

Increased sequencing depth does not increase captured diversity of arbuscular mycorrhizal fungi

Martti Vasar¹  · Reidar Andreson^{2,3} · John Davison¹ · Teele Jairus¹ · Mari Moora¹ · Mairo Remm² · J. P. W. Young⁴ · Martin Zobel¹ · Maarja Öpik¹

Received: 5 May 2017 / Accepted: 12 July 2017 / Published online: 20 July 2017
© Springer-Verlag GmbH Germany 2017

Abstract The arrival of 454 sequencing represented a major breakthrough by allowing deeper sequencing of environmental samples than was possible with existing Sanger approaches. Illumina MiSeq provides a further increase in sequencing depth but shorter read length compared with 454 sequencing. We explored whether Illumina sequencing improves estimates of arbuscular mycorrhizal (AM) fungal richness in plant root samples, compared with 454 sequencing. We identified AM fungi in root samples by sequencing amplicons of the SSU rRNA gene with 454 and Illumina MiSeq paired-end sequencing. In addition, we sequenced metagenomic DNA without prior PCR amplification. Amplicon-based Illumina sequencing yielded two orders of magnitude higher sequencing depth per sample than 454 sequencing. Initial analysis with minimal quality control recorded five times higher AM fungal richness per sample with Illumina sequencing. Additional quality control of Illumina samples, including restriction of the marker region to the most variable amplicon fragment, revealed AM fungal richness values close to those produced by 454 sequencing. Furthermore, AM fungal richness estimates were not correlated

with sequencing depth between 300 and 30,000 reads per sample, suggesting that the lower end of this range is sufficient for adequate description of AM fungal communities. By contrast, metagenomic Illumina sequencing yielded very few AM fungal reads and taxa and was dominated by plant DNA, suggesting that AM fungal DNA is present at prohibitively low abundance in colonised root samples. In conclusion, Illumina MiSeq sequencing yielded higher sequencing depth, but similar richness of AM fungi in root samples, compared with 454 sequencing.

Keywords Arbuscular mycorrhiza · Illumina sequencing · 454 Sequencing · Community diversity

Introduction

Arbuscular mycorrhizal (AM) fungi (Phylum Mucoromycota, Subphylum Glomeromycotina, Spatafora et al. 2016) are widespread soil microorganisms that form a symbiosis with the roots of the great majority of land plants (Smith and Read 2008). The fungal partner provisions host plants with soil nutrients and improved resistance to abiotic and biotic stresses; in exchange, host plants supply the fungi with carbon derived from photosynthesis (Pozo et al. 2015; Garcia et al. 2016). Thus, AM fungi play critical roles in global nutrient and carbon cycling and ecosystem functioning (van der Heijden et al. 2015; Treseder 2016). Consequently, considerable effort has already been made to understand patterns of AM fungal diversity at different spatial scales (e.g. Davison et al. 2015; Öpik and Davison 2016) and how AM fungal diversity relates to ecosystem functioning (van der Heijden et al. 2015; Treseder 2016).

Detecting and identifying AM fungi in natural systems is far from straightforward because the organisms are

Electronic supplementary material The online version of this article (doi:10.1007/s00572-017-0791-y) contains supplementary material, which is available to authorized users.

✉ Martti Vasar
martti.vasar@ut.ee

¹ Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai Str, 51005 Tartu, Estonia

² Institute of Molecular and Cell Biology, University of Tartu, 23b Riia Str, 51010 Tartu, Estonia

³ Estonian Biocentre, 23b Riia Str, 51010 Tartu, Estonia

⁴ Department of Biology, University of York, York YO10 5DD, UK

microscopic and at best culturable together with a host plant (Smith and Read 2008). The long-established method for taxonomic identification on the basis of spore morphology has been accompanied, and in some fields replaced, by DNA-based methods targeting fungal structures in plant roots and soil (Öpik and Davison 2016). Since the 1990s, cloning and Sanger sequencing has been used to identify naturally occurring AM fungi, with studies typically generating tens to thousands of sequences and tens to hundreds of sequences per sample (e.g. Öpik et al. 2008). The early 2000s saw the emergence of 454 sequencing, as well as other next-generation sequencing (NGS) platforms, which allowed sequencing depth (number of sequences per sample) to be increased to hundreds to thousands of sequences per sample (Shokralla et al. 2012). This technological development has allowed a substantial increase in sample throughput, finally making it feasible to study variation in communities of AM fungi and other fungi (Lindahl et al. 2013; Hart et al. 2015; Taylor et al. 2017). Estimates of AM fungal richness increased concomitantly with increasing sample sizes and sequencing depth per sample, indicating that the new approaches allowed infrequent fungi to be detected and the proportions of diversity captured to increase (Öpik et al. 2006, 2009).

454 Sequencing as a tool is now being replaced by several high-throughput sequencing platforms, including Illumina, Ion Torrent and PacBio SMRT sequencing in fungal (Brown et al. 2013; Glassman et al. 2015; Tedersoo et al. 2015; Cline et al. 2017) and AM fungal diversity research (Johansen et al. 2016; Schlaeppi et al. 2016). Illumina sequencing represents a particularly useful approach because of its lower error rate than 454 sequencing and the further increase in sequencing depth per sample—by up to two orders of magnitude—it offers (Shokralla et al. 2012). Consequently, it is becoming a favoured approach in fungal community ecology. However, the relatively short maximum sequence length returned by Illumina platforms has been a limitation for AM fungal diversity studies, where commonly used amplicon lengths range from 540 to 1500 bp (Öpik and Davison 2016). The use of alternative platforms that allow longer reads, such as PacBio (Schlaeppi et al. 2016), or the increase in Illumina MiSeq read length to 2×300 bp and approaches, such as tagmentation, that permit sequencing longer DNA fragments, are partly overcoming this limitation. Tagmentation incorporates universal tags into DNA fragments of user-selected size via transposon cleavage (Lee et al. 2016), allowing longer fragments, such as the full ribosomal small subunit RNA genes of bacteria, to be sequenced (Burke and Darling 2016).

Illumina sequencing has been used to study AM fungal diversity in field and experimental systems (Cui et al. 2016; Johansen et al. 2016; Xu et al. 2016; Liu et al. 2017; Orchard et al. 2017; Wang et al. 2016). In some cases, relatively low AM fungal diversity has been detected in comparison with the levels typically returned by 454 sequencing-based studies,

though this could reflect a real effect of the extreme environments being studied (Cui et al. 2016). On the other hand, a study of AM fungal communities in dune grass roots showed that Illumina sequencing can recover AM fungi from most root samples while providing broad phylogenetic coverage of clades (Johansen et al. 2016). Results from arid ecosystems in Northern China (Liu et al. 2017) and from maize roots (Wang et al. 2016) also showed that the techniques can generate large quantities of AM fungal reads (>70% of quality-controlled Illumina paired-end sequences were from Glomeromycotina). However, it has not yet been assessed how the Illumina sequencing approach, with increased sequencing depth per sample, a change in amplicon length and associated differences in bioinformatic procedures, affects the capture of AM fungal diversity and the accuracy of AM fungal identification compared with the relatively well tested 454 sequencing approach.

Platforms such as Illumina also allow a metagenomic (i.e. PCR-free) approach to be used when describing microbial communities. This allows direct identification of the microbiome without potential PCR-introduced bias, whereby certain groups are favoured or disfavoured in the amplification process, for example due to primer bias (Bradley et al. 2016). Furthermore, metagenomic sequence data includes entire genomes and is largely quantitative proportionally to the template DNA (Thomas et al. 2012), potentially allowing precise identification, functional profiling and quantification (Randle-Boggis et al. 2016). While community surveys of AM fungi still rely almost exclusively on markers derived from the nuclear ribosomal operon, it would be informative to assess the feasibility of applying a metagenomic approach to study the ribosomal gene diversity of AM fungi in environmental samples.

Here, we use the Illumina MiSeq sequencing platform using tagmentation with PCR-based and metagenomic approaches to describe AM fungal diversity in natural plant root samples and compare the results with those generated using 454 sequencing. We aim to (1) determine whether the choice of sequencing approach (Illumina MiSeq and 454 sequencing) affects measurements of AM fungal diversity, (2) determine whether amplicon-based and metagenomic Illumina MiSeq sequencing capture AM fungal diversity differently, and (3) identify the most variable region of SSU rRNA gene.

Materials and methods

Plant root sampling

Plant root sampling was conducted in the Järvselja forest reserve in south-eastern Estonia (58° 17.916' N, 27° 15.744' E), as described in Saks et al. (2014). We used different plant individuals from the same sampling event as in Saks et al.

(2014). Samples were collected from two plots of 30 × 30 m, ca 50 m apart, in June 2009. Two arbuscular mycorrhizal plant species that were abundant in the plots were sampled for this study: *Convallaria majalis* L. and *Rubus saxatilis* L. Three random individuals of each plant species per plot were excavated (a total of 12 samples). Plant roots were then carefully cleaned of soil and other adhering materials as quickly as possible, wrapped in tissue paper and stored in a plastic bag with silica gel.

DNA extraction

DNA was extracted from each plant individual from 30 to 80 mg of dried roots (fragments from the root system) using the PowerSoil-htp™ 96 Well Soil DNA Isolation Kit (MO BIO Laboratories, Inc., Carlsbad, CA, USA) with some modifications as in Öpik et al. (2013). In short, in order to include a sufficient quantity of roots in each DNA extraction, roots were milled to powder in 2-ml tubes with one or two 3-mm tungsten carbide beads per tube with Mixer Mill MM400 (Retsch GmbH, Haan, Germany), instead of milling in the bead plate as suggested by the manufacturer. Seven hundred fifty microliters of bead solution was added to the tubes, mixed, and the slurry transferred to the bead plate. Also, to increase the DNA yield, bead plates were shaken at 60 °C (as suggested by the manufacturer as a modification of the standard protocol, which uses room temperature) for 10 min at 150 rpm in a shaking incubator. Finally, in order to increase DNA yield but maintain DNA concentration, final elution was performed twice with 75 µl of Solution C6.

Illumina sequencing

DNA from root samples was subjected to amplicon-based and metagenomic-based Illumina sequencing on a MiSeq platform, using a 2 × 250 bp paired-read sequencing approach.

Amplicons of the nuclear SSU rRNA gene were generated with primers NS31 and AML2 (Simon et al. 1992; Lee et al. 2008) for identification of AM fungi as in Öpik et al. (2013). Amplified DNA was purified with the Qiagen QIAquick Gel Extraction kit (Qiagen GmbH, Hilden, Germany) following the manufacturer's protocol. The concentration of purified DNA was quantified with a Qubit® 2.0 Fluorometer (Invitrogen, Grand Island, USA).

For both amplicon-based and metagenomic approaches, 1 ng of sample DNA was processed for the sequencing libraries using the Illumina Nextera XT sample preparation kit (Illumina, San Diego, USA) following the manufacturer's protocol. This approach uses tagmentation to enzymatically fragment and tag the sample DNA with adapters in random positions. This step is followed by a 12-cycle PCR in order to add the sequencing indices, PCR cleanup and normalisation.

Normalised libraries were pooled and sequenced on Illumina MiSeq with 2 × 250 bp paired-end reads.

454 Sequencing

AM fungal amplicons were generated with the same primers and PCR conditions as described above. 454 Sequencing was performed as in Öpik et al. (2013) using the primers NS31 and AML2 linked to 454 sequencing adaptors A and B, respectively. In order to identify sequences originating from different samples, we used a set of 8 bp barcodes designed following Parameswaran et al. (2007). The barcode sequences were inserted between the A primer and NS31 primer sequences and between the B primer and AML2 primer sequences. Thus, the composite forward primer was 5' *GTCT CCGACTCAG* (NNNNNNNN) *TTGGAGGGCAAGTC TGGTGCC* 3' and the reverse primer 5' *TTGG CAGTCTCAG* (NNNNNNNN) *GAACCCAAACAATT TGGTTCC* 3', where the A and B adaptors are underlined, the barcode is indicated by N-s in parentheses, and the specific primers NS31 and AML2 are shown in italics. The ten-times diluted product of the first PCR reaction was used in the second PCR with primers A (5'-CCATCTCATCCCTGCGTGCTCCGACTCAG-3') and B (5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAG-3'). The PCR conditions were as in Öpik et al. (2013). The resulting DNA mix was sequenced on a Genome Sequencer FLX System, using Titanium Series reagents (Roche Applied Science, Mannheim, Germany) at GATC Biotech (Konstanz, Germany).

Bioinformatics

Database and bioinformatics scripts to prepare, clean and analyse the data are available in a GitHub repository (<http://github.com/ut-planteco/ssu-pipeline>). Scripts were written in Python 2.7 using a minimal approach with no dependencies and external libraries needed. Users need to install the following third-party software: MAFFT (v7.306, Katoh and Standley 2013) to generate multiple sequence alignment, FLASH (v1.2.10, Magoč and Salzberg 2011) to combine paired-end reads, USEARCH (v7.0.1090, Edgar et al. 2011) to remove chimeric reads, legacy BLAST with blastclust (BLAST v2.2.26, Altschul et al. 1990) to cluster sequences and BLAST (BLAST+ v2.5.0, Camacho et al. 2009) to identify sequences. The workflow for analysing 454 and Illumina sequences is summarised in Fig. S1 (Electronic Supplementary Material).

Quality control and preparation of read sets Illumina and 454 sequences were divided into four sets to incorporate the most common analysis and cleaning protocols for both sequencing methods. For Illumina amplicon reads, we generated the following sets: (1) Ill-qf, quality-filtered single reads (R1)

representing a random collection of forward and reverse direction reads, and (2) Ill-t, paired-end reads that were combined using quality-filtered forward (R1) and reverse (R2) reads and then trimmed to retain only the most variable section of the amplicon (170–300 bp). For 454 reads, we generated partly analogous set: (3) 454-qf, quality-filtered full-length reads, and (4) 454-t, quality-filtered reads that were trimmed to retain only the most variable amplicon section.

The Illumina amplicon read set was quality-filtered by removing sequences exhibiting Nextera adapter contamination (Fig. S2a, S2b) and where average quality was <30 (maximum 41, Cock et al. 2010). The majority of reads contained Nextera tagmentation-based adapter fragments (Figs. S2, S3, S4). Adapter fragments were revealed at random locations (Fig. S2d) dependent on insert size, which sometimes was smaller than MiSeq read length, resulting in palindrome reads where adapter read-through occurs (Bolger et al. 2014). Such paired-end reads containing fragments of adapters were discarded. Sequences were not trimmed based on quality scores as this results in the loss of combined reads if overlapping sections are removed, while the criteria required for read combination themselves provide a quality filter. In our case, R2 reads were of lower quality compared to R1 (average quality scores: amplicon R1 29.9% and R2 27.6%; metagenomic R1 35% and R2 31.9%; Fig. S5). After adapter removal and quality filtering, reads without the respective paired read were omitted, as they could not be combined.

Quality-filtered paired-end reads were combined using FLASH with the default parameters (minimum overlap between paired reads = 10 bp; maximum mismatch density in the overlap alignment = 0.25). In case of mismatch in the overlapping segments of paired-end reads, FLASH compares the quality scores for both reads and picks the nucleotide with the highest quality score. The resultant combined paired-end reads exhibited lengths ranging from 251 to 492 nt (Fig. S2c). We also tested pairing the reads with Pear (Zhang et al. 2014) and PANDAseq (Masella et al. 2012). Chimeric reads were removed using USEARCH in reference database mode (the MaarjAM database, Öpik et al. 2010) with the default parameters. The MaarjAM database contains representative sequences covering the NS31/AML2 amplicon from published environmental Glomeromycotina sequence groups and morphologically described taxa, classified into virtual taxa (VT, see Öpik et al. 2009, 2014). The full MaarjAM database (status October 2016) contains 20,399 SSU rRNA gene sequences distributed in 352 VT.

454 Sequencing reads were subjected to the bioinformatic procedures described in Öpik et al. (2013) and Saks et al. (2014). In short, 454 sequencing reads were retained for analysis only if they carried the correct barcode (8 bp) and forward primer sequences (21 bp), and were ≥ 170 bp long (excluding the barcode and primer sequence). Barcode and primer sequences were trimmed from reads. Sequences longer than

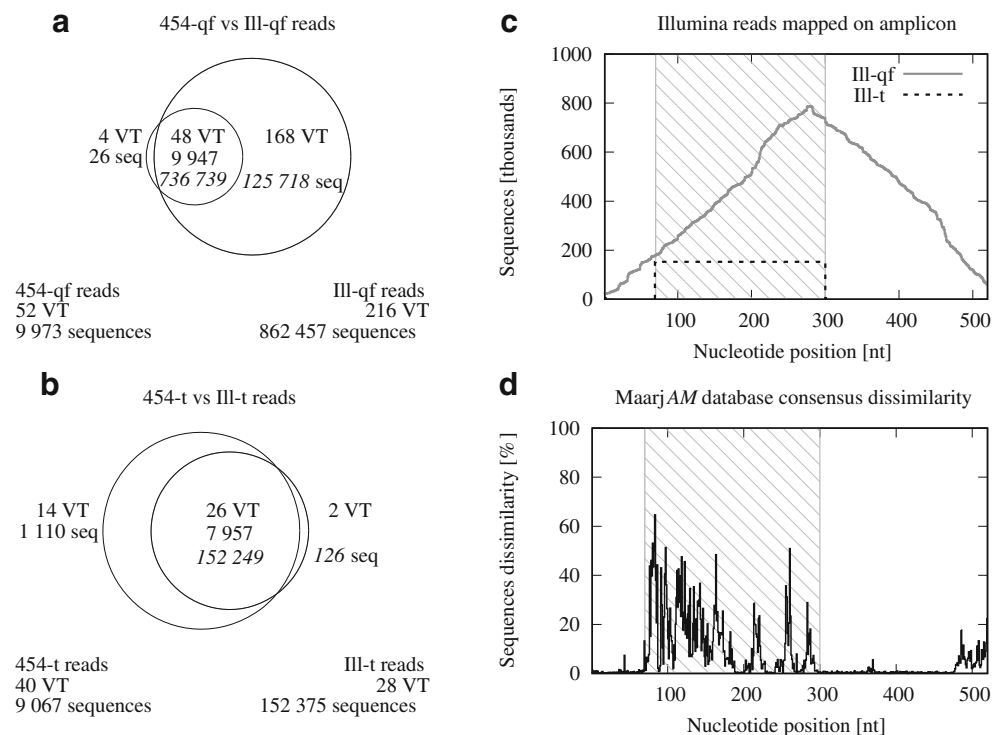
520 bases were trimmed to exclude reverse primer sequences. Chimeric reads were detected and removed from the data using USEARCH in reference database mode (MaarjAM database) with the default settings.

To validate BLAST search identification of SSU rRNA amplicon using the reference database, we conducted an *in silico* experiment. Two sets of sequences derived from the MaarjAM database (6405 sequences fitting the full amplicon, 352 VT) with different trimming lengths and starting positions were produced to test the effect of sequencing the variable versus the invariable part of the amplicon for 454-based and Illumina-based approaches (Fig. S6). Sequences simulating tagmentation-based Illumina reads (251 bp long) with a starting position 250 bases along the amplicon (after NS31 primer) showed a significant drop in correct identification, only identifying 60% of fixed length sequences correctly. For 454 reads the same experiment showed that using the 170 bases following the NS31 primer location is sufficient to correctly identify AM fungal sequences, confirming findings made by Davison et al. (2012).

To identify paired Illumina reads and 454 reads spanning the most variable fraction of the NS31-AML2 amplicon, we first calculated nucleotide dissimilarities along the NS31-AML2 amplicon, using sequences from the MaarjAM database. We identified database sequences spanning the entire amplicon by subjecting all sequences to BLAST searches against the NS31 and AML2 primer sequences. Only sequences without ambiguous nucleotides and receiving a perfect match against both primers were retained totalling 3693 database sequences (all 352 VT represented) that included the full-length amplicon (520 bases). The BLAST output was analysed and sequences trimmed to exclude primers and sections falling outside the NS31-AML2 amplicon region. A multiple alignment using MAFFT with default settings was used to identify misaligned or poor-quality reads. This yielded an alignment containing 1187 high-quality reference sequences. Trimmed MaarjAM sequences were compared to a consensus sequence from the alignment to calculate mismatch for each base. Nucleotide dissimilarities varied considerably along the amplicon, and the most variable fragment was located between positions 70 and 300, with the section from nucleotides 70 to 170 exhibiting particularly high variation (Fig. 1d). Variable region filtering was applied to 454 and Illumina reads by conducting a BLAST search against the consensus sequence mentioned above to retrieve the read position on the amplicon, and each read was trimmed accordingly to include only the variable region (positions 70–300 on NS31-AML2 amplicon).

Metagenomic Illumina reads were quality-filtered to remove reads with $q < 30$ and orphan reads. We then combined the remaining paired-end reads using FLASH with the same parameters as described above. We further filtered paired reads based on GC% content between 20 and 42% following

Fig. 1 Summary of AM fungal virtual taxa (VT) identified in root samples by 454 sequencing or Illumina MiSeq paired-end sequencing. **a** Taxon distribution between quality-filtered 454 (454-qf) and Illumina single-end reads (Ill-qf, R1). **b** Taxon distribution between amplicon-filtered 454 (454-t) and Illumina paired-end reads (Ill-t). Illumina reads are presented in *italic*. **c** Location of quality-filtered Illumina single-end (Ill-qf, R1, *solid line*) and amplicon-filtered paired-end reads (Ill-t, *dashed line*) on the NS31-AML2 amplicon with most variable fragment shaded with hatched lines. **d** MaarjAM database multiple alignment consensus dissimilarity for each base with most variable region *shaded*



published AM fungal genome data (Tisserant et al. 2013, Lin et al. 2014) and assembled sequences into contigs with Newbler (v2.6, the 454 Life Science de novo assembler) using the default parameters (minimum identity 96%, minimum length 40 bases, and parameter *-large* to speed up assembly and reduce memory footprint). Following assembly, we predicted potential genes in the contigs and translated these into proteins using GeneMark-ES (v4.33, Ter-Hovhannisyanyan et al. 2008) with the prebuilt Hidden Markov Model for eukaryotes.

Sets of representative filtered paired-end Illumina and 454 sequence reads have been deposited in the EMBL nucleotide collection (accession numbers KY685085–KY685341 and KY685342–KY685519 for 454 and Illumina reads, respectively; each set consists of up to two of the longest reads representing taxa found in each of the two studied plant species). Raw metagenomic paired-end sequences are stored in the Sequence Read Archive (SRA) (accession number PRJNA377617).

Identification of reads We identified amplicon reads using BLAST searches (soft masking with the DUST filter) against the MaarjAM database (3693 sequences, 352 VT) with sequences fitting the full amplicon (520 bases) between NS31 and AML2 primers (excluding both primers). Sequence reads were assigned to VT with the following criteria required for a match: sequence similarity $\geq 97\%$; the alignment length not differing from the length of the shorter of the query (454 or Illumina read) and subject (reference database sequence) sequences by more than 5%; and a BLAST e-value $< 1e-50$

(Thiéry et al. 2016). The best hit on the basis of the BLAST score was recorded for each read. Those reads (no hits) that BLAST did not match against the MaarjAM database were investigated by conducting a further BLAST search against the NCBI (status October 2016, Benson et al. 2013) non-redundant nucleotide database.

Putative proteins translated from metagenomic Illumina genes were subjected to a BLAST search against the NCBI non-redundant protein database (status December 2016) using protein BLAST with 40% identity (Rost 1999) and a 50% alignment filter of the shorter read (query or database sequence) giving rough estimation of organisms found in samples. The genome of a *Gigaspora rosea* (Tang et al. 2016) and a *Gigaspora margarita* (Salvioli et al. 2016) were missing from the protein database as they existed only in the nucleotide database, so an additional protein BLAST with tblastn (BLAST+ v2.5.0, protein against nucleotide) was conducted to add the missing information. The results were visualised using Krona (v2.7, Ondov et al. 2011).

Statistics

It is common practice to exclude singleton taxa (i.e. taxa represented by a single DNA read) or rare taxa (i.e. taxa below a specified abundance) prior to conducting diversity analyses using DNA-based community data. This is because singleton or rare taxa may represent sequencing artefacts (Tedersoo et al. 2010) or may be of minor interest if a description of dominant community patterns is intended. We removed

singleton taxa from 454 results, where each singleton hit represented 0.01% of all the hits. Using the same proportion, we removed rare taxa from Illumina reads when they represented less than 0.01% of total hits. Sampling efficacy was assessed using individual-based rarefaction (function *rarefy* from R package *vegan*; Oksanen et al. 2016). Richness estimates were calculated for each sample in each sequence set (Ill-qf, Ill-t, 454-qf, 454-t), and differences between sets and host plant species were estimated using a mixed-effects model (function *lmer* from R package *lme4*, Bates et al. 2015). Richness was log-transformed prior to inclusion in the model. Sample identity was included in the model as a random effect. The Kenward-Roger approximation was used to estimate model degrees of freedom (Kenward and Roger 1997), and Tukey contrasts were used to compare group means. Richness estimates were also regressed against read number within each sequence set and against each other between sets. The relative abundances of different VT in the 454-t and Ill-t sets (samples pooled within each set) were compared using Pearson's correlation. Pearson's correlation was also used to test the relationships between sample weight and recorded VT richness and sequence depth.

Results

Amplicon-based Illumina MiSeq sequencing

We obtained a total of 8,461,400 raw reads from 12 samples, with 164,385 to 1,610,065 reads per sample (Table S1). Following adapter removal and quality filtering, we detected and removed 19,469 chimeras among single-end reads (R1, 251 bases), leaving 937,658 quality-filtered sequences (Ill-qf) distributed randomly along the amplicon (Fig. 1c). Among paired-end reads, after adapter removal, quality filtering and combining the reads (R1 + R2, 251–492 bp), we detected and removed 23,628 chimeric reads.

MaarjAM database SSU rRNA gene amplicon sequences containing full NS31 and AML2 primers were aligned together in order to compare dissimilarities. Nucleotide dissimilarities varied considerably along the amplicon, with the section from nucleotide positions 70 to 170 exhibiting particularly high variation (Fig. 1d). Paired reads were quality filtered and trimmed to the 70 to 300 bp portion of the amplicon (Fig. 1c), retaining 159,068 sequences (Ill-t) that fully span variable region of the targeted marker (Fig. 1d). The length distribution of Illumina reads is shown in Fig. S2.

BLAST against the MaarjAM database identified 862,457 Ill-qf reads (92.0%) as belonging to 216 VT (Fig. 1a) and 152,375 Ill-t reads (95.8%) as 28 VT (Fig. 1b, Table 1, S2). Mapping Ill-qf reads onto the NS31-AML2 amplicon showed that reads of the 168 VTs detected only among the Ill-qf reads, but not in the 454-qf reads, were positioned in the 200 to

520 bp part of the amplicon (Fig. S7). Alternative identification of these reads (141,634 reads, 168 VT) by clustering using *blastclust* with 97% identity resulted in 29 clusters, where the largest cluster contained 139,667 reads (96.6% of the reads). This indicates that BLAST-based identity assignments of similar score alignment matches within the invariable region (200–520 nt) were largely random. Illumina amplicon reads were also paired with Pear and PANDAseq using same filtering parameters as Ill-t reads, but achieved the same VT (28) count compared to FLASH with slight changes in sequences identified (FLASH 152,375 sequences, Pear 159,289 sequences, PANDAseq 156,910 sequences).

Among Ill-t reads, all sequences (30,726 nohits) that did not receive a match against the MaarjAM database were identified by BLAST against the NCBI non-redundant nucleotide collection with 90% identity threshold: of the 29,798 sequences, 53% were identified as Metazoa, 44% Fungi (34% of all reads as *Glomeraceae*) and 3% Viridiplantae (the host plant families: 2% *Rosaceae* and <1% *Asparagaceae*).

Amplicon-based 454 sequencing

A total of 13,611 reads were successfully demultiplexed using the 8 bp barcode with 100% identity, generating 372 to 1790 reads per sample (Fig. S8, Table S1). Further quality filtering by primer identity, average read quality and read length retained 12,354 sequences. Chimera checking identified and removed a further 525 sequences, leaving a total of 11,829 cleaned 454 reads (454-qf). Sequence length and quality distributions are shown in Fig. S9.

A total of 9973 reads (454-qf) were identified as 52 VT (Fig. 1a, Table 1), including 48 VT also detected in the Ill-qf set (Table S2). When the 454 reads were trimmed to the 70–300 bp fragment of the amplicon, 9067 sequences (454-t) in 40 VT were retained, including 26 VT in common with Ill-t set (Fig. 1b). The 2079 unidentified sequences from 454-t were further identified against INSDC with 90% identity threshold: of the 1670 sequences, 39% were identified as Metazoa, 30% as Viridiplantae (the host plant families: 17% of all reads as *Rosaceae* and <1% *Asparagaceae*) and 26% as Fungi (10% *Glomeraceae*).

Comparison of AM fungal diversity detected by Illumina MiSeq and 454 sequencing

Rarefaction analysis suggested that the number of AM fungal reads per sample for 454 sequencing and Illumina sequencing was generally sufficient to produce asymptotic estimates of VT richness per sample (Fig. S10). AM fungal richness per sample differed between sequence data types ($F_{3,33} = 331.15$, $P < 0.001$; Fig. 2), but not between host plant species ($F_{1,10} = 0.02$, $P = 0.90$). Specifically, the mean number of VT per sample was ca 5–10 times higher among Ill-qf reads

Table 1 Quantitative aspects of amplicon-based AM fungal SSU rRNA sequencing using 454 and Illumina MiSeq sequencing platforms

	454 reads 454-qf			454-t			Illumina amplicon reads Ill-qf			Ill-t		
	Quality filtered			Filtered 70–300 bp			Quality filtered R1			Paired reads filtered 70–300 bp		
	Reads	Hits	%	Reads	Hits	%	Reads	Hits	%	Reads	Hits	%
Plot 1												
CM1	514	450	<i>87.55%</i>	499	425	<i>85.17%</i>	48,177	45,965	<i>95.41%</i>	10,122	10,066	<i>99.45%</i>
CM2	781	709	<i>90.78%</i>	757	629	<i>83.09%</i>	24,906	21,861	<i>87.77%</i>	3484	3200	<i>91.85%</i>
CM3	1290	1018	<i>78.91%</i>	1207	935	<i>77.46%</i>	62,551	47,898	<i>76.57%</i>	9062	8263	<i>91.18%</i>
RS1	1380	1186	<i>85.94%</i>	1335	1099	<i>82.32%</i>	90,021	83,488	<i>92.74%</i>	17,903	16,821	<i>93.96%</i>
RS2	1255	981	<i>78.17%</i>	1234	953	<i>77.23%</i>	73,915	62,765	<i>84.92%</i>	13,498	12,646	<i>93.69%</i>
RS3	1200	992	<i>82.67%</i>	1119	872	<i>77.93%</i>	50,861	46,312	<i>91.06%</i>	9128	8959	<i>98.15%</i>
Plot 2												
CM4	1131	1093	<i>96.64%</i>	1120	1024	<i>91.43%</i>	48,899	44,039	<i>90.06%</i>	5882	5609	<i>95.36%</i>
CM5	1328	1070	<i>80.57%</i>	1250	894	<i>71.52%</i>	61,249	52,768	<i>86.15%</i>	7523	6889	<i>91.57%</i>
CM6	294	236	<i>80.27%</i>	277	210	<i>75.81%</i>	72,126	65,651	<i>91.02%</i>	11,402	11,204	<i>98.26%</i>
RS4	465	393	<i>84.52%</i>	434	326	<i>75.12%</i>	163,006	151,810	<i>93.13%</i>	26,842	26,157	<i>97.45%</i>
RS5	1199	1031	<i>85.99%</i>	1139	967	<i>84.90%</i>	146,471	135,712	<i>92.65%</i>	23,862	23,715	<i>99.38%</i>
RS6	992	814	<i>82.06%</i>	957	733	<i>76.59%</i>	114,945	104,188	<i>90.64%</i>	20,360	18,846	<i>92.56%</i>
Total	11,829	9973	<i>84.31%</i>	11,328	9067	<i>80.04%</i>	957,127	862,457	<i>90.11%</i>	159,068	152,375	<i>95.79%</i>

For each data set, the number of quality-filtered and position-filtered reads are shown, with the number of reads identified as AM fungi by BLAST hits in the MaarjAM database

Proportions of reads getting a hit (expressed as a percentage) are shown in italic type

CM host plant *Convallaria majalis*, RS host plant *Rubus saxatilis*, R1 Illumina raw forward reads, 454-qf quality-filtered 454 reads, 454-t trimmed 454 reads, Ill-qf quality-filtered Illumina reads, 454-t trimmed Illumina reads

than among Ill-t reads (Fig. 2), reflecting inflated richness estimation in the former set. Ill-t had higher sequencing depth but lower observed richness compared with 454 sequencing (Fig. 2), and neither approach exhibited a positive relationship between sequencing depth and richness (Fig. S11). There was no correlation between sample richness estimated using the 454-t and Ill-t approaches (Pearson's $r = 0.19$, $P = 0.55$). However, the relative abundance of individual VT in 454-t

and Ill-t sets was highly correlated (Pearson's $r = 0.92$, $P < 0.001$; Fig. 3) and the most abundant VT was VT74 (*Glomus*) according to both sequencing methods (Table S2). The ten most abundant shared VT for both sequencing methods represented 74.8 and 88% of reads for 454-t and Ill-t, respectively. There was no correlation between dry root weight and sequence depth or recorded AM fungal richness (Fig. S12).

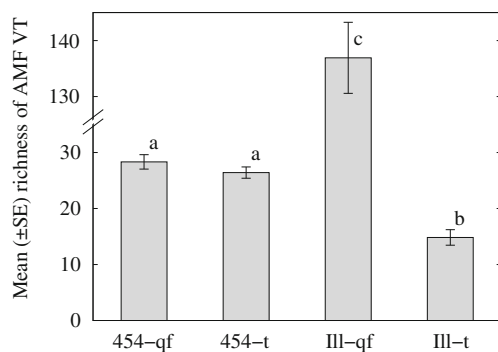


Fig. 2 Average number of AM fungal virtual taxa (VT) per sample when amplified by 454 sequencing or Illumina MiSeq paired-end sequencing after quality filtering (454-qf, Ill-qf) or quality filtering and trimming to the most variable fragment (70–300 bp) of the NS31-AML2 amplicon (454-t, Ill-t). Bars topped by the same letter do not differ significantly at $P \leq 0.05$ by Tukey's HSD test

Metagenomic Illumina MiSeq sequencing

We obtained a total of 63,371,233 raw metagenomic reads in both directions from 12 root samples, with 376,428 to 13,057,950 reads per sample. After adapter removal and quality filtering, combining the forward and reverse reads, a total of 21,640,048 paired reads were retained (34.2% of raw reads), with 32,924 to 4,548,256 paired reads per sample (Fig. S4). Identification against the MaarjAM database resulted in only 309 hits, on average 26 sequences per sample, to 16 VT, including 5 singleton VT.

We further filtered paired reads for GC content between 20 and 42%, retaining 8,598,087 reads, assembled into 46,653 contigs among 1,334,793 sequences. Contig length ranged from 7682 to 30,345 bp. Contigs longer than 500 bp were subjected to gene prediction with GeneMark-ES, resulting in 41,732 potential genes and 2728 protein sequences (Fig. S4).

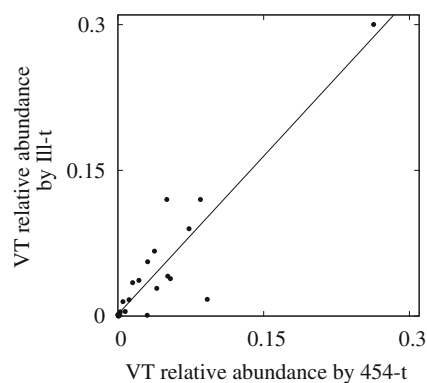


Fig. 3 Pearson's correlation between proportional abundances of shared VT between trimmed 454 reads (454-t) and Illumina MiSeq paired-end sequences (Ill-t) (Pearson's correlation $r = 0.92$, $P < 0.001$; omitting the outlier (VT74): $r = 0.73$, $P < 0.001$)

Of the latter, 88% were of plant origin (Fig. 4a), mostly *Rosaceae* (28% of all reads) and *Asparagaceae* (12%), and 11% were fungal (Fig. 4b), of which most had their closest hit to *Rhizophagus irregularis* (7%).

Discussion

In this study, AM fungal communities were described from the same set of plant root samples using Illumina MiSeq sequencing of tagged amplicons (SSU rRNA gene, NS31-AML2 amplicon), 454 sequencing of the same amplicons and metagenomic (i.e. PCR-free) shotgun sequencing of DNA templates on the Illumina MiSeq platform. We obtained relatively similar estimates of AM fungal diversity (richness) with Illumina MiSeq amplicon sequencing and 454 sequencing of the same samples, once reads were carefully quality filtered. This is notable given that Illumina MiSeq sequencing provides two orders of magnitude greater sequencing depth than 454 sequencing. These results show that the increase in sequencing depth per sample with the Illumina approach did not further increase AM fungal diversity (richness) estimates per sample or total diversity estimates. This is in strong contrast with the major improvement in the capture of AM fungal diversity that accompanied replacement of cloning-Sanger sequencing with 454 sequencing (Öpik et al. 2009). Lastly, metagenomic Illumina sequencing yielded very low numbers of AM fungal SSU reads, and hence very low estimates of AM fungal diversity in the samples, which were dominated by host-plant DNA. This suggests that amendments to the metagenomic sequencing approach are required for PCR-free description of AM fungal assemblages to become feasible.

Sequencing depth vs observed richness

An appropriate number of observations (here, sequencing depth per sample) is required to adequately describe the diversity of organisms in a sample (Hart et al. 2015; Bálint et al. 2016). This number depends on the complexity of the community. In the case of AM fungal assemblages, earlier results have indicated that a few hundreds to thousands of AM fungal sequences are sufficient to capture most diversity within a root or soil sample (Öpik et al. 2009; Hart et al. 2015). Here, the increase of sequencing depth from several hundreds of reads to over 30,000 reads per sample did not result in a further increase in the number of detected AM fungal species. Thus, approximately 300 reads would appear to provide sufficient sequencing depth to adequately describe AM fungal communities containing ca 5 to 30 species or VT. However, this was the case only after careful quality filtering of Illumina MiSeq reads.

When using tagmentation-based Illumina sequencing and BLAST-based taxonomy assignment, appropriate quality filtering of data is crucial in order to avoid inflated richness estimates. We obtained, on average, a five times higher estimate of AM fungal species richness per sample when the reads were not trimmed to the variable part of the SSU rRNA gene amplicon. This was due to those tagged reads that were located in the less variable part of the amplicon getting very similar BLAST matches to a broad range of reference sequences. It is notable that after trimming the reads to the most variable portion of the NS31-AML2 amplicon, Illumina and 454 read sets showed similar relative abundances of individual VT. We did not find a correlation between AM fungal richness estimates produced by the two groups, perhaps reflecting the fairly low variability within each group.

Sequencing errors and appropriate read length for correct AM fungal identification with the SSU rRNA gene marker

Available read lengths from next-generation sequencing (NGS) platforms have increased with advances in sequencing technology, but are generally (including in the case of the Illumina MiSeq platform) still too short to cover the full amplicon most commonly used in AM fungal ecology, a ca 520 bp amplicon of the SSU rRNA gene (Öpik and Davison 2016). This raises the question of the sufficient sequence length for correct identification of the target amplicon. An *in silico* test using the sequences in the MaarjAM database confirmed the importance of identifying reads against the variable section of the NS31-AML2 amplicon located at positions 70–300 nt. Identification of relatively short reads against sufficiently variable reference reads is particularly crucial when BLAST-based closed reference OTU picking is used (*sensu* Bik et al. 2012). BLAST hits located in the invariant part of reference sequences would result in imprecise taxonomy

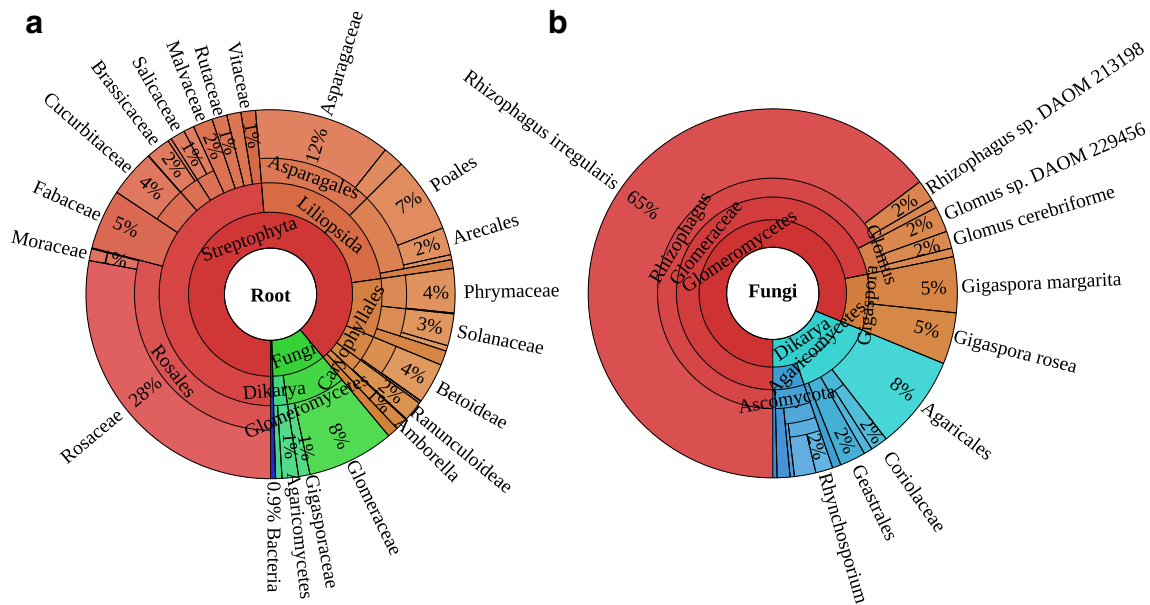


Fig. 4 Metagenomic Illumina sequences translated into genes and protein BLAST by graph drawing tool Krona. Taxonomic ranks are shown as nested sectors with higher ranks at the centre. **a** All identified data. **b** Fungi

assignments and inflated taxon richness estimates. The minimal read length to obtain correct taxonomy assignment with the same amplicon has been tested earlier with 170 to 400 bp long (from the NS31 end) 454 reads, i.e. located in the same variable portion of the amplicon, and indicated that reads ≥ 170 bp are sufficient for robust AM fungal identification with this amplicon (Davison et al. 2012). The findings of this study further confirm earlier conclusion.

Another source of misidentification of reads is sequencing errors, which accumulate towards the end of reads, as the quality of the called bases decreases. This phenomenon has been reported for both Illumina and 454 reads (Perkins et al. 2013; Niklas et al. 2013). 454 Reads are prone to errors in long homopolymeric regions, where the probability of indel errors increases with sequence length (Niklas et al. 2013). In contrast, substitution errors are the most frequent type of errors for Illumina platform, constituting 79 to 88% of all errors (Nguyen et al. 2011). Both sequencing technologies have a mean error rate of 1% among unfiltered reads (Niklas et al. 2013; Gilles et al. 2011). Errors in 454 reads occur more often after bases 200–300 (Niklas et al. 2013), where substitutions and ambiguous base calls accumulate because of a loss of synchronism during the sequencing process on the multi-templated beads (Gilles et al. 2011). As the most variable part of the amplicon used in this study is located in the first 300 bases, the invariant part of the amplicon (300–520 nt of the NS31-AML2 amplicon) is more error prone, and in our case, this was removed. To overcome sequencing errors with Illumina reads, Nguyen et al. (2011) reported a 5-fold decrease of the error rate (reducing the likelihood of false base call to

$<0.1\%$) by eliminating low-quality ($q < 30$) sequences, resulting in loss of 24 to 35% sequences.

Combining Illumina paired-end reads can act as a quality filter by resulting in the omission of poor quality reads and improvement of low quality regions, thus vastly improving diversity estimates (Bokulich et al. 2013). With paired-end Illumina sequencing, the forward read usually has slightly better quality than the reverse read (Minoche et al. 2011). The reads are designed to overlap at the end of both sequences where the quality is lowest. This helps to improve the quality of poor regions (the best-supported nucleotides are retained in the combined reads), but raises the question of how well the algorithms can combine overlapping regions of such quality (Parikh et al. 2016). Tagmentation introduced sequences containing Nextera adapters that were removed from the analyses. However, these removed reads can be used to find further taxon richness in samples by trimming adapters from the sequences. After trimming of the adapter sequences (Bolger et al. 2014), the remaining sequences are shorter than the original single-end read, 251 bp. Shorter fragments mean that the target region in the reference database has to be carefully selected to ensure correct species identification. Allowing sequences that are too short may introduce misidentification of species, but on the other hand, use of too strict parameters may lead to rare species being overlooked, thus affecting species richness estimates. Nextera tagmentation-based Illumina reads increase the overall complexity of the analysis, as further study is needed to understand the reference dataset and underlying variable region properties when using BLAST-based closed reference OTU picking.

PCR-based vs metagenomic view on AM fungal communities

Caution should be exercised when drawing biological meaning from PCR-amplified data sets (Medinger et al. 2010), as PCR bias (Bradley et al. 2016) and technical sequencing errors can strongly inflate diversity estimates (Reeder and Knight 2009) and shift the abundances of individual organisms (Schmidt et al. 2013). Furthermore, the choice of the PCR polymerase can have a strong impact on detected richness (Oliver et al. 2015). Sequencing errors can be partly detected and removed by high quality filtering and by using chimera detection tools (Edgar et al. 2011). Nonetheless, to overcome PCR and primer bias, metagenomic sequencing that yields data largely proportional to the template DNA would be highly desirable (Thomas et al. 2012).

The metagenomes of plant roots present challenges of size and complexity. In this study, our workflow detected only a small fraction of AM fungal sequences in plant root samples, resulting in a considerably lower detected richness than with PCR-based approaches. Metagenomic Illumina has also been reported to have uneven read coverage (Dohm et al. 2008, Hillier et al. 2008); this bias can result in gaps in the metagenome without any reads and affect quantitative assessments. Discarding low-quality reads can also result in uneven read coverage, introducing potential bias in quantitative studies (Nakamura et al. 2011). The great majority of the metagenomic sequences in our dataset were of host-plant origin, with AM fungi apparently present at low relative abundance in terms of DNA amount in the plant root samples. It may be worth testing available approaches for enrichment of the template DNA or targeted removal of unwanted DNA, e.g. of the plants (Motley et al. 2014). Additionally, multiple displacement amplification (MDA) can be used by annealing random hexamer primers to the template, where DNA synthesis is carried out by a high fidelity enzyme at a constant temperature (Picher et al. 2016). Non-target sequences can also be filtered out bioinformatically, e.g. if the GC% content of the target group is distinct or by aligning reads against a targeted reference genome.

A further limitation of the metagenomic approach is that reference databases are insufficiently populated with genomic data (Santamaria et al. 2012; del Campo et al. 2014). In our data, AM fungal reads were mostly identified as belonging to an AM fungal species, *R. irregularis* for which genome sequences (Tisserant et al. 2013; Lin et al. 2014) are available in INSDC. Although PCR-based data show that many more taxa are present in these samples, metagenomic approaches for detecting them will be limited as long as further AM fungal genomes are not available in public sequence repositories. Thus, PCR-based approaches for AM fungal detection remain the viable choice.

Conclusion

Amplicon-based sequencing using Illumina and 454 platforms recorded similar AM fungal taxon richness in samples and highly correlated relative abundances of individual VT. Thus, the two approaches yield comparable AM fungal diversity estimates. Although Illumina sequencing using tagmentation is not common in community surveys, it could yield valuable data when sequencing and genotyping longer genomic fragments, such as the entire SSU rRNA gene, or nuclear ribosomal operon, which are too long for most current sequencing platforms.

Metagenomic reads from plant root samples included very few AM fungal sequences. This is because the proportion of AM fungal DNA in root samples is very low against the overwhelming background of plant (and other organism) DNA, but also due to difficulties in assigning sequences to reference taxa, as representation of genomic AM fungal sequences in reference databases is very limited. For these reasons, metagenomic sequencing seems unlikely to replace amplicon-based analyses of AM fungal diversity in the near future. Nonetheless, DNA enrichment or advances in sequencing approaches (Taylor et al. 2017) may provide opportunities to better capture AM fungal DNA using PCR-free approaches.

Acknowledgements This research was funded by grants from the Estonian Research Council (grant IUT 20-28), the European Regional Development Fund (Centre of Excellence EcolChange) and ERA-NET Cofund BiodivERSA project SoilMan. Preparatory procedures for 454 sequencing were performed by BiotaP Ltd. (Tallinn, Estonia). RA and MR were supported by the EU ERDF through the Estonian Centre of Excellence in Genomics and Translational Medicine (project no. 2014-2020.4.01.15-0012) and by the Estonian Ministry of Education and Research (institutional grant IUT34-11). We are grateful to the Estonian Genome Centre for technical support.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Bálint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, O'Hara RB, Öpik M, Sogin ML, Unterseher M, Tedersoo L (2016) Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiol Rev* 40: 686–700
- Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. *Nucleic Acids Res* 41:D36–D42

- Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol Evol* 27:233–243
- Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 10:57–59
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120
- Bradley IM, Pinto AJ, Guest JS (2016) Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl Environ Microbiol* 82:5878–5891
- Brown SP, Callahan MA, Oliver AK, Jumpponen A (2013) Deep Ion Torrent sequencing identifies soil fungal community shifts after frequent prescribed fires in a southeastern US forest ecosystem. *FEMS Microbiol Ecol* 86:557–566
- Burke CM, Darling AE (2016) A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *Peer J* 4:e2492
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinf* 10:421
- Cline LC, Zak DR, Upchurch RA, Freedman ZB, Peschel AR (2017) Soil microbial communities and elk foraging intensity: implications for soil biogeochemical cycling in the sagebrush steppe. *Ecol Lett* 20:202–211
- Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38:1767–1771
- Cui XC, Hu JL, Wang JH, Yang JS, Lin XG (2016) Reclamation negatively influences arbuscular mycorrhizal fungal community structure and diversity in coastal saline-alkaline land in Eastern China as revealed by Illumina sequencing. *Appl Soil Ecol* 98:140–149
- Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, Johnson NC, Kane A, Koorem K, Kochar M, Ndiaye C, Pärtel M, Reier Ü, Saks Ü, Singh R, Vasar M, Zobel M (2015) Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* 349:970–973
- 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
- del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol Evol* 29:252–259
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200
- García K, Doidy J, Zimmermann SD, Wipf D, Courty PE (2016) Take a trip through the plant and fungal transportome of mycorrhiza. *Trends Plant Sci* 21:937–950
- Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12:245
- Glassman SI, Peay KG, Talbot JM, Smith DP, Chung JA, Taylor JW, Vilgalys R, Bruns TD (2015) A continental view of pine-associated ectomycorrhizal fungal spore banks: a quiescent functional guild with a strong biogeographic pattern. *New Phytol* 205:1619–1631
- Hart MM, Aleklett K, Chagnon PL, Egan C, Ghignone S, Helgason T, Lekberg Y, Öpik M, Pickles BJ, Waller L (2015) Navigating the labyrinth: a guide to sequence-based, community ecology of arbuscular mycorrhizal fungi. *New Phytol* 207:235–247
- Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 5:183–188
- Johansen RB, Johnston P, Mieczkowski P, Perry GLW, Robeson MS, Burns BR, Vilgalys R (2016) A native and an invasive dune grass share similar, patchily distributed, root-associated fungal communities. *Fungal Ecol* 23:141–155
- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780
- Kenward MG, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53:983–997
- Lee HK, Lee CK, Tang JW, Loh TP, Koay ES (2016) Contamination-controlled high-throughput whole genome sequencing for influenza A viruses using the MiSeq sequencer. *Sci Rep* 6:33318
- Lee J, Lee S, Young JPW (2008) Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. *FEMS Microbiol Ecol* 65:339–349
- Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, Pang E, Cao H, Cha H, Lin T, Zhou Q, Shang Y, Li Y, Sharma T, van Velzen R, de Ruijter N, Aanen DK, Win J, Kamoun S, Bisseling T, Geurts R, Huang S (2014) Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genet* 10:e1004078
- Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjoller R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J, Kauterud H (2013) Fungal community analysis by high-throughput sequencing of amplified markers—a user’s guide. *New Phytol* 199:288–299
- Liu M, Zheng R, Bai SL, Bai Y, Wang JG (2017) Slope aspect influences arbuscular mycorrhizal fungus communities in arid ecosystems of the Daqing Mountains, Inner Mongolia, North China. *Mycorrhiza* 27:189–200
- Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27:2957–2963
- Masella AP, Bartram AK, Trzaskowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinf* 13:31
- Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenigk J (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol* 19:32–40
- Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 12:R112
- Motley ST, Picuri JM, Crowder CD, Minich JJ, Hofstadler SA, Eshoo MW (2014) Improved multiple displacement amplification (iMDA) and ultra-clean reagents. *BMC Genomics* 15:443
- Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 39:e90
- Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL (2011) Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* 12:106
- Niklas N, Pröll J, Danzer M, Stabenheiner S, Hofer K, Gabriel C (2013) Routine performance and errors of 454 HLA exon sequencing in diagnostics. *BMC Bioinf* 14:176
- Oksanen J, Blanchet GF, Kindt R, Legendre P, Minchin PR, O’Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2016) Vegan: community ecology package. R package version 2:3–5 <https://CRAN.R-project.org/package=vegan>
- Oliver AK, Brown SP, Callahan MA, Jumpponen A (2015) Polymerase matters: non-proofreading enzymes inflate fungal community richness estimates by up to 15%. *Fungal Ecol* 15:86–89

- Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. *BMC Bioinform* 12:385
- Öpik M, Davison J, Moora M, Zobel M (2014) DNA-based detection and identification of Glomeromycota: the virtual taxonomy of environmental sequences. *Botany* 92:135–147
- Öpik M, Davison J (2016) Uniting species- and community-oriented approaches to understand arbuscular mycorrhizal fungal diversity. *Fungal Ecol* 24B:106–113
- Öpik M, Metsis M, Weedon JT, 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
- Öpik M, Moora M, Liira J, Zobel M (2006) Composition of root-colonizing arbuscular mycorrhizal fungal communities in different ecosystems around the globe. *J Ecol* 94:778–790
- Öpik M, Moora M, Zobel M, Saks Ü, Wheatley R, Wright F, Daniell T (2008) High diversity of arbuscular mycorrhizal fungi in a boreal herb-rich coniferous forest. *New Phytol* 179:867–876
- Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytol* 188:223–241
- Öpik M, Zobel M, Cantero JJ, Davison J, Facelli JM, Hiiesalu I, Jairus T, Kalwij JM, Koorem K, Leal ME, Liira J, Metsis M, Neshataeva V, Paal J, Phosri C, Pölmel S, Reier Ü, Saks Ü, Schimann H, Thiéry O, Vasar M, Moora M (2013) Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. *Mycorrhiza* 23:411–430
- Orchard S, Hilton S, Bending GD, Dickie IA, Standish RJ, Gleeson DB, Jeffery RP, Powell JR, WALKER C, Bass D, Monk J, Simonin A, Ryan MH (2017) Fine endophytes (*Glomus* Tenue) are related to *Mucoromycotina*, not *Glomeromycota*. *New Phytol* 213:481–486
- Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire AZ (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 35:e130
- Parikh HI, Koparde VN, Bradley SP, Buck GA, Sheth NU (2016) MeFiT: merging and filtering tool for illumina paired-end reads for 16S rRNA amplicon sequencing. *BMC Bioinform* 17:491
- Perkins TT, Tay CY, Thirriot F, Marshall B (2013) Choosing a benchtop sequencing machine to characterise *Helicobacter pylori* genomes. *PLoS One* 8:e67539
- Picher AJ, Budeus B, Wafzig O, Krüger C, García-Gómez S, Martínez-Jiménez MI, Díaz-Talavera A, Weber D, Blanco L, Schneider A (2016) TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. *Nat Commun* 7:13296
- Pozo MJ, Lopez-Raez JA, Azcon-Aguilar C, Garcia-Garrido JM (2015) Phytohormones as integrators of environmental signals in the regulation of mycorrhizal symbioses. *New Phytol* 205:1431–1436
- Randle-Boggis RJ, Helgason T, Sapp M, Ashton PD (2016) Evaluating techniques for metagenome annotation using simulated sequence data. *FEMS Microbiology Ecology* 92:fw095
- Reeder J, Knight R (2009) The 'rare biosphere': a reality check. *Nat Methods* 6:636–637
- Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng Des Sel* 12:85–94
- 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 understory. *Botany* 92:277–285
- Salvioli A, Ghignone S, Novero M, Navazio L, Bagnaresi P, Bonfante P (2016) Symbiosis with an endobacterium increases the fitness of a mycorrhizal fungus, raising its bioenergetic potential. *ISME J* 10:130–144
- Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemán D, Valiente G, Pesole G (2012) Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform* 13:682–695
- Schlaeppli K, Bender SF, Mascher F, Russo G, Patrignani A, Camenzind T, Hemple S, Rilling MC, van der Heijden MGA (2016) High-resolution community profiling of arbuscular mycorrhizal fungi. *New Phytol* 212:780–791
- Schmidt PA, Bálint M, Greshake B, Bandow C, Römbke J, Schmitt I (2013) Illumina metabarcoding of a soil fungal community. *Soil Biol Biochem* 65:128–132
- Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 21:1794–1805
- Simon L, Lalonde M, Bruns TD (1992) Specific amplification of 18S fungal ribosomal genes from vesicular-arbuscular endomycorrhizal fungi colonizing roots. *Appl Environ Microbiol* 58:291–295
- Smith SE, Read DJ (2008) *Mycorrhizal symbiosis*. Academic Press, Amsterdam
- Spatafora JW, Chang Y, Benny GL, Lazarus KL, Smith ME, Berbee ML, Bonito G, Corradi N, Grigoriev I, Gryganskyi A, James TY, O'Donnell K, Roberson RW, Taylor TN, Uehling J, Vilgalys R, White MM, Stajich JE (2016) A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia* 108:1028–1046
- Tang N, San Clemente H, Roy S, Bécard G, Zhao B, Roux C (2016) A survey of the gene repertoire of *Gigaspora rosea* unravels conserved features among Glomeromycota for obligate biotrophy. *Front Microbiol* 7:233
- Taylor JD, Helgason T, Öpik M (2017) Molecular community ecology of arbuscular mycorrhizal fungi. In: Dighton J, White JF, eds. *The fungal community: its organization and role in the ecosystem*, 4th edn. CRC Press, 00
- Tedersoo L, Anslan S, Bahram M, Pölmel S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson H, Hildebrand F, Bork P, Abarenkov K (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding of fungi. *MycKeys* 10:1–43
- Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, 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 Phytol* 188:291–301
- Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 18:1979–1990
- Thiéry O, Vasar M, Jairus T, Davison J, Roux C, Kivistik PA, Metspalu A, Milani L, Saks Ü, Moora M, Zobel M (2016) Sequence variation in nuclear ribosomal small subunit, internal transcribed spacer and large subunit regions of *Rhizophagus irregularis* and *Gigaspora margarita* is high and isolate-dependent. *Mol Ecol* 25:2816–2832
- Thomas T, Gilbert J, Meyer F (2012) *Metagenomics—a guide from sampling to data analysis*. *Microb Inf Exp* 2:3
- Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charon P, Duensing N, Frei dit Frey N, Gianinazzi-Pearson V, Gilbert LB, Handa Y, Herr JR, Hijri M, Koul R, Kawaguchi M, Krajinski F, Lammers PJ, Masclaux FG, Murat C, Morin E, Ndikumana S, Pagni M, Petitpierre D, Requena N, Rosikiewicz P, Riley R, Saito K, San Clemente H, Shapiro H, van Tuinen D, Becard G, Bonfante P, Paszkowski U, Shachar-Hill Y, Tuskan GA, JPW Y, Sanders IR, Henrissat B, Rensing SA, Grigoriev IV, Corradi N, Roux C, Martin F (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci U S A* 110:20117–20122
- Treseder KK (2016) Model behavior of arbuscular mycorrhizal fungi: predicting soil carbon dynamics under climate change. *Botany* 94:417–423

- van der Heijden MGA, de Bruin S, Luckerhoff L, van Logtestijn RSP, Schlaeppli K (2015) A widespread plant-fungal-bacterial symbiosis promotes plant biodiversity, plant nutrition and seedling recruitment. *ISME J* 10:389–399
- Wang C, White PJ, Li CJ (2016) Colonization and community structure of arbuscular mycorrhizal fungi in maize roots at different depths in the soil profile respond differently to phosphorus inputs on a long-term experimental site. *Mycorrhiza* 27:369–381
- Xu TL, Veresoglou SD, Chen YL, Rillig MC, Xiang D, Ondrej D, Hao ZP, Liu L, Deng Y, Hu YJ, Chen WP, Wang JT, He JZ, Chen BD (2016) Plant community, geographic distance and abiotic factors play different roles in predicting AMF biogeography at the regional scale in northern China. *Environ Microbiol Rep* 8:1048–1057
- Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina paired-end reAd mergeR. *Bioinformatics* 30:614–620