Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline

Charlie Pauvert, Marc Buée, Valerie Laval, Véronique Edel-Hermann, Laure Fauchery, Angelique Gautier, Isabelle Lesur, Jessica Vallance, Corinne Vacher

Bioinformatics matters: the accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline

Charlie Pauvert[1], Marc Buée[2], Valérie Laval[3], Véronique Edel-Hermann[4], Laure Fauchery[2], Angélique Gautier[3], Isabelle Lesur[1,5], Jessica Vallance[6], Corinne Vacher[1]*

1. BIOGECO, INRA, Univ. Bordeaux, 33615 Pessac, France

2. INRA, Université de Lorraine, UMR IAM 1136, Laboratoire d'Excellence ARBRE, Centre INRA-Lorraine, 54280 Champenoux, France

3. BIOGER, INRA, 78850 Thiverval Grignon, France

4. Agroécologie, AgroSup Dijon, CNRS, INRA, Univ. Bourgogne Franche-Comté, 21000 Dijon, France

5. HelixVenture, 33700 Mérignac, France

6. SAVE, Bordeaux Sciences Agro, INRA, ISVV, Univ. Bordeaux, 33882 Villenave d'Ornon, France

* Corresponding author: Corinne Vacher

E-mail: corinne.vacher@inra.fr

Address: UMR BIOGECO, Université de Bordeaux, Allée Geoffroy St-Hilaire, Bât. B2, 33615 Pessac, France

Abstract

Fungal communities associated with plants and soil influence plant fitness and ecosystem functioning. They are frequently studied by metabarcoding approaches targeting the ribosomal internal transcribed spacer (ITS), but there is no consensus concerning the most appropriate bioinformatic approach for the analysis of these data. We sequenced an artificial fungal community composed of 189 strains covering a wide range of Ascomycota and Basidiomycota, to compare the performance of 360 software and parameter combinations. The most sensitive approaches, based on the USEARCH and VSEARCH clustering algorithms, detected almost all fungal strains but greatly overestimated the total number of strains. By contrast, approaches using DADA2 to detect amplicon sequence variants were the most effective for recovering the richness and composition of the fungal community. Our results suggest that analyzing single forward (R1) sequences with DADA2 and no filter other than the removal of low-quality and chimeric sequences is a good option for fungal community characterization.

Keywords

Bioinformatics; Environmental DNA; Fungi; Illumina MiSeq; Internal transcribed spacer (ITS); Metabarcoding; DADA2; USEARCH; VSEARCH; LULU

Introduction

Fungal communities associated with soil and plant tissues have a significant impact on plant fitness and ecosystem function (Dighton et al. 2005; Buée et al. 2009a; Rodriguez et al. 2009; Hacquard and Schadt 2015; Vandenkoornhuyse et al. 2015; Vacher et al. 2016a; Baldrian 2017). Identification of the fungal species present is a prerequisite for understanding these complex communities, but this task is challenging, due to the cryptic nature, microscopic characters and morphological variability of many fungal species (Hibbett and Taylor 2013; Yahr et al. 2016). Sequence-based taxonomic identification of fungal community members, or metabarcoding, has been the standard technique for the last 10 years (Buée et al. 2009b; Hibbett et al. 2009; Jumpponen and Jones 2009; Öpik et al. 2009; Cordier et al. 2012; Hibbett & Taylor 2013; Schmidt et al. 2013; Hibbett et al. 2016). The internal transcribed spacer (ITS) region is now recognized as the universal barcode for fungi (Schoch et al. 2012), and is conventionally used to sequence fungal communities (Lindahl et al. 2013; Bálint et al. 2014), in combination with large taxonomic reference databases, such as UNITE (Abarenkov et al. 2010; Kõljalg et al. 2013).

Despite their widespread use, metabarcoding approaches suffer from various biases due to the sampling process, molecular biology steps and bioinformatic analyses used (Lindahl et al. 2013; Schmidt et al. 2013; Bálint et al. 2016; Sommeria-Klein et al. 2016; Palmer and Jusino et al. 2018). These biases can prevent accurate recovery of the fungal community. For instance, the fungi identified may differ according to the barcode region chosen and the primers used for its amplification (Tedersoo et al. 2015), the sequencing platform (Motooka et al. 2017), the method used to assemble reads (Nguyen et al. 2015), the sequence clustering method (Cline et al. 2017; Halwachs et

al. 2017) and the filters subsequently applied to the operational taxonomic unit (OTU) table (Bokulich et al. 2013; Brown et al. 2015). Fungal ecologists thus face difficult decisions at every stage in metabarcoding studies (Alberdi et al. 2018).

Many pipelines have been developed that can be used for processing fungal ITS sequence data. These pipelines include MOTHUR (Schloss et al. 2009), QIIME (Caporaso et al. 2010), SCATA (Durling et al. 2011), CLOVR-ITS (White et al. 2013), VSEARCH (Rognes et al. 2016), FROGS (Escudié et al. 2017), PIPITS (Gweon et al. 2015) and DADA2 (Callahan et al. 2016). However, the variable length of fungal ITS sequences between taxa and high levels of sequence variability render analysis and interpretation of the data particularly difficult (Tedersoo et al. 2015; Halwachs et al. 2017; Palmer and Jusino et al. 2018). Fortunately, comprehensive guidelines have been developed, to help fungal community ecologists to make the most appropriate choices (Lindahl et al. 2013; Bálint et al. 2014; Bálint et al. 2016). These guidelines suggest, for example, that sequence clustering yields the best results with ITS extraction tools such as ITSx (Bengtsson-Palme et al. 2013; Lindahl et al. 2013; Bálint et al. 2014). The removal of rare OTUs, which may be artifacts, is also generally recommended (Bálint et al. 2016), but there is no consensus concerning the threshold number of sequences below which an OTU can be considered rare. The proposed thresholds range from 1 to 10 sequences (Brown et al. 2015) or depend on the relative abundance of OTUs (Bokulich et al. 2013). Fungal mock communities have also recently been used for the development of guidelines (Nguyen et al. 2015; Cline et al. 2017; Bakker 2018). Nguyen et al. (2015) showed, for example, that single forward reads could be used to recover all of the 25 well-amplified species of their mock community, whereas only 23 of these species were recovered with assembled paired-end reads. The clustering algorithm of USEARCH (Edgar 2010) has also been

recommended, based on the demonstration that it recovered the expected number of mock species (Cline et al. 2017).

There is currently no clear consensus in the scientific community concerning the most appropriate bioinformatic approach for analysis of the fungal ITS regions sequenced on Illumina MiSeq platforms. We aimed to fill this gap, by creating and sequencing a mock community of 189 Dikarya strains commonly found in agricultural and forest soils and in plant tissues. As advised by Nguyen et al. (2015), this mock community had a large taxonomic breadth and some genera were represented by several closely related strains (Fig. 1). We compared the ability of 360 combinations of bioinformatic softwares and parameters to recover the fungal strains present in this mock community, in the expected proportions. In particular, we investigated whether clustering-free software packages that identify exact sequence variants of amplicons (ASVs) rather than clustering similar sequences into OTUs (Callahan et al. 2016) outperformed conventional clustering approaches, by fully exploiting molecular barcode resolution (Callahan et al. 2017). We also tested novel post-clustering curation tools (Frøslev et al. 2017). We provide new guidelines, based on our results, for researchers using metabarcoding approaches for the analysis of fungal community richness and composition.

Materials and methods

Fungal mock community

The mock community consisted of an equimolar mixture of DNA extracted from 189 pure fungal strains isolated from soils, sporocarps or plant tissues. All the strains belonged to the superkingdom Holomycota (Tedersoo et al. 2018): 87 Ascomycota

strains, 99 Basidiomycota strains and 3 Mucoromycota strains, corresponding to 181 different species, 97 genera, 67 families, 30 orders and 11 classes. Altogether, 30 genera were represented by several species and 4 species were represented by several strains (Fig. 1 and Table S1).

Fungal DNA was obtained from the inner flesh of sporocarps, or from aerial mycelium scraped aseptically from the surface of pure cultures grown on PDA (Potato Dextrose Agar), MA (Malt Agar) or Pachlewski's medium (Martin et al. 1983). The mycelium was lyophilized for 24 h in an Edwards Modulyo 4K lyophilizer (Edwards, United Kingdom) and 100 mg of lyophilized mycelium was then placed in a Fast-Prep tube (2 mL) containing 130 mg glass beads (4.5 mm in diameter; Dutscher, France) and ground with a FastPrep® machine (MP Biomedicals, France) for 30 s at maximum shaking frequency. DNA was extracted with a DNeasy Plant Minikit (Qiagen, France), in accordance with the manufacturer's instructions, except that the incubation time was extended to 1 h at 65°C, and the volumes of buffers AP1 and P3 were doubled. DNA from all strains was quantified with a Qubit® 2.0 Fluorometer (Life Technologies, USA) and pooled in an equimolar mixture. Pooling was performed three times (replicates A, B and C). The fungal ITS1 region was amplified from each replicate with the ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3', Gardes and Bruns 1993) and ITS2 (5'-GCTGCGTTCTTCATCGATGC-3', White et al. 1990) primers. These primers are considered as universal fungal primers and are commonly used in fungal community analyses (Buée et al. 2009b; Cordier et al. 2012; Nguyen et al. 2015; Palmer and Jusino et al. 2018). They are known to amplify Ascomycota, Basidiomycota and Mucoromycota (Bellemain et al. 2010; Schoch et al. 2012). PCR was performed with a GeneAmp PCR System 2700 (Applied Biosystems, USA). The reaction mixture (20 µL final volume) consisted of 1x of PCR buffer, 0.56 mg mL^{-1} of bovine serum albumin

(A2153-10G, Sigma, USA), 0.2 mM of each dNTP, 0.2 µM of each primer, 0.05 U μL^{-1} *Taq* DNA polymerase (D1806, Sigma-Aldrich) and 5 ng of DNA template. The following cycling parameters were then used for amplification: enzyme activation at 94°C for 3 min; 35 cycles of denaturation at 94°C for 30 s, annealing at 53°C for 30 s, extension at 72°C for 45 s, and a final extension at 72°C for 10 min. The quality of the PCR products was checked by electrophoresis on 2% agarose gels. PCR products were purified (CleanPCR, MokaScience), multiplex identifiers and sequencing adapters were added, and library sequencing on an Illumina MiSeq platform (v3 chemistry, 2x250 bp) and sequence demultiplexing (with exact index search) were performed at the Get-PlaGe sequencing facility (Toulouse, France).

Full-length ITS sequences were also obtained by Sanger sequencing for the 189 fungal strains. PCR was performed with the ITS1F and ITS4 (5'-TCCTCCGCTTATTGATATGC-3', White et al. 1990) primers, with the same PCR mixture as described above. The PCR program consisted of an initial denaturation at 95°C for 3 min, followed by 30 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 30 s, and elongation at 72°C for 45 s. Sequencing reactions were performed by Genewiz (Takeley Essex, UK) and sequences from both strands were assembled with MultAlin (Corpet 1988) and manually curated.

Bioinformatic approaches

We analyzed the MiSeq sequences with 360 combinations of bioinformatic softwares and parameters (referred to hereafter as bioinformatic approaches). These approaches differed in (1) the paired-end read assembly algorithm used (*Assembly*), (2) the fungal ITS1 extraction method (*Extraction*), (3) the method of sequence variation analysis (*Variation*), (4) the treatment of chimeric sequences (*Chimeras*) and

(5) the final filtering of the community (*Filtering*). The various steps of the bioinformatic analyses are described below and in Figure 2.

(1) Paired-end read assembly algorithm (*Assembly*)

Two paired-end read assembly algorithms were compared (Fig. 2). Paired-end sequences were joined with the FASTQ-JOIN function of QIIME v1.8.0 (*Assembly*=FASTQ-JOIN_OL) or with PEAR v0.9.10 (*Assembly*=PEAR_OV) (Caporaso et al. 2010; Zhang et al. 2014). For each algorithm, three minimum overlapping lengths (OLs) between forward and reverse sequences were tested (50 bp, 100 bp and 150 bp). For the FASTQ-JOIN algorithm, no mismatch was allowed in the overlap region. We also considered the use of single forward (R1) sequences (*Assembly*=QUALITY_R1) (Fig. 2).

(2) Fungal ITS1 extraction (*Extraction*)

The ITS1 region was either extracted (*Extraction*=YES) from the high-quality sequences with ITSx v1.0.10 (Bengtsson-Palme et al. 2013), or not extracted (*Extraction*=NO) (Fig. 2). ITSx uses an alignment of conserved ribosomal genes to identify and delineate highly variable regions, such as the ITS1 region, accurately. The minimum length of the region between the binding sites for the ITS1F-ITS2 primers is about 100 bp (Motooka et al. 2017; Palmer and Jusino et al. 2018), but the ITS1 region *sensu stricto* (as defined by ITSx) is shorter, as it does not include portions of the 18S and 5.8S flanking regions. We thus discarded sequences of less than 100 bp in length in cases in which the ITS1 region was not extracted, or 50 bp in cases in which it was extracted.

(3) Sequence variation analysis (*Variation*)

Two clustering algorithms were compared (Fig. 2). Fungal ITS1 sequences displaying more than 97% similarity were clustered into OTUs with the popular USEARCH v7.0 program (*Variation*=USEARCH) (Edgar 2010) or with the open-source alternative VSEARCH v2.5.2 (*Variation*=VSEARCH) (Rognes et al. 2016). A similarity threshold of 97% was chosen as this threshold is commonly used in fungal metabarcoding studies (e.g. Bakker 2018; Durand et al. 2017) and has been shown to perform well on a mock community (Tedersoo et al. 2015), despite its tendency to aggregate closely related species (Ryberg 2015; Bálint et al. 2016). All other settings were left to default. Sequences with a Phred score greater than 30 over 75% of the read length were included in the clustering process. Quality filtering was performed with the QIIME script *split_libraries_fastq.py* (Fig. 2).

We also used the R package DADA2 (Callahan et al. 2016) to correct sequencing errors and to infer exact amplicon sequence variants (*Variation*=DADA2) (Fig. 2). We retained only reads with less than one expected error (given the quality scores; Edgar and Flyvbjerg 2015). Quality filtering was performed with the *fastqFilter* function. Quality data were lost during the extraction step. We therefore applied DADA2 only to the fungal ITS1 sequences not extracted with ITSx. This analysis strategy is different from that recommended in the DADA2 tutorial (http://benjjneb.github.io/dada2/tutorial.html). We, therefore, also included an approach adhering to the strategy described in the tutorial (*Assembly*=CUTADAPT_MERGED) (Fig. 2). Primers were removed from both forward and reverse reads, with Cutadapt v1.13 (Martin 2011). The forward and reverse reads were then truncated (at 200 bp and 180 bp, respectively) and we retained only reads with fewer than two expected errors (as in the default parameters of the *filterAndTrim*

function). Reads were merged after the inference of sequence variation as described in the tutorial.

(4) Treatment of chimeric sequences (*Chimeras*)

The chimeric sequences identified were either removed (*Chimeras*=Removed) or retained in the dataset (*Chimeras*=Retained) (Fig. 2). Detection with QIIME script *identify_chimeric_seqs.py* was performed on the demultiplexed reads before USEARCH clustering (Fig. 2), as recommended in the QIIME tutorial (http://qiime.org/scripts/identify_chimeric_seqs.html). Following VSEARCH clustering, we combined *de novo* and reference-based strategies for chimera detection. The *de novo* strategy used the UCHIME (Edgar et al. 2011) algorithm implemented in VSEARCH. The reference-based strategy used the ITS1-only UNITE-UCHIME dataset v7.2 (as of 2017-10-10) as a reference (Nilsson et al. 2015). Following DADA2 sequence variation analysis, chimeric sequences were removed with the *removeBimeraDeNovo* function, using the consensus option.

(5) Final filtering of the community (*Filtering*)

Finally, we filtered the OTU and ASV tables (Fig. 2). Five filtering methods were compared: *Filtering*=1 involved removing OTUs (or ASVs) composed of a single sequence; *Filtering*=10 involved removing OTUs (or ASVs) for which less than 10 sequences were obtained when all three replicates were considered; *Filtering*=RA involved removing OTUs (or ASVs) with a relative abundance lower than 0.005% of the total number of sequences; *Filtering*=LULU used the LULU curation algorithm (Frøslev et al. 2017) to collapse erroneous OTUs (or ASVs) into their parent OTUs; *Filtering*=All involved keeping all OTUs or ASVs, regardless of the number of sequences obtained. Representative sequences were assigned to taxa with the QIIME

script *assign_taxonomy.py,* with BLAST v2.2.22 (Altschul et al. 1990) and default QIIME parameters (e-value < 0.001; identity ≥ 90%) against the local database of Sanger sequences for the fungal strains of the mock community. The LULU curation algorithm was applied with both default settings and a set of parameters adjusted to the features of the mock community. Three parameters can be tuned in LULU: the minimum sequence similarity between a 'potential daughter' and its 'potential parent' (default 84%), the minimum ratio of parent OTU abundance to daughter OTU abundance in all samples (default 1) and their minimum co-occurrence rate across samples (default 95%). Increasing the first parameter is only advised when the barcode region has little variation or when few PCR and sequencing errors are expected, and changing the second parameter is generally not recommended (Frøslev et al. 2017). Therefore we tuned the third parameter. We lowered its value to 66.6% to account for the small number of samples in our study (3 replicate samples per bioinformatic approach).

Comparison criteria

We defined three criteria for comparisons of the ability of the bioinformatic approaches to recover the mock community: sensitivity, precision and compositional similarity. Sensitivity and precision were defined as the true positive rate *TP/(TP+FN)* and the positive predictive value *TP/(TP+FP)*, respectively, where *TP* is the number of true-positive OTUs (or ASVs), *FN* is the number of false-negative OTUs (or ASVs) and *FP* is the number of false-positive OTUs (or ASVs). True-positive OTUs corresponded to fungal strains present in the mock community and identified by the bioinformatic approach considered. False-negative OTUs corresponded to fungal strains present in the mock community but not detected by the bioinformatic approach considered. False-

positive OTUs corresponded to all other OTUs. If several OTUs were assigned to the same fungal strain of the mock community (i.e. 'split' OTUs), only the most abundant was considered to be a true-positive OTU, the others being considered false-positive OTUs. Compositional similarity was defined as the Bray-Curtis similarity (Odum 1950) between the community recovered and the mock community. It was calculated as 1-BC, where BC is the Bray-Curtis dissimilarity obtained from the *vegdist* function of the R *vegan* package (Oksanen et al. 2017), assuming a uniform distribution of sequences between the fungal strains in the mock community. The expected number of sequences per fungal strain in the mock community was calculated for each replicate and each bioinformatic approach as the total number of high-quality sequences (obtained after the *Filtering* step) divided by the total number of fungal strains in the mock community. Ribosomal RNA gene copy number information was not available for each strain and was not used to adjust the expected number of sequences.

All three criteria theoretically range between 0 and 1. They equal 1 when the algorithm successfully identifies all members of the mock community. However, maximum sensitivity may be below 1 if the sequences of some fungal strains are absent from the raw Illumina dataset. We, therefore, estimated the total number of strains present in the raw dataset, by aligning the forward and reverse MiSeq sequences with the ITS1 Sanger sequences, with a similarity threshold of 100% and an alignment length threshold of 90% of the length of the shorter sequence. Alignments were performed with VSEARCH (*--usearch_global)* (Rognes et al. 2016), using the following parameters: *--id 1 –userout --userfields query+target+qcov+tcov+id --maxaccepts 20 –top_hits_only*.

Results

Assessment of maximum sensitivity

The manually curated Sanger database contained the ITS1 sequences of the 189 fungal strains of the mock community (Fig. 1 and Table S1). Several strains had identical Sanger sequences for ITS1: two strains from the genus *Alternaria*, six from the genus *Botrytis* (*B. calthae*, *B. pseudocinerea*, *B. ranunculi* and three strains of *B. cinerea*), two strains from two different species of *Colletotrichum* (*C. destructivum* and *C. higginsianum*), two strains of *Craterellus cornucopioides*, four pairs of strains from the genus *Fusarium* (*F. acuminatum* and *F. avenaceum*, *F. langsethiae* and *F. sporotrichioides*, *F. oxysporum* and *F. commune*, *F. verticillioides* and another species of the *F. fujikuroi* species complex), two strains from two different species of *Lepista* (*L. irina* and *L. nuda*) and two strains from the genus *Zymoseptoria*. The Sanger database, therefore, contained 175 unique ITS1 sequences.

Only 160 of these 175 unique ITS1 sequences were detected in the raw Illumina MiSeq dataset (Table S1), suggesting that the other 15 strains were either not amplified by the ITS1F-ITS2 primer pair, not sequenced or were sequenced with errors. Eleven of these strains were detected in the Illumina data when the similarity threshold between Sanger and Illumina sequences was lowered to 93.5% (Table S1), suggesting that their apparent absence was caused by mismatches between the Sanger sequence and the Illumina sequences. Four of these eleven strains had ambiguous bases in the Sanger sequence, preventing a perfect match with Illumina sequences. These ambiguous bases might be due to the within-strain polymorphism of the ITS region that exists for some fungi (Fiers et al. 2011). The other four ITS1 sequences absent from the raw Illumina dataset came from the following species: *Lepiota clypeolaria*, *Mycena abramsii, M. galopus and Panellus stipticus*. The first species was successfully

amplified with the ITS1F-ITS4 primer pair before Sanger sequencing and possessed the exact sequence of the ITS2 primer, suggesting that its absence from the Illumina dataset was caused by DNA pooling biases rather than lack of amplification. In contrast, the sequence of the ITS2 primer was detected with some mismatches for the three last species. Their absence could be due to a lack of amplification by the ITS1F-ITS2 primer pair (Table S1). The maximum sensitivity attainable by any bioinformatic approach (that is, the maximum proportion of fungal strains that could actually be found) therefore ranged from 84.7% to 90.1%. The lower bound was obtained by considering that the raw Illumina dataset contained the ITS sequences of 160 strains (of 189), while the upper bound also took into account the nine strains sequenced with some errors.

Influence of *Assembly*, *Extraction*, *Variation*, *Chimera* and *Filtering* on sequencing data

(1) Influence of *Assembly*

In total, we obtained 143873 paired-end Illumina reads of 250 bp each (Table S2). We obtained 43352, 56406 and 44115 sequences for the three replicates. The mean quality of the forward (R1) reads was slightly higher than that of the reverse (R2) reads (35.59 *versus* 34.31, respectively) (Table S2).

The choice of paired-end read assembly algorithm strongly influenced the number, length and quality of the consensus sequences (Table S2 and Figure S1). FASTQ-JOIN retained on average 53.1% of the raw reads (whatever the minimum overlapping length), whereas PEAR retained on average 96.3% of the raw reads (Table S2). Mean sequence quality was also higher for PEAR than for FASTQ-JOIN (Table S2 and Figure S1). More than 90% of assembled reads passed the quality filter, whatever the

assembly algorithm used (Table S3). Thus, PEAR generated twice as many high-quality assembled reads as FASTQ-JOIN (Table S3).

(2) Influence of *Extraction*

The extraction of the ITS region with ITSx (Bengtsson-Palme et al. 2013) retained 97% to 99% of the assembled reads, but only 59% of the forward reads (Table S4). Reads were 118 nucleotides shorter, on average, after extraction. Most reads were between 200 and 300 bp long before extraction (Table S2), versus 100 to 200 bp after extraction (Table S4).

(3) Influence of *Variation and Chimera*

The total number of OTUs (or ASVs) varied by several orders of magnitude, depending on the method used to analyze sequence variation. For example, USEARCH identified 878 non-chimeric OTUs and 71 chimeric OTUs on average with the following parameters, *Assembly*=QUALITY_R1 and *Extraction*=No. VSEARCH identified 577 non-chimeric OTUs and 315 chimeric OTUs with the same parameters. DADA2 identified 157 non-chimeric ASVs and 40 chimeric ASVs. These striking differences were found for all mock replicates (Figure S2).

(4) Influence of *Filtering*

The final filtering step also strongly influenced the number of OTUs. For instance, removing OTUs with less than 10 sequences (*Filtering*=10) reduced the number of non-chimeric OTUs identified by USEARCH from 878 to 329, and the number of non-chimeric OTUs identified by VSEARCH from 577 to 257, for *Assembly*=QUALITY_R1 and *Extraction*=No. The LULU curation algorithm reduced even more the number of non-chimeric OTUs but, in contrast to other filtering methods, it did not lose any

sequence (Table S5). ASV tables were more robust than OTU tables to variations in the filtering methods (Table S5).

Comparison of the bioinformatic approaches on the basis of sensitivity, precision and compositional similarity criteria

Bioinformatic analyses generated 360 matrices containing the number of sequences per OTU (or ASV) for the three replicates (Fig. 2). The matrices differed considerably. For example, the mean number of OTUs (or ASVs) per replicate ranged from 57 to 1562, depending on the bioinformatic approach used (Table S6). Sensitivity, precision and compositional similarity values were calculated at replicate level. The ranking of approaches according to these criteria differed between the three mock replicates (Table S6), but the approaches that performed very well according to a given criterion for one replicate generally also performed well for the other two replicates. We, therefore, used the mean value of the criteria over the three replicates to rank the bioinformatic approaches.

The sensitivity of the 360 bioinformatic approaches ranged from 22% to 87% (Table S6). The 10 most sensitive approaches are listed in Table 1. All used USEARCH or VSEARCH to cluster sequences into OTUs and did not extract the fungal ITS1 region with ITSx. All these approaches produced very large numbers of OTUs, up to seven times more than the actual number of fungal strains in the mock community (Table 1 and Fig. 3A). They, thus, recovered most of the fungal strains of the mock community but also generated many false-positive OTUs (Fig. 3B). The precision of these approaches was, therefore, very low (Fig. 4). However, they displayed a high degree of compositional similarity to the mock community despite the large number of false-positive OTUs (Table 1).

Precision (the proportion of OTUs (or ASVs) corresponding to true strains), ranged from 9% to 98% (Table S6). The 10 most precise approaches are listed in Table 2. All used DADA2 to identify amplicon sequence variants (Fig. 4), did not extract the ITS region with ITSx and used the LULU curation algorithm. The 8 first approaches used assembled reads as input data (Table 2). Removing chimeras with DADA2 before LULU curation appeared to be unfavourable, as it slightly reduced the sensitivity of all these top-ranking approaches (Table 2). Adjusting the minimum co-occurrence threshold of the LULU algorithm did not influence the results (data not shown). Unlike the most sensitive approaches, the most precise approaches yielded fewer ASVs than there were fungal strains in the mock community (Table 2) and produced very few false-positive ASVs (Fig. 3B). However, they did not recover all mock strains. The most precise approach, P1, recovered only 36% of the fungal strains of the mock community (Table 2 and Fig. 3B).

Compositional similarity to the mock community ranged from 0.15 to 0.396 (Table S6). The 10 best approaches according to this criterion are listed in Table 3. Like the most precise approaches, these 10 approaches used DADA2 to identify amplicon sequence variants (Fig. 2) and did not extract the ITS region with ITSx. However, unlike the most precise approaches, most used non-assembled reads as input data and they did not use the LULU curation algorithm. The approach with the best performance according to the similarity criterion (Si1) used R1 reads as input data, retained chimeras and applied no filters to the final ASV table. This approach recovered 77.4% of the fungal strains from the mock community but had a relatively low precision (Table 3). In contrast, the Si5 approach, which used the same options but with the removal of chimeras, had a precision increased by 17%. This is because chimera removal efficiently discarded false positive ASVs, lowering their number from 51 to 13 (Fig. 3B).

As a side effect, chimera removal triggered the loss of 3 true positive ASVs, slightly reducing the sensitivity of the Si5 approach (Fig. 3B). The removal of primers (as recommended in the DADA2 tutorial) did not improve the performance of these two top-ranking approaches. It slightly lowered the precision of the Si1 approach (Table S7). The compositional similarities of the Si1 and Si5 approaches were 0.396 and 0.393, respectively (Table 3). These values were among the highest obtained, but were far from the maximal value of 1 indicating an exact match between the observed and expected community. This difference resulted from the huge variability in the number of sequences per ASV, contrasting with the expected uniform distribution of reads between fungal strains (Fig. S3). The expected number of reads for each fungal strain was then multiplied by the number of fungal strains with an identical ITS1 sequence (Fig. S3), which increased compositional similarity values (Table S6) but did not change the ranking of the bioinformatic approaches (Spearman $\rho = 0.99$; $p < 2.2e\text{-}16$).

Finally, comparison of the bioinformatic approaches revealed that some steps that are commonly recommended, such as ITS extraction and chimera removal, can have positive effects but also negative ones. For instance, the extraction of the ITS1 region with ITSx before USEARCH and VSEARCH clustering increased significantly precision but it decreased sensitivity (Fig. S4), suggesting that ITS extraction discarded some false-positive OTUs but also some true-positive OTUs. Similarly, bioinformatic approaches that kept chimeras after DADA2 sequence correction (Si1-Si4 in Table 3) were slightly more sensitive than approaches that remove chimeras (Si5-Si8 in Table 3), indicating that chimera removal discarded some true-positive OTUs. This negative effect of chimera removal also occurred in the USEARCH and VSEARCH pipelines, but to a lower extent (Fig. S2).

Discussion

Metabarcoding approaches have revolutionized fungal ecology over the last decade (Hibbett et al. 2009) and have become the gold standard for describing the richness and composition of communities and the networks of associations between community members (Bálint et al. 2016). They have been so successful that fungal ecologists are struggling to cope with the boom in sequencing platforms, bioinformatic pipelines, taxonomic databases and community analysis tools. Benchmark studies and methodological reviews are required to help them make the most appropriate choices (e.g. Lindahl et al. 2013; Bálint et al. 2016; Weiss et al. 2016; Pollock et al. 2018). In this study, we focused on one aspect of the metabarcoding approach, bioinformatic analysis, assessing its effect on the recovery of community richness and composition. We compared the ability of 360 bioinformatic approaches to recover a mock community of fungal strains commonly found in soils and plants and including 97 genera from subkingdom Dikarya. This mock community was much larger than the fungal mock communities analyzed in previous studies (Amend et al., 2010; Ihrmark et al., 2012; Nguyen et al. 2015; Taylor et al., 2016; Cline et al. 2017; Bakker 2018) and covered both the Ascomycota and Basidiomycota clades (Fig. 1).

We selected three criteria for comparing bioinformatic approaches: sensitivity, precision and compositional similarity to the mock community. The first two criteria are related to the number of OTUs (or ASVs) recovered and are commonly used in benchmark studies (see Weiss et al. 2016). The third takes relative abundance into account and has been used by Bakker (2018). We believe that this third criterion is very important, particularly if the fungal metabarcoding data are to be used to reconstruct fungal association or interaction networks for biocontrol (Poudel et al.

2016; Vacher et al. 2016b; Hassani et al. 2018) or biomonitoring applications (Bohan et al. 2017; Karimi et al. 2017; Derocles et al. 2018). Indeed, network inference requires the most accurate possible recovery of microbial species and their abundances (Faust and Raes 2012; Friedman and Alm 2012; Berry and Widder 2014; Weiss et al. 2016).

Our comparison revealed huge discrepancies between bioinformatic approaches, thereby confirming the importance of carefully selecting the most appropriate method for the analysis of fungal metabarcoding data (Nguyen et al. 2015; Cline et al. 2017; Anslan et al. 2018). The number of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) identified by the bioinformatic approaches compared ranged from 57 to 1562, even though there were only 189 strains in the mock community. These results confirm that fungal community analyses should not focus on absolute values of richness estimated from metabarcoding data, but rather on the relative changes in richness between samples (Cline et al. 2017). The percentage of fungal strains recovered by the bioinformatic approaches ranged from 22% to 87.5%. This second value may be considered a very good result, because we estimated the maximum sensitivity attainable by a bioinformatic approach, given our data, at 90.1%. Indeed, not all the strains in the mock community could be distinguished on the basis of their ITS1 sequences, and several strains were either not amplified at all or not accurately amplified. Our analyses revealed that four fungal species (*Lepiota clypeolaria*, *Mycena abramsii, M. galopus,* and *Panellus stipticus*) were absent from the sequence dataset (Fig. 1), suggesting a lack of amplification by the so-called "universal" primers (Bellemain et al. 2010; Tedersoo and Lindahl 2016) or a sequencing failure (Nguyen et al. 2015; Palmer and Jusino et al. 2018).

DADA2, a clustering-free software package (Callahan et al. 2016), effectively recovered the composition of the mock community. The top ten bioinformatic

approaches in terms of performance for the compositional similarity criterion all used DADA2 to identify amplicon sequence variants. The total number of ASVs generated by these 10 approaches ranged from 148 to 197, which was therefore of the same order of magnitude as the total number of strains in the mock fungal community (i.e. 189). We highlighted several options for increasing the efficiency of DADA2 for fungal metabarcoding datasets. Firstly, our results confirm that the use of single forward (R1) reads as input data is a good option (Nguyen et al. 2015). This made it possible to ensure that strains with longer ITS regions (such as those of the genus *Cantharellus*, for instance; Feibelman et al. 1994) were not excluded. Based on our results, we also recommend retaining the primers for fungal communities amplified with the ITS1F-ITS2 primer pair. Indeed, we found that primer removal did not improve the recovery of mock community composition. These findings may be accounted for by the absence of degenerate nucleotides in the ITS1F-ITS2 primer pair. Primer retention may be relevant in this case, because non-degenerate primers have no impact on the denoising step of DADA2. The merging of reads after sequence variation inference, as recommended in the DADA2 tutorial (http://benjjneb.github.io/dada2/tutorial.html), did not improve the recovery of the mock community either.

The Si5 approach represented one of the best trade-offs between the three selection criteria among the 360 bioinformatic approaches compared. The Si5 approach used single forward (R1) reads as input. Quality filtering, sequence variation analysis and chimera removal were performed with DADA2 (Callahan et al. 2016). ITS1 extraction (Bengtsson-Palme et al. 2013) and downstream OTU table filtering were not required. The Si5 approach recovered the ITS1 regions of 80 out of 87 Ascomycota strains, 83 out of 99 Basidiomycota strains and all 3 Mucoromycota strains (Fig. 1), suggesting that there was no detection bias against Ascomycota strains despite the intron insert

downstream of the ITS1F primer binding site that might impair their amplification (see Taylor et al. 2016). We recommend the use of this simple bioinformatic approach in ecological studies of fungal communities, for the following reasons: (i) it did not overestimate the number of fungal strains, (ii) it was among the ten best bioinformatic approaches in terms of recovery of the composition of the mock community and (iii) it performed very well according to the two other criteria used for comparison (precision and sensitivity). Based on these results, the Si5 approach appears to be an appropriate bioinformatic approach for studies involving whole-community profiling and network inference.

By contrast, the clustering algorithms of USEARCH (Edgar 2010) and VSEARCH (Rognes et al. 2016) should be favored in studies in which species detection is the main goal. These clustering algorithms generally overestimated the actual number of fungal strains, but were able to retrieve almost all detectable strains. Their sensitivity was close to the maximum value. The most sensitive approach, Se1, used single forward (R1) reads as input and clustered them with the VSEARCH algorithm. ITS1 extraction (Bengtsson-Palme et al. 2013), chimera removal and downstream OTU table filtering were not required. In general, our comparison revealed that the steps of ITS extraction and chimera removal can eliminate fungal strains that are actually present in the community and should not be systematically used. The second most sensitive approach, Se2, used the USEARCH clustering algorithm. These two highly sensitive bioinformatic approaches are potentially useful for the early detection of invasive species (Comtet et al. 2015), including fungal pathogens (Munck and Bonello 2018), for the detection of emerging pathogens accounting for the decline or death of host populations (Ricciardi et al. 2017), and for exploring environmental reservoirs of pathogens (Agtmaal et al. 2017). On the other hand, if the purpose of a study is to

focus only on fungal species present with high certainty (i.e. on a precise but incomplete community), then DADA2 and LULU (Frøslev et al. 2017) should be combined and applied to assembled sequences. The Pi3 approach, that merges reads after sequence variation inference as recommended in the DADA2 tutorial, seems to be a good compromise in this case.

Overall, our study highlights the importance of carefully selecting the bioinformatic approach to be used according to the objective of the metabarcoding study. Indeed, the ability of bioinformatic approaches to recover fungal strains and the relative abundances of the strains recovered varied greatly. Some approaches detected almost all strains of the mock community but overestimated community richness, whereas others retrieved the actual richness and composition of the mock community more accurately. The former are more appropriate for the detection of target species, whereas the latter are more appropriate for community ecology studies. However, none of the bioinformatic approaches compared recovered the mock community perfectly. In particular, none of the approaches found the expected distribution of sequences between fungal strains. This may be due to differences in the number of ribosomal RNA gene repeats between fungal species (Ganley and Kobayashi 2007), and imperfections in equimolar pooling of DNA samples, together with biased amplification for pooled species (Palmer and Jusino et al. 2018). Because of these biases, current fungal community analyses should not focus on the within-sample distribution of taxa abundance, but rather on the changes in taxa abundance between samples. Future methodological developments should focus on reducing biases caused by molecular biology steps (Nichols et al. 2018; Porter and Hajibabaei 2018) and on improving the bioinformatics pipelines to better recover the abundances of fungal strains. Our comparison of bioinformatics approaches could be extended, since

the 360 bioinformatic approaches compared here constitute only a small fraction of the approaches that could be used to analyze fungal metabarcoding data. Other approaches may give better results, and their ranking may vary with sequence data quality (Nguyen et al. 2015). Future bioinformatic approach comparisons should therefore be based on multiple mock communities sequenced independently. They could also include error-correction methods alternative to that of DADA2, such as UNOISE2 (Edgar 2016), or recent clustering approaches, such as OptiClust (Westcott and Schloss 2017) or SeekDeep (Hathaway et al. 2018), or consider reference-based clustering approaches (Cline et al. 2017; Halwachs et al. 2017; but see Westcott and Schloss, 2015). All the data required for the extension of our methodological comparison are provided.

Data Availability

The raw sequence data were deposited in Dataverse and are available in the FASTQ format at https://doi.org/10.15454/8CVWRR. The code is available as an archive at https://doi.org/10.15454/VKTWKR.

Acknowledgments

We thank Matthieu Barret, Martial Briand, Lucas Auer, Gregory Gambetta, Guilherme Martins, Frédéric Barraquand and Tania Fort for useful discussions on the first draft of the manuscript. We also thank the editor, Peter Kennedy and one anonymous reviewer for helpful comments of the submitted version. We are grateful to the INRA MEM metaprogramme (Meta-Omics of Microbial Ecosystems) for financial and scientific support. The mock community sequencing was funded by the INRA MEM MetaBAR project (PI: MB) and the bioinformatic analyses were performed as part of the INRA

MEM Learn-biocontrol project (PI: CV). Additional funding was received from the LABEX COTE (ANR-10-LABX-45) and the LABEX CEBA (ANR-10-LABX-25-01). CP's PhD grant was funded by the INRA and Bordeaux Sciences Agro (BSA). We thank the Genotoul sequencing facility (Get-PlaGe) for sequencing the mock community and the Genotoul bioinformatics facility (Bioinfo Genotoul) for providing computing and storage resources. We also thank Julie Sappa from Alex Edelman & Associates for English language revision.

Author contributions

CP performed the bioinformatic work, analyzed the results in accordance with the recommendations of IL and CV, and wrote the first draft of the article in collaboration with JV and CV. MB coordinated the design and sequencing of the mock community. MB, VL, VEH and LF provided fungal DNA for the mock community and performed the molecular biology work. AG and VL provided the Sanger sequence database. CV conceived the study in collaboration with MB, supervised the work and made a major contribution to the writing of the manuscript. All authors revised the manuscript.

Supplementary Materials

Table S1 - List of the fungal strains of the mock community and their ITS Sanger sequences.

Table S2 - Sequence summary statistics for the raw data and the assembled reads, before quality filtering.

Table S3 - Sequence summary statistics for the raw data and the assembled reads, after quality filtering.

Table S4 - Number and length of sequences after extraction of the ITS1 region with ITSx.

Table S5 - Number of OTUs and sequences in the final OTU table according to the method of sequence variation analysis and the filtering option.

Table S6 - Sensitivity, precision and similarity values for all bioinformatic approaches.

Table S7 - Effect of primer removal on the sensitivity, precision and similarity values for the Si1 and Si5 approaches.

Figure S1 - Mean quality score along raw reads and assembled reads before quality filtering.

Figure S2 - Total number of OTUs per replicate as a function of the method for analyzing sequence variation.

Figure S3 - Variability of sequence counts among ASVs for the Si5 approach.

Figure S4 – Sensitivity, precision and compositional similarity as a function of the ITS extraction step.

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U, 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* **186**: 281–285.

Agtmaal M van, Straathof A, Termorshuizen A, Teurlincx S, Hundscheid M, Ruyters S, Busschaert P, Lievens B, Boer W de, 2017. Exploring the reservoir of potential fungal plant pathogens in agricultural soil. *Applied Soil Ecology* **121**: 152–160.

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K, 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution* **9**: 134–147.

Altschul SF, Gish W, Miller W, Myers, E.W. and Lipman, DJ, 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–10.

Amend AS, Seifert KA, Bruns TD, 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**: 5555–5565.

Anslan S, Nilsson H, Wurzbacher C, Baldrian P, Tedersoo L, Bahram M, 2018. Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *PeerJ Preprint.*

Bakker MG, 2018. A fungal mock community control for amplicon sequencing experiments. *Molecular Ecology Resources* **18**: 541–556.

Baldrian P, 2017. Forest microbiome: diversity, complexity and dynamics. *FEMS Microbiology Reviews* **41**: 109–130.

Bálint M, Schmidt P-A, Sharma R, Thines M, Schmitt I, 2014. An Illumina metabarcoding pipeline for fungi. *Ecology and Evolution* **4**: 2642–2653.

Bálint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, O'Hara RB, Öpik M, Sogin ML, Unterseher M, Tedersoo L, 2016. Millions of reads, thousands of taxa:

microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews* **40**: 686–700.

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**: 189.

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger I, de Sousa F, Amend AS, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson KM, Vik U, Veldre V, Nilsson RH, 2013. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution* **4**: 914–919.

Berry D, Widder S, 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* **5**: 219.

Bohan DA, Vacher C, Tamaddoni-Nezhad A, Raybould A, Dumbrell AJ, Woodward G, 2017. Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in Ecology and Evolution* **32**: 477–487.

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG, 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* **10**: 57–59.

Brown SP, Veach AM, Rigdon-Huss AR, Grond K, Lickteig SK, Lothamer K, Oliver AK, Jumpponen A, 2015. Scraping the bottom of the barrel: are rare high throughput sequences artifacts? *Fungal Ecology* **13**: 221–225.

Buée M, Boer W, Martin F, Overbeek L, Jurkevitch E, 2009a. The rhizosphere zoo: An overview of plant-associated communities of microorganisms, including phages, bacteria, archaea, and fungi, and of some of their structuring factors. *Plant and Soil* **321**: 189–212.

Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F, 2009b. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449–456.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP, 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**: 581–583.

Callahan BJ, McMurdie PJ, Holmes SP, 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* 11: 2639–43.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R, 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335–336.

Cline LC, Song Z, Al-Ghalith GA, Knights D, Kennedy PG, 2017. Moving beyond *de novo* clustering in fungal community ecology. *New Phytologist* **216**: 629-634.

Comtet T, Sandionigi A, Viard F, Casiraghi M, 2015. DNA (meta)barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens. *Biological Invasions* **17**: 905–922.

Cordier T, Robin C, Capdevielle X, Desprez-Loustau M-L, Vacher C, 2012. Spatial variability of phyllosphere fungal assemblages: genetic distance predominates over geographic distance in a European beech stand (*Fagus sylvatica*). *Fungal Ecology* **5**: 509–520.

Corpet F, 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* **16**: 10881–10890.

Derocles SAP, Bohan DA, Dumbrell AJ, Kitson JJN, Massol F, Pauvert C, Plantagenest M, Vacher C, Evans DM, 2018. Biomonitoring for the 21st Century: integrating next-generation sequencing into ecological network analysis. *Advances in Ecological Research* **58**: 1–62.

Dighton J, White JF, Oudemans P, 2005. *The Fungal community: its organization and role in the ecosystem*. Taylor & Francis, Boca Raton, FL.

Durand A, Maillard F, Foulon J, Gweon HS, Valot B, Chalot M, 2017. Environmental metabarcoding reveals contrasting belowground and aboveground fungal communities from poplar at a Hg phytomanagement site. *Microbial Ecology* **74**: 795–809.

Durling MB, Clemmensen KE, Stenlid J, Lindahl B, 2011. SCATA - An efficient bioinformatic pipeline for species identification and quantification after high-throughput sequencing of tagged amplicons. Available from https://scata.mykopat.slu.se/.

Edgar RC, 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R, 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Edgar RC, Flyvbjerg H, 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**: 3476–3482.

Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv, https://doi.org/10.1101/081257

Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G, 2017. FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics* **34**: 1287-1294.

Faust K, Raes J, 2012. Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10**: 538–50.

Feibelman T, Bayman P, Cibula WG, 1994. Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. *Mycological Research* **98**: 614-618

Fiers M, Edel-Hermann V, Héraud C, Gautheron N, Chatot C, Hingrat YL, Bouchek-Mechiche K, Steinberg C, 2011. Genetic diversity of *Rhizoctonia solani* associated with potato tubers in France. *Mycologia* **103**: 1230–1244.

Friedman J, Alm EJ, 2012. Inferring correlation networks from genomic survey data. *PLoS Computational Biology* **8**: e1002687.

Ganley ARD, Kobayashi T, 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* **17**: 184–191.

Gardes M, Bruns TD, 1993. ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Molecular Ecology* **2**: 113–118.

Gweon HS, Oliver A, Taylor J, Booth T, Gibbs M, Read DS, Griffiths RI, Schonrogge K, 2015. PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods in Ecology and Evolution* **6**: 973–980.

Hacquard S, Schadt CW, 2015. Towards a holistic understanding of the beneficial interactions across the *Populus* microbiome. *New Phytologist* **205**: 1424–30.

Halwachs B, Madhusudhan N, Krause R, Nilsson RH, Moissl-Eichinger C, Högenauer C, Thallinger GG, Gorkiewicz G, 2017. Critical issues in mycobiota analysis. *Frontiers in Microbiology*, online. https://doi.org/10.3389/fmicb.2017.00180

Hassani MA, Durán P, Hacquard S, 2018. Microbial interactions within the plant holobiont. *Microbiome* **6**: 58.

Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA, 2018. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Research* **46**: e21.

Hibbett DS, Ohman A, Kirk PM, 2009. Fungal ecology catches fire. *New Phytologist* **184**: 279–282.

Hibbett DS, Taylor JW, 2013. Fungal systematics: is a new age of enlightenment at hand? *Nature Reviews Microbiology* **11**: 129–133.

Hibbett D, Abarenkov K, Kõljalg U, Öpik M, Chai B, Cole J, Wang Q, Crous P, Robert V, Helgason T, Herr JR, Kirk P, Lueschow S, O'Donnell K, Nilsson RH, Oono R, Schoch C, Smyth C, Walker DM, Porras-Alfaro A, Taylor JW, Geiser DM, 2016. Sequence-based classification and identification of fungi. *Mycologia* **108**: 1049–1068.

Ihrmark K, Bödeker ITM, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandström-Durling M, Clemmensen KE, Lindahl BD, 2012. New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology* **82**: 666–677.

Jumpponen A, Jones KL, 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* **184**: 438–448.

Karimi B, Maron PA, Chemidlin-Prevost Boure N, Bernard N, Gilbert D, Ranjard L, 2017. Microbial diversity and ecological networks as indicators of environmental quality. *Environmental Chemistry Letters* **15**: 265–281.

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns TD, Bengtsson□Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lücking R, Martín MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott JA, Senés C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiss M, Larsson K-H, 2013. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* **22**: 5271–5277.

Lindahl BD, Nilsson RH, 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**: 288–299.

Martin F, Msatef Y, Botton B. 1983. Nitrogen assimilation in mycorrhizas. I. Purification and properties of the nicotinamide adenine dinucleotide phosphate□specific glutamate dehydrogenase of the ectomycorrhizal fungus *Cenococcum graniforme*. *New Phytologist*, **93**: 415-422.

Martin M, 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10.

Motooka D, Fujimoto K, Tanaka R, Yaguchi T, Gotoh K, Maeda Y, Furuta Y, Kurakawa T, Goto N, Yasunaga T, Narazaki M, Kumanogoh A, Horii T, Iida T, Takeda K, Nakamura S, 2017. Fungal ITS1 deep-sequencing strategies to reconstruct the composition of a 26-species community and evaluation of the gut mycobiota of healthy Japanese individuals. *Frontiers in Microbiology*, online. https://doi.org/10.3389/fmicb.2017.00238

Munck IA, Bonello P, 2018. Modern approaches for early detection of forest pathogens are sorely needed in the United States. *Forest Pathology* **0**: e12445.

Nguyen NH, Smith D, Peay K, Kennedy P, 2015. Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* **205**: 1389–1393.

Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE, Shapiro B, 2018. Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources* **18**: 927–939.

Nilsson RH, Tedersoo L, Ryberg M, Kristiansson E, Hartmann M, Unterseher M, Porter TM, Bengtsson-Palme J, Walker DM, de Sousa F, Gamper HA, Larsson E, Larsson K-H, Kõljalg U, Edgar RC, Abarenkov K, 2015. A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequencing efforts. *Microbes and Environments* **30**: 145–150.

Odum E, 1950. Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology* **31**: 587–605.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H, 2017. vegan: Community Ecology Package.

Öpik M, Metsis M, Daniell TJ, Zobel M, Moora M, 2009. Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist* **184**: 424–437.

Palmer JM, Jusino MA, Banik MT, Lindner DL, 2018. Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ* **6**: e4925.

Pollock J, Glendinning L, Wisedchanwet T, Watson M, 2018. The madness of microbiome: attempting to find consensus "best practice" for 16s microbiome studies. *Applied and Environmental Microbiology* **84**: e02627-17.

Porter TM, Hajibabaei M, 2018. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology* **27**: 313–338.

Poudel R, Jumpponen A, Schlatter DC, Paulitz TC, Gardener BBM, Kinkel LL, Garrett KA, 2016. Microbiome networks: a systems framework for identifying candidate microbial assemblages for disease management. *Phytopathology* **106**: 1083-1096.

Ricciardi A, Blackburn TM, Carlton JT, Dick JTA, Hulme PE, Iacarella JC, Jeschke JM, Liebhold AM, Lockwood JL, MacIsaac HJ, Pyšek P, Richardson DM, Ruiz GM, Simberloff D, Sutherland WJ, Wardle DA, Aldridge DC. 2017. Invasion science: a horizon scan of emerging challenges and opportunities. *Trends in Ecology & Evolution* **32** (6): 464-474,

Rodriguez RJ, White Jr JF, Arnold AE, Redman RS, 2009. Fungal endophytes: diversity and functional roles: *New Phytologist* **182**: 314–330.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F, 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

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**: 5770–5777.

Schmidt P-A, Bálint M, Greshake B, Bandow C, Römbke J, Schmitt I, 2013. Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry* **65**: 128–132.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJV, Weber CF, 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**: 7537–7541.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Consortium FB, List FBCA, Bolchacova E, Voigt K, Crous PW, Miller AN, Wingfield MJ, Aime MC, An K-D, Bai F-Y, Barreto RW, Begerow D, Bergeron M-J, Blackwell M, Boekhout T, Bogale M, Boonyuen N, Burgaz AR, Buyck B, Cai L, Cai Q, Cardinali G, Chaverri P, Coppins BJ, Crespo A, Cubas P, Cummings C, Damm U, Beer ZW de, Hoog GS de, Del-Prado R, Dentinger B, Diéguez-Uribeondo J, Divakar PK, Douglas B, Dueñas M, Duong TA, Eberhardt U, Edwards JE, Elshahed MS, Fliegerova K, Furtado M, García MA, Ge Z-W, Griffith GW, Griffiths K, Groenewald JZ, Groenewald M, Grube M, Gryzenhout M, Guo L-D, Hagen F, Hambleton S, Hamelin RC, Hansen K, Harrold P, Heller G, Herrera C, Hirayama K, Hirooka Y, Ho H-M, Hoffmann K, Hofstetter V, Högnabba F, Hollingsworth PM, Hong S-B, Hosaka K, Houbraken J, Hughes K, Huhtinen S, Hyde KD, James T, Johnson EM, Johnson JE, Johnston PR, Jones EBG, Kelly LJ, Kirk PM, Knapp DG, Kõljalg U, Kovács GM, Kurtzman CP, Landvik S, Leavitt SD, Liggenstoffer AS, Liimatainen K, Lombard L, Luangsa-ard JJ, Lumbsch HT, Maganti H, Maharachchikumbura SSN, Martin MP, May TW, McTaggart AR, Methven AS, Meyer W, Moncalvo J-M, Mongkolsamrit S, Nagy LG, Nilsson RH, Niskanen T, Nyilasi I, Okada G, Okane I, Olariaga I, Otte J, Papp T, Park D, Petkovits T, Pino-Bodas R, Quaedvlieg W, Raja HA, Redecker D, Rintoul TL, Ruibal C, Sarmiento-Ramírez JM, Schmitt I, Schüßler A, Shearer C, Sotome K, Stefani FOP, Stenroos S, Stielow B, Stockinger H, Suetrong S, Suh S-O, Sung G-H, Suzuki M, Tanaka K, Tedersoo L, Telleria MT, Tretter E, Untereiner WA, Urbina H, Vágvölgyi C, Vialle A, Vu TD, Walther G, Wang Q-M, Wang Y, Weir BS, Weiß M, White MM, Xu J, Yahr R, Yang ZL, Yurkov A, Zamora J-C, Zhang N, Zhuang W-Y, Schindel D, 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences* **109**: 6241–6246.

Sommeria-Klein G, Zinger L, Taberlet P, Coissac E, Chave J, 2016. Inferring neutral biodiversity parameters using environmental DNA data sets. *Scientific Reports* **6** : 35644.

Taylor DL, Walters WA, Lennon NJ, Bochicchio J, Krohn A, Caporaso JG, Pennanen T, 2016. Accurate Estimation of Fungal Diversity and Abundance through Improved Lineage-Specific Primers Optimized for Illumina Amplicon Sequencing. *Applied and Environmental Microbiology* **82**: 7217–7226.

Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson H, Hildebrand F, Bork P, Abarenkov K, 2015. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* **10**: 1–43.

Tedersoo L, Lindahl B, 2016. Fungal identification biases in microbiome projects. *Environmental Microbiology Reports* **8**: 774–779.

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**: 135–159.

Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE, 2016a. The phyllosphere: microbial jungle at the plant-climate interface. *Annual Reviews in Ecology Evolution and Systematics* **47** : 1–24.

Vacher C, Tamaddoni-Nezhad A, Kamenova S, Peyrard N, Moalic Y, Sabbadin R, Schwaller L, Chiquet J, Alex Smith M, Vallance J, Fievet V, Jakuschkin B, Bohan, DA, 2016b. Learning ecological networks from next-generation sequencing data. *Advances in Ecological Research* **54**: 1–39.

Vandenkoornhuyse P, Quaiser A, Duhamel M, Le Van A, Dufresne A, 2015. The importance of the microbiome of the plant holobiont. *New Phytologist* **206**: 1196–1206.

Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ, Birmingham A, Cram JA, Fuhrman JA, Raes J, Sun F, Zhou J, Knight R, 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal* **10**: 1669–1681.

Westcott SL, Schloss PD, 2015. *De novo* clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**: e1487.

Westcott SL, Schloss PD, 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2: e00073-17.

White JR, Maddox C, White O, Angiuoli SV, Fricke WF, 2013. CloVR-ITS: Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. *Microbiome* **1**: 6.

White TJ, Bruns T, Lee S, Taylor J, 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protocols: A Guide to Methods and Applications* 315–322.

Yahr R, Schoch CL, Dentinger BTM, 2016. Scaling up discovery of hidden diversity in fungi: impacts of barcoding approaches. *Philosophical Transactions of the Royal Society of London Series B* **371**: 20150336.

Zhang J, Kobert K, Flouri T, Stamatakis A, 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**: 614–620.

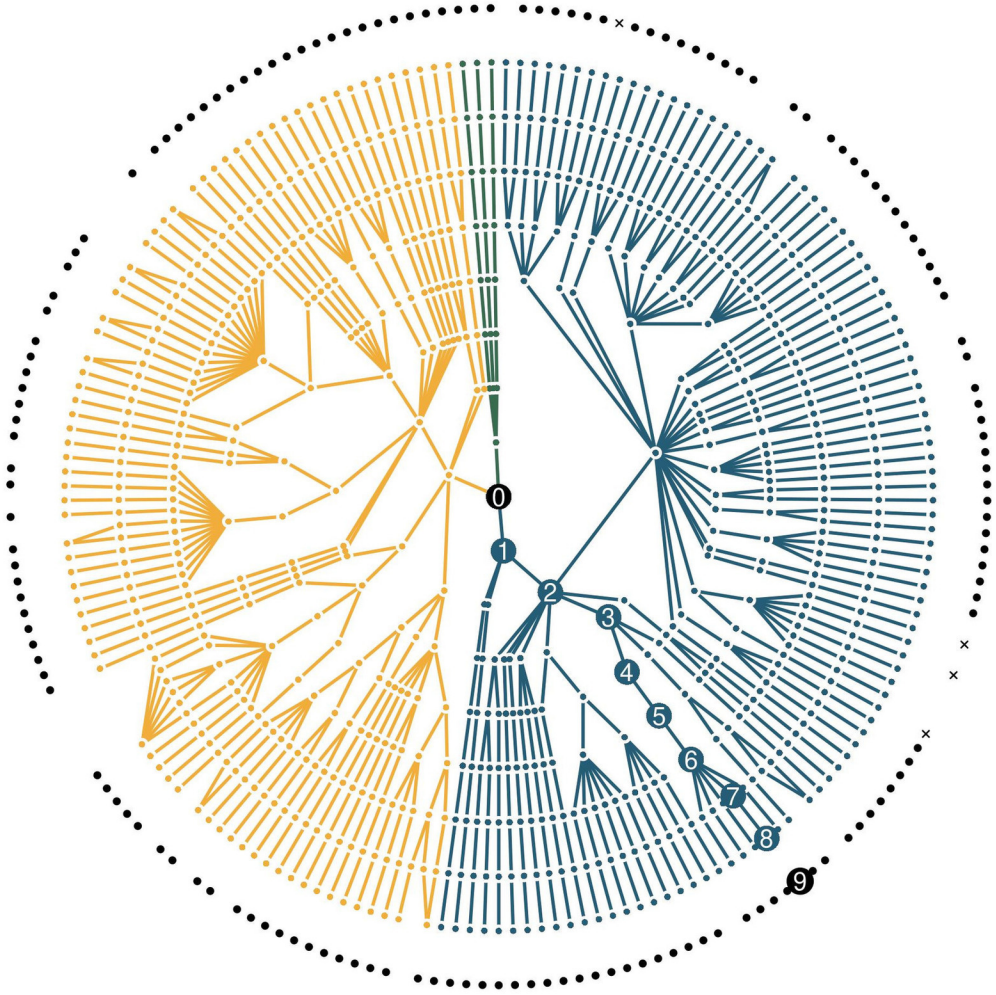
Figure Legends

Figure 1 – Taxonomic composition of the artificial fungal community. The terminal nodes of the tree are the ITS1 sequences (n = 175) of the fungal strains (n = 189) that constitute the mock community. The ITS1 sequences that were not present in the raw Illumina dataset are indicated with a cross and those that were recovered by the recommended bioinformatic approach (Si5, Table 3) are indicated with a black circle.

Figure 2 – Overview of the 360 bioinformatic approaches compared in this study. The Illumina MiSeq sequences were (1) assembled with FASTQ-JOIN (Caporaso et al. 2010) or PEAR (Zhang et al. 2014) with three minimum overlapping lengths (50 bp, 100 bp or 150 bp), or not assembled. In this latter case, single forward (R1) reads were used. After quality filtering, (2) the ITS1 region was extracted from the reads with ITSx (Bengtsson-Palme et al. 2013), or not extracted. (3) Sequence variations were then analyzed with DADA2 (Callahan et al. 2016), USEARCH (Edgar 2010) or VSEARCH (Rognes et al. 2016) and (4) chimeras were either retained or removed. (*: for USEARCH, chimera detection was performed before clustering). Finally, (5) the datasets were either filtered by removing rare or erroneous OTUs (or ASVs), or left unfiltered. Filtering thresholds (T) were based on the number of sequences per OTU, or on their relative abundance (RA), or OTU curation was performed using the LULU algorithm (Frøslev et al. 2017). §: When DADA2 was used, an alternative method of read processing (CUTADAPT_MERGED) was included.

Figure 3 – Richness estimates for the top four approaches. (A) Total number of OTUs (or ASVs) retrieved by the most sensitive approach (Se1; Table 1), the most precise approach (P1; Table 2), the approach with the best performance in terms of compositional similarity to the mock community (Si1; Table 3) and the bioinformatic approach recommended in this study (Si5; Table 3). The black horizontal line indicates the expected richness. (B) Total number of OTUs (or ASVs) per bioinformatic approach depending on OTU (or ASV) category. TP = true positives, FN = false negatives and FP = false positives. Results were averaged over the three replicates and rounded for clarity.

Figure 4 – Values of precision and (A) sensitivity or (B) compositional similarity to the mock fungal community, for all 360 bioinformatic approaches. Each dot corresponds to the mean value obtained for an approach over the three replicates. The methods used to analyze sequence variation (DADA2, USEARCH or VSEARCH) are highlighted with different colors and symbols. Se1 (Table 1), P1 (Table 2) and Si1 (Table 3) correspond to the most sensitive approach, the most precise approach and the approach with the best performance in terms of compositional similarity to the mock community, respectively. The bioinformatic approach recommended in this study is Si5 (Table 3).



Phylum:

Ascomycota

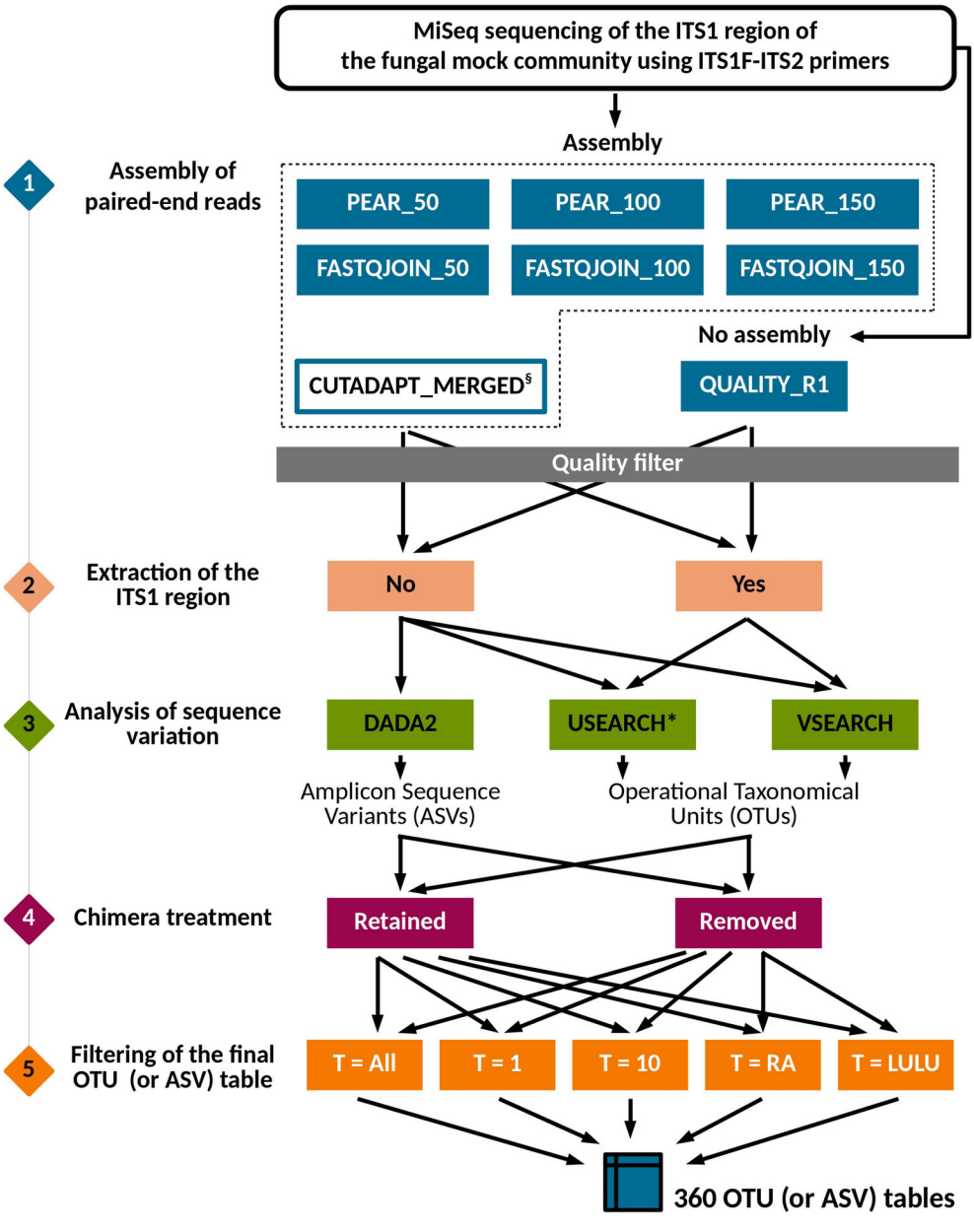
Basidiomycota

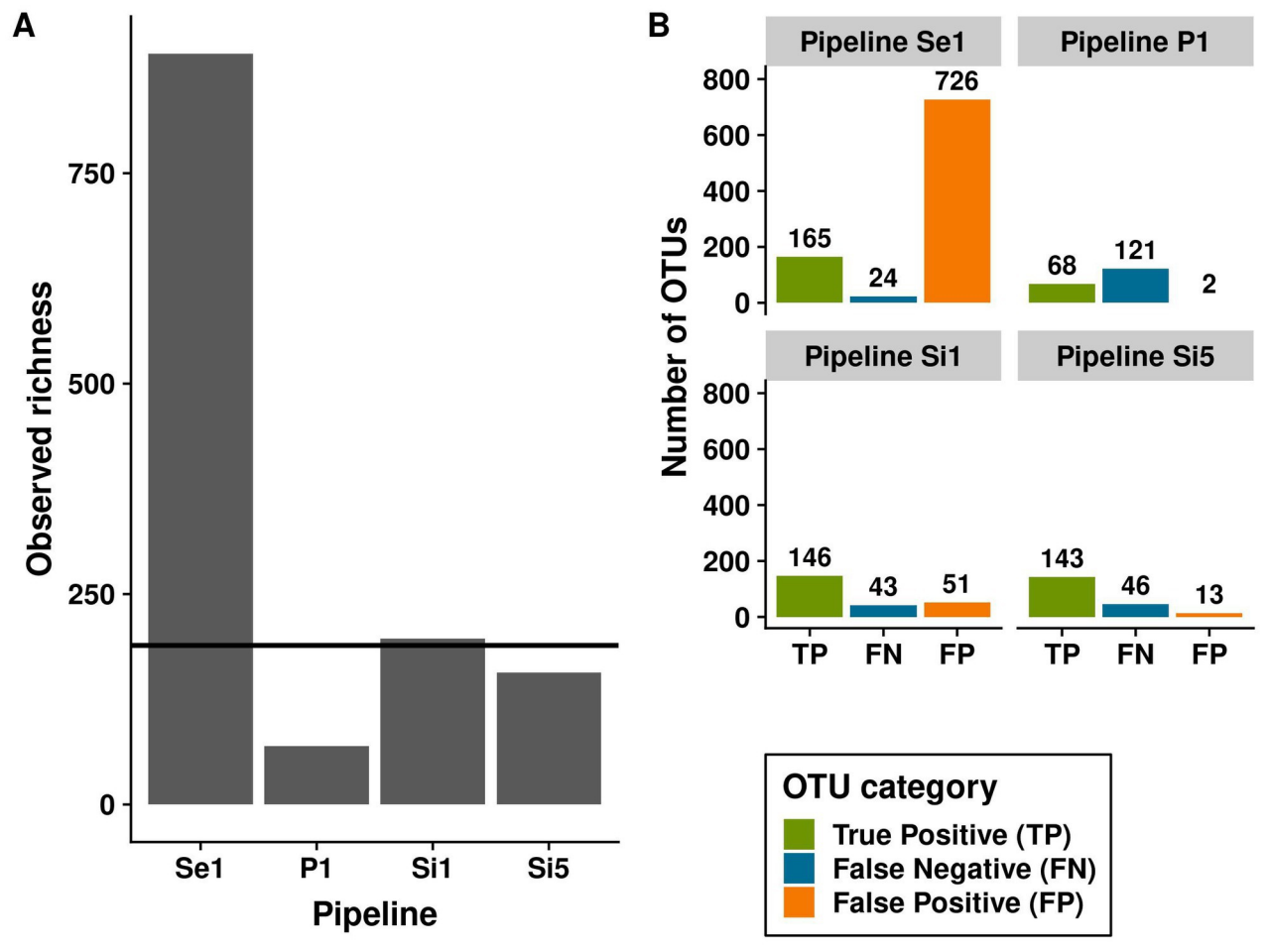
Mucoromycota

0 *Kingdom* 1 *Phylum* 2 *Class* 3 *Order* 4 *Family* 5 *Genus* 6 *Species* 7 *Strain*

8 *ITS1 sequences of the mock community*

9 *ITS1 sequences recovered by the recommended approach*


MiSeq sequencing of the ITS1 region of the fungal mock community using ITS1F-ITS2 primers
Assembly
1
Assembly of paired-end reads
PEAR_50
PEAR_100
PEAR_150
FASTQJOIN_50
FASTQJOIN_100
FASTQJOIN_150
No assembly
CUTADAPT_MERGED§
QUALITY_R1
Quality filter
2
Extraction of the ITS1 region
No
Yes
3
Analysis of sequence variation
DADA2
USEARCH*
VSEARCH
Amplicon Sequence Variants (ASVs)
Operational Taxonomical Units (OTUs)
4
Chimera treatment
Retained
Removed
5
Filtering of the final OTU (or ASV) table
T = All
T = 1
T = 10
T = RA
T = LULU
360 OTU (or ASV) tables


A
Observed richness
750
500
250
0
Se1
P1
Si1
Si5
Pipeline
B
Number of OTUs
Pipeline Se1
Pipeline P1
Pipeline Si1
Pipeline Si5
800
600
400
200
0
165
24
726
68
121
2
146
43
51
143
46
13
TP
FN
FP
OTU category
True Positive (TP)
False Negative (FN)
False Positive (FP)

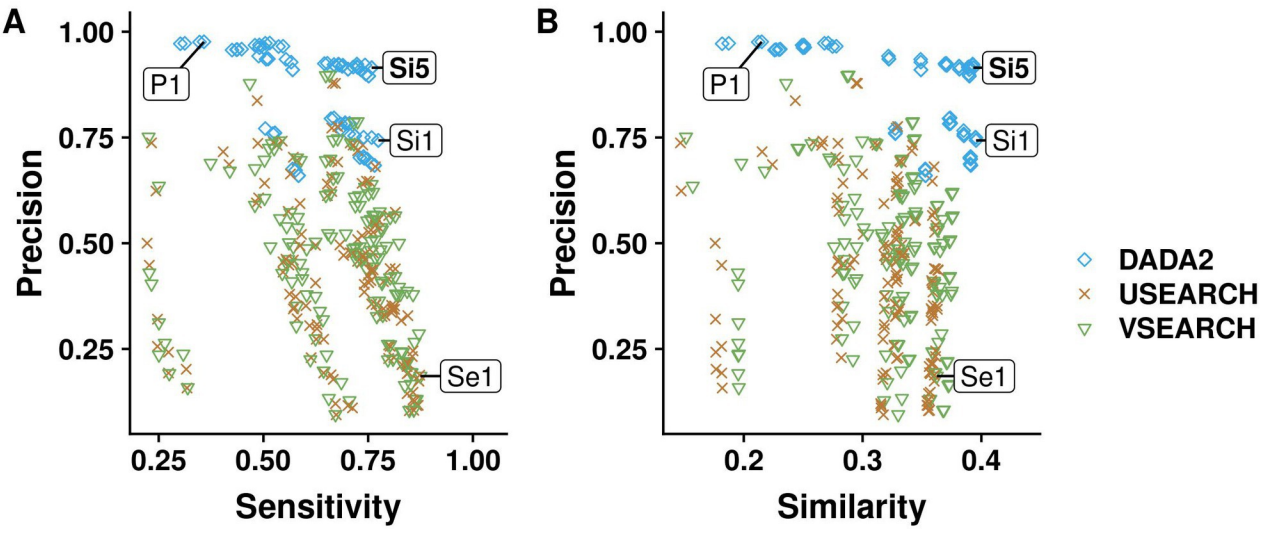

A
1.00
0.75
0.50
0.25
Precision
P1
Si5
Si1
Se1
0.25
0.50
0.75
1.00
Sensitivity
B
1.00
0.75
0.50
0.25
Precision
P1
Si5
Si1
Se1
0.2
0.3
0.4
Similarity
DADA2
USEARCH
VSEARCH

Table 1 – List of the 10 most sensitive approaches. Sensitivity, precision and similarity values were averaged over the three replicates for each bioinformatic approach. Richness is defined as the mean number of OTUs identified by the bioinformatic approach.

Approach	***Assembly***	***Extraction***	***Variation***	***Chimeras***	***Filtering***	**Richness**	**Sensitivity**	**Precision**	**Similarity**
Se1	QUALITY_R1	No	VSEARCH	Retained	All	892	0.875	0.186	0.362
Se2	QUALITY_R1	No	USEARCH	Removed	All	878	0.873	0.188	0.360
Se3	QUALITY_R1	No	USEARCH	Retained	All	949	0.871	0.174	0.3608
Se4	QUALITY_R1	No	VSEARCH	Removed	All	577	0.871	0.286	0.366
Se5	PEAR_50	No	USEARCH	Retained	All	1410	0.869	0.117	0.354
Se6	PEAR_100	No	USEARCH	Retained	All	1413	0.866	0.116	0.355
Se7	PEAR_50	No	VSEARCH	Retained	All	1257	0.866	0.131	0.357
Se8	PEAR_100	No	VSEARCH	Retained	All	1246	0.862	0.131	0.357
Se9	PEAR_50	No	USEARCH	Removed	All	1287	0.861	0.127	0.354
Se10	QUALITY_R1	No	VSEARCH	Retained	1	612	0.861	0.266	0.364

Table 2 – List of the 10 most precise approaches. Sensitivity, precision and similarity values were averaged over the three replicates for each bioinformatic approach. Richness is defined as the mean number of ASVs. LULU was applied with default settings.

Approach	***Assembly***	***Extraction***	***Variation***	***Chimeras***	***Filtering***	***Richness***	***Sensitivity***	***Precision***	***Similarity***
P1	PEAR_150	No	DADA2	Retained	LULU	69	0.358	0.976	0.215
P2	PEAR_150	No	DADA2	Removed	LULU	67	0.347	0.976	0.212
P3	CUTADAPT_MERGED	No	DADA2	Retained	LULU	100	0.515	0.973	0.271
P4	CUTADAPT_MERGED	No	DADA2	Removed	LULU	98	0.504	0.973	0.268
P5	FASTQJOIN_150	No	DADA2	Retained	LULU	61	0.312	0.973	0.187
P6	FASTQJOIN_150	No	DADA2	Removed	LULU	59	0.302	0.972	0.182
P7	PEAR_100	No	DADA2	Retained	LULU	96	0.49	0.969	0.251
P8	PEAR_100	No	DADA2	Removed	LULU	94	0.48	0.968	0.249
P9	QUALITY_R1	No	DADA2	Retained	LULU	107	0.547	0.966	0.278
P10	QUALITY_R1	No	DADA2	Removed	LULU	105	0.536	0.965	0.275

Table 3 – List of the 10 approaches with the best performances in terms of compositional similarity to the mock community. Sensitivity, precision and similarity values were averaged over the three replicates for each bioinformatic approach. Richness is defined as the mean number of ASVs. The bioinformatic approach recommended in this study (Si5) is shown in bold.

Approach	***Assembly***	***Extraction***	***Variation***	***Chimeras***	***Filtering***	**Richness**	**Sensitivity**	**Precision**	**Similarity**
Si1	QUALITY_R1	No	DADA2	Retained	All	197	0.774	0.743	0.396
Si2	QUALITY_R1	No	DADA2	Retained	1	197	0.774	0.743	0.396
Si3	QUALITY_R1	No	DADA2	Retained	RA	192	0.758	0.75	0.396
Si4	QUALITY_R1	No	DADA2	Retained	10	187	0.739	0.751	0.396
Si5	**QUALITY_R1**	**No**	**DADA2**	**Removed**	**All**	**157**	**0.758**	**0.915**	**0.395**
Si6	QUALITY_R1	No	DADA2	Removed	1	157	0.758	0.915	0.395
Si7	QUALITY_R1	No	DADA2	Removed	RA	152	0.743	0.921	0.395
Si8	QUALITY_R1	No	DADA2	Removed	10	148	0.723	0.924	0.394
Si9	PEAR_50	No	DADA2	Retained	All	212	0.765	0.684	0.391
Si10	PEAR_50	No	DADA2	Retained	1	212	0.765	0.684	0.391