environmental microbiology

Environmental Microbiology (2014) **16**(9), 2764–2776 doi:10.1111/1462-2920.12338

Identification accuracy and diversity reproducibility associated with internal transcribed spacer-based fungal taxonomic library preparation

Naomichi Yamamoto,[1] Karen C. Dannemiller,[2] Kyle Bibby[3,4] and Jordan Peccia[2]*

[1]*Department of Environmental Health, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.*

[2]*Department of Chemical and Environmental Engineering, Yale University, New Haven, CT 06520, USA.*

[3]*Department of Civil and Environmental Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA.*

[4]*Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA.*

Summary

This study investigated analytical parameters that are inherently relevant to identifying and quantifying fungal communities based on polymerase chain reaction amplicons. Specifically, we evaluated the accuracy of the BLASTn-based identification for internal transcribed spacer (ITS) sequences generated from pure cultures, and quantified the reproducibility of relative abundances as well as α and β diversity measurements using duplicated environmental samples. The BLASTn-based method produced accurate fungal identification for the pure culture sequences at the genus rank. Percentages of the sequences assigned to correct genera were 99.8%, 99.8% and 99.9% for *Alternaria alternata*, *Cladosporium cladosporioides* and *Epicoccum nigrum* respectively. These fractions were smaller for *Aspergillus fumigatus* and *Penicillium chrysogenum*, which have dual nomenclatures or sibling species that are indistinguishable by ITS sequences. Our duplicate environmental analyses demonstrated that α diversity and relative abundance levels were reproducible ($r^2 > 0.9$), that variability decreases with increased sequence quantity, and that the differences in distinct environmental samples were larger than differences in replicate samples (β diversity). These results serve to better characterize the identification and quantification limits of ITS-based fungal taxonomic studies, and demonstrate that while diversity quantification is reproducible, limitations in ITS-based taxonomic identification and dual nomenclature conventions are current barriers to identification accuracy.

Introduction

The study of fungal diversity is integral to environmental science and health. Many fungi are saprophytic, and their role in the decomposition of organic substances in terrestrial environments is a significant contributor to global carbon cycling (Hodge *et al.*, 2001; Zhu and Miller, 2003). Selected fungi are also clinically important as human allergens (Crameri *et al.*, 2006), are associated with chronic inflammatory disease (Ege *et al.*, 2011; Iliev *et al.*, 2012), and cause primary or opportunistic infections (Pitman *et al.*, 2011). Improvements in methods used to study fungal diversity may enhance our ability to determine the role of fungi in human health and the environment.

Similar sequencing methods, primer sets and BLASTn-based approaches have been applied for nearly all fungal ecological studies using next-generation DNA sequencing (Buée *et al.*, 2009; Amend *et al.*, 2010b; Blaalid *et al.*, 2012; Yamamoto *et al.*, 2012; Adams *et al.*, 2013). The internal transcribed spacer (ITS) region is considered the primary barcode marker for fungal identification (Schoch *et al.*, 2012), and due to its hypervariability, identification to the rank of species has been considered possible when using the appropriate primers (Buée *et al.*, 2009; Amend *et al.*, 2010b; Yamamoto *et al.*, 2012). Prior studies have utilized the BLASTn programme to taxonomically assign individual sequence reads using a database with named fungal ITS sequences (Nilsson *et al.*, 2009; Yamamoto *et al.*, 2012), or have applied 'training-type' classifiers such as the Ribosomal Database Project, MEtaGenome Analyzer or BLASTClust to cluster the sequences into operational taxonomic units (OTUs), and subsequently annotate these clustered sequences by BLASTn (O'Brien *et al.*, 2005; Buée *et al.*, 2009; Fröhlich-Nowoisky *et al.*,

Received 11 April, 2013; accepted 19 August, 2013. *For correspondence. E-mail jordan.peccia@yale.edu; Tel. (+1) 203 432 4385; Fax (+1) 203 432 4387.

2009; Amend *et al.*, 2010b; Blaalid *et al.*, 2012; Adams *et al.*, 2013).

It has been previously recognized that impediments exist to highly accurate ITS identification at the rank of species. BLASTn often results in ambiguity due to tying top taxonomic hits with equivalent E-values, the nomenclature of fungal species allows for different names for genetically similar species (Taylor, 2011) and fungal ITS databases are known to contain incorrectly classified fungal species (Nilsson *et al.*, 2006). Setting informed expectations for using next-generation DNA sequencing in fungal ecology requires information on the quantitativeness of the methods, the accuracy with which current databases and database search methods annotate amplicons, and quantitative estimates of the reproducibility associated with these methods. Notably, Amend and colleagues (2010a) examined the performance of fungal abundance quantification by 454 pyrosequencing of ITS amplicons and demonstrated that the stringency of read quality processing impacted annotation outcome. However, further studies are needed to define annotation accuracy and the reproducibility of relative abundance and diversity measures for characterizing fungal communities.

Studies to define identification accuracy and reproducibility are not available for ITS-based fungal ecological analyses. The goal of this study is to evaluate the accuracy of BLASTn-based fungal taxonomic assignments and the reproducibility of characterizing fungal communities based on next-generation DNA sequencing of the fungal ITS region. To define accuracy, ITS taxonomic assignments of sequences from five medically important fungal pure-cultures were analysed. To characterize the overall reproducibility for environmental fungal communities by amplicon sequencing, reads from five duplicated environmental samples were processed and the resulting α and β diversity results and relative abundances were statistically compared. Together, these results characterize the current annotation and reproducibility limits of fungal amplicon sequencing and demonstrate areas for potential improvements.

Results

Summary statistics

A total of 14 065 ITS sequences, including 8452 and 5613 sequences from the pure culture and environmental samples, respectively, were obtained (Tables 1,2 and Supporting Information Table S2). The median lengths of the trimmed sequences of *Alternaria alternata*, *Aspergillus fumigatus*, *Cladosporium cladosporioides*, *Epicoccum nigrum*, *Penicillium chrysogenum* and the environmental samples were 491, 514, 533, 496, 535, and 472 bp respectively (Fig. 1 and Supporting Information Table S1). These lengths are approximately 87%, 87%, 97%, 92% and 92% of the lengths of the amplicons respectively (Fig. 1).

For the pure-cultured fungi, ITS1 sequences were also extracted from the original ITS sequences. A total of 8445 ITS1 sequences were obtained (Tables 3 and 4). The median lengths of the extracted ITS1 sequences of *A. alternata*, *A. fumigatus*, *C. cladosporioides*, *E. nigrum* and *P. chrysogenum* were 164, 184, 153, 142 and 175 bp respectively.

Alpha diversity metrics, including the number of observed OTUs, the Chao1 estimator and Shannon index, were calculated without and with normalizing the number of sequence reads. Without normalizing the sequence reads, the number of the observed OTUs ranged from 179 to 545 for the environmental samples (Supporting Information Table S2). Chao1 estimator predicted 1010–5157 OTUs, whereas Shannon indices ranged from 4.5 to 8.7 (Supporting Information Table S2). When α diversity indices were based on 250 randomly chosen sequences in each sample, diversity metrics decreased, with the number of the observed OTUs ranging from 92 to 228 for the environmental samples (Supporting Information Table S3). Chao1 estimator predicted 756–2912 OTUs, whereas Shannon indices ranged from 4.2 to 7.8 (Supporting Information Table S3).

For the environmental samples, 4713 and 3998 sequences were identified to the genus and species ranks, respectively, including 371 different genera and 642 different species (Supporting Information Table S4). A full list of fungal species identified from the environmental samples is summarized in Supporting Information Table S5.

Accuracy

For the pure culture accuracy tests, taxonomic assignments at the species rank were often ambiguous due to tying top BLASTn hits (E-value). Supporting Information Fig. S1–5 show fractions of the sequences by types of taxonomic placements in case sequences of the entire ITS were used for BLASTn. Large fractions of the sequences were assigned to correct species either solely or ambiguously, that is, 99.6%, 76.6%, 99.2%, 99.9% and 95.8% for *A. alternata*, *A. fumigatus*, *C. cladosporioides*, *E. nigrum*, and *P. chrysogenum* respectively (Supporting Information Fig. S1–5). At the genus rank, the fractions were 99.9%, 99.8%, 99.9%, 99.9% and 99.8% respectively (Supporting Information Fig. S1–5).

In case sequences of the entire ITS were used for BLASTn, true identification ratios (TIRs) were 2.1%, 30.9%, 0.3%, 99.9% and 8.2% for *A. alternata*, *A. fumigatus*, *C. cladosporioides*, *E. nigrum* and *P. chrysogenum* at the species rank respectively (Table 1). Some

Table 1. Accuracy of BLASTn-based identification for pure-culture fungi at the species rank using the full-length ITS sequences generated by 454 pyrosequencing.

Tested microorganisms	Assigned taxa[a]	Number of sequences	Percentage of sequences (%)	True / False	TIR (%)[b]	FIR (%)[c]	ER (%)[d]
Alternaria alternata	*Alternaria alternata*	38	2.14	T	2.1	0.4	17.4
	Alternaria spp.	1730	97.30	n.d.			
	Alternaria citri	5	0.28	F			
	Alternaria tenuissima	1	0.06	F			
	Aspergillus fumigatus	1	0.06	F			
	Cladosporium spp.	1	0.06	F			
	Unidentified	2	0.11	n.d.			
	Total	1778	100				
Aspergillus fumigatus	*Aspergillus fumigatus*	161	30.90	T	30.9	8.1	20.7
	Aspergillus spp.	1	0.19	n.d.			
	Aspergillus lentulus	4	0.77	F			
	Aspergillus ustus	8	1.54	F			
	Neosartorya fischeri	30	5.76	F			
	Neosartorya spp.	30	5.76	n.d.			
	Unidentified	287	55.09	n.d.			
	Total	521	100				
Cladosporium cladosporioides	*Cladosporium cladosporioides*	8	0.26	T	0.3	0.8	74.2
	Cladosporium spp.	2989	98.81	n.d.			
	Cladosporium cucumerinum	21	0.69	F			
	Davidiella tassiana	1	0.03	F			
	Glomerella cingulata	1	0.03	F			
	Unidentified	5	0.17	n.d.			
	Total	3025	100				
Epicoccum nigrum	*Epicoccum nigrum*	2068	99.90	T	99.9	0.0	0.0
	Pycnopeziza sympodialis	1	0.05	F			
	Unidentified	1	0.05	n.d.			
	Total	2070	100				
Penicillium chrysogenum	*Penicillium chrysogenum*	87	8.22	T	8.2	2.8	25.6
	Penicillium spp.	5	0.47	n.d.			
	Penicillium allii	1	0.09	F			
	Penicillium brevistipitatum	3	0.28	F			
	Penicillium commune	1	0.09	F			
	Penicillium concentricum	1	0.09	F			
	Penicillium coprobium	2	0.19	F			
	Penicillium granulatum	13	1.23	F			
	Penicillium griseofulvum	4	0.38	F			
	Penicillium lanosum	3	0.28	F			
	Penicillium vinaceum	1	0.09	F			
	Talaromyces leycettanus	1	0.09	F			
	Unidentified	936	88.47	n.d.			
	Total	1058	100				

a. Species name is denoted as 'spp.' in case sequences are ambiguous at the species but not genus rank due to multiple top hits by BLASTn. Sequence is denoted as 'unidentified' in case they are ambiguous at the genus rank due to multiple top hit by BLASTn. Trueness of unidentified sequences is not determined because they might be assigned with a correct taxon along with incorrect taxa.
b. TIR is calculated according to Eq. (1).
c. FIR is calculated according to Eq. (2).
d. ER is calculated according to Eq. (3).
n.d., trueness is not determined due to ambiguous identity at this taxonomic rank.

sequences were falsely assigned to an incorrect taxon. False identification ratios (FIRs) were 0.4%, 8.1%, 0.8%, 0.0% and 2.8% for *A. alternata*, *A. fumigatus*, *C. cladosporioides*, *E. nigrum* and *P. chrysogenum* at the species rank respectively (Table 1). *A. fumigatus* sequences were assigned to the teleomorphic relative *Neosartorya fischeri*, in 5.8% of the cases. *C. cladosporioides* sequences were assigned to *C. cladosporioides* and *C. cucumerinum* with tying top hits in 98.8% of the cases (Supporting Information Fig. S3).

Greater accuracy in taxonomic placements was achieved at the genus rank. TIRs were 99.8%, 44.9%, 99.8%, 99.9% and 11.4% for *A. alternata*, *A. fumigatus*, *C. cladosporioides*, *E. nigrum* and *P. chrysogenum* at the genus rank respectively (Table 2). At the genus rank, 11.5% of the *A. fumigatus* sequences were assigned to *Neosartorya* spp., teleomorphic states of *Aspergillus* spp. (Nierman *et al.*, 2005; O'Gorman *et al.*, 2009) (Table 2). Notably, 88.5% of the *P. chrysogenum* sequences were ambiguous even at the genus level (Table 2) due to tying

Table 2. Accuracy of the BLASTn-based identification for pure culture fungi at the genus rank using the full-length ITS sequences generated by 454 pyrosequencing.

Tested microorganisms	Assigned taxa[a]	Number of sequences	Percentage of sequences (%)	True / False	TIR (%)[b]	FIR (%)[c]	ER (%)[d]
Alternaria alternata	*Alternaria*	1774	99.78	T	99.8	0.1	0.1
	Aspergillus	1	0.06	F			
	Cladosporium	1	0.06	F			
	Unidentified	2	0.11	n.d.			
	Total	1778	100				
Aspergillus fumigatus	*Aspergillus*	174	33.40	T	44.9	0.0	0.0
	Neosartorya	60	11.52	T			
	Unidentified	287	55.09	n.d.			
	Total	521	100				
Cladosporium cladosporioides	*Cladosporium*	3018	99.77	T	99.8	0.1	0.1
	Davidiella	1	0.03	F			
	Glomerella	1	0.03	F			
	Unidentified	5	0.17	n.d.			
	Total	3025	100				
Epicoccum nigrum	*Epicoccum*	2068	99.90	T	99.9	0.0	0.0
	Pycnopeziza	1	0.05	F			
	Unidentified	1	0.05	n.d.			
	Total	2070	100				
Penicillium chrysogenum	*Penicillium*	121	11.44	T	11.4	0.1	0.8
	Talaromyces	1	0.09	F			
	Unidentified	936	88.47	n.d.			
	Total	1058	100				

a. Sequence is denoted as 'unidentified' in case they are ambiguous at the genus rank due to multiple top hit by BLASTn. Trueness of unidentified sequences is not determined because they might be assigned with a correct taxon along with incorrect taxa.
b. TIR is calculated according to Eq. (1).
c. FIR is calculated according to Eq. (2).
d. ER is calculated according to Eq. (3).
n.d., trueness is not determined owing to ambiguous identity at this taxonomic level.

top hits with fungal genera *Botryosphaeria*, *Eupenicillium*, *Hypocrea* and *Scopulariopsis* (Supporting Information Fig. S5). FIRs as well as error ratios (ERs) were less than 0.8% for any of the pure cultures considered at the genus rank (Table 2).

Overall, identification accuracy for the pure-cultured fungi tended to be higher with the full-length ITS sequences than with the shorter extracted ITS1 sequences (Tables 1–4). At the genus rank, for instance, higher TIRs were observed for four of the five pure-cultured fungi considered when tested with sequences of the entire ITS region (Tables 2 and 4). ERs were smaller than 1% for all five species when tested with sequences of the entire ITS regions, whereas ERs were greater than 1% for *A. alternata* and *C. cladosporioides* if examined with sequences of the shorter extracted ITS1 (Tables 2 and 4).

Reproducibility

Figure 2 displays relationships of the duplicate environmental measurements to quantify the α diversity metrics

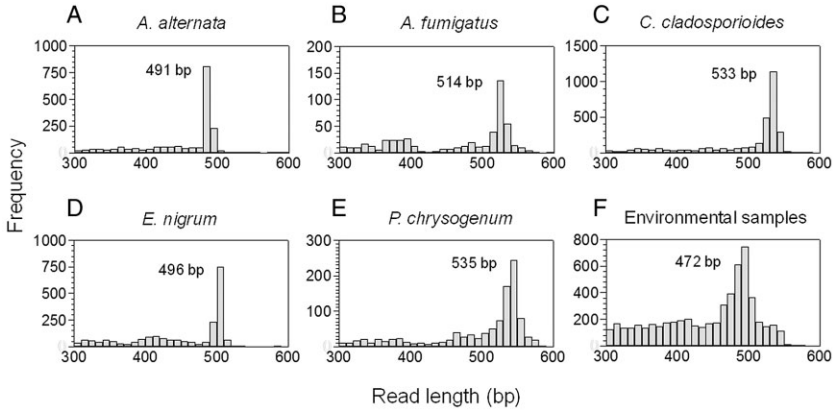

Fig. 1. Histograms of fungal ITS sequences generated by 454 pyrosequencing. The values indicate median lengths. The estimated sizes of the trimmed amplicons for *Alternaria alternata* (A), *Aspergillus fumigatus* (B), *Cladosporium cladosporioides* (C), *Epicoccum nigrum* (D) and *Penicillium chrysogenum* (E) are 566, 593, 547, 540 and 581 bp respectively. (F) Environmental samples.

Table 3. Accuracy of BLASTn-based identification for pure-culture fungi at the species rank using the ITS1 sequences extracted by the ITS1/ITS2 extractor (Nilsson *et al*., 2010).

Tested microorganisms	Assigned taxa[a]	Number of sequences	Percentage of sequences (%)	True / False	TIR (%)[b]	FIR (%)[c]	ER (%)[d]
Alternaria alternata	*Alternaria alternata*	6	0.34	T	0.3	1.2	78.6
	Alternaria spp.	259	14.57	n.d.			
	Alternaria brassicae	3	0.17	F			
	Alternaria citri	5	0.28	F			
	Alternaria compacta	3	0.17	F			
	Alternaria longipes	4	0.22	F			
	Alternaria mali	1	0.06	F			
	Alternaria tenuis	1	0.06	F			
	Aspergillus spp.	1	0.06	F			
	Botryosphaeria berengeriana	1	0.06	F			
	Cladosporium spp.	1	0.06	F			
	Septoria lycopersici	2	0.11	F			
	Unidentified	1491	83.86	n.d.			
	Total	1778	100				
Aspergillus fumigatus	*Aspergillus fumigatus*	12	2.31	T	2.3	0.6	20.0
	Aspergillus spp.	123	23.65	n.d.			
	Neosartorya fischeri	2	0.38	F			
	Simonyella variegata	1	0.19	F			
	Unidentified	382	73.46	n.d.			
	Total	520	100				
Cladosporium cladosporioides	*Cladosporium cladosporioides*	32	1.06	T	1.1	1.2	52.9
	Cladosporium spp.	2876	95.11	n.d.			
	Cladosporium cucumerinum	6	0.20	F			
	Cladosporium tenuissimum	1	0.03	F			
	Curvularia pallescens	1	0.03	F			
	Pleurotus ostreatus	28	0.93	F			
	Unidentified	80	2.65	n.d.			
	Total	3024	100				
Epicoccum nigrum	*Epicoccum nigrum*	2064	99.95	T	100.0	0.0	0.0
	Unidentified	1	0.05	n.d.			
	Total	2065	100				
Penicillium chrysogenum	*Penicillium chrysogenum*	32	3.02	T	3.0	5.6	64.8
	Penicillium spp.	11	1.04	n.d.			
	Penicillium adametzii	8	0.76	F			
	Penicillium brevicompactum	5	0.47	F			
	Penicillium citrinum	1	0.09	F			
	Penicillium commune	1	0.09	F			
	Penicillium coprobium	2	0.19	F			
	Penicillium crustosum	4	0.38	F			
	Penicillium granulatum	12	1.13	F			
	Penicillium oxalicum	23	2.17	F			
	Unidentified	959	90.64	n.d.			
	Total	1058	100				

a. Species name is denoted as 'spp.' in case sequences are ambiguous at the species but not genus rank due to multiple top hits by BLASTn. Sequence is denoted as 'unidentified' in case they are ambiguous at the genus rank due to multiple top hit by BLASTn. Trueness of unidentified sequences is not determined because they might be assigned with a correct taxon along with incorrect taxa.
b. TIR is calculated according to Eq. (1).
c. FIR is calculated according to Eq. (2).
d. ER is calculated according to Eq. (3).
n.d., trueness is not determined due to ambiguous identity at this taxonomic rank.

with and without normalizing the number of the sequences. Without normalizing sequences, a linear relationship between replicates was not observed for the number of OTUs and Chao1 estimator ($r^2 < 0.00001$), but existed for the Shannon indices ($r^2 = 0.7862$). The coefficients of determination increased when the number of the sequences were normalized to 250 reads. Coefficients of determination for the normalized number of OTUs, Chao1 estimator and Shannon index were 0.9162, 0.4299 and 0.9406 respectively. Additionally, the relative standard deviation (RSD)-based reproducibility of the observed OTU counts, Chao1 estimator and Shannon indices decreased when estimates were based on normalized counts (Supporting Information Tables S4 and S5).

Figure 3 illustrates the quantitative statistical relationships between the duplicate environmental measurements for relative abundance. Coefficients of determination (r^2) were 0.9092, 0.8855, 0.9560 and 0.9865 at the species,

Table 4. Accuracy of the BLASTn-based identification for pure culture fungi at the genus rank using the ITS1 sequences extracted by the ITS1/ITS2 extractor (Nilsson *et al.*, 2010).

Tested microorganisms	Assigned taxa[a]	Number of sequences	Percentage of sequences (%)	True/False	TIR (%)[b]	FIR (%)[c]	ER (%)[d]
Alternaria alternata	*Alternaria*	282	15.86	T	15.9	0.3	1.7
	Aspergillus	1	0.06	F			
	Botryosphaeria	1	0.06	F			
	Cladosporium	1	0.06	F			
	Septoria	2	0.11	F			
	Unidentified	1491	83.86	n.d.			
	Total	1778	100				
Aspergillus fumigatus	*Aspergillus*	135	25.96	T	26.4	0.2	0.7
	Neosartorya	2	0.38	T			
	Simonyella	1	0.19	F			
	Unidentified	382	73.46	n.d.			
	Total	520	100				
Cladosporium cladosporioides	*Cladosporium*	2915	96.40	T	96.4	1.0	1.0
	Curvularia	1	0.03	F			
	Pleurotus	28	0.93	F			
	Unidentified	80	2.65	n.d.			
	Total	3024	100				
Epicoccum nigrum	*Epicoccum*	2064	99.95	T	100.0	0.0	0.0
	Unidentified	1	0.05	n.d.			
	Total	2065	100				
Penicillium chrysogenum	*Penicillium*	99	9.36	T	9.4	0.0	0.0
	Unidentified	959	90.64	n.d.			
	Total	1058	100				

a. Sequence is denoted as 'unidentified' in case they are ambiguous at the genus rank due to multiple top hit by BLASTn. Trueness of unidentified sequences is not determined because they might be assigned with a correct taxon along with incorrect taxa.
b. TIR is calculated according to Eq. (1).
c. FIR is calculated according to Eq. (2).
d. ER is calculated according to Eq. (3).
n.d., trueness is not determined owing to ambiguous identity at this taxonomic level.

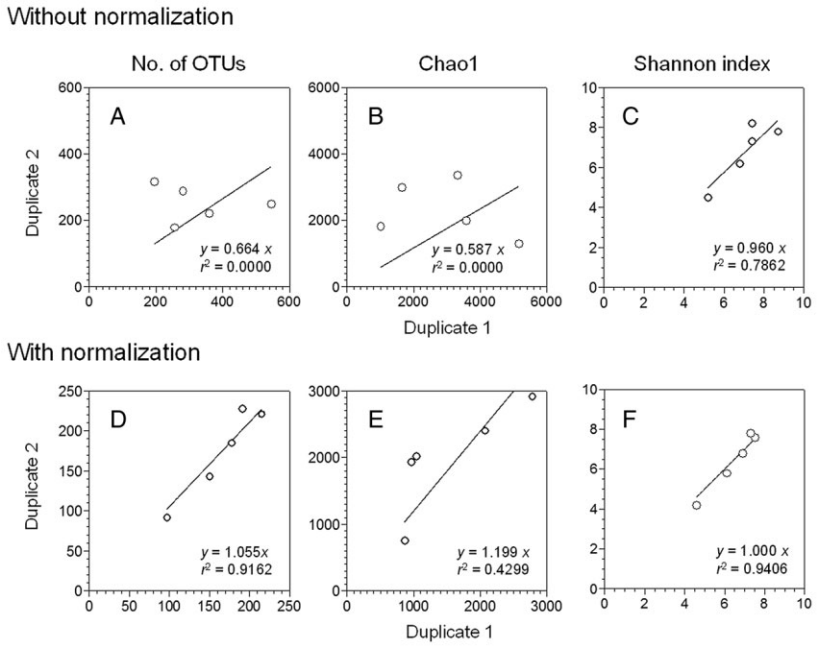

Fig. 2. Relationships between duplicate environmental α diversity metrics including the numbers of the observed 97% OTUs (A, D), Chao1 estimator (B, E) and Shannon index (C, F). The α diversity metrics were calculated with and without normalizing the number of sequence reads.

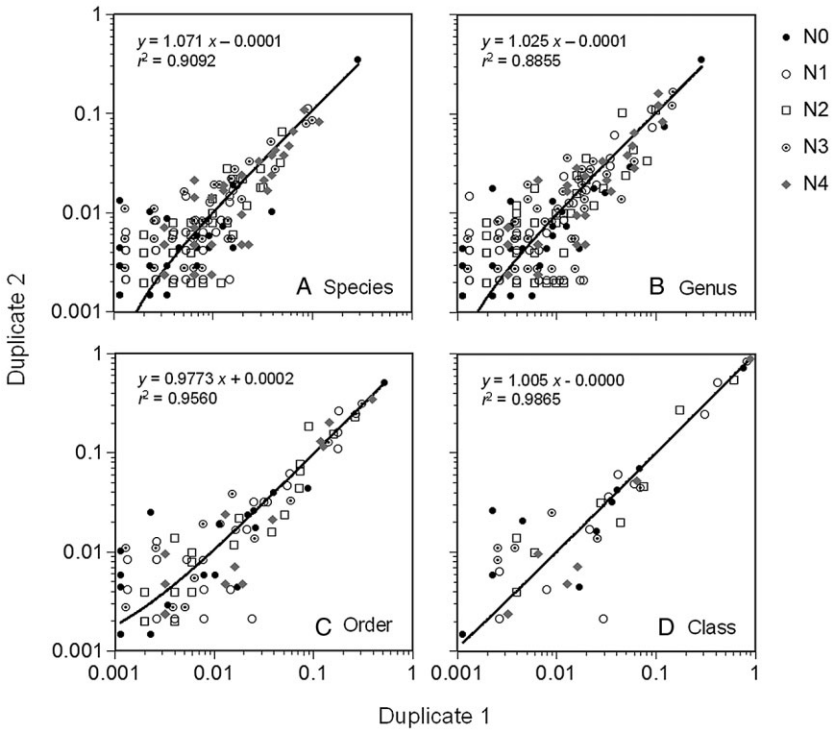

Fig. 3. Relationships between duplicate environmental relative abundances of fungal species (A), genera (B), order (C) and class (D). Each data point indicates relative abundances of different unique taxa found in each library. Linear regression lines are calculated with all the libraries combined.

genus, order and class rank respectively. Results from Fig. 3 indicated an increase in r^2 value as taxonomic rank decreases from species to class, and also suggest that reproducibility decreases as a function of sequence abundance. Figure 4 statistically shows the reproducibility of the relative abundances of the observed taxa as a function of the number of sequences per taxa. Larger numbers of the sequences produced more reproducible relative abundance measurements (smaller RSD). The observed tendency follows a theoretical estimate of standard error (SE) calculation $\left(\mathrm{SE} = 1/\sqrt{N}\right)$.

Regarding β diversity, the principle coordinate analysis plot in Fig. 5 demonstrates the reproducibility of environmental fungal composition by ITS amplicon sequencing.

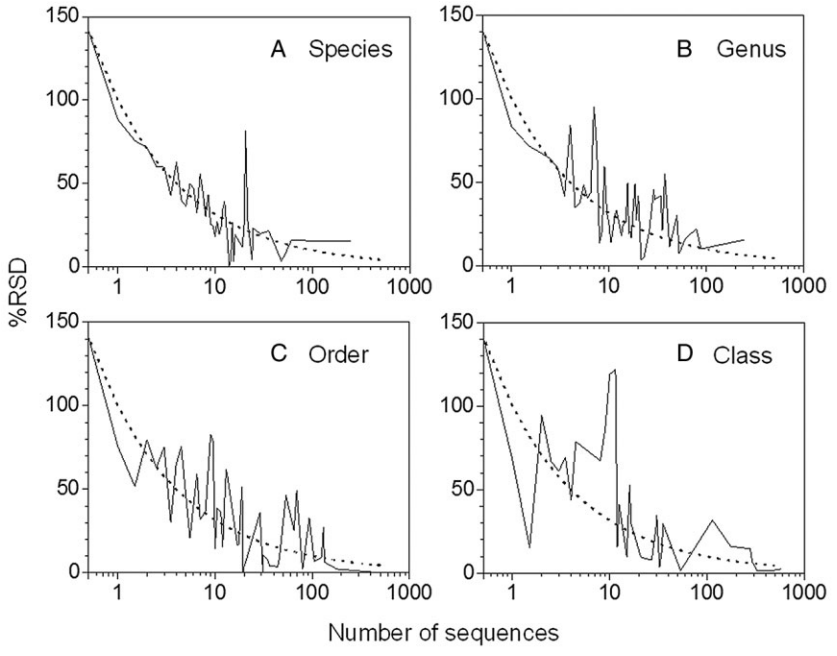

Fig. 4. The effect of number of sequences on relative abundance reproducibility for species (A), genera (B), order (C) and class (D). Reproducibility is calculated by Eq. (4) and reported in percentage relative standard deviation (%RSD). If multiple taxa have the same number of the sequences with different %RSDs, a mean of %RSDs is taken to represent %RSD for this number of the sequences. The dotted lines indicate a theoretical estimate based on SE calculation $\left(\mathrm{SE} = 1/\sqrt{N}\right)$.

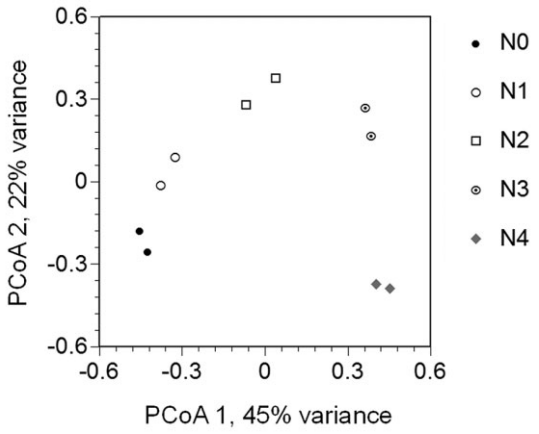

Fig. 5. PCoA plot demonstrating the population similarity of duplicates from five environmental samples that correspond to five different size fractions of an aerosol sample. OTUs defined by 97% sequence similarity were used for calculating the non-phylogenetic Morisita Horn distance.

Analysis of similarity testing indicates that the variability within a sample's replicates is significantly smaller that the differences in microbial populations across the five different airborne particle size samples (ANOSIM, $P < 0.005$).

Discussion

The quantitative analysis of microbial populations via next-generation DNA-sequencing has biases rooted in the biology of the target organisms, the sequencing technology and fungal ITS databases (Nilsson *et al.*, 2006; Herrera *et al.*, 2009; Amend *et al.*, 2010a; Bellemain *et al.*, 2010; Dannemiller *et al.*, 2013). Although some of these biases are unavoidable, we expect that more reliable and realistic assessments of microbial communities will be achieved with well-defined analytical parameters. Here, we focus on evaluating accuracy and reproducibility of fungal identification and quantification for the sequences generated by ITS amplicon sequencing.

Accuracy

We found large fractions of the sequences assigned to correct taxa by BLASTn and through using a database that contained only named fungal ITS sequences (Nilsson *et al.*, 2009). However, ambiguity caused by multiple tying top hits is a significant concern, especially for identifications at the species level (Supporting Information Fig. S1–5). For *A. fumigatus*, identity was sometimes ambiguous with its teleomorphic relative *N. fischeri* (Table 1 and Supporting Information Fig. S2). *A. fumigatus* and *N. fischeri* belong to the *Aspergillus* section *Fumigati*, and their genetic variations are extremely low (Rydholm *et al.*, 2006). Thus, the use of the ITS regions is not recommended to distinguish these sibling species (Balajee *et al.*, 2007). For *A. alternata*, 96.7% of the sequences were ambiguous due to tying top hits with *Alternaria arborescens*, *Alternaria longipes*, *Alternaria mali* and *Alternaria tenuissima* (Aa5 in Supporting Information Fig. S1), most of which are considered sibling species (Pryor and Gilbertson, 2000). For *C. cladosporioides*, 98.8% of the sequences were ambiguous due to tying top hits with *Cladosporium cucumerinum* (Cc5 in Supporting Information Fig. S3). *C. cucumerinum* is a close relative of *C. cladosporioides* (Braun *et al.*, 2003). Of the five species considered, only *E. nigrum* had a TIR near 100% at the species level. Thus, we note that additional scrutiny, including use of phylogenetic-based approaches (Porter and Golding, 2011), may be applied to the annotation process if species of interest have sibling species that are indistinguishable by BLASTn. However, inherent limitations, such as ITS gene conservation among sibling species, may limit the resolution of these additional approaches. Additional DNA markers, such as β-tubulin locus for fungi *Aspergillus* section *Fumigati* (Balajee *et al.*, 2007), needs to be further sequenced to differentiate these closely related fungal taxa.

The results were more accurate when identification was made at the genus rank (Table 2). ERs were reduced to less than 0.8%. TIRs were increased to > 99% except for *A. fumigatus* (44.9%) and *P. chrysogenum* (11.4%). Of the *Aspergillus* genus, 11.5% of sequences were identified as *Neosartorya* and the remainders were mostly ambiguous due to *Aspergillus* and *Neosartorya* top tying hits. For the *Penicillium* genus, 88.5% of the *P. chrysogenum* sequences were ambiguous (Table 2). Part of the *P. chrysogenum* ambiguity was due to tying top hits with *Eupenicillium* (Supporting Information Fig. S5), which is a teleomorphic state of the genera *Penicillium*. Current convention of fungal taxonomy that permit dual nomenclatures for a genetically identical microorganism is a cause of ambiguity in molecular-based identifications.

The fungal ITS region consists of ITS1, 5.8S and ITS2, and removal of 5.8S sequences has been suggested as an improvement for fungal identification accuracy (Nilsson *et al.*, 2010). This study compared identification accuracy with sequences of the whole ITS and the extracted ITS1 regions to determine the potential impacts caused by inclusion of 5.8S sequences. The results showed that identification accuracy was generally better with the whole ITS sequences than with the shorter extracted ITS1 sequences (Tables 1–4), indicating no substantive impact caused by inclusion of 5.8S sequences when reads were long enough to capture the entire ITS region. As shown in Fig. 1, most reads that we produced by the 454 GS FLX were long enough to cover the entire ITS region, i.e., 491–535 bp length. As previously concluded, 5.8S sequences might be problematic for shorter sequence

reads (~ 250 bp) as BLAST alignment might skew with increased relative fractions of 5.8S sequences in query sequences (Nilsson *et al*., 2010). Accuracy of future ITS-based fungal ecological studies may also be improved by development of more accurately curated databases. For instance, future databases should avoid dual nomenclature of genetically identical fungal pleomorphs such as *A. fumigatus* and *N. fumigata* (O'Gorman *et al*., 2009), or future annotation programmes should allow for recognition of fungal pleomorphs (Taylor, 2011). Another factor that can cause ambiguous and/or false identification by BLASTn is incorrectly annotated sequences uploaded to the database. It is known that up to 20% of fungal sequences in databases are incorrectly annotated at the species rank (Bridge *et al*., 2003; Nilsson *et al*., 2006). However, efforts are under way to improve database quality (Abarenkov *et al*., 2010). Beyond databases, factors such as errors that are associated with all DNA sequencing platforms, the production of different sequence lengths and intergenome variability in ITS sequences (Simon and Weiss, 2008), all dictate that identification at the genus rank is more robust than at the species rank.

Reproducibility

As previously reported (Soetaert and Heip, 1990; Gihring *et al*., 2012), α diversity metrics such as Chao1 estimators and Shannon indices can vary with sample size. Normalizing the number of the sequence reads prior to OTU clustering may help to circumvent this issue. Our results indicated that sequence count normalization substantially improved reproducibility. The results were particularly reproducible for the sequence normalized numbers of the observed OTUs ($r^2 = 0.9162$) and Shannon indices ($r^2 = 0.9406$) (Fig. 2).

Our duplicate environmental measurements suggested that the results of taxonomic compositions were reproducible (Fig. 3). In particular, the results were more reproducible for taxa with a greater number of observed sequences (Fig. 4). The observed tendency follows a theoretical estimate based on SE calculation $\left(\mathrm{SE} = 1/\sqrt{N}\right)$, suggesting that variability of the relative abundance measurements was attributable to a random sampling process. Technological advances allowing for increased number of sequences per sample for lower cost, time and effort will improve reproducibility.

Finally, we note that the accuracy and reproducibility results reported here are specific for primer choice. ITS1F-ITS4 primers were used to amplify fungal sequences, which are widely used in sequencing-based fungal ecological studies (Amend *et al*., 2010b; Yamamoto *et al*., 2012). There are several advantages of selecting ITS1F-ITS4. First, ITS1F allows for fungal specific polymerase chain reaction (PCR) without amplifying plant sequences (Bellemain *et al*., 2010). Among many forward universal fungal primers, only ITS1F is thought to be fungal specific (Bellemain *et al*., 2010). Second, ITS1F-ITS4 allows for amplification of the entire ITS region. Our results indicate that 454 pyrosequencing technology is capable of sequencing approximately 90% of the entire ITS regions (Fig. 1). Third, we expect that ITS1F-ITS4 did not cause large omissions of commonly found airborne fungi. In this study, more than 600 species were identified (Supporting Information Table S4), and the α diversity (Supporting Information Tables S2 and S3) was as high or higher than previously reported fungal ecology studies (O'Brien *et al*., 2005; Buée *et al*., 2009; Fröhlich-Nowoisky *et al*., 2009). For these reasons, we conclude that ITS1F-ITS4 primers are preferable for fungal diversity studies that utilize the next-generation DNA sequencing technology that can produce long sequence reads (~ 500 bp). Important examples of such platforms are the 454 GS FLX Titanium Platform, and the latest version of Illumina sequencers that can produce longer sequence reads (2×250 bp moving to 2×300 bp).

Conclusions

Here we investigated the accuracy of identification and reproducibility of diversity measurements associated with amplicon sequencing of the fungal ITS region. The BLASTn-based method generally produced accurate fungal identification when the identification was made at the genus level. Care must be taken when the species of interest have dual nomenclatures and/or sibling taxa that have indistinguishable ITS sequences. Our duplicate environmental measurements showed that amplicon sequencing was reproducible especially with increasing sequencing depth for specific taxa. Alpha diversity metrics were significantly more reproducible between replicates when sequence quantity was normalized prior to rarefaction. This more comprehensive understanding of accuracy and reproducibility should improve experimental design and data interpretation in future fungal taxonomy studies that utilize next-generation DNA sequencing technology.

Experimental procedures

Sample preparation

Annotation accuracy was determined by sequencing and annotating ITS reads derived from pure fungal cultures. Five medically important allergenic fungi including *Alternaria alternata* (PEM 01043), *Aspergillus fumigatus* (ATCC 34506), *Cladosporium cladosporioides* (ATCC 16022), *Epicoccum nigrum* (TU BL-3) and *Penicillium chrysogenum* (CAES PC-1) were analysed. These fungi were cultured on malt extract agar at room temperature for 1 month. Approximately 10^9 conidia of

A. fumigatus, *C. cladosporioides* and *P. chrysogenum* each were harvested by cotton swabs. For *A. alternata* and *E. nigrum*, tissues were isolated from a 4 cm^2 area of the agar media (Yamamoto *et al.*, 2011).

Reproducibility was determined by comparing the annotations of replicated taxonomic libraries produced from environmental aerosol samples. Five duplicate environmental fungal samples were obtained in a previous aerosol sampling field campaign (Yamamoto *et al.*, 2012). Air samples were collected on glass fibre substrates (New Star Environmental, Roswell, GA, USA) using duplicate multi-stage, non-viable Andersen samplers (New Star Environmental) operated at a flow rate of 28.3 l min^{-1}. The samples were collected on five different stages of the impactor corresponding to aerodynamic diameters of 2.1–3.3, 3.3–4.7, 4.7–5.8, 5.8–9.0 and > 9.0 μm (denoted as N4, N3, N2, N1 and N0 respectively). These samples were collected from October 15 to November 12 in 2009 on the rooftop of a five-storey building in New Haven, CT, USA. The duplicate air samplers were located side by side, and the two separate filters were recovered from each stage of the duplicate air samplers, processed and analysed by the exact same protocol.

DNA extraction

For DNA extraction, a one-eighth section of each filter (~ 6.3 cm^2) used in the Andersen sampler was analysed. DNA extraction was performed with the PowerMax® Soil DNA Isolation Kit (Mobio Laboratory, Carlsbad, CA, USA). The filter section was enclosed in a 2 ml microcentrifuge tube along with the kit's power beads (1.0 g), kit's extraction solution (750 μl), 0.1 mm diameter glass beads (300 mg) and 0.5 mm diameter glass beads (100 mg). The samples were then homogenized for 5 min by a bead beater (Model 607; BioSpec Products, Bartlesville, OK, USA). After bead beating, the extracted DNA were purified in accordance with the kit protocol and eluted with 50 μl of TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH = 8.0).

Pyrosequencing of the ITS region of rDNA

Sequencing was performed with universal fungal primers ITS1F (5′-Adaptor A-Key-MID-CTTGGTCATTTAGAGGAAG TAA-3′) and ITS4 (5′-Adapter B-Key-TCCTCCGCTT ATTGATATGC-3′) (Larena *et al.*, 1999; Manter and Vivanco, 2007), where Key represents bar-coding sequences, and MIDs are the sequences for multiplex identifiers. For PCR amplification, a 50 μL reaction mixture containing the template DNA (2 μL sample extract), 1 × PCR Master Mix (Roche Applied Science, Madison, WI, USA) and 0.3 μM of each primer was used. Cycling conditions were: 95°C for 5 min of initial denaturation and 36 cycles of 95°C for 30 s of dissociation, 55°C for 30 s of annealing, 72°C for 1 min for each cycle extension and 72°C for 10 min of final extension step. PCR amplicons were purified and concentrations normalized with SequalPrep™ Normalization Plate (96 well) Kit (Invitrogen, Grand Island, NY, USA). The normalized amplicons were pooled and then further purified by agarose gel extraction (QIAquick Gel Extraction Kit, Qiagen, Valencia, CA, USA). Sequencing of amplicons was performed using the 454 GS FLX Titanium Platform (454 Life Sciences, Branford, CT, USA) at the Yale University Center for Genome Analysis. Raw sequencing data have been deposited in the MG-RAST (metagenomics.anl.gov) archive under accession number 447 6444.3.

DNA sequence processing and analyses

Sequence FASTA and QUAL files were extracted from the machine output file, and then trimmed, and parsed by sequence tag using the Ribosomal Database Project pyrosequencing pipeline (Cole *et al.*, 2005). Trimming removed primers, sequences with one or more undefined base, sequences below a minimum machine quality score of 20, and sequences with a trimmed read length of fewer than 300 base pairs. Trimmed sequences were taxonomically placed using BLASTn ver. 2.2.19 (Altschul *et al.*, 1990) and a maximum E-score of 10^{-5} using a database that contained only named fungal ITS sequences identified down to the species rank (Nilsson *et al.*, 2009). Taxonomic placement was assigned to the top hit using FHiTINGS ver. 1.1 (Dannemiller *et al.*, 2013). In cases of multiple top hits with tying E-values, identity was assigned to the taxonomic rank that could be unambiguously determined in accordance with the least common ancestor (LCA) method (Amend *et al.*, 2010b). For instance, a sequence with tying top hits '*Alternaria alternata*' and '*Alternaria tenuissima*' was ambiguous at the species rank but reported as *Alternaria* at the genus rank. The sequences were not denoised at this stage because the homopolymer insertions and deletions that are the major causes of the noise are not primary concerns for taxonomic assignment (Reeder and Knight, 2010).

For the pure-cultured fungi, accuracy of taxonomic assignment was also tested with shorter ITS1 sequences extracted from the original ITS sequences. To extract ITS1 sequences, the ITS1/ITS2 extractor for fungal ITS sequences (Nilsson *et al.*, 2010) was used. Taxonomic assignment for the extracted ITS1 sequences was done with the same methodologies used for the aforementioned full-length ITS sequences.

Prior to calculation of diversity metrics, the sequence libraries for each sample were split within QIIME 1.3.0 (Caporaso *et al.*, 2010) and denoised using the denoise_wrapper.py script (Quince *et al.*, 2009; 2011; Reeder and Knight, 2009). Following denoising, OTUs were assigned using UCLUST 1.2.22 (Edgar, 2010) implemented within QIIME 1.3.0 based on 97% sequence similarity to characterize reproducibility of fungal diversity metrics (O'Brien *et al.*, 2005; Buée *et al.*, 2009; Kunin *et al.*, 2010; Yamamoto *et al.*, 2012), and the samples were rarified through the QIIME α diversity workflow script. These OTUs were then used to calculate Chao1 and Shannon indices. Singletons, interpreted to mean sequences that were not grouped into clusters, were included for calculating diversity parameters (e.g. O'Brien *et al.*, 2005; Buée *et al.*, 2009; Yamamoto *et al.*, 2012). The α diversity metrics were calculated both with and without normalizing the number of sequence reads among the libraries. To calculate the α diversity with a normalized number of sequences, 250 sequences were randomly subsampled. For the β diversity analysis, non-phylogenetic Morisita Horn distance among the 10 populations (five samples, duplicated) was used in a principal coordinate analysis (PCoA) without normalizing for read number. The ANOSIM program in QIIME 1.5.0 was used to calculate statistical significance based on *P* values.

Statistical analysis

This study evaluated both the accuracy of the BLASTn-based taxonomic assignments for the pure culture fungal ITS sequences generated by 454 pyrosequencing, and the reproducibility of characterizing fungal communities. To test accuracy, we analysed the ITS sequences from the pure fungal cultures with known identities and compared them with the BLASTn outputs with assignment by the LCA method. The sequences were classified based on types of identification: (i) true, (ii) false and (iii) ambiguous. The sequences that were unambiguously assigned to a correct fungal taxon were defined as true. For example, if an *A. alternata* sequence was unambiguously assigned to *A. alternata*, it was classified as true. The sequences that were unambiguously assigned to an incorrect fungal taxon were defined as false. For example, if an *A. alternata* sequence was unambiguously assigned to an incorrect fungal taxon such as *A. arborescens*, it was classified as false. The sequences that were ambiguously assigned to multiple taxa with tying top hits were defined as ambiguous sequences. For example, if an *A. alternata* sequence was assigned to *A. alternata* and *A. tenuissima* with tying top hits, it was classified as ambiguous at the rank of species.

The same classification methodology was applied to genus-rank identifications. For example, if an *A. alternata* sequence was unambiguously assigned to *Alternaria*, it was classified as true at the genus rank, even though this sequence may have been assigned to incorrect taxa such as *A. tenuissima* or *A. arborescens* at the species rank. Some fungal species belonging to the genera *Aspergillus* and *Penicillium* have teleomorphic states named as the *Neosartorya* and *Eupenicillium* genera respectively (O'Gorman *et al*., 2009; Houbraken and Samson, 2011). In the current study, taxonomic assignments were considered to be true if sequences from *A. fumigatus* were assigned to the teleomorph *Neosartorya fumigata* (O'Gorman *et al*., 2009). At the genus rank, assignments of *A. fumigatus* sequences to any *Neosartorya* spp. were considered correct. For instance, if an *A. fumigatus* sequence was assigned to *N. fischeri*, it was classified as a false sequence at the species level, but reported as a true sequence at the genus level.

For the pure culture data, TIR, FIR and ER were calculated. TIR, a measure true identification frequency, was calculated as follows:

$$\mathrm{TIR} = N_{\mathrm{true}}/(N_{\mathrm{true}} + N_{\mathrm{false}} + N_{\mathrm{ambiguous}}) = N_{\mathrm{true}}/N_{\mathrm{total}} \quad (1)$$

where N_{true}, N_{false} and $N_{\mathrm{ambiguous}}$ are numbers of sequences with true, false and ambiguous identification, respectively, and N_{total} is the total number of sequences. FIR, a measure false identification frequency, was calculated as follows:

$$\mathrm{FIR} = N_{\mathrm{false}}/(N_{\mathrm{true}} + N_{\mathrm{false}} + N_{\mathrm{ambiguous}}) = N_{\mathrm{false}}/N_{\mathrm{total}} \quad (2)$$

ER, the percentage of non-ambiguously identified sequences with a false identification, was calculated as follows:

$$\mathrm{ER} = N_{\mathrm{false}}/(N_{\mathrm{true}} + N_{\mathrm{false}}) \quad (3)$$

The number of ambiguous sequence numbers are excluded in the ER calculations because the ER measures the false identification in terms of unambiguous identifications. Including ambiguous identification in this value would artificially deflate the ER. TIR, FIR and ER were calculated at both species and genus ranks.

Reproducibility is defined here as variations and errors derived from sample and sequence processing (Hospodsky *et al*., 2010). This study analysed several duplicate environmental samples to evaluate the reproducibility. The reproducibility was characterized at the species, genus, class and order ranks. The relative abundance was defined as a ratio of number of the sequences belonging to a particular fungal taxon to total number of the sequence reads. The equation is given by:

$$x_{\mathrm{i}} = N_{\mathrm{i}}/N_{\mathrm{total}} \quad (4)$$

where x_{i} is relative abundance of a fungal taxon i, and N_{i} is number of sequences of a fungal taxon i. The reproducibility is expressed by RSD of duplicate measurements:

$$\mathrm{RSD}_{\mathrm{i}} = \frac{1}{\sqrt{2}} \frac{|x_{\mathrm{i,dup1}} - x_{\mathrm{i,dup2}}|}{x_{\mathrm{i,average}}} \quad (5)$$

where $x_{\mathrm{i,dup1}}$ and $x_{\mathrm{i,dup2}}$ are relative abundances of a fungal taxon i in duplicates 1 and 2, respectively, and $x_{\mathrm{i,average}}$ is a mean of relative abundances of a fungal taxon i in duplicates 1 and 2.

Acknowledgements

Primary funding for this project was provided by the Alfred P. Sloan Foundation. Additionally; this work was supported by Research Settlement Fund for the new faculty of SNU (N.Y.).

References

Abarenkov, K., Henrik Nilsson, R., Larsson, K.H., Alexander, I.J., Eberhardt, U., Erland, S., *et al*. (2010) The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytol* **186:** 281–285.

Adams, R.I., Miletto, M., Taylor, J.W., and Bruns, T.D. (2013) Dispersal in microbes: Fungi in indoor air are dominated by outdoor air and show dispersal limitation at short distances. *ISME J* **7:** 1262–1263.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Amend, A.S., Seifert, K.A., and Bruns, T.D. (2010a) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol Ecol* **19:** 5555–5565.

Amend, A.S., Seifert, K.A., Samson, R., and Bruns, T.D. (2010b) Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *Proc Natl Acad Sci USA* **107:** 13748–13753.

Balajee, S.A., Houbraken, J., Verweij, P.E., Hong, S.B., Yaghuchi, T., Varga, J., and Samson, R.A. (2007) Aspergillus species identification in the clinical setting. *Stud Mycol* **59:** 39–46.

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

Blaalid, R., Carlsen, T., Kumar, S., Halvorsen, R., Ugland, K.I., Fontana, G., and Kauserud, H. (2012) Changes in the root-associated fungal communities along a primary succession gradient analysed by 454 pyrosequencing. *Mol Ecol* **21:** 1897–1908.

Braun, U., Crous, P.W., Dugan, F., Groenewald, J.Z., and De Hoog, G.S. (2003) Phylogeny and taxonomy of *Cladosporium*-like hyphomycetes, including *Davidiella* gen. nov., the teleomorph of *Cladosporium s. str. Mycol Prog* **2:** 3–18.

Bridge, P.D., Roberts, P.J., Spooner, B.M., and Panchal, G. (2003) On the unreliability of published DNA sequences. *New Phytol* **160:** 43–48.

Buée, M., Reich, M., Murat, C., Morin, E., Nilsson, R.H., Uroz, S., and Martin, F. (2009) 454 pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytol* **184:** 449–456.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335–336.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., *et al*. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33:** D294–D296.

Crameri, R., Weichel, M., Fluckiger, S., Glaser, A.G., and Rhyner, C. (2006) Fungal allergies: A yet unsolved problem. *Chem Immunol Allergy* **91:** 121–133.

Dannemiller, K., Reeves, D., Bibby, K., Yamamoto, N., and Peccia, J. (2013) Fungal high-throughput taxonomic identification tool for use with next-generation sequencing (FHiTINGS). *J Basic Microbiol*. doi:10.1002/jobm.201200507.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.

Ege, M.J., Mayer, M., Normand, A.C., Genuneit, J., Cookson, W.O.C.M., Braun-Fahrländer, C., *et al*. (2011) Exposure to environmental microorganisms and childhood asthma. *New Engl J Med* **364:** 701–709.

Fröhlich-Nowoisky, J., Pickersgill, D.A., Després, V.R., and Pöschl, U. (2009) High diversity of fungi in air particulate matter. *Proc Natl Acad Sci USA* **106:** 12814–12819.

Gihring, T.M., Green, S.J., and Schadt, C.W. (2012) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol* **14:** 285–290.

Herrera, M.L., Vallor, A.C., Gelfond, J.A., Patterson, T.F., and Wickes, B.L. (2009) Strain-dependent variation in 18S ribosomal DNA copy numbers in *Aspergillus fumigatus*. *J Clin Microbiol* **47:** 1325–1332.

Hodge, A., Campbell, C.D., and Fitter, A.H. (2001) An arbuscular mycorrhizal fungus accelerates decomposition and acquires nitrogen directly from organic material. *Nature* **413:** 297–299.

Hospodsky, D., Yamamoto, N., and Peccia, J. (2010) Accuracy, precision, and method detection limits of quantitative PCR for airborne bacteria and fungi. *Appl Environ Microbiol* **76:** 7004–7012.

Houbraken, J., and Samson, R.A. (2011) Phylogeny of *Penicillium* and the segregation of *Trichocomaceae* into three families. *Stud Mycol* **70:** 1–51.

Iliev, I.D., Funari, V.A., Taylor, K.D., Nguyen, Q., Reyes, C.N., Strom, S.P., *et al*. (2012) Interactions between commensal fungi and the C-type lectin receptor dectin-1 influence colitis. *Science* **336:** 1314–1317.

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

Larena, I., Salazar, O., Gonzalez, V., Julian, M.C., and Rubio, V. (1999) Design of a primer for ribosomal DNA internal transcribed spacer with enhanced specificity for ascomycetes. *J Biotechnol* **75:** 187–194.

Manter, D.K., and Vivanco, J.M. (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. *J Microbiol Methods* **71:** 7–14.

Nierman, W.C., Pain, A., Anderson, M.J., Wortman, J.R., Kim, H.S., Arroyo, J., *et al*. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438:** 1151–1156.

Nilsson, R.H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.H., and Kõljalg, U. (2006) Taxonomic reliability of DNA sequences in public sequence databases: A fungal perspective. *PLoS ONE* **1:** e59.

Nilsson, R.H., Bok, G., Ryberg, M., Kristiansson, E., and Hallenberg, N. (2009) A software pipeline for processing and identification of fungal ITS sequences. *Source Code Biol Med* **4:** 1.

Nilsson, R.H., Veldre, V., Hartmann, M., Unterseher, M., Amend, A., Bergsten, J., *et al*. (2010) An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecol* **3:** 284–287.

O'Brien, H.E., Parrent, J.L., Jackson, J.A., Moncalvo, J.M., and Vilgalys, R. (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Appl Environ Microbiol* **71:** 5544–5550.

O'Gorman, C.M., Fuller, H., and Dyer, P.S. (2009) Discovery of a sexual cycle in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Nature* **457:** 471–474.

Pitman, S.K., Drew, R.H., and Perfect, J.R. (2011) Addressing current medical needs in invasive fungal infection prevention and treatment with new antifungal agents, strategies and formulations. *Expert Opin Emerg Drugs* **16:** 559–586.

Porter, T.M., and Golding, G.B. (2011) Are similarity- or phylogeny-based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? *New Phytol* **192:** 775–782.

Pryor, B.M., and Gilbertson, R.L. (2000) Molecular phylogenetic relationships amongst *Alternaria* species and related fungi based upon analysis of nuclear ITS and mt SSU rDNA sequences. *Mycol Res* **104:** 1312–1321.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6:** 639–641.

Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12:** 38.

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

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

Rydholm, C., Szakacs, G., and Lutzoni, F. (2006) Low genetic variation and no detectable population structure in aspergillus fumigatus compared to closely related Neosartorya species. *Eukaryot Cell* **5:** 650–657.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., and Chen, W. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. *Proc Natl Acad Sci USA* **109:** 6241–6246.

Simon, U.K., and Weiss, M. (2008) Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Mol Biol Evol* **25:** 2251–2254.

Soetaert, K., and Heip, C. (1990) Sample-size dependence of diversity indexes and the determination of sufficient sample-size in a high-diversity deep-sea environment. *Mar Ecol Prog Ser* **59:** 305–307.

Taylor, J.W. (2011) One fungus = one name: DNA and fungal nomenclature twenty years after PCR. *IMA Fungus* **2:** 113–120.

Yamamoto, N., Shendell, D.G., and Peccia, J. (2011) Assessing allergenic fungi in house dust by floor wipe sampling and quantitative PCR. *Indoor Air* **21:** 521–530.

Yamamoto, N., Bibby, K., Qian, J., Hospodsky, D., Rismani-Yazdi, H., Nazaroff, W.W., and Peccia, J. (2012) Particle-size distributions and seasonal diversity of allergenic and pathogenic fungi in outdoor air. *ISME J* **6:** 1801–1811.

Zhu, Y.G., and Miller, R.M. (2003) Carbon cycling by arbuscular mycorrhizal fungi in soil-plant systems. *Trends Plant Sci* **8:** 407–409.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Fractions of *Alternaria alternata* sequences by types of taxonomic placements. The full-length ITS sequences were used for BLSATn. The categories labelled Aa1 through Aa10 each correspond to a fraction of the total *A. alternata* sequences and the taxa that represent top BLASTn hit(s) to this fraction are listed in the adjacent map. Nearly all (99.6%) of the sequences were assigned to *A. alternata* either solely or ambiguously (Aa1-Aa6). At the genus rank, 99.9% of the sequences were assigned correctly (Aa1-Aa8). The truly identified sequences that were assigned solely to *A. alternata* constitute 2.1% of the sequences (Aa1).

Fig. S2. Fractions of *Aspergillus fumigatus* sequences by types of taxonomic placements. The full-length ITS sequences were used for BLASTn. The categories labelled Af1 through Af19 each correspond to a fraction of the total *A. fumigatus* sequences and the taxa that represent top BLASTn hit(s) to this fraction are listed in the adjacent map. There were 76.6% of the sequences assigned to *A. fumigatus* either solely or ambiguously (Af1-Af9). At the genus rank, 99.8% of the sequences were assigned correctly either to *Aspergillus* or *Neosartorya* (Af1-Af16, Af18 and Af19). The truly identified sequences that were assigned solely to *A. fumigatus* constitute 30.9% of the sequences (Af1).

Fig. S3. Fractions of *Cladosporium cladosporioides* sequences by types of taxonomic placements. The full-length ITS sequences were used for BLASTn. The categories labelled Cc1 through Cc9 each correspond to a fraction of the total *Cladosporium cladosporioides* sequences and the taxa that represent top BLASTn hit(s) to this fraction are listed in the adjacent map. There were 99.2% of the sequences assigned to *C. cladosporioides* either solely or ambiguously (Cc1-Cc6). At the genus rank, 99.9% of the sequences were assigned correctly (Cc1-Cc7). The truly identified sequences that were assigned solely to *C. cladosporioides* constitute 0.3% of the sequences (Cc1). There were 98.8% of the sequences ambiguous with *Cladosporium cucumerinum* (Cc5).

Fig. S4. Fractions of *Epicoccum nigrum* sequences by types of taxonomic placements. The full-length ITS sequences were used for BLASTn. The categories labelled En1 through En3 each correspond to a fraction of the total *Epicoccum nigrum* sequences and the taxa that represent top BLASTn hit(s) to this fraction are listed in the adjacent map. The truly identified sequences that were assigned solely to *E. nigrum* constitute 99.9% of the sequences (En1).

Fig. S5. Fractions of *Penicillium chrysogenum* sequences by types of taxonomic placements. The full-length ITS sequences were used for BLASTn. The categories labelled Pc1 through Pc48 each correspond to a fraction of the total *P. chrysogenum* sequences and the taxa that represent top BLASTn hit(s) to this fraction are listed in the adjacent map. 95.8% of the sequences were assigned to *P. chrysogenum* either solely or ambiguously (Pc1-Pc29). At the genus rank, 99.8% of the sequences were assigned correctly to either *Penicillium* or *Eupenicillium* (Pc1-Pc29, Pc31-Pc45, Pc47 and Pc48). The truly identified sequences that were assigned solely to *P. chrysogenum* constitute 8.2% of the sequences (Pc1).

Table S1. Summary statistics of fungal ITS sequences obtained with ITS1F and ITS4 primers by 454 pyrosequencing.

Table S2. Diversity parameters for sampled atmospheric fungi based on 97% OTU similarity without normalizing the number of sequence reads among the libraries.

Table S3. Diversity parameters for sampled atmospheric fungi based on 97% OTU similarity with normalizing the number of sequences to 250 reads for each library.

Table S4. Number of fungal taxa identified by 454 pyrosequencing.

Table S5. Number of the ITS sequences and identified fungal species. Genus names are denoted as ambiguous if multiple genera are assigned by BLASTn. Species names are denoted as spp. if multiple species are assigned by BLASTn.