MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 5555–5565 doi: 10.1111/j.1365-294X.2010.04898.x

Quantifying microbial communities with 454 pyrosequencing: does read abundance count?

ANTHONY S. AMEND,* KEITH A. SEIFERT† and THOMAS D. BRUNS*

*Department of Plant and Microbial Biology, University of California Berkeley, 111 Koshland Hall, Berkeley, CA 94720, USA, †Agriculture and Agri-Food Canada, 960 Carling Ave., Ottawa, ON, Canada K1A 0C6

Abstract

Pyrosequencing technologies have revolutionized how we describe and compare complex microbial communities. In 454 pyrosequencing data sets, the abundance of reads pertaining to taxa or phylotypes is commonly interpreted as a measure of genic or taxon abundance, useful for quantitative comparisons of community similarity. Potentially systematic biases inherent in sample processing, amplification and sequencing, however, may alter read abundance and reduce the utility of quantitative metrics. Here, we examine the relationship between read abundance and biological abundance in a sample of house dust spiked with known quantities and identities of fungi along a dilution gradient. Our results show one order of magnitude differences in read abundance among species. Precision of quantification within species along the dilution gradient varied from R^2 of 0.96–0.54. Read-quality based processing stringency profoundly affected the abundance of one species containing long homopolymers in a read orientation-biased manner. Order-level composition of background environmental fungal communities determined from pyrosequencing data was comparable with that derived from cloning and Sanger sequencing and was not biased by read orientation. We conclude that read abundance is approximately quantitative within species, but between-species comparisons can be biased by innate sequence structure. Our results showed a trade off between sequence quality stringency and quantification. Careful consideration of sequence processing methods and community analyses are warranted when testing hypotheses using read abundance data.

Keywords: 454 Life Sciences pyrosequencing, beta diversity, fungal community analysis, quantitative PCR, ribosomal DNA

Received 19 May 2010; revision received 18 September 2010; accepted 21 September 2010

Introduction

Culture independent analysis of sequences derived from samples of environmental genomic nucleic acids has revolutionized our understanding of microbial diversity, function and processes (Stahl *et al.* 1984; Hugenholtz & Pace 1996). Technological advances such as pyrosequencing enable rapid characterization of microbial communities that are faster and at greater sequence depth than was deemed possible via cloning and Sanger sequencing (Sogin *et al.* 2006). Concurrent with these technological advances are new ideas and methods for statistical comparison of complex microbial communities (Schloss 2008).

A common assumption in targeted sequence analysis of environmental samples is that read abundance correlates with genic or taxon abundance. *Strictly* quantitative measures, in which absolute abundances are inferred, can be discounted outright for current pyrosequencing technologies, as upper (and effectively lower) constraints on the number of total reads produced are imposed. A looser definition of a 'quantitative' measure (used herein) is the correlation between proportional read abundance and the proportional abundance of a given organism in relation to its neighbours. Intuitively,

Correspondence: Anthony S. Amend, Fax: (510) 642 4995; E-mail: a.amend@berkeley.edu

a dominant community member should concomitantly dominate a pyrosequencing data set.

Unfortunately, the ability to predict the prevalence of an organism in an environment based on the abundance of gene copies in a metagenomic DNA template is a tenuous proposition. Portions of the ribosomal operon (rDNA) that serve as *de facto* 'barcodes' for prokaryotic and fungal diversity studies are generally present in multiple tandem copies within a genome. This adds to the region's usefulness for polymerase chain reaction (PCR)-based detection methods where DNA is in short supply or inhibitory substances in a substrate require template dilution. Ribosomal gene copy numbers, however, are known to vary by an order of magnitude among species of Bacteria (Lee *et al.* 2009) and Fungi (Rooney & Ward 2005), probably distorting any semblance of a constant copy number to individual relationship.

Although rDNA gene copy number variation may preclude direct quantification of biological abundance from environmental samples, several well-known downstream biases probably skew the relative abundance of a gene detected in a pyrosequencing read count further. Such biases associated with nucleic acid extraction (DeSantis *et al.* 2005; Feinstein *et al.* 2009), primer selection (Jumpponen 2007; Engelbrektson *et al.* 2010) and PCR (Polz & Cavanaugh 1998) are well enumerated, although biases can be altered or mitigated by carefully selected protocols or alternative approaches (Williams *et al.* 2006; Hori *et al.* 2007; Li *et al.* 2008). Additional steps innate to 454 pyrosequencing preparation and analyses, which often include use of long concatenated primers with varying multiplex sequences, an additional (emulsion) PCR and well-documented systematic base-calling errors (Huse *et al.* 2007; Quinlan *et al.* 2008; Rozera *et al.* 2009; Kunin *et al.* 2010), presumably exacerbate these biases.

A less wishful assumption is that read counts are *semi*-quantitative: producing metrics meaningful for comparing the proportional abundance of a given species with itself across samples. Theoretically, variables that differentially affect the processing of individuals in an environment into a count of 454 pyrosequencing reads (such as priming site homology or cell wall durability) are consistent within species. Therefore, the differences in the proportional abundance of a given species across samples should be biologically meaningful and reflect the actual proportional abundance of that species in the environment.

Both quantitative and semi-quantitative data can be informative for comparing communities using quantitative distance metrics such as weighted UniFrac (Lozupone *et al.* 2007), Bray–Curtis index (Bray & Curtis 1957) or quantitative versions of the Jaccard and Sorenson indices (Chao *et al.* 2005). Although studies estimated abundance in community studies using previous technologies (i.e. fluorescence intensity of T-RFLP peaks, or species counts in clone libraries), the sampling depth derived from 454 data sets has made quantification much more common. Statistical methods based on these abundance data can have significant impact on results compared with presence/absence analyses (Arnold *et al.* 2002; Lozupone *et al.* 2007; Taylor *et al.* 2007; Costello *et al.* 2009).

Community dissimilarity inferred using quantitative metrics tend to be particularly pronounced in microbial communities with a 'long-tailed' or 'J-shaped' distribution, in which a minority of taxa are represented by relatively numerous reads, and the rest are represented by relatively few. In quantitative analyses the significance that these rare taxa play in determining community similarity is diminished or excluded altogether. The results of several studies suggest that a substantial percentage of rare taxa in pyrosequencing studies are the result of methodological artefacts such as sequencing error and/or overly divisive methods of determining operational taxonomic units (OTUs; Quince *et al.* 2009; Reeder & Knight 2009; Kunin *et al.* 2010). Although the magnitude of uncultivated microbial community diversity remains a topic for debate, the long-tailed distributional pattern of those communities appears to be consistent and real (Morales *et al.* 2009). Knowing little about the functional role or true physical abundance of the 'rare' organisms *in situ*, it is unclear how to account for them statistically.

Here, we assess the feasibility of using 454 pyrosequencing read abundance as a quantitative metric in studies of fungal diversity. Because quantification depends on taxonomic assignment and binning, we also address the topics of within-individual sequence variance (both spurious and biological) and sequence orientation with regards to their affects on quantitative studies of fungal communities.

We spiked an environmental dust sample with known quantities and identities of fungal spores in a dilution gradient, and subject the samples to typical processing steps including DNA extraction, PCR amplification and multiplexed 454 FLX Titanium sequencing. We consciously chose not to control for variance in template quantity among species because taxon, not genic abundance, is typically what we are interested in comparing among communities, and the relationship between the two is usually unknown. Although this experimental design does not allow us to partition sources of bias, it allows a realistic assessment of quantification encountered in studies of environmental fungal samples.

Materials and methods

A schematic of the experimental design is included in Fig. S1 (Supporting Information).

Sample preparation

A large house dust sample was collected from the storage container of a central vacuum system from a home in Stittsville, Ontario, Canada in September 2008. Large fragments were removed using a 2-mm sieve, and the sample was refrigerated at 4 °C until extracted the following month.

Fungal spores were isolated from single sporocarps of *Rhizopogon salebrosus* A.H. Sm., *Laccaria proxima* (Boud.) Pat., *Tricholoma imbricatum* (Fr.) P. Kumm., *Russula sanguinea* (Bull.) Fr. and *Thelephora terrestris* Ehrh., collected from Pt. Reyes National Seashore, and yeast cells from a culture of *Metschnikowia noctiluminum* N.H. Nguyen, S.O. Suh, Erbil & M. Blackw (USDA ARS collection: NRRL Y-27753). Duplicate 10 μL aliquots of spore suspensions were mounted on an improved Neubauer Brightline hemacytometer (Hausser Scientific) and examined under 400× magnification. Spore concentrations were quantified by averaging the number of spores contained within ten 0.004-mm^3 boxes from each aliquot. Added species were selected to be probabilistically absent from house dust (all species are biotrophic), and distinguishable based on the internal transcribed spacer (ITS) region of the rDNA. Initial Sanger sequencing of clone pools detected none of these taxa in the prespiked samples. Duplicate aliquots containing no added spores were also subsequently 454-sequenced and did not contain any of the added species. Dilutions of spore slurries containing 10^6, 10^5, 10^4, 10^3, and 10 spores from each species were pooled and added to duplicate 100-mg aliquots of the dust sample for each concentration prior to DNA extraction. Duplicates of five separate dust sample aliquots were augmented with spores, and subject to DNA extraction and PCR amplification.

DNA extraction

Genomic DNA was extracted from dust sample aliquots by mechanical cell lysis with a bead mill, followed by a phenol/chloroform extraction, a chloroform wash and an additional purification using a MoBIO PowerSoil DNA Isolation Kit (MoBio) as previously published (Amend *et al.* 2010). Extracted DNA was eluted with 50 μL of 1X TE (10 mM Tris pH 8, 1 mM EDTA). DNA concentration was quantified on a NanoDrop spectrophotometer (Thermo Scientific).

PCR and sequencing

Approximately 2.5 ng of each DNA extract [the mass of *c.* 2 (10^9) ITS copies or 76 500 fungal genomes containing 30 mb] was added to a PCR cocktail containing 1.2 units of HotStarTaq polymerase (Qiagen), 1× PCR buffer supplied by the manufacturer, 200 μM of each DNTP, 0.5 μM of each primer, and H_2O to a final concentration of 25 μL. ITS amplification used concatemers containing (in order from 3′ to 5′) the fungal-specific primer ITS1f (Gardes & Bruns 1993) or ITS4 (White *et al.* 1990), an 8-bp multiplex tag, and the 454 'A' adaptor CCATCTCATCCCTGCGTGTCTCCGACTCAG (in the forward direction) and the complimentary primer ITS1f or ITS4 and the 454 'B' adaptor CCTATCCC CTGTGTGCCTTGGCAGTCTCAG (in the reverse direction). Each aliquot was amplified twice to allow sequencing in two orientations. The resulting amplicons contained either the ITS1 or ITS2, and depending on gene and read length, either portions or entire reads of the intercalary 5.8s gene and the other ITS region. Following an initial denaturation step at 95 °C for 15 min, PCR was cycled 34 times at 95 °C for 1 min, 51 °C for 1 min, 72 °C for 1 min, and a final extension at 72 °C for 7 min. Negative controls were run on both DNA extractions and PCR.

Polymerase chain reaction products were cleaned using the QIAquick PCR Purification Kit (Qiagen) following the manufacturer's instructions, quantified using a Qubit Fluorometer (Invitrogen), pooled into equimolar concentrations and pyrosequenced at the Duke University ISGP Sequencing Facility on three of eight regions of a 454 Life Sciences FLX Titanium Shotgun plate (454 Life Sciences) in multiplex with another study (Amend *et al.* 2010). In total, 14 997 reads were generated for this study.

DNA from the dust sample with no additional spores was PCR-amplified as above, and inserted into *Escherichia coli* competent cells using the TOPO-TA pCR 2.1 TOPO kit (Invitrogen) as per the manufacturer's instructions. Ninety-five colonies were picked, PCR-amplified using ITS1F and ITS4 primers, and sequenced using the ITS1F primer following standard protocols. DNA from spores of each of the six added individuals was extracted, PCR amplified and directly sequenced without cloning in two directions using the ITS1F and ITS4 primers using standard protocols.

Pyrosequence processing

Sequences were culled if they (i) were shorter than 300 bp, (ii) contained an ambiguous base call, or (iii) imperfectly matched the priming site or multiplex tag

sequences. Reads were sorted by multiplex tags and were screened in the CLC Genomics Workbench 2 using the modified Mott trimming algorithm (comparable to Phred score processing), which evaluates sequences based on 454 quality score. Where base calls did not meet a cumulative error probability threshold of 0.01 they were trimmed from the ends of reads (details of the algorithm and its implementation are available in the CLC Genomics Workbench user manual). Following sequence processing, 3596 forward sequences, 4379 reverse sequences and 67 Sanger sequences remained. 454 pyrosequencing reads were deposited in the NCBI Short Read Archive under accession SRS010093, and Sanger sequences were submitted to GenBank under Accession nos GU931704–GU931774.

Identifying and analysing spiked reads for quantification analyses

Sanger sequences derived from each of the six added species were queried against the reads from each dilution pool using BLASTN (Altschul *et al.* 1997). Reads matching the target sequences aligned over the entire read length (>300 bp as per our sequence processing parameters). Nonmatching reads also partially aligned to target sequences, but these alignments were generally aligned at only the *c.* 170-bp region of the 5.8s gene (Fig. 1a). BLAST alignments were also used to screen for chimeric sequences, resulting from erroneous union of DNA from two individuals. Reads with alignment lengths intermediate of 170 and 300 bp, or with alignment lengths <30 bp of the total read length were flagged as potentially chimeric. This method effectively detects chimeric sequences in which the breakpoint is located in the 5.8s. Putative chimeras were manually confirmed by comparing the 5′ and 3′ halves of the read against the NCBI nt database. Matching reads were counted and summed across all pools (Fig. 2).

To calculate the relative abundance of each added species per pool, matching reads were divided by the total number of processed reads in the pool (N = 369–1350; Fig. 3). Although data were constrained to relative proportions, and thus bounded by upper limits, we choose a logarithmic regression model to examine read abundance as regression residuals were not clustered near 0 or 1.

Extracting and comparing within-species ITS sequence variation

The 18s, 5.8s and 28s transcribed regions of the ribosomal operon are many times more conserved than either ITS1 or ITS2, which are more useful for delineating taxonomy at a resolution near the species level for most fungal lineages (Ryberg *et al.* 2008). Due to the short (*c.* 400 bp) read lengths of 454 data, however, the ITS regions, ranging from *c.* 83 bp in *M. noctiluminum* ITS1 to *c.* 244 bp in *T. imbricatum* ITS1 comprise varying proportions of the total read. Therefore, read length and the relative proportion of conserved ribosomal subunit to variable spacer DNA in any read may affect compar-

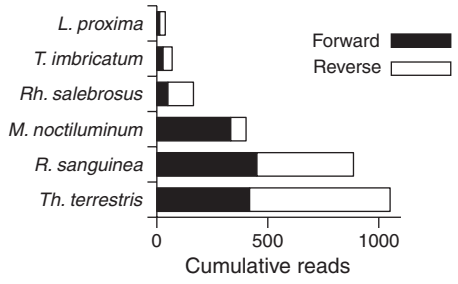

Fig. 2 Read abundance differed by an order of magnitude between the most (*Thelephora terrestris*) and least (*Laccaria proxima*) abundant taxa. The same number of spores was added for each species. Read abundance did not correlate with PCR fragment length or spore size.

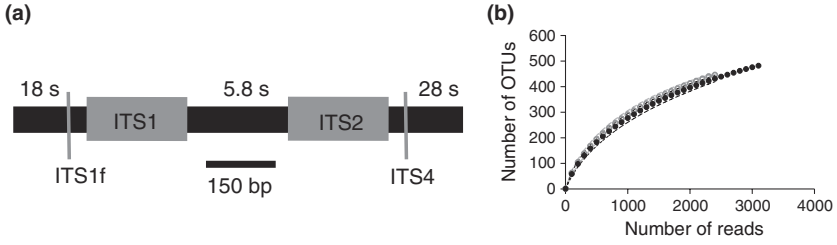

Fig. 1 Map of the ITS region of the ribosomal operon (a) shows the relative positions of the conserved regions (black bars) to the variable regions (grey bars). Scale is approximate because length of the variable regions varies among species. The 400-bp reads originating from the ITS1f (forward) or ITS4 (reverse) will contain both the variable and conserved regions in varying proportions. Rarefaction analysis of OTUs based on 97% identity (b) between dust fungal community diversity inferred from forward- (black circles) and reverse- (grey circles) sequenced communities are nearly identical. Error bars denote 95% confidence intervals.

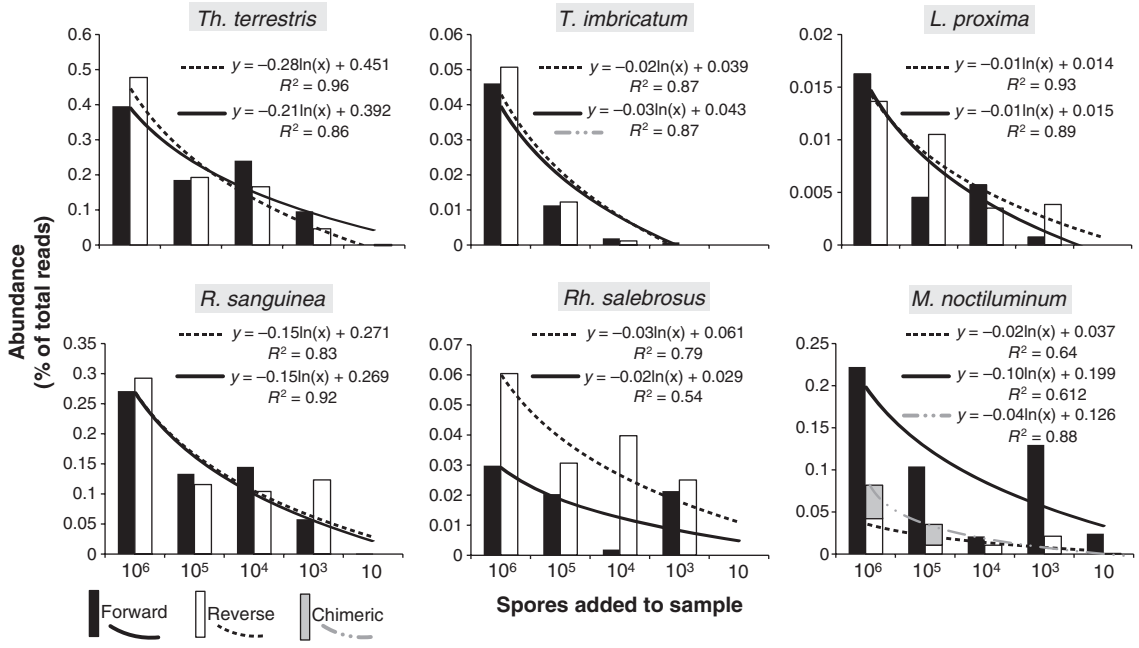

Fig. 3 Read abundance semi-quantification varies among samples. Slopes and R^2-statistics are based on logarithmic regressions of the proportion of reads in the pool to the number of spores added. Note the varying *Y*-axis scale.

ative sequence identity (Nilsson *et al.* 2008). To calculate a more standardized value of within-individual variation at this locus, we excluded the 18s, 5.8s and 28s regions using the Fungal ITS Extractor 1.01 (Nilsson *et al.* 2010), so as to compare the sequence identity of only the variable ITS regions. Extracted sequences were aligned separately for each added species using MAFFT 6.703b (Katoh *et al.* 2009) and pairwise distances and distance trees were computed in PAUP 4.0d (Swofford 2001) using uncorrected DNA/RNA distance settings.

Comparing quality scores among species

To compare quality scores among species, we constructed a 'dirty' data set in which reads were processed by minimum length and homology to primer/multiplex code, but not by Phred scores. For each 454 read matching one of the six added species, the average quality scores across all of the bases was averaged into a read average. The read averages were compiled and imported into the R programming environment. Significant differences between the mean read-quality scores for each species were calculated via analysis of variance (ANOVA) and Tukey–Kramer post hoc significance tests.

Fungal community analyses

To compare potential phylogenetic biases of read orientation, we compiled all reads derived from the dust sample that did not match any of the six added species. These background fungal community sequences were compared to a database of all identified fungal sequences in GenBank using BLASTN, determined to 'last common ancestor' (LCA) using the program MEGAN 3.6 (Huson *et al.* 2007; LCA parameters: minimum support: 1, minimum score: 300, score/length ratio: 1.97, Top per cent: 10) and cross-referenced to order-level taxonomy with the Dictionary of Fungi (Kirk *et al.* 2008) as an authority in an SQL database. The tree and taxon heat map (Fig. 4) were constructed with the web-based Interactive Tree of Life program (Letunic & Bork 2007) according to the NCBI taxonomy scheme. The tree was input into the online version of UniFrac (Lozupone & Knight 2005) for global and pairwise 'P-test' analysis of presence/absence community phylogenetic composition as in (Martin 2002). Briefly, this method compares the number of changes required to explain covariance between community membership and phylogeny using randomized tree topologies as a null distribution.

The ITS region, the typical barcode target for species-level taxonomy in Fungi (Ryberg *et al.* 2009; Seifert 2009), is not amenable to unambiguous multiple-sequence alignments, even within some genera, due to the high rate of evolution and high frequency of indels (Ryberg *et al.* 2008). For this reason, our sequence processing is based on pairwise comparisons. Reads from each added species were pooled by sequencing orientation, and OTUs were calculated by comparing pairwise sequence alignment identities in the program CD-HIT-EST

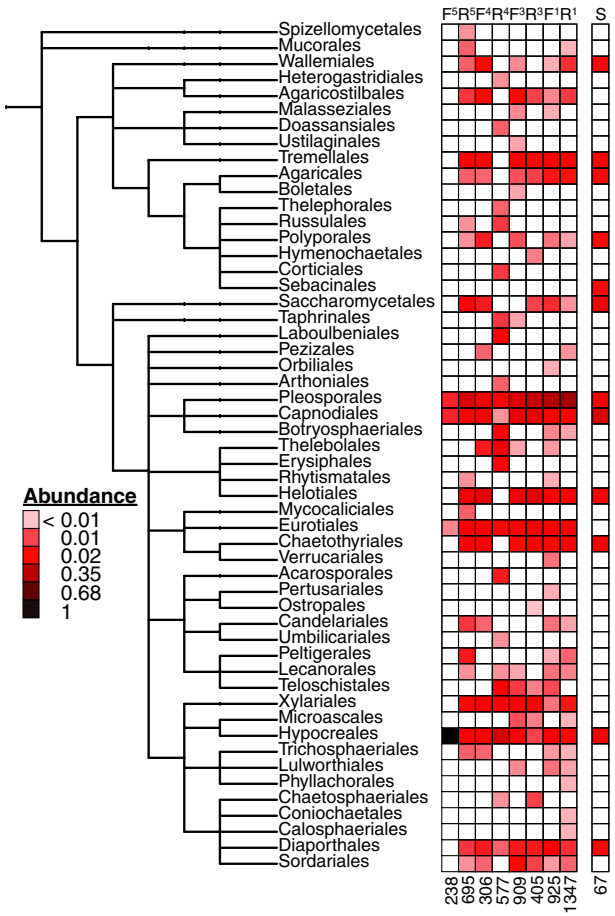

Fig. 4 Enumeration of fungal community composition from dust sample did not differ based on read orientation, and was concordant with community composition derived from cloning and Sanger sequencing. F, forward sequencing orientation; R, reverse sequencing orientation; S, Sanger sequences. Superscript numbers denote the number of spores added to the pool (i.e. F^5 is the pool amplified with the tagged primers used for the experiment with 10^5 added spores and sequenced in the forward orientation). Pools sharing a superscript number were derived from the same dust aliquot DNA extraction. Only environmental sequences are included here, those derived from added spores were removed from this analysis. The tree is based on NCBI taxonomic scheme. Proportional abundances are rounded to the nearest colour-coded values, and do not necessarily sum to 100% within samples. Heatmap scale is skewed towards smaller values with 2% set as the median value. Neither the pooled Sanger, forward- or reverse-sequenced communities were significantly differentiated according to 'P-test' analysis. Sample size is denoted below.

3.1.2 (Li & Godzik 2006) using the 'slow and thorough' setting. Pairwise comparisons in which at least 80% of the shorter read aligned at ≥97% identity to the longer read were clustered into OTUs. Rarefaction of OTUs (Fig. 1b) was computed using the program MOTHUR v. 1.7.2 (Schloss *et al.* 2009) with 1000 iterations.

Results

Quantification of added spores

Quality score-trimmed read abundance among added species varied by approximately an order of magnitude between the most abundant species: (*Thelephora terrestris*, $N = 1051$) and the least abundant: (*Laccaria proxima*, $N = 38$; Fig. 2). Neither spore size ($r = 0.46$) nor ITS fragment length ($r = 0.01$) significantly correlated with read abundance among the six species tested (two-tailed Pearson's, d.f. = 5, $P \geq 0.1$).

The strength of the relationship between read abundance and spore abundance (Fig. 3) varied among species and between sequence orientations. The strongest relationship was for *Th. terrestris* ($R^2 = 0.96$), and the weakest was for *Rhizopogon salebrosus* ($R^2 = 0.54$). Some of the variance in the less-encountered species may be attributed to a relatively few number of reads.

A high incidence of chimeric *Metschnikowia noctiluminum* reads were detected in the reverse sequencing orientation (Fig. 3). Sequence orientation resulted in significantly different read abundance within *M. noctiluminum* between forward and reverse sequencing orientations whether chimeric sequences were excluded (two sample homoscedastic *T*-test, one-tailed, $P \leq 0.03$) or included ($P = 0.05$; neither test was significant at $\alpha = 0.05$ with sequential Bonferroni correction). No other species showed significant bias based on read orientation. Within the 'dirty' data set where reads were not culled due to low quality scores, a nearly 15-fold increase of *M. noctiluminum* reads were identified, and there was no significant difference between read abundance based on sequencing orientation. Relaxation of read-quality criteria resulted in 17% more reads matching the *Rh. salebrosus* reference sequence.

Fungal community analysis

Background environmental reads were assigned to 53 orders in total (Fig. 4). Pools with 10^6 added spores per species were excluded because very few reads were from environmental sequences. Rarefaction analysis of environmental fungal sequences showed near identical numbers of OTUs derived from the two sequence orientations (Fig. 1b). Results of the P-test showed no significant differences among order-level phylogenetic community structure of forward, reverse, and clone/Sanger-sequenced communities. Because forward- and reverse-oriented reads were often only partially overlapping, direct analyses of shared OTUs between these data sets were not possible.

Sequence variance

For added species, with the exception of *M. noctiliuminum*, mean pairwise sequence alignment identity between extracted ITS regions was >99% in both sequencing orientations (Table 1). Dendrograms of the ITS (including the conserved 18s, 5.8s and 28s regions as is standard) derived from reads of added species demonstrate the autapomorphic nature of sequence variation (Fig. 5), in which relatively few deviations from a consensus sequence were shared.

Sequencing orientation affected OTU clustering for our added species. For forward-oriented sequences, the correct number of six species-specific OTU clusters

Table 1 Mean pairwise per cent identity of aligned reads from each of the added species

	Laccaria proxima	*Metschnikowia noctiluminum*	*Rhizopogon salebrosus*	*Russula sanguinea*	*Thelephora terrestris*	*Tricholoma imbricatum*
R	0.997 ± 0.007	0.992 ± 0.016	0.998 ± 0.003	0.996 ± 0.006	0.993 ± 0.009	0.998 ± 0.004
F	0.999 ± 0.002	0.981 ± 0.015	0.997 ± 0.003	0.998 ± 0.004	0.997 ± 0.003	0.999 ± 0.002

R, reverse orientation (ITS2); F, forward orientation (ITS1). Error is SD. Number of reads compared in each alignment is displayed in Fig. 2.

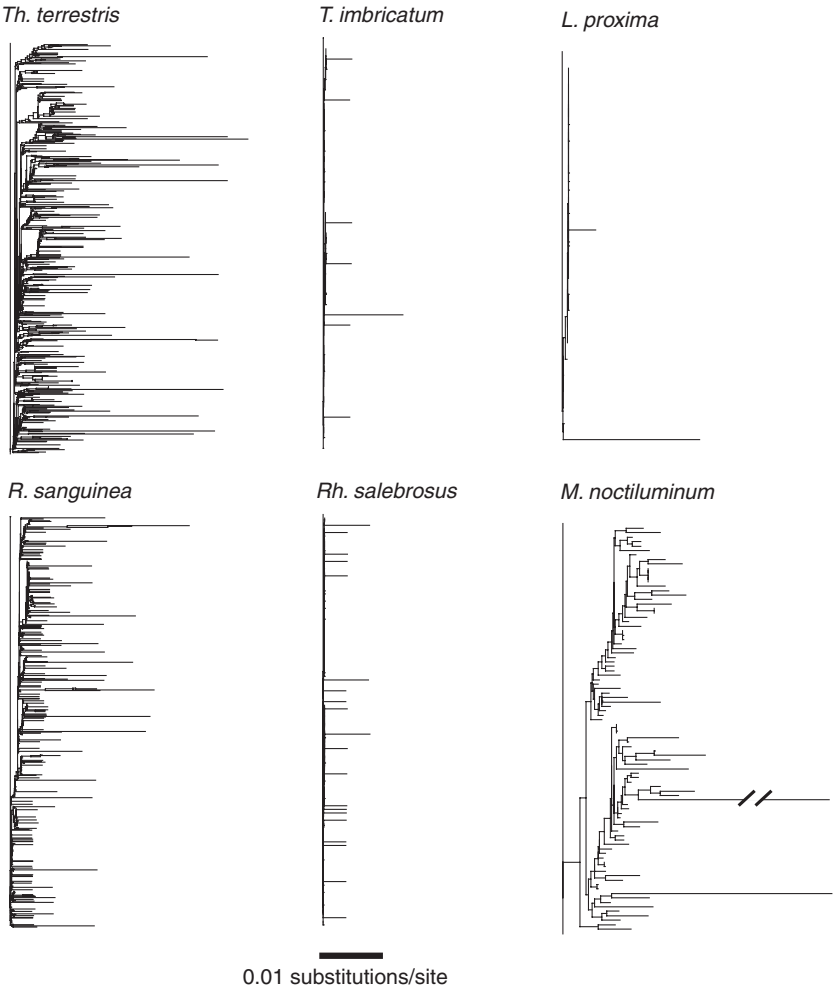

Fig. 5 Distance trees of reads from the extracted ITS2 regions of each of the six added species show largely autapomorphic variance probably generated by a combination of sequence variation among gene copies and sequencing error. Scale bar denotes 0.01 substitutions per site. Conserved regions of the rDNA operon flanking this variable region contribute to alignment identity calculations for OTU clustering.

were recovered at 95% identity, whereas 9, 10, 19 and 74 OTUs were recovered at 96%, 97%, 98% and 99% respectively. In the reverse-oriented sequences, the correct number of OTUs were determined at sequence identity ≤97%, whereas 15 and 63 OTUs were recovered at 98% and 99% respectively. The discrepancy between OTUs based on sequencing orientation is concordant with the fact that ITS1 is generally more variable across kingdom Fungi than the ITS2 spacer (Nilsson *et al.* 2008).

Metschnikowia noctiluminum was more variable in ITS1 compared to ITS2. Much of this variance is attributable to varying numbers of 'A's called in two 9-bp homopolymer runs present in the ITS1. Additionally, two single nucleotide polymorphisms: a C–T transition variant present in 32 reads (9.6%) and an A–T transversion variant present in 23 reads (6.9%) contributed to this variance. We presume these latter two variants are true biological ribotypes and not merely PCR or sequencing artefacts. Rare sequences containing 4–6 bp insertions were detected in other species. Such variants are not consistent with typical errors associated with polymerase or 454 sequencing and may be pseudo-genes or examples of failed concerted evolution.

We found significant differences between mean quality scores of the six species using the 'dirty' data set (Fig. 6), with the lowest scores attributed to sequences of *M. noctiluminum*. Within species, significant differences were found between forward and reverse sequences of *Th. terrestris* and *M. noctiluminum* (not shown).

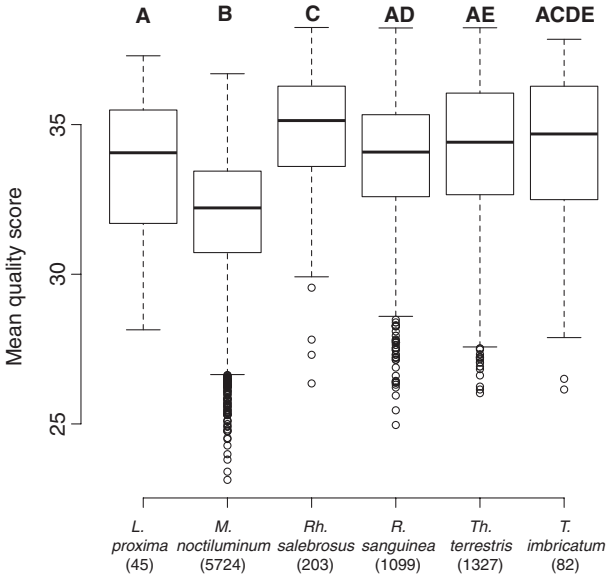

Fig. 6 Box and whisker plots of mean quality scores from reads from each of the six added species. Reads are from the non quality-score-culled 'dirty' data set. The means of species that do not share a letter (A–E above) were determined to differ significantly (corrected $P \leq 0.01$) as calculated using the post hoc Tukey–Kramer test on an ANOVA. Boxes contain the upper and lower quartiles, and the median is displayed as the band within. Outliers (open circles) are more than 1.5 times the interquartile range, and whiskers display the extent of nonoutlier values. Number of reads is noted below.

Discussion

Despite the increasing use of pyrosequencing read abundance to calculate dissimilarity metrics among environmental samples of microbial communities, we know of no studies in which this technique has been validated. Although it is difficult to generalize based on only six fungal species, our results show that sequence abundance can vary by an order of magnitude among species that were known to be present in equal quantity of spores. We made no attempt to normalize for gene copy number variation among species; however, nuclear conditions tend to be consistent within genera and would predict that all of the added spores contain a single nucleus with the exception of *Tricholoma imbricatum* which possibly contains two (Horton 2006).

With the possible exception of aerobiological systems, the fungal composition of most substrates will probably contain a high percentage of filamentous hyphae compared to spores. Thus, quantitative biases originating from DNA extraction efficiency in particular may be exaggerated in this study. Other variables, such as varying proportions of live vs. dead cells may have contributed to differences in read abundance. Despite this, there were no significant correlations between read abundance and spore size or ITS fragment length. Although not explicitly measured, cell wall thickness did not appear to be a significant correlate either.

Surprisingly, few studies have examined rDNA copy number across kingdom Fungi (but see Rooney & Ward 2005), and we do not know the number of copies contained in the genomes of the individuals examined here. However, a previous study showed fourfold copy number variations among yeast species (Maleszka & Clarkwalker 1993). Several studies have shown that rDNA gene copy numbers can vary within species. For example, Liti *et al.* (2009) estimate rDNA copy numbers among strains of *Saccharomyces cerivisieae* range from 54 to 511 copies. More striking, meiotic segregants of individuals frequently demonstrate different rDNA gene copy numbers (see lengthy discussion of mechanisms and references in Zolan 1995). Mitotic instability of fungal rDNA copy numbers is also frequently reported, and has been shown to correlate with shifts in physiology or growth environment (Russell & Rodland 1986; Pukkila & Skrzynia 1993; Zolan 1995; Herrera *et al.* 2009). Copy numbers of rDNA in Bacteria may be may

more stable within species due to the lack of meiotic recombination and the relatively few repeats found in those genomes.

Semi-quantification of spores derived from a single individual varied in our study, and was much less precise compared with a study (Manter & Vivanco 2007) in which the same primer pair (ITS1f and ITS4) was used in a quantitative PCR (SYBR fluorescence) analysis of known quantities of mixed and single fungal templates along a dilution gradient. The authors demonstrate near-perfect within-species regressions based on template abundance, but poor across-species quantification. Quantitative PCR, unlike pyrosequencing, enables absolute quantification of a taxon based on a standard curve. It should be noted that Manter and Vivanco's study, as well as our own, tests semi-quantification among relatively homogeneous substrates and background environmental communities, and we urge caution when attempting to compare the abundance of taxa among different environmental samples. When using universal primers, the relative read abundance of individual species will always be affected by other species in the pool. As reaction reagents (such as PCR primer, or 454 DNA capture beads) will be limiting in studies of high diversity environments, taxa with high template abundance or PCR affinity will diminish the read abundance of other taxa.

A final consideration when using read abundance as a quantitative metric is that biases result from innate nucleotide sequence structure. In our study, chimeras were only detected in the reverse sequences of *Metschnikowia noctiluminum*. As with the other added fungal species, both priming sites in *M. noctiluminum* are perfectly complimentary to the ITS1f and ITS4 primers, and there do not appear to be secondary binding sites in or near the locus. We hypothesize that the existence of the two 9-bp homopolymer runs in the ITS1 region may potentially cause polymerase slippage and incomplete extension, although it is unclear why this appears to be limited to a single orientation.

The two homopolymer runs may also explain why there were fewer sequences in the forward orientation. 454 sequencing technology theoretically incorporates all bases in a homopolymer during a single flow cycle, and the intensity of light generated by the luciferase reaction determines how many bases are called. Because longer homopolymers are more prone to mis-calls, each successive base in a homopolymer run will receive a lower quality score (Brockman *et al.* 2008; Quinlan *et al.* 2008). For this reason, reads from taxa with long homopolymers will be statistically more likely to contain ambiguous base calls or sections with low quality scores. If read-quality criteria are used to screen errant reads, taxa with homopolymers will be disproportionately culled from subsequent analyses as occurred in our study. We found significant differences among mean quality scores of only six taxa, indicating that many more such biases probably exist.

Read orientation resulted in significantly different abundances for *M. noctiluminun* (and to a lesser extent that of *Rhizopogon salebrosus*). These differences were mitigated when quality score stringency was relaxed, and low quality sequences were tolerated. Relaxing quality scores, however, results in greater sequence variance and an increased probability of inflated diversity estimates (Kunin *et al.* 2010). The inclusivity of OTU thresholds will have obvious impacts on quantification and OTU frequency distributions. 454 Life Science base-calling software is rapidly evolving, and the FLX Titanium Amplicon platform, as well as alternative software (Quince *et al.* 2009) all promise future improvements, if not solutions, to this problem.

This study lacks technical replicates among separate 454 runs, so it is not possible to deduce whether the biases we encounter here are representative of pyrosequencing in general. A recent review (Prosser 2010) poignantly describes the pitfalls of inferences based on idiosyncratic unreplicated samples. In the case that replication is not feasible, researchers may consider adding positive controls to their studies to evaluate the quantitative and compositional biases their data and analyses may contain. This study, for example, was conducted in conjunction with another study (Amend *et al.* 2010) examining global biogeography of indoor fungi. By evaluating OTU clustering thresholds, quantification precision and quality score distributions of a known subset of our data, we were able to determine which types of analyses would be appropriate in a specific system of interest, as well as estimates of error. As a result, we chose not to use quantitative measures of ß-diversity and used phylogenetic-, rather than OTU-based multivariate comparisons.

Conclusion

Pyrosequencing technologies have revolutionized our understanding of complex microbial communities. Delineation of some of the most taxonomically diverse communities yet discovered is becoming increasingly possible, as are novel hypotheses about the processes that govern and shape them. Although the ability to distinguish communities based on the frequency distribution of its membership offers an appealing insight into community similarity, we show that both technological artefacts and innate biological traits bias relative quantification of biological abundance by 454 read counts. Our results show that analyses based on semi-quantitative data may be warranted, although careful

validation of read abundance should be undertaken within the study system of interest. Although the possibility exists that relative quantification of microbial rDNA gene copies is a more tractable exercise than quantification of cell abundance or biomass, copy number instability within and among fungal species may limit the utility and relevance of such comparisons.

Acknowledgements

We are grateful to Henrik Nilsson and three anonymous referees for comments on earlier versions of this manuscript and to John Taylor for useful input into our discussion of rDNA copy number variation. This work was supported by a grant to the Indoor Mycobiota Barcode of Life initiative from the Alfred P. Sloan Foundation.

References

Altschul S, Madden T, Schaffer A *et al.* (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389.

Amend A, Samson R, Seifert K, Bruns T (2010) Deep sequencing reveals diverse and geographically structured assemblages of fungi in indoor dust. *Proceedings of the National Academy of Sciences, USA*, **107**, 13748–13753.

Arnold A, Maynard Z, Gilbert G (2002) Fungal endophytes in dicotyledonous neotropical trees: patterns of abundance and diversity. *Mycological Research*, **105**, 1502–1507.

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**, 326–349.

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, **18**, 763–770.

Chao A, Chazdon RL, Colwell RK, Shen TJ (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.

Costello EK, Lauber CL, Hamady M *et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.

DeSantis T, Stone C, Murray S, Moberg J, Andersen G (2005) Rapid quantification and taxonomic classification of environmental DNA from both Prokaryotic and Eukaryotic origins using a microarray. *FEMS Microbiology Letters*, **245**, 271–278.

Engelbrektson A, Kunin V, Wrighton K *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal*, **4**, 642–647.

Feinstein LM, Sul WJ, Blackwood CB (2009) Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Applied and Environmental Microbiology*, **75**, 5428–5433.

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

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.

Hori M, Fukano H, Suzuki Y (2007) Uniform amplification of multiple DNAs by emulsion PCR. *Biochemical and Biophysical Research Communications*, **352**, 323–328.

Horton TR (2006) The number of nuclei in basidiospores of 63 species of ectomycorrhizal Homobasidiomycetes. *Mycologia*, **98**, 233–238.

Hugenholtz P, Pace NR (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology*, **14**, 190–197.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Huson DH, Auch AF, Qi J, Schuster SC (2007) Megan analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Jumpponen A (2007) Soil fungal communities underneath willow canopies on a primary successional glacier forefront: RDNA sequence results can be affected by primer selection and chimeric data. *Microbial Ecology*, **53**, 233–246.

Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology*, **537**, 39–64.

Kirk P, Cannon P, Minter D, Stalpers J (2008) *Dictionary of the Fungi*, 10th edn. CAB International, Oxon, UK.

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

Lee ZMP, Bussema C, Schmidt TM (2009) Rrndb: documenting the number of rRNA and tRNA genes in Bacteria and Archaea. *Nucleic Acids Research*, **37**, D489–D493.

Letunic I, Bork P (2007) Interactive tree of life (ITOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127.

Li W, Godzik A (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li J, Wang LL, Mamon H *et al.* (2008) Replacing PCR with cold-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nature Medicine*, **14**, 579–584.

Liti G, Carter DM, Moses AM *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.

Lozupone C, Knight R (2005) Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.

Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, **73**, 1576–1585.

Maleszka R, Clarkwalker GD (1993) Yeasts have a 4-fold variation in ribosomal DNA copy number. *Yeast*, **9**, 53–58.

Manter DK, Vivanco JM (2007) Use of the ITS primers, ITS1f and ITS4, to characterize fungal abundance and diversity in mixed-template samples by qPCR and length heterogeneity analysis. *Journal of Microbiological Methods*, **71**, 7–14.

Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology*, **68**, 3673–3682.

Morales SE, Cosart TF, Johnson JV, Holben WE (2009) Extensive phylogenetic analysis of a soil Bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. *Applied and Environmental Microbiology*, **75**, 668–675.

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH (2008) Intraspecific its variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics*, **4**, 193–201.

Nilsson RH, Veldre V, Hartmann M *et al.* (2010) An open source software package for rapid, automated extraction of ITS1 and ITS2 from fungal its sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology*, **3**, 284–287.

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

Prosser JI (2010) Replicate or lie. *Environmental Microbiology*, **12**, 1806–1810.

Pukkila PJ, Skrzynia C (1993) Frequent changes in the number of reiterated ribosomal-rna genes throughout the life-cycle of the Basidiomycete *Coprinus cinereus*. *Genetics*, **133**, 203–211.

Quince C, Lanzen A, Curtis TP *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, **6**, 639–641.

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **5**, 179–181.

Reeder J, Knight R (2009) The 'rare biosphere': a reality check. *Nature Methods*, **6**, 636–637.

Rooney AP, Ward TJ (2005) Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *Proceedings of the National Academy of Sciences, USA*, **102**, 5084–5089.

Rozera G, Abbate I, Bruselles A *et al.* (2009) Massively parallel pyrosequencing highlights minority variants in the hiv-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology*, **6**, 15.

Russell PJ, Rodland KD (1986) Magnification of ribosomal-RNA gene number in a *Neurospora crassa* strain with a partial deletion of the nucleolus organizer. *Chromosoma*, **93**, 337–340.

Ryberg M, Nilsson RH, Kristiansson E *et al.* (2008) Mining metadata from unidentified its sequences in genbank: a case study in *Inocybe* (Basidiomycota). *BMC Evolutionary Biology*, **8**, 50–64.

Ryberg M, Kristiansson E, Sjokvist E, Nilsson RH (2009) An outlook on the fungal internal transcribed spacer sequences in genbank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytologist*, **181**, 471–477.

Schloss PD (2008) Evaluating different approaches that test whether microbial communities have the same structure. *The ISME Journal*, **2**, 265–275.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing MOTHUR: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Seifert KA (2009) Progress towards DNA barcoding of fungi. *Molecular Ecology Resources*, **9**, 83–89.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored ''Rare biosphere''. *Proceedings of the National Academy of Sciences, USA*, **103**, 12115–12120.

Stahl DA, Lane DJ, Olsen GJ, Pace NR (1984) Analysis of hydrothermal vent-associated symbionts by ribosomal-RNA sequences. *Science*, **224**, 409–411.

Swofford D (2001) *Paup* 4.0 beta 5: Phylogenetic analysis using parsimony (and other methods)*. Sinauer Associates, Sunderland, MA.

Taylor DL, Herriott IC, Long J, O'Neill K (2007) TOPO ta is a-ok: a test of phylogenetic bias in fungal environmental clone library construction. *Environmental Microbiology*, **9**, 1329–1334.

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

Williams R, Peisajovich SG, Miller OJ *et al.* (2006) Amplification of complex gene libraries by emulsion PCR. *Nature Methods*, **3**, 545–550.

Zolan ME (1995) Chromosome-length polymorphism in Fungi. *Microbiological Reviews*, **59**, 686–698.

A. A. is a post-doctoral researcher who is broadly interested in inferring patterns in fungal and microbial biodiversity and the processes that shape it. K. S. is a fungal taxonomist specializing in the identification and phylogenetic relationships of anamorphic fungi, in particular those producing mycotoxins. T. B. studies the ecology of fungi and is especially interested in how ectomycorrhizal species interact with butter and garlic.

Supporting information

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

Fig. S1 Experimental scheme.

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 corresponding author for the article.