WESTERDIJK FUNGAL BIO DIVERSITY INSTITUTE

Large-scale generation and analysis of filamentous fungal DNA barcodes boosts coverage for kingdom fungi and reveals thresholds for fungal species and higher taxon delimitation

D. Vu[1*], M. Groenewald[1], M. de Vries[1], T. Gehrmann[1], B. Stielow[1], U. Eberhardt[2], A. Al-Hatmi[1], J.Z. Groenewald[1], G. Cardinali[3], J. Houbraken[1], T. Boekhout[1,4], P.W. Crous[1,5,6], V. Robert[1], and G.J.M. Verkley[1*]

[1]*Westerdijk Fungal Biodiversity Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands;* [2]*Staatliches Museum f. Naturkunde Stuttgart, Abt. Botanik, Rosenstein 1, D-70191 Stuttgart, Germany;* [3]*University of Perugia, Dept. of Pharmaceutical Sciences, Via Borgo 20 Giugno 74, I 06121 Perugia, Italy;* [4]*Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, The Netherlands;* [5]*Wageningen University and Research Centre (WUR), Laboratory of Phytopathology, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands;* [6]*Department of Genetics, Biochemistry and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0028, South Africa*

**Correspondence*: D. Vu, d.vu@westerdijkinstitute.nl; G.J.M. Verkley, g.verkleij@westerdijkinstitute.nl

Abstract: Species identification lies at the heart of biodiversity studies that has in recent years favoured DNA-based approaches. Microbial Biological Resource Centres are a rich source for diverse and high-quality reference materials in microbiology, and yet the strains preserved in these biobanks have been exploited only on a limited scale to generate DNA barcodes. As part of a project funded in the Netherlands to barcode specimens of major national biobanks, sequences of two nuclear ribosomal genetic markers, the Internal Transcribed Spaces and 5.8S gene (ITS) and the D1/D2 domain of the 26S Large Subunit (LSU), were generated as DNA barcode data for ca. 100 000 fungal strains originally assigned to ca. 17 000 species in the CBS fungal biobank maintained at the Westerdijk Fungal Biodiversity Institute, Utrecht. Using more than 24 000 DNA barcode sequences of 12 000 ex-type and manually validated filamentous fungal strains of 7 300 accepted species, the optimal identity thresholds to discriminate filamentous fungal species were predicted as 99.6 % for ITS and 99.8 % for LSU. We showed that 17 % and 18 % of the species could not be discriminated by the ITS and LSU genetic markers, respectively. Among them, ~8 % were indistinguishable using both genetic markers. ITS has been shown to outperform LSU in filamentous fungal species discrimination with a probability of correct identification of 82 % vs. 77.6 %, and a clustering quality value of 84 % vs. 77.7 %. At higher taxonomic classifications, LSU has been shown to have a better discriminatory power than ITS. With a clustering quality value of 80 %, LSU outperformed ITS in identifying filamentous fungi at the ordinal level. At the generic level, the clustering quality values produced by both genetic markers were low, indicating the necessity for taxonomic revisions at genus level and, likely, for applying more conserved genetic markers or even whole genomes. The taxonomic thresholds predicted for filamentous fungal identification at the genus, family, order and class levels were 94.3 %, 88.5 %, 81.2 % and 80.9 % based on ITS barcodes, and 98.2 %, 96.2 %, 94.7 % and 92.7 % based on LSU barcodes. The DNA barcodes used in this study have been deposited to GenBank and will also be publicly available at the Westerdijk Institute's website as reference sequences for fungal identification, marking an unprecedented data release event in global fungal barcoding efforts to date.

Key words: Automated curation, Biological resource centre, Fungi, ITS, LSU, Taxonomic thresholds.

Available online 30 May 2018; https://doi.org/10.1016/j.simyco.2018.05.001.

INTRODUCTION

Species identification lies at the heart of biodiversity studies, and biodiversity efforts have in recent years favoured DNA-based approaches over morphology-based approaches (Hebert *et al.* 2003, De Queiroz 2007, Verkley *et al.* 2013, Woudenberg *et al.* 2013, Liu *et al.* 2016, Wang *et al.* 2016a, b). The success of DNA barcoding has given rise to characterisation of bacterial biodiversity in every nook and cranny on the planet (Huttenhower *et al.* 2012, Mau *et al.* 2014, Afshinnekoo *et al.* 2015). Knowledge of the microbial composition of such communities aide knowledge of host-microbe interactions and their environmental function, and revealed a complex and sensitive balance that may be easily disturbed (Fuhrman 2009, Garza *et al.* 2016, Levy *et al.* 2017). Such metagenomics studies have been limited in fungi due to the complexity of fungal genomes and the lack of validated databases cataloguing sufficient biodiversity. Fungal genomes are generally larger than prokaryotic genomes with a size ranging from 9 Mb to 178 Mb because they contain large amounts of non-coding and repetitive DNA (Mohanta & Bae 2015). They have been shown to be divergent, even between the members of the same genus (Galagan *et al.* 2005), and dynamic in nature (Dujon *et al.* 2004, Kellis *et al.* 2004, Strope *et al.* 2015). The studies already conducted into fungal metagenomic communities have often been limited to generic or higher levels (Cui *et al.* 2013, Geml *et al.* 2014, Tedersoo *et al.* 2014, Nguyen *et al.* 2015, Botschuijver *et al.* 2017), limiting our understanding of the role of fungal species in our surrounding ecosystems.

The Internal Transcribed Spacer (ITS) nrDNA region has been proposed as a universal barcode for fungi (Schoch *et al.* 2012, Blaalid *et al.* 2013, Koljalg *et al.* 2013). Despite criticism (Kiss 2012), ITS sequences have been shown to be useful in delineating many fungal species (Irinyi *et al.* 2015, Vu *et al.* 2016). Together with the D1/D2 domain of the Large Subunit (LSU, 26S) nrDNA sequences, ITS sequence analysis is frequently used as a method to discriminate fungal species (Kurtzman & Robnett 1998, Fell *et al.* 2000, Boon *et al.* 2010, Kooij *et al.* 2015, Woudenberg *et al.* 2015). Unfortunately less than 23 000 (0.6 %; Vu *et al.* 2014) of the estimated 3.8 million

Peer review under responsibility of Westerdijk Fungal Biodiversity Institute.

Studies in Mycology

species (Hawksworth & Lücking 2017) of fungi have ITS sequences available at GenBank and many of the sequences are often of poor quality (Nilsson *et al.* 2006, Vu *et al.* 2016). Furthermore, the optimal similarity range to make informed decisions about species delineation is still highly uncertain compared to bacterial species (Stackebrandt & Ebers 2006, Edgar 2018). In most fungal ecology studies, a 97 % threshold is given as default (Geml *et al.* 2014, Gweon *et al.* 2015). Although there have been efforts to define a range of taxonomic thresholds from 97 % to 100 % for fungal species identification, such as the UNITE species hypothesis (Koljalg *et al.* 2013), threshold choice remains subjective.

Mycologists studying fungal strains have deposited reference materials in public culture collections for over a century. Microbial Biological Resource Centres (MBRC) preserve this heritage which constitutes an invaluable source for well-defined and high-quality materials for microbiology. At the Westerdijk Fungal Biodiversity Institute (WI, previously CBS-KNAW), Utrecht, The Netherlands, more than 100 000 living strains of yeasts and filamentous fungi are preserved that were originally assigned to ca. 17 000 species. Information on the publically available strains, are accessible via the website http://www.westerdijkinstitute.nl/Collections/. This includes data on morphology, physiology, molecular markers sequences, growth conditions and safety measures. The CBS filamentous fungal collection currently has approximately 70 000 strains including ca. 8 000 ex-type strains representing more than 12 000 currently recognized species, with the oldest strains isolated around 1895. When accessioned, each identified strain is assigned a taxon name from MycoBank (Robert *et al.* 2013), an online registration system for fungal species and higher level taxon names. Strain identification is normally achieved with the knowledge and methods available at the time of accession. While names of strains other than ex-type strains are updated following modern taxonomic concepts, it has not been possible to re-evaluate the original identification of each strain in the CBS collection. For most of the older strains, not all characters necessary for identification are expressed in culture or modern molecular studies have not been conducted yet.

In the WI DNA barcoding project, ITS and LSU barcode sequences were generated based on DNA extracted from cultures for the majority of CBS strains. In a previous study, the barcoding project resulted in the release of 8 669 barcode sequences of manually validated CBS strains representing 1 351 yeast species (Vu *et al.* 2016). In this study, over 24 000 sequences of manually validated strains that belong to 7 300 filamentous fungal species, have been made available by depositing them in GenBank and at the website of the Westerdijk Institute to improve fungal identification in online public databases. Similarly to the results obtained in Vu *et al.* (2016), we showed that ITS and LSU can be used to classify a large portion of all fungi to species level. Additionally, ITS and LSU sequences were complementary, and could be used together to achieve a better identification performance. Based on this large number of ITS and LSU barcodes, thresholds were predicted for circumscription of species and higher taxa of filamentous fungi. The newly generated ITS barcodes were also compared with the "Top 50 Most Wanted Fungi" (Nilsson *et al.* 2016, UNITE Community 2017) dataset to reveal the most frequently sampled environmental sequence types that have been difficult to be assigned to meaningful taxonomic levels.

MATERIAL AND METHODS

Generation and management of barcode sequences

The protocols of genomic DNA extraction and generation of the ITS and LSU barcode sequences were given in Eberhardt (2012) and Stielow *et al.* (2015). The ITS sequences were generated using the forward and backward primers ITS5 and ITS4 for PCR reactions for amplification, containing partial 18S, complete ITS1-5.8S-ITS2, and partial LSU sequences. The LSU sequences were generated in the D1/D2 domain employing the forward and backward primers LROR and LR5 (Stielow *et al.* 2015). To be able to manage a large amount of sequence data and to keep track of the experimental procedures, a laboratory information management system (LIMS; Vu *et al.* 2012) was used for the WI DNA barcoding workflow as a module of BioloMICS (Robert *et al.* 2011), a software package to manage biological databases. The trace files obtained from the sequencer were aligned. The consensus sequences were imported automatically into the WI database, and edited manually by replacing or removing all ambiguous characters. After that, they were checked, compared with the existing sequences from local and public databases such as GenBank, and validated by the curators. For the analyses presented in this paper, only validated sequences that were added to the database until December 2016 were included, and for each strain, only one sequence was selected.

Computing a similarity score between DNA sequences

The similarity score of two DNA sequences was computed using our own implementation (Robert *et al.* 2011) of the BLAST algorithm (Altschul *et al.* 1997) as the percentage identity of the most similar region (coverage) between the two sequences.

Selecting representative sequences for species

The representative sequence of a species was taken as the sequence of the ex-type strain if it was available, otherwise as the central representative sequence of all the sequences of the strains associated with the species (Vu *et al.* 2016). Here, the central representative sequence of a group was the one that had the highest similarity score to the other sequences in the group. The strain associated with the (central) representative sequence was considered to represent the (central) representative strain of the species. The similarity score of two strains was the similarity score of their reference sequences.

Probability of correct identification (PCI)

Given a genetic marker, the identification of a species is correct if, for every strain of the species, there is no other strain from another species for which the similarity score between them is greater or equal to the minimum similarity score between the strains of the species under consideration. The PCI of the given genetic marker is the fraction of species correctly identified. To evaluate the resolving power of multiple genetic markers for species discrimination, the similarity score of two strains using