Fungal identification biases in microbiome projects

Leho Tedersoo¹* and Björn Lindahl²
¹Natural History Museum, University of Tartu, 14a Ravila, Tartu 50411, Estonia.
²Department of Soil and Environment, Swedish University of Agricultural Sciences, Box 7014, Uppsala, SE 75007, Sweden

SUMMARY

Fungi are the key players in ecosystems as well as in plant and human health. High-throughput molecular identification of fungi has greatly progressed our understanding about the diversity of mutualists, saprotrophs, and pathogens. We argue that the methods promoted by the microbiome consortia are suboptimal for detection of the most important fungal pathogens and ecologically important degraders. We recommend several sets of optimized primers for analysis of fungi or all eukaryote groups based on either short or long amplicons that cover the ITS region as well as part of 18S and 28S rDNA.

Introduction

Microbes play a key role in nutrient cycling and disease, yet the diversity and community composition of prokaryotes, protists, and fungi are poorly known compared with plants and animals. Large-scale initiatives such as the Human Microbiome Project (Turnbaugh et al., 2007), the Earth Microbiome Project (Gilbert et al., 2010), and the Microbial Earth Project [http://genome.jgi-psf.org/programs/bacteria–archaea/MEP/index.jsf] serve to provide deeper insights into the diversity and functions of microorganisms. These consortia have established standardized laboratory protocols for marker-based identification and metagenomics. While the standards for prokaryote identification have been widely accepted, those for eukaryotes are subject to debate (Lindahl et al., 2013; Geisen et al., 2015). Although the whole Internal Transcribed Spacer (ITS) region has been nominated a formal barcode for fungi (Schoch et al., 2012; Bates et al., 2013), there is no consensus about the selection of ITS subregions (ITS1 or ITS2) and primers for identification based on high-throughput sequencing (HTS) technologies (Fig. 1). The majority of mycology and microbiology laboratories focus on the ITS1 subregion, but the microbiome projects utilize and endorse the primers ITS1F (Gardes and Bruns, 1993) and ITS2 (White et al., 1990), which target the ITS1 subregion (Prober et al., 2015; Walters et al., 2015; Metcalf et al., 2016). As judged from thousands of citations, these decades-old primers have outstandingly served the scientific community in the DNA fingerprinting and Sanger sequencing era. However, these primers have one or more mismatches to several broad fungal groups (Bellemain et al., 2010). For another set of ITS primers, a single central mismatch reduced the relative abundance of amplicons >10-fold in artificial communities (Ihrmark et al., 2012).

Here, we highlight the critical mismatches of ITS1F and ITS2 primers to ecologically and clinically important fungal taxa and provide explicit evidence for their discrimination in soil fungal communities. Our ultimate purpose is to provide alternative options for improving fungal HTS-based molecular identification to be globally comparable and applicable in all disciplines.

Results and discussion

In silico tests (see Experimental procedures section) revealed that both the ITS1F and ITS2 primers completely miss Microsporidia, which are common pathogens of a wide variety of organisms. Similarly, both primers have multiple mismatches to Tulasnellaceae that constitutes the most common group of orchid mycorrhizal symbionts world-wide (Dearnaley et al., 2012). The same problem is inherent to all forward primers within 5.8S rDNA, indicating that both ITS1 and ITS2 regions separately are unsuited for identification of these taxa (Oja et al., 2015). The ITS2 primer possesses 3’ terminal mismatches to nearly all species in five early diverging fungal (sub)phyla (Fig. 2). Of these, Neocallimastigomycotina are gut symbionts of artiodactyls, Zoopagomycotina and Entomophthoromycotina are obligate animal parasites, while Rozellomyctota (syn. Cryptomycota) and Chytridiomycota represent the most diverse and abundant groups of aquatic fungi (Richards et al., 2012; Hassett and Gradinger, 2016). Besides chytrids, members of the Venturiales (Ascomycota), Exobasidiales, Ustilaginales and Entylomatales (all Basidiomycota) are important plant pathogens that are all missed...
by the ITS2 primers. Most species within the Omphalotaceae family and the related genus *Mycena*, which are among the globally most diverse and abundant saprotrophic decomposers of plant litter (Tedersoo et al., 2014; Sterkenburg et al., 2015), have a terminal mismatch at the 3’ end of the ITS2 primer. In many environments, exclusion of these groups by the ITS2 primer may potentially lead to wrong conclusions about the distribution of fungal functional guilds and their contribution to ecosystem processes, such as carbon turnover during litter decomposition. Terminal and near-terminal ITS2 primer mismatches are also alarming for human gut and skin microbiota, including opportunistic pathogens, such as true yeasts *Galactomyces geotrichum* and *Yarrowia lipolytica* and basidio-yeasts belonging to Malasseziales and Tremellales s.lat. (LaTuga et al., 2011; Huffnagle and Noverr, 2013; Hallen-Adams et al., 2015). Among the latter group, *Cryptococcus neoformans* causes lethal cryptococcosis in immunosuppressed patients. *Emmonsia parva* and *Ajellomyces capsulatus* (Ajellomyceataceae) are causal agents of adiaspiromycosis and histoplasmosis, respectively. The ITS1F primer has multiple central mismatches to nearly all taxa in the subphylum Mucoromycotina (Fig. 2) that includes molds and opportunistic human pathogens. This primer has also near-terminal mismatches to much of the Saccharomycetes, including *Saccharomyces cerevisiae* (baker yeast) and *Candida* spp. – abundant members of the normal microflora and/or pathogens (Hoffmann et al., 2013; Hallen-Adams et al., 2015).

To test the differential performance of primers in the mismatching taxa, we re-analyzed a soil fungal data from natural habitats of Papua New Guinea that was generated by Illumina MiSeq sequencing (Tedersoo et al., 2015a). We selected this data set, because it was the only one generated with >1 primer combination for both ITS1 and ITS2 regions. Taxonomic analysis of the soil HTS data revealed that nearly half of the groups with terminal mismatches (Ustilaginales, Entylomatales, Malasseziales, Exobasidiales, and Ascosphaeraceae) are completely missing in the ITS1 data subset, while others are reduced 15-fold on average (Fig. 3). Except for Rozellomycota, taxa with near-terminal mismatches are less affected (three-fold on average). Near-terminal and central mismatches have even stronger effect on the performance of the ITS1F primer. For example, relative abundance of Saccharomycetales, Mucoromycota (excl. Umbelopsidaceae) and Umbelopsidaceae is reduced by a factor of 9.0, 117.7, and 16.8, respectively. Taken together, our analyses indicate that the diversity of many ecologically and functionally important fungal groups is severely underestimated by the primers used by microbiome consortia. Moreover, the relative performance of mismatching primers is unpredictable, because primer annealing depends on PCR conditions, type of polymerase, the nature of mismatching

Fig. 1. Primer map of the ITS region indicating the discussed primers. Primers in red denote those used by the microbiome consortia; [I] indicates a common intron site.

![Primer map of the ITS region indicating the discussed primers.](https://example.com/primer_map.png)

Fig. 2. Template mismatches of the ITS1F and ITS2 primers. Only groups with >10% mismatching accessions are indicated. Taxa with single mismatches in the central area and 5’ end are not shown for the ITS2 primer; taxa with single mismatches in the 5’ end are not shown for the ITS1F primer.

ITS1/ITS2 ratio.

across fungi (Nilsson et al., 2011), compared with the ITS2 region (Mello et al., 2012). Fewer OTUs and lower phylogenetic richness of fungi were statistically analyzed by testing the differences in relative abundance of sequences belonging to targeted taxa across 26 soil samples using a series of two-way ANOVAs, treating samples as blocking factors. Error bars represent standard error. Zoopagomyces, a portion of 18S would enable to improve accuracy results of amplicons ITS1ngs-ITS2 and ITS1Fngs-ITS2 against longer templates. For this option, an optimized version of a universal ITS barcode (600–700 bases) in fungi. For longer ampli-

and neighboring nucleotides, sample complexity, etc. (Kanagawa, 2003).

Previous studies indicate that the ITS1 region yields fewer OTUs and lower phylogenetic richness of fungi compared with the ITS2 region (Mello et al., 2011; Op de Beeck et al., 2014; Tedersoo et al., 2015a,b). While the overall variability of ITS1 and ITS2 regions is comparable across fungi (Nilsson et al., 2008), PCR and sequencing biases may recover lower fungal diversity using the ITS1 marker. Besides primer bias, multiple species from several ascomycete classes possess an intron downstream of the ITS1F priming site (Bhattacharya et al., 2000). In addition, the ITS1 region has greater length heterogeneity among eukaryote phyla and fungal taxa (Tedersoo et al., 2015a; Wang et al., 2015). Among ecologically and economically important fungi, Cantharellus and Boletaceae species exhibit excessively long ITS1 regions, but this is also true for the ITS2 region of some Boletaceae. Both the intron and oversized ITS1 may render these taxa out-competed in amplification and emPCR steps that discriminate against longer templates.

The capacity of HTS technologies to cover species diversity to saturation (Caporaso et al., 2012) calls into question the necessity of kingdom-specific primers for fungi. Fungi are highly dominant in soil (90–93% of eukaryote ITS sequences; Tedersoo et al., 2015a; 2016a), wild boar feces (89%; L. Tedersoo, unpublished), leaves (39%; Remmelgas, 2016), and non-ectomycorrhizal roots (7%; Oja et al., 2015), based on nearly universal eukaryote primers. Therefore, fungal diversity may be captured using broader, universal primers, which also enable rough estimates of relative abundance across broad eukaryote groups (Tedersoo et al., 2016a). Furthermore, additional peptide-nucleic acid PCR clamps can be used to reduce amplification of other organisms such as the host species (Lundberg et al., 2013), but this approach is seldom applied in environmental microbiology.

For the above reasons, we advocate the use of the ITS2 subregion and a degenerate version of the ITS4 primer (White et al., 1990) in combination with modified gITS7 (Ihrmark et al., 2012) or ITS3ngs mixes (Tedersoo et al., 2014; Table 1) when applying current second-generation sequencing technologies that produce short reads. The proposed primer ITS4ngsUni accounts for nearly all mismatches in fungi and most eukaryotes, except Foraminifera and Microsporidia. Because HTS technologies evolve rapidly in terms of read length, quality, and throughput, there is a great potential to cover the entire ITS region for more accurate species-level identification (Schoch et al., 2012; Köljalg et al., 2013). The combination of modified ITS1 and ITS1F forward primers (Oja et al., 2015) and the ITS4ngsUni reverse primer serve as strong candidates for HTS analysis of the full ITS barcode (600–700 bases) in fungi. For longer ampli-

Experimental procedures

We used the aligned 18S, 5.8S, and 26S rDNA sequence data set (Bengtsson-Palme et al., 2013) for initial
monitoring of primer-template mismatches across the fungal kingdom and identified specific taxonomic groups (genus to phylum level) where mismatches were common (>5% of accessions) and not attributable to incomplete or erroneous data. Compromised quality was assigned to sequences that had the primer sites only partly covered and that were obviously amplified and sequenced using the primers we evaluated (as seen from matching start and end of query and reference sequences). Ambiguous nucleotides, unique single nucleotide indels and unique substitutions (occurring once or < 0.1% for a taxonomic group) in conserved regions were also considered indicative of low sequence quality. These sequences were excluded from primer-template mismatch evaluation.

To determine additional mismatching taxa, we performed blastN searches (parameters: word size = 7; match score = 1; mismatch score = -3; gap opening cost = 5; gap extension cost = 2) using all single-mismatched variants of the primer sequence against all fungi in the entire International Nucleotide Sequence Databases collaboration (INSDC) (accessed 2016-01-20). The relative frequency of mismatching entries in these problematic taxa was determined by calculating the proportion of mismatching primer variants among all hits using taxonomically filtered blastN searches against these particular taxonomic groups. We used the above described criteria to disregard references of low quality. Inclusion of low-quality, non-fungal and mitochondrial reference sequences in in silico tests of primer suitability for fungi seem to have severely overestimated the true primer mismatches in previous studies and blurred differences among primers, especially those in 3’ 18S and 5’ 28S rDNA. By including problematic reference sequences, 5–30% of entries were mismatched that fall into the range reported in most previous studies. Removal of inappropriate sequences reduced the mismatch frequency five-fold to 10-fold and provided clear and consistent taxonomic patterns for primer bias. In reporting mismatches, we define positions in the primers as terminal (two last 3’ nucleotides), near-terminal (third and fourth nucleotide from 3’ end), central (fifth in 3’ end to fifth in 5’ end), and 5’ (four nucleotides in 5’ end).

Based on the above rationale of sequence data quality filtering, we removed compromised sequences in data sets of all eukaryote kingdoms (Bengtsson-Palme et al., 2013) for designing improved primers. Based on the partial alignments of the 18S, ITS, and 26S rDNA, we designed modifications to the existing primers to cover either all fungi or all eukaryotes (still, with certain exceptions), depending on the level of primer site conservation.

Acknowledgements

We thank R.H. Nilsson for providing the 18S-ITS-28S data sets of eukaryotes and discussion about ITS sequence length; T.D. Bruns, J.W. Taylor, and J. Gilbert for constructive comments on an earlier version of the manuscript. L.T. acknowledges funding from the Estonian Science Foundation grants PUT171, RMK, KIK, and EcolChange.
References


Tedersoo, L., Ansalan, S., Bahram, M., Pölmle, S., Riit, T.,Liiv, I., et al. (2015a) Shotgun metagenomes and multiple primer


