m. Separate correlograms were drawn for samples taken during the same year, and samples separated in time by one year, in order to assess the degree to which the soil community changes over the course of one year.

Additionally, empirical distance-decay curves were generated by plotting mean community dissimilarity as a function of spatial distance, and fit to an exponential model of the form given by Legendre and Legendre (2012) using the `nls` function in R. Points in the empirical distance-decay curve were weighted by the number of comparisons within the distance class and the inverse of the distance for the purposes of model fitting. For datasets where the Mantel correlogram indicated spatial correlation between samples taken in separate years, the model was re-fit with an additional term to represent temporal correlation:

$$D = C_0 + C_1 \left[1 - \exp\left(-3 \left(\frac{d}{a_d} + \frac{t}{a_t} \right) \right) \right]$$

where D, d, and t represent the community dissimilarity, spatial distance, and time lag between samples, respectively, and the parameters are C_0, the community dissimilarity from replicate samples ("nugget"); $C_0 + C_1$, the community dissimilarity at long distances ("sill"); a_d the spatial range at which the community dissimilarity has moved 95% of the way from "nugget" to "sill"; and a_t, the equivalent temporal range. The 95% confidence intervals were calculated for the spatial and temporal range parameters by profiling using the `MASS` package in R.

3 Results

DNA concentrations after extraction and PCR, as well as sequencing reads for PacBio and Illumina, are shown per sample in Figure S7. Samples from Ang in 2015 yielded low quantities of DNA, poor PCR performance, and ultimately very few sequencing reads, especially in the long amplicon library, where only one sample produced more than 100 reads. Consequently, Ang samples were excluded from spatial analysis, although they were retained for denoising, phylogenetic reconstruction, and taxonomic assignment.

The number of sequencing reads and ASVs at each stage in the bioinformatics pipeline are shown in Table S2. Sequencing with PacBio RS II yielded more than twice as many raw reads for long amplicons as for short amplicons, with approximately 125 thousand and 50 thousand reads, respectively. Ion Torrent Ion S5 and Illumina MiSeq yielded substantially more reads, with 20.7 million and 10.8 million, respectively. Demultiplexing, primer trimming, and quality filtering reduced these totals by 64% for PacBio long reads, but only by 21% for PacBio short reads, resulting in a similar number of filtered reads for the two strategies. Losses in demultiplexing, trimming, and quality filtering were intermediate for Ion Torrent and Illumina, with 41% and 28% loss, respectively. In contrast, extraction of only the ITS2 region before quality filtering resulted in the loss of 29% of trimmed long amplicon PacBio reads, 21% of trimmed short amplicon PacBio reads, and 34% of trimmed Ion Torrent reads. This represented greater loss of PacBio short reads, but less loss of PacBio long reads and Ion Torrent reads.

Almost all of the short amplicons from all three technologies were between 240 and 375 bp long (Figure S8a). Although the length profile of the three sequencing runs were similar, Illumina MiSeq had the largest fraction of reads near the top of the range, followed by Ion Torrent Ion S5 and PacBio RS II (Figure S8b). The difference in length distributions was statistically significant due to the large sample size (Kruskal-Wallis statistic $= 8.5735947 \times 10^4$, $p <$

405 2.2×10^{-16}), but the difference between means was fairly small, with mean amplicon lengths
406 of 276, 281, and 286 bp for PacBio, Ion Torrent, and Illumina, respectively. In contrast, the
407 length of the long amplicon reads varied widely, from 696 to 1638 bp, with a mean of 1431
408 bp.

409 The length distribution of the different regions extracted from the long amplicon are shown
410 in Figure S9. ITS1 showed the greatest length variability (mean $\pm$ standard deviation: 193
411 $\pm$ 55 bp), followed by ITS2 (184 $\pm$ 41 bp) and the variable regions in LSU (D2: 227 $\pm$ 36 bp;
412 D3: 108 $\pm$ 10 bp; D1: 159 $\pm$ 5 bp). Approximately 2% of reads included an intron of 40–60
413 bp in the LSU4 region, not visible in Figure S9 due to rarity. Except for these sequences,
414 all conserved regions of LSU, as well as 5.8S, displayed very little size variation, as expected,
415 with standard deviations < 2 bp.

416 *Agaricus bisporus*, the positive control, was represented by a single ASV in the positive control
417 samples for both long- and short-amplicon PacBio datasets, and in the Ion Torrent dataset.
418 *A. bisporus* was represented by two ASVs in the Illumina dataset, which differed at one base
419 pair (99.5% similarity in ITS2). The abundance of the second ASV was 1.1% and 1.0% that
420 of the primary *A. bisporus* ASV in the two Illumina positive controls. The consistency of this
421 ratio across replicate positive controls suggests that it represents true inter-copy variation
422 within the specimen, rather than sequencing or PCR error. Despite higher total sequencing
423 depth, this ASV was not identified from the Ion Torrent dataset.

424 *A. bisporus* sequences represented 0.01%, 0.09%, 0.09%, and 0.09% of non-control reads,
425 in the PacBio long, PacBio short, Illumina, and Ion Torrent datasets, respectively, giving
426 similar estimates for the rate of tag-switching for all technologies. These reads were excluded
427 from community analyses.

3.1 Reproducibility of sequence capture using different technologies

The majority of abundant ASVs and OTUs were captured by all sequencing strategies used (Figure 1). ASVs shared between all datasets represented 56–80% of the reads for the long and short PacBio datasets, Illumina dataset, and Ion Torrent dataset, respectively. These fractions increased to 73–89% when differences at the intra-species scale were removed by clustering the ASVs into 97% OTUs. In particular, 100%, 93% , and 89% of reads in the PacBio, Illumina, and Ion Torrent short-amplicon datasets belonged to OTUs shared between all three datasets. In contrast, 21% of reads in the long PacBio dataset belonged to ASVs which were unique to that dataset, and the fraction only reduced to 20% after OTU clustering. Complete tabulations of the number of ASVs and OTUs shared between the different sequencing strategies are shown in Supplementary Tables S3 and S4, respectively.

Figure 2 shows the correspondence between the read count for different ASVs (2a) and OTUs(2b) in the different technologies, where shared ASVs/OTUs are plotted as circles, and unshared OTUs are plotted as lines along the axes. In all cases, the read counts for shared ASVs and OTUs were correlated, with a minimum R^2 value of 0.47. Correlations between read counts for the three technologies using the short amplicon library were increased by OTU clustering (0.69 to 0.72, 0.49 to 0.74, and 0.74 to 0.82, for PacBio vs. Illumina, PacBio vs. Ion Torrent, and Illumina vs. Ion Torrent, respectively), but not between the long amplicon library and short amplicon library (0.65 to 0.62, 0.58 to 0.57, and 0.47 to 0.49, for PacBio long amplicon reads vs. PacBio, Illumina, and Ion Torrent short reads, respectively).

3.2 Taxonomic assignment

For all sequencing datasets and taxonomic assignment protocols, a higher proportion of reads was assigned than of ASVs, indicating that common ASVs were more likely to be identified than rare ASVs (Figure 3). A greater fraction of ITS reads and ASVs were assigned using the Unite database than the Warcup database across sequencing technologies, amplicons, algorithms, and taxonomic ranks. At most taxonomic ranks, the RDPC algorithm assigned the greatest fraction of reads and ASVs, followed by SINTAX, and then IDTAXA.

Taxonomic composition of the sequenced soil fungal community at the class level is is summarized in Figure 4 and as a heat tree (Foster et al., 2017) in Figure S10. The ML tree for fungal ASVs, along with taxonomic assignments, is shown in Supplementary File 3. According to the PHYLOTAX assignments, Fungi represented 76% of the ASVs and 90% of the reads in the long amplicon library, compared to 89.8%–94.9% of the ASVs and 98.4%–99.0% of the reads in the short amplicon library. Measured fungal community composition at the class level varied significantly between amplicons (PERMANOVA with 9999 permutations, $p < 0.0001$, $R^2 = 0.047$), but only marginally between sequencing technologies ($p = 0.0532$, $R^2 = 0.002$). The majority of variation was spatiotemporal (i.e., between samples; $p < 0.0001$, $R^2 = 0.90$), but once this variation was removed, the remaining effect consisted of a clear bias against Sordariomycetes in the long amplicon dataset (Figures 4 and S12).

Fungi categorized as ECM made up 8.5% of ASVs and 39.4% of reads in the long amplicon library, and 6.3%–13.8% of the ASVs and 36.7%–47.4% of the reads in the short amplicon library (Figure S11). Although amplicon length had a significant effect on ECM community composition at the family level, the explained variation was very low (PERMANOVA with 9999 permutations, $p = 0.0019$, $R^2 = 0.002$), and the majority of variation was again spatiotemporal ($p < 0.0001$, $R^2 = 0.98$). Variation between sequencing technologies was not significant ($p = 0.76$, $R^2 = 0.0002$).

3.3 Spatial analysis

Results of spatial analysis based on the Bray-Curtis dissimilarity were qualitatively similar between the two amplicon libraries and between PacBio and Illumina sequencing, with significant autocorrelation at $p < 0.05$ for ranges of up to 2–3 m for the total fungal community, and 1–2 m for the ECM fungal community (Figure S13). In both cases, the greatest correlation magnitudes were found with Illumina, followed by long amplicon PacBio. The least spatial structure was detected with PacBio short amplicon sequencing.

The Bray-Curtis metric showed significant ($p < 0.05$) positive correlation when resampling at the same locations one year later (i.e., spatial distance of 0 m, time lag of 1 year), for both the total fungal and ECM fungal communities in the long amplicon library. For the short amplicon library, although the general profile of the correlograms was similar, correlation at 0 m and 1 year was not significant, but there was a negative correlation at time lag of 1 year and a distance of 1 m for both sequencing technologies. This puzzling negative correlation was significant in all correlograms based on short amplicon sequencing irrespective of technology.

In contrast to the Bray-Curtis distance, the weighted UNIFRAC distance showed very little spatial structure, with only the total fungal community in the 1 m distance class showing a significant correlation at $p < 0.05$. No temporal correlation was found for the weighted UNIFRAC distance.

The best fit spatial ranges based on distance-decay curves vary between the different datasets by a factor of about 3, but there is overlap of the 95% confidence intervals for all of the Bray-Curtis spatial ranged in both the total fungal and ECM fungal communities, across amplicon libraries and sequencing technologies (Figure 5, Table S5). Although a distance-decay model was fit for the weighted UNIFRAC distance applied to the total fungal community, the result was very poorly constrained, and a range of 0 m, indicating no spatial structure, was included in the 95% confidence interval.

4 Discussion

4.1 Reconstruction of long amplicons from denoised subregions

ASV recovery for long amplicons using DADA2 was dramatically improved (12% to 76% of reads) by denoising homologous subregions independently using the `LSUx` and `tzara` packages. Although newer sequencing platforms from PacBio (Sequel and Sequel II) feature increased sequencing depth and lower error rate compared to the RS II, long sequences inherently require much more sampling depth to identify ASVs. Thus, `tzara` should increase ASV recovery from these platforms as well. It may also be adaptable to Oxford Nanopore sequencing, which has hitherto posed difficulties for application to complex community metabarcoding (Loit et al., 2019).

4.2 Comparison of sequencing strategies

The three sequencing technologies gave similar results for the short amplicon library, the major difference being in sequencing depth. Although a greater fraction of PacBio raw reads were ultimately mapped to ASVs (75%) compared to Illumina (63%) or Ion Torrent (65%), the latter two technologies provided much greater sequencing depth for a similar cost, allowing a greater diversity of rare ASVs to be recovered.

DADA2 denoising may perform differently on different technologies (or perhaps sequencing runs), indicated by the fact that clustering ASVs at 97% led to substantially higher correspondence between both the set of sequences recovered from the same library by different technologies (Figure 1) and the read counts for each sequence (Figure 2). The large number of ASVs unique to Ion Torrent, while only the Illumina dataset recovered an apparent intragenomic variant in the positive control sample, suggests that DADA2 may not control sequencing error as effectively in Ion Torrent sequences as in Illumina, for which it was

developed (Callahan et al., 2016).

Although the longer read length capabilities of PacBio would allow recovery of longer ITS2 sequences than the other two technologies, PacBio did not recover any ITS2 fragments longer than those recovered by Illumina and Ion Torrent. Notably, neither long nor short amplicon sequencing recovered any sequences identifiable to *Cantharellus*, an ECM genus which is commonly observed at the study sites as fruitbodies (personal observation), but which is also known to have accelerated evolution in the rDNA (Moncalvo et al., 2006) and longer ITS regions than other fungi (Feibelman et al., 1994), making it an especially difficult target for metabarcoding. Contrary to expectations, Illumina showed a slightly higher fraction of longer ITS2 sequences than Ion Torrent, which in turn showed slightly longer sequences than PacBio (Figures S8 and S14).

Of long amplicon reads, 21% belonged to ASVs which occurred only in the long amplicon dataset, and clustering at 97% similarity only reduced this fraction to 20%. Additionally, ITS2 sequences extracted from the long amplicon dataset included some sequences that were much shorter than those recovered from the short amplicon datasets (Figure S14). Taxonomic assignments revealed that the majority of these non-shared sequences fall outside kingdom Fungi (Figure S15), and that in particular the short ITS2 sequences are mostly Alveolates (Figure S16). Within Fungi, the short amplicon datasets recovered more Sordariomycetes (Figures 4, S12, and S15). Additionally, several smaller groups showed increased detection in either the long or short datasets, such as Tulasnellaceae and Pyronemataceae in the long amplicon dataset, and *Myerozyma* in the short amplicon datasets (Figures S15 and S17). These differences may be due to primer mismatches in these taxa.

4.3 Taxonomic identification

The RDP fungal training set and Unite performed comparably at taxonomic placement of long amplicon sequences. The Warcup database placed notably fewer sequences at all taxonomic levels for all datasets, probably in part due to the fact that only fungal sequences are included. However, even with this considered, IDTAXA performed very poorly with the Warcup database, placing <25% of ASVs to kingdom in all datasets. IDTAXA placed fewer sequences than RDPC or SINTAX even with the other databases, but this is expected given its more conservative assignment of confidence scores (Murali et al., 2018a).

Gdanetz et al. (2017) showed that a majority-rule consensus of three assignment algorithms can improve the fraction of sequences assigned as well as decrease the false assignment rate. Strict consensus rejects assignments whenever there is conflict between methods and should therefore provide more conservative taxonomic assignments than majority-rule consensus. Here, we found that strict consensus also usually increases the number of assigned sequences relative to any single method, except at family and genus level identifications. This suggests that different assignment algorithms and databases bring mostly complementary, non-contradictory information at higher taxonomic levels. However, contradictory assignments between different methods is more common at lower taxonomic levels, which can be problematic because accurate assignment at the family or genus level is generally required for ecological guild assignment using FUNGuild.

For ASVs where a long amplicon sequence is available, PHYLOTAX uses phylogenetic relationships to resolve these disagreements in a principled manner. For instance, 56% and 87% of Illumina reads were assigned to genus and family, respectively, by the strict consensus of methods, but PHYLOTAX increased this fraction to 75% and 97%. This led to a corresponding increase in the fraction of reads assigned to a functional guild (Figure S11).

4.4 Turnover rate

Weighted UNIFRAC did not reliably detect spatial structure within this relatively ecologically homogeneous community. Although the Mantel test did show a small but significant positive autocorrelation in the fungal community at the smallest size category (1 m; Figure S13), the distance-decay plot in Figure 5 does not show any clear relationship. The functional fit showed poor convergence, with a 95% confidence interval for spatial range of 0–5700 m, indicating little evidence of spatial structure. This is probably due to the majority of weighted branch length in the community being between the Pezizomycotina and Agaricomycetes (Figure S10), which are both well represented in the majority of samples. UNIFRAC would be more suited at larger spatial scales and/or larger ecological gradients.

Mantel correlograms based on the Bray-Curtis dissimilarity (Figure S13) revealed spatial autocorrelation in the soil fungal community at distance classes ≤ 3 m for both Illumina and PacBio using long and short amplicons, and in the ECM fungal community at distance classes ≤ 2 m for Illumina and PacBio long amplicons, and ≤ 1 m for the PacBio short amplicons. These results are similar to autocorrelation ranges found in previous work based on ECM root tips in temperate forests (Lilleskov et al., 2004; Pickles et al., 2012). Lilleskov et al. (2004) found autocorrelation only at ranges <2.6 m at most sites using Sanger sequencing. Similarly, Pickles et al. (2012) found autocorrelation at distances <3.4 m based on T-RFLP analysis. However, previous work in Miombo woodland, a similar ecosystem to the Soudanian woodland in this study, found autocorrelation at ranges <10 m using Sanger sequencing of ECM root tips (Tedersoo et al., 2011), which was their smallest distance class.

Distance-decay plots (Figure 5, Table S5) gave substantially longer autocorrelation distances. There was little variation in the results between the Illumina and long-amplicon PacBio datasets for both the total fungal community and the ECM community, with best fit estimates ranging from 13–18 m. The 95% confidence interval was substantially wider than this

variation, generally covering a range of 5–41 m. All of these values are smaller than the 65 m reported by Bahram et al. (2013), also based on distance-decay curves from an ECM woodland habitat in Benin.

The PacBio short amplicon dataset shows a longer spatial range, of 25 m for the total fungal community and 38–38 m for the ECM community, in both cases with wide confidence intervals spanning 11–214 m. It is possible that the weaker fit for this dataset, which also showed weaker autocorrelation in the Mantel correlogram, is due to low sequencing depth.

The Bray-Curtis Mantel correlogram for both the total fungal and ECM communities from the long amplicon dataset show a significant positive correlation at 0 m and 1 year. The spatiotemporal distance-decay fit estimated the temporal turnover range as 3.3 years for the total fungal community and 4.1 years for the ECM community, but with overlapping confidence intervals. Both datasets from the short amplicon library showed a puzzling pattern with no autocorrelation at 0 m and 1 year, but a weak negative correlation at 1 m and 1 year. The general shape of the correlograms were similar for long and short amplicon datasets. We hypothesize that two different processes may be at work with differing spatiotemporal scales, whose superposition result in this pattern.

4.5 Conclusion

The choice of amplicon and sequencing technology did not seem to affect the results of the spatial analysis, provided sufficient sequencing depth. However, the addition of long amplicon reads did allow the construction of a phylogenetic tree from the metabarcoding reads, which allowed refinement of taxonomic assignments. DADA2 ASV yield was initially poor for long reads, but this was improved by developing a workflow for extraction of subregions, separate denoising, and then reconstruction of full-length unique sequences. Together these approaches provide a hybrid approach using long-read sequencing to acquire long amplicon

sequences for the local species pool, and cost-effective short-read sequencing to provide high sampling depth and sample number.

Acknowledgements

This project was funded by the Swedish research council FORMAS grant number 2014-01109.

Laboratory work including PCR and library pooling was performed by Dr. Ylva Strid.

The authors would like to acknowledge support of the National Genomics Infrastructure (NGI) / Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI / VR and and Science for Life Laboratory, Sweden.

Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

References

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., Campo, J. del, Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., … Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119. https://doi.org/10.1111/jeu.12691

Ainsworth, G. C. (2008). *Ainsworth & Bisby's dictionary of the fungi*. Cabi.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3)pmid 2231712, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*, *2*(2)pmid 28289731, e00191–16. https://doi.org/10.1128/mSystems.00191-16

Bahram, M., Koljalg, U., Courty, P.-E., Diedhiou, A. G., Kjøller, R., Polme, S., Ryberg, M., Veldre, V., & Tedersoo, L. (2013). The distance decay of similarity in communities of ectomycorrhizal fungi in different ecosystems and scales. *Journal of Ecology*, *101*(5), 1335–1344.

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., Wit, P. D., Sánchez-García, M., Ebersberger, I., Sousa, F. de, Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., … Nilsson, R. H. (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

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology*, *60*(3), 291–302. https://doi.org/10.1093/sysbio/syr010

Brundrett, M. C. (2017). Global Diversity and Importance of Mycorrhizal and Nonmycorrhizal Plants. In L. Tedersoo (Ed.), *Biogeography of Mycorrhizal Symbiosis* (pp. 533–556). Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-56363-3_21

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., & Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research*, *47*(18), e103–e103. https://doi.org/10.1093/nar/gkz569

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & 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

Deshpande, V., Wang, Q., Greenfield, P., Charleston, M., Porras-Alfaro, A., Kuske, C. R., Cole, J. R., Midgley, D. J., & Tran-Dinh, N. (2016). Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia*, *108*(1), 1–5. https://doi.org/10.3852/14-293

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, *22*(11), 2079–2088. https://doi.org/10.1093/nar/22.11.2079

Edgar, R. C. (2016a). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161. https://doi.org/10.1101/074161

Edgar, R. C. (2016b). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, 081257. https://doi.org/10.1101/081257

Feibelman, T., Bayman, P., & Cibula, W. G. (1994). Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. *Mycological Research*, *98*(6), 614–618. https://doi.org/10.1016/s0953-7562(09)80407-3

Foster, Z. S. L., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLOS Computational Biology*, *13*(2), e1005404. https://doi.org/10.1371/journal.pcbi.1005404

[Dataset] Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2020). *Data for "Long- and short-read metabarcoding reveal similar spatio-temporal structures in fungal communities"*. Dryad. https://doi.org/XXXX

Gdanetz, K., Benucci, G. M. N., Vande Pol, N., & Bonito, G. (2017). CONSTAX: A tool for improved taxonomic resolution of environmental fungal ITS sequences. *BMC Bioinformatics*, *18*(1), 538. https://doi.org/10.1186/s12859-017-1952-x

Glöckner, F. O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Bruns, G., Yarza, P., Peplies, J., Westram, R., & Ludwig, W. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology*, *261*, 169–176. https://doi.org/10.1016/j.jbiotec.2017.06.1198

Holst-Jensen, A., Vaage, M., Schumacher, T., & Johansen, S. (1999). Structural characteristics and possible horizontal transfer of group I introns between closely related plant pathogenic fungi. *Molecular Biology and Evolution*, *16*(1), 114–126. https://doi.org/10.1093/oxfordjournals.molbev.a026031

Horton, T. R., & Bruns, T. D. (2001). The molecular revolution in ectomycorrhizal ecology: Peeking into the black-box. *Molecular Ecology*, *10*(8), 1855–1871. https://doi.org/10.1046/j.0962-1083.2001.01333.x

Ihrmark, K., Bödeker, I. T. M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K. E., & Lindahl, B. D. (2012). New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology*, *82*(3), 666–677. https://doi.org/10.1111/j.1574-6941.2012.01437.x

Jehl, P., Sievers, F., & Higgins, D. G. (2015). OD-seq: Outlier detection in multiple sequence alignments. *BMC Bioinformatics*, *16*(1), 269. https://doi.org/10.1186/s12859-015-0702-1

Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., & Petrov, A. I. (2018). Non-Coding RNA Analysis Using the Rfam Database. *Current Protocols in Bioinformatics*, *62*(1), e51. https://doi.org/10.1002/cpbi.51

Kennedy, P. G., Cline, L. C., & Song, Z. (2018). Probing promise versus performance in longer read fungal metabarcoding. *New Phytologist*, *217*(3), 973–976.

Legendre, P., & Legendre, L. F. J. (2012, July 21). *Numerical Ecology*. Elsevier.

Lilleskov, E. A., Bruns, T. D., Horton, T. R., Taylor, D. L., & Grogan, P. (2004). Detection of forest stand-level spatial structure in ectomycorrhizal fungal communities. *FEMS Microbiology Ecology*, *49*(2), 319–332. https://doi.org/10.1016/j.femsec.2004.04.004

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., et al. (2013). Fungal community analysis

by high-throughput sequencing of amplified markers–a user's guide. *New Phytologist*, *199*(1), 288–299.

Loit, K., Adamson, K., Bahram, M., Puusepp, R., Anslan, S., Kiiker, R., Drenkhan, R., & Tedersoo, L. (2019). Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens. *Applied and Environmental Microbiology*, *85*(21)pmid 31444199. https://doi.org/10.1128/AEM.01368-19

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

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538

Michot, B., Hassouna, N., & Bachellerie, J.-P. (1984). Secondary structure of mouse 28S rRNA and general model for the folding of the large rRNA in eukaryotes. *Nucleic Acids Research*, *12*(10), 4259–4279. https://doi.org/10.1093/nar/12.10.4259

Moncalvo, J.-M., Nilsson, R. H., Koster, B., Dunham, S. M., Bernauer, T., Matheny, P. B., Porter, T. M., Margaritescu, S., Weiß, M., Garnica, S., Danell, E., Langer, G., Langer, E., Larsson, E., Larsson, K.-H., & Vilgalys, R. (2006). The cantharelloid clade: Dealing with incongruent gene trees and phylogenetic reconstruction methods. *Mycologia*, *98*(6), 937–948. https://doi.org/10.1080/15572536.2006.11832623

Murali, A., Bhargava, A., & Wright, E. S. (2018a). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(1), 140. https://doi.org/10.1186/s40168-018-0521-5

Murali, A., Bhargava, A., & Wright, E. S. (2018b). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(1), 140. https://doi.org/10.1186/s40168-018-0521-5

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/bioinformatics/btt509

Nguyen, N. H., Song, Z., Bates, S. T., Branco, S., Tedersoo, L., Menke, J., Schilling, J. S., & Kennedy, P. G. (2016). FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology*, *20*, 241–248.

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2019). *Vegan: Community ecology package* [R package version 2.5-6]. R package version 2.5-6. https://CRAN.R-project.org/package=vegan

Pacific Biosciences. (2019, February 22). *Consensus library and applications. Contribute to PacificBiosciences/unanimity development by creating an account on GitHub*. Retrieved March 11, 2019, from https://github.com/PacificBiosciences/unanimity

Pickles, B. J., Genney, D. R., Anderson, I. C., & Alexander, I. J. (2012). Spatial analysis of ectomycorrhizal fungi reveals that root tip communities are structured by competitive interactions. *Molecular Ecology*, *21*(20), 5110–5123. https://doi.org/10.1111/j.1365-294X.2012.05739.x

Raué, H. A., Klootwijk, J., & Musters, W. (1988). Evolutionary conservation of structure and function of high molecular weight ribosomal RNA. *Progress in Biophysics and Molecular Biology*, *51*(2), 77–129. https://doi.org/10.1016/0079-6107(88)90011-9

Rinaldi, A., Comandini, O., & Kuyper, T. W. (2008). Ectomycorrhizal fungal diversity: Seperating the wheat from the chaff. *Fungal Diversity*, *33*, 1–45.

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

Ryberg, M. (2015). Molecular operational taxonomic units as approximations of species in the light of evolutionary models and empirical data from Fungi. *Molecular Ecology*, *24*(23), 5770–5777. https://doi.org/10.1111/mec.13444

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., & Consortium, F. B. (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)pmid 22454494, 6241–6246. https://doi.org/10.1073/pnas.1117018109

Smith, S. E., & Read, D. J. (2010). *Mycorrhizal symbiosis*. Academic press.

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, *32*(19), 2920–2927. https://doi.org/10.1093/bioinformatics/btw346

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Tedersoo, L. (2017a). Proposal for practical multi-kingdom classification of eukaryotes based on monophyly and comparable divergence time criteria. *bioRxiv*, 240929. https://doi.org/10.1101/240929

[Dataset] Tedersoo, L. (2017b). *Proposed practical classification of the domain Eukarya based on the NCBI system and monophyly and comparable divergence time criteria*. PlutoF. https://doi.org/10.15156/BIO/587483

Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., Kõljalg, U., Kisand, V., Nilsson, H., Hildebrand, F., Bork, P., & 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., Bahram, M., Jairus, T., Bechem, E., Chinoya, S., Mpumba, R., Leal, M., Randrianjohany, E., Razafimandimbison, S., Sadam, A., et al. (2011). Spatial structure and the effects of host and soil environments on communities of ectomycorrhizal fungi in wooded savannas and rain forests of Continental Africa and Madagascar. *Molecular*

Ecology, *20*(14), 3071–3080. Retrieved March 10, 2016, from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2011.05145.x/full

Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., Bahram, M., Döring, M., Schigel, D., May, T., Ryberg, M., & Abarenkov, K. (2018). High-level classification of the Fungi and a tool for evolutionary ecological analyses. *Fungal Diversity*. https://doi.org/10.1007/s13225-018-0401-0

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

Urbina, H., Scofield, D. G., Cafaro, M., & Rosling, A. (2016). DNA-metabarcoding uncovers the diversity of soil-inhabiting fungi in the tropical island of Puerto Rico. *Mycoscience*, *57*(3), 217–227. https://doi.org/10.1016/j.myc.2016.02.001

Vesterinen, E. J., Ruokolainen, L., Wahlberg, N., Peña, C., Roslin, T., Laine, V. N., Vasko, V., Sääksjärvi, I. E., Norrdahl, K., & Lilley, T. M. (2016). What you need is what you eat? Prey selection by the bat Myotis daubentonii. *Molecular Ecology*, *25*(7), 1581–1594. https://doi.org/10.1111/mec.13564

Vilgalys, R., & Hester, M. (1990). Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several Cryptococcus species. *Journal of Bacteriology*, *172*(8)pmid 2376561, 4238–4246. https://doi.org/10.1128/jb.172.8.4238-4246.1990

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.*, *73*(16)pmid 17586664, 5261–5267. https://doi.org/10.1128/AEM.00062-07

White, T., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In M. Innis, D. Gelfand, J. Sninsky, & T. White (Eds.), *PCR Protocols: A Guide to Methods and Applications* (pp. 315–322). Academic Press, Inc.

Wong, R. G., Wu, J. R., & Gloor, G. B. (2016). Expanding the UniFrac Toolbox. *PLOS ONE*, *11*(9), e0161196. https://doi.org/10.1371/journal.pone.0161196

Wright, E. S. (2015). DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics*, *16*(1), 322. https://doi.org/10.1186/s12859-015-0749-z

Data Accessibility

- Trimmed, demultiplexed sequencing reads have been deposited at the European Nucleotide Archive (ERA) under Project accession number PRJEB37385. Accession numbers are given in Supplementary Files 1 and 2.

- Consensus ASV sequences will also be deposited at ENA prior to publication.
- Nucleotide alignment and ML tree will be deposited at Dryad prior to final publication (Furneaux et al., 2020).
- R packages `LSUx`, `tzara`, `phylotax`, and `FUNGuildR` are available on Github at https://github.com/brendanf/LSUx, https://github.com/brendanf/tzara, https://github.com/brendanf/phylotax, and https://github.com/brendanf/FUNGuildR. These packages are currently being prepared for submission to CRAN/Bioconductor. If they are not accepted prior to final publication, snapshots will be archived at Dryad.
- FASTA-format files for the RDP, Warcup, and Unite reference databases with unified classifications, as well as scripts used to generate them, are available at https://github.com/brendanf/reannotate. The versions used in this paper will be archived at Dryad prior to publication.
- Bioinformatics pipeline and analysis scripts are available at https://github.com/oueme-fungi/oueme-fungi-transect.

Author Contributions

Sampling was planned and carried out by BF, NSY, and MR. Bioinformatics and data analysis were performed by BF with input from MB, AR, and MR. Scripts and R packages were written by BF. The manuscript was drafted by BF and MR. All authors contributed to and approved the final version of the manuscript.

List of Figures

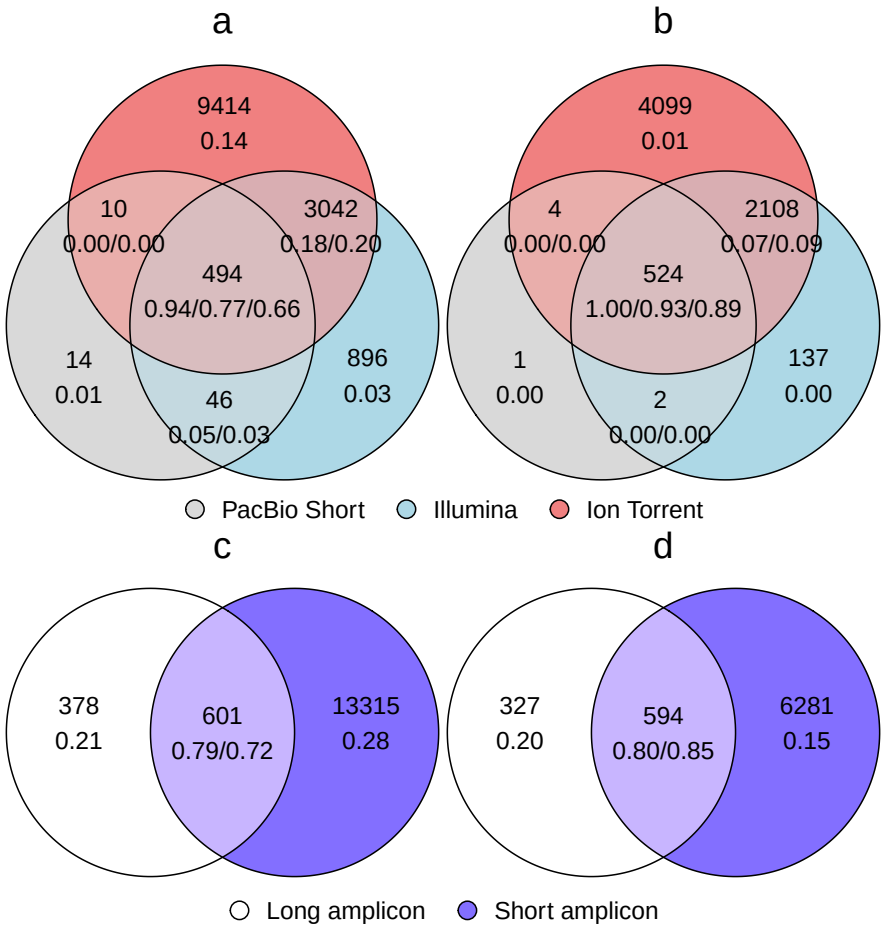

Figure 1: Venn diagrams showing shared ITS2-based ASVs (*a*, *c*) and 97% OTUs (*b*, *d*) between different sequencing technologies from the same short amplicon library (*a*, *b*), and between long and short amplicon libraries (*c*, *d*). In each region, the number of ASVs/OTUs is given above, while the fractions of reads for each sequencing strategy are shown below. For short amplicons in *c* and *d*, ASV/OTU counts reflect detection by any of the three technologies, and read counts represent the mean fraction of reads across the three technologies.

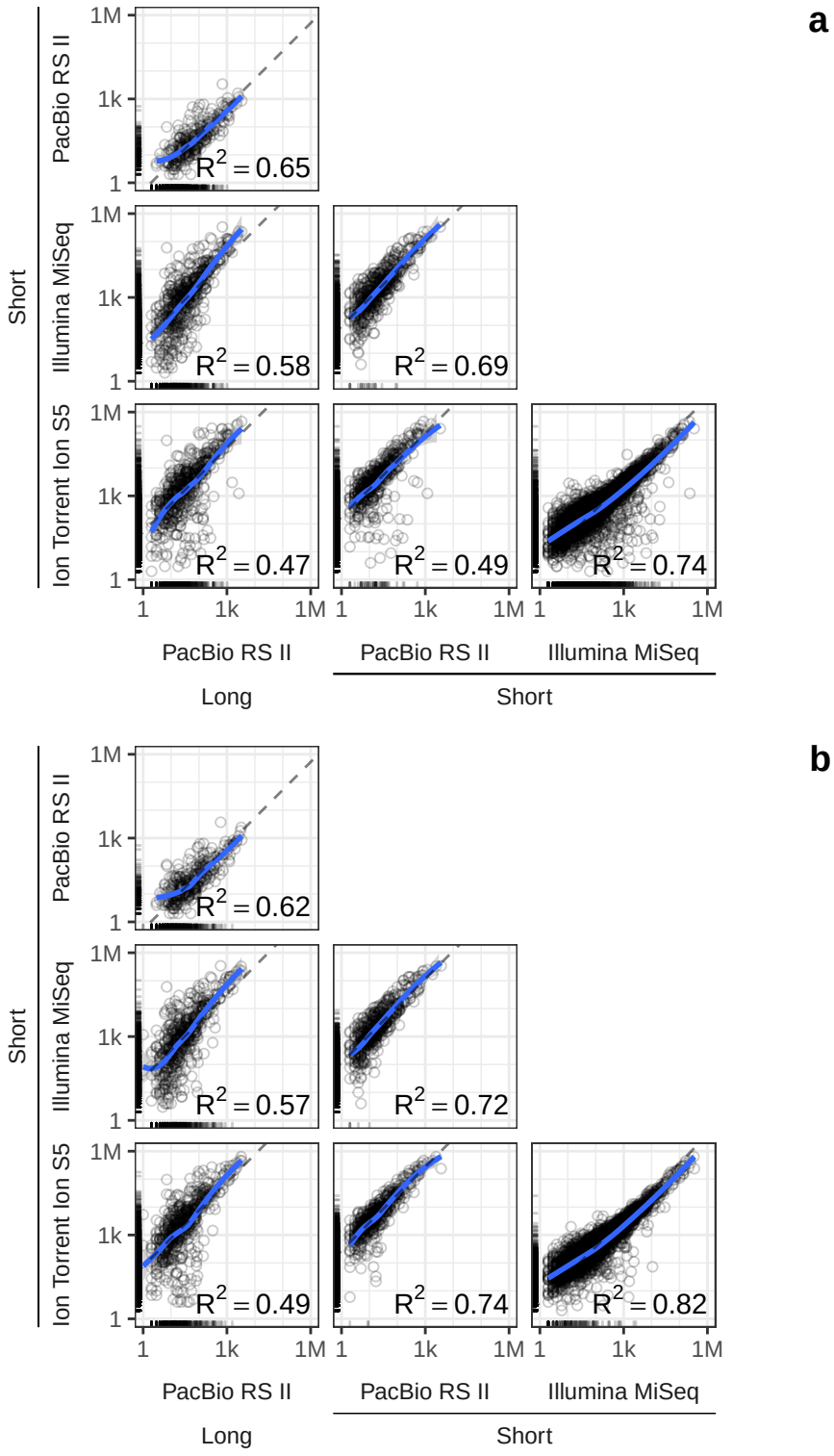

Figure 2: Comparison between read numbers for different sequencing strategies, by ASV (**a**) and 97% OTU (**b**). ASVs/OTUs which were detected by one sequencing strategy but not the other are plotted as tick marks along the axes. Dashed line represents a constant ratio of read numbers. The blue line is a LOESS smooth of the data, with associated uncertainty in grey shading. R^2 value displayed is for log-transformed non-zero read numbers.

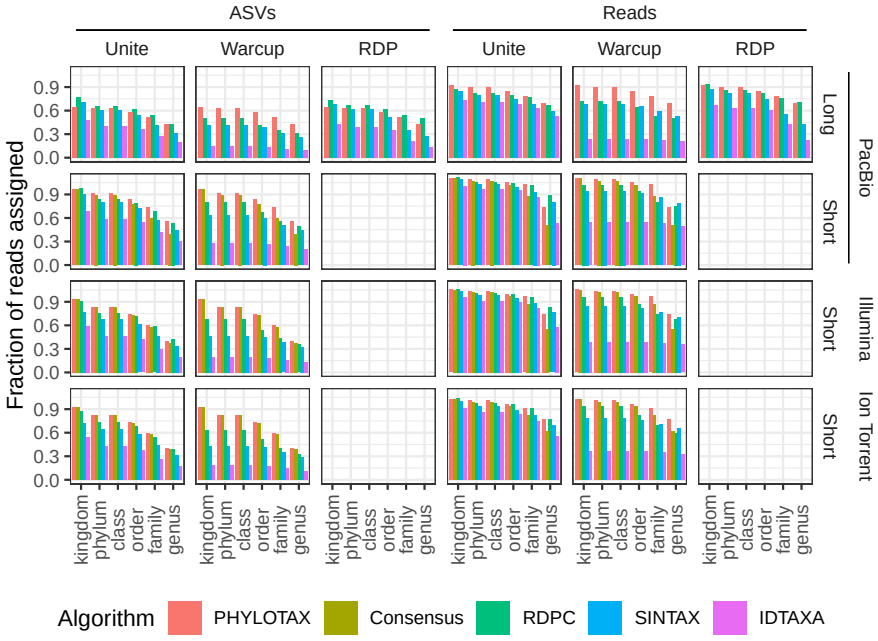

Figure 3: Fraction of ASVs (left) and reads (right) assigned to each taxonomic rank, for different sequencing technologies (PacBio RS II, Illumina MiSeq, Ion Torrent Ion S5), amplicons (Long, Short), reference databases (Unite, Warcup, RDP), and assignment algorithms (PHYLOTAX, Consensus, RDPC, SINTAX, IDTAXA). Consensus and PHYLOTAX assignments are based on the consensus of RDPC, SINTAX, and IDTAXA, using all available databases and, in the case of PHYLOTAX, phylogenetic information. These two methods are plotted in each column to compare with results for the individual databases.

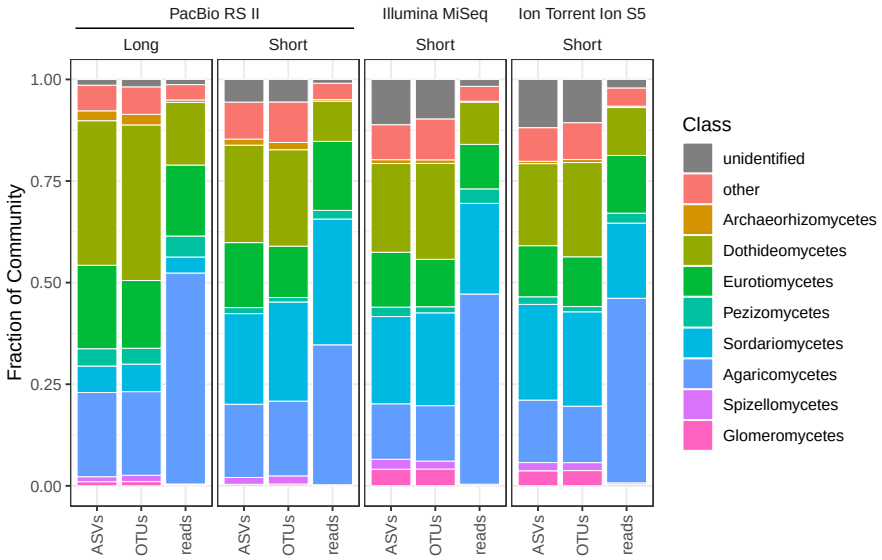

Figure 4: Taxonomic composition of fungal community at the class level. Values represent the fraction of all ASVs, OTUs, or reads which were assigned to kingdom Fungi. Assignments based on PHYLOTAX. Classes which represented less than 2% of reads, OTUs, and ASVs in all datasets are grouped together as "other".

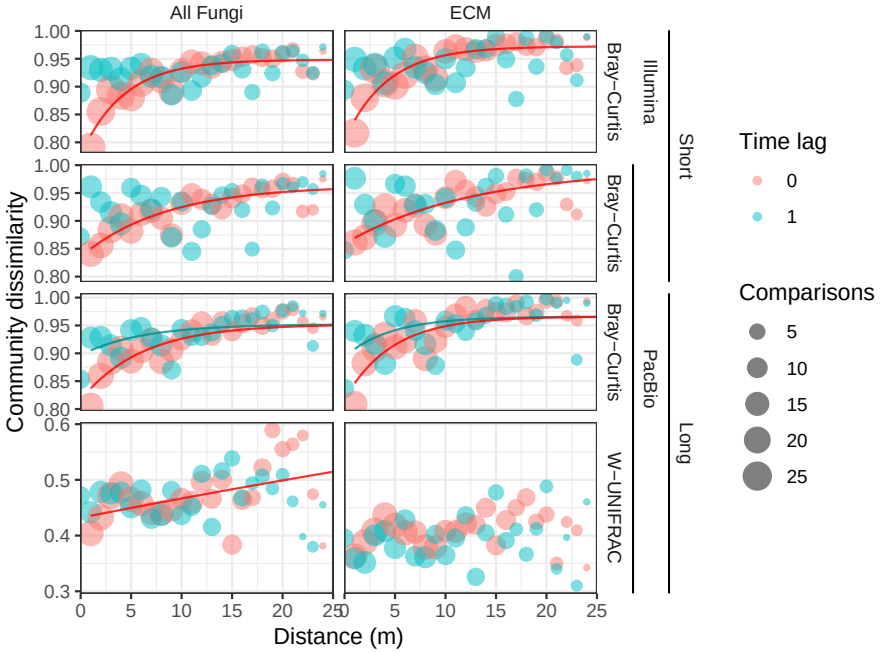

Figure 5: Distance-decay plot for community dissimilarities and spatio-temporal distance. Circles represent community data from short (top two rows) and long (bottom two rows) amplicon libraries, sequenced by Illumina MiSeq (top row) or PacBio RS II (bottom three rows). Community dissimilarities are calculated using the Bray-Curtis dissimilarity for all datasets (top three rows) and using the weighted UNIFRAC dissimilarity for the long amplicon library, for which a phylogenetic tree could be constructed (bottom row). The left column represents the full fungal community, and the right column only sequences identified as ECM. The color of each circle represents the time lag between samples being compared (0 or 1 year), and the size represents the number of comparisons for that spatial distance and time lag. Lines are the best-fit lines for an exponential decay to max model. The model was only fit for datasets where the Mantel test indicated a significant relationship between community dissimilarity and spatial (for the 0 year timelag) or spatiotemporal (for the 1 year time lag) distance.