New Phytologist

Research

Methods

Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities

Carles Castaño[1], Anna Berlin[1], Mikael Brandström Durling[1], Katharina Ihrmark[1], Björn D. Lindahl[2], Jan Stenlid[1], Karina E. Clemmensen[1]* and Åke Olson[1]*

[1]Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala SE-75007, Sweden; [2]Department of Soil and Environment, Swedish University of Agricultural Sciences, Uppsala SE-75007, Sweden

Author for correspondence:
Carles Castaño
Tel: +46 18 672376
Email: carles.castanyo@slu.se

Received: *14 February 2020*
Accepted: *29 May 2020*

New Phytologist (2020) **228:** 1149–1158
doi: 10.1111/nph.16731

Key words: high-throughput sequencing, Illumina MiSeq, ITS, metabarcoding, metabarcoding biases, mock community, PacBio.

Summary

- Recent studies have questioned the use of high-throughput sequencing of the nuclear ribosomal internal transcribed spacer (ITS) region to derive a semi-quantitative representation of fungal community composition. However, comprehensive studies that quantify biases occurring during PCR and sequencing of ITS amplicons are still lacking.
- We used artificially assembled communities consisting of 10 ITS-like fragments of varying lengths and guanine-cytosine (GC) contents to evaluate and quantify biases during PCR and sequencing with Illumina MiSeq, PacBio RS II and PacBio Sequel I technologies.
- Fragment length variation was the main source of bias in observed community composition relative to the template, with longer fragments generally being under-represented for all sequencing platforms. This bias was three times higher for Illumina MiSeq than for PacBio RS II and Sequel I. All 10 fragments in the artificial community were recovered when sequenced with PacBio technologies, whereas the three longest fragments (> 447 bases) were lost when sequenced with Illumina MiSeq. Fragment length bias also increased linearly with increasing number of PCR cycles but could be mitigated by optimization of the PCR setup. No significant biases related to GC content were observed.
- Despite lower sequencing output, PacBio sequencing was better able to reflect the community composition of the template than Illumina MiSeq sequencing.

Introduction

The use of high-throughput DNA sequencing technologies in the field of fungal ecology has increased our understanding of how fungi affect processes in soils and ecosystems (Lindahl *et al.*, 2013; Nilsson *et al.*, 2019). One of the main strengths of these technologies is that communities can be taxonomically profiled at relatively low costs and effort. However, overcoming biases and errors introduced during amplification and sequencing remains a major challenge (Nilsson *et al.*, 2019). Thus, the interpretation of internal transcribed spacer (ITS) sequencing data as semi-quantitative (i.e. correctly representing relative abundances of members) in fungal community studies has been questioned based on observations of biases and errors when sequencing artificially assembled ('mock') communities (Amend *et al.*, 2010; Bakker, 2018; Palmer *et al.*, 2018). Such biases and errors affect perceived differences in composition and diversity among communities, and should be minimized (Frøslev *et al.*, 2017; Nilsson *et al.*, 2019). Therefore, it is necessary to establish laboratory protocols and data handling strategies that are optimized to preserve qualitative and quantitative representation of community members throughout DNA amplification and sequencing. In addition, any remaining effects of various biases should be assessed, quantified and compensated for in downstream analyses.

Marker amplification by polymerase chain reaction (PCR) is recognized as one of the most important sources of community distortion in metabarcoding studies (Tedersoo *et al.*, 2018). The PCR may be biased towards preferential amplification of specific DNA templates with lower cytosine-guanine (GC) content or of shorter lengths (Benjamini & Speed, 2012; Ihrmark *et al.*, 2012). Similarly, preferential amplification or lack of amplification of specific taxa is influenced by primer as well as marker choice (Tedersoo & Lindahl, 2016; Nilsson *et al.*, 2019). When degenerated primers are used, excessive PCR cycles may bias amplification towards templates matching nondepleted primers (Lindahl *et al.*, 2013). In addition, the PCR amplification process, typically multiplying the template by a factor of 10^6–10^{10} is, *per se*, a

*Shared last authorship.

source of error and chimera formation (Edgar *et al.*, 2011), which may be reduced by the use of a proof-reading polymerase (Tedersoo *et al.*, 2018) and optimization of the PCR setup (Lindahl *et al.*, 2013; Clemmensen *et al.*, 2016). Another source of bias affecting PCR is the presence of inhibiting compounds in the DNA extracts of environmental samples, which can decrease PCR sensitivity (Schrader *et al.*, 2012).

Bias may also occur during sequencing (Quail *et al.*, 2012; D'Amore *et al.*, 2016; Kennedy *et al.*, 2018). For example, sequencing may introduce errors (i.e. base mismatches, deletions/insertions) and tag jumps that generate sequences with different sample tags than the ones originally applied (Carlsen *et al.*, 2012). In addition, several sequencing technologies favour shorter amplicons with balanced adenine-thymine : cytosine-guanine (AT : GC) ratios (Tedersoo *et al.*, 2018; Nilsson *et al.*, 2019). Other errors occurring during sequencing include the creation of 'daughter' operational taxonomic units (OTUs), which consistently co-occur with more abundant 'parent' OTUs (Frøslev *et al.*, 2017). Hence, the magnitude and importance of sequencing errors and biases depend on both the PCR setup and the sequencing platform (Quail *et al.*, 2012; Wagner *et al.*, 2016).

Illumina MiSeq is currently the most widely used sequencing platform for metabarcoding of fungi (Nilsson *et al.*, 2019) due to the high sequencing output and the relatively low cost per read. Recently, Tedersoo *et al.* (2018) and Tedersoo & Anslan (2019) used Pacific Biosciences (PacBio) to sequence longer fragments (1200–2500 bases), which can provide better taxonomic resolution (Singer *et al.*, 2016), but the PacBio platform has also been used to sequence shorter ITS (800–900 bases: Redondo *et al.*, 2018) and ITS2 markers (200–450 bases: Kyaschenko *et al.*, 2017a,b; Varenius *et al.*, 2017; Castaño *et al.*, 2018; Sterkenburg *et al.*, 2018). Despite higher error rates of single reads for PacBio (*c.* 11%) as compared with Illumina MiSeq (*c.* 0.1–2.6%) (May *et al.*, 2015; Pfeiffer *et al.*, 2018), the PacBio SMRT technology enables correction of random errors by calculating a consensus sequence based on typically > 30 separate sequencing passes, reducing errors to < 1% when short fragments are sequenced (Reuter *et al.*, 2015). However, one constraint of the PacBio technology is the lower sequencing output compared to MiSeq – the output of Illumina MiSeq is around 20–50 times higher than that of PacBio RS II (at the same cost). Low sequencing output might reduce diversity resolution and compromise ecological inference (Kennedy *et al.*, 2018). However, while high output is generally desirable, sequencing costs for various technologies decrease over time, allowing generally larger sequencing depths. Nevertheless, no comprehensive studies have yet quantified the biases introduced by different sequencing platforms in relation to biases introduced during PCR. When platforms were compared using 16S markers, D'Amore *et al.* (2016) found that communities were always distorted in relation to the template, but that data were still semi-quantitative to some extent. Based on a synthetic mock community, Palmer *et al.* (2018) showed that read abundances using Illumina MiSeq and Ion Torrent deviated strongly from expected values and concluded that distortion occurred mainly during PCR. Bakker (2018) also found that community composition based on MiSeq sequencing deviated strongly from the template, even after optimization of the PCR protocol and the bioinformatics pipeline.

Mock communities have been used to improve and optimise PCR protocols (Ihrmark *et al.*, 2012; Gohl *et al.*, 2016; Palmer *et al.*, 2018), to evaluate bioinformatics parameters (Bakker, 2018), and to detect potentially chimeric sequences (Aas *et al.*, 2016). If biases are consistent and quantified by sequencing of known standards (e.g. a mock community) together with the study samples, they may be accounted for during downstream bioinformatics or data analyses. Hence, the use of spike-in mock communities is a promising approach to measure and potentially account for biases and errors introduced during PCR and sequencing of fungal communities (Lindahl *et al.*, 2013; Gohl *et al.*, 2016).

Here we constructed an amplicon mock community, 'ITS2 mock', consisting of 10 artificial ITS2-like fragments differing in length (142–591 bases) and GC content (45–63%), which can be used as an internal control in metabarcoding studies where variation in length and GC content across fragments might distort final community composition. We used various mixtures of this ITS mock community to separate, quantify and characterize biases and errors introduced during PCR and sequencing. In a first experiment, we used the 'ITS2 mock' assembled post-PCR to compare biases introduced by three sequencing platforms (Illumina MiSeq, PacBio RS II and PacBio Sequel I). In a second experiment, we constructed complex communities pre-PCR, to assess the combined PCR and sequencing biases. In a third experiment, we used the least biased sequencing method and explored options to optimize PCR conditions towards minimal bias.

Materials and Methods

Construction of the mock community

The sequenced genome of *Heterobasidion irregulare* isolate TC32-1 (Olson *et al.*, 2012) was used to select 10 artificial fragments of different lengths and GC contents (Supporting Information Fig. S1a; Table 1). Primers were designed in the software Primer3 v.0.4.0 (http://primer3.ut.ee) and are available in Table S1. The fragments aimed to be close to 50, 100, 150, 200, 250, 300, 350, 400, 450, 500 bases (without primers), were unique and did not match any known ITS sequences. The fITS7 (Ihrmark *et al.*, 2012) and ITS4 (White *et al.*, 1990) primer target sequences plus *c.* 92 bases sequence from *Lophium mytilinum* strain CBS 114111 (GenBank, Benson *et al.*, 2017, accession no. EF596819) were added to the fragments using the MultiSite Gateway Three-Fragment Vector Construction Kit (Invitrogen), and the constructs were cloned in *Escherichia coli* using the One Shot TOP10 chemical transformation protocol (Invitrogen) according to the manufacturer's instructions.

To inspect the constructed fragments, including the intactness of the ITS primer sites, the cloned fragments were amplified using M13 primers with target sites in the plasmid vector outside of the ITS construct in a 2720 Thermal Cycler (Life Technologies). The 50 µl PCR reactions contained final concentrations of

Table 1 Setup of the experiments performed in this study.

Experiment	Biases assessed	PCR cycles	Template starting conditions		Other tests
			Total copies reaction^{-1}	Fragment mix	
Single fragments	Sequencing	25	750 000*	Single fragment	
Even community	Sequencing + PCR	25	7500 000; 750 000; 75 000, 7500; 750; 75	Even	Sequencing depth
Uneven community	Sequencing + PCR	25	167 475	Uneven	
PCR test	PCR	22,25,28,31,35	400 000; 4000 000	Even	Richness (Extra OTUs)

* 750 000 copy numbers in 'single fragments' experiment refers to post-PCR purified amplicon copies used for sequencing, whereas in all the other experiments, 'total copies' refers to the number of copies used pre-PCR. OTUs, operational taxonomic units.

200 μM of each nucleotide, 2.75 mM $MgCl_2$, 200 nM of each primer and 0.025 U μl^{-1} of polymerase (DreamTaq Green, Thermo Fisher Scientific, Waltham, MA) in the buffer supplied by the manufacturer. Cycling conditions were as follows: 5 min at 94°C, followed by 35 cycles of 30 s at 94°C, 30 s at 55°C, 30 s at 72°C, and a final extension of 7 min at 72°C. Amplicons were purified with the EZNA Cycle Pure kit (Omega Biotech, Norcross, GA, USA) and subjected to Sanger sequencing from both ends using the M13 primers at Macrogen Inc., South Korea. Including the ITS primer sequences, the final sizes of the artificial ITS2 fragments in the plasmid vectors were 180, 227, 281, 333, 382, 485, 518, 537, 586 and 629 bases. Excluding the primer sites, the fragments were 142, 189, 243, 295, 344, 395, 447, 499, 548, 591 bases long and had GC contents of 45–63% (Fig. S1a).

The 10 cloned plasmids were purified with the Nucleobond Xtra Maxi kit (Macherey-Nagel, Duren, Germany), diluted 1 : 100 and quantified with the high sensitivity kit on a Qubit Fluorometer (Life Technologies). Plasmids can be obtained from the authors upon request.

Experiment 1: assessment of sequencing biases

In this experiment, we tested community distortion during sequencing (Table 1) of an artificial post-PCR amplicon community. Each of the 10 fragments were amplified in three separate technical replicates using the fITS7-ITS4 primers (as detailed in the Construction of the mock community section, above), both of which were extended by a linker base (T), a unique 8-base identification tag (differing from each other in at least three positions) and a terminal base (C) (Clemmensen *et al.*, 2016). Three negative controls (sterile water) with unique tags were also included (Fig. S1b). Amplification was performed in a 2720 Thermal Cycler (Life Technologies) using a previously optimized PCR mix, according to Ihrmark *et al.* (2012), that contained 750 000 plasmid copies as template in 50 μl reactions with 200 μM of each nucleotide, 2.75 mM $MgCl_2$, primers at 500 nM (fITS7) and 300 nM (ITS4) and 0.025 U μl^{-1} of DreamTaq Green polymerase (Thermo Fisher Scientific) in the buffer supplied by the manufacturer. Cycling conditions were as follows: 5 min at 95°C, followed by 25 cycles of 30 s at 95°C, 30 s at 57°C, 30 s at 72°C, and a final extension of 7 min at 72°C. Amplicons were purified using the AMPure kit (Beckman Coulter Inc., Brea, CA, USA) and quantified with the Qubit high sensitivity (0.01–100 ng μl^{-1}) DNA quantification kit on a Qubit Fluorometer. Each amplified fragment was checked for quality and length on a BioAnalyzer 7500 chip (Agilent Technologies, Santa Clara, CA, USA). PCR products were adjusted to equal molarity of amplified fragments and pooled in a final library with a total of 750 000 copies of each fragment (Fig. S1b; Table 1).

Experiment 2: PCR and sequencing biases in artificial communities

In this experiment, we tested combined community distortion resulting from PCR and sequencing of artificially assembled communities (Fig. S1c; Table 1). We constructed artificial template communities by distinct pre-PCR combinations of plasmids with the 10 different fragments. In a first setup, all plasmids were mixed in equal copy numbers, but with different amounts of total template DNA (7500 000, 750 000, 75 000, 7500, 750, 75 copies) in different PCRs (Fig. S1c, 'even communities'), yielding different PCR product concentrations. The same volumes of final PCR reactions were pooled, except that volumes of the three samples with the highest starting amounts were downscaled somewhat, giving pooling ratios for the six template levels of 0.1 : 0.3 : 0.8 : 1 : 1 : 1. We intended to investigate the effects of sequencing depth on how closely the sequenced amplicons reflected the original template. In a second setup, the plasmids were mixed in 10 distinct combinations with varying copy numbers (75 000, 7500, 750, 75) of each fragment, but with all samples containing the same total number of copies (167 475), in order to resemble more complex templates (Fig. S1c, 'uneven communities'). An equal amount of DNA from each PCR reaction was pooled before sequencing to investigate the magnitude of combined PCR and sequencing biases and errors.

All community templates were amplified using the same primers and PCR conditions as in Expt 1. Amplifications were performed in three technical replicates, each fitted with distinct 8-base identification tags (as detailed in the Experiment 1: assessment of sequencing biases section, above), resulting in $n = 6 \times 3 = 18$ reactions for the 'even communities' and $n = 10 \times 3 = 30$ PCR reactions for the 'uneven communities'. Amplicons were purified using the AMPure kit (Beckman Coulter) and quantified with a Qubit Fluorometer. The resulting libraries were purified using the EZNA Cycle Pure kit (Omega Bio-Tek, Norcross, GA, USA). The two resulting pools were checked for quality and size distribution using

BioAnalyzer DNA 7500 (Agilent Technologies), and pooled in a molarity ratio of *c.* 0.5 : 1.

Experiment 3: assessing PCR biases in relation to sequencing biases

In this experiment, different PCR conditions were tested on template communities containing the 10 plasmids pooled in equal copy numbers (Fig. S2). We tested starting quantities of 400 000 and 4000 000 total copies per reaction (40 000 and 400 000 copies of each fragment in the two starting mixtures, respectively) and five PCR cycle numbers (21, 25, 28, 31 and 35 cycles) in triplicate reactions ($n = 2 \times 5 \times 3 = 30$ reactions). In each PCR reaction, primers were fitted with distinct ID tags and three negative control samples of sterile water were included. PCR conditions, as well as methods for product purification, quantification of individual amplicons, and final quantification and quality control of the library were the same as in Expts 1 and 2.

High-throughput sequencing

All libraries were sequenced at SciLifeLab, NGI-Uppsala, Sweden, after addition of relevant sequencing adapters by ligation. Libraries from Expts 1 and 2 were combined (1 : 10) and prepared with the ThruPLEX library preparation kit (Rubicon Genomics), and cluster generation and paired-end sequencing were conducted with 300 cycles on a single lane on the MiSeq system using the v3 chemistry (Illumina Inc., San Diego, CA, USA). The same libraries were prepared using the SMRTbell Template Prep Kit 1.0 and sequenced on 8 PacBio RS II SMRT cells (Pacific Biosciences, Menlo Park, CA, USA). The library from Expt 3 was sequenced together with another experiment (1 : 1) with PacBio Sequel using 1 SMRT cell. A standard sample corresponding to the post-PCR equalized community in Expt 1 was also included (among other samples) in three separate runs of PacBio Sequel I.

Bioinformatics

Sequences were quality filtered and clustered using the SCATA pipeline (https://scata.mykopat.slu.se/, accessed 5 July 2019). Sequences with a length of < 100 bases were removed, after which the remaining sequences were screened for primers (requiring 90% match) and sample tags (100% match). Quality scores (ranging between 0 and 40) provided by the sequencing facility indicate the probability of each base to be correctly called. For PacBio datasets, sequences with an average amplicon quality score of < 20 or with a score of < 7 at any position were removed. For the Illumina MiSeq dataset, sequences with an average amplicon quality score of < 20 or with a score of < 10 at any position were removed and paired reads were merged. After collapsing homopolymers to 3 bases (as all current sequencing platforms are sensitive to homopolymers, Laehnemann *et al.*, 2016), sequences were compared pair-wise using USEARCH (Edgar *et al.*, 2011), with all data sets from all sequencing platforms included in a single run. Pairwise alignments were scored using a mismatch penalty of 1, gap open penalty of 0 and a gap extension penalty of 1. Genotypes occurring only once in the global data set were removed before clustering. Sequences were clustered using single linkage clustering with a minimum similarity of 98.5% to the closest neighbour required to enter clusters. Since the Illumina MiSeq technology generates daughter OTUs closely related to the parent OTUs, we also merged the daughter OTUs with their corresponding parent OTUs by applying the post-clustering algorithm LULU, defining minimum_ratio_type = 'min' and minimum_match = 98.5% (Frøslev *et al.*, 2017). Read counts for all OTUs were converted to relative proportions out of the total count of each sample, except in Expt 1 (in which single fragments were amplified), where relative proportions were obtained by dividing the read count of each OTU by the total read number of all the samples in the experiment.

Statistical analysis

Community data was subjected to multivariate analyses using the VEGAN package (Oksanen *et al.*, 2015) in the R software environment (v.2.15.3; R Development Core Team, 2015). The NLME package (Pinheiro *et al.*, 2016) was used to apply linear mixed effects (LME) models.

To test and quantify the effects of fragment length and GC content on community bias in individual samples pooled post-PCR (Expt 1) or pre-PCR (Expt 2), we used linear regressions of relative abundances of mock community members in the sequenced communities against their lengths or GC contents. The correlation between relative abundances and lengths (but not GC contents) of mock community members was always significant (with $P < 0.001$ and $R^2 > 0.7$), and we used the slope of these regressions as a representation of the 'fragment length bias' of each sample. The effect of sequencing platform on fragment length bias was tested by one-way ANOVAs. Relative abundances of mock fragments were uncorrelated with fragment GC contents, which was further confirmed using LMEs, in which length of each community member was considered as a random effect, and GC content as a fixed factor.

In the 'uneven communities' of Expt 2, we used variation partitioning analyses of two Hellinger transformed community combinations (one including 'initial template' and 'MiSeq' and one including 'initial template' and 'PacBio RS II') to test for the relative importance of the 'sequencing' ('initial template' vs 'sequenced') and the pre-PCR initial template composition (10 sample types; Fig. S1). Effects were tested by canonical correlation analysis using 999 permutations. We also used Procrustes analyses ('protest' function) over the same community combinations to test which sequencing platform yielded sequenced amplicon communities that best resembled the template communities.

In Expt 3, PCR bias was disentangled from sequencing bias by subtracting the sequencing bias (i.e. the slope from Expt 1 for PacBio Sequel) from the total bias (i.e. regression slope of observed community representing combined sequencing and PCR bias). The effect of PCR cycle number and template concentration on PCR bias was tested by linear regressions. To test if PCR at different cycle numbers was free from bias, we used

Tukey specific contrasts over LME models in which template concentration was defined as a random factor and cycle number as a fixed factor. To test whether PCR cycle numbers and template concentration affected the prevalence of OTUs other than the initial template, we calculated Hill's N0 diversity index (= richness) using the iNEXT package (Hsieh *et al.*, 2016). N0 was calculated from the asymptotic estimates of detected OTUs against read numbers using the 'iNEXT' function. N0 values corresponding to 4000 reads were derived from all samples, and ANOVA was used to test the effect of PCR cycle number and template concentration on N0 values.

Data availability

Sequence data are archived at NCBI's Sequence Read Archive under accession no. PRJNA604970. Full sequences of the mock community members are available in the Supporting Information file Methods S1. Experimental design, community data and associated data of this study can be found in Mendeley Data, v.2, doi: 10.17632/tw6d38fzyy.2.

Results

Sequencing output

Illumina MiSeq sequencing of the communities obtained from the Expt 1, Expt 2 'even communities' and Expt 2 'uneven communities' resulted in 277 321, 1.3 million and 2.9 million sequences respectively, whereas PacBio RS II sequencing resulted in 7250, 27 313 and 70 502 sequences, respectively (Table S2). In Expt 3, sequencing using PacBio Sequel I resulted in 79 678 ITS mock sequences (Table S2). We detected 2111 OTUs, but *c.* 97.1% of the reads belonged to the 10 mock community members. In total, 698 daughter OTUs of the Illumina MiSeq data set were merged to their corresponding parent OTUs. We identified and deleted sequences of contaminants (*c.* 2.2% of the reads, 971 OTUs) that were randomly distributed across samples.

Experiment 1: assessing sequencing technology biases

All of the 10 mock community fragments were recovered when samples were sequenced with PacBio RS II. However, when the same sample pool was sequenced with Illumina MiSeq, the three longest fragments (> 447 bases) were not detected (Fig. 1). As the 10 fragments constituted equimolar proportions of the amplicon pool, we expected that all fragments would be equally represented in the sequenced communities. However, the relative abundances of the 10 mock members were negatively correlated with the mock member lengths, when using Illumina MiSeq ($P = 0.008$, $R^2 = 0.55$), PacBio RS II ($P = 0.007$, $R^2 = 0.57$) and PacBio Sequel I ($P < 0.001$, $R^2 = 0.88$). By contrast, no significant effect of GC content on the relative proportions of fragments was observed, even after accounting for length bias, for any of the sequencing technologies ($P > 0.2$ for all). Fragment length bias was almost 3 times higher in samples sequenced with Illumina MiSeq (regression slope: − 0.029) than in samples sequenced with PacBio RS II (regression slope: − 0.009) and PacBio Sequel I (regression slope: − 0.009).

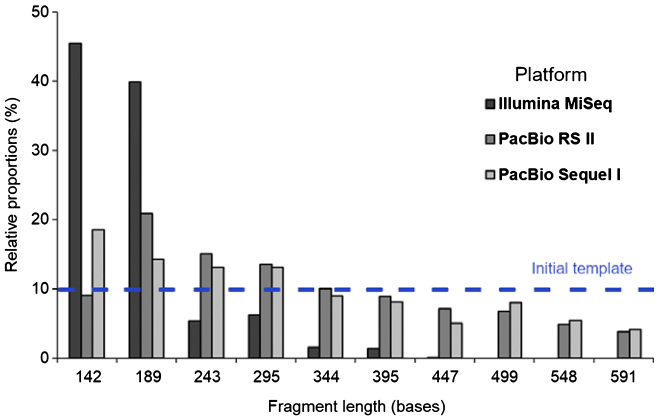

Fig. 1 Relative abundances of 10 internal transcribed spacer 2 (ITS2)-like fragments of different lengths obtained by sequencing with Illumina MiSeq, PacBio RS II and PacBio Sequel I. The fragments were pooled in equal proportions prior to sequencing (i.e. post-PCR), as indicated by the blue dashed line. For PacBio Sequel, data were obtained by including an equimolar mix of the 10 fragments (with one ID-tag) in three separate sequencing runs, whereas RS II and MiSeq data were obtained by tag-multiplexing within a single pool (see Supporting Information Fig. S1, Expt 1).

Experiment 2: PCR and sequencing technology biases in complex artificial communities

In the 'even community', PacBio RS II recovered all of the expected 10 mock fragments in samples with > 750 template copies per PCR reaction (corresponding to at least 247 sequence reads per sample), but several fragments were missing in individual replicates of samples with < 750 template copies (corresponding to 3–38 final reads per sample) (Fig. 2). The two shortest ITS fragments dominated the composition of all samples sequenced with Illumina MiSeq (Fig. 2). The PacBio-based community composition was closer to the initial template composition than the MiSeq-based community, and a reasonably good picture of the template communities could be obtained by PacBio sequencing, even when output was as low as 10–100 reads per sample (Fig. 2).

In the 'uneven community', fragments larger than 447 bases were either not detected or represented < 0.1% of the communities obtained with Illumina MiSeq, even when the longer fragments contributed > 40% of the template (Fig. 3). Variation partitioning revealed that the experimental design (i.e. the distinct initial mock combinations) explained a lower proportion of variation for communities sequenced with Illumina MiSeq (52.8% attributed to initial design) than for communities sequenced with PacBio RS II (97.6% attributed to initial design). Procrustes analyses also showed that the PacBio-based communities resembled the template communities to a greater extent than Illumina MiSeq-based sequence communities did (Fig. 4). The longest ITS fragments, however, occurred at lower abundances than expected with both platforms, although this was more pronounced with Illumina MiSeq (Fig. 3, S3).

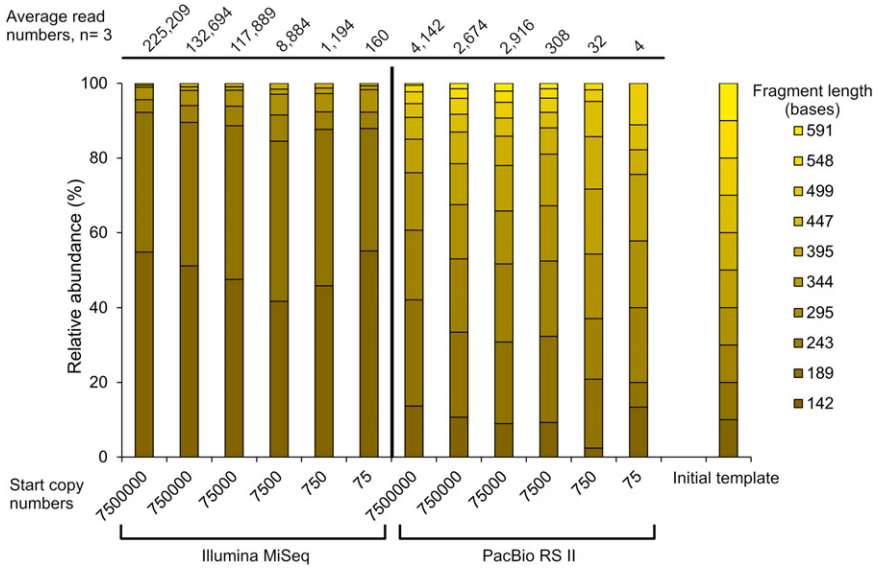

Fig. 2 Results from Expt 2, 'even communities'. Relative abundances of 10 mock community members varying in fragment size (142–591 bp) based on community sequencing with Illumina MiSeq or PacBio RS II. The ITS2 mock community consisted of 10 ITS2-like fragments, added in equal proportions to the initial PCR (initial template), but with different total starting quantities in the PCR (75–7500 000 copies). Above the graph, the averaged observed sequencing output is given for each sample ($n = 3$).

Experiment 3: assessing PCR biases in relation to sequencing biases

Fragment length bias (the slope of linear regressions of mock fragment proportions against their lengths) caused by PCR alone was significantly correlated with the number of PCR cycles ($F = 10.1$, $P < 0.001$, Fig. 5a) but not with the template concentration in the PCR reaction ($F = 0.9$, $P = 0.345$). There was no significant effect of GC content on the relative abundances of mock fragments, even after accounting for length bias ($P > 0.05$ for all samples). The length bias of samples amplified with 22 cycles was not significantly different from the sequencing bias (t-value = 1.84, $P = 0.438$, Table S3), but already PCRs using 25 cycles added significant bias in relation to the biases expected from sequencing alone (t-value = 3.61, $P = 0.004$, Fig. 5a; Table S3).

Furthermore, richness (N0) was positively correlated with both the number of PCR cycles ($F = 6.65$, $P < 0.001$) and the amount of template ($F = 5.27$, $P = 0.032$) (Fig. S4). PCRs with up to 28 cycles resulted in the expected richness of 10 community members, but after 35 cycles, richness was *c.* 50% higher than after 28 cycles.

Discussion

We developed a set of 10 fungal ITS2-like fragments, which can be combined in an ITS mock community and used in DNA metabarcoding studies to assess potential biases during PCR amplification and DNA sequencing. Based on ITS mock communities constructed pre- or post-PCR, we demonstrate that shorter fragments are overrepresented in sequenced communities, whereas variation in GC content had little importance. The magnitude of observed fragment length bias, however, strongly depended on the sequencing platform and the PCR protocol. Sequenced communities from Illumina MiSeq deviated more strongly from the initial template community than sequence communities from PacBio RS II and Sequel I instruments, although PacBio-sequencing was not free from length bias. All fragments over 447 bases completely escaped sequencing on the Illumina platform, whereas fragments shorter than 200 bases were strongly over-represented. Further, fragment length biases were linearly exacerbated with increasing numbers of PCR cycles. Optimization of the PCR protocol with the aim of reducing the number of cycles can result in negligible biases and errors during PCR.

These results apply especially to communities in which large fragment length variation is expected, such as when the ITS region is used as marker (Bellemain *et al.*, 2010). Forest soil samples amplified using the gITS7-ITS4 primer combination typically contain fungal ITS2 amplicons with lengths ranging from *c.* 200 bases (e.g. *Archaeorhizomyces* spp.) up to *c.* 450 bases (e.g. *Gymnopus* spp.), 600 bases (certain *Entoloma* spp.) and even > 1000 bases (*Cantharellus* spp.), while amplicon lengths of other common cosmopolitan fungal taxa such as *Mortierella* spp. or *Leccinum* spp. are around 400 bases (See figure 6 in Clemmensen *et al.,* (2016) for an example of an amplicon size profile in a multiplexed ITS2 mix). Although the variation in length for most taxa is thus expected to be 200–500 bases, that is, less than the variation of the ITS2 mock used here (142–591 bases), the fact that length bias was obvious for fragments of 189 to 499 bases suggests that there is also a major risk of significant length bias when sequencing natural communities. This risk may be particularly large for certain sample types, and if long primer constructs (adding further length to the amplicon) are used in the PCR.

Recent studies based on artificial communities have shown that length variation across fragments is an important source of bias

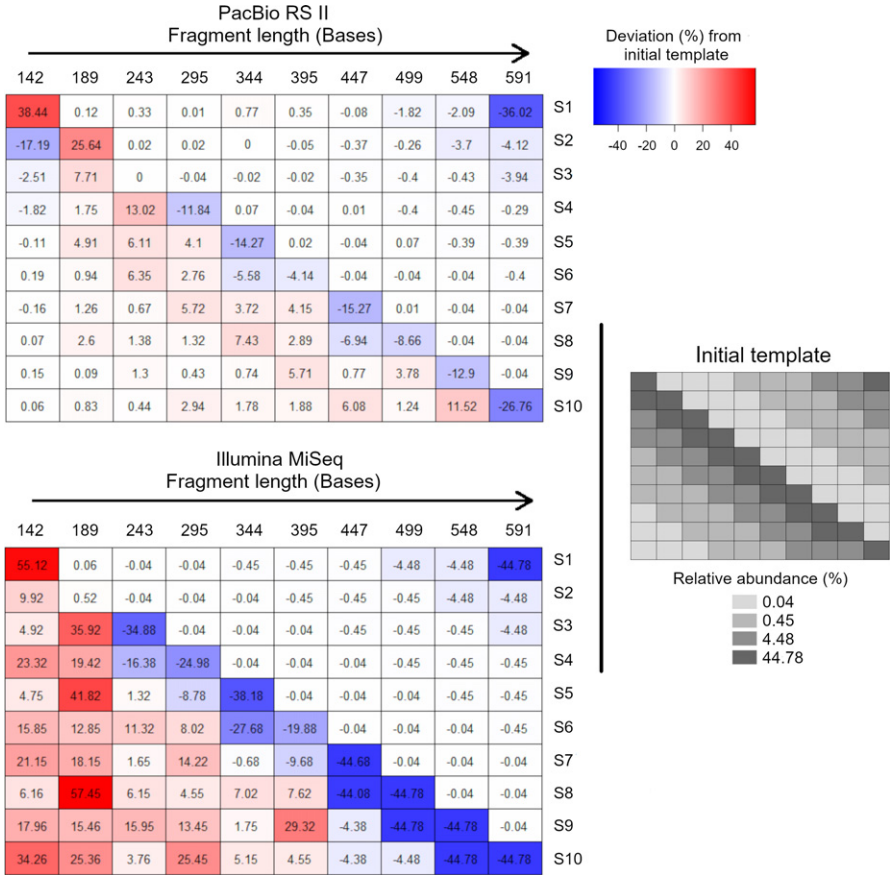

Fig. 3 Results of Expt 2, 'uneven communities'. (a) Heat maps represent the deviation in relative abundances of 10 mock members of varying fragment length (142–591 bp) between the initial template and the observed communities obtained by PacBio RS II or Illumina MiSeq sequencing. The initial template was specific to each sample (S1–S10), with mock members mixed in four orders of magnitude abundance levels according to the greyscale legend, further specified in Supporting Information Fig. S1(c). For each sample, the average of three replicates is given. Colours from blue to red indicate a gradient from negative deviations (observed values lower than template values) to positive deviations (observed values higher than template values), while values close to zero (white) indicate little deviation from the template.

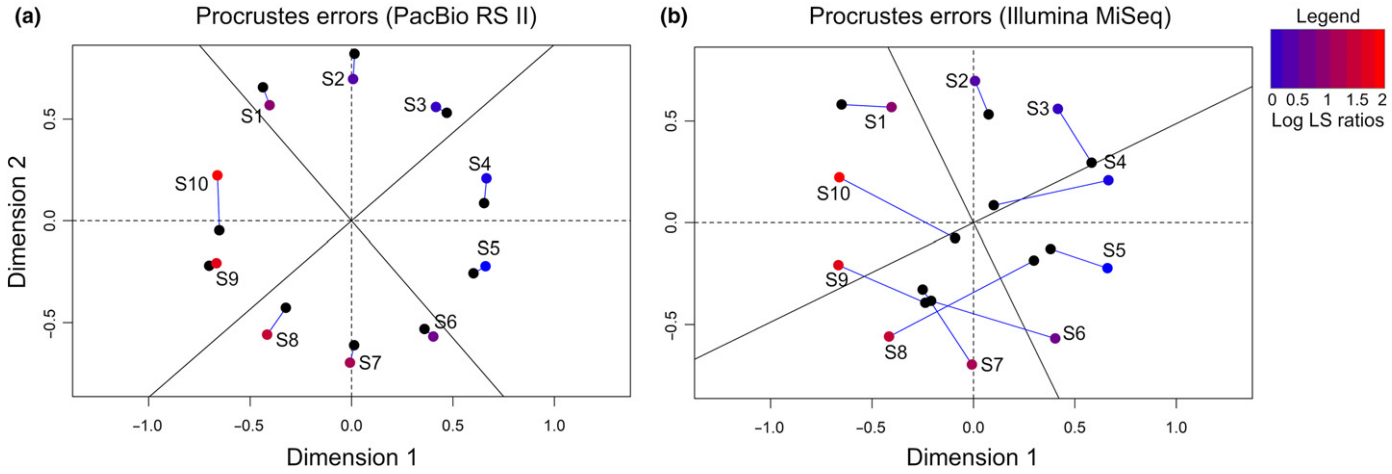

Fig. 4 Procrustes analysis of results from Expt 2, 'uneven communities', showing the degree of match between two ordinations: (a) template community (i.e. pre-PCR) vs community obtained by PacBio RS II sequencing, and (b) template community vs community obtained by Illumina MiSeq. Black points show the position of communities obtained after sequencing, and lines connect to the corresponding template community, with dots coloured according to the ratio in abundance between the three longest and the three shortest fragments (L/S). Each dot represents averaged values from three replicates ($n = 3$). The two continuous lines crossing the plot show the rotation between the two ordinations giving the optimal match between ordinations.

when sequencing is performed using MiSeq or IonTorrent (Bakker, 2018; Palmer *et al.*, 2018). These authors found highly distorted communities after PCR and sequencing and argued for the interpretation of sequencing data as merely binary (presence–absence) rather than semi-quantitative (Palmer *et al.*, 2018). However, while length bias appears to be highly problematic when using Illumina MiSeq, we show here that length bias can be almost eliminated using the PacBio sequencing platforms and that PacBio sequencing data can be interpreted as semi-quantitative (> 97% of variation across sequence communities in the 'uneven communities' experiment was attributed to template design; Figs 3, 4). In addition, length biases with PacBio Sequel I

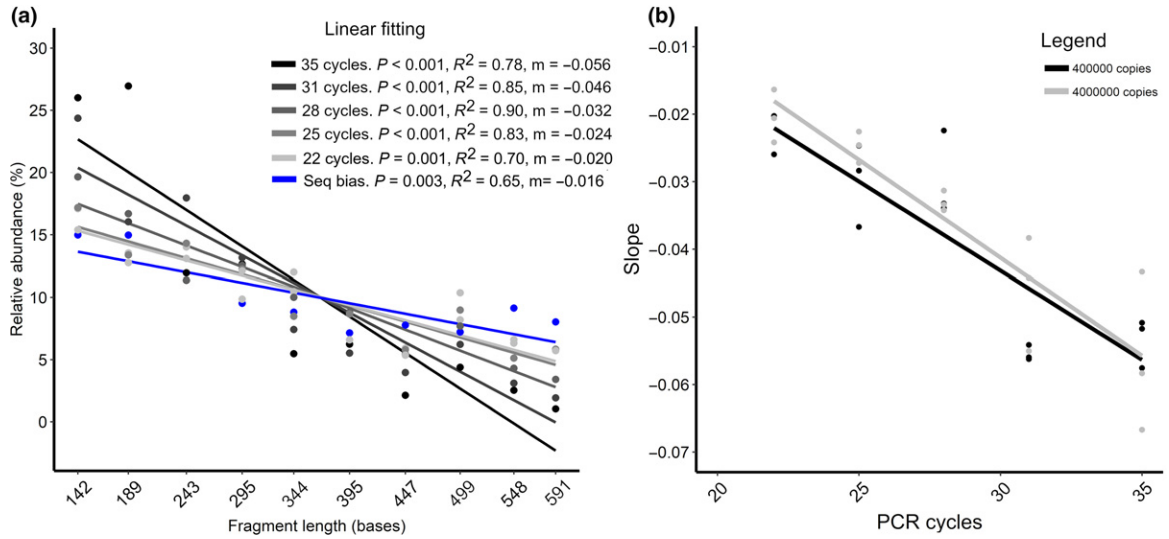

Fig. 5 (a) Linear regressions representing the effect of fragment length on relative abundances of 10 mock members in communities that were evenly mixed and PCR amplified using an initial total quantity of 400 000 copies and 22–35 cycles ($n = 3$). Slope estimates ('m') are used to represent length bias in each sample. The blue line represents the sequencing bias (Seq bias) as estimated in communities assembled post-PCR. (b) Correlations between PCR cycles and slope estimates, with lower slope estimates indicating larger fragment length bias during PCR amplification. All samples were sequenced with PacBio Sequel I.

sequencing were predictable and feasible to correct for to obtain a sequence community composition that more closely reflects the template composition, provided that a mock-community is included in the sequencing run. In contrast to what we expected, differences in GC content did not have any effect on the relative abundances of the sequenced fragments. This contrasts with previous findings (Benjamini & Speed, 2012), but in our study we used a lower range in GC content than in previously published studies.

Although the choice of sequencing platform was the main factor that contributed to length bias, we observed that biases were exacerbated by PCR amplification. Thus, our results emphasize the advice to minimize the number of PCR cycles (Polz & Cavanaugh, 1998; Lindahl *et al.*, 2013; Palmer *et al.*, 2018; Nilsson *et al.*, 2019). Despite this, many studies still use excessive numbers of cycles (*c.* 35 cycles), and in some cases > 40 cycles are used when two-step PCRs are performed. For example, during Illumina MiSeq library preparation, there are two PCR steps involved: one for library preparation and one during the cluster generation on the glass (https://support.illumina.com), which may contribute to the perceived sequencing bias. Chimeric sequences may also be created if excessive cycles are used (Haas *et al.*, 2011). In our study, all samples that contained < 3 ng DNA μl^{-1} in the final PCR products had very low PCR-related fragment length bias, highlighting that the aim should be to obtain rather weak final PCR products, as assessed on an agarose gel. In fact, in addition to the linearly increasing length biases with increasing numbers of cycles, we observed that satellite clusters generated by error accumulation in the amplicons were higher in samples amplified with the highest cycle numbers. Template concentration did not affect bias directly, but higher template concentrations may reduce bias indirectly by permitting a lower number of PCR cycles. Thus, we advise that PCR bias can be reduced by reducing the number of PCR cycles and that the possibility of using PCR-free kits for Illumina library preparation should be explored.

We did not intend to assess other, previously acknowledged, sources of bias, such as primer selection or the choice of bioinformatics pipeline (Anslan *et al.*, 2018; Nilsson *et al.*, 2019; Pauvert *et al.*, 2019) – rather, we intended to compare the two sequencing approaches at their optimum performance. Single linkage clustering, as implemented in the SCATA pipeline (https://scata.mykopat.slu.se/), was able to assemble all the expected fragments generated with PacBio RS II and Sequel, with no extra satellite clusters generated besides the ones created during amplification with an excessive number of PCR cycles or through lab contamination. This was not the case for Illumina data, which contained many 'daughter' OTUs closely related to the expected mock fragments. Most of these daughter OTUs were merged with their parents using LULU (Frøslev *et al.*, 2017). However, we noted that some daughter OTUs were detected without their corresponding parent mock fragment, presumably because the parent fragment exceeded the maximum length permitted by Illumina MiSeq. The appearance of artificial OTUs in Illumina data should be considered, particularly when richness values are compared.

Although we were able to derive semi-quantitative fungal community descriptions, which accurately reflected the relative composition of the template, based on amplification and sequencing of fungal ITS2 markers, it is still not possible to extrapolate such data to absolute abundances in terms of biomass. For example, it is known that the nucleus to biomass ratio as well as the ITS copy numbers per genome differ across fungal taxa (Baldrian *et al.*, 2013; Löfgren *et al.*, 2018), and even between strains from the same species (Herrera *et al.*, 2009). This can lead to erroneous conversions of read abundances to fungal biomass, and other markers, such as ergosterol or single copy genes, should be quantified by complementary quantitative approaches to enable more reliable interpretation of ITS-based community data.

Illumina MiSeq is the most widely used sequencing platform in fungal metabarcoding studies (Nilsson *et al.*, 2019). However, the large length variation of the fungal ITS region makes this sequencing platform prone to vastly over-represent shorter fragments (*c.* 200 bases) and underrepresent fragments of intermediate sizes, while longer fragments are not sequenced at all. Although length biases were detected for all platforms, PacBio Sequel I had the least length biases, and PacBio-based estimates of community composition were much closer to the template, even at sequencing depths as low as 30 sequence reads per sample (or < 5 reads per mock member). The higher sequencing depth (at similar cost) is one of the main advantages of Illumina MiSeq sequencing, especially for samples with a low proportion of fungal sequences. However, a more specific primer pair, for example fITS7-ITS4 for fungi, could be an alternative for obtaining sufficient sequencing output for the organism group of interest (Ihrmark *et al.*, 2012). Our results indicate that the higher sequencing depth of Illumina MiSeq does not compensate for its inherent length bias and low accuracy. Our conclusion is that Illumina sequencing of ITS markers yields a detailed but very distorted view of fungal communities, while PacBio sequencing yields less distorted results, even at low sequence output.

Acknowledgements

This work was funded by Swedish University of Agricultural Sciences (SLU); seed money grant from the plant protection platform to CC, KEC, ÅO, AB. We further acknowledge financial support from the Department of Forest Mycology and Plant Pathology, SLU, to ÅO and AB and from the Swedish Research Council's FORMAS (grant 2011-1747) to BDL and KEC and VR to KEC (grant 2015-04411). The authors would like to acknowledge support from Olga Vinnere Pettersson and the National Genomic Infrastructure (NGI)/Uppsala Genome Centre for providing assistance in massive parallel sequencing. Work performed at Uppsala Genome Centre was funded by RFI/VR (Swedish Research Council) and Science for Life Laboratory, Sweden. Three anonymous reviewers are acknowledged for their valuable comments on the manuscript.

Author contributions

CC, AB, MBD, KI, BDL, JS, KEC and ÅO contributed to the design of the experiments. The laboratory work for Expts 1 and 2 was performed by ÅO, AB and KI and for Expt 3 by CC. Bioinformatics analyses were carried out by CC, with support from MBD and BDL. Data analyses and manuscript preparation were executed by CC and KEC with inputs from all authors.

ORCID

Anna Berlin https://orcid.org/0000-0002-9518-5719
Mikael Brandström Durling https://orcid.org/0000-0001-6485-197X
Carles Castaño https://orcid.org/0000-0002-2403-7006
Karina E. Clemmensen https://orcid.org/0000-0002-9627-6428
Björn D. Lindahl https://orcid.org/0000-0002-3384-4547
Åke Olson https://orcid.org/0000-0001-8998-6096
Jan Stenlid https://orcid.org/0000-0002-5344-2094

References

Aas AB, Davey ML, Kauserud H. 2016. ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources* **17**: 730–741.

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

Anslan S, Nilsson RH, Wurzbacher C, Baldrian P, Tedersoo L, Bahram M. 2018. Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *MycoKeys* **39**: 29–40.

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

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

Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H. 2010. ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *BMC Microbiology* **10**: 189.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**: 1–14.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW. 2017. GenBank. *Nucleic Acids Research* **46**: 41–47.

Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kauserud H. 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology* **5**: 747–749.

Castaño C, Lindahl BD, Alday JG, Hagenbo A, Martínez de Aragón J, Parladé J, Pera J, Bonet JA. 2018. Soil microclimate changes affect soil fungal communities in a Mediterranean pine forest. *New Phytologist* **220**: 1211–1221.

Clemmensen KE, Ihrmark K, Brandström Durling M, Lindahl BD. 2016. Sample preparation for fungal community analysis by high-throughput sequencing of barcode amplicons. *Methods in Molecular Biology* **1399**: 61–88.

D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**: 55.

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

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications* **8**: 1188.

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R *et al.* 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology* **34**: 942–949.

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.

Herrera ML, Vallor AC, Gelfond JA, Patterson TF, Wickes BL. 2009. Strain-dependent variation in 18S ribosomal DNA copy numbers in *Aspergillus fumigatus*. *Journal of Clinical Microbiology* **47**: 1325–1332.

Hsieh KH, Ma KH, Chao A. 2016. *iNEXT: iNterpolation and EXTrapolation for species diversity. R package v.2.0.8.* [WWW document] URL https://cran.r-project.org/web/packages/iNEXT/index.html.

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

Kennedy PG, Cline LC, Song Z. 2018. Probing promise versus performance in longer read fungal metabarcoding. *New Phytologist* **217**: 973–976.

Kyaschenko J, Clemmensen KE, Hagenbo A, Karltun E, Lindahl BD. 2017a. Shift in fungal communities and associated enzyme activities along an age gradient of managed *Pinus sylvestris* stands. *ISME Journal* **11**: 863–874.

Kyaschenko J, Clemmensen KE, Karltun E, Lindahl BD. 2017b. Below-ground organic matter accumulation along a boreal forest fertility gradient relates to guild interaction within fungal communities. *Ecology Letters* **20**: 1546–1555.

Laehnemann D, Borkhardt A, McHardy AC. 2016. Denoising DNA deep sequencing data – high-throughput sequencing errors and their correction. *Briefings in Bioinformatics* **17**: 154–179.

Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjøller R, Köljalg 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.

Löfgren LA, Uehling JK, Branco S, Bruns TD, Martin F, Kennedy PG. 2018. Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Molecular Ecology Resources* **28**: 721–730.

May A, Abeln S, Buijs MJ, Heringa J, Crielaard W, Brandt BW. 2015. NGS-eval: NGS error analysis and novel sequence variant detection tool. *Nucleic Acids Research* **43**: W301–W305.

Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L. 2019. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nature Review Microbiology* **17**: 95–109.

Oksanen FG, Blanchet R, Friendly M, Kindt P, Legendre PR, McGlinn D, Minchin B, O'Hara RB, Simpson GL, Solymos P, et al. 2015. *VEGAN: community ecology package. R package v.2.2-1* [WWW document] URL https://cran.r-project.org/web/packages/vegan/index.html. [accessed 14 August 2019].

Olson Å, Aerts A, Asiegbu F, Belbahri L, Bouzid O, Broberg A, Canbäck B, Coutinho PM, Cullen D, Dalman K *et al.* 2012. Insight into trade-off between wood decay and parasitism from the genome of a fungal forest pathogen. *New Phytologist* **194**: 1001–1013.

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

Pauvert C, Buée M, Laval V, Edel-Hermann V, Fauchery L, Gautier A, Lesur I, Vallance J, Vacher C. 2019. Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology* **41**: 23–33.

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports* **8**: 10950.

Pinheiro J, Bates D, DebRoy S, Sarkar D; R Core Team. 2016. *nlme: Linear and nonlinear mixed effects models. R package v.3.1-128.* [WWW document] URL http://CRAN.R-project.org/package=nlme [accessed 17 August 2019].

Polz MF, Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* **64**: 3724–3730.

Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

R Core Team. 2015. *R: a language and environment for statistical computing, v.3.5.0.* Vienna, Austria: R Core Team. [WWW document] URL http://cran.r-project.org/ [accessed 17 August 2019].

Redondo MA, Boberg J, Stenlid J, Oliva J. 2018. Contrasting distribution patterns between aquatic and terrestrial *Phytophthora* species along a climatic gradient are linked to functional traits. *ISME Journal* **12**: 2967–2980.

Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Molecular Cell* **58**: 586–97.

Schrader C, Schielke A, Ellerbroek L, Johne R. 2012. PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology* **113**: 1014–1015.

Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, Gies EA, Cheng J-F, Copeland A, Klenk H-P *et al.* 2016. High resolution phylogenetic microbial community profiling. *ISME Journal* **10**: 2020–2032.

Sterkenburg E, Clemmensen KE, Ekblad A, Finlay RD, Lindahl BD. 2018. Contrasting effects of ectomycorrhizal fungi on early and late stage decomposition in a boreal forest. *ISME Journal* **12**: 2187–2197.

Tedersoo L, Anslan S. 2019. Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. *Environmental Microbiology Reports* **11**: 659–668.

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

Tedersoo L, Tooming-Klunderud A, Anslan S. 2018. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist* **217**: 1370–1385.

Varenius K, Lindahl BD, Dahlberg A. 2017. Retention of seed trees fails to lifeboat ectomycorrhizal fungal diversity in harvested Scots pine forests. *FEMS Microbiology Ecology* **93**: fix105.

Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiology* **16**: 274.

White TJ, Burns T, Lee S, Taylor J. 1990. Amplification and sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, eds. *PCR protocols. A guide to methods and applications.* San Diego, CA, USA: Academic Press, 315–322.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Experimental design of Expt 1 and 2.

Fig. S2 Experimental design of Expt 3.

Fig. S3 Composition of the mock communities before and after sequencing.

Fig. S4 Relationship between richness and PCR cycles.

Table S1 Primers used to amplify the selected fragments in the *Heterobasidion irregulare* genome.

Table S2 Bioinformatic statistics of the quality control and clustering of the mock community data.

Table S3 Specific slope value contrasts for distinct PCR cycles.

Methods S1 Full sequences of the mock community members (as file all_ref_in_one.fasta).

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.