Methods

Navigating the labyrinth: a guide to sequence-based, community ecology of arbuscular mycorrhizal fungi

Miranda M. Hart[1], Kristin Aleklett[1], Pierre-Luc Chagnon[2], Cameron Egan[1], Stefano Ghignone[3], Thorunn Helgason[4], Ylva Lekberg[5], Maarja Öpik[6], Brian J. Pickles[1] and Lauren Waller[7]

[1]Biology University of British Columbia Okanagan, Kelowna, BC, Canada; [2]Département de Biologie, Université de Sherbrooke, 2500 Boulevard de l'université, Sherbrooke, QC, Canada; [3]Istituto per la Protezione Sostenibile delle Piante (UOS Torino), C.N.R., Torino, Italy; [4]Department of Biology, University of York, Heslington, York, YO10 5DD, UK; [5]MPG Ranch and Department for Ecosystem and Conservation Sciences, University of Montana, Missoula, MT, USA; [6]Department of Botany, University of Tartu, 40 Lai St, 51005 Tartu, Estonia; [7]Division of Biological Sciences, University of Montana, Missoula, MT, USA

Author for correspondence:
Miranda M. Hart
Tel: +1 250 807 9398
Email: Miranda.hart@ubc.ca

Received: *13 October 2014*
Accepted: *18 January 2015*

New Phytologist (2015)
doi: 10.1111/nph.13340

Key words: arbuscular mycorrhizal fungi (AMF), bioinformatics, environmental sequencing, next generation sequencing (NGS), PCR.

Summary

- Data generated from next generation sequencing (NGS) will soon comprise the majority of information about arbuscular mycorrhizal fungal (AMF) communities. Although these approaches give deeper insight, analysing NGS data involves decisions that can significantly affect results and conclusions. This is particularly true for AMF community studies, because much remains to be known about their basic biology and genetics.
- During a workshop in 2013, representatives from seven research groups using NGS for AMF community ecology gathered to discuss common challenges and directions for future research. Our goal was to improve the quality and accessibility of NGS data for the AMF research community. Discussions spanned sampling design, sample preservation, sequencing, bioinformatics and data archiving.
- With concrete examples we demonstrated how different approaches can significantly alter analysis outcomes. Failure to consider the consequences of these decisions may compound bias introduced at each step along the workflow.
- The products of these discussions have been summarized in this paper in order to serve as a guide for any researcher undertaking NGS sequencing of AMF communities.

Introduction

The development of molecular techniques has made field-based community ecology of arbuscular mycorrhizal fungi (AMF) truly feasible. Although cloning and Sanger sequencing permitted detection and identification of AMF *in situ* without the need of recognizable morphological features, next generation sequencing (NGS) approaches have given us unprecedented insight into AMF communities with sufficient depth to recover even rare taxa. This shift brings about a labyrinth of decisions spanning a breadth of scientific expertise that can be daunting for both students and established researchers.

In December 2013, we held a workshop in Kelowna, Canada, to discuss major obstacles specific to AMF molecular ecology, involving sampling, sample processing, amplification, sequencing protocols and bioinformatic analyses. Here we present the outcome of these discussions, provide recommendations where appropriate (Fig. 1), and highlight the consequences of some decisions using published data. We also discuss the importance of databases, both reference and sequence repositories, to the AMF research community. It is our hope that this paper will serve as a navigation guide to the NGS labyrinth, and inspire us to improve the tools and approaches used in AMF molecular ecology.

Sampling design

Sampling design must be driven by research questions and hypotheses grounded in ecological theory (Prosser *et al.*, 2007). Although studies that aim to inventory AMF taxa for descriptive or exploratory means continue to play an important role in understanding the biology of Glomeromycota, their evolution, ecology and biogeography, here we concentrate on sampling design for experimental approaches. Studies characterizing rare taxa will require different sampling strategies from those that aim to quantify the responses of abundant taxa, or those looking at ecological gradients, such as host plant identity, soil abiotic factors or seasonality. Later we discuss sampling considerations for statistically robust conclusions.

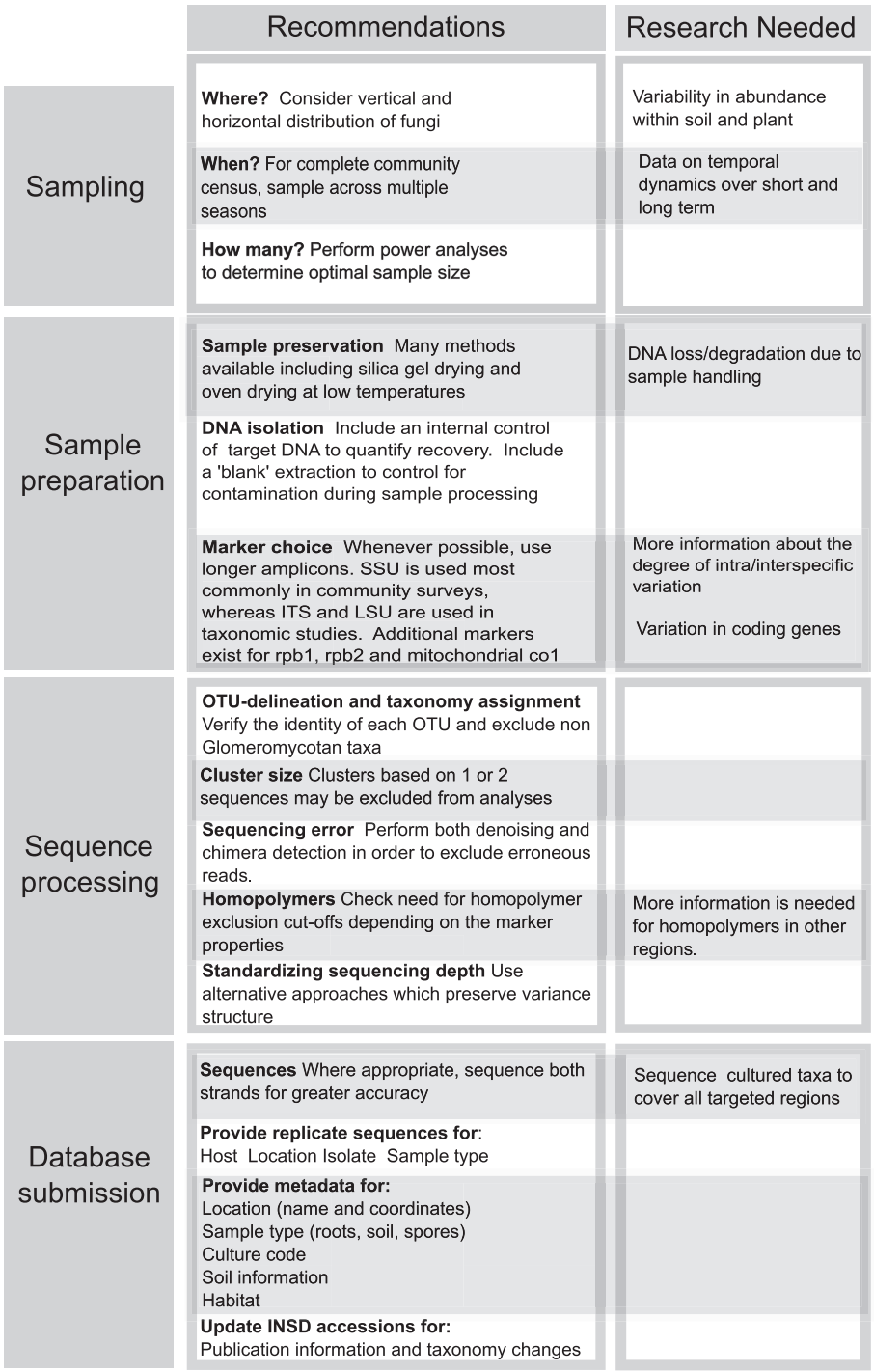

Fig. 1 Overview of the workflow for arbuscular mycorrhizal fungi (AMF) next generation sequencing (NGS), highlighting current recommendations and areas requiring further research along the work flow. SSU, small subunit; ITS, internal transcribed spacer; LSU, large subunit; OTU, operational taxonomic unit; INSD, International Nucleotide Sequence Database.

Where, what type, when, how many and how much to sample?

Where With an explicit sampling design, *where* to sample must be carefully considered. AMF communities show geographic patterns that are not entirely explained by plant community composition or soil abiotic conditions (Dumbrell *et al.*, 2010; Lekberg *et al.*, 2011). For example, AMF communities also change with soil depth and can be surprisingly rich in deeper soil horizons (Oehl *et al.*, 2003). Sampling only the upper 10–15 cm of the soil profile may substantially underestimate mycorrhizal fungal diversity (Pickles & Pither, 2014). Both arbuscular mycorrhizal (AM) and ectomycorrhizal (EM) fungal distribution has been shown to vary

significantly with depth along soil profile (Bahram *et al.*, 2015). Thus consideration of both vertical and horizontal distribution should be an important consideration for every study.

What type AMF communities observed in different types of samples (soil, root, spores) can be distinct (Clapp *et al.*, 1995; Hempel *et al.*, 2007), possibly due to differences in life history strategies (Sýkorová *et al.*, 2007) and intraradical vs extraradical biomass allocation among AMF (Hart & Reader, 2002). Sampling roots of individual plant species is the most common approach, as it enables detection of actively colonizing fungi, and can reveal host effect, if any (Vandenkoornhuyse *et al.*, 2002; Öpik *et al.*, 2009). Spores extracted from soil permit detection of the sporulating fraction of the AMF community (Clapp *et al.*, 1995; Sikes *et al.*, 2012), but as AMF taxa do not sporulate consistently, and the lifespan of spores in natural environment is not known (Smith & Read, 2008), spore-based sampling may underestimate total diversity (Sanders, 2004). It is also possible to target only hyphae extracted from soil (Jakobsen *et al.*, 1992) or hyphal in-growth compartments (Wallander *et al.*, 2001; Neumann & George, 2005). Alternatively, soil containing spores and hyphae can be sampled as a whole when the goal is to thoroughly describe the community in soil, regardless of differences in sporulation (Saks *et al.*, 2014). Finally, composite root samples (as opposed to sampling roots from individual species) capture AMF regardless of host-related co-occurrence patterns (Hiiesalu *et al.*, 2014).

When Due to seasonal and interannual temporal variation in AMF communities (Bever *et al.*, 1997; Davison *et al.*, 2011; Dumbrell *et al.*, 2011; Cotton *et al.*, 2015), the timing of sampling is important, particularly in experiments with a large spatial extent. However, little is known about AMF temporal patterns across different climates and habitats, and even in better-studied temperate zones, our understanding of AMF seasonal dynamics is still rudimentary. To address the research gap of temporal variability in AMF community dynamics, time-course sampling across seasons and years is needed.

How many The optimal replication level (number of samples per treatment or sample group) depends on the expected variation within the system. If it is high, more samples per treatment need to be analysed to detect patterns of interest. A surprising number of microbial diversity studies, including AMF studies, are unreplicated (Prosser, 2010). We cannot overemphasize that large sequence numbers do not compensate for low sample numbers due to among-sample variation. Therefore, taxon accumulation (rarefaction) curves of samples per study rather than sequences per study are needed to assess sufficiency of the sampling effort for capturing the taxa present in the study system (Fig. 2). Expected effect sizes (the magnitude of differences between groups) and required replication level can be calculated on the basis of preliminary surveys or earlier publications.

Power analysis is well established for determining adequate sampling effort to detect treatment differences in studies with single response variables. It is now possible to perform power analysis for multivariate datasets as well (La Rosa *et al.*, 2012; 2015) because statistical power is determined by both sample numbers (replication level) and number of sequences per sample (sequencing depth). Estimates of statistical power may be highly sensitive to sequencing decisions. Depending on the complexity of the system and expected variance, different replication levels and sequencing depths are required to detect patterns of interest in different study systems (Table 1; Supporting Information Methods S1). Overall, power analyses from preliminary surveys are a cost-effective way to develop sampling strategies before any large-scale sampling.

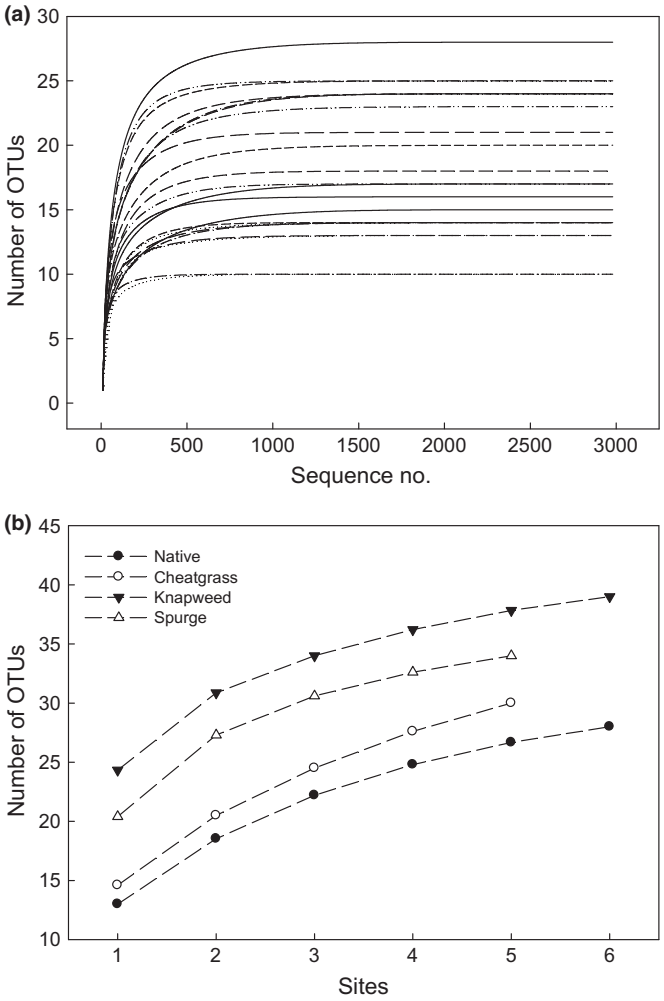

Fig. 2 Comparisons of rarefaction and sampling effort curves (taken from Lekberg *et al.*, 2012, 2013). (a) Operational taxonomic unit (OTU) accumulation curves of individual samples sequenced to depth of 3000 reads suggest that over 500 reads are required to adequately characterize the arbuscular mycorrhizal fungi (AMF) community in each sample. (b) OTU accumulation curves of four plant communities sampled at replication level 5–6 suggest that over six samples per habitat are required to fully characterize AMF communities. Whereas (a) depends on AMF richness in each sample, (b) is driven by AMF community heterogeneity among samples. It can be concluded that the sample level diversity was fully captured and additional sequencing depth per sample is not necessary (a), whilst sampling additional samples and sites could still increase detected diversity for each plant community type.

Table 1 Experimental power calculations (based on $\alpha = 0.05$) using the Dirichlet-multinomial (DM) model as described in La Rosa *et al.* (2012, 2013)

	Number of sequence reads (%)*							
	Denmark				Montana			
Samples	100	1000	5000	10 000	100	1000	5000	10 000
2	6.7	9.5	11.1	8.0	56.9	72.5	72.7	75.0
4	7.1	11.4	10.4	10.2	95.9	99.0	99.3	98.3
6	26.5	37.9	38.0	37.9	100	100	100	100
8	64.1	70.9	71.1	72.9	100	100	100	100
10	89.8	90.7	92.9	93.1	100	100	100	100
12	97.3	98.6	98.9	98.3	100	100	100	100

Using two published datasets, we estimated taxon counts across a number of samples and sequence reads. The Denmark dataset (Lekberg *et al.*, 2012) reported no differences among four disturbance treatments ($n = 11$), and the Montana dataset (Lekberg *et al.*, 2013) showed different arbuscular mycorrhizal fungi (AMF) communities among four plant community types ($n = 6$). Our power analyses show that the lack of treatment effects in the Denmark study was not due to low power. Importantly, the number of samples required per treatment to provide statistical power of 90% was only slightly different for 100 vs 10 000 sequence reads per sample, suggesting that deep sequencing would have been superfluous. Also, the large differences among plant communities in the Montana study indicate that as few as four replicates would have sufficed to detect differences. See Supporting Information Methods S1 for a more detailed description of the power analysis.
*Based on 1000 Monte-Carlo randomizations of DM vector data.

Avoiding pseudoreplication requires knowledge of the processes responsible for spatial autocorrelation of AMF communities (see Legendre & Legendre, 1998; Ettema & Wardle, 2002). Individual sample size, distance between samples, and the spatial extent of sampling should be based on AMF traits, such as hyphal density and extent of individual fungi. However, there is limited knowledge about AMF traits in nature. Given the indeterminate mycelial growth of AMF it is not a simple procedure to confirm that samples are independent. Research into spatial autocorrelation in EMF root tip communities has recommended a 3–4-m distance between samples to minimize resampling of the same fungal community patch (Lilleskov *et al.*, 2004; Pickles *et al.*, 2012). Yet results from AMF studies are limited and inconclusive: there is evidence for community-wide spatial autocorrelation and overdispersion occurring at similar resolutions (< 0.4 and > 0.5 m, respectively; Mummey & Rillig, 2008). However, AMF genets may measure as much as 10 m across (Rosendahl & Stukenbrock, 2004) so these community patterns may be nested inside the distribution of an individual fungus. Recently, Bahram *et al.* (2015) showed larger vertical than horizontal spatial variability in AMF communities. Clearly, further research on spatial properties of AMF diversity patterns across ecosystems is needed in order to sample AMF communities so that spatial variation is appropriately taken into account.

How much The *size* of individual samples depends on the density and distribution of AMF species within sites. In practice, the effect of sample volume on detected diversity patterns has not been tested. For root samples, *c.* 20 cm root length (Öpik *et al.*, 2008) or *c.* 70 mg DW are common choices (Öpik *et al.*, 2013). Alternatively, multiple samples of 0.5–1 cm per plant individual have been used (Kjøller & Rosendahl, 2000). For soil samples, common kits require the use of 250 mg of soil, which can yield low and variable results (Lumini *et al.*, 2010; Davison *et al.*, 2012), possibly due to low biomass and/or nuclear concentration of AMF in soil (Saks *et al.*, 2014). It has been suggested that larger soil or root samples can considerably increase the consistency of PCR and sequencing success (Janoušková *et al.*, 2015). Clearly, more research is needed to determine optimal individual sample sizes, and distances between samples for AMF communities in different ecosystems.

Sample preservation, DNA extraction and handling

Regardless of sample type, AMF DNA is typically rare compared with bacterial and general fungal DNA that is coextracted (O'Brien *et al.*, 2005; Karst *et al.*, 2013). Sample handling methods can destroy much of this DNA, as well as introduce or enhance contaminating DNA from other organisms. Below, we outline important decisions pertaining to AMF samples and DNA extraction.

Sample preservation

The impact of sample preservation methods on subsequent recovery of AMF DNA can be substantial (Bainard *et al.*, 2010; Janoušková *et al.*, 2015). The main considerations when choosing the suitable approach aim to halt physiological processes in the sample, and optimize practicalities of storage and transportation. Snap-freezing in liquid nitrogen is fast and convenient in the lab, but limits transportation of samples from field to lab. Silica-gel drying of root and soil samples permits indefinite sample storage, and transportation at ambient temperature has been used when sampling even from remote areas (Davison *et al.*, 2012; Öpik *et al.*, 2013). Freeze-drying (Hiiesalu *et al.*, 2014), oven-drying at low heat (50–60°C; Janoušková *et al.*, 2015) and sample storage in DNA extraction buffer are other commonly used sample preservation approaches for AMF. Although freezing at −80°C or −20°C may not affect DNA quality compared with control DNA extracts from fresh samples, other sample preservation methods (storage in ethanol, lyophilizing) may result in a

significant loss of DNA (Bainard *et al.*, 2010). Oven drying – the cheapest and simplest preservation method, but requiring access to equipped laboratory – may in fact be best at preserving AMF DNA from roots (Janoušková *et al.*, 2015), but it is not clear if this is also true for soil samples. In general, time from sampling to sample preservation should be kept to minimum in order to avoid DNA degradation, changes in the fungal community and degradation by saprotrophs.

DNA isolation

A wide range of approaches have been used to extract Glomeromycotan DNA, ranging from in-house protocols based on cetyltrimethylammonium bromide (CTAB; Clapp *et al.*, 1999) or phenol/chloroform extractions (Daniell *et al.*, 2001), to the now prevailing use of single-sample or high-throughput kits. DNA from single spores or small numbers of spores can be extracted with simplified protocols (Schwarzott & Schüßler, 2001), or protocols dedicated to recovering minute quantities of DNA (Manian *et al.*, 2001; Thiéry *et al.*, 2012).

Even small differences in sample handling can result in differential recovery of DNA. Sample processing order and position on high throughput extraction plates may also introduce unintended variation among samples, and, thus, randomization is recommended to avoid bias against particular treatments. This is also true for timing: researcher fatigue may result in later samples being handled less efficiently. Because high-throughput extractions can take several hours, samples processed early in the protocol will be exposed to variable conditions longer than those processed near the end.

Despite meticulous handling and perfectly optimized protocols, extracted DNA will constitute a fraction of that present in the original sample (Bainard *et al.*, 2010; Janoušková *et al.*, 2015). Internal standards, as simple as 'spiking' sterile material (soil or root extraction) with a known quantity of AMF DNA, will provide a measure of DNA yield (Nguyen *et al.*, 2014). Although some researchers include this control, it is often overlooked. Such controls are especially important for samples that originate from different soil types or host plants, which might impose differential constraints on extraction efficiency. Finally, contamination of samples during handling can result in a significant proportion of 'nontarget' reads (Lindahl *et al.*, 2013). To assess the degree of handling contamination, it is imperative to include a 'blank' sample (negative control) with every step of the protocol (Nguyen *et al.*, 2014). Including independent validation means is also valuable – for example, estimating AMF biomass in root or soil samples by microscopic methods or fatty acid quantification, composing artificial communities of known species, or estimating variation among replicates of the same sample in DNA extraction, PCRs or sequencing runs (Schmidt *et al.*, 2013).

Sequencing

Ideally, a single marker would be used to recognise organisms at organisational levels from genotype to kingdom. In reality, there is no *de facto* best sequence target that would achieve all aims; the ideal marker is the one that provides appropriate data for the hypothesis being tested. Researchers embarking on a DNA-based study of AMF communities should therefore ask the following questions: is it necessary that each operational taxonomic unit be given a specific identification? This currently limits target selection to ribosomal genes that have database representation. If change across communities is more important than operational taxonomic unit (OTU) identity, then different sequence targets may be used. To what taxonomic level do the OTUs need to be differentiated? A target sequence (marker) should have appropriate variability. Discriminating taxa at the species level requires a more variable sequence than at the genus or family level. Is it necessary to sample across the phylum Glomeromycota? Primer sets vary in the extent to which all AMF groups are amplified. Is presence or function more important? Most studies have focused only on identifying taxa, but protein-encoding genes with known functions may become important functional markers for future community surveys.

Marker choice

The nuclear ribosomal operon comprises the small subunit (SSU or 18S) rRNA gene, the internal transcribed spacer (ITS) and the large subunit (LSU) rRNA gene (Table S1). This operon has multiple copies per nucleus, and is thus easier to amplify. Although it has been the most widely used marker for Glomeromycota, no marker within the operon possesses a clear barcode gap for all Glomeromycotan lineages. Thus, the preferred marker region remains open to debate.

The general fungal barcoding marker ITS has been found suboptimal as the sole marker for Glomeromycota (Schoch *et al.*, 2012; Stockinger *et al.*, 2010; Thiéry *et al.*, 2012). This problem is not unique to Glomeromycota. For example, the same applies to basal clades of Fungi (Schoch *et al.*, 2012) and several other fungal groups such as *Fusarium* where other markers are used (O'Donnell *et al.*, 2010). Although ITS can be reliable for AMF taxon identification (Krüger *et al.*, 2011; Schoch *et al.*, 2012), it is generally regarded as hyper-variable within AMF species, has low resolution among species for some groups and carries little direct information about a taxon's evolutionary relationships within the fungi (Lloyd-MacGilp *et al.*, 1996). ITS and LSU can be used to discriminate among taxa that are poorly resolved by SSU, and it may become more useful as databases are populated with further Glomeromycotan ITS sequences.

Many ecological studies have used ribosomal operon markers, particularly SSU and LSU rRNA gene targets (Helgason *et al.*, 1998; Öpik *et al.*, 2014). Although the quantity of LSU-based studies is increasing, the majority of sequence-based AMF community surveys use SSU targets (Öpik *et al.*, 2014), even though there is support for LSU being phylogenetically somewhat more informative. Although SSU gives good species resolution for most lineages, resolution of species is poor for some groups, especially the Diversisporaceae and Gigasporaceae (Öpik *et al.*, 2013). Three practical aspects have contributed to the broader use of SSU in AMF community surveys in comparison with LSU and ITS that prevail in taxonomic studies of

AMF (Öpik *et al.*, 2014): presence of general Glomeromycota specific primers, an amplicon that is of suitable length for many applications and can be aligned for phylogenetic analysis over the phylum, which is particularly critical for validating the identity of novel OTUs.

The ribosomal operon offers the greatest resolution when used as a whole. Together, these genes and regions permit alignment over all Glomeromycota (Krüger *et al.*, 2011). Phylogenetic analysis can be used to obtain family and higher rank classification as well as species-level taxon delimitation and identification, which is particularly important in environmental surveys where detection of novel clades is common (Öpik *et al.*, 2013, 2014). Unfortunately, the ribosomal operon is in excess of 5500 bp, which is intractable for Sanger sequencing, and challenging for current NGS technology, although this will improve with time. Further considerations for marker choice have been reviewed elsewhere for fungi (Lindahl & Kuske, 2013; Lindahl *et al.*, 2013) and Glomeromycota (Kohout *et al.*, 2014; Öpik *et al.*, 2014; Van Geel *et al.*, 2014).

Other markers?

Genome data (Tisserant *et al.*, 2013; Lin *et al.*, 2014) present the AMF community ecologist with the opportunity of selecting gene targets that offer information about AMF activity and function in addition to identity and abundance. *Rhizophagus irregularis* (DAOM 197198) has a large, haploid genome of 153 Mb and is estimated to have > 28 000 coding genes (Tisserant *et al.*, 2013; Lin *et al.*, 2014). As other species and isolates of AMF are sequenced, comparative genomics will allow the identification of a range of targets that can reveal genetic variation at any scale – from among isolates within a species, to higher taxonomic ranks, and everything else in between. Future studies may combine data from functional genes with taxonomic markers such as nuclear ribosomal genes or ITS, in order to determine both functional and taxonomic diversity. Transcribed genes that reliably distinguish variation at an appropriate level (i.e. mitochondrial COI; Borriello *et al.*, 2014) would allow parallel assessment of taxonomic composition and activity of fungal communities. Recently, Stockinger *et al.* (2014) used the DNA sequences of the largest subunit of RNA polymerase II gene (RPB1) as an alternative taxonomic marker to those of nuclear ribosomal operon in order to demonstrate a shift in AMF community structure in response to tillage. Although AMF community shifts have also been shown with transcribed nuclear ribosomal markers (RNA of LSU rRNA gene; Verbruggen *et al.*, 2012), functional gene markers, such as phosphate transporters (Burleigh *et al.*, 2002), aquaporins (Li *et al.*, 2013) or others, may be a more reliable measure of activity than rRNA (Anderson & Parkin, 2007; Blazewicz *et al.*, 2013); however, these have yet to be fully realized for AMF research (van der Heijden & Scheublin, 2007).

Length and quality of sequence

It is axiomatic that longer sequences provide better species-discriminating information than shorter sequences. The sequencing method used, however, places a limit on the length of sequence that can be analysed. Sanger sequencing can deliver long – *c.* 1 kb – sequences of high quality, but not in high throughput. NGS currently delivers up to 600 bp at best, but in very large numbers. However, both approaches are valuable for fully characterizing communities: Sanger sequencing is best used to populate reference databases with high quality sequences for identification, whereas shorter and more error-prone reads from NGS can be used for in-depth characterization of community patterns.

Bioinformatics

Rather than present a comprehensive description of the workflow for processing data generated with NGS platforms, we highlight issues that can affect AMF studies using our own data. These include: processing raw data (denoising, homopolymers, chimera detection, sequence trimming), OTU determination and dealing with unequal sequencing depth per sample.

Denoising

Given the large number of reads in NGS datasets, the intrinsic error rate can affect overall diversity estimates (Kunin *et al.*, 2010). The term 'noise' covers a variety of errors producing nontarget sequences including sequencing errors, PCR errors and chimeras (a comparison of denoising methods is discussed in Gaspar & Thomas, 2013). Denoising appears to always reduce alpha diversity by decreasing the number of sequences usable for analysis and OTUs detected. We show here that it can also reduce beta diversity (i.e. the 'spread' of samples) substantially, and thus may also affect final patterns detected (Fig. 3).

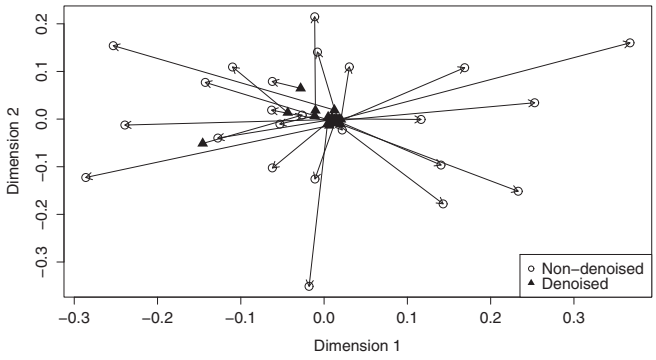

Fig. 3 Procrustes super imposition plot showing the differences between non-denoised and denoised data where the same samples are connected by vector arrows (nonmetric multidimensional scaling (NMDS) ordinations of Bray-Curtis distances). Vector length reflects the change in ordination space between non-denoised and denoised samples. Denoising resulted in a shift of all samples in ordination space, which may change the interpretation of results. Denoising reduced the total number of sequences by 46%, as well as the number of operational taxonomic units (OTUs) (400 vs 222 OTUs). Denoising was performed using a modified Needleman–Wunsch global alignment algorithm with a custom scoring function based on signal intensities developed by Reeder & Knight (2010) using data from Holland *et al.* (2014).

Homopolymers

Although primarily an artifact of the Roche 454 and Ion Torrent sequencing platforms, homopolymers introduced as error can compromise data quality. During the bioinformatic workflow, it is possible to set a maximum allowed homopolymer length, but this step may exclude valid sequences. For example, some SSU AMF reference sequences in the Maarj*AM* database contain adenine (A) homopolymers of up to 18 bp, and members of the Acaulosporaceae and Glomeraceae commonly contain homopolymers in excess of 10 thymine (T) and adenine (A) bp, respectively. Thus, relying on default parameters may result in the unintentional exclusion of sequences from certain taxonomic groups over others. Increasing the maximum homopolymer limit in the analysis will result in more *bona fide* OTUs, albeit with inflated risk of spurious OTUs.

Chimeras

Chimeras, or sequences composed of more than one organism, are another source of error in NGS datasets that can significantly inflate diversity estimates (Fonseca *et al.*, 2012). Several approaches have been developed to identify chimeras and remove them from the analysis (i.e. ChimeraSlayer, Haas *et al.*, 2011; Perseus, Quince *et al.*, 2011; UCHIME, Edgar *et al.*, 2011), but a common obstacle for AMF studies is that these approaches require additional materials from sequence databases such as global alignments and lane masks which may be difficult to obtain for commonly used targets of the nuclear ribosomal operon. Some researchers choose to use general fungal alignments whereas others create in-house databases for personal use. Ultimately, a curated database with regularly updated and publicly available global alignments is needed for Glomeromycota.

OTU clustering methods

Strategies for generating OTUs from sequences (reviewed by Lindahl *et al.*, 2013) can be categorized into three approaches: (1) cluster reads against one another without external reference sequence collection (*de novo* OTU picking); (2) group reads using a reference dataset and exclude reads that do not match a sequence in the reference database (closed-reference OTU picking); or (3) group reads as in (2) but recluster unmatched reads *de novo* (open-reference OTU picking) (Bik *et al.*, 2012).

De novo strategies produce more OTUs than reference-based methods, as some OTUs will be defined by unconstrained clustering, which may include erroneous reads. Reference-based approaches avoid this because new sequences are queried against known, reference sequences. For this process to be effective, the reference dataset must contain only high-quality, trusted sequences, and should also contain reference sequences for all known taxa. Denoising results in fewer OTUs, in particular when using *de novo* OTU picking (Fig. 4a), but proportionally less so in reference-based OTU picking. When using virtual taxon (VT) sequence identities from the Maarj*AM* database (Öpik *et al.*, 2010), the denoising effect is smallest (Fig. 4b), likely because VTs represent preclustered sequences.

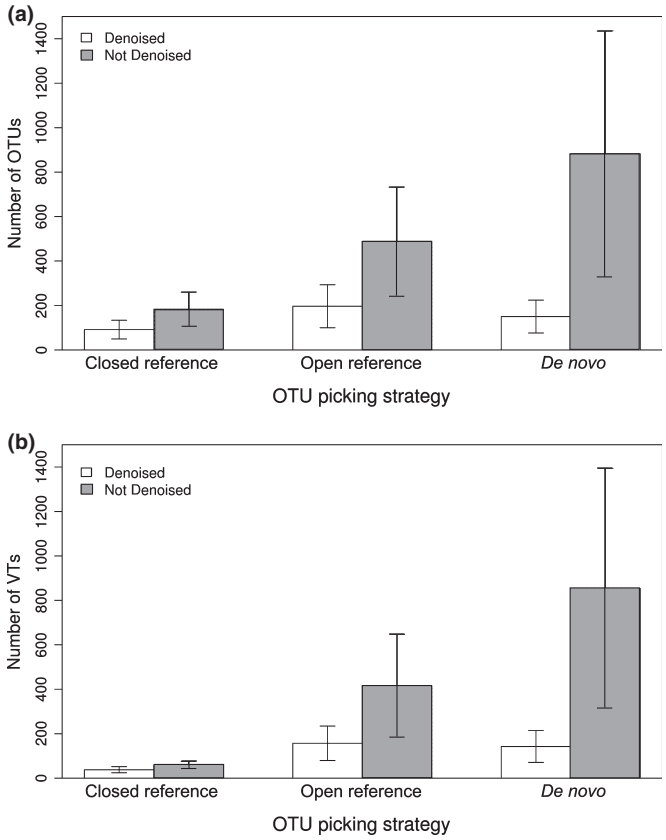

Fig. 4 The effect of sequence denoising and operational taxonomic unit (OTU) picking strategy on the number of Glomeromycotan OTUs generated. Reference databases for taxonomic assignment of OTUs was the Maarj*AM* database (a) using sequence identities as provided by original authors and (b) using Virtual Taxon (VT) identities of the sequences. When using original sequence identities (a), denoising significantly reduced the number of OTUs regardless of OTU picking strategy ($P = 0.049$). There was no effect of OTU picking strategy on the number of OTUs ($P = 0.30$). When using VT identities (b), there were fewer OTUs and whereas denoising reduced the number of OTUs, this was not significant at 5% ($P = 0.072$). Data were generated from three 454 studies, for a total of seven libraries, and 109 samples, targeting 18S rDNA, using nested primers AMDGR (Sato *et al.*, 2005) and AMADF (A. Desirò, PhD thesis, unpublished), following AML1-AML2 (Lee *et al.*, 2008) amplification. OTUs reflect 97% sequence identity. Values plotted are mean ± SE.

Reference-based OTU picking provides immediate taxonomic identity for newly generated sequences. *Open and Closed reference* OTU picking is hampered by the availability of databases for AMF sequences compared with 16S analyses, which are well-supported by databases (i.e. Greengenes, McDonald *et al.*, 2012; ARB, Ludwig *et al.*, 2004; SILVA, Quast *et al.*, 2013).

Cluster (OTU) size

Smaller sequence clusters in NGS datasets are more likely to be spurious due to read errors, undetected chimeras and other artefacts. At the same time, larger clusters may represent OTU

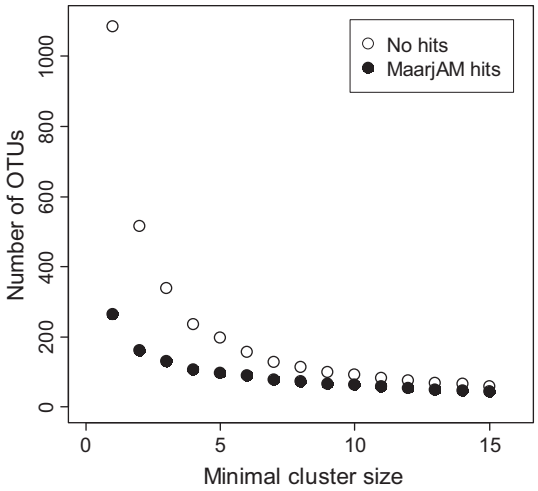

Fig. 5 Total operational taxonomic unit (OTU) richness with minimum cluster sizes ranging from 1 to 15 sequences. Small cluster thresholds resulted in a higher proportion of sequences remaining unidentified ('no hits') with Maarj*AM* database as reference. Using a previously published dataset (Hart *et al.*, 2014), we created an OTU table using an open reference OTU picking approach with a 0.97 sequence similarity threshold (UCLUST, Edgar *et al.*, 2011) and Maarj*AM* database as a reference set (Öpik *et al.*, 2010). We constructed a phylogenetic tree (MUSCLE, Edgar, 2004), and removed obvious nontarget sequences. Finally, we trimmed the dataset for different minimum cluster sizes (1, 2, 3, 4, 5, 7, 9 and 15 sequences per OTU).

'complexes', either closely related or cryptic taxa, and may be similarly uninformative. Do the generated clusters represent valid, rare OTUs (small clusters) or valid, dominant OTUs (large clusters)? By setting a minimum cluster size, fewer spurious OTUs may be created, but this practice could underestimate true levels of diversity (Fig. 5).

OTU delineation

OTU delineation is also affected by the threshold for sequence similarity. Fungal studies targeting ITS generally use a 97% identity threshold (Smith *et al.*, 2007; Bjorbækmo *et al.*, 2010; Mohamed & Martiny, 2011; Tedersoo *et al.*, 2010), though multiple OTU (species hypothesis) delineation thresholds are available in the UNITE database (Kõljalg *et al.*, 2013). AMF display extensive inter- and intraspecific divergence in ITS, LSU and possibly also SSU (Stockinger *et al.*, 2010; Thiéry *et al.*, 2012). Although the level of intraspecific variation differs among these markers, there is also variation within a single marker among AMF lineages (Stockinger *et al.*, 2010; Öpik *et al.*, 2014). The *MaarjAM* database delineates AMF VTs on the basis of phylogenetically supported clades of SSU rRNA gene sequences of 97% or higher similarity, resulting in many VT with *c.* 99% sequence identity (Öpik *et al.*, 2010). Recently, Lekberg *et al.* (2014) used a predictive model to validate two OTU delineation models (monophyletic clade vs 97% universal approach) for AMF communities based on the LSU rRNA gene. Their findings support a previous observation that fewer OTUs are observed when delineated phylogenetically rather than on the basis of sequence similarity thresholds, but this has little impact on detected AMF community patterns (Powell *et al.*, 2011).

To rarefy or not?

NGS often results in large variation in obtained sequence abundance between samples. Typically these inequalities are dealt with by standardising samples through rarefaction to a common sequencing depth per sample (e.g. minimum or median sequence numbers per sample; de Cárcer *et al.*, 2011; Hiiesalu *et al.*, 2014) or by using proportional composition (relative abundance of reads per sample; Davison *et al.*, 2012). Rarefying is so commonly applied that most analysis packages – including QIIME (Caporaso *et al.*, 2010), MOTHUR (Schloss *et al.*, 2009) and phyloseq (McMurdie & Holmes, 2013) – incorporate rarefaction code, and few researchers consider the biological and statistical implications. However, there is increasing evidence that these techniques may result in data that do not represent the original community (McMurdie & Holmes, 2014). Specifically, it has been suggested that rarefying alters the variance structure of the data and increases uncertainty through randomisation. This can lead to the following problems: increased over dispersion or decreased sensitivity, causing inflation of both Type I and Type II errors; and difficulty in developing optimum sampling protocols for previously unstudied locations and ecosystems, due to the loss of information about underlying variance patterns.

Although there is no consensus on this issue, we provide three recommendations. First, we advocate exploration of the approaches currently used in the 'RNA-Seq' pipelines (Wang *et al.*, 2009), which are optimised for sequence count data in the R-packages phyloseq (McMurdie & Holmes, 2013) and DESeq2 (Love *et al.*, 2014). These packages employ a variety of statistical solutions including noise modelling, variance stabilisation, independent filtering and Bayesian techniques, which enable analysis without increasing the error rate. Second, it is important to stay up to date with the latest statistical methods for microbial analysis, which often requires looking beyond the typical ecological journals where AMF research is published and into the rapidly advancing statistical and computational biology literature. Third, researchers should acknowledge the limitations of rarefying while we as a community work to incorporate new statistical methods into our analyses. If possible it would be useful for future studies to show the results of their analyses using both rarefaction and alternative approaches.

Sequence database(s)

An essential requirement for DNA-based identification of organisms is reference sequence datasets against which to identify the newly obtained sequences. For Glomeromycota, of the currently known *c.* 250 morphospecies (http://schuessler.userweb.mwn.de/amphylo/), only *c.* 30% have been sequenced for SSU, and *c.* 35% for the fungal barcode ITS, LSU or a combination of both (Krüger *et al.*, 2011). On the one hand, this means that the

majority of the morphologically known AMF can be detected as OTUs or VT, but cannot be identified from natural samples on the basis of DNA sequences as long as expert-identified representative cultures of these species are not sequenced. On the other hand, the majority of the environmentally detected sequences belong to taxa known only on the basis of DNA and for which morphology remains unknown (Hibbett *et al.*, 2011; Ohsowski *et al.*, 2014; Öpik *et al.*, 2014).

Species-level variation among AMF

Ostensibly, matching DNA-only species to existing AMF morphospecies could greatly improve our ability to characterize environmental sequences using databases. However, it is not a straightforward practice, as we do not know the extent of intraspecific genetic variation of Glomeromycotan ribosomal markers. This information is essential when creating guidelines for new sequence identification; more effort is needed to close this knowledge gap. Until then, pragmatic approaches must be used, including clusters defined on the basis of sequence similarity and clade support in phylogenetic analysis (AMF VT, Öpik *et al.*, 2010, 2014), on the basis of clustering algorithms (e.g. fungal species hypotheses (SH), Kõljalg *et al.*, 2013) or on the basis of computationally expensive phylogenetic modelling (Powell *et al.*, 2011).

Public sequence databases

Ideally, reference sequences should be maintained in a dedicated reference sequence database, because public sequence repositories (e.g. International Nucleotide Sequence Database Collaboration, INSDC) are not designed for identification purposes and are known to contain sequences with inaccurate identity (Bidartondo, 2008; Tedersoo *et al.*, 2011). For example, specifically curated subsets of public data repositories such as RefSeq (Pruitt *et al.*, 2007), and curated databases for specific markers or groups of organisms (e.g. UNITE, Abarenkov *et al.*, 2010) can provide high-quality sequences for reference or identification purposes. For AMF specifically, ITS sequences are available in the UNITE database, which is automatically updated twice a year from INSDC, and curated manually (Kõljalg *et al.*, 2013). Currently, UNITE + INSD contain > 8000 AMF ITS sequences. However, the majority of these sequences (67%) are identified only to genus level (http://unite.ut.ee; accessed November 2014). Full-length SSU rRNA gene sequences are available in the PHYMYCO-DB database of fungal SSU and EF1-alpha (Mahé *et al.*, 2012). The SILVA database (Quast *et al.*, 2013) contains curated and aligned sets of rRNA gene sequences from various organisms, including Glomeromycota, but the numbers of sequences are limited.

Glomeromycotan databases

The Maarj*AM* database (http://maarjam.botany.ut.ee/) is currently the only curated database dedicated to AMF sequences and associated metadata. Its sequences serve as a reference dataset for new sequence identification and the occurrence data repository provides information about AMF biogeography. Maarj*AM* database originally contained only the SSU rRNA gene sequence data, because it is the most commonly used marker for AMF diversity surveys (Öpik *et al.*, 2014), but by now it is equally complete for published sequence data from the remainder of the nuclear ribosomal cistron (ITS region, LSU rRNA gene), and also contains protein-encoding genes (actin, beta-tubulin, EF1-alpha, phosphate transporters, RPB1, RPB2) and mitochondrial marker sequences (mtLSU) (Öpik *et al.*, 2014; Maarj*AM* database status November 2014). Only sequences of central fragment of SSU rRNA gene in Maarj*AM* are classified into VT.

Metadata

Metadata accompanying sequence data allow others to ask questions which may go beyond the scope of the original study. For example, the metadata associated with VT records in the Maarj*AM* database revealed distinct distribution patterns of AMF taxa in different habitats (Öpik *et al.*, 2009), geographic regions (Moora *et al.*, 2011) and in relation to host range (Merckx *et al.*, 2012). Therefore, to provide added value to the observations of individual studies, submissions of metadata to public repositories should include a minimum number of items (Fig. 1). Further discussion of metadata guidelines can be found in Tedersoo *et al.* (2011) and Nilsson *et al.* (2011).

Data archiving

Currently, there are many data repositories for environmental community surveys, but few standard approaches for its archival. Ideally, the optimal sharing platform should facilitate the sharing of sequence data, metadata and annotations with other users. The main features of such a resource should: allow data transfer directly from a sequencing centre; be open to community or user contributions; allow data exchange with public data resources; allow integration of reference datasets; provide geolocalization of samples; and enable access to voucher specimens. At the time of writing, there are no archives matching all of these criteria, and only the NCBI SRA archive provides limited analytic tools (www.ncbi.nlm.nih.gov/sra) in addition to raw reads storage.

Conclusion

There is no doubt that NGS has transformed the way we approach AMF community research. The challenge is to tailor it to the research community's needs and tackle outstanding issues. It is essential for the research community to recognize that there is no single 'right' way of conducting innovative AMF molecular ecology. Rather, there are many alternative approaches that can be equally informative, and carry their own *caveats*. It is important for reviewers and authors to acknowledge the plurality of choices, and document clear rationale for their decisions. Given the variety of approaches and platforms currently being used, we suggest that future datasets will have greater value to the research community if researchers consider the following points. (1) *Provide sufficient detail about OTU occurrence*: Submitting a single, representative sequence per OTU without frequency information omits valuable

ecological context. In order to maximize the impact of data from individual studies, it is essential to record OTU occurrence per location, host and sample type. (2) *Target more geographic regions and ecosystems*: Sequence-based occurrence data of Glomeromycota are strongly dominated by studies from Europe and North America (Öpik *et al.*, 2010), although focused sampling effort in less targeted regions and habitats has provided a large increase in global Glomeromycota species richness (Öpik *et al.*, 2013). Therefore, sequences from more locations, habitats and host species will further improve the reference datasets in public data repositories, and contribute to a more complete biogeographic knowledge about AMF. (3) *Generate more informative sequences*: Long sequences that link different markers within a gene region (e.g. SSU, ITS and LSU markers of AMF) would improve comparability of data obtained with alternative markers. (4) *Sequence existing morphospecies*: Systematic generation of reference sequences via deeper sequencing per sample, sequencing multiple spores per culture, cultures per species and multiple markers would help link morphospecies with sequence data. These goals would most easily be fulfilled by targeting existing accessions in culture collections (e.g. INVAM, BEG etc.), using cloning and double-stranded Sanger sequencing, with nucleotide variation mapped by deep sequencing using NGS approaches.

Although NGS has been a boon to AMF community ecology, it is not the only vehicle for understanding these communities. As technologies are increasingly transient, it is even more important to remember the approaches that came before, and explore those used in other fields of ecology, because these may hold the key to answering the questions at hand. Indeed, the most meaningful way forward may require incorporating old approaches, such as autecology, to make molecular fungal ecology truly relevant (Peay, 2014).

Acknowledgements

M.M.H. would like to thank the University of British Columbia Okanagan (Internal Grants) for funding the workshop. Thanks also to Alicia Tymstra, John Klironomos and Hiltrud Volger for administrative help. S.G. thanks RISINNOVA Project, grant no. 2010-2369, AGER Foundation; PURE, grant no. FP7-265865, European Community's Seventh Framework Programme (FP7/2007-2013), and ECOFLOR, Piemonte Region. Y.L. thanks MPG Ranch for funding. M.Ö. is funded by the Estonian Science Foundation (grants 9050, IUT20-28) and the European Regional Development Fund (Center of Excellence FIBIR). T.H. is funded by the Natural Environment Research Council (NERC) grant NE/M004864/1 and The Biotechnology and Biosciences Research Council (BBSRC) grant BB/L026007/1. B.J.P. thanks the Simard and Mohn labs at UBC Vancouver for funding. We also thank anonymous reviewers for constructive comments on the earlier versions of the paper.

References

Abarenkov K, Nilsson RH, Larsson K, Alexander IJ, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R *et al.* 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* **186**: 281–285.

Anderson IC, Parkin PI. 2007. Detection of active soil fungi by RT-PCR amplification of precursor rRNA molecules. *Journal of Microbiological Methods* **68**: 248–253.

Bahram M, Peay KG, Tedersoo L. 2015. Local-scale biogeography and spatiotemporal variability in communities of mycorrhizal fungi. *New Phytologist* **205**: 1454–1463.

Bainard LD, Klironomos JN, Hart MM. 2010. Differential effect of sample preservation methods on plant and arbuscular mycorrhizal fungal DNA. *Journal of Microbiological Methods* **82**: 124–130.

Bever JD, Westover KM, Antonovics J, Westover M. 1997. Incorporating the soil community into plant population dynamics: the utility of the feedback approach. *Journal of Ecology* **85**: 561–573.

Bidartondo AMI, Arnosti C, Rothman DH, Forney DC. 2008. Preserving accuracy in Genbank. *Science* **319**: 1616.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. 2012. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* **27**: 233–243.

Bjorbækmo MFM, Carlsen T, Brysting A, Vrålstad T, Høiland K, Ugland KI, Geml J, Schumacher T, Kauserud H. 2010. High diversity of root associated fungi in both alpine and arctic *Dryas octopetala*. *BMC Plant Biology* **10**: 244.

Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. 2013. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME Journal* **7**: 2061–2068.

Borriello R, Bianciotto V, Orgiazzi A, Lumini E, Bergero R. 2014. Sequencing and comparison of the mitochondrial COI gene from isolates of arbuscular mycorrhizal fungi belonging to Gigasporaceae and Glomeraceae families. *Molecular Phylogenetics and Evolution* **75**: 1–10.

Burleigh SH, Cavagnaro T, Jakobsen I. 2002. Functional diversity of arbuscular mycorrhizas extends to the expression of plant genes involved in P nutrition. *Journal of Experimental Botany* **53**: 1593–1601.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335–336.

de Cárcer DA, Denman SE, McSweeney C, Morrison M. 2011. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. *Applied and Environmental Microbiology* **77**: 8795–8798.

Clapp JP, Fitter AH, Young JPW. 1999. Ribosomal small subunit variation within spores of an arbuscular mycorrhizal fungus, *Scutellospora* sp. *Molecular Ecology* **8**: 915–921.

Clapp JP, Young JP, Merryweather JW, Fitter AH. 1995. Diversity of fungal symbionts in arbuscular mycorrhizas from a natural community. *New Phytologist* **130**: 259–265.

Cotton TEA, Fitter AH, Miller RM, Dumbrell AJ, Helgason T. 2015. Fungi in the future: inter-annual variation and effects of atmospheric change on arbuscular mycorrhizal fungal communities. *New Phytologist* **205**: 1598–1607.

Daniell TJ, Husband R, Fitter AH, Young JPW. 2001. Molecular diversity of arbuscular mycorrhizal fungi colonizing arable crops. *FEMS Microbiology Ecology* **36**: 203–209.

Davison J, Öpik M, Daniell TJ, Moora M, Zobel M. 2011. Arbuscular mycorrhizal fungal communities in plant roots are not random assemblages. *FEMS Microbiology Ecology* **78**: 103–115.

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, Alastair H. 2011. Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *New Phytologist* **190**: 794–804.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH. 2010. Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME Journal* **4**: 337–345.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

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

Ettema CH, Wardle DA. 2002. Spatial soil ecology. *Trends in Ecology & Evolution* **17**: 177–183.

Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, Creer S. 2012. Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Research* **40**: e66.

Gaspar JM, Thomas WK. 2013. Assessing the consequences of denoising marker-based metagenomic data. *PLoS ONE* **8**: e60458.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E *et al.* 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21**: 494–504.

Hart MM, Gorzelak M, Ragone D, Murch SJ. 2014. Arbuscular mycorrhizal fungal succession in a long-lived perennial. *Botany-Botanique* **320**: 313–320.

Hart MM, Reader RJ. 2002. Taxonomic basis for variation in the colonization strategy of arbuscular mycorrhizal fungi. *New Phytologist* **153**: 335–344.

van der Heijden MGA, Scheublin TR. 2007. Functional traits in mycorrhizal ecology: their use for predicting the impact of arbuscular mycorrhizal fungal communities on plant growth and ecosystem functioning. *New Phytologist* **174**: 244–250.

Helgason T, Daniell TJ, Husband R, Fitter AH, Young JPW. 1998. Ploughing up the wood-wide web? *Nature* **394**: 431.

Hempel S, Renker C, Buscot F. 2007. Differences in the species composition of arbuscular mycorrhizal fungi in spore, root and soil communities in a grassland ecosystem. *Environmental Microbiology* **9**: 1930–1938.

Hibbett DS, Ohman A, Glotzer D, Nuhn M, Kirk P, Nilsson RH. 2011. Progress in molecular and morphological taxon discovery in fungi and options for formal classification of environmental sequences. *Fungal Biology Reviews* **25**: 38–47.

Hiiesalu I, Meelis P, Davison J, Gerhold P, Metsis M, Moora M, Öpik M, Vasar M, Zobel M, Wilson SD. 2014. Species richness of arbuscular mycorrhizal fungi: associations with grassland plant richness and biomass. *New Phytologist* **203**: 233–244.

Holland TC, Bowen P, Bogdanoff C, Hart MM. 2014. How distinct are arbuscular mycorrhizal fungal communities associating with grapevines? *Biology and Fertility of Soils* **50**: 667–674.

Jakobsen I, Abbott LK, Robson AD. 1992. External hyphae of vesicular-arbuscular mycorrhizal fungi associated with *Trifolium subterraneum* L. 1. Spread of hyphae and phosphorus inflow into roots. *New Phytologist* **120**: 371–380.

Janoušková M, Püschel D, Hujslová M, Slavíková R, Jansa J. 2015. Quantification of arbuscular mycorrhizal fungal DNA in roots: how important is material preservation? *Mycorrhiza*. doi:10.1007/s00572-014-0602-7.

Karst J, Piculell B, Bringham C, Booth M, Hoeksema J. 2013. Fungal communities in soils along a vegetative ecotone. *Mycologia* **105**: 61–70.

Kjøller R, Rosendahl S. 2000. Detection of arbuscular mycorrhizal fungi (Glomales) in roots by nested PCR and SSCP (single stranded conformation polymorphism). *Plant and Soil* **226**: 189–196.

Kohout P, Sudová R, Janoušková M, Čtvrtlíková M, Hejda M, Pánková H, Slavíková R, Štajerová K, Vosátka M, Sýkorová Z. 2014. Comparison of commonly used primer sets for evaluating arbuscular mycorrhizal fungal communities: is there a universal solution? *Soil Biology and Biochemistry* **68**: 482–493.

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Mohammad B, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B *et al.* 2013. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* **22**: 5271–5277.

Krüger M, Krüger C, Walker C, Stockinger H, Schüßler A. 2011. Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. *New Phytologist* **193**: 907–984.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**: 118–123.

La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD. 2012. Hypothesis testing and power calculations for taxonomic-based human microbe data. *PLoS ONE* **7**: e52078.

La Rosa PS, Zhou Y, Sodergren E, Weinstock G, Shannon WD. 2015. Hypothesis testing of metagenomic data. In: Izard J, Rivera M, eds. *Metagenomics for microbiology*. Waltham, MA, USA: Academic Press, 81–96.

Lee J, Lee S, Young JPW. 2008. Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. *FEMS Microbiology Ecology* **65**: 339–349.

Legendre P, Legendre L. 1998. *Numerical ecology*. Amsterdam, the Netherlands: Elsevier Science.

Lekberg Y, Gibbons SM, Rosendahl S. 2014. Will different OTU delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns ? *New Phytologist* **202**: 1101–1104.

Lekberg Y, Gibbons SM, Rosendahl S, Ramsay PW. 2013. Severe plant invasions can increase mycorrhizal fungal abundance and diversity. *ISME Journal* **7**: 1424–1433.

Lekberg Y, Meadow J, Rohr JR, Redecker D, Zabinski CA. 2011. Importance of dispersal and thermal environment for mycorrhizal communities: lessons from Yellowstone National Park. *Ecology* **96**: 1292–1302.

Lekberg Y, Schnoor T, Kjøller R, Gibbons SM, Hansen LH, Al-Soud WA, Sørensen SJ, Rosendahl S. 2012. 454-sequencing reveals stochastic local reassembly and high disturbance tolerance within arbuscular mycorrhizal fungal communities. *Journal of Ecology* **100**: 151–160.

Li T, Hu YJ, Hao ZP, Li H, Wang YS, Chen BD. 2013. First cloning and characterization of two functional aquaporin genes from an arbuscular mycorrhizal fungus *Glomus intraradices*. *New Phytologist* **197**: 617–630.

Lilleskov EA, Bruns TD, Horton TR, Taylor D, Grogan P. 2004. Detection of forest stand-level spatial structure in ectomycorrhizal fungal communities. *FEMS Microbiology Ecology* **49**: 319–332.

Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, Pang E, Cao H, Cha H, Lin T *et al.* 2014. Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genetics* **10**: e1004078.

Lindahl BD, Kuske CR. 2013. Metagenomics for study of fungal ecology. In: Martin F, ed. *The ecological genomics of fungi*. Chichester, UK: John Wiley & Sons, Inc., 281–303.

Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjøller R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J *et al.* 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist* **199**: 288–299.

Lloyd-MacGilp SA, Chambers SM, Dodd JC, Fitter AH, Walker C, Young JP. 1996. Diversity of the ribosomal internal transcribed spacers within and among isolates of *Glomus mosseae* and related mycorrhizal fungi. *New Phytologist* **133**: 103–111.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2". *Genome Biology* **15**: 550.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar , Buchner A, Lai T, Steppi S, Jobb G *et al.* 2004. ARB: a software environment for sequence data. *Nucleic Acids Research* **32**: 1363–1371.

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. *Environmental Microbiology* **12**: 2165–2179.

Mahé S, Duhamel M, Le Calvez T, Guillot L, Sarbu L, Bretaudeau A, Collin O, Dufresne A, Kiers ET, Vandenkoornhuyse P. 2012. PHYMYCO-DB: a curated database for analyses of fungal diversity and evolution. *PLoS ONE* **7**: e43117.

Manian S, Sreenivasaprasad S, Mills PR. 2001. DNA extraction method for PCR in mycorrhizal fungi. *Letters in Applied Microbiology* **33**: 307–310.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal* **6**: 610–618.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**: e61217.

McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology* **10**: e1003531.

Merckx VSFT, Janssens SB, Hynson NA, Specht CD, Bruns TD, Smets EF. 2012. Mycoheterotrophic interactions are not limited to a narrow phylogenetic range of arbuscular mycorrhizal fungi. *Molecular Ecology* **21**: 1524–1532.

Mohamed DJ, Martiny JBH. 2011. Patterns of fungal diversity and composition along a salinity gradient. *ISME Journal* **5**: 379–388.

Moora M, Berger S, Davison J, Öpik M, Bommarco R, Bruelheide H, Kühn I, Kunin WE, Metsis M, Rortais A *et al.* 2011. Alien plants associate with widespread generalist arbuscular mycorrhizal fungal taxa: evidence from a continental-scale study using massively parallel 454 sequencing. *Journal of Biogeography* **38**: 1305–1317.

Mummey DL, Rillig MC. 2008. Spatial characterization of arbuscular mycorrhizal fungal molecular diversity at the submetre scale in a temperate grassland. *FEMS Microbiology Ecology* **64**: 260–270.

Neumann E, George E. 2005. Extraction of extraradical arbuscular mycorrhizal mycelium from compartments filled with soil and glass beads. *Mycorrhiza* **15**: 533–537.

Nguyen NH, Smith D, Peay K, Kennedy P. 2014. Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* **205**: 1389–1393.

Nilsson RH, Tedersoo L, Lindahl BD, Kjøller R, Carlsen T, Quince C, Abarenkov K, Pennanen T, Stenlid J, Bruns T *et al.* 2011. Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. *New Phytologist* **191**: 314–318.

O'Brien HE, Parrent JL, Jackson JA, Moncalvo J, Vilgalys R. 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology* **71**: 5544–5550.

O'Donnell K, Sutton DA, Rinaldi MG, Sarver BA, Balajee SA, Schroers H, Summerbell RC, Robert VARG, Crous PW, Zhang N *et al.* 2010. Internet-accessible DNA sequence database for identifying *Fusaria* from human and animal infections. *Journal of Clinical Microbiology* **48**: 3708–3718.

Oehl F, Sieverding E, Ineichen K, Ris E, Boller T, Wiemken A. 2003. Community structure of arbuscular mycorrhizal fungi at different soil depths in extensively and intensively managed agroecosystems. *New Phytologist* **165**: 273–283.

Ohsowski BM, Zaitsoff PD, Öpik M, Hart MM. 2014. Where the wild things are: looking for uncultured Glomeromycota. *New Phytologist* **204**: 171–179.

Öpik M, Davison J, Moora M, Zobel M. 2014. DNA-based detection and identification of Glomeromycota : the virtual taxonomy of environmental sequences. *Botany-Botanique* **92**: 135–147.

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.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database Maarj AM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist* **188**: 223–241.

Öpik M, Zobel M, Cantero JJ, Davison J, Facelli JM, Hiiesalu I, Jairus T, Kalwij JM, Koorem K, Leal ME *et al.* 2013. Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. *Mycorrhiza* **23**: 411–430.

Peay KG. 2014. Back to the future: natural history and the way forward in modern fungal ecology. *Fungal Ecology* **12**: 4–9.

Pickles BJ, Genney DR, Anderson IC, Alexander IJ. 2012. Spatial analysis of ectomycorrhizal fungi reveals that root tip communities are structured by competitive interactions. *Molecular Ecology* **21**: 5110–5123.

Pickles BJ, Pither J. 2014. Still scratching the surface: how much of the "black box" of soil ectomycorrhizal communities remains in the dark? *New Phytologist* **201**: 1101–1105.

Powell JR, Monaghan MT, Öpik M, Rillig MC. 2011. Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology* **20**: 655–666.

Prosser JI. 2010. Replicate or lie. *Environmental Microbiology* **12**: 1806–1810.

Prosser JI, Bohannan BJM, Curtis TP, Ellis RJ, Firestone MK, Freckleton RP, Green JL, Green LE, Killham K, Lennon JJ *et al.* 2007. The role of ecological theory in microbial ecology. *Nature* **5**: 384–392.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**(Suppl. 1): D61–D65.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**: D590–D596.

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

Reeder J, Knight R. 2010. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nature Methods* **7**: 668–669.

Rosendahl S, Stukenbrock EH. 2004. Community structure of arbuscular mycorrhizal fungi in undisturbed vegetation revealed by analyses of LSU rDNA sequences. *Molecular Ecology* **13**: 3179–3186.

Saks Ü, Davison J, Öpik M, Vasar M, Moora M, Zobel M. 2014. Root-colonizing and soil-borne communities of arbuscular mycorrhizal fungi in a temperate forest understorey. *Botany-Botanique* **92**: 277–285.

Sanders IR. 2004. Plant and arbuscular mycorrhizal fungal diversity – are we looking at the relevant levels of diversity and are we using the right techniques? *New Phytologist* **164**: 415–418.

Sato K, Suyama Y, Saito M, Sugawara K. 2005. A new primer for discrimination of arbuscular mycorrhizal fungi with polymerase chain reaction-denature gradient gel electrophoresis. *Grassland Science* **51**: 179–181.

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

Schmidt PA, Balint M, Greshake B, Blandow C, Römbke J, Schmitt I. 2013. Illumina metabarcoding of a soil fungal community. *Soil Biology & Biochemistry* **65**: 128–132.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences, USA* **109**: 6241–6246.

Schwarzott D, Schüßler A. 2001. A simple and reliable method for SSU rRNA gene DNA extraction, amplification, and cloning from single AM fungal spores. *Mycorrhiza* **10**: 203–207.

Sikes BA, Maherali H, Klironomos JN. 2012. Arbuscular mycorrhizal fungal communities change among three stages of primary sand dune succession but do not alter plant growth. *Oikos* **121**: 1791–1800.

Smith SE, Read DJ. 2008. *Mycorrhizal symbiosis, 3rd edn.* Waltham, MA, USA: Academic Press.

Smith ME, Douhan GW, Rizzo DM. 2007. Ectomycorrhizal community structure in a xeric *Quercus* woodland based on rDNA sequence analysis of sporocarps and pooled roots. *New Phytologist* **174**: 847–863.

Stockinger H, Krüger M, Schüßler A. 2010. DNA barcoding of arbuscular mycorrhizal fungi. *New Phytologist* **187**: 461–474.

Stockinger H, Peyret-Guzzon M, Koegel S, Bouffaud M-L, Redecker D. 2014. The largest subunit of RNA Polymerase II as a new marker gene to study assemblages of arbuscular mycorrhizal fungi in the field. *PLoS ONE* **9**: e107783.

Sýkorová Z, Ineichen K, Wiemken A, Redecker D. 2007. The cultivation bias: different communities of arbuscular mycorrhizal fungi detected in roots from the field, from bait plants transplanted to the field, and from a greenhouse trap experiment. *Mycorrhiza* **18**: 1–14.

Tedersoo L, Abarenkov K, Nilsson RH, Schüssler A, Grelet G-A, Kohout P, Oja J, Bonito GM, Veldre V, Jairus T *et al.* 2011. Tidying up international nucleotide sequence databases: ecological, geographical and sequence quality annotation of its sequences of mycorrhizal fungi. *PLoS ONE* **6**: e24940.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Mohammad B, Bechem E, Chuyong G, Kõljalg U. 2010. 454 Pyrosequencing and Sanger

sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* **188**: 291–301.

Thiéry O, Moora M, Vasar M, Zobel M, Öpik M. 2012. Inter- and intrasporal nuclear ribosomal gene sequence variation within one isolate of arbuscular mycorrhizal fungus, *Diversispora* sp. *Symbiosis* **58**: 135–147.

Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charron P, Duensing N, Frei dit Frey N, Gianinazzi-Pearson V *et al.* 2013. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences, USA* **110**: 20 117–20 122.

Van Geel M, Busschaert P, Honnay O, Lievens B. 2014. Evaluation of six primer pairs targeting the nuclear rRNA operon for characterization of arbuscular mycorrhizal fungal (AMF) communities using 454 pyrosequencing. *Journal of Microbiological Methods* **106**: 93–100.

Vandenkoornhuyse P, Husband R, Daniell TJ, Watson IJ, Duck JM, Fitter AH, Young JP. 2002. Arbuscular mycorrhizal community composition associated with two plant species in a grassland ecosystem. *Molecular Ecology* **11**: 1555–1564.

Verbruggen E, Kuramae EE, Hillekens R, de Hollander M, Kiers ET, Röling WFM, Kowalchuk GA, van der Heijden MGA. 2012. Testing potential effects of maize expressing the *Bacillus thuringiensis* Cry1Ab endotoxin (Bt maize) on mycorrhizal fungal communities via DNA- and RNA-based pyrosequencing and molecular fingerprinting. *Applied and Environmental Microbiology* **78**: 7384–7392.

Wallander H, Nilsson LO, Hagerberg D, Bååth E. 2001. Estimation of the biomass and seasonal growth of external mycelium of ectomycorrhizal fungi in the field. *New Phytologist* **151**: 753–760.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57–63.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Published primers most frequently used for arbuscular mycorrhizal fungi (AMF) studies

Methods S1 Power analysis for multivariate data.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.