RESOURCE ARTICLE

Long- and short-read metabarcoding technologies reveal similar spatiotemporal 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

Correspondence
Brendan Furneaux, Program in Systematic Biology, Department of Organismal Biology, Uppsala University, Uppsala, Sweden.
Email: brendan.furneaux@ebc.uu.se

Funding information
Swedish research council FORMAS, Grant/Award Number: 2014-01109; Science for Life Laboratory, Sweden; National Genomics Infrastructure Sweden; Swedish Research Council; Knut and Alice Wallenberg Foundation

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 spatiotemporal 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, all from the same PCR library, 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, 2008). 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. Amherstiae (Brundrett, 2017), together representing approximately 60% of tree stems globally (Steidinger et al., 2019).

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*) fruitbodies. Even when fruitbodies are large, they are ephemeral, so study of ECM communities

is facilitated by sampling of 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 such as 454 Pyrosequencing, Illumina and Ion Torrent have usually targeted the rDNA internal transcribed spacer regions ITS1 or ITS2, which are the standard molecular barcode for fungi, providing 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). In some groups such as arbuscular mycorrhizal fungi, variable regions of the rDNA small subunit (SSU) are the barcode of choice (Öpik et al., 2010), and variable regions of the rDNA large subunit (LSU) have also been used for barcoding (e.g. House et al., 2016; Kurtzman & Robnett, 1998; Tedersoo et al., 2015). 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) with public databases such as GenBank (Benson et al., 2013) and Unite (Nilsson et al., 2019). However, there is potential to improve this approach at several stages, including sequencing technology, amplicon choice, clustering and taxonomic assignment.

Different sequencing technologies have different capabilities in terms of sequencing depth and read length, as well as differing quality profiles and potential biases (Yang et al., 2013). The rapid development of new HTS technologies, as well as subsequent iterative improvements in sequencing chemistry and read capacity, means that the technologies used in metabarcoding studies, along with any associated biases, change frequently. As an example, the first study using HTS metabarcoding of soil fungi was published in 2009 (Buée et al., 2009) using 454 Pyrosequencing; production of 454 sequencers was subsequently discontinued in 2015, and sales of reagents stopped in 2016 (Hollmer, 2013). This brings into question the comparability of studies conducted only a few years apart. Existing studies that sequenced the same environmental samples using different HTS technologies (e.g. Claesson et al., 2010; Divoll et al., 2018; Kennedy et al., 2018; D. P. Smith & Peay, 2014; Speranskaya et al., 2018; Tedersoo et al., 2018) have found that most differences in results seem to be attributable to differences in sequencing depth or different primer biases, rather than differences in the technologies themselves. Only a few of these studies have controlled for primer biases by using the same primer pairs in each technology (Claesson et al., 2010; Divoll et al., 2018; Speranskaya et al., 2018), and to our knowledge, none have sequenced the same PCR products using multiple HTS technologies.

ITS1 and ITS2 often have suitable variation to distinguish species, although closely related species may share identical ITS sequences in certain groups such as various Pezizomycotina (Schoch et al., 2012), but this variability means that they cannot be reliably aligned over the fungal kingdom (Lindahl et al., 2013; Tedersoo et al., 2018). Additionally, the wide range of length variation of these regions may introduce bias in recovery of different taxa (Ihrmark et al., 2012; Palmer et al., 2018; Tedersoo et al., 2015). 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 intraspecies variation and sequencing error (Lindner & Banik, 2011; Nilsson et al., 2008), and results are data set-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 intraspecies 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, although alpha and beta diversity estimates based on ASVs and OTUs at different clustering thresholds are highly correlated (Glassman & Martiny, 2018; Botnen et al., 2018). Amplicon sequence variants have been suggested to be less data set specific than cluster-based OTUs (Callahan et al., 2017). Support for PacBio has recently been added to DADA2 (Callahan et al., 2019), but its application requires greater sequencing depth for longer reads, especially in high diversity samples.

Because both OTU clustering and denoised ASVs may 'clump' different species into a single unit and 'split' a single species into multiple units (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; Munch et al., 2008; Munch et al., 2008), 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 et al., 2018) or to create hybrid trees using ITS and a more conserved marker such as SSU or LSU based on matching taxonomic annotations in reference databases (Fouquier et al., 2016), but these approaches are only applicable to sequences of known taxonomic affiliation.

Assignment of taxonomic identities to environmental sequences is dependent on both the reference database and the algorithm used. Although the public INSDC databases (Karsch-Mizrachi et al., 2018) are often used for sequence identification, the open nature of submission to these databases results in a substantial fraction of incorrect taxonomic annotations (Bidartondo, 2008; Nilsson et al., 2006;

Steinegger & Salzberg, 2020) as well as sequences of poor technical quality (Ashelford et al., 2006; Nilsson et al., 2012). Consequently, taxonomic assignments based on these databases may be incorrect or inconsistent (Nilsson et al., 2005). Several curated databases also exist which attempt to address these issues and which cover the whole fungal kingdom. The Unite database is an attempt to include all publicly available high-quality ITS sequences (800,000 as of release 8.0), originally limited to fungi but now expanded to include all eukaryotes (Nilsson et al., 2019), where efforts have been made to correct incorrect annotations and exclude low-quality sequences (Abarenkov et al., 2018). The Ribosomal Data Project (RDP, Cole et al., 2014) hosts two additional manually curated fungal barcode sequence databases, which are specifically intended for use in taxonomic assignment of sequences: the Warcup ITS training set, containing 18,000 manually curated fungal ITS sequences (Deshpande et al., 2016), and the RDP fungal LSU training set, containing 8000 manually curated LSU sequences from fungi and 3000 from other eukaryotic groups (RDP-LSU, Liu et al., 2012). Although the quality of sequences and taxonomic annotations is undoubtedly higher in these more curated databases, they are inherently limited in taxonomic coverage and do not include the most recently published sequences.

Assigning taxonomy to unknown sequences using BLAST requires *a priori* choice of similarity thresholds for different taxonomic ranks. Several algorithms specifically designed for taxonomic assignment have been published which instead use information about variability within different taxa in the reference database to assign unknown sequences, along with confidence estimates for these assignments, including the RDP Classifier (RDPC, Wang et al., 2007), SINTAX (Edgar, 2016a) and IDTAXA (Murali et al., 2018) among others. In addition, methods have been published which integrate predictions from multiple algorithms to increase the reliability of assignments (Gdanetz et al., 2017; Palmer et al., 2018; Somervuo et al., 2016). However, all sequence similarity-based approaches are dependent on high taxonomic coverage in the reference database, making the placement of novel or undersampled groups problematic (Nilsson et al., 2016; Tedersoo et al., 2018).

Recent long-read HTS technologies such as Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing enable sequencing longer amplicons, which include both the ITS regions and the flanking, more highly conserved SSU and/or LSU regions. This can improve taxonomic placement of sequences that lack close database matches and allow the alignment of metabarcoding reads for subsequent phylogenetic analysis (Tedersoo et al., 2018). Information from phylogenetic trees produced from long-amplicon metabarcoding has the potential to both improve taxonomic assignment and provide alternative measures of community alpha and beta diversity. PacBio sequencing has also been shown to recover longer variants of variable-length regions such as ITS1 and ITS2, which are excluded by other technologies (Castaño et al., 2020). 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).

Because of the variety of sequencing platforms and analytical pipelines, which have been used in metabarcoding studies, comparisons between studies may be difficult. Here, we investigated the effects of different sequencing strategies and postanalysis on biological conclusions using measurement of the spatiotemporal turnover rate of the fungal community in an ECM-dominated woodland in Benin by metabarcoding of bulk soil, sampled at narrow intervals, over two years. Turnover scale is the distance at which two communities can be considered to be independent samples of the local species pool (Legendre & Legendre, 2012). Knowledge of turnover scale is important when planning studies of local diversity and its environmental correlates. Turnover scale varies between different ecosystems and taxonomic groups, and can be 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).

We compare three different sequencing platforms (PacBio RS II, Illumina MiSeq, Ion Torrent Ion S5), long and short amplicons, three different taxonomic assignment algorithms (RDPC, SINTAX, IDTAXA) with three different reference databases (Unite, Warcup, RDP-LSU), and both nonphylogenetic and phylogenetic community distance measures. We also present new algorithms for dividing the LSU into domains, combining denoising results from multiple domains as a strategy to capture more ASVs from long amplicons in diverse communities, and incorporating phylogenetic information into taxonomic assignments. We hypothesize that (a) PacBio sequencing of the short amplicon gives less bias against longer ITS2 amplicons than Illumina and Ion Torrent, both qualitatively (recovering amplicons missed by the others) and quantitatively (greater fraction of reads in longer amplicons); (b) our long amplicons (ITS1-LR5) recover a more complete view of the fungal community than our short amplicons (gITS7-ITS4), due to reduced length and primer biases; (c) these differences lead to differing results for ecological metrics, specifically OTU/ASV richness and turnover distance; and (d) incorporating LSU in the long amplicon allows for better taxonomic assignment.

2 | MATERIALS AND METHODS

2.1 | Sampling

Sampling was conducted at two sites, near the villages of Angaradebou (Ang: 9.75456°N 2.14064°E) and Gando (Gan: 9.75678°N 2.31058°E) approximately 30 km apart in the *Forêt Classée de l'Ouémé Supérieur* (Upper Ouémé Forest Reserve) in central Benin. Both sites were located in West Sudanian savannah woodlands (Yorou et al., 2014; Olson et al., 2001) dominated by the ECM host tree *Isoberlinia doka* (Fabaceae tr. Amherstiae). At each site, 25 soil samples were collected along a single 24-m 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 a total of 67 samples. For each sample, coarse organic debris was removed from the soil surface and a sample of approximately 5 cm × 5 cm × 5 cm

was extracted with an ethanol sterilized knife blade. Each sample was sealed in a plastic zipper bag and homogenized by shaking and manually breaking apart soil aggregations. A subsample of approximately 250 mg total of soil was collected from two locations in the homogenized soil sample and stored in a DNA preservation buffer before return to the laboratory for extraction (see Methods S1.1, as well as Figures S1–S3, for preservation and extraction methods).

2.2 | DNA amplification and sequencing

DNA extracts were sequenced using four distinct strategies, with two different amplicon lengths (long and short; Figure 1a) and three different technologies (PacBio, Ion Torrent and Illumina) for the short amplicon. Due to length limitations of Ion Torrent and Illumina sequencing, long amplicons were only sequenced with PacBio. 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 an equimolar mix of ITS4 (White et al., 1990) and ITS4a (Urbina et al., 2016) as the reverse primer. 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 1a), 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*.

For the short amplicon, forward primers included sample-specific indexes and adapters for multiplexing (File S1). PCR amplification was performed in 20-µl reactions containing 200 µM dNTP mix, 250 µM indexed gITS7 primer, 150 µM ITS4 m, 2 mM MgCl$_2$, 0.1 U *Taq* polymerase (Dream *Taq*, Thermo Fisher Scientific) 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.

For the long amplicon, both forward and reverse primers included indexes for combinatorial multiplexing (File S2). 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.

After pooling of technical replicates, amplicons were purified using SPRI beads (Vesterinen et al., 2016) and quantified fluorometrically using Quant-iT™ PicoGreen™ dsDNA (Thermo Fisher Scientific) fluorescent indicator dye on a Infinite F200 plate spectrofluorometer (Tecan Trading AG) according to the manufacturer's
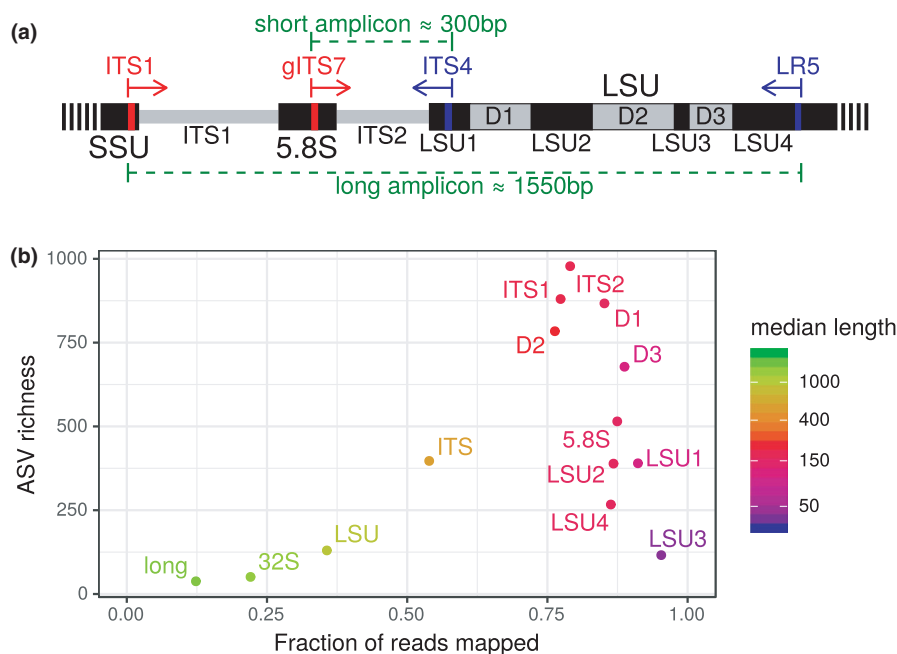

FIGURE 1 rDNA regions. (a) Partial map of rDNA showing the 5.8S rDNA, partial SSU and LSU rDNA, and internally transcribed spacer (ITS) regions. D1–3 represent the first three variable regions in LSU, while LSU1–4 represent the conserved regions. Primer sites used in this study are indicated in red (forward primers) and blue (reverse primers), and the resulting amplicons are shown with green braces. (b) Total number of DADA2 ASVs vs. fraction of demultiplexed reads successfully mapped to ASVs for different rDNA regions extracted from a set of long PacBio amplicon sequences using LSUx. Data from all samples were analysed as a single pool. *long*: entire long amplicon, including ITS1, 5.8S, ITS2 and partial LSU; *32S*: partial 32S precursor to LSU, including 5.8S, ITS2 and partial LSU; *LSU*: section of LSU rDNA included in the long amplicon, from the 5′ end to the LR5 primer site; *ITS*: full ITS region, including ITS1, 5.8S and ITS2. Colour indicates median region length. Shorter and more conserved regions yielded a greater fraction of successfully mapped reads. At a given fraction of mapped reads, more variable regions yield a greater number of unique ASVs

protocol. 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 on a 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 same short-amplicon PCR 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) using one half of a lane. Platform-specific library preparation, including adapter ligation, was performed at the sequencing facilities according to their standard protocols.

2.3 | Bioinformatics

Circular consensus sequence (CCS) basecalls for PacBio sequences were made using CCS version 3.4 (Pacific Biosciences, 2016, 13 July 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, resulting in two samples sharing each tag. These reads were thus either analysed as a pool, or comparisons were made to equivalently combined samples in the other data sets. 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 and clustering

All amplicons were denoised using DADA2 version 1.12.1 according to the ITS pipeline workflow (Callahan, 2020a; Callahan et al., 2016), with technology-specific modifications for Ion Torrent (Callahan, 2020b) and PacBio (Callahan et al., 2019). Although this

was successful for the short amplicons on all technologies, only 38 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 amplicon length and low sequencing depth relative to community diversity, which lead to most biological variants being represented only by a cluster of reads differing by a small number of unique sequencing errors (for calculations, see Methods S1.2.1). We therefore developed a new workflow to assemble ASVs from the long amplicons by splitting the reads into homologous domains, including the two ITS regions, 5.8S, the variable D1–3 regions of LSU (Michot et al., 1984) and the conserved LSU regions between the D regions, here referred to as LSU1–4 (Figure 1a). We then independently denoised reads from each domain and concatenated the denoised domains for each read. Finally, denoised reads were clustered based on 100% ITS2 identity, and a full-length consensus ASV was calculated for each cluster. This method, implemented in the new R packages LSUx (splitting reads into homologous regions; https://github.com/brendanf/LSUx) and TZARA (reassembling regions and generating full-length consensus ASVs; https://github.com/brendanf/tzara) and detailed in Methods S1.2 and Table S1, was used for all of the PacBio and Ion Torrent data sets. Because the LSUx plus TZARA method as currently implemented is not applicable to Illumina paired-end reads, the ASVs generated from the Illumina data set according to the standard DADA2 workflow were used. The ITS2 region was extracted from the ASVs using LSUx for comparison to the results from the other technologies. To account for intraspecies variation and the possibility of different denoising performance between the different sequencing strategies, the pooled ITS2-ASVs from all sequencing strategies were also clustered into operational taxonomic units (OTUs) at 97% similarity using VSEARCH version 2.9.1 (Rognes et al., 2016).

2.3.2 | Phylogenetic inference and taxonomic assignment

Full-length long-amplicon ASVs were aligned using DECIPHER (Wright, 2015) with up to 10 iterations of alternating progressive alignment and conserved RNA secondary structure calculation, followed by 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. Sequences confidently (i.e. by at least five of the primary taxonomic identification methods) assigned outside kingdom Fungi were used to root the tree, and sequences outside the clade defined by confidently identified Fungi were removed (see below and Methods S1.3.2).

Taxonomic annotations of the RDP-LSU training set version 11.5 (Cole et al., 2014; Liu et al., 2012) and Warcup ITS training set (Deshpande et al., 2016) were mapped to a uniform taxonomic classification system (see Methods S1.3.1). Primary taxonomic assignment was performed to genus level separately on the ITS region using Unite and Warcup and on the LSU region using RDP-LSU, respectively, as taxonomic references. For each region/reference combination, taxonomy was assigned using three popular algorithms: the RDPC (Wang et al., 2007) as implemented in DADA2; SINTAX (Edgar, 2016a) as implemented in VSEARCH version 2.9.1 (Rognes et al., 2016); and IDTAXA (Murali et al., 2018). A relatively lax confidence threshold of 50% was used for all three algorithms, in order to increase the amount of input for consensus algorithms. Each full-length ASV was thus given up to nine primary taxonomic assignments (three references × three algorithms). Amplicon sequence variants from the short-amplicon data sets 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.

For full-length long-amplicon ASVs, the primary taxonomic assignments were refined based on the ML phylogenetic tree generated above using the new algorithm PHYLOTAX. The PHYLOTAX algorithm resolves conflicts among one or more primary assignment methods using a supplied phylogenetic tree (see Figure S4 and Methods S1.3.3). It is available in the new R package PHYLOTAX at https://github.com/brendanf/phylotax.

Amplicon sequence variants which were not present in the tree, either because they were not represented in the long-amplicon data set, or because full-length ASV reconstruction failed, were given refined taxonomic assignments using a strict consensus of the different primary assignments at each rank, resulting in a consensus assignment equivalent to the 'last common ancestor' of the primary assignments (Huson et al., 2007). This algorithm has been used to assign a consensus taxonomy based on a list of top BLAST hits (e.g. MEGAN and LCAClassifier, Huson et al., 2016; Lanzén et al., 2012) or k-mer similarity scores (mothur's k-nearest neighbour method, Schloss et al., 2009), but here is used to resolve conflicts between assignments from different algorithms and databases. Strict consensus assignments were also generated for all ASVs, as a comparison to the PHYLOTAX assignments, and are referred to as 'Consensus'.

2.4 | Effect of sequencing strategy on recovered community

We compared alpha diversity estimates by the different sequencing strategies by calculating ASV and OTU accumulation curves, as well as comparing richness estimates after rarefaction for each sample (Methods S1.4). We also compared the effect of sequencing technology and amplicon length on the recovered fungal community composition (i.e. after removal of nonfungi), as assessed by the Bray–Curtis dissimilarity, using PERMANOVA and heat tree visualizations (Methods S1.4).

2.5 | Spatiotemporal analysis

To estimate turnover scale, ecological community dissimilarity matrices were calculated using the ASV/OTU-based Bray–Curtis metric (Bray & Curtis, 1957, for both long and short amplicons) and the phylogenetically based weighted UniFrac metric (Lozupone & Knight, 2005; Lozupone et al., 2007, for only long amplicons) in Phyloseq version 1.26.0. Dissimilarities were based on relative read abundance within each sample. Samples were not rarefied to a standard sequencing depth within data sets, as both the Bray–Curtis dissimilarity and the UniFrac metric are robust to unequal sampling depths (McMurdie & Holmes, 2014). In addition, we did not standardize sequencing depth between data sets, because this would remove one of the potential benefits of Illumina and Ion Torrent relative to PacBio.

Each of the distance matrices was used to calculate a Mantel correlogram with a 1 m bin size for distances in the range of 0–12 m, that is half the maximum separation present in the data set. 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 spatiotemporal distance-decay curves were generated by plotting mean community dissimilarity as a function of spatial distance and time lag, and fit to an exponential model of the form given by Legendre and Legendre (2012) using the nls() function in R (Methods S1.5). Spatiotemporal analyses were performed on the full recovered fungal community after removal of nonfungal sequences and on the ECM community. Sequences were assigned as ECM based on taxonomic assignments using the FUNGuild database (as of 20 Feburary 2020; 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.

3 | RESULTS

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 (Figure S5). Consequently, Ang samples were excluded from spatial analysis, although they were retained for denoising, phylogenetic reconstruction, taxonomic assignment and all nonspatial analyses. Spatial analyses were based on the remaining 34 samples for Illumina and 30 samples each for the PacBio long and short amplicons.

The number of sequencing reads and ASVs at each stage in the bioinformatics pipeline differed between sequencing strategies (Table S2). Sequencing with PacBio 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 and Illumina yielded substantially more reads, with 20.7 million and 10.8 million, respectively. PacBio sequencing of the short-amplicon

library yielded the highest fraction of high-quality reads (≤1 expected error), followed by Illumina, with Ion Torrent yielding the lowest quality (Figure S6b). Although the per-base read quality of the long-amplicon PacBio sequences was similar to that of Illumina (Figure S6a), this translated to a greater number of expected errors per read due to the amplicon length (Figure S6b). Demultiplexing, primer trimming and quality filtering reduced the read totals by 64% for PacBio long amplicons, but only by 21% for PacBio short amplicons, 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. Extraction of only the ITS2 region before quality filtering (Figure S6c) reduced the loss of long-amplicon PacBio reads to only 29%, comparable to Illumina. Application of TZARA resulted in 708 reconstructed long-amplicon ASVs, representing 97% of denoised ITS2 reads from the long-amplicon PacBio data set. Mapping identical ITS2 ASVs from the short- and long-amplicon data sets allowed 58%, 71% and 81% of denoised reads from the Ion Torrent, Illumina and PacBio short-amplicon data sets, respectively, to be assigned to a long-amplicon ASV (Table S2).

Almost all of the short-amplicon sequences from all three technologies were between 240 and 375 bp long (Figure S7a). Although the length profile of the three sequencing runs was similar, Illumina had the largest fraction of reads near the top of the range, followed by Ion Torrent and PacBio (Figure S7b). The difference in length distributions was statistically significant due to the large sample size (Kruskal–Wallis statistic =8.57e+04, $p < 2.2 \times 10^{-16}$), but the difference between means was fairly small, with mean amplicon lengths of 276, 281 and 286 bp for PacBio, Ion Torrent and Illumina, respectively. The length of the long-amplicon reads varied widely, from 696 to 1638 bp, with a mean of 1431 bp (Figure S7c).

Among the different regions extracted from the long amplicon (Figure S8), ITS1 showed the greatest length variability (mean ± standard deviation: 193 ± 55 bp), followed by ITS2 (184 ± 41 bp) and the variable regions in LSU (D2: 227 ± 36 bp; D3: 108 ± 10 bp; D1: 159 ± 6 bp). Approximately 2% of reads included an intron of 40–60 bp in the LSU4 region, not visible in Figure S8 due to rarity. Except for these sequences, all conserved regions of LSU, as well as 5.8S, displayed very little size variation, as expected, with standard deviations <2 bp. Around 12% of ITS2 sequences extracted from the long-amplicon data set were shorter than 140 bp, a much greater fraction than the 0.26%–0.44% from the short-amplicon data sets (Figure S9). The taxonomic identity of these sequences is discussed below.

Agaricus bisporus, the positive control, was represented by a single ASV in the positive control samples for both long- and short-amplicon PacBio data sets, and in the Ion Torrent data set. *Agaricus bisporus* was represented by two ASVs in the Illumina data set, which differed at one base pair (99.5% similarity in ITS2). The abundance of the second ASV was 1.1% and 1.0% that of the primary *A. bisporus* ASV in the two Illumina positive controls. The consistency of this ratio across replicate positive controls suggests that it represents true intercopy variation within the specimen, rather than sequencing

or PCR error. Despite higher total sequencing depth, this ASV was not identified from the Ion Torrent data set.

Agaricus bisporus sequences represented 0.01%, 0.09%, 0.09% and 0.09% of noncontrol reads, in the PacBio long, PacBio short, Illumina and Ion Torrent data sets, respectively, giving similar estimates for the rate of tag-switching for all technologies. These reads were excluded from community analyses.

3.1 | Reproducibility of sequence detection using different technologies

We compared the unique ASVs and OTUs shared between data sets from different sequencing strategies, and the number of reads represented by these ASVs and OTUs in each strategy. The majority of abundant ASVs and OTUs were detected by all sequencing strategies used (Tables S3 and S4). The short-amplicon ASVs shared between all sequencing technologies represented 95%, 76% and 66% of the reads for PacBio, Illumina and Ion Torrent, respectively (Figure 2a). When differences at the intraspecies scale were removed by clustering the ASVs into 97% OTUs, the number of OTUs shared between all three technologies increased to 524, representing 100%, 93% and 89% of reads, respectively (Figure 2b). In particular, the majority of the 9418 unique Ion Torrent ASVs were found to be shared with other sequencing technologies upon OTU clustering. Amplicon sequence variants unique to the Ion Torrent data set made up 14% of reads in that data set, but only 1% belonged to a unique OTU after clustering. In contrast, 21% of reads in the long PacBio data set belonged to ASVs whose ITS2 region was unique to that data set (Figure 2c), and the fraction only reduced to 20% after clustering the ITS2 regions into OTUs (Figure 2d). The taxonomic identity of these ASVs is discussed below.

Read counts for shared ASVs and OTUs were highly correlated between strategies, with a minimum R^2 value of .47 (Figure S10). Correlations between read counts for the three technologies using the short-amplicon library were increased by OTU clustering (0.69–0.72, 0.49–0.74 and 0.74–0.82, 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 (0.65–0.62, 0.58–0.57 and 0.47–0.49, for PacBio long-amplicon reads vs. PacBio, Illumina and Ion Torrent short-amplicon reads, respectively; Figure S10).

Amplicon sequence variant richness estimates after rarefaction were strongly correlated between the three sequencing technologies applied to the short-amplicon library (R^2 = .91–.94; Figure 3). The slope of the relationship between PacBio and Illumina richness estimates was only slightly different from 1, indicating that these two technologies give highly comparable rarefied richness estimates, despite the approximately 200× difference in original sequencing depth. Ion Torrent resulted in rarefied richness estimates, which were 24%–31% greater than the other technologies, an effect which is also visible in ASV accumulation curves (Figure 4). Amplicon sequence variant richness estimates were somewhat less strongly correlated between the PacBio long-amplicon data set and the three short-amplicon data sets
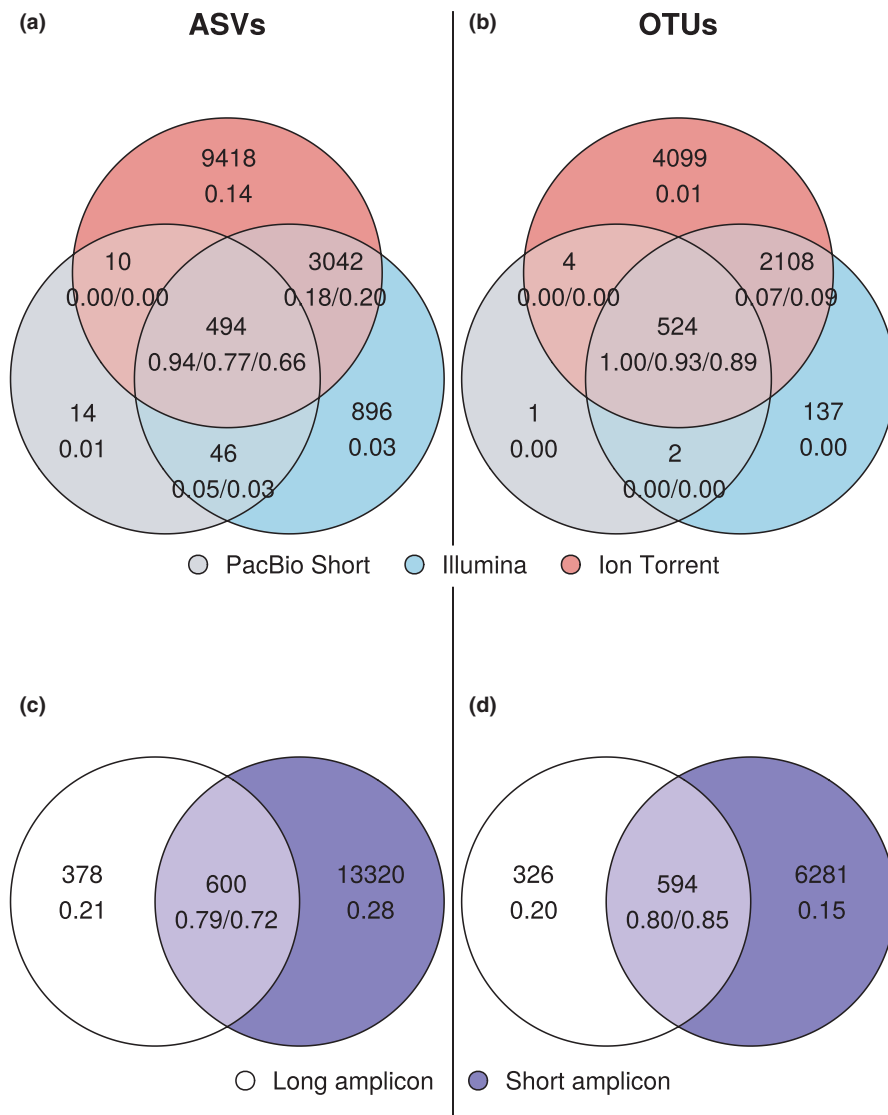
FIGURE 2 Shared richness and abundance of ITS2-based ASVs (a, c) and 97% OTUs (b, d) between different sequencing technologies from the same short-amplicon library (a, b), and between long- and short-amplicon libraries (c, d). In each region, the ASV/OTU richness is given above, while the relative abundance of reads represented by these ASVs/OTUs in each sequencing strategy is shown below in the order PacBio/Illumina/Ion Torrent (a, b), or long/short (c, d). 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. Analyses performed on pooled ASVs/OTUs from all samples

(R^2 = .65–.72; Figure 3). Total least squares regression indicated that the long-amplicon data set resulted in richness estimates which were intermediate between the short-amplicon results from Ion Torrent and the other two technologies. Despite the fact that experiment-wide OTU richness was lower than ASV richness (Figure 2), OTU accumulation curves for each sample (Figure S11) and rarefied OTU richness relationships between sequencing strategies (Figure S12) were highly similar to those for ASVs.

3.2 | Taxonomic assignment

For all sequencing data sets and taxonomic assignment protocols, a higher proportion of reads were assigned than of ASVs, indicating that common ASVs were more likely to be taxonomically identified than rare ASVs (Figure 5). 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 summarized in Figure 6 and as a heat tree (Foster et al., 2017) in Figure S13. The ML tree for fungal ASVs, along with taxonomic assignments, is shown in File S3. According to the PHYLOTAX assignments, fungi represented 88% of the ASVs and 81% of the reads in the long-amplicon library, compared to 92.4%–96.4% of the ASVs and 97.9%–98.3% of the reads in the short-amplicon library. Many of the ASVs that were unique to the long-amplicon library thus fall outside kingdom Fungi (Figure S14). In particular, a large fraction of ITS2 sequences with length less than 140 (Figure S9) were identified as Alveolates (Figure S15).

Measured fungal community composition at the class level varied significantly between long and short amplicons (PERMANOVA with 9999 permutations, $p < .0001$, $R^2 = .048$), but only marginally between sequencing technologies ($p = .0346$, $R^2 = .002$). The

FIGURE 3 Comparison of ASV richness between sequencing technologies. Each point represents the richness of one or two pooled samples, as determined by two different sequencing strategies. All values represent the average of 100 replicate rarefactions with a sample depth of 100 reads. Because samples in the same well on different plates could not be demultiplexed in the Ion Torrent data set, these samples were also bioinformatically pooled in the other data sets prior to rarefaction. Blue lines are total least squares fit with 95% confidence interval, with the given slope (and 95% confidence interval) and R^2 value. Dashed diagonal line indicates 1:1 line
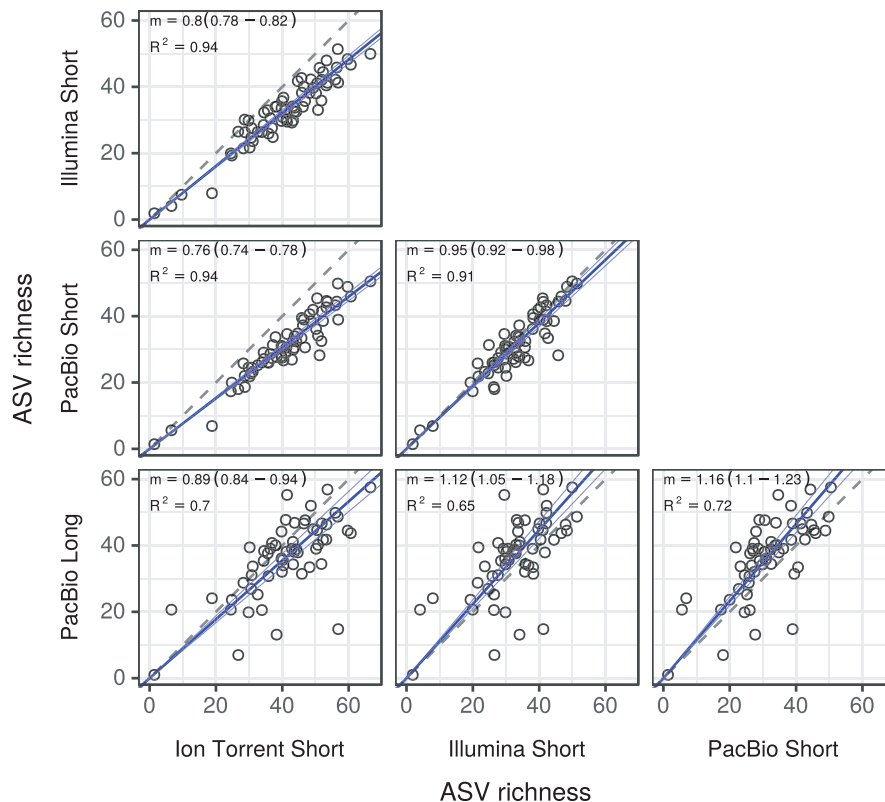

FIGURE 4 Amplicon sequence variant (ASV) accumulation curves. Each curve represents rarefaction of a single sample or pooled pair of samples. Points at the end of each curve represent the actual read depth and observed ASV richness. Panel at left has enlarged scale to show PacBio more clearly. Vertical dashed line at 100 reads indicates rarefaction level for ASV richness comparisons
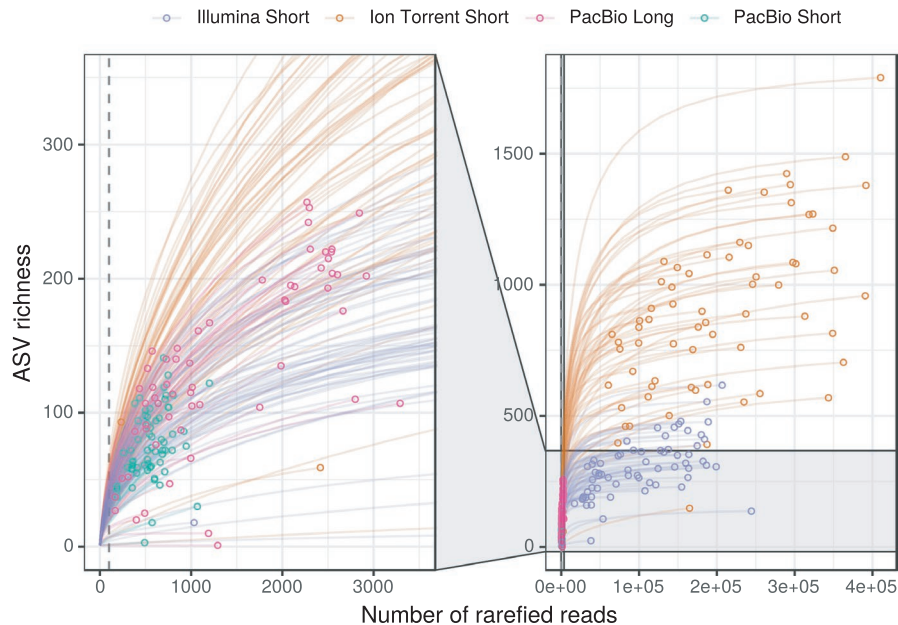
majority of variation was spatiotemporal (i.e. between samples; $p < .0001$, $R^2 = .90$), but once this variation was removed, the remaining effect consisted of a clear bias against Sordariomycetes in the long-amplicon data set (Figures 6, S14 and S16). Additionally, several lower-rank taxonomic groups showed increased detection in either the long or short data sets, such as Tulasnellaceae (Agaricomycetes) and Pyronemataceae (Pezizomycetes) in the long-amplicon data set, and *Meyerozyma* (Saccharomycetes) in the short-amplicon data sets (Figures S14 and S17).

Fungi categorized as ECM made up 9.0% of ASVs and 39.2% of reads in the long-amplicon library, and 5.4%–13.2% of the ASVs and 36.4%–46.4% of the reads in the short-amplicon library (Figure S18). Although amplicon length had a significant effect on ECM community composition at the family level (Figure S17), the explained variation was very low (PERMANOVA with 9999 permutations, $p = .0040$, $R^2 = .002$), and the majority of variation was again spatiotemporal ($p < .0001$, $R^2 = .98$). Variation between sequencing technologies was not significant ($p = .76$, $R^2 = .0002$).
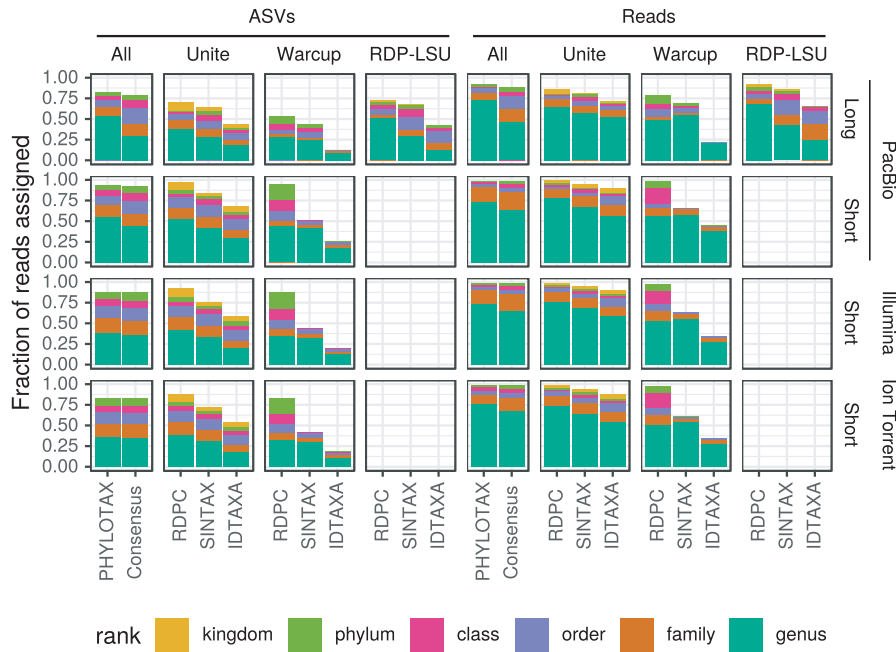
FIGURE 5 Summary of taxonomic assignments. Fraction of ASVs (left) and reads (right) assigned to each taxonomic rank, for different sequencing technologies (PacBio RS II, Illumina MiSeq, Ion Torrent Ion S5), amplicons (Long, Short), reference databases (Unite, Warcup, RDP-LSU) and assignment algorithms (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
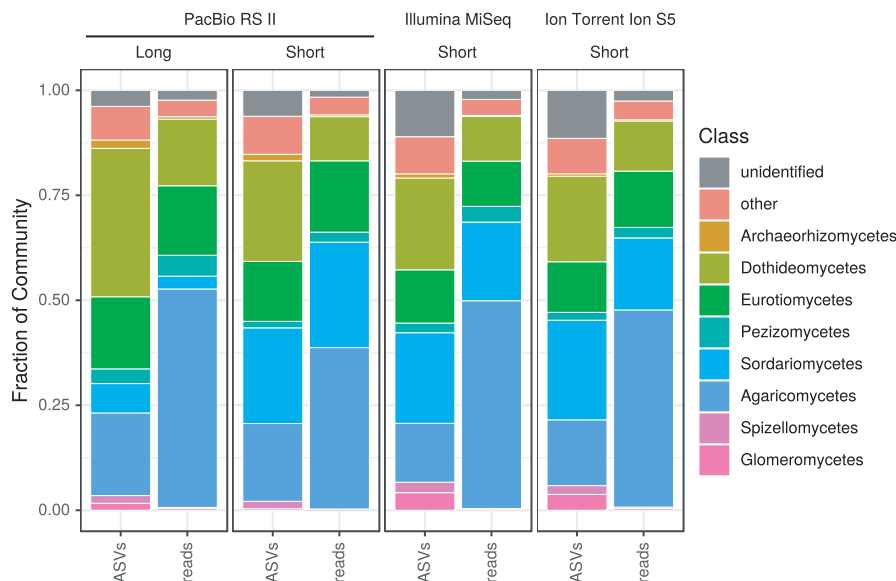

FIGURE 6 Taxonomic composition of fungal community at the class level. Values represent the fraction of all ASVs and reads which were assigned to kingdom Fungi. Assignments based on PHYLOTAX. Classes that represented <2% of reads and ASVs in all data sets are grouped together as 'other'

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 $p < .05$ for ranges of up to 2–3 m for the total recovered fungal community, and 1–2 m for the ECM fungal community (Figure S19). 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 positive correlation when re-sampling at the same locations one year later (i.e. spatial distance of 0 m, time lag of 1 year), for both the total recovered fungal and ECM fungal communities, although this result did not reach statistical significance for all sequencing strategies. This spatiotemporal correlation did not extend to a range of 1 m, and in fact, correlation was negative at a time lag of 1 year and distance of 1 m, indicating that samples collected 1 m apart in different years were more different than randomly selected pairs of samples. This negative correlation, which reached marginal statistical significance in the PacBio short-amplicon data set, was probably a statistical artefact.

In contrast to the Bray–Curtis distance, the weighted UniFrac distance showed very little spatial structure, with only the total recovered fungal community in the 1 m distance class showing a significant correlation at $p < .05$. No temporal correlation was found for the weighted UniFrac distance.

The best-fit spatial turnover ranges based on Bray–Curtis distance-decay curves calculated from different sequencing strategies range widely from 13 to 31 m for the total recovered fungal community and 12–42 m for the ECM fungal community (Figure 7,

Table S5). However, there was overlap of the 95% confidence intervals for all of the Bray–Curtis spatial ranges in both the total recovered fungal and ECM fungal communities, across amplicon libraries and sequencing technologies (Table S5), so no strong conclusion of variability between methods can be drawn. Although a distance-decay model was fit for the weighted UniFrac distance applied to the total recovered 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 (Table S5).

4 | DISCUSSION

4.1 | Reconstruction of long amplicons from denoised subregions

Sequencing depth in the long-amplicon PacBio data set was not sufficient to successfully denoise using standard protocols, given the amplicon length and diversity of the samples. Amplicon sequence variant recovery for long amplicons using DADA2 was dramatically improved from 12% to 76% of reads by denoising homologous subregions independently using our new 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 recovery of rare ASVs 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, and were much closer to saturation of their respective species accumulation curves (Figure 4). Operational taxonomic unit read counts were strongly correlated between technologies ($R^2 = .72$–$.82$), and even between primer pairs ($R^2 = .49$–$.62$, Figure S10b). This lends some support for the technical repeatability of abundance-based beta diversity measures in metabarcoding, although bias at the amplification stage still presents issues (Bellemain et al., 2010; Castaño et al., 2020; Kanagawa, 2003; Polz & Cavanaugh, 1998).
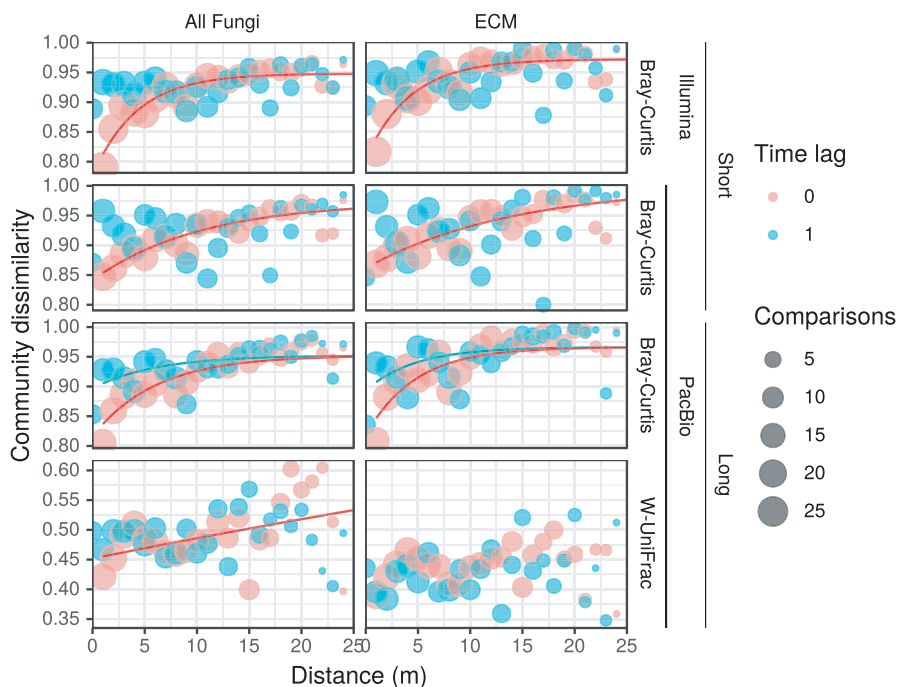

FIGURE 7 Distance-decay plot for community dissimilarities and spatiotemporal distance. Circles represent community data from short- (top two rows) and long- (bottom two rows) amplicon libraries, sequenced by Illumina MiSeq (top row) or PacBio RS II (bottom three rows). Community dissimilarities are calculated using the Bray–Curtis dissimilarity for all data sets (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 colour 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 data sets where the Mantel test indicated a significant relationship between community dissimilarity and spatial (for the 0 year time lag) or spatiotemporal (for the 1 year time lag) distance

DADA2 denoising may perform differently on different technologies (or perhaps sequencing runs) and indicated by the fact that clustering ASVs at 97% led to substantially higher correspondence between both the set of OTUs recovered from the same library by different technologies and the read counts for each OTU (Figure S10). ASV diversity appears to be artificially inflated in the Ion Torrent data set relative to the Illumina and PacBio data sets, which gave remarkably similar ASV richness after rarefaction, despite a difference of around 200× in unrarefied sequencing depth (Figures 2 and 3). This may be a result of the lower fraction of very high-quality reads in the Ion Torrent data set (Figure S6). We used options for DADA2 intended to improve performance on technologies, like Ion Torrent, with higher rates of homopolymer indel errors (Callahan, 2020b), but our results suggest that this still does not result in performance comparable to that which DADA2 achieves on Illumina sequences, for which it was developed (Callahan et al., 2016).

Although the longer read length capabilities of PacBio allow recovery of longer ITS2 sequences than the other two technologies, as has recently been demonstrated in mock communities (Castaño et al., 2020), in our data set from a natural community PacBio did not recover any reads from the short-amplicon library which were 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 observations by BF and NSY), 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 S7 and S9).

The long-amplicon data set included 20% unique taxa, even after clustering at 97% ITS2 similarity, indicating that the differences in the communities recovered are not due to small sequencing errors, but rather that the different primers amplify different parts of the community. The ITS4 primer used in the short-amplicon data set has known mismatches to Tulasnellaceae and Alveolata, while gITS7 also has mismatches for Tulasnellaceae (Tedersoo et al., 2015). ITS1 and LR5 match a much broader range of fungal and other eukaryote groups (Tedersoo et al., 2015). The alternate LR5-F primer (Tedersoo et al., 2008) would select against the nontarget Alveolata, at the expense of also having mismatches for the Tulasnellaceae. We assert that, for studies targeting ECM fungi in particular, more complete detection of groups with high rDNA variability such as *Tulasnella* (and ideally other Cantharellales) is worth the read depth spent on nontarget groups.

4.3 | Taxonomic identification

Assignment of ecological function to environmental fungal sequences is dependent on accurate taxonomic identification, especially at the genus level or below (Nguyen et al., 2016). However, different combinations of algorithms and reference data sets vary in their performance at confidently assigning taxonomy to sequences. Although RDP-LSU and Unite performed comparably at taxonomic placement of long-amplicon sequences, the Warcup database placed notably fewer sequences at all taxonomic levels for all data sets (Figure 5). This is probably due to two factors. First, the Warcup database does not include any nonfungi, so it cannot place any nonfungal sequences. Second, due to its low-density coverage of the fungal kingdom (18,000 sequences vs. 800,000 for Unite), it is likely that many ITS sequences, especially from uncultured tropical soil fungi, have no close match in the Warcup database, and so cannot be placed. RDP-LSU, which has even fewer sequences (8000 fungi plus 3000 other eukaryotes), is probably more successful due to higher sequence conservation in LSU. Heeger et al. (2019) also found that a more conserved region, in their case 5.8S, outperformed ITS at placing sequences without close database matches. Of the three algorithms tested, IDTAXA placed fewer sequences than RDPC or SINTAX with all databases, as expected given its more well-calibrated and conservative confidence scores (Murali et al., 2018), but this was particularly dramatic when paired with the Warcup database, where IDTAXA placed <25% of ASVs even to phylum.

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. AMPtk (Palmer et al., 2018) uses a strict consensus taxonomy between UTAX and SINTAX as an alternative when an initial BLAST search failed to give a hit with at least 97% sequence identity, but did not present results assessing the results of this approach. 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 (Figure 5). Inconsistent family- and genus-level assignments are particularly problematic because accurate assignment at these ranks is generally required for ecological guild assignment using FUNGuild.

For ASVs where a long-amplicon sequence is available, our novel PHYLOTAX algorithm uses relationships from a provided phylogenetic tree to resolve these disagreements. The effect was most pronounced for the PacBio long-amplicon data set, where 46% and 62% of reads were assigned to genus and family, respectively, by the strict consensus of methods, but PHYLOTAX increased this fraction to 73% and 81%. This led to a corresponding increase in the fraction of fungal reads assigned to a functional guild from 71% to 90% (Figure S18). For short-amplicon sequencing strategies, the improvement was more modest, because PHYLOTAX could only be applied for ITS2 ASVs with a match to one of the long-amplicon ASVs (last row of Table S2). Deeper long-amplicon sequencing would improve the coverage of long amplicons, allowing a greater fraction of short-amplicon ASVs to also be placed phylogenetically.

Because our data set was generated from environmental samples whose true taxonomic affinity is unknown, we were not able to

assess the accuracy of taxonomic assignments by any of the methods used here. Accuracy has been assessed using leave-one-out validation for the primary assignment algorithms (e.g. Edgar, 2018; Murali et al., 2018) and other consensus methods (e.g. Gdanetz et al., 2017; Somervuo et al., 2016), and similar work could be carried out in the future for PHYLOTAX.

4.4 | Turnover rate

Mantel correlograms based on the Bray–Curtis dissimilarity (Figure S19) 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. Previous work in Miombo woodland, a similar ecosystem to the Sudanian woodland in this study, found autocorrelation at ranges <10 m using Sanger sequencing of ECM root tips (Tedersoo, Sánchez-Ramírez, et al., 2018), which was their smallest distance class.

Distance-decay plots (Figure 7, Table S5) gave substantially longer autocorrelation distances. There was little variation in the results between the Illumina and long-amplicon PacBio data sets for both the total recovered fungal community and the ECM community, with best-fit estimates ranging from 12–18 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 a similar ECM woodland habitat in Benin, but based on Sanger sequencing of ECM root tips rather than HTS metabarcoding of bulk soil. We speculate that this discrepancy is due to an increased ability to detect spatially variable rare species using HTS.

For the short-amplicon data set, PacBio showed a spatial turnover range more than twice as long as showed by Illumina (Table S5) for both the total fungi and ECM communities, with wide confidence intervals. It is possible that the weaker fit for this data set, which also showed weaker autocorrelation in the Mantel correlogram, is due to low sequencing depth in the PacBio short-amplicon data set. The long-amplicon PacBio data set, with more than twice the read depth of the short-amplicon PacBio data set, gave spatial turnover distance results much closer to those from Illumina. This is consistent with our speculation that the longer spatial turnover range found by Bahram et al. (2013) is related to sequence sampling depth.

Year-to-year correlation was found for both the total recovered fungal and ECM communities in the long-amplicon data set (Figure S19). The spatiotemporal distance-decay fit estimated the temporal turnover range as 3.3 years for the total recovered fungal

community and 4.2 years for the ECM community, but with overlapping confidence intervals. This corresponds to a space-for-time substitution rate (i.e. ratio of a spatial distance to a time delay which results in equivalent community dissimilarity) of 5.4 and 3.3 month/year for the total recovered fungal community and ECM community, respectively. In a recent study, Kivlin and Hawkes (2020) reported a space-for-time substitution rate of 81 month/year (reported as 6.8 day/1.5 m) in the soil fungal community of a nonseasonal tropical forest in Costa Rica. However, comparison is obscured by different spatial and temporal sampling scales between the two studies. Year-to-year variation in ECM fungal communities, which we sampled, has been shown to be less than intra-annual variation (Bahram et al., 2015), as sampled by Kivlin and Hawkes (2020). Neither data set from the short-amplicon library showed significant temporal autocorrelation.

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 S19), the distance-decay plot in Figure 7 does not show any clear relationship. The functional fit showed poor convergence, with a 95% confidence interval for spatial range of 0–5470 m, indicating little evidence of spatial structure. This is probably because the majority of community turnover in this system, especially among ECM fungi, is between closely related species or individuals of the same species, while the presence of major clades (e.g. ECM lineages sensu Tedersoo et al., 2010) are more spatially constant. This is also reflected in the generally smaller sample-to-sample dissimilarities measured by UniFrac (0.4–0.6) as compared to Bray–Curtis (0.8–1.0) in Figure 7. UniFrac analysis would be more suited at larger spatial scales and/or larger ecological gradients.

5 | CONCLUSION

Contrary to our hypothesis, we found that Illumina and Ion Torrent sequencing of real environmental samples resulted in neither qualitative nor quantitative bias against longer ITS2 amplicons, relative to PacBio. Furthermore, although we did find an increased ability to detect certain fungal groups using the more universal ITS1-LR5 primer pair, the choice of amplicon and sequencing technology did not affect the results of the spatial analysis, provided sufficient sequencing depth. Alpha diversity estimates were strongly correlated between methods, but somewhat inflated for Ion Torrent relative to the other technologies. However, the addition of long-amplicon reads did allow the construction of a phylogenetic tree directly from the metabarcoding reads, which allowed refinement of taxonomic assignments using our new tool PHYLOTAX. DADA2 ASV yield was initially poor for long amplicons, 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 in order to improve taxonomic assignments, and cost-effective short-read sequencing to provide high sampling depth and sample number.

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.

DATA AVAILABILITY STATEMENT

- Trimmed, demultiplexed sequencing reads are deposited at the European Nucleotide Archive (ENA) under Project Accession no. PRJEB37385. Accession numbers are given in Files S1 and S2.
- Consensus ASV sequences from the PacBio and Illumina data sets are deposited at ENA under Project Accession no. PRJEN37385, Accession nos. HG995461–HG996435 (ASVs detected in the long amplicon data set) and FR984199–FR988025 (ASVs only detected in the short amplicon data set). Because of concern that the Ion Torrent data set may contain many ASVs with uncorrected sequencing errors, we elected not to submit ASVs, which were only detected in the Ion Torrent data set. However, these sequences, along with all other consensus ASVs, are archived in Dryad (Furneaux et al., 2021, https://doi.org/10.5061/dryad.6wwpzgmvf).
- Nucleotide alignment and ML tree are archived at Dryad (Furneaux et al., 2021, https://doi.org/10.5061/dryad.6wwpzgmvf).
- 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. Snapshots of the versions used in this study are archived in Dryad (Furneaux et al., 2021, https://doi.org/10.5061/dryad.6wwpzgmvf).
- FASTA-format files for the RDP fungal LSU training set, 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 study are archived in Dryad (Furneaux et al., 2021, https://doi.org/10.5061/dryad.6wwpzgmvf).

- Bioinformatics pipeline and analysis scripts are available at https://github.com/ouemefungi/oueme-fungi-transect and archived in Dryad (Furneaux et al., 2021, https://doi.org/10.5061/dryad.6wwpzgmvf).

ORCID

Brendan Furneaux 🄳 https://orcid.org/0000-0003-3522-7363
Mohammad Bahram 🄳 https://orcid.org/0000-0002-9539-3307
Anna Rosling 🄳 https://orcid.org/0000-0002-7003-5941
Nourou S. Yorou 🄳 https://orcid.org/0000-0001-6997-811X
Martin Ryberg 🄳 https://orcid.org/0000-0002-6795-4349

REFERENCES

Abarenkov, K., Somervuo, P., Nilsson, R. H., Kirk, P. M., Huotari, T., Abrego, N., & Ovaskainen, O. (2018). Protax-fungi: A web-based tool for probabilistic taxonomic placement of fungal internal transcribed spacer sequences. *New Phytologist*, 220(2), 517–525. https://doi.org/10.1111/nph.15301

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), 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), e00191-16. https://doi.org/10.1128/mSystems.00191-16

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2006). New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Applied and Environmental Microbiology*, 72(9), 5734–5741. https://doi.org/10.1128/AEM.00556-06

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. https://doi.org/10.1111/1365-2745.12120.

Bahram, M., Peay, K. G., & Tedersoo, L. (2015). Local-scale biogeography and spatiotemporal variability in communities of mycorrhizal fungi. *New Phytologist*, 205(4), 1454–1463. https://doi.org/10.1111/nph.13206

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: An in silico approach reveals potential PCR biases. *BMC Microbiology*, 10(1), 1–9. https://doi.org/10.1186/1471-2180-10-189

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

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

Bidartondo, M. I. (2008). Preserving accuracy in GenBank. *Science*, 319(5870), 1616. https://doi.org/10.1126/science.319.5870.1616a

Botnen, S. S., Davey, M. L., Halvorsen, R., & Kauserud, H. (2018). Sequence clustering threshold has little effect on the recovery of microbial community structure. *Molecular Ecology Resources*, 18(5), 1064–1076. https://doi.org/10.1111/1755-0998.12894

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4), 325–349. https://doi.org/10.2307/1942268

Brundrett, M. C. (2017). Global diversity and Importance of mycorrhizal and nonmycorrhizal plants. In L. Tedersoo (Ed.), *Biogeography of mycorrhizal symbiosis. Ecological studies* (Vol. *230*, pp. 533–556). Springer. https://doi.org/10.1007/978-3-319-56363-3_21

Buée, M., Reich, M., Murat, C., Morin, E., Nilsson, R. H., Uroz, S., & Martin, F. (2009). 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, *184*(2), 449–456. https://doi.org/10.1111/j.1469-8137.2009.03003.x.

Callahan, B. J. (2020a). *DADA2 ITS pipeline workflow (1.8)*. Retrieved September 18, 2020, from https://benjjneb.github.io/dada2/ITS_workflow.html

Callahan, B. J. (2020b). Frequently Asked Questions: Can I use dada2 with my 454 or Ion Torrent data? Retrieved September 18, 2020, from https://benjjneb.github.io/dada2/faq.html#can-i-use-dada2-with-my-454-or-ion-torrent-data

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. https://doi.org/10.1093/nar/gkz569

Castaño, C., Berlin, A., Durling, M. B., Ihrmark, K., Lindahl, B. D., Stenlid, J., Clemmensen, K. E., & Olson, Å. (2020). Optimized metabarcoding with Pacific Biosciences enables semi-quantitative analysis of fungal communities. *New Phytologist*, *228*(3), 1149–1158. https://doi.org/10.1111/nph.16731

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., & O'Toole, P. W. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, *38*(22), e200. https://doi.org/10.1093/nar/gkq873.

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

Divoll, T. J., Brown, V. A., Kinne, J., McCracken, G. F., & O'Keefe, J. M. (2018). Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows. *Molecular Ecology Resources*, *18*(3), 590–601. https://doi.org/10.1111/1755-0998.12770

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

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, *6*, e4652. https://doi.org/10.7717/peerj.4652

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

Fouquier, J., Rideout, J. R., Bolyen, E., Chase, J., Shiffer, A., McDonald, D., Knight, R., Caporaso, J. G., & Kelley, S. T. (2016). Ghost-tree: Creating hybrid-gene phylogenetic trees for diversity analyses. *Microbiome*, *4*(1), 11. https://doi.org/10.1186/s40168-016-0153-6

Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2021). Data for Long- and short-read metabarcoding reveal similar spatiotemporal structures in fungal communities. *Dryad*, https://doi.org/10.5061/dryad.6wwpzgmvf

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

Glassman, S. I., & Martiny, J. B. H. (2018). Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *mSphere*, *3*(4), e00148–18. https://doi.org/10.1128/mSphere.00148-18

Heeger, F., Wurzbacher, C., Bourne, E. C., Mazzoni, C. J., & Monaghan, M. T. (2019). Combining the 5.8S and ITS2 to improve classification of fungi. *Methods in Ecology and Evolution*, *10*(10), 1702–1711. https://doi.org/10.1111/2041-210X.13266

Hollmer, M. (2013). *Roche to close 454 Life Sciences as it reduces gene sequencing focus*. FierceBiotech. Retrieved September 18, 2020, from https://www.fiercebiotech.com/medical-devices/roche-to-close-454-life-sciences-as-it-reduces-gene-sequencing-focus

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

House, G. L., Ekanayake, S., Ruan, Y., Schütte, U. M. E., Kaonongbua, W., Fox, G., Ye, Y., & Bever, J. D. (2016). Phylogenetically structured differences in rRNA gene sequence variation among species of Arbuscular mycorrhizal fungi and their implications for sequence clustering. *Applied and Environmental Microbiology*, *82*(16), 4921–4930. https://doi.org/10.1128/AEM.00816-16

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., & Tappu, R. (2016). MEGAN community edition – Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology*, *12*(6), e1004957. https://doi.org/10.1371/journal.pcbi.1004957

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.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, *96*(4), 317–323. https://doi.org/10.1016/s1389-1723(03)90130-7

Karsch-Mizrachi, I., Takagi, T., Cochrane, G., & on behalf of the International Nucleotide Sequence Database Collaboration. (2018). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *46*(D1), D48–D51. https://doi.org/10.1093/nar/gkx1097

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

Kivlin, S. N., & Hawkes, C. V. (2020). Spatial and temporal turnover of soil microbial communities is not linked to function in a primary tropical forest. *Ecology*, *101*(4), e02985. https://doi.org/10.1002/ecy.2985

Kurtzman, C. P., & Robnett, C. J. (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie Van Leeuwenhoek*, *73*(4), 331–371. https://doi.org/10.1023/A:1001761008817

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., Øvreås, L., & Urich, T. (2012). CREST – Classification resources for environmental sequence tags. *PLoS One*, *7*(11), e49334. https://doi.org/10.1371/journal.pone.0049334

Legendre, P., & Legendre, L. F. J. (2012). *Numerical ecology* (3rd edn.). 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., & Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers–a user's guide. *New Phytologist*, *199*(1), 288–299.

Lindner, D. L., & Banik, M. T. (2011). Intragenomic variation in the ITS rDNA region obscures phylogenetic relationships and inflates estimates of operational taxonomic units in genus Laetiporus. *Mycologia*, *103*(4), 731–740. https://doi.org/10.3852/10331

Liu, K.-L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. A., & Xie, G. (2012). Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Applied and Environmental Microbiology*, *78*(5), 1523–1533. https://doi.org/10.1128/AEM.0682611

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), https://doi.org/10.1128/AEM.01368-19

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, *73*(5), 1576–1585. https://doi.org/10.1128/AEM.01996-06

Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.

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.

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

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

Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., & Nielsen, R. (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, *57*(5), 750–757. https://doi.org/10.1080/10635150802422316

Munch, K., Boomsma, W., Willerslev, E., & Nielsen, R. (2008). Fast phylogenetic DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3997–4002. https://doi.org/10.1098/rstb.2008.0169

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

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. https://doi.org/10.1016/j.funeco.2015.06.006

Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N., & Larsson, K.-H. (2008). Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for Molecular Species identification. *Evolutionary Bioinformatics*, *4*, EBO.S653. https://doi.org/10.4137/EBO.S653

Nilsson, R. H., Kristiansson, E., Ryberg, M., & Larsson, K.-H. (2005). Approaching the taxonomic affiliation of unidentified sequences in public databases – an example from the mycorrhizal fungi. *BMC Bioinformatics*, *6*(1), 1–7. https://doi.org/10.1186/14712105-6-178

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

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.-H., & Kõljalg, U. (2006). Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective. *PLoS One*, *1*(1), e59. https://doi.org/10.1371/journal.pone.0000059

Nilsson, R. H., Tedersoo, L., Abarenkov, K., Ryberg, M., Kristiansson, E., Hartmann, M., Schoch, C. L., Nylander, J. A. A., Bergsten, J., Porter, T. M., Jumpponen, A., Vaishampayan, P., Ovaskainen, O., Hallenberg, N., Bengtsson-Palme, J., Eriksson, K. M., Larsson, K.-H., Larsson, E., & Kõljalg, U. (2012). Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. *MycoKeys*, *4*, 37–63. https://doi.org/10.3897/mycokeys.4.3606.

Nilsson, R. H., Wurzbacher, C., Bahram, M., R. M. Coimbra, V., Larsson, E., Tedersoo, L., Eriksson, J., Duarte, C., Svantesson, S., Sánchez-García, M., Ryberg, M. K., Kristiansson, E., & Abarenkov, K. (2016). Top 50 most wanted fungi. *MycoKeys*, *12*, 29.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, *51*(11), 933–938. https://doi.org/10.1641/00063568(2001)051[0933:TEOTWA]2.0.CO;2

Öpik, M., Vanatoa, A., Vanatoa, E., Moora, M., Davison, J., Kalwij, J. M., Reier, Ü., & Zobel, M. (2010). The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist*, *188*(1), 223–241. https://doi.org/10.1111/j.1469-8137.2010.03334.x

Pacific Biosciences. (2019). *Consensus library and applications*. Retrieved March 11, 2019, from https://github.com/PacificBiosciences/unanimity

Palmer, J. M., Jusino, M. A., Banik, M. T., & Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*, *6*, e4925. https://doi.org/10.7717/peerj.4925

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.1365294X.2012.05739.x

Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multi-template PCR. *Applied and Environmental Microbiology*, 64(10), 3724–3730. https://doi.org/10.1128/AEM.64.10.3724-3730.1998

Rinaldi, A., Comandini, O., & Kuyper, T. W. (2008). Ectomycorrhizal fungal diversity: Separating 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

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. V., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing Microbial communities. *Applied and Environmental Microbiology*, 75(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

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

Smith, D. P., & Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS One*, 9(2), e90234. https://doi.org/10.1371/journal.pone.0090234

Smith, S. E., & Read, D. J. (2008). *Mycorrhizal Symbiosis* (3rd edn.). 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

Speranskaya, A. S., Khafizov, K., Ayginin, A. A., Krinitsina, A. A., Omelchenko, D. O., Nilova, M. V., Severova, E. E., Samokhina, E. N., Shipulin, G. A., & Logacheva, M. D. (2018). Comparative analysis of Illumina and Ion Torrent high-throughput sequencing platforms for identification of plant components in herbal teas. *Food Control*, 93, 315–324. https://doi.org/10.1016/j.foodcont.2018.04.040

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

Steidinger, B. S., Crowther, T. W., Liang, J., Van Nuland, M. E., Werner, G. D. A., Reich, P. B., Nabuurs, G. J., De-Miguel, S., Zhou, M., Picard, N., Herault, B., Zhao, X., Zhang, C., Routh, D., & Peay, K. G. (2019). Climatic controls of decomposition drive the global biogeography of forest-tree symbioses. *Nature*, 569(7756), 404–408. https://doi.org/10.1038/s41586-019-1128-0

Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology*, 21(1), 115. https://doi.org/10.1186/s13059-020-02023-1

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 metabar coding 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., Naadel, T., & Kõljalg, U. (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. https://doi.org/10.1111/j.1365-294X.2011.05145.x

Tedersoo, L., Jairus, T., Horton, B. M., Abarenkov, K., Suvi, T., Saar, I., & Kõljalg, U. (2008). Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. *The New Phytologist*, 180(2), 479–490.

Tedersoo, L., May, T. W., & Smith, M. E. (2010). Ectomycorrhizal lifestyle in fungi: Global diversity, distribution, and evolution of phylogenetic lineages. *Mycorrhiza*, 20(4), 217–263. https://doi.org/10.1007/s00572-009-0274-x

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*, 90(1), 135–159. 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), 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. *Applied and Environment Microbiology*, 73(16), 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.

Wong, R. G., Wu, J. R., & Gloor, G. B. (2016). Expanding the UniFrac Toolbox. *PLoS One*, 11(9), 315–322. 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

Yang, X., Chockalingam, S. P., & Aluru, S. (2013). A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*, 14(1), 56–66. https://doi.org/10.1093/bib/bbs015

Yorou, N. S., Koné, N. A., Guissou, M.-L., Guelly, A. K., Maba, D. L., Ekué, M. R. M., & De Kesel, A. (2014). Biodiversity and sustainable use of wild edible fungi in the Soudanian center of endemism: A plea for valorisation. In A. M. Bâ, K. L. McGuire, & A. G. Diédhiou (Eds.), *Ectomycorrhizal symbioses in tropical and neotropical forests* (pp. 241–269). CRC Press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.