Biol Fertil Soils (2013) 49:1027–1037
DOI 10.1007/s00374-013-0801-y

ORIGINAL PAPER

Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the introduction of the pipeline SEED

Tomáš Větrovský · Petr Baldrian

Received: 1 December 2012 / Revised: 11 March 2013 / Accepted: 25 March 2013 / Published online: 16 April 2013

Abstract Fungi are important in soils as both decomposers and plant symbionts, and an understanding of the composition of their complex communities is thus indispensable to answer a variety of ecological questions. 454 Pyrosequencing is currently the method of choice for the in-depth analysis of fungal communities. However, the interpretation of its results is complicated by differences in data analysis approaches that make inter-study comparisons difficult. The pyrosequencing studies published so far have also used variable molecular targets in fungal rDNA. Although the ITS region and, in particular, ITS1 appear to be the most frequent sequencing targets, the use of various primers with different coverages of fungal groups remains a serious problem. Sequence length limits also vary widely across studies, and in many studies, length differences may negatively affect sequence similarity clustering or identification. Unfortunately, many studies neglect the need to correct for method-dependent errors, such as pyrosequencing noise or chimeric sequences. Even when performed, error rates in sequences may be high, and consensus sequences created by sequence clustering therefore better represent operational taxonomic units. We recommend a data analysis workflow that includes sequence denoising, chimera removal, sequence trimming before clustering and random resampling before calculating diversity parameters. The newly developed free pipeline (SEED) introduced here can be used to perform all the required analytical steps. The improvement and unification of data analysis procedures should make future studies both more reliable and comparable and allow meta-studies to be performed to provide more general views on fungal diversity, biogeography or ecology.

T. Větrovský · P. Baldrian (✉)
Laboratory of Environmental Microbiology,
Institute of Microbiology of the ASCR, v.v.i., Vídeňská 1083,
14220, Prague 4, Czech Republic
e-mail: baldrian@biomed.cas.cz

Keywords Fungal community · Internal transcribed spacer · Pyrosequencing pipeline · Ribosomal DNA · Sequencing errors · Soil microbial ecology

Introduction

Fungi are important in soils as both decomposers and plant symbionts. Traditional surveys based on macroscopic or microscopic features, such as fruit body surveys, microscopy of plant roots or isolation techniques, despite considerable progress, have been insufficient to describe fungal communities inhabiting soil environments. Molecular methods have recently greatly overcome these limitations to allow detection of unculturable community members. Since its first applications in 2009 (Buéé et al. 2009; Jumpponen and Jones 2009), amplicon pyrosequencing studies have focused on the diversity of fungal communities (e.g. Buéé et al. 2009; Jumpponen and Jones 2009; Öpik et al. 2009), the activity of fungal communities (Baldrian et al. 2012; Štursová et al. 2012) or functional genes (e.g. Baldrian et al. 2013; Voříšková and Baldrian 2013) of both total fungi and specific groups like the Glomeromycota (e.g. Dumbrell et al. 2011; Lekberg et al. 2012; Öpik et al. 2009). Pyrosequencing has become the method of choice for the in-depth analysis of fungal community composition.

Data accumulate with increasing numbers of studies, but the experimental approaches for data collection and analysis widely differ. This unfortunately greatly limits our ability to compare among studies and draw general conclusions regarding important questions such as estimating community diversity, evenness and composition or identifying important taxa. Data analysis appears to be an important area where further improvements and unification of experimental procedures are necessary. Past experience derived from published studies indicates which steps are

important and should be considered when designing data analysis workflows.

Most importantly, the complexity of the data and the specifics of the methods may cause several biases that affect the quality of the resulting sequence dataset and any subsequent statistical analyses or ecological considerations. These include pyrosequencing-specific errors, sometimes termed "sequencing noise" (Quince et al. 2009, 2011), unclear quality of low abundance sequences, the presence of chimeric sequences (Taylor and Houston 2011; Tedersoo et al. 2010) and PCR target-associated biases (Bellemain et al. 2010; Krüger et al. 2012).

Despite the development of alternative sequencing platforms (Shokralla et al. 2012), pyrosequencing will likely remain widely used in the near future. For this reason, we believe that the standardisation of methods for fungal community analysis is highly desirable because it will soon allow us to exploit the wealth of individual studies to deliver general statements regarding fungal diversity, biogeography or ecology.

Standards of data reporting that include information regarding the sampling site and its corresponding metadata, laboratory processing steps and data analysis were previously suggested (Nilsson et al. 2011). The aim of this paper was to describe the data analysis procedures previously used, indicate the limiting steps and suggest a simple data analysis workflow that can avoid potential problems. Because the processing of large-scale pyrosequencing-derived data may represent a methodological limitation, a newly developed software pipeline, SEED, is introduced in this paper that allows researchers to perform the required data analysis steps with a single, easy-to-use user interface.

Materials and methods

Meta-analysis of studies using amplicon pyrosequencing to explore fungal communities

Scientific publications using amplicon pyrosequencing to analyse fungal communities were retrieved. The source of sample material, molecular target (gene and primer pair) and number of sequences that were used for community analysis were recorded. With respect to the experimental methodology used for sequence processing, the minimum sequence length and the presence or absence of sequence processing steps (removal of pyrosequencing noise, removal of chimeric sequences, creation of similarity clusters, diversity analysis and sequence annotation) were recorded. The bioinformatic tools used for data cleanup, sequence clustering and annotation were also recorded (Table 1). The data retrieved from publications were used to analyse the approaches used in fungal community amplicon pyrosequencing.

Development of pipeline to analyse sequences obtained by amplicon pyrosequencing

Based on the previously applied approaches to amplicon pyrosequencing data analysis, the necessary steps were identified and the analysis workflow was proposed. The development of the optimized workflow was based on both the available knowledge from previous papers about the effects of certain data analysis steps on the resulting dataset quality (Schloss et al. 2009, Edgar et al. 2011) and on our own analysis of a sample dataset. For this purpose, the publicly available dataset deposited in MG Rast 4497081.3 that contains sequences of fungal internal transcribed spacer (ITS) region from oak leaves at different stages of decomposition (Voříšková and Baldrian 2013) was used. The aim was to analyse the effects of certain data analysis steps on the fungal diversity estimates and identification of operational taxonomic units (OTUs; defined as sequences clustered at a 97 % probability level). Specifically, we analysed (1) for each sample (n=21, 1129 sequences per sample) the effects of clustering sequences of original length (380–560 bases) versus sequences truncated to the same length on OTU richness, the Chao estimate and the number of singletons; (2) for each sample the effects of chimera removal on OTU richness, Chao and singletons; and (3) for the 150 most abundant OTUs, the quality of OTU identification was compared with OTUs represented either by random sequences or consensus sequence. The quality of identification was defined as the similarity of the query sequence and the most similar Sanger sequencing-derived sequence deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/). This is based on the assumption that sequences containing errors are less similar to real sequences and that consensus construction should correct random errors in sequences. Sequence clustering and chimera removal were performed using default Usearch and Uchime settings (Edgar 2010, Edgar et al. 2011), and nucleotide BLAST (Altschul et al. 1990) was used to retrieve the closest hits from GenBank. Wilcoxon pair test was used to analyse the differences among dataset pairs. Differences at $P<0.01$ were regarded as statistically significant. The optimized data analysis workflow was used in the course of the development of a user-friendly data analysis pipeline.

The pipeline, SEED (http://www.biomed.cas.cz/mbu/lbwrf/seed/main.html), was created that enables users to perform the entire bioinformatic analysis of PCR amplicons according to the suggested workflow. The same pipeline was also used to perform all workflow testing steps outlined above. The functionality of the pipeline was tested with datasets from previous pyrosequencing projects with amplicon sequences for the fungal ITS region, bacterial 16S rDNA and the fungal *cbhI* exocellulase gene (Baldrian et al. 2012; Štursová et al. 2012; Větrovský and Baldrian 2013;

Table 1 Overview of studies using amplicon-based 454 pyrosequencing to analyse fungal communities along with important parameters of data processing

Sample[a]	Target	Primer pair(s)	Length[b]	Denoising	Chimera removal	Data cleanup[c]	Clustering[d]	Percentage	Rep. sequence[e]	Singleton removal	Diversity[f]	Annotation[g]	Size[h]	Reference
Soil	ITS	ITS1f/ITS2	100			EMBOSS	BLASTCLUST	97	R		YES	MEGAN	OOO	Buéé et al. (2009)
Plant	ITS	ITS1f/ITS2	200				CAP3	95	R	YES		BLASTN	OO	Jumpponen and Jones (2009)
Roots (G)	18S	NS31/AM1	160								YES	BLASTN	OOO	Öpik et al. (2009)
Sediment	ITS, 28S	ITS1f/ITS4, LRORf/LR5f	300			RDP	CD-HIT	97				MEGAN	OOO	Amend et al. (2010b)
Soil	ITS	ITS1f/ITS4	300			CLC Gen. W.	CD-HIT	97			YES (R)	MEGAN	O	Amend et al. (2010a)
Roots	ITS	ITS1f/ITS2	200				CAP3	95			YES	BLASTN	OO	Jumpponen et al. (2010b)
Plant	ITS	ITS1f/ITS4	200				CAP3	95		YES		BLASTN	OO	Jumpponen and Jones (2010)
Soil	ITS	ITS5/ITS4	200				CAP3	95		YES		BLASTN	OO	Jumpponen et al. (2010a)
Soil (G)	18S	AMV4.5NF /AMDGR, NS31/ Ammix	230					97					O	Lumini et al. (2010)
Plant	ITS	ITS1f/ITS2	150									BLASTN	OO	Ovaskainen et al. (2010)
Roots	ITS	ITS5/ITS2	140		YES	NEWBLER	Other	97	L	YES		BLASTN	OO	Tedersoo et al. (2010)
Soil	ITS	ITS1f/ITS4	200			SCATA	SCATA	98.5	C			SCATA	OO	Wallander et al. (2010)
Roots (G)	18S	WANDA/AM1	100									BLASTN	OOO	Dumbrell et al. (2011)
Soil, roots	28S	LROR/LR3	150			MOTHUR	Other	95		YES	YES	BLASTN	OOO	Gottel et al. (2011)
Soil	ITS	ITS1f/ITS2	200				CAP3	95	R	YES	YES	BLASTN	O	Hui et al. (2011)
Soil	ITS	ITS5/ITS2	196				Other	98			YES		OOO	Lentendu et al. (2011)
Soil	ITS	ITS1f/ITS2, ITS3/ITS4	60			Sequencher	Other	97			YES	BLASTN	OO	Mello et al. (2011)
Roots (G)	18S	NS31/AM1	160								YES	BLASTN	OOO	Moora et al. (2011)
Soil	ITS	ITS1/58A2R	200			CLOTU	BLASTCLUST	97		YES	YES	BLASTN	O	Xu et al. (2011)
Coral reef	28S	LROR/LR5	250	YES	YES	MOTHUR	Other	97		YES		MEGAN	OO	Amend et al. (2012)
Plant	ITS, 18S	ITS1f/ITS4, SSU817f/ SSU1536r	200		YES	PANGEA	CD-HIT	98	L	YES	YES	MEGAN	OOO	Arfi et al. (2012a)
Soil	ITS	ITS1f/ITS4	200		YES		CD-HIT	97–98	L	YES	YES		OO	Arfi et al. (2012b)
Soil	ITS	ITS1/ITS4	380	YES	YES	MOTHUR	CD-HIT (S)	97	C		YES (R)	PlutoF	OO	Baldrian et al. (2012)
Soil	18S	SSU817f/ SSU1536r	195				CAP3	97	R	YES	YES	BLASTN	?	Becklin et al. (2012)
Sediment	18S	SSU_F04/ SSU_R22, NF1/18Sr2b	200	YES	YES	QIIME	UCLUST	95–99	C			BLASTN	OOO	Bik et al. (2012)
Roots	ITS	ITS5/ITS2	150			CLOTU	BLASTCLUST	98.5		YES		CLOTU	OOO	Blaalid et al. (2012)
Plant	ITS	ITS1f/ITS2	100			RDP	UCLUST	97	A			BLASTN	OO	Cordier et al. (2012)
Plant, soil	ITS	ITS3/ITS4	200	YES		QIIME	UCLUST	97	A	YES		BLASTN	OOO	Davey et al. (2012)
Soil (G)	18S	NS31/AML2	170		YES					YES	YES	BLASTN	OOO	Davison et al. (2012)
Soil	ITS	ITS3/ITS4	ITS2			MOTHUR	CRUNCHCLUST	97			YES	MOTHUR	OOO	Hartmann et al. (2012)

Table 1 (continued)

Sample[a]	Target	Primer pair(s)	Length[b]	Denoising	Chimera removal	Data cleanup[c]	Clustering[d]	Percentage	Rep. sequence[e]	Singleton removal	Diversity[f]	Annotation[g]	Size[h]	Reference
Soil, plant	ITS	ITS1f/ITS4, fITS7/ITS4, gITS7/ITS4, fITS9/ITS4	ITS2			SCATA	SCATA	97.5	C			PlutoF	OOO	Ihrmark et al. (2012)
Roots	ITS	ITS5/ITS2	150			CLOTU	BLASTCLUST	97	L			BLASTN	OO	Kauserud et al. (2012)
Wood	ITS	ITS1f/ITS4	200			SCATA	SCATA	98.5	A	YES		SCATA	OOO	Kubartova et al. (2012)
Sediment	ITS	ITS1f/ITS4	200		YES	MOTHUR	CRUNCHCLUST	97				PlutoF	OOO	La Duc et al. (2012)
Roots (G)	28S	glo454/NDL22	150			RDP	Other	97		YES		BLASTN	OOO	Lekberg et al. (2012)
Soil (G)	18S	AMV4.5NF/ AMDGR	200			QIIME	Other	97	A		YES (R)	BLASTN	OOO	Lin et al. (2012)
Soil	18S	SSU817f/ SSU1196r	?			QIIME	Other	97			YES (R)	BLASTN	OO	McGuire et al. (2012)
Soil, roots	ITS	ITS9/ITS4	200			SEQMAN PRO	Other	97			YES	BLASTN	O	Menkis et al. (2012)
Soil	ITS	ITS1/ITS4	300	YES	YES	MOTHUR	CD-HIT (S)	97	C		YES (R)	PlutoF	O	Šťursová et al. (2012)
Roots	ITS	ITS1f/ITS2	150			CLOTU	BLASTCLUST	97	L	YES	YES	BLASTN	OO	Yu et al. (2012a)
Roots	ITS	ITS1f/ITS4	150			CLOTU	BLASTCLUST	97		YES	YES	BLASTN	OO	Yu et al. (2012b)
Soil, plant	ITS	ITS1/ITS4	380	YES	YES	MOTHUR	CD-HIT (S)	97	C		YES (R)	PlutoF	OO	Voříšková and Baldrian (2013)

[a] Soil also includes litter; (G)—only Glomeromycota

[b] ITS2: the length of the entire ITS2 region

[c] CLC Gen. W. (http://www.clcbio.com/products/clc-genomics-workbench/), CLOTU (http://www.bioportal.uio.no/appinfo/show.php?app=CLOTU), EMBOSS (http://emboss.sourceforge.net/), MOTHUR (http://www.mothur.org/), NEWBLER (http://en.wikipedia.org/wiki/Newbler), PANGEA (http://pangea-16s.sourceforge.net/), QIIME (http://qiime.org/), RDP (http://rdp.cme.msu.edu/), SCATA (http://scata.mykopat.slu.se/), SEQMAN PRO (http://www.dnastar.com/t-sub-products-lasergene-seqmanpro.aspx), Sequencher (http://genecodes.com/)

[d] BLASTCLUST (http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html), CAP3 (http://seq.cs.iastate.edu/), CD-HIT (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi), CRUNCHCLUST (http://code.google.com/p/crunchclust/), SCATA (http://scata.mykopat.slu.se/), UCLUST (http://www.drive5.com/usearch/). (S)—sequences resized to identical lengths

[e] A—most abundant, C—cluster consensus, L—longest, R—random

[f] R—resampling to the same depth

[g] BLASTN (blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn), CLOTU (http://www.bioportal.uio.no/appinfo/show.php?app=CLOTU), MEGAN (http://ab.inf.uni-tuebingen.de/software/megan/), MOTHUR (http://www.mothur.org/), PlutoF (http://elurikkus.ut.ee/plutof.php?lang=eng), SCATA (http://scata.mykopat.slu.se/)

[h] O<10,000 sequences, OO>10 000 sequences, OOO>100 000 sequences

Voříšková and Baldrian 2013). In the last paper, the data analysis workflow recommended here was used for data processing.

The SEED pipeline is a workbench that runs in the Microsoft Windows environment with internal functions and functions performed by external programmes that must be installed for full functionality. The removal of pyrosequencing noise is performed using Pat Schloss's translation of Chris Quince's PyroNoise algorithm implemented within the Mothur package (Schloss et al. 2009). The removal of chimeras created during PCR amplification is performed using Uchime (Edgar et al. 2011), and Usearch (Edgar 2010) is used for sequence clustering. Sequence alignment is performed by calling MAFFT (Katoh et al. 2009), and BLAST searching and the creation of local databases are dependent on the National Center for Biotechnology Information (NCBI) tools (http://www.ncbi.nlm.nih.gov/; Altschul et al. (1990)). Internet connection is required for searching online databases, e.g. the NCBI nucleotide database.

The SEED pipeline is freely available for non-commercial use and can be downloaded along with documentation from the SEED project webpage: http://www.biomed.cas.cz/mbu/lbwrf/seed/main.html. The installation of external programmes may require the consent of their authors: more information can be found at the web pages of these projects, accessible by hyperlink from the above address.

Results

In total, 42 published studies were analysed (Table 1). The number of papers using amplicon pyrosequencing to analyse fungal communities increased rapidly from 3 in 2009 to 22 in 2012. Soil fungal communities were the most common target of amplicon pyrosequencing (21 studies), along with fungal communities in plant roots (12 papers). Other environments (sediments, aboveground plant tissues, corals or wood) were only rarely addressed. Although most studies were designed to cover the entire fungal community, six papers targeted specifically arbuscular mycorrhizal fungi. In addition to analysing entire fungal communities, amplicon sequencing was also applied to analyse the diversity of the fungal *cbhI* exocellulase gene, a proxy for the community of cellulose-decomposing fungi (Baldrian et al. 2012; Štursová et al. 2012; Voříšková and Baldrian 2013). There is only one single study to date in which RNA-derived amplicons were used to specifically analyse metabolically active fungal taxa (Baldrian et al. 2012; Purahong and Krüger 2012).

The ITS region was by far the most frequently analysed region of fungal rDNA: only four and three papers analysed various regions of the 18S and 28S rRNA genes, respectively (Fig. 1). Within the ITS, ITS1 was mainly targeted with several primer pairs to amplify only this region; in additional studies, both the ITS1 and ITS2 regions were amplified, but because of the limiting lengths of pyrosequencing-derived sequences and the fact that sequencing mostly occurred from primers within the 18S, the sequence data also covered predominantly the ITS1 region. Only recently, studies analysing the ITS2 region specifically have been conducted (Davey et al. 2012; Hartmann et al. 2012; Ihrmark et al. 2012; Menkis et al. 2012).

The initial steps of sequence data processing typically consisted of sequence quality filtering and reduction of PCR or sequencing errors. A wide set of tools was used for data cleanup, which resulted in the removal of sequences of insufficient length or quality, but the minimal length of sequences retained in the cleaned dataset varied considerably (Table 1 and Fig. 2). Typically, between 10 and 40 % of sequences were removed in this step. Pyrosequencing-derived errors, typically the variable lengths of longer homopolymer regions, were corrected by clustering pyrosequencing flowgrams, termed "denoising", and PCR-derived errors were removed by chimera-cleaning tools. Despite the high rate of occurrence of both types of errors, only <30 % of all studies used one of these approaches and only 12 % used both (Fig. 2).

Sequences that passed filtering steps were used to create virtual taxa, i.e. the sequence similarity clusters most often termed operational taxonomic units. Despite the inconsistency of clustering sequences of variable lengths, only a handful of studies truncated sequences to identical lengths or extracted particular DNA regions before clustering. CD-HIT, BLASTCLUST and CAP3 were most frequently used for clustering. For annotation, OTUs were represented either by a randomly selected sequence or by the longest or most abundant sequence. In six studies, consensus sequences were constructed to represent OTUs (Table 1). Approximately one half of the studies only considered non-singleton sequences for community analysis (Fig. 2), and BLAST against the NCBI database was the most frequent approach to assign taxonomic identity to OTUs. In 55 % of studies, diversity parameters were calculated for individual samples. Among these, only 26% performed resampling to the same depth before calculating diversity (Fig. 2).

After considering the previous data analysis protocols, we suggest the following workflow (Table 2). Quality trimming should first exclude sequences of low base quality and length. The minimal length of sequences to be analysed should be at least above 150 bases because both the ITS1 and ITS2 are longer than that for many fungi. The quality of taxonomic assignments based on the 18S or 28S region analyses also greatly increases with sequence length. Both denoising and chimera removal should be performed to reduce the sequence error rate to a minimum. Because clustering algorithms compare sequences in a pairwise manner, the regions to be clustered should optimally be defined as the same DNA region, i.e. with defined primer positions at

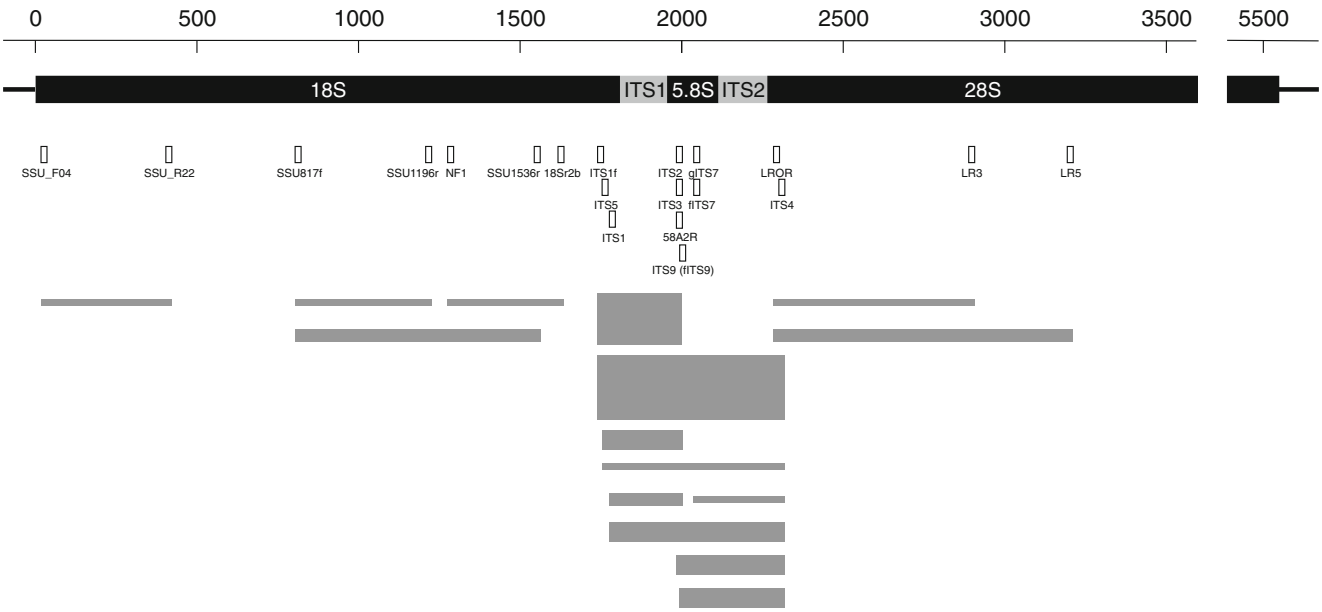

Fig. 1 Primers and PCR amplicons used in amplicon pyrosequencing analyses of the community composition of general fungi. The thickness of grey bars indicates the number of studies using the respective amplicons. Numbers indicate the positions in the rDNA of *Fusarium oxysporum*

both ends (if the amplicons are shorter than pyrosequencing read length), or using a defined sequence (e.g. ITS1, ITS2, ITS1+5.8S+ITS2) that can be extracted easily (Nilsson et al. 2010), or at least defined by the same length of all sequences. Consensus sequences best represent individual sequences within an OTU.

Depending on the aim of the study, sequence identification may be requested either based on the identity of the closest database hit or through multiple alignment of OTU sequences with known sequences. In the studies targeting the diversity of fungal communities, community richness, evenness or other parameters may also be derived. To obtain comparable data, the sequence database has to be randomly resampled to obtain identical numbers of sequences from each sample.

Clustering of sequences truncated at 380 bases gave lower OTU counts, numbers of singletons and Chao estimates of total community richness than the clustering of

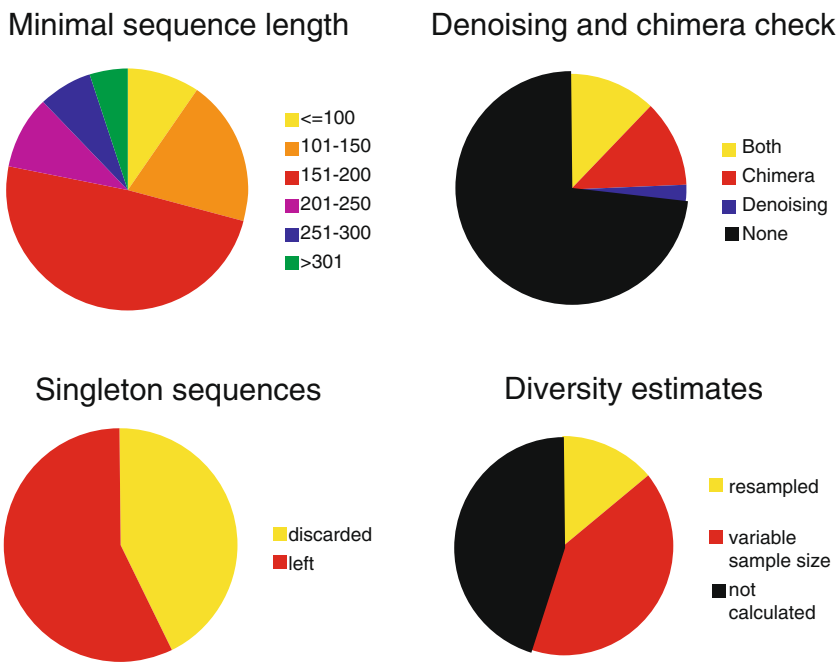

Fig. 2 Overview of approaches used to analyse sequences derived by amplicon pyrosequencing of fungal rDNA based on 42 recently published studies

Table 2 Workflow of the analysis of sequences derived by amplicon pyrosequencing of fungal communities

Step	Description and comments
Quality trimming	• Removal of sequences of inferior quality (e.g. quality score <20) or length (e.g. length <100)
	Trimming before denoising helps reduce the size of the dataset and, thus, the length of the denoising procedure.
Sample identification	• Distribution of sequences into samples based on barcode sequences
Denoising	• Removal of pyrosequencing noise
	Denoising helps reduce the amount of method-dependent errors (e.g. \homopolymers, point mistakes) by comparing pyrosequencing flow grams between sets of highly similar sequences.
Removal of chimeric sequences	• Deletion of potentially chimeric sequences from the dataset
	Chimeric sequences arise during PCR at various rates (often >10 %) depending on PCR conditions, such as the number of cycles or template concentrations. The large numbers of sequences from each PCR reaction allow the detection of chimeric sequences based on their multiple comparisons.
Selection of region for clustering	• Trimming of sequences to contain the same part of the template
	For clustering, sequences should contain identical regions of DNA because length differences make clustering algorithms unreliable because of the uneven similarity of short and long sequences. This can be achieved by trimming to identical lengths (setting a lower sequence length limit, e.g. 300 bases, and truncating long sequences) or by the identification of sequence boundaries within rDNA (e.g. the start and end of ITS1 or ITS2). It may be desirable to exclude primer sequence(s).
Clustering	• OTUs are created by grouping sequences based on their similarity
	Typically, clustering is based on sequential comparison of individual sequences with sequences used for cluster establishment (seed sequences, i.e. those sequences that show similarity lower than the defined threshold with all seed sequences of clusters established so far).
Creation of consensus sequences	• Consensus sequences for each OTU are created by sequence alignment.
	Consensus sequences better represent the OTU than individual sequences because consensus creation removes random sequencing errors that survived denoising. As a result, closer hits to known taxa are found for consensus sequences than for individual sequences. Consensus sequences can also be used to represent OTUs in phylogenetic analyses.
Sequence identification	• Best-identified hits are retrieved for each OTU consensus sequence. The full taxonomy of the best hit may also be retrieved.
	The best hits are retrieved along with the similarity values (*E*, per cent similarity) that help assign the OTU to a specific taxon.
Community composition analysis	• Based on the full taxonomic placement of best hits, the abundance of taxa of various taxonomic levels can be calculated.
	The abundance of individual taxa (e.g. OTUs, genera, orders, phyla) in each sample can be expressed as a percentage of all sequences. Community composition data are usually used for statistical purposes (comparison of samples by correlation, analysis of variance, multivariate methods or sample similarity clustering).
Estimation of diversity parameters	• Based on the sequence counts for each OTU of a sample, diversity and evenness parameters of the community are calculated.
	Because diversity estimates tend to scale up with increasing sampling depth, it is essential to randomly resample the sequence database to include identical numbers of sequences from each sample. For this subsampled database, OTUs must be newly created.

sequences of their original lengths of 380–560 bases. The numbers of OTUs, singletons and Chao estimates were lower by 13.4±1.6, 12.7±2.4 and 7.4±3.3 %, respectively, all differences being statistically significant at $P<0.003$. This shows that the OTU counts are inflated when sequences of different lengths are clustered together. The

application of chimera removal on sequences truncated to 380 bases decreased the numbers of OTUs, singletons and Chao estimates further by 20.1±1.2, 16.7±2.0 and 17.5±3.4 %, respectively, all differences being statistically significant at $P<0.001$. This shows that a significant part of the apparent diversity in the dataset may be due to the presence of chimeric sequences. Consensus sequences of the 150 most abundant OTUs in the dataset showed significantly higher ($P<0.0001$) sequence similarity to the closest BLAST hit in GenBank than random sequences, with 69 % consensus sequences showing higher similarity, 27 % showing the same similarity and 4 % showing lower similarity. Moreover, 13 % OTU consensus sequences showed 100 % similarity to the GenBank sequence, whilst the corresponding random sequences were less similar. On average, consensus sequences were by 0.29±0.04 % more similar to the closest GenBank hits than randomly selected sequences.

The SEED pipeline makes it possible to perform all steps of the sequence analysis workflow from a single, user-friendly interface (Fig. 3). The features of the pipeline are summarised in Table 3, and more information can be found on the project webpage (http://www.biomed.cas.cz/mbu/lbwrf/seed/main.html) that contains full documentation of the functions and a step-by-step introduction to the data processing workflow. Importantly, in addition to sequence grouping, SEED makes it possible to perform batch operations with groups, such as chimera removal from individual samples, calculation of consensus sequences for individual OTUs, resampling of all samples at a specific depth, etc. SEED can be used to analyse PCR amplicons of any type, e.g. bacterial

Table 3 Features of the amplicon pyrosequencing pipeline SEED

Sequence editing and sorting
Extraction of sequences and sequence qualities from *.sff files
Quality trimming
Grouping sequences based on sequence motifs or sequence titles
Sequence batch processing
Sequence denoising (using the PyroNoise algorithm translation within Mothur)[a]
Chimera removal (using Uchime)[a]
Sequence alignment (using MAFFT)[a]
Sequence clustering (using Usearch)[a]
Construction of consensus sequences
Searching for best hits in a local database or the NCBI (using nucleotide BLAST)[a]
Retrieval of taxonomical classification of best hits from the NCBI
Creation of local databases for searching by nucleotide BLAST
Calculation of diversity parameters

[a] Mothur (Schloss et al. 2009), Uchime (Edgar et al. 2011), MAFFT (Katoh et al. 2009), Usearch (Edgar 2010), BLAST (Altschul et al. 1990), NCBI (http://www.ncbi.nlm.nih.gov/)

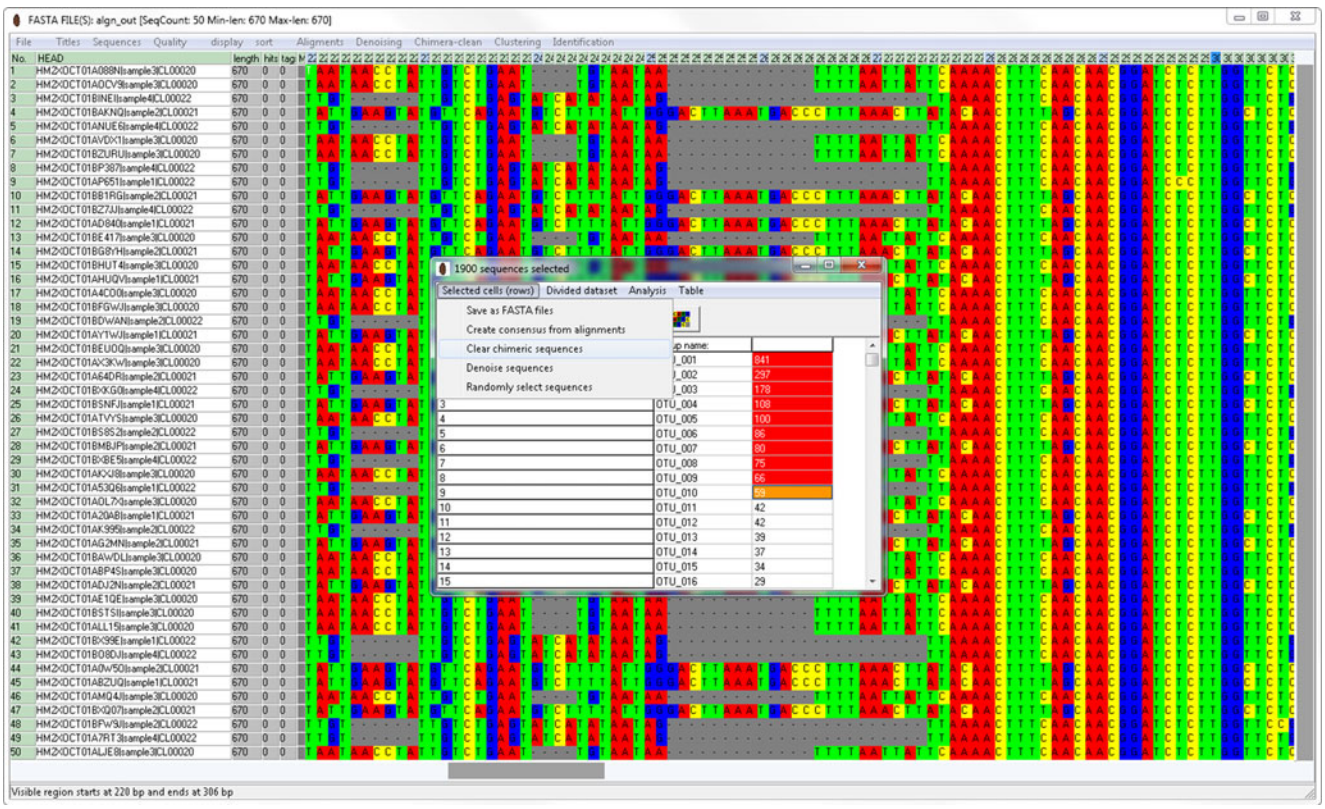

Fig. 3 Screenshot of the amplicon pyrosequencing pipeline SEED

16S rDNA or functional genes, or to analyse gene sequences obtained by other means (e.g. batch download from the NCBI nucleotide or genome database).

Discussion

The methods of next-generation sequencing have revolutionised microbial ecology, allowing researchers to explore complex communities at unprecedented depths. Despite the first applications of the Illumina (Caporaso et al. 2012) or Ion Torrent (Whiteley et al. 2012) technologies to explore bacterial communities, pyrosequencing remains the method of choice for fungal and bacterial amplicon sequencing, offering the advantages of reasonable sequence length, easy multiplexing and sufficient sequencing depth for most studies (Glenn 2011). Nevertheless, successful applications of pyrosequencing approaches are dependent on a number of methodological considerations, including sampling strategies and metadata collection, the choice of suitable molecular marker and approaches for data analysis. Because of the diversity of all of the above methodologies in the published studies, it is extremely difficult to use the wealth of information derived by pyrosequencing for inter-study comparisons or meta-studies. Furthermore, published papers differ widely in the level of method descriptions and data availability. We strongly agree with the previous paper by Nilsson et al. (2011) in that full description of the experimental procedures and public data availability should be a standard.

Here we show that despite some general preferences, many different molecular targets are used to study both general fungi and arbuscular mycorrhizal fungi. Without exception, fungal rDNA was targeted despite widely varying relationships between its copy numbers and fungal cell counts or biomasses (Amend et al. 2010a; Baldrian et al. 2013). The ITS region amplified using various sets of primers was the preferred target, consistent with the dominant current opinion (Schoch et al. 2012).

Although ITS1 was frequently sequenced, it is notable that the results obtained with various primers cannot be easily compared because of their variable coverage of the fungal tree of life (Anderson et al. 2003). Unfortunately, there are only a few papers in which various primers were compared. The recent paper by Ihrmark et al. (2012) demonstrates that PCR amplification can be highly uneven among primer pairs as well as diversity estimates. More work is still required in this direction.

The data analysis procedures used in past amplicon pyrosequencing studies indicate many potential limitations of data quality. Studies using sequences of <150 bases length covered less than the entire ITS1 or ITS2 regions of certain fungi because of the differences in the regions' lengths, and this seems to be unsuitable. In our in silico study considering the region between the ITS1/ITS4 primers, fungal sequence assignment quality increased with increasing sequence length up to the length of 350–380 bases (data not shown). Such sequence lengths are easily available with current technologies and may be desirable when reliable OTU classification is required. Furthermore, clustering algorithms work best with sequences of identical boundaries (or lengths), a fact that is usually not considered. Here, we show that clustering of sequences of uneven length significantly increases the diversity estimates.

PCR and pyrosequencing have been shown to cause method-dependent sequencing errors (Quince et al. 2009; Tedersoo et al. 2010). In PCR amplification, chimeric sequences are formed with frequencies at or above 3 %, depending on the number of cycles (Taylor and Houston 2011). Because these sequences are most often singletons, the presence of chimeric sequences may result in an overestimation of diversity. This was also clearly demonstrated here in the comparison of diversity estimates among the original and chimera-cleaned dataset. Chimera-cleaning procedures should therefore always be applied. When choosing a minimal length, one should also consider that the probability of detecting chimeric sequences rapidly increases with sequence length, and shorter sequences are more likely to contain undiscovered chimeras. In addition, the increase of sequence error counts associated with increasing sequence lengths and the frequency of sequencing errors in homopolymeric regions that stem from the techniques of pyrosequencing should be reduced by applying denoising (i.e. error correction) procedures (Quince et al. 2011). Unfortunately, error-correcting procedures have been rarely applied so far. Given the error rate of pyrosequencing-derived reads and the random distribution of such errors, the creation of OTU consensus sequences should further improve the representation of an OTU. This was demonstrated here by the fact that the consensus sequences are significantly more similar to the Sanger sequences deposited in GenBank than individual OTU sequences.

To explore fungal diversity, the analysis of identical numbers of sequences from all samples is essential because diversity estimates always scale up with sampling depth. This fact has also been frequently neglected in past studies.

We here outline a workflow of data analysis that aims to reflect all of the considerations required for obtaining high-quality data for community analysis and offer the SEED pipeline to accomplish this task. We hope that the unification of data analysis procedures represents an important step towards better comparability of individual studies and justification of their conclusions. The SEED pipeline should offer ecologists a tool that is easy to use, even for those with no preliminary experience with amplicon pyrosequencing, the method that will likely continue to dominate microbial community analysis in the coming years.

Acknowledgments This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (LD12048, LD12050), by the Czech Science Foundation (P504/12/0709) and by the Research concept of the Institute of Microbiology ASCR (RVO61388971).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Amend AS, Barshis DJ, Oliver TA (2012) Coral-associated marine fungi form novel lineages and heterogeneous assemblages. ISME J 6:1291–1301

Amend AS, Seifert KA, Bruns TD (2010a) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Mol Ecol 19:5555–5565

Amend AS, Seifert KA, Samson R, Bruns TD (2010b) Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. Proc Nat Acad Sci U S A 107:13748–13753

Anderson IC, Campbell CD, Prosser JI (2003) Potential bias of fungal 18S rDNA and internal transcribed spacer polymerase chain reaction primers for estimating fungal biodiversity in soil. Environ Microbiol 5:36–47

Arfi Y, Buéé M, Marchand C, Levasseur A, Record E (2012a) Multiple markers pyrosequencing reveals highly diverse and host-specific fungal communities on the mangrove trees *Avicennia marina* and *Rhizophora stylosa*. FEMS Microbiol Ecol 79:433–444

Arfi Y, Marchand C, Wartel M, Record E (2012b) Fungal diversity in anoxic–sulfidic sediments in a mangrove soil. Fungal Ecol 5:282–285

Baldrian P, Kolarik M, Stursova M, Kopecky J, Valaskova V, Vetrovsky T, Zifcakova L, Snajdr J, Ridl J, Vlcek C, Voriskova J (2012) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J 6:248–258

Baldrian P, Větrovský T, Cajthaml T, Dobiášová P, Petránková M, Šnajdr J, Eichlerová I (2013) Estimation of fungal biomass in forest litter and soil. Fungal Ecol 6:1–11

Becklin KM, Hertweck KL, Jumpponen A (2012) Host identity impacts rhizosphere fungal communities associated with three alpine plant species. Microb Ecol 63:682–693

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 Microbiol 10:189

Bik HM, Sung W, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A, Thomas WK (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. Mol Ecol 21:1048–1059

Blaalid R, Carlsen T, Kumar S, Halvorsen R, Ugland KI, Fontana G, Kauserud H (2012) Changes in the root-associated fungal communities along a primary succession gradient analysed by 454 pyrosequencing. Mol Ecol 21:1897–1908

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

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624

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

Davey ML, Heegaard E, Halvorsen R, Ohlson M, Kauserud H (2012) Seasonal trends in the biomass and structure of bryophyte-associated fungal communities explored by 454 pyrosequencing. New Phytol 195:844–856

Davison J, Öpik M, Zobel M, Vasar M, Metsis M, Moora M (2012) Communities of arbuscular mycorrhizal fungi detected in forest soil are spatially heterogeneous but do not vary throughout the growing season. PLoS One 7:e41938

Dumbrell AJ, Ashton PD, Aziz N, Feng G, Nelson M, Dytham C, Fitter AH, Helgason T (2011) Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. New Phytol 190:794–804

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

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Res 11:759–769

Gottel NR, Castro HF, Kerley M, Yang ZM, Pelletier DA, Podar M, Karpinets T, Uberbacher E, Tuskan GA, Vilgalys R, Doktycz MJ, Schadt CW (2011) Distinct microbial communities within the endosphere and rhizosphere of *Populus deltoides* roots across contrasting soil types. Appl Environ Microbiol 77:5934–5944

Hartmann M, Howes CG, Vaninsberghe D, Yu H, Bachar D, Christen R, Nilsson HR, Hallam SJ, Mohn WW (2012) Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests. ISME J 6:2199–2218

Hui N, Jumpponen A, Niskanen T, Liimatainen K, Jones KL, Koivula T, Romantschuk M, Strommer R (2011) EcM fungal community structure, but not diversity, altered in a Pb-contaminated shooting range in a boreal coniferous forest site in Southern Finland. FEMS Microbiol Ecol 76:121–132

Ihrmark K, Bödeker IT, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandstrom-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 Microbiol Ecol 82:666–677

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

Jumpponen A, Jones KL (2010) Seasonally dynamic fungal communities in the *Quercus macrocarpa* phyllosphere differ between urban and nonurban environments. New Phytol 186:496–513

Jumpponen A, Jones KL, Blair J (2010a) Vertical distribution of fungal communities in tallgrass prairie soil. Mycologia 102:1027–1041

Jumpponen A, Jones KL, Mattox D, Yaege C (2010b) Massively parallel 454-sequencing of fungal communities in *Quercus* spp. ectomycorrhizas indicates seasonal dynamics in urban and rural sites. Mol Ecol 19:41–53

Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. In: Posada D (ed) Bioinformatics for DNA sequence analysis. Humana, Totowa, pp 39–64

Kauserud H, Kumar S, Brysting AK, Nordén J, Carlsen T (2012) High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal plant root samples. Mycorrhiza 22:309–315

Krüger D, Kapturska D, Fischer C, Daniel R, Wubet T (2012) Diversity measures in environmental sequences are highly dependent on alignment quality-data from ITS and new LSU primers targeting basidiomycetes. PLoS One 7:e32139

Kubartova A, Ottosson E, Dahlberg A, Stenlid J (2012) Patterns of fungal communities among and within decaying logs, revealed by 454 sequencing. Mol Ecol 21:4514–4532

La Duc MT, Vaishampayan P, Nilsson HR, Torok T, Venkateswaran K (2012) Pyrosequencing-derived bacterial, archaeal, and fungal diversity of spacecraft hardware destined for Mars. Appl Environ Microbiol 78:5912–5922

Lekberg Y, Schnoor T, Kjøller R, Gibbons SM, Hansen LH, Al-Soud WA, Sorensen SJ, Rosendahl S (2012) 454-Sequencing reveals stochastic local reassembly and high disturbance tolerance within arbuscular mycorrhizal fungal communities. J Ecol 100:151–160

Lentendu G, Zinger L, Manel S, Coissac E, Choler P, Geremia RA, Melodelima C (2011) Assessment of soil fungal diversity in different alpine tundra habitats by means of pyrosequencing. Fungal Divers 49:113–123

Lin XG, Feng YZ, Zhang HY, Chen RR, Wang JH, Zhang JB, Chu HY (2012) Long-term balanced fertilization decreases arbuscular mycorrhizal fungal diversity in an arable soil in North China revealed by 454 pyrosequencing. Environ Sci Technol 46:5764–5771

Lumini E, Orgiazzi A, Borriello R, Bonfante P, Bianciotto V (2010) Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach. Environ Microbiol 12:2165–2179

McGuire KL, Fierer N, Bateman C, Treseder KK, Turner BL (2012) Fungal community composition in neotropical rain forests: the influence of tree diversity and precipitation. Microb Ecol 63:804–812

Mello A, Napoli C, Murat C, Morin E, Marceddu G, Bonfante P (2011) ITS-1 versus ITS-2 pyrosequencing: a comparison of fungal populations in truffle grounds. Mycologia 103:1184–1193

Menkis A, Burokiene D, Gaitnieks T, Uotila A, Johannesson H, Rosling A, Finlay RD, Stenlid J, Vasaitis R (2012) Occurrence and impact of the root-rot biocontrol agent *Phlebiopsis gigantea* on soil fungal communities in *Picea abies* forests of northern Europe. FEMS Microbiol Ecol 81:438–445

Moora M, Berger S, Davison J, Öpik M, Bommarco R, Bruelheide H, Kuhn I, Kunin WE, Metsis M, Rortais A, Vanatoa A, Vanatoa E, Stout JC, Truusa M, Westphal C, Zobel M, Walther GR (2011) Alien plants associate with widespread generalist arbuscular mycorrhizal fungal taxa: evidence from a continental-scale study using massively parallel 454 sequencing. J Biogeogr 38:1305–1317

Nilsson RH, Tedersoo L, Lindahl BD, Kjoller R, Carlsen T, Quince C, Abarenkov K, Pennanen T, Stenlid J, Bruns T, Larsson KH, Koljalg U, Kauserud H (2011) Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. New Phytol 191:314–318

Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, Abarenkov K (2010) An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. Fungal Ecol 3:284–287

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 Phytol 184:424–437

Ovaskainen O, Nokso-Koivista J, Hottola J, Rajala T, Pennanen T, Ali-Kovero H, Miettinen O, Oinonen P, Auvinen P, Paulin L, Larsson KH, Mäkipää R (2010) Identifying wood-inhabiting fungi with 454 sequencing—what is the probability that BLAST gives the correct species? Fungal Ecol 3:274–283

Purahong W, Krüger D (2012) A better understanding of functional roles of fungi in the decomposition process: using precursor rRNA containing ITS regions as a marker for the active fungal community. Ann Forest Sci 69:659–662

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Meth 6:639–641

Quince C, Lanzén A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12:38

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Mol Ecol 21:1794–1805

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, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. Proc Nat Acad Sci U S A 109:6241–6246

Štursová M, Žifčáková L, Leigh MB, Burgess R, Baldrian P (2012) Cellulose utilization in forest litter and soil: identification of bacterial and fungal decomposers. FEMS Microbiol Ecol 80:735–746

Taylor DL, Houston S (2011) A bioinformatics pipeline for sequence-based analyses of fungal biodiversity. In: Xu JR, Bluhm BH (eds) Fungal genomics. Humana, Totowa, pp 141–155

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Koljalg U (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytol 188:291–301

Větrovský T, Baldrian P (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One 8:e57923

Voříšková J, Baldrian P (2013) Fungal community on decomposing leaf litter undergoes rapid successional changes. ISME J 7:477–486

Wallander H, Johansson U, Sterkenburg E, Durling MB, Lindahl BD (2010) Production of ectomycorrhizal mycelium peaks during canopy closure in Norway spruce forests. New Phytol 187:1124–1134

Whiteley AS, Jenkins S, Waite I, Kresoje N, Payne H, Mullan B, Allcock R, O'Donnell A (2012) Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. J Microbiol Methods 91:80–88

Xu LH, Ravnskov S, Larsen J, Nicolaisen M (2011) Influence of DNA extraction and PCR amplification on studies of soil fungal communities based on amplicon sequencing. Can J Microbiol 57:1062–1066

Yu L, Nicolaisen M, Larsen J, Ravnskov S (2012a) Molecular characterization of root-associated fungal communities in relation to health status of *Pisum sativum* using barcoded pyrosequencing. Plant Soil 357:395–405

Yu L, Nicolaisen M, Larsen J, Ravnskov S (2012b) Succession of root-associated fungi in *Pisum sativum* during a plant growth cycle as examined by 454 pyrosequencing. Plant Soil 358:216–224