MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 275–285 doi: 10.1111/j.1365-294X.2010.04948.x

Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology

MARTIN UNTERSEHER,* ARI JUMPPONEN,† MAARJA ÖPIK,‡ LEHO TEDERSOO,‡ MARI MOORA,‡ CARSTEN F. DORMANN§ and MARTIN SCHNITTLER*

**University Greifswald, Institute of Botany and Landscape Ecology, Grimmer Str. 88, 17487 Greifswald, Germany, †Kansas State University, Ecological Genomics Institute, 433 Ackert Hall, Manhattan, KS 66506, USA, ‡Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., Tartu 51005, Estonia, §Department Computational Landscape Ecology, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318 Leipzig, Germany*

Abstract

Results of diversity and community ecology studies strongly depend on sampling depth. Completely surveyed communities follow log-normal distribution, whereas power law functions best describe incompletely censused communities. It is arguable whether the statistics behind those theories can be applied to voluminous next generation sequencing data in microbiology by treating individual DNA sequences as counts of molecular taxonomic units (MOTUs). This study addresses the suitability of species abundance models in three groups of plant-associated fungal communities – phyllosphere, ectomycorrhizal and arbuscular mycorrhizal fungi. We tested the impact of differential treatment of molecular singletons on observed and estimated species richness and species abundance distribution models. The arbuscular mycorrhizal community of 48 MOTUs was exhaustively sampled and followed log-normal distribution. The ectomycorrhizal (153 MOTUs) and phyllosphere (327 MOTUs) communities significantly differed from log-normal distribution. The fungal phyllosphere community in particular was clearly undersampled. This undersampling bias resulted in strong sensitivity to the exclusion of molecular singletons and other rare MOTUs that may represent technical artefacts. The analysis of abundant (core) and rare (satellite) MOTUs clearly identified two species abundance distributions in the phyllosphere data – a log-normal model for the core group and a log-series model for the satellite group. The prominent log-series distribution of satellite phyllosphere fungi highlighted the ecological significance of an infrequent fungal component in the phyllosphere community.

Keywords: ectomycorhizal fungi, fungal diversity, Glomeromycota, high-throughput parallel sequencing, microbial ecology, phyllosphere fungi

Received 25 May 2010; revision received 29 October 2010; revision accepted 4 November 2010

Introduction

Biodiversity is a principal ecological concept of habitat quality that represents both intraspecific variation (genetic diversity) and community variation (richness, abundance and evenness of species) at different spatial scales (Whittaker 1977; Bisby *et al*. 1995; Jost 2007). All these measures rely on species concepts that are relatively straightforward in macroscopic organisms such as plants and animals, but are strongly debated in microbial taxa. At small spatial scales, fungi and other microbes generally exceed the diversity of macroscopic organisms by orders of magnitude, but data sets are mostly small and incomplete due to laborious isolation, cultivation and identification procedures. Furthermore, a large proportion of the fungal taxa is not culturable and prevailing mitotic reproduction renders definitions of

Correspondence: Martin Unterseher, Fax: +49 3834 864114; E-mail: martin.unterseher@uni-greifswald.de

individuals and species arguable (Taylor *et al.* 2000). With respect to exhaustiveness of species inventories, detection of microbes from the environment has greatly benefited from the next generation sequencing (NGS) approach that however also has shed further light on the poverty of applicable taxonomic concepts (Sogin *et al.* 2006; Keijser *et al.* 2008; Buée *et al.* 2009; Stoeck *et al.* 2009). Nevertheless, these NGS studies have linked genetic diversity directly to α-diversity (species richness) by statistically separating molecular operational taxonomic units (MOTUs; proxies for species) from the genetic diversity data. The measure of abundance is more difficult to use in microbial ecology because of underlying PCR bias (Bellemain *et al.* 2010) and poor correlation of the amount of genes to biomass or cell frequency (von Wintzingerode *et al.* 1997). Nevertheless, sequence reads per MOTU are currently used as distinct counts (comparable to individuals of macroscopic organisms) thus forming the basis for diversity statistics such as observed and predicted species richness, species abundance distributions (SADs), etc. (Buée *et al.* 2009; Jumpponen & Jones 2009; Öpik *et al.* 2009; Sun *et al.* 2009; Dumbrell *et al.* 2010; Unterseher & Schnittler 2010).

Species richness is the basic measure of biodiversity at any spatial scale. Based on abundance information and distribution of each species, species-accumulation curves (SACs; Colwell & Coddington 1994; Gotelli & Colwell 2001; Ugland *et al.* 2003), species-abundance distributions (SADs; Fisher *et al.* 1943; Preston 1948; MacArthur 1957; May 1975; Tokeshi 1990; Hubbell 2001; Williamson & Gaston 2005; McGill *et al.* 2007; McPherson & Jetz 2007; Coddington *et al.* 2009) and diversity indices are usually calculated to compare species richness among communities or treatments. Many SAD models have been developed to understand the statistical structure of biological communities and to be able to predict unsampled parts of the communities. For example, geometric series (May 1975) predict extremely uneven abundances of organisms; broken-stick distributions (MacArthur 1957) represent extremely even abundances; and log-normal (Preston 1948) and log-series (Fisher *et al.* 1943) models predict very low and very high proportions of rare species, respectively. Despite the general interest of ecologists in SADs, lesser importance has been attached to the discrimination of exhaustiveness and insufficiency of sampling. A recent meta-analysis of plant and animal communities revealed clear impacts of sampling intensity on the observed SAD (Ulrich *et al.* 2010). Complete surveys typically follow log-normal types of SADs, whereas incompletely sampled communities significantly deviate from log-normality, irrespective of spatio-temporal scales, geographic positions and species richness.

Here, three voluminous plant-related fungal NGS data sets of the phyllosphere (Jumpponen & Jones 2010) and rhizosphere (Öpik *et al.* 2009; Tedersoo *et al.* 2010) were tested against three recently established hypotheses in community ecology. (i) The log-normal SAD is a general model for completely sampled communities whereas a power-law type of SADs best describes undersampled communities. (ii) Exclusion of the rarest MOTUs (molecular singletons, doubletons, tripletons, quadrupletons and quintupletons are given the rank of biological singletons) distinctly influences observed and extrapolated species richness. (iii) Partitioning the fungal data sets into core (abundant) and satellite (rare) MOTUs unravel overlapping SADs that otherwise complicate the interpretation of observed SADs. Finally we evaluated the performance of different richness estimators with an emphasis on Ugland's function (Ugland *et al.* 2003) and Chao 2 that were considered more accurate than others in predicting total fungal species richness (Unterseher *et al.* 2008).

Material and methods

Underlying sequence data

The three pyrosequencing data sets analysed here differ in habitat type, species richness and ecology of the target organisms. The present data set of arbuscular mycorrhizal fungi (AMF) comprises sequences from Glomeromycota that colonized 10 understorey plant species in a boreonemoral forest in Estonia (Öpik *et al.* 2009). The data set includes 11 samples with 48 MOTUs from 126 654 sequences of the rRNA small subunit (SSU) gene. The ectomycorrhizal fungal (EcMF) data set comprises sequences of ectomycorrhizal Basidio- and Ascomycota from root tips that were collected in Korup National Park, Cameroon (Tedersoo *et al.* 2010). This data set of 41 587 sequences of rRNA internal transcribed spacer (ITS) region contains 153 MOTUs from 24 different samples (in that case approaches of DNA extraction and PCR). The phyllosphere fungal (PpF) data set represents 12 samples of foliar endophytes, saprobes and parasites of *Quercus macrocarpa* in Kansas, USA (Jumpponen & Jones 2010). The data set comprises 327 fungal MOTUs from 16 541 sequences of the ITS region. The exclusion of a large part of the original PpF data was necessary because both mycorrhizal data relied on one sampling event, whereas the original PpF data were gathered at six different sampling events during an entire vegetation period (Jumpponen & Jones 2010). For the present analysis only the last sampling event was considered. Sampling, molecular analyses, bioinformatics and taxon frequency are described in the original publications (Öpik *et al.* 2009; Jumpponen & Jones 2010; Tedersoo *et al.* 2010).

Statistics of species richness

Molecular singletons were removed from the original data sets to minimize the inclusion of sequencing artefacts (Kunin *et al.* 2010; Tedersoo *et al.* 2010). Because the absence of singletons violates some fundamental assumptions of species richness analysis, one 454 sequence per MOTU was removed randomly, causing all MOTUs with originally two reads (molecular doubletons) to become the rarest taxa. Thus, doubletons in the data sets were considered biological singletons. The effect of an even more conservative interpretation of 454 reads on species richness analysis was evaluated by consecutively omitting molecular doubletons, tripletons and quadrupletons from the data sets. The five data sets – (i) singletons retained, (ii) singletons excluded, (iii) doubletons excluded, (iv) tripletons excluded and (v) quadrupletons excluded – were used for comparisons of SAC and the richness estimator Chao 2.

EstimateS version 8.2 (Colwell 2009) was used with default settings to calculate the rarefied SACs and four non-parametric minimum richness estimators (ACE, Chao 2, Jackknife 2 and Bootstrap; Colwell & Coddington 1994). The relative performance of non-parametric richness estimators were compared to the method of Ugland *et al.* (2003).

To separate fungal taxa into core (abundant) and satellite (occasional or rare) MOTUs, persistence-abundance plots (*sensu* Magurran & Henderson 2003) were drawn: the number of samples in which each taxon was observed was plotted against the maximum abundance of each taxon in any one sample. In the case of the fish community described in Magurran & Henderson (2003), a discontinuity of plotted species was visible at approx. 50% persistence. Those species which occurred at less than 50% were termed satellite (occasional) species. On an entirely molecular basis, Pedrós-Alió (2006) and Galand *et al.* (2009) separated abundant from rare MOTUs according to different criteria (i.e. probability of successful amplification) without giving concrete thresholds.

Our choice of splitting the molecular assemblages was different from that of Pedrós-Alió (2006) and Galand *et al.* (2009) who relied on abundances (sequences per MOTU) rather than on persistence (samples per MOTU). Here, we followed the approach described in Magurran & Henderson (2003) and used a persistence threshold of ≥50% for the core MOTUs.

The R Package (R Development Core Team 2010) was used for the basic species abundance distribution models, goodness of fit tests (chi–squared, Shapiro–Wilk, Kolmogoroff–Smirnoff & Anderson–Darling) and richness estimators that assume a truncated log-normal distribution of MOTU abundances. The R source file and the data files are given as Supporting information.

Results

Species (MOTU) abundance distribution

The rank abundance plot (Whittaker plot) of the PpF displayed a power law-like SAD (Fig. 1a) that deviated significantly from log-normal (Fig. 1d, Table 1). The EcMF assemblage also differed significantly from log-normal (Fig. 1b,e, Table 1). By contrast, the AMF community fitted significantly to a log-normal type of SAD (Fig. 1c,f, Table 1).

The core and satellite approach for the PpF suggested that 38 MOTUs belong to the core taxa (Fig. 2a). The core taxa accounted for 12% of the MOTU richness and 73% of all sequences. The ratio of core to satellite MOTUs was 1:7.6. Analysis of the respective SADs dissected the power-law SAD of the entire data set (Fig. 1a) into two different distributions. The core phyllosphere fungi approached a log-normal SAD (Fig. 2d) that was statistically supported by goodness-of-fit tests (Table 1). By contrast, the satellite group significantly followed a log-series SAD (Fig. 2g; two-sample K–S test: $D = 0.26$, $P = 0.07$).

The EcMF core taxa accounted for 46% of the MOTU richness (Fig. 2b) and for 96% of all sequences. The ratio of core to satellite MOTUs was 1:4. Different SADs were also observed for the EcMF core and satellite taxa. Goodness of fit tests for log-normality of the EcMF core group were significant (Table 1), whereas visual and statistical fit of the satellite group to the predicted log-series was not significant (Fig. 2h; two-sample K–S test: $D = 0.5$, $P < 0.05$). The AMF core taxa accounted for 40% of the total observed MOTU richness (Fig. 2c) and comprised 95% of all sequences. The ratio of core to satellite MOTUs decreased further to 1:1.5. Goodness of fit tests did not reject log-normality of the AMF core group, however visual inspection of the SADs indicated poor fit to log-normality of the core group (Fig. 2f). The satellite group did not follow a logseries SAD (Fig. 2i; two-sample K–S test: $D = 0.84$, $P < 0.05$).

Observed and estimated MOTU richness

The analysis of observed MOTU richness displayed a steeply rising rarefied SAC in the case of PpF, a clear convergence to saturation for EcMF and a long and stable plateau for AMF (Fig. 3).

Species richness estimators differed in their efficiency to predict total MOTU richness (Fig. 4, Table 2). None of the four richness estimators of PpF (Fig. 4a) reached plateau at 100% data coverage and the functions provided substantially different estimates. Chao 2 and Jackknife2 estimated the largest number of unseen PpF

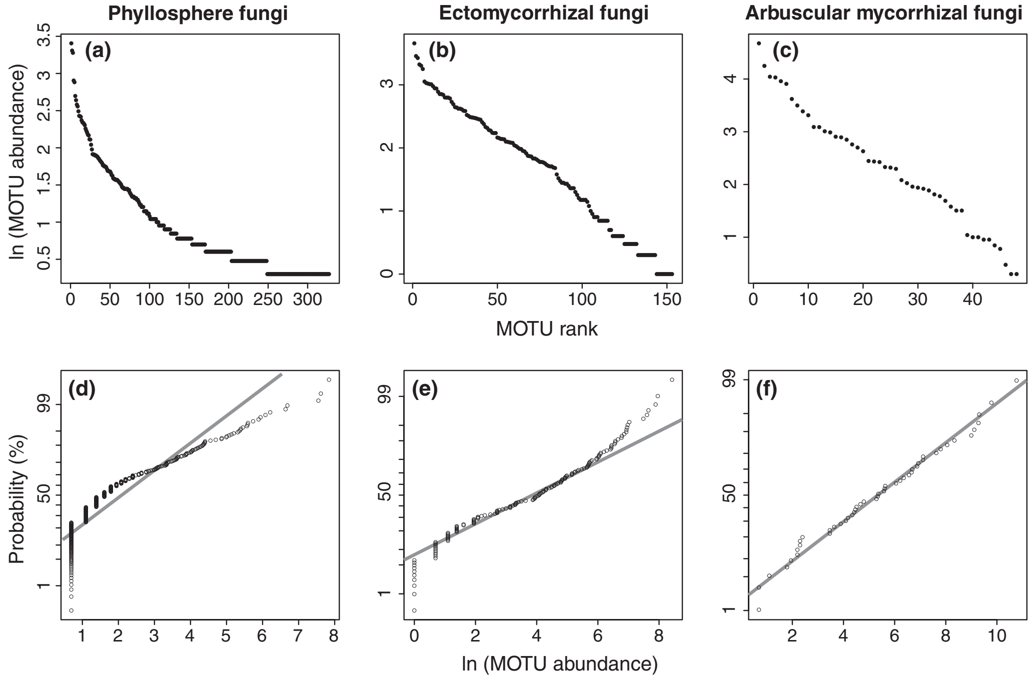

Fig. 1 Rank abundance plots (a–c) illustrate the relationships between rare and abundant MOTUs. The probability plots (d–f) visually test for log-normality of the data. The better the grey line in a probability plot fits to the plotted abundances (open circles) the better is the fit to a log-normal species abundance distribution.

Table 1 Tests for fit to log-normal species abundance distributions of MOTU-sequences data sets, giving the number of MOTUs, and test statistics for the chi-square, Kolmogorov–Smirnov, Shapiro–Wilk and Anderson–Darling test

		Goodness-of-fit tests for log-normality			
Data set	No. MOTUs	Chi-square	Kolmogorov–Smirnov	Shapiro–Wilk	Anderson–Darling
PpF					
All taxa	327	$P = 173.6$, $P < 0.001$	$D = 0.18$, $P < 0.001$	$W = 0.85$, $P < 0.001$	$A = 15.8$, $P < 0.001$
Core group	38	$P = 10.8$, $P = 0.05$	$D = 0.13$, $P = 0.55$	$W = 0.96$, $P = 0.24$	$A = 0.48$, $P = 0.22$
Satellite group	289	$P = 167.1$, $P < 0.001$	$D = 0.18$, $P < 0.001$	$W = 0.85$, $P < 0.001$	$A = 13.7$, $P < 0.001$
EcMF					
All taxa	153	$P = 14.3$, $P = 0.01$	$D = 0.09$, $P = 0.15$	$W = 0.96$, $P < 0.001$	$A = 1.76$, $P < 0.001$
Core group	71	$P = 4.6$, $P = 0.47$	$D = 0.08$, $P = 0.76$	$W = 0.98$, $P = 0.27$	$A = 0.38$, $P = 0.39$
Satellite group	82	$P = 13.6$, $P = 0.02$	$D = 0.12$, $P = 0.19$	$W = 0.94$, $P = 0.001$	$A = 1.33$, $P = 0.002$
AMF					
All taxa	48	$P = 8$, $P = 0.16$	$D = 0.08$, $P = 0.89$	$W = 0.98$, $P = 0.6$	$A = 0.22$, $P = 0.83$
Core group	19	$P = 2.5$, $P = 0.78$	$D = 0.12$, $P = 0.94$	$W = 0.97$, $P = 0.86$	$A = 0.2$, $P = 0.86$
Satellite group	29	$P = 8.2$, $P = 0.14$	$D = 0.13$, $P = 0.66$	$W = 0.96$, $P = 0.25$	$A = 0.36$, $P = 0.43$

MOTUs (Fig. 4a, Table 2). Richness estimators behaved similarly in the EcMF data set, but the Chao 2 function seemed to level off (Fig. 4b). All species richness estimators suggested the true richness of AMF to be nearly the same as the number of observed MOTUs (Fig. 4c, Table 2), although two estimators (ACE and Bootstrap) failed to reach a stable plateau. Jackknife2 reached a stable value at 55 ± 2 MOTUs whereas Chao 2 equalled the observed richness of 48 MOTUs.

Semi-log transformation of Ugland's T–S curve resulted in nearly perfect fits to linear regression functions for the PpF and EcMF data sets (Fig. 4a, b; $R^2 = 0.99$ for both). The linear fit was somewhat lower for the AMF assemblage (Fig. 4c; $R^2 = 0.95$). The parameters of the regression lines allowed both evaluating the observed richness by downscaling to parts of the data set (10% and 50% data coverage in Table 2) and predicting fungal richness of larger areas, or host

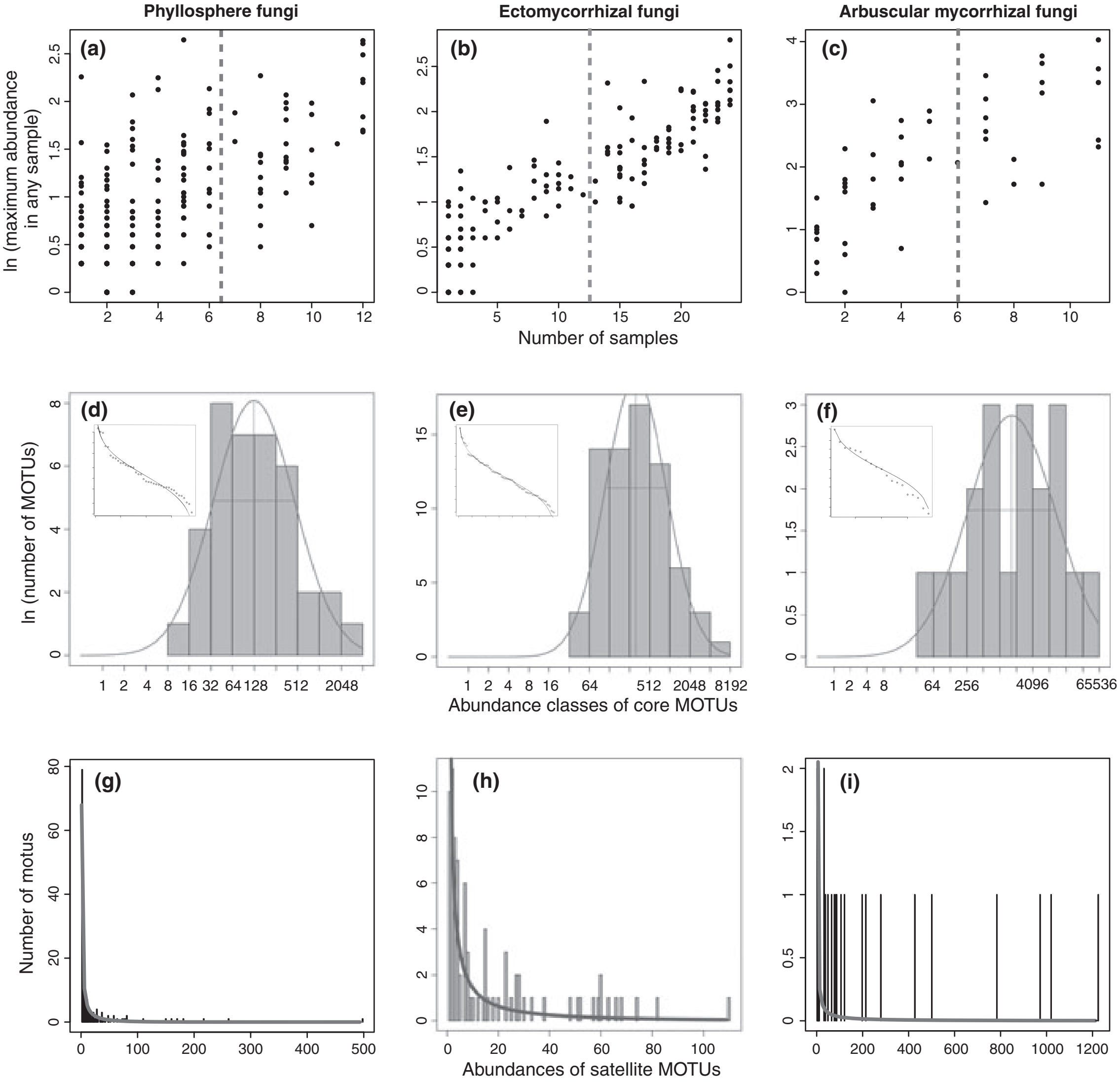

Fig. 2 Delimitation of core and satellite taxa for the three fungal data sets. (a–c) Patterns of abundance and persistence in the fungal assemblages. The 50% persistence threshold (grey dashed lines, cf. 'Materials and methods') allowed to separate the abundant core MOTUs (>50% persistence) from the satellite MOTUs (<50% persistence). (d–f) 'Binned' MOTU abundances of the core group overlaid with a fitted truncated log-normal SAD (grey curve). The insert shows the fit to a log-normal SAD as displayed by a Whittaker plot. (g–i) Visual test for a fit to log-series distribution of the satellite groups.

plants in the case of mycorrhizal fungi, by upscaling of calculations (200% and 500% data coverage in Table 2).

The effect of omitting the rarest MOTUs for species richness analysis depended on the data set used (Fig. 5). Removal of molecular singletons from the PpF data had the most pronounced effect on both observed and estimated MOTU richness based on Chao 2 estimator (Fig. 5a, d). Deleting further MOTUs resulted in gradually decreasing differences. Differences in Chao 2 estimations were non-significant based on the respective lower and upper 95% confidence interval when molecular tripletons (curve 3), quadrupletons (curve 4) and quintupletons (curve 5) were given the rank of biological singletons (Fig. 5d). The same treatment had less pronounced impacts on observed and predicted EcMF richness (Fig. 5b,e). The successive exclusion of rarest AMF MOTUs resulted in negligible changes for both observed and estimated MOTU richness (Fig. 5c,f).

Discussion

Species abundance distributions, accumulation curves and estimators

The present statistics of the three empirical fungal data sets differed strongly in sampling intensity. The assessment of the AMF species richness was comprehensive, as the SAC displays a long and stable plateau and most estimators predicted the presence of just one to two additional species. The AMF community clearly fits a log-normal type of SADs. Thus, we propose the log-normal SAD as a working hypothesis to elucidate MOTU richness and biodiversity of AMF communities with low to medium sampling coverage that agrees with suggestions of Dumbrell *et al.* (2010). Our results also corroborate the findings of Ulirich *et al.* (2010) that completely sampled animal and plant communities follow a log-normal SAD. By contrast, the EcMF and PpF communities were slightly to strongly undersampled as the richness estimators predicted 85–97% and 58–90% inventory exhaustiveness, respectively. These two communities failed to fit log-normal distributions.

Non-parametric richness estimators are traditionally considered more accurate than parametric estimators, but their performance may strongly suffer from undersampling bias. To overcome these problems, Coddington *et al.* (2009) suggested the use of permutation procedures and parametric estimators that relied on the 'veiled line' or truncated log-normal concept of Preston (1948). The choice of 0.5 as the default 'veiled line' threshold is an arbitrary but common choice without foundation in statistics (Dewdney 1998; Golicher *et al.* 2006). Here, Preston's concept was applied and showed

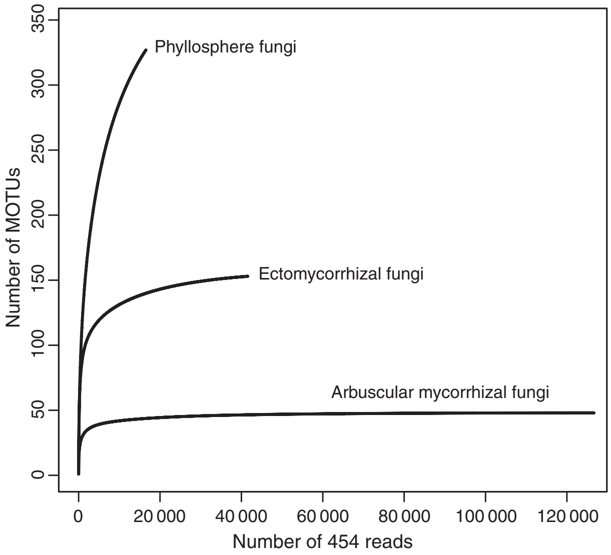

Fig. 3 Rarefied taxon (MOTU) accumulation curves of the three plant-related fungal communities obtained by 454 sequencing.

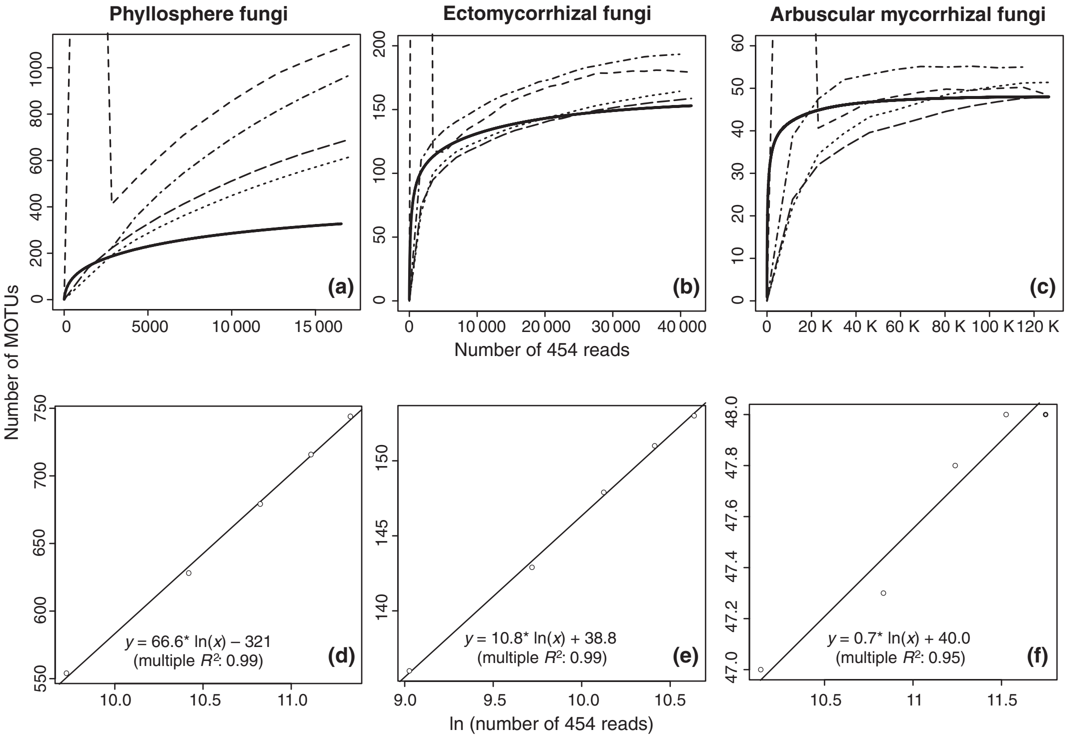

Fig. 4 Observed and estimated MOTU richness for the three fungal communities. (a–c): rarefied species accumulation curves (thick solid line) and the estimators Jack 2 (– - –), Chao 2 (– – –), Bootstrap (- - -) and ACE (□ □ □). (d–f): linear fit of Ugland's semi-log transformed T-S curve (Ugland *et al.* 2003) and the corresponding regression function, that allows calculation of observed and predicted MOTU richness. According to the equation of (d) a sequencing effort of for example 5 million reads would result in 1,225 phyllosphere MOTUs.

Table 2 Observed and estimated MOTU richness. The values of observed MOTUs are based on the Coleman rarefaction curve calculated with EstimateS. The four estimators ACE, Chao 2, Jack 2 and Bootstrap were also calculated with EstimateS. Ugland's estimator is based on a regression line (see log-linear functions in Fig. 5d–f) and additionally allowed the calculations of species richness at lesser data coverage. The parametric estimator in the rightmost column is based on the 'veiled-line' concept of Preston's log-normal distribution. Different veiled-line thresholds (trunc =) were used, only the upper and lower limits are shown. Plus-minus values are standard deviations

Data set	MOTUs observed	ACE	Chao 2	Jack 2	Bootstrap	Ugland	Log-normal
PpF							
100% data coverage	327	854 ± 9	821 ± 85	1276 ± 12	830 ± 5	326 ± 2	352 (trunc = 0.5) – 642 (trunc = 0.9)
10%	155					173 ± 3	
50%	270					280 ± 4	
200%	–					372 ± 3	
500%	–					433 ± 5	
EcMF							
100% data coverage	153	159	158 ± 13	181 ± 9	164 ± 2	154 ± 5	160 (trunc = 0.5) – 171 (trunc = 0.9)
10%	116					128 ± 5	
50%	144					146 ± 5	
200%	–					161 ± 5	
500%	–					171 ± 6	
AMF							
100% data coverage	48	48	48 ± 3	55 ± 2	51 ± 1	48 ± 2	49 (trunc = 0.5–0.9)
10%	42					47 ± 2	
50%	45					47 ± 3	
200%	–					49 ± 2	
500%	–					49 ± 3	

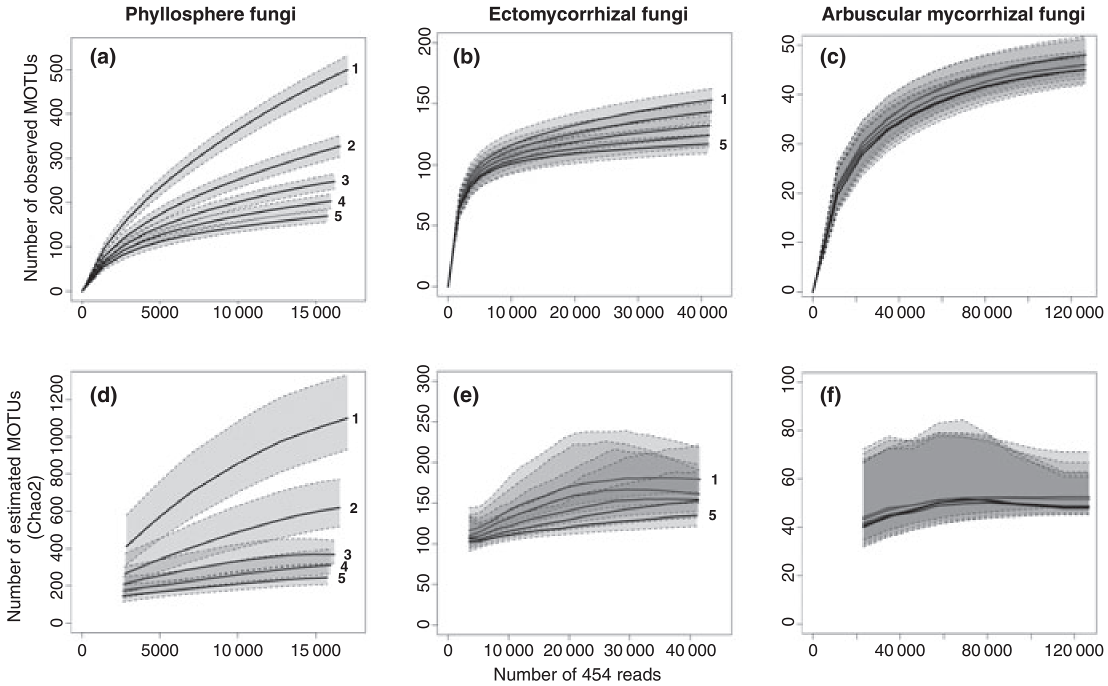

Fig. 5 Different handling of rarest MOTUs influenced observed (a–c) and estimated (d–f) species richness. The effect of keeping molecular singletons (curve 1) and deleting molecular singletons (2), doubletons (3), tripletons (4) and quadrupletons (5) varied both within (a, d) as well as between the three data sets.

that changes in the threshold from 0.5 to 0.9 may have a large effect on the fitted log-normal distribution and thus on the estimated species richness (Table 2). The more complete the sampling, such as in the EcMF and AMF data, the more robust was this estimator, irrespective of the threshold applied. Interestingly, recent studies caution about the uncritical use of Preston plots or other histograms that are based on grouping species abundances (Ulrich *et al.* 2010). Instead they advocate the use of raw data plots such as Whittaker plots or probability plots to fit SAD models in order to retain precise abundance information. The presented descriptions of molecular fungal community structure corroborate these recommendations.

The regression function obtained from the semi-log transformed T–S curve (Fig. 4d–f; Ugland *et al.* 2003) gave almost identical numbers of observed richness (100% data coverage in Table 2), but overestimated observed MOTU richness at lower sampling effort compared with the 'Coleman curve' of EstimateS (Table 2). Extrapolations beyond 100% data coverage yielded richness predictions that fell within the range of the other estimators. Given that Ugland's method was previously considered as a reliable predictor of fungal species richness (Unterseher *et al.* 2008) and that it had also proved superior to other estimators, in particular for voluminous data sets and heterogeneous environments (Ugland *et al.* 2003), this estimator deserves serious consideration in future predictions of fungal species/MOTU richness.

The core and satellite species approach

Despite generating tens of thousands of sequences, undersampling remains problematic in environmental NGS datasets, especially addressing hyperdiverse organisms in transient habitats (Jumpponen & Jones 2010). In the present study the core satellite concept (Hanski 1982; Magurran & Henderson 2003) was applied to describe and compare the fungal communities despite undersampling of one (PpF) and oversampling of another (AMF). The proportion of core to satellite taxa was up to five times higher for mycorrhizal than for phyllosphere fungi. It can be concluded that the relative proportion of core species increases with increased stability of the system at hand. However, more data have to be analysed to generally validate this assumption. Assignment of MOTUs to core and satellite groups may further prove useful in studies of effects of climate, land use or habitat change on microbial diversity. Ecologically relevant shifts in abundances probably occur predominantly among the core members that are by definition well established in the system (Magurran & Henderson 2003).

Differences in the SADs of core and satellite groups were most prominent for PpF data. The PpF core group fitted log-normal SAD and can be viewed as a comparatively characteristic assemblage specific to the phyllosphere of deciduous trees, comprising mostly Ascomycota (Jumpponen & Jones 2009, 2010; Supporting information). In contrast, the PpF satellite group approached a log-series SAD. It comprised many ubiquitous taxa of both Ascomycota and Basidiomycota (Supporting information). The basidiomycete taxa belonged almost entirely to Tremellomycetes (Filobasidiales, Tremellales, Erythrobasidiales) that may live as epiphyllous yeasts (Fonseca & Inácio 2006). Many satellite taxa in general represent transient inhabitants (migrants) from the surrounding species pool that are physiologically inactive or contribute little to functions of a particular ecosystem. The distinction between SADs of core and satellite assemblages became less clear for EcMF and in particular for AMF data, but it has to be noted that MOTUs belonging to non-mycorrhizal taxa were all effectively removed from these analyses. Nevertheless, on the basis of the evaluated data this lack of distinctiveness again points to relative stability of the entire mycorrhizal systems.

Sequencing artefacts influence observed and estimated MOTU richness

It is widely acknowledged that thresholds of sequence variability influence the interpretation of biological species richness based on molecular data. Sequence errors add even more uncertainty to the analyses. Tedersoo *et al.* (2010) showed that *c.* 75% of 454 pyrosequencing singletons are technical artefacts in agreement with recent studies (Quince *et al.* 2008, 2009; Kunin *et al.* 2010). Currently, the removal of singletons is recommended for eukaryotic microorganisms to avoid overestimation of species richness (Dickie 2010; Medinger *et al.* 2010; Tedersoo *et al.* 2010).

Exclusion of the MOTUs with less than five reads significantly influenced both observed and predicted MOTU richness of the undersampled PpF community. Giving molecular doubletons the rank of biological singletons (i.e. all molecular singletons deleted from the data set) significantly lowered the numbers of observed and predicted MOTUs, whereas further exclusion of rare taxa had gradually lower impacts. For the AMF data, richness analyses were not affected by the removal of the four rarest MOTUs. These results again indicate the importance of survey intensity for any biodiversity inventory and cannot be overstated in the light of environmental NGS studies. The more complete the representation of the parent community in empirical data, the more robust species richness analysis is with respect to sequencing artefacts. The general problem of predicting species richness from undersampled community data (O'Hara 2005) could be overcome for NGS by applying a more conservative interpretation of 454 reads (i.e. removal of molecular singletons) and by the application of a less conservative richness estimator, e.g. Ugland's function (Ugland *et al.* 2003).

Reasons for the observed hyperdiversity of microbial communities and methodological implications

High species richness and a power-law SAD as observed for the fungal phyllosphere assemblage is an inherent characteristic of most bacterial and fungal communities. In contrast to sexually reproducing macro-organisms, for which reproductive constraints limit the number of extremely rare species (Coddington *et al.* 2009), micro-organisms can survive without a compatible partner and thus successfully colonize a habitat from a single spore or asexual propagule (Schnittler & Tesmer 2008). In addition, spatial and temporal alternation of life cycles may strongly affect composition of fungal communities in soil and phyllosphere (Schadt *et al.* 2003; Unterseher *et al.* 2007). Therefore, 'snapshot samples' may not always reflect natural properties of hyperdiverse communities.

The three data sets compared for the present study targeted different loci and different lineages of fungi which presumably evolve at different rates. The less restrictive bioinformatics pipeline for the phyllosphere data (Jumpponen & Jones 2010) which could have split rather than combine, MOTUs, may be an additional argument for the observed divergence in the phyllosphere species richness (Fig. 3).

However, once technical and bioinformatic problems of NGS have been solved successfully and reasonable thresholds of intraspecific sequence variability have been agreed upon for biologically meaningful MOTUs, we may be able to derive sound richness extrapolations and estimates of species abundance distribution from the observed data sets. If additionally an appropriate sampling design is used – microhabitats rather than large and highly variable systems – one could especially profit from NGS so that undersampling becomes the exception rather than the rule.

The preference for microhabitats in microbial biodiversity assessment instead of surveys at the level of ecosystems is even more consequential in the light of the recent findings that local SADs are downscaled versions of the respective metacommunity SADs (Connolly *et al.* 2005; Ulrich *et al.* 2010). With (near-)complete surveys and sufficient statistical power at hand, environmental NGS will reveal the true underlying community in a particular microhabitat allowing also the development of universally valid models for larger metacommunities.

Conclusions

It still remains unclear what information the molecular abundance measures of NGS data carry. Even if PCR biases are considered unimportant, sequence abundance at most shows the amount of undegraded genetic information in the environment that may or may not reflect the number of species or biomass. NGS provides large amounts of data, but undersampling remains a problem in transient, species-rich communities. We developed a conservative method of treating the rarest MOTUs in NGS biodiversity surveys and demonstrate the applicability of community models in large sequence data sets. However, there are no one-size-fits-all solutions. Instead, one should start with an inherent understanding of the system, consider what the pertinent hypotheses are and how to best address them. It may not always be necessary to saturate the richness but this should be an a priori decision. If the investigator's first and foremost aim is to unravel true species richness of microbial communities, a highly important task in biodiversity research, he/she should be prepared for a thorough and intensive sampling that will far exceed average sampling efforts as well for a pluralistic way of dealing with SADs and species richness estimators.

Authors' contribution

MU designed the study and performed statistical analyses; AJ, MÖ, LT, MM provided data; CFD, MS performed statistical analyses; all authors contributed to writing the manuscript.

Acknowledgements

The phyllosphere data were collected with support from Kansas State University, Division of Biology and Ecological Genomics Institute. Research on mycorrhizal fungi was funded by the Estonian Science Foundation grants 7371, 7434, JD-0092J, 7738, SF0180098s08 (University of Tartu), Marie Curie European Reintegration Grant within the 7th European Community Framework Programme (GLOBAM, PERG03-GA-2008-231034, MÖ) and by the European Regional Development Fund (Centre of Excellence FIBIR). CFD acknowledges funding by the Helmholtz Association (VH-NG 247), MU and MS acknowledge support from the Faculty of Mathematics and Natural Science, University of Greifswald.

References

Bellemain E, Carlsen T, Brochmann C *et al.* (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Mircobiology*, **10**, 189.

Bisby FA, Coddington J, Thorpe JP *et al.* (1995) Characterization of biodiversity. In: *Global Biodiversity Assessment* (eds Heywood VH, Gardener K). pp. 22–106, Cambridge University Press for UNEP, Cambridge.

Buée M, Reich M, Murat C *et al.* (2009) 454 pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, **184**, 449–456.

Coddington JA, Agnarsson I, Miller JA, Kuntner M, Hormiga G (2009) Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. *Journal of Animal Ecology*, **78**, 573–584.

Colwell RK (2009) *Estimates: statistical estimation of species richness and shared species from samples. Version 8.2*. User's guide and application published at: http://purl.oclc.org/estimates. (assessed 27/10/2010)

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society London B*, **345**, 101–118.

Connolly S, Hughes T, Bellwood D, Karlson R (2005) Community structure of corals and reef fishes at multiple scales. *Science*, **309**, 1363.

Dewdney AK (1998) A general theory of the sampling process with applications to the ''veil line''. *Theoretical Population Biology*, **54**, 294–302.

Dickie IA (2010) Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist*, **188**, 916–918.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Idiosyncrasy and overdominance in the structure of natural communities of arbuscular mycorrhizal fungi: is there a role for stochastic processes? *Journal of Ecology*, **98**, 419–428.

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.

Fonseca A, Inácio J (2006) Phylloplane yeasts. In: *Biodiversity and Ecophysiology of Yeasts* (eds Rosa CA, Péter G). pp. 263–301, Springer-Verlag, Berlin, Heidelberg.

Galand P, Casamayor E, Kirchman D, Lovejoy C (2009) Ecology of the rare microbial biosphere of the arctic ocean. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 22427–22432.

Golicher D, O'Hara R, Ruíz-Montoya L, Cayuela L (2006) Lifting a veil on diversity: a Bayesian approach to fitting relative-abundance models. *Ecological Applications*, **16**, 202–212.

Gotelli N, Colwell R (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Hanski I (1982) Dynamics of regional distribution: The core and satellite species hypothesis. *Oikos*, **38**, 210–221.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogegraphy*. Princeton University Press, New Jersey.

Jost L (2007) Partitioning diversity into independent aopha and beta components. *Ecology*, **88**, 2427–2439.

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

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

Keijser BJF, Zaura E, Huse SM *et al.* (2008) Pyrosequencing analysis of the oral microflora of healthy adults. *Journal of Dental Research*, **87**, 1016–1020.

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.

MacArthur RH (1957) On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, **43**, 293–295.

Magurran A, Henderson P (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714–716.

May RM (1975) Patterns of species abundance and diversity. In: *Ecology and Evolution of Communities* (eds Cody ML, Diamond JM). pp. 81–120, Harvard University Press, Cambridge, MA.

McGill B, Etienne R, Gray J *et al.* (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995–1015.

McPherson J, Jetz W (2007) Effects of species' ecology on the accuracy of distribution models. *Ecography*, **30**, 135–151.

Medinger R, Nolte V, Pandey RV *et al.* (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, **19**, S1, 32–40.

O'Hara RB (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, **74**, 375–386.

Öpik M, Metsis M, Daniell TJ, Zobel M, Moora M (2009) Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist*, **184**, 424–437.

Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? *Trends in Microbiology*, **14**, 257–263.

Preston FW (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

Quince C, Curtis T, Sloan W (2008) The rational exploration of microbial diversity. *ISME Journal*, **2**, 997–1006.

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

R Development Core Team (2009) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schadt C, Martin A, Lipson D, Schmidt S (2003) Seasonal dynamics of previously unknown fungal lineages in tundra soils. *Science*, **301**, 1359–1361.

Schnittler M, Tesmer J (2008) A habitat colonisation model for spore-dispersed organisms – does it work with eumycetozoans? *Mycological Research*, **112**, 697–707.

Sogin M, Morrison H, Huber J *et al.* (2006) Microbial diversity in the deep sea and the underexplored ''rare biosphere''. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–12120.

Stoeck T, Behnke A, Christen R *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology*, **7**, 72.

Sun Y, Cai Y, Liu L *et al.* (2009) ESPRIT: Estimating species richness using large collections of 16s rRNA pyrosequences. *Nucleic Acids Research*, **37**, e76.

Taylor JW, Jacobson DJ, Kroken S *et al.* (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genetics and Biology*, **31**, 21–32.

Tedersoo L, Nilsson RH, Abarenkov K *et al.* (2010) 454 pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, **188**, 291–301.

Tokeshi M (1990) Niche apportionment or random assortment: species abundance patterns revisited. *Journal of Animal Ecology*, **59**, 1129–1146.

Ugland KI, Gray JS, Ellingsen KE (2003) The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology*, **72**, 888–897.

Ulrich W, Ollik M, Ugland KI (2010) A meta-analysis of species–abundance distributions. *Oikos*, **119**, 1149–1155.

Unterseher M, Schnittler M (2010) Species richness analysis and its rDNA phylogeny revealed the majority of cultivable foliar endophytes from beech (*Fagus sylvatica*). *Fungal Ecology*, **3**, 366–378.

Unterseher M, Reiher A, Finstermeier K, Otto P, Morawetz W (2007) Species richness and distribution patterns of leaf-inhabiting endophytic fungi in a temperate forest canopy. *Mycological Progress*, **6**, 201–212.

Unterseher M, Schnittler M, Dormann C, Sickert A (2008) Application of species richness estimators for the assessment of fungal diversity. *FEMS Microbiology Letters*, **282**, 205–213.

Whittaker RH (1977) Evolution of species diversity in land communities. *Evolutionary Biology*, **10**, 1–67.

Williamson M, Gaston K (2005) The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *Journal of Animal Ecology*, **74**, 409–422.

von Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, **21**, 213–229.

M.U. investigates the biodiversity of fungi, especially of foliar endophytes and focuses on species richness and species concepts in the light of molecular identification. A.J. is focusing on various topics ranging from theoretical work on succession of mycorrhizal fungi, to defining niche concepts for fungi and to assessing fungal diversity with next-generation sequencing. M.Ö. has specialized on arbuscular mycorrhizal fungi and their influence on plant performance with a special emphasis on molecular and functional diversity. M.M. is researcher in plant ecology and investigates mechanisms behind plant community structure with a special emphasis on mycorrhizal symbiosis. L.H. is researcher in fungal molecular ecology and has expertise in root-inhabiting fungi and in bioinformatical management and analyses of ecological pyrosequencing data. C.F.D. investigates the mechanistic functioning of ecological communities, in order to be able to predict effects of Global Environmental Change. He has expertise in statistical ecology and mechanistic modeling. M.S. has expertise in the diversity of myxomycetes, other protists and true fungi and in population genetics of higher plants.

Supporting information

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

Data S1 Data zip: This contains all data used for the analyses.

Data S2 Phyllosphere_fungi.txt: This file contains taxonomic information of the phyllosphere MOTUs.

Data S3 Source. R: This file contains all R commands used for the present study, some additional explanations and further analyses not presented in the paper.

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.