Fungal Ecology 36 (2018) 75–80

Contents lists available at ScienceDirect

Fungal Ecology

journal homepage: www.elsevier.com/locate/funeco

Coming up short: Identifying substrate and geographic biases in fungal sequence databases

Maryia Khomich [a, b, *], Filipa Cox [c, d], Carrie J. Andrew [e, b], Tom Andersen [a], Håvard Kauserud [b], Marie L. Davey [b, f]

[a] Section for Aquatic Biology and Toxicology, Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, 0316, Oslo, Norway
[b] Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, 0316, Oslo, Norway
[c] School of Earth and Environmental Sciences, University of Manchester, Manchester, M13 9PL, United Kingdom
[d] British Antarctic Survey, Natural Environment Research Council, Cambridge, CB3 0ET, United Kingdom
[e] Swiss Federal Institute for Forest, Snow, and Landscape Research WSL, Zürcherstrasse 111, 8903, Birmensdorf, Switzerland
[f] Norwegian Institute of Bioeconomy Research, Department of Soil Quality and Climate Change, Høgskoleveien 7, 1430, Ås, Norway

ARTICLE INFO

Article history:
Received 30 March 2018
Received in revised form
31 July 2018
Accepted 2 August 2018

Corresponding Editor: Björn D. Lindahl

Keywords:
Fungi
Mycobiome
ITS region
GenBank
UNITE
RDP Bayesian classifier
Diversity
Metabarcoding
Substrate

ABSTRACT

Insufficient reference database coverage is a widely recognized limitation of molecular ecology approaches which are reliant on database matches for assignment of function or identity. Here, we use data from 65 amplicon high-throughput sequencing (HTS) datasets targeting the internal transcribed spacer (ITS) region of fungal rDNA to identify substrates and geographic areas whose underrepresentation in the available reference databases could have meaningful impact on our ability to draw ecological conclusions. A total of 14 different substrates were investigated. Database representation was particularly poor for the fungal communities found in aquatic (freshwater and marine) and soil ecosystems. Aquatic ecosystems are identified as priority targets for the recovery of novel fungal lineages. A subset of the data representing soil samples with global distribution were used to identify geographic locations and terrestrial biomes with poor database representation. Database coverage was especially poor in tropical, subtropical, and Antarctic latitudes, and the Amazon, Southeast Asia, Australasia, and the Indian subcontinent are identified as priority areas for improving database coverage in fungi.

1. Introduction

Fungi encompass one of the most functionally and ecologically diverse kingdoms of eukaryotes, maintaining ecosystem functioning on a global scale and playing fundamental roles as decomposers, mutualists and pathogens of animals and plants (Peay et al., 2016). Estimates of global fungal diversity range from 0.6 to 5.1 million species of fungi (Hawksworth, 2001, 2012; Bass and Richards, 2011; Blackwell, 2011; Hawksworth, 2012). However, to date, only a tiny fraction of them (ca. 140 000 species) have been classified, although some 1200 new fungal species are described each year (Kirk et al., 2008; Hibbett et al., 2011).

The advent of massively parallel high-throughput sequencing (HTS) has enabled the exploration of fungal diversity on a previously impossible scale (Hibbett et al., 2009). As a result, fungal barcoding of environmental samples is increasingly driving the exploration of the processes structuring fungal diversity, the identification of ecosystem functions linked to fungal diversity and the discovery of novel fungal biodiversity, especially for understudied geographic regions and substrates (Schoch et al., 2012; Öpik et al., 2016). Fungal barcoding approaches largely focus on the internal transcribed spacer (ITS) region, which is the standard barcode for Fungi (Schoch et al., 2012). The establishment of large-scale public reference ITS databases is therefore crucial to allow reliable sequence-based identification of fungal species in HTS approaches (Coissac et al., 2016).

* Corresponding author. Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, 0316, Oslo, Norway.
E-mail addresses: maryia.khomich@ibv.uio.no, marykhomich@gmail.com (M. Khomich).

https://doi.org/10.1016/j.funeco.2018.08.002
1754-5048/

Database-dependent HTS approaches suffer from several biases and limitations directly related to the quality and breadth of the databases. For example, there are only a relatively small fraction of reference database sequences for which a specimen or culture is readily available (Bridge et al., 2003) and consequently large proportions of environmental sequences typically are not represented in the sequence databases. In the case of Fungi, the three public repositories in the International Nucleotide Sequence Database Collaboration (INSDC), namely the DNA Data Bank of Japan (DDBJ), European Nucleotide Archive (ENA) and GenBank, have become a default resource of taxonomic annotation for newly generated environmental sequences (Karsch-Mizrachi et al., 2018). However, it has been reported that 10–21% of fungal sequences deposited in INSD can be either chimeric, of poor quality or contain incorrect and insufficient taxonomic information (Bridge et al., 2003; Nilsson et al., 2006). To improve the annotation of fungal ITS sequences from NCBI databases, the ITS RefSeq Targeted Loci project has been initiated to develop a separate, curated database representing sequences from type material and stored in public archives (Schoch et al., 2014; Robbertse et al., 2017). By contrast, UNITE (unite.ut.ee) provides highly filtered, curated ITS reference sequences for molecular identification of fungi (Kõljalg et al., 2013). The geographic representation in both databases is strongly skewed towards Europe, North America, China, and Japan (Ryberg et al., 2009; Kõljalg et al., 2013). As a result, satisfactory taxonomic assignment remains problematic in the kingdom Fungi due to the lack of reliable and correctly annotated reference sequences, and coverage related biases.

Here, we assess the impact of unbalanced database representation by geographic locale and substrate on our ability to discern and identify the components of fungal communities using data from 65 amplicon HTS datasets targeting the ITS region of rDNA. We attempt to identify both substrates and geographic regions in which underrepresentation in the available fungal ITS databases could have meaningful impact on our interpretation of HTS amplicon sequencing data.

2. Materials and methods

The data analysed represent 65 next generation ITS amplicon sequencing datasets from 14 different substrates, including terrestrial, aquatic, and marine environments, as well as plant and animal hosts (Table S1). Data were gleaned from published materials, through personal communication with the authors, or from public data archives (e.g. ENA or NCBI Sequence Read Archive (SRA)). Among these datasets, 30 were derived from soil substrates representing 625 sites from 14 biomes with global distribution across all continents (Table S1, Table S2). Sites were assigned to biomes following the classification of the World Wildlife Foundation (http://worldwildlife.org) with the following modifications: (i) temperate deciduous forests in the Northern and Southern hemispheres were treated separately; (ii) montane forests were separated from lowland forests in the tropics; (iii) grasslands and shrublands were considered as a single unit globally, and (iv) vegetated subantarctic sites were differentiated from unvegetated maritime Antarctic sites. For all datasets, sequences were error-corrected and quality-filtered prior to clustering into operational taxonomic units (OTUs) at a 97% similarity threshold (Table S1).

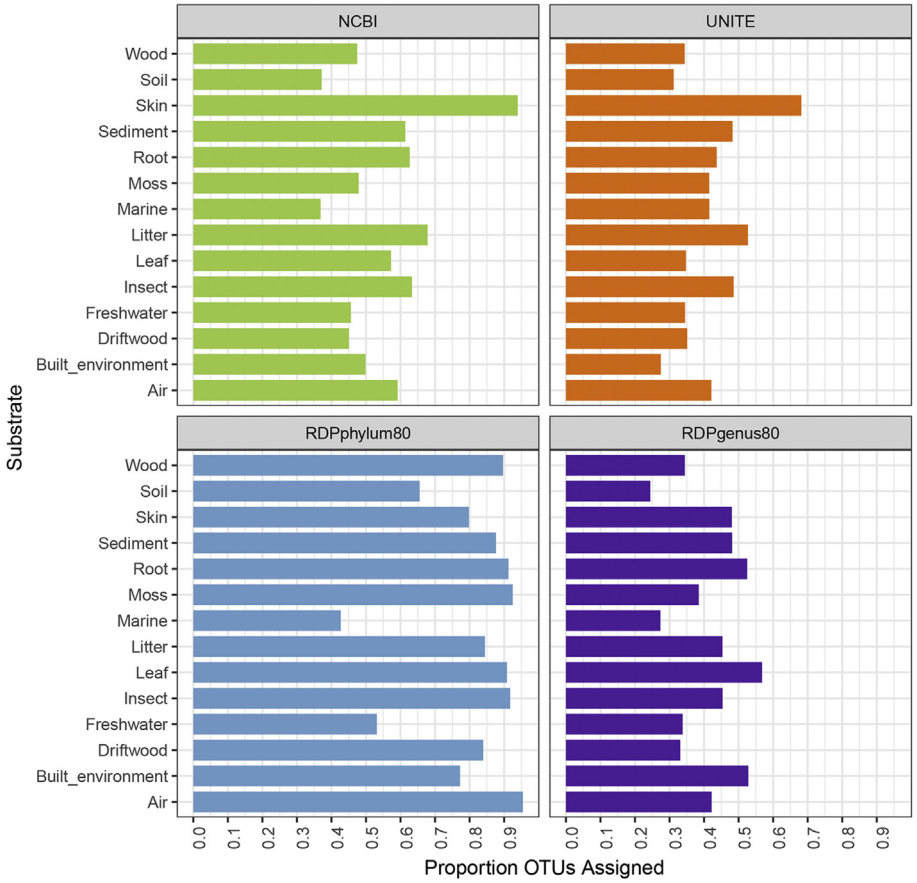

Fig. 1. A barplot representing proportion of assigned fungal OTUs across 14 different substrates in NCBI (green), UNITE (orange) and RDP (phylum level: blue; genus level: purple) databases.