British Mycological Society promoting fungal science

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

Review

Opportunities for diversified usage of metabarcoding data for fungal biogeography through increased metadata quality

Mathew Andrew HARRIS[a,b,*], *Bernard SLIPPERS*[b,c], *Martin KEMLER*[d], *Michelle GREVE*[a,b]

[a]*Department of Plant and Soil Science, University of Pretoria, Pretoria, South Africa*
[b]*Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa*
[c]*Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa*
[d]*Organismic Botany and Mycology, Universität Hamburg, Hamburg, Germany*

ARTICLE INFO

Article history:
Received 22 March 2023
Received in revised form
4 May 2023
Accepted 14 June 2023

Keywords:
High-throughput sequencing
Metadata
Fungi
Conservation
Big data
INSDC
Macroecology
Species

ABSTRACT

The widely adopted use of metabarcoding techniques and the ability to sequence microbial communities directly from environmental samples have advanced the field of fungal ecology. The growth of publicly available big data offers opportunities for collating data from different sources to explore biogeographical and macroecological patterns of fungal groups over large spatial scales. This requires reliable and high-quality metadata associated with the raw sequencing data. We assessed the accuracy of submitted metadata linked to terrestrial plant-associated fungal genetic marker sequences, extracted from NCBI's BioProject web-portal. The amount of correctly captured, missing, and incorrectly supplied metadata was determined. The quality of submitter-defined metadata was of a variable quality, with some adhering to metadata standards, and others not capturing metadata for certain attributes or, when metadata was captured, duplicating metadata across samples, or only partially meeting metadata requirements. This ultimately limits the ability to find, and subsequently re-use, sequence data. The rapid accumulation of metabarcoding data and the ability to directly compare samples taken from different studies holds opportunities for gaining a deeper understanding of fungal biogeographical patterns and their drivers. Standardised vocabularies for metadata attributes during submission to public repositories like NCBI's Sequence Read Archive, coupled with adequate incentives for the data providers, would facilitate the Findability, Accessibility, Interoperability, and Reusability (FAIR) data principles and ultimately enable metabarcoding sequence data to be readily utilized to perform large scale global biogeographical studies on the kingdom Fungi.

* Corresponding author. Department of Plant and Soil Science, University of Pretoria, Pretoria, South Africa.
E-mail address: mathew.harris@fabi.up.ac.za (M. A. Harris).
https://doi.org/10.1016/j.fbr.2023.100329

1. Introduction

Metabarcoding techniques have become the *de facto* standard used for almost all microbial ecological studies (Cline et al., 2017; Sun et al., 2021; Tedersoo et al., 2022a). The decreasing costs and widely adopted use of high-throughput sequencing technologies have resulted in an exponential growth in microbial metabarcoding studies since the technology's initial commercial availability in 2005 with the Roche GS20 454 sequencing machine (Katz et al., 2022; White III et al., 2016) (Fig. 1). Such studies are revealing how important factors such as climate, geographic isolation and host identity are in shaping fungal species distribution patterns across fungal guilds (Cowan et al., 2022; Davison et al., 2015; Steidinger et al., 2019; Talbot et al., 2014; U'Ren et al., 2019; Větrovský et al., 2019). Additionally, the creation of databases to store raw metabarcoding data (e.g. the National Center for Biotechnology Information (NCBI) Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/) and associated BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) and BioSample (https://www.ncbi.nlm.nih.gov/biosample/) metadata-databases) has meant that the number of publicly available datasets containing georeferenced microbial community information and associated habitat information has expanded rapidly, particularly over the last decade as sequencing costs have continued to decrease (Katz et al., 2022; Nemergut et al., 2013; Peay et al., 2016).

As the amount of metabarcoding data increases, so do opportunities to re-use these data to answer new questions for which these datasets were initially not intended; as has been done for biological collections of macro-organisms e.g. plants and animals in herbaria and museums (Greve et al., 2016; Pyke and Ehrlich, 2010). Such opportunities, amongst others, include obtaining valuable baseline data for determining spatially explicit patterns in species distributions or diversity, assessments of community changes through time or changes in phenology (e.g. leaf senescence, flower emergence and bud break) (Elith and Leathwick, 2007; Greve et al., 2016; Heberling and Isaac, 2017; Zani et al., 2020).

For macro-organisms these opportunities have been widely explored, with the collation of disparate datasets collected across different geographic areas and multiple decades, which have especially led to the advancement of the field of biogeography (Chalmers and Henderson, 1998; Elith and Leathwick, 2007; Greve et al., 2016; Lavoie, 2013; Romeiras et al., 2014). Biogeography considers the role of historic and contemporary factors in shaping the spatial and temporal distributions of all levels of biological diversity, from genes to ecosystems (Lomolino et al., 2010). In the rapidly changing anthropogenic world the field of biogeography seeks to generate predictive capacities to gauge how biodiversity will respond. Such information is vital to set well-considered conservation priorities, limit biological invasions and truly appreciate the ecosystem services that biodiversity provides the growing global population (Ackerly et al., 2010; De Kort et al., 2021; Greve et al., 2016; Liu et al., 2018; van Wilgen et al., 2021).

Initially, fungal biogeographical studies were restricted to fungi that produced conspicuous but ephemeral sporocarps, fungal pathogens of economic importance, long-lived lichens or a subset of culturable cryptic fungi (Arnold and Lutzoni, 2007; Ellis et al., 2007; Klich, 2002; Mueller et al., 2007). This meant that the true dimensions of Earth's fungal diversity and patterns of their composition remained largely unknown (Peay et al., 2016). The rapid adoption of metabarcoding

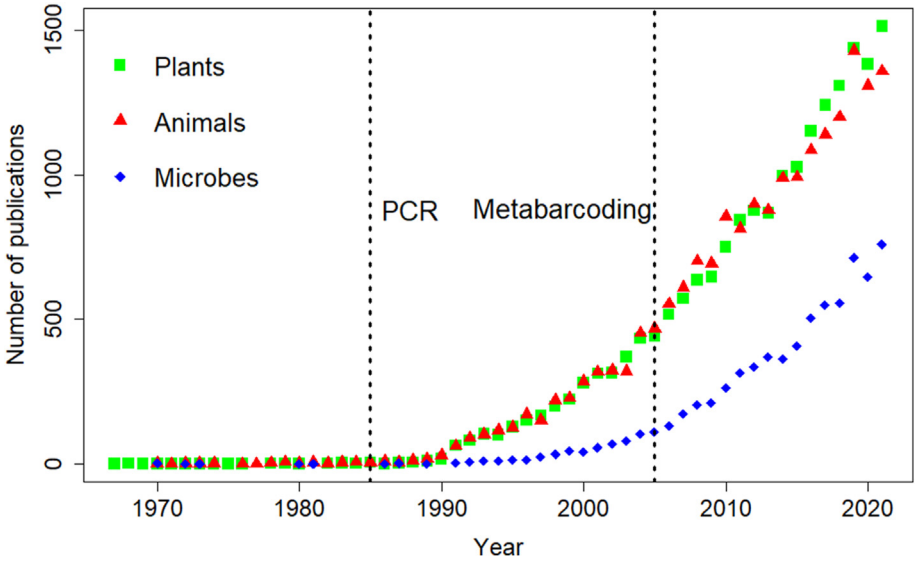

Fig. 1 – Count of publications per year that focus on biogeography OR macroecology of animals (mammal*, reptile*, bird*, insect*, fish, amphibian* and animal*), plants (tree*, grass*, forb*, moss*, fern*, plant* and vegetation) and microbes (fung*, bacteria*, virus, protist* and archaea). The 1985 vertical dashed line marks the year in which PCR became commercially available. The 2005 vertical line represents the year when the first high-throughput sequencing platform (Roche GS20 454 sequencing machine), and by extension metabarcoding, became commercially available. All searches were performed on the Web of Science core collection database on August 31, 2022 with data displayed until 2021.

techniques by the mycological community since the commercial availability of high-throughput sequencing technologies (Fig. 1) has undoubtedly begun to unravel continental to global-scale biogeographical patterns of fungi (Bahram et al., 2018, 2012; Cowan et al., 2022; Davison et al., 2015; Tedersoo et al., 2022b, 2014; U'Ren et al., 2019). However, amongst these studies, a clear bias exists towards soil-dwelling fungi, with other fungal groups receiving less attention. However, the technology has also resulted in many studies assessing local-scale patterns of less studied fungal groups, e.g. plant-associated, indoor, or insect-associated (to name a few) (Ceballos-Escalera et al., 2022; Harris et al., 2023; Korpelainen et al., 2016). To fully make use of existing fungal metabarcoding data for exploring global biogeographical patterns, raw sequence data must be accompanied by good quality metadata that enables easy reuse of said data for purposes for which the data were not initially intended, including global-scale analyses on fungal biogeography.

1.1. *Public databases: opportunities and limitations*

For conclusive biogeographical and macroecological insights, many samples collected over vast areas spanning different environments are required (Greve et al., 2016). Since the cost to travel around the world, and the logistics and costs of research permits, guides, and sample collection to collect large quantities of widely distributed samples is simply not feasible for (most) researchers, alternative methods to obtain biological species information across large spatial scales are required (Maldonado et al., 2015; Wüest et al., 2020). One option is to collate data that have been collected by different collectors in different geographic areas to perform large-scale biogeographical and macroecological analyses (Greve et al., 2016; Wüest et al., 2020). For plant and animal taxa, such datasets have been compiled using, for example, herbarium accessions and museum specimens and, more recently, through citizen science platforms (Brown and Williams, 2019; Chandler et al., 2017; Greve et al., 2012; Lavoie, 2013; Meineke et al., 2018). While these platforms do not exclude fungal data, they are biased towards the compilation of macro-organisms that can be visually identified. In the animal and plant world, species concepts covering much of the diversity of these groups have been in existence for centuries (Linnaeus, 1758), not least because species could be well-defined due to their large size and thus easily quantifiable morphological features (relative to microbes). As a result, the collection of specimens of macro-organisms has had a long history, with much of this history being reflected in existing biodiversity databases (Meineke et al., 2018). Obtaining such historic data for most fungal groups has been more challenging. This is due to the fact that the majority of fungal diversity is microscopic in nature (Hawksworth, 2001), and species concepts for microbes has since the end of the 20th century been almost exclusively based on the phylogenetic species concept and not the more traditional morphological species concept used to define many macro-organisms (Taylor et al., 2000), ultimately requiring different databases to store said microbial taxonomic data. It is primarily large fruit-producing saprophytes and ectomycorrhiza, or long-lived lichens, which can be photographed and/or found in herbaria and fungaria which have historical records (Andrew et al., 2017).

Microbial biogeography, including fungal biogeography, is an emerging field (Cowan et al., 2022; Tedersoo et al., 2022b, 2014; U'Ren et al., 2019). The number of studies that have utilized high-throughput sequencing of genetic markers by PCR amplification has grown dramatically since the technology's inception in 2005 (Katz et al., 2022; Tedersoo et al., 2022a). As the number of studies has increased, so has the size of the sequencing datasets (Sun et al., 2021). While efforts have been made to utilise metabarcoding datasets to map and monitor fungal occurrences across the globe, e.g. GlobalFungi and UNITE (Nilsson et al., 2019; Větrovský et al., 2020), few studies have looked to re-utilise existing metabarcoding data for global biogeographical studies (see Tedersoo et al., 2022b for a good example). Thus, existing metabarcoding datasets represent a wealth of information that has the potential to unlock a deeper understanding of the factors responsible for shaping fungal biogeographical patterns around the globe for both free-living and host-associated fungi.

Metabarcoding data has accumulated quickly, and public data repositories that store this raw sequencing data and the associated metadata have needed to evolve equally rapidly (Barrett et al., 2012; Katz et al., 2022; Yilmaz et al., 2011). Theoretically, these public repositories should handle raw sequence data and its metadata in a structured manner so that it enables FAIR data principles not just for humans, but also for computers (GO FAIR, 2022; Wilkinson et al., 2016). Examples of such public data repositories include databases hosted by the members of the International Nucleotide Sequence Database Collaboration (INSDC) (Arita et al., 2021), which consists of the DNA DataBank of Japan (https://www.ddbj.nig.ac.jp/index-e.html), the European Nucleotide Archive (https://www.ebi.ac.uk/ena), and NCBI (https://www.ncbi.nlm.nih.gov).

Historically, a major factor that has limited the effective pooling of microbial metabarcoding datasets has been the way in which operational taxonomic units (OTUs) (Blaxter et al., 2005) of different studies are quantified (Bolyen et al., 2019; Callahan et al., 2017; Peay et al., 2016). In some cases, OTUs defined from one study are not comparable to OTUs defined in another study if *de-novo* OTU clustering methods were used to assign sequences to OTUs. In other cases, OTUs that are defined using closed-reference methods are comparable to one another, but any biological sequence variation that is not present in the reference database from which the OTUs are assigned is simply lost (Amir et al., 2017; Callahan et al., 2017). Indeed, OTUs which are clustered using open-reference methods overcome this issue by retaining unique sequences which would be lost if closed-reference clustering was used (Tedersoo et al., 2022a). Newer methods which resolve metabarcoding datasets to amplicon sequence variants (ASVs), which are exact sequence variants, were developed to improve taxonomic resolution and reproducibility of identified taxa (Amir et al., 2017; Bolyen et al., 2019; Tedersoo et al., 2022a). While the use of ASVs has advantages over that of OTUs, as they can be directly compared between studies, can be reproduced in future datasets, and are not hindered by incomplete reference databases, they do have drawbacks, e.g. they bias richness estimates and perform poorly for

certain groups of fungi (Runnel et al., 2022; Tedersoo et al., 2022a). Therefore, the use of open-reference clustering which takes into account the taxonomic structure of reference databases has been suggested as the best practice for generating the most taxonomically stable OTUs (Cline et al., 2017; Tedersoo et al., 2022a). Whether researchers opt to use open-reference OTUs or ASVs, the time is ripe for publicly available metabarcoding datasets to be pooled such that global scale biogeographical analyses of fungi can be performed.

A further limitation to collating metabarcoding datasets is the quality of metadata deposited in public repositories. Therefore, we explore in this review the utility of publicly available data stored in repositories like NCBI for studying fungal biogeographical patterns and processes. We assess whether the metadata associated with raw metabarcoding sequence data is of a high enough quality for performing large-scale biogeographical analyses. We encourage continued efforts by, and open dialog between the mycological community to promote the practice of submitting metadata that adheres to current metadata standards associated with raw metabarcoding datasets (Abarenkov et al., 2016; Nilsson et al., 2022; Penev et al., 2017; Tedersoo et al., 2015). However, we acknowledge that the responsibility of providing correct and complete metadata cannot solely be placed on the data providers but requires a holistic approach which includes all stakeholders. Finally, we propose that 1) stricter requirements and checks of metadata submitted to public repositories housed by members of the INSDC (NCBI, European Nucleotide Archive and DNA DataBank of Japan) (Arita et al., 2021) could limit issues of data heterogeneity and incomparability, and ultimately quicken advances made to understand the diversity and distribution of fungi globally; 2) incentives, in the form of citations, are given to those who supply raw sequencing data with correct metadata; and 3) that citations for datasets become comparative to citations of research articles (Abarenkov et al., 2022; Nilsson et al., 2022; Penev et al., 2017).

2. Methods

We explored the metadata of terrestrial plant-associated fungal community data from NCBIs' BioProject and BioSample databases to assess whether the user-supplied metadata could facilitate the re-use of raw metabarcoding data housed on NCBI's Sequence Read Archive for answering large-scale biogeographical questions. We selected terrestrial plant-associated fungi as an example group on which metadata can be assessed. Plant-associated fungi have varied relationships with their hosts, ranging from mutualism (e.g. mycorrhiza) to parasitism (i.e. pathogens) (Compant et al., 2019). Therefore, understanding their spatial distributions and host ranges provide valuable information about the ecosystem processes and services host-associated fungi can deliver (Steidinger et al., 2019) and have applications for agriculture and forestry, such as pre-empting fungal pandemics (Ristaino et al., 2021).

Data were downloaded from the BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) and BioSample (https://www.ncbi.nlm.nih.gov/biosample/) metadata databases through the NCBI's interactive web portal. The BioProject database stores information pertaining to research projects that have deposited their metabarcoding data to one of the members of the INSDC's portals (Clark et al., 2013). Data between the three members of the INSDC is continually shared. Thus, the same data and its associated metadata should theoretically be obtained through each platform (Barrett et al., 2011). The BioProject database describes how and why each study was conducted, i.e. it provides project title and project description attributes. Linked to each BioProject accession are multiple BioSample accessions which provide metadata associated with each sample within one BioProject (Barrett, 2013; Barrett et al., 2012). Since the metadata uploaded onto the BioProject and BioSample databases are user-supplied and since the metadata-databases store descriptions for many different primary data archives on NCBI, the supplied metadata is highly heterogeneous in nature (Barrett, 2013). This makes it difficult to reliably obtain all relevant data, even with complex search terms (Gonçalves and Musen, 2019; Klie et al., 2021) and it is unlikely that we obtained all terrestrial plant-associated fungal metabarcoding data. Nevertheless, we believe the patterns we observe in the quality of the user-supplied metadata are representative.

To download the data from the BioProject online database within the NCBI web-portal, a text search was performed in NCBI's BioProject database. Boolean operators and filters were used to narrow down the search before extracting the data. The search was as follows: ((("Fungi" [Organism] OR fungi [All Fields]) OR fungal [All Fields]) OR endophyte [All Fields]) OR epiphyte [All Fields] AND (("targeted locus loci" [Filter] OR metagenome [Filter]) AND "bioproject sra" [Filter]). This search was conducted on June 29, 2019, yielding 1050 unique BioProject accessions linked to NCBI's Sequence Read Archive. Each of these 1050 BioProject accessions was manually assessed based on the BioProject attribute "project description", to check whether the project specifically targeted terrestrial plant-associated fungi. All BioSample accessions that targeted terrestrial plant-associated fungi were downloaded in. xml format and parsed from. xml to. csv using the XML package version 3.99–0.13 (Lang and CRAN Team, 2019) in R version 3.6.0 (R Core Team, 2019). The parsed dataset was inspected for any evidence that the downloaded metadata was not related to fungal sequences. In some instances, projects targeted both fungal (e.g. ITS) and plant genetic (e.g. rbcL and trnH) markers. In such instances, the accessions which targeted plant genetic markers were removed. The final dataset consisted of 17,921 unique BioSample accessions from 142 different BioProject accessions. This dataset can be downloaded at: doi. org/10.6084/m9. figshare.21,387,363. v2 and a list of publications which are associated with the downloaded BioProject accessions can be viewed in the supporting document - Supplementary Table S1.

The amount of missing, correctly supplied and incorrectly supplied values for eight of the BioSample attributes from the raw downloaded metadata were assessed against the Genomic Standards Consortium's definitions set out in the Minimum Information about any x Sequence (MIxS) standards (Genomic Standards Consortium, 2022). The eight BioSample attributes we assessed were: host (i.e. plant scientific

name), geographic coordinates in decimal degrees (i.e. latitude and longitude), elevation (i.e. height above mean sea level), geographic location (i.e. country and/or sea region), collection date (i.e. date when sample was collected), broad-scale environmental context (i.e. the biome or large scale environment from where samples were collected), local-scale environmental context (i.e. entity or entities which have causal influence on sample and smaller spatial grain than broad-scale environmental context) and environmental medium (i.e. the tissue of the symbiotic host organism from which DNA was extracted) (see Supplementary Table S2 for definitions and examples).

Datasets are rarely published immediately after collection and sequencing, which can lead to duplication of research efforts, missed opportunities to detect biosecurity risks earlier and slows down knowledge generation and innovation. The delay in sample availability was calculated by working out the number of months from when samples were collected (i.e. collection date) until BioSample accessions became publicly available on NCBI (i.e. BioProject "Registration date"). When only the year was provided for collection date, we assumed that these BioSample accessions were collected in the middle of that year (i.e. June). Since we were interested to examine how the trend in delayed data availability changed through time, we repeated our text search (detailed above) on NCBI's BioProject web-portal on August 30, 2022 to obtain all additional collection date metadata available until 2022. This yielded data from an additional 48 BioProject accessions over and above the 142 already obtained in June 2019.

To determine how much of user-supplied metadata was missing for each of the eight attributes assessed here (i.e. host, geographic coordinates, elevation, geographic location, collection date, broad-scale environmental context, local-scale environmental context and environmental medium), we counted the number of the 17,921 BioSample accessions per attribute which did not contain values (i.e. cell was blank) or had the text "missing", "not-applicable", "NA", "unknown" or something similar.

For each of the eight attributes, standards exist for what should be entered as an appropriate value (Genomic Standards Consortium, 2022) (Supplementary Table S2). Within the definition of each attribute, criteria are often specified for what constitutes complete and correct metadata capturing. Therefore, to assess the quality of the metadata we calculated the percentage of BioSample accessions which met all the criteria set out in each definition for the eight attributes included in this study (Supplementary Table S2, see also Table 1). For the coordinates attribute, we additionally mapped the coordinates to ensure they fell within the country information provided in the country and/or sea region attribute. The remaining BioSample accessions for each attribute which did not meet all the criteria set out in their respective definitions were placed into categories of common mistakes made for each of the eight attributes assessed here (Table 1). These mistakes are discussed in terms of lost resolution to the metadata, particularly with regards to geography, taxonomy and collection dates and what this means for performing large-scale biogeographical studies.

To assess biases in the downloaded plant-associated fungal metadata, the plant organs, hosts, sequencing machines, gene regions and plant growth forms of each accession were scored and this manually curated dataset can be obtained: doi. org/10.6084/m9. figshare.21,387,351. v2. Where possible the pertinent information was obtained from the downloaded BioSample accessions. However, this information was not always available from the BioSample accessions and thus, the appropriate information was obtained by combing through the linked BioProject and Sequence Read Archive accessions for each project. Pertinent information was mainly obtained by examining the longer text descriptors like "project description" which often contained information such as the targeted gene region or sequencing machine used in the project. As host information was provided at different taxonomic resolutions, all the host information provided in the downloaded BioSample accessions were assigned to a consistent taxonomic hierarchy: from order to species, where possible. Biases in host taxonomy (i.e. whether some plant families are more often represented than others) were assessed at the level of plant family. The growth form of each taxonomic unit was assigned by performing a Google search and deciding whether host growth form fell into one of five categories: grass, bamboos or sedges; herbaceous forbs; trees, woody shrubs or lianas; ferns, mosses or liverworts or unknown/unassignable.

3. Results and discussion

3.1. *Delay in sample availability*

On average, there was a 2.74-year delay from when samples were collected to when the corresponding sequencing data became publicly available on NCBI (Fig. 2a). The delay in sequence availability appears to be getting longer (Fig. 2b) although this trend was not significant ($F_{1,184} = 1.938$, $p = 0.166$, $R^2 = 0.011$). The length of the delay until sample availability on public repositories is likely due to researchers placing an embargo on their data until their datasets are published due to a fear of being "scooped". However, we observed that many of the BioProject accessions became publicly available before research articles were published on said data, sometimes up to five years before articles were published (Supplementary Table S1) highlighting that being "scooped" may happen less than researchers fear. Additionally, the delay between data collection and publishing means that valuable data cannot be re-used timeously, possibly leading to a duplication of research efforts and posing an impediment to knowledge generation and innovation (Penev et al., 2017; Wilkinson et al., 2016). Therefore, we posit that the swift publication of metabarcoding datasets (before analyses for further publications), in journals which specifically handle biological and other datasets (e.g. Biodiversity Data Journal or Scientific Data) will ultimately provide a means for credit to be afforded to data-generators, i.e. a citable reference and DOI. Moreover, the ability for credit to be afforded to data-generators may entice researchers to upload their sequencing data quicker and ultimately facilitate the timeous "reusability" encompassing the FAIR data principles (GO FAIR, 2022; Wilkinson et al., 2016). Additionally, delayed availability of plant-associated fungal sequence data on public repositories poses serious

Table 1 – The percentage of missing, correctly, and incorrectly supplied metadata, compared against the Genomics Standards Consortium (GSC) definitions for each of the eight attributes considered in this study. Only studies of fungal communities associated with a plant host were considered. Percentages are calculated as the proportion of the 17,921 accessions in each category. MIxS complaint attribute values are bold.

MIxS ID and Term name															
MIxS:0000009 Geographic location (latitude and longitude)		MIxS:0000010 Geographic location (country and/or sea region)		MIxS:0000011 Collection date		MIxS:0000012 Broad-scale environmental context		MIxS:0000013 Local-scale environmental context		MIxS:0000014 Environmental medium		MIxS:0000029 Host scientific name		MIxS:0000093 Elevation	
No coordinates	5.2%	No locality information	0.0%	No collection date	0.5%	No broad-scale environmental context	25.5%	No local-scale environmental context	25.6%	No environmental medium provided	1.2%	No host information provided	4.1%	No elevation	70.8%
one decimal	0.1%	Region only	18.4%	Range of dates	4.3%	Broad- and local-scale context the same	3.9%	Broad- and local-scale context the same	3.9%	Environmental medium the same as broad- or local-scale context	0.4%	Pooled host information (i.e., collected from several hosts)	0.3%	**Elevation provided**	29.2%
two decimals	21.2%	Country only	12.5%	Year only	7.8%	Context more specific than local-scale context	11.1%	Context broader than broad-scale context	11.2%	Too broad, i.e., plant, tree, or host name	49.2%	Nondescript host information (i.e., tree, plant, vegetation)	5.4%	-	-
three decimals	23.8%	**Country and region**	69.1%	Month and year	61.6%	Context too focused	0.5%	Context too broad	0.1%	**Plant tissue provided as environmental medium**	49.2%	Host common name	1.0%	-	-
four decimals	18.3%	-	-	**Day month and year**	25.9%	**"Biome" - MIxS suggested**	59.0%	**Local-scale context provided and of a finer spatial grain**	59.2%	-	-	Only host order	0.2%	-	-
five decimals	7.3%	-	-	-	-	-	-	-	-	-	-	Only host family	4.8%	-	-
six decimals	12.4%	-	-	-	-	-	-	-	-	-	-	Only host genus	14.9%	-	-
seven decimals	2.1%	-	-	-	-	-	-	-	-	-	-	**Scientific name**	69.3%	-	-
eight decimals	9.6%	-	-	-	-	-	-	-	-	-	-	-	-	-	-

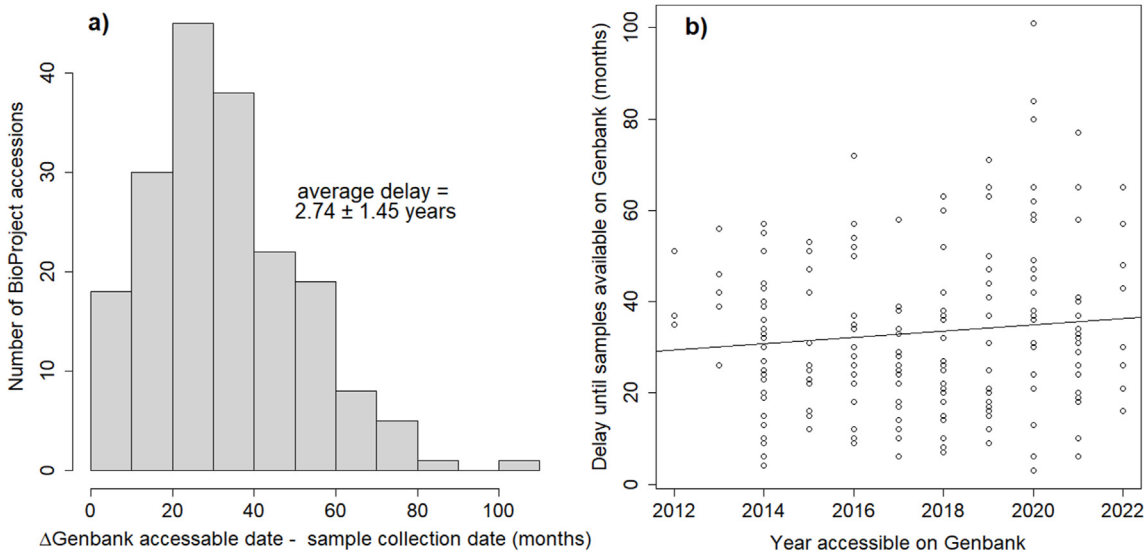

Fig. 2 – a) Histogram showing the delay, in months, from when samples are collected to when samples are publicly available on GenBank. b) The trend of the delay until samples are publicly available on GenBank over time.

problems for the biosecurity of many countries as early detection and action against potentially devastating pathogens is essential to managing outbreaks. If data are not available, responses to natural or agricultural pandemics are delayed which may result in severe economic losses.

3.2. *Missing metadata*

The BioSample attribute with the most missing values was elevation, with 71% of these BioSample accessions missing information on this attribute (Table 1) echoing findings from other studies where elevation was provided less than 17% of the time (Durkin et al., 2020). This is unsurprising as elevation is the only attribute assessed in this study which is not specified as a "mandatory" field during BioSample submission. If accurate locality data are provided, obtaining elevation data from digital elevation models, of which several are freely available at fine resolutions (e.g. 30 m - https://gdemdl.aster.jspacesystems.or.jp/), is fairly easy. For the other seven attributes assessed here, less than 26% of the accessions had missing values (Table 1): local-scale environmental context had the most missing information – 25.6%, followed by broad-scale environmental context – 25.5%, coordinates – 5.2%, host – 4.1%, environmental medium – 1.2% and collection date 0.5% (Table 1). Only country and/or sea region had no missing values (Table 1).

One reason why some metadata may be missing for certain attributes relates to when samples were added to NCBI. For example, broad-scale environmental context and local-scale environmental context, only became mandatory metadata attributes after the publication of the Minimum Information about a MARKer gene Sequence (MIMARKS) packages in 2011 (Yilmaz et al., 2011). Therefore, any sequences and their associated metadata added to NCBI before this time likely used environmental packages which did not require these attributes to be submitted. In many cases when attributes are not mandatory they are often not provided by users (Gonçalves and Musen, 2019; Wang et al., 2019). Alternatively, in the "generic environmental package" no attribute fields are mandatory and users who select this environmental package generally have more incomplete metadata than those researchers who select environmental packages encompassed by MIMARKS (e.g. plant-associated) where certain attributes are theoretically mandatory (Gonçalves and Musen, 2019). However, it is important to note that even though these metadata attributes are theoretically mandatory, whether they are included or not is not strictly enforced by NCBI, and therefore explains why values for some attributes are missing (Gonçalves and Musen, 2019). In a positive development, INSDC recently announced that, from June 2023, it will become mandatory for country and collection date (spatiotemporal) metadata to be provided for all BioSample submissions, with these requirements being strictly enforced (https://www.insdc.org/news/insdc-spatiotemporal-metadata-missing-values-update-03-04-2023/). This is a promising start, hopefully paving the way for other metadata attributes to also be strictly enforced by INSDC members in future.

3.3. *Correctly submitted metadata*

On average 57% of the user-supplied metadata was of the so-called gold-standard (i.e. met all the criteria set out in their respective MIxS definitions; Supplementary Table S2) for the eight attributes assessed here (Table 1). However, this ranged significantly between attributes, from as high as 94.8% (geographic location – latitude and longitude) to as low as 25.9% (collection date) (Table 1). It is important to note here that while the "coordinates" appear well captured, the criterion was considered met based on the MIxS definition regardless of whether only one, or whether eight decimal places were provided (Supplementary Table S2). What is promising is that 73.5% of the BioSample accessions had coordinates

with three decimal places or more – meaning that for most of the accessions, information is to within a kilometre accuracy. All coordinates provided mapped within the boundaries of the country specified in the geographic location (country and/or sea region) attribute. The percentage of correctly specified metadata for the remaining six attributes assessed here were: host – 69.3% (i.e. scientific name), followed by geographic location – 69.1% (i.e. country and region provided), local-scale environmental context – 59.2% (i.e. is of smaller spatial grain than broad-scale environmental context and specifies an entity that realistically has influence on the samples collected), followed by broad-scale environmental context – 59% (i.e. "biome"), environmental medium – 49.2% (i.e. plant tissue from which fungal DNA was extracted) and elevation – 29.2% (i.e. elevation provided) (Table 1).

3.4. *Common mistakes and loss of resolution in metadata*

3.4.1. *Common mistakes*

For six of the eight attributes common mistakes were found, which did not adhere to the MixS standard definitions provided for each of the metadata attributes (Supplementary Table S2). Here we highlight the mistakes for six attributes: geographic location [country or sea region], collection date, broad-scale environmental context, local-scale environmental context, environmental medium, and host [scientific name] (Table 1). For geographic location 12.5% only provided country information and 18.4% provided only region information (Table 1), where the MIxS definition sets out that country and local region be provided (Supplementary Table S2). The majority (61.6%) of collection date accessions provided only month and year, 7.8% provided only year and 4.3% provided a range of dates (Table 1). None of these collection dates are ISO8601 compliant and are therefore considered incomplete. For the broad-scale environmental context attribute, 0.5% of the accessions were too narrowly focused (e.g. "phyllosphere" instead of biome), 11.1% of the accessions were more focused than the local-scale environmental context and 3.9% of the accessions were the same from broad- and local-scale contexts (Table 1). This was similar for local-scale environmental context accessions with 0.1% having a context too broad (e.g. "forest" instead of something like forest canopy or forest understorey as suggested in the MIxS definition for this attribute), 11.2% of the accessions had a context broader than the broad-scale environmental context (e.g. "forest" when broad-scale environmental context was "evergreen pine forest") and 3.9% of the accessions were identical for broad- and local-scale environmental contexts (Table 1). For the environmental medium attribute, 49.2% of the accessions were too broad (e.g. tree/plant/host instead of roots/leaves/stem, etc.) and 0.4% of the accessions had the same values as broad- or local-scale environmental context (Table 1). Lastly, for host 26.7% of the accessions only provided host genus, family, order, common names, or pooled host information (Table 1).

3.4.2. *Lost resolution in the metadata*

For several studies, the value of supplied metadata was reduced due to poor metadata capturing practices for individual BioSample accessions within one BioProject. Below we highlight how poor metadata practices results in lost resolution of geographic, taxonomic and date metadata attributes and what it implies for data reuse.

3.4.2.1. *Geographic*

Although most accessions provided geographic coordinates, 62% of BioProjects provided identical geographic coordinates for all their BioSample accessions within one BioProject. Of the 17,921 BioSample accessions we obtained, only 8% of the coordinates were representative of unique or two BioSample accessions, while 86% of the coordinates were duplicates, i.e. identical coordinates were provided for multiple accessions. The number of BioSample accessions per BioProject with identical coordinates ranged between three and 2013. This lost resolution has implications for when data are pooled to try answering large-scale biogeographical questions. First, the diversity of a particular area may be overestimated, as it is well known that the species richness of fungi increases with increasing sampling depth (Peay et al., 2016). Additionally, loss in spatial resolution of the samples results in uncertainty in models assessing patterns and drivers of diversity (Daru et al., 2018; Greve et al., 2016). All in all, the loss of geographic resolution in coordinates, along with the uncertainty in the accuracy that this creates, may make the data of little use for spatial biogeographical applications (Dobson et al., 2020).

3.4.2.2. *Taxonomic*

Some loss in taxonomic resolution of host plants was also observed. While more than two thirds (69.3%) of the BioSample accessions had host information at the species level, 14.9% of the accessions only had genus information, 4.8% only provided host information to the family level and 0.2% gave host information to the level of order. Some 6.7% of the BioSample accessions provided host information as common names, pooled hosts, or nondescript host information like plant/tree/vegetation. For the 1.0% of accessions which provided only common names in the host attribute field, most hosts were agricultural plants like "wheat" or "apple tree", for which at least the genus should be determinable. Information on host at as high a taxonomic resolution as possible is extremely important in host-based fungal studies as host identity is often the most important determinant of fungal composition and diversity (Harris et al., 2023; U'Ren et al., 2019).

3.4.2.3. *Date*

In terms of collection dates, 25.9% of the BioSample accessions provided the year, month and day that samples were collected, 61.6% provided only year and month and 7.8% of accessions provided the year alone. A further 4.3% of the accessions provided a range of dates on which samples were collected (Table 1). Date information is vital for understanding temporal patterns, both seasonal patterns and long-term trends and their drivers, and is essential in monitoring the effects of global change drivers such as climate change (Greve et al., 2016). Therefore, we recommend that researchers should include at the very least month and year that samples were collected (Genomic Standards Consortium, 2022).

3.5. *Collection biases*

More than 80% of plant associated fungal BioSample accessions came from the northern hemisphere (Fig. 3), mostly from temperate zones of North America, Europe and Asia with far fewer accessions from southern hemisphere regions and from the tropics. Africa was particularly poorly sampled (Fig. 3). These geographic biases are similar to those that have been found for macro-organism taxa, but nevertheless represent a major gap in our current knowledge for the kingdom Fungi (Daru et al., 2018; Peay et al., 2016). Therefore, more effort to sample the global South and particularly the African continent will be essential not only to perform global biogeographical studies on plant-associated fungi, but to obtain valuable baseline data from the continent.

Fungal communities differ between above- and belowground domains and many taxa are restricted to one plant organ (Compant et al., 2019; Peay et al., 2016). Most sample accessions isolated fungi from the root material of plants, followed by leaf material. Much fewer accessions contained data from branches, stems, seeds, flowers and pollen (Fig. 4a). Since there are major differences between the fungal communities of different plant organs, these biases would need to be considered when pooling data to perform any large-scale analyses (Compant et al., 2019; Peay et al., 2016).

Fungal communities were not sampled equally across different plant families. Some plant families were sampled disproportionately compared to others, with the plant families Fagaceae and Poaceae representing >12% and Pinaceae, Ericaceae and Nothofagaceae representing >4% of the 17,921 BioSample accessions (Fig. 4b). These families represent commercially important species and it is therefore unsurprising that they are well represented.

Sequencing instruments influence the communities that are recovered during sequencing (Ramirez et al., 2018). The effects of using different sequencing instruments should thus be considered when pooling sequence data as diversity estimates and species composition differ based on sequencing instrument used (Callahan et al., 2017). Most of the plant-associated fungal sequencing data was sequenced using Illumina instruments, followed by Roche 454 instruments and few samples were sequenced using Ion torrent and PacBio sequencing machines (Fig. 4c). However, we do expect that 454 sequencing is largely considered to be redundant, and that longer read PacBio and Oxford nanopore sequencing will become more widely used over time as longer reads have been found to produce more stable and robust taxonomic assignments (Heather and Chain, 2016; Runnel et al., 2022).

Pooling samples from different datasets requires communities to be quantified using the same genomic region, as only samples amplified using the same genomic region can be directly compared when assigning sequences to OTUs or ASVs (Amir et al., 2017; Callahan et al., 2016). The ITS region, which is the universal barcoding region for fungi (Schoch et al., 2012), was the most frequently used to amplify fungal communities with the 18S (SSU) and 28S (LSU) gene regions used more infrequently. (Fig. 4d).

The growth form of the host of each BioSample accession was assigned to determine whether biases existed in which growth forms have typically had their fungal communities quantified. Approximately 66% of the BioSample accessions were taken from species with woody growth forms i.e. trees, woody shrubs and lianas (Fig. 4e). Samples from the other growth forms all represented less than 12% of the BioSample accessions.

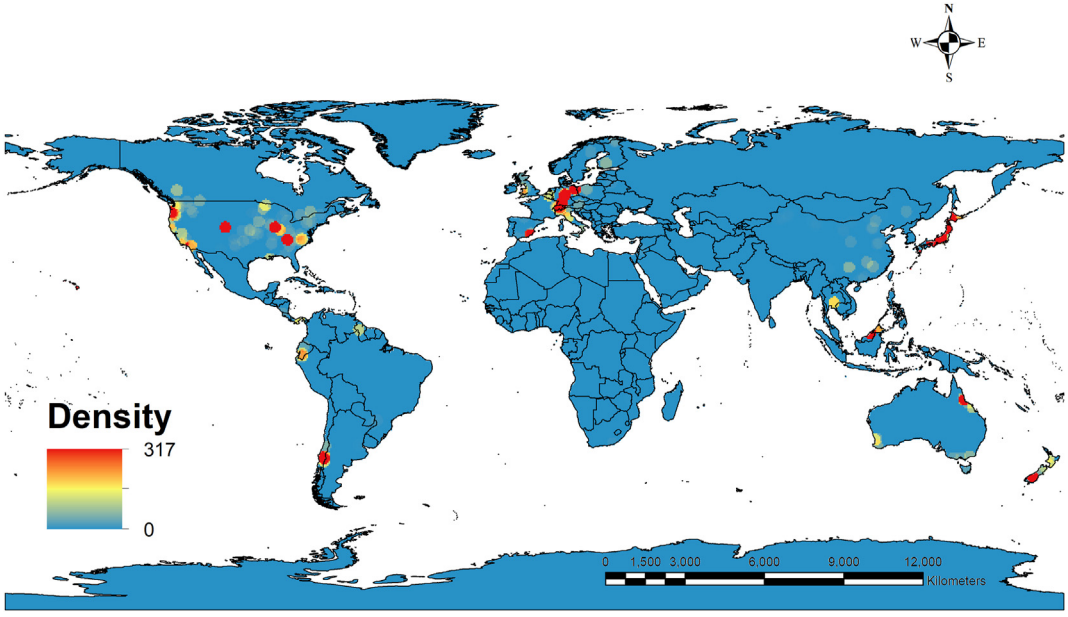

Fig. 3 – The density of plant associated BioSample accessions per 2 by 2° cell globally. Density refers to the number of plant-associated BioSample accessions that have been sampled within a 2° latitude and 2° longitude cell of the earth's surface.

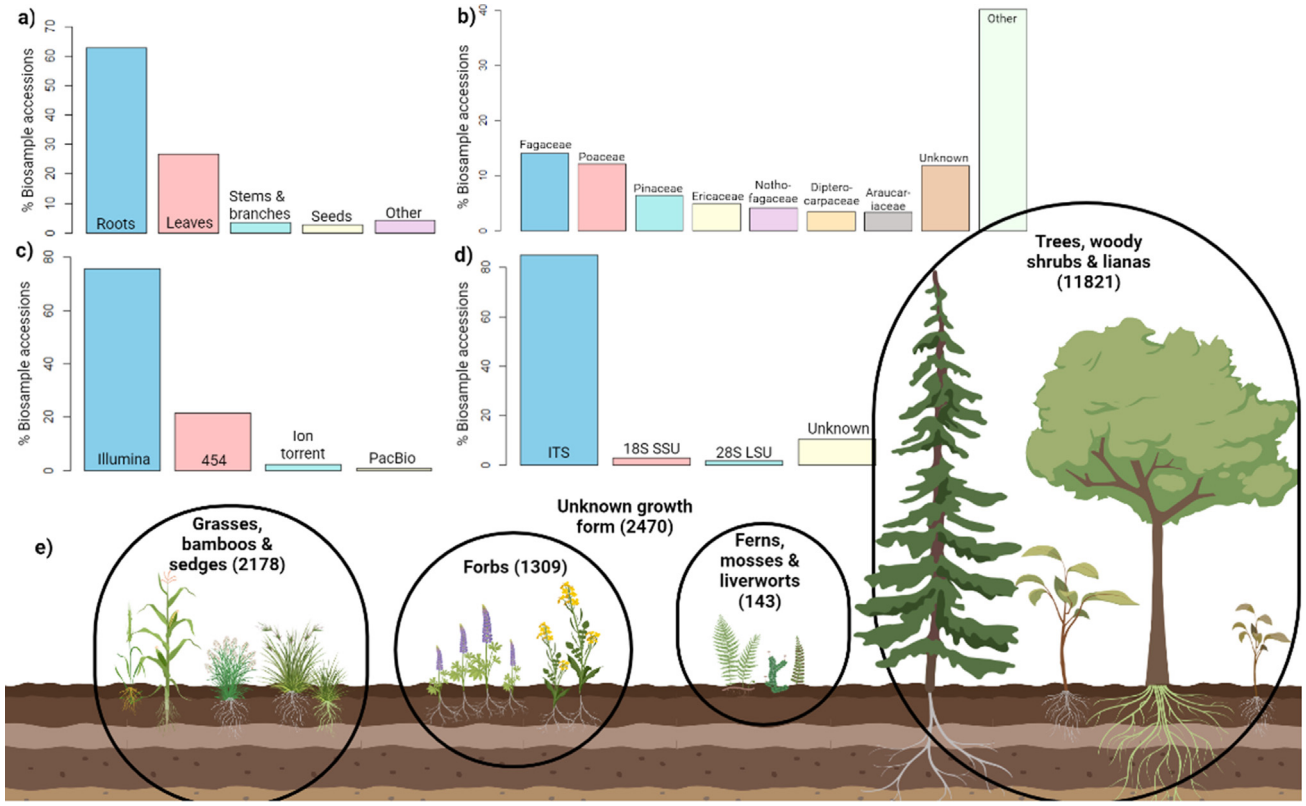

Fig. 4 – The bar charts represent the percentage of the 17,921 downloaded plant associated fungal BioSample accessions that could be assigned a) to the different plant organs from which fungi were extracted, b) the host plant family, c) the sequencing instrument on which the samples were quantified, d) the genomic region used to amplify the fungal communities from a sample, and e) the number of BioSample accessions that could be assigned to different growth forms.

4. Conclusions

The major hindrance of fungal sequencing data re-use was, until recently, the inability to pool sequence data from different datasets because of the way that OTUs were quantified. But now the use of open-reference clustering which produce stable OTUs or the use of ASVs have largely resolved this issue (Tedersoo et al., 2022a). The rapid accumulation of fungal metabarcoding data that are stored on public repositories like NCBI can potentially be used to explore large scale patterns of fungal biogeography and form valuable baseline data to inform policies on conservation initiatives in the face of rising species extinctions (Greve et al., 2016; Nilsson et al., 2019; Větrovský et al., 2020). However, the ability to pool data also requires well-captured metadata. Data collated from haphazardly-collected datasets always comes with biases and uncertainties, but methods and decision trees have been developed to guide usage of such data (e.g. Dobson et al., 2020). Before blindly using data from such haphazardly-collected datasets, it is essential for researchers to assess whether the benefits of using said data outweighs the costs required to obtain, manage, curate and address biases inherent in these data (Dobson et al., 2020; Pocock et al., 2014). Therefore, continued efforts are required to ensure consistently recorded metadata attributes, standardised methods for how samples are collected and studies that confront biases (Chiarello et al., 2022; Gonçalves and Musen, 2019).

As fungal datasets grow, it will be essential to maintain well managed public repositories like members of INSCD, in such a manner that they are able to facilitate FAIR data principles (GO FAIR, 2022; Wilkinson et al., 2016). The FAIRness of such data will not only be governed by the quality of the metabarcoding data, but by the quality and completeness of the associated metadata (Nilsson et al., 2022). While much time and effort has gone into creating standards with well-defined requirements for metadata submitted to databases of the INSCD members (Genomic Standards Consortium, 2022), these well specified requirements are not strictly enforced upon submission to these databases (Gonçalves and Musen, 2019; Klie et al., 2021). The consequence is that metadata stored in INSCD databases are not standardised, leading to difficulties when searching for related datasets. This ultimately reduces the FAIRness of the data landscape, and is something the mycological community has put emphasis on trying to avoid going forward (Nilsson et al., 2022).

Here we provide further arguments and information to support the calls by other researchers to ensure that future datasets comply with the FAIR principles. Firstly, data producers working on metabarcoding data from environmental samples should try familiarise themselves with the definitions provided

on the minimum information about any (x) sequence (MIxS) standards published by the Genomic Standards Consortium (http://www.gensc.org/pages/standards-intro.html) (Miralles et al., 2020; Nilsson et al., 2022; Penev et al., 2017). Additionally, a concerted effort by users to try to utilise standardised vocabularies which are MIxS or Darwin Core standard compliant for supplied metadata will vastly improve the FAIRness of their submitted data (Nilsson et al., 2022; Penev et al., 2017; Tedersoo et al., 2015). Public repositories like those housed by INSCD members may need to put into place a system where non-compliant metadata attributes are automatically flagged during submission, and only once these flagged accessions are rectified will the supplied metabarcoding data and linked metadata be accepted (Miralles et al., 2020). Lastly, as the power and usability of AI grows, it seems only logical that researchers will be able to retrospectively fix metadata already uploaded onto databases which did not meet metadata standards.

Correctly supplied metadata that adheres to all standards is a complex issue to implement at present that cannot solely be seen as the responsibility of data providers. To entice the rapid uploading of data, incentives to researchers who supply correct, well-defined metadata to public repositories, particularly in the form of citations, may be necessary. These citations that give credit to datasets will need to be viewed more equivalently to research article citations (especially by hiring committees, scientific boards, and funding agencies) if we hope to see true change in the quality and speed of data availability (Abarenkov et al., 2022; Durkin et al., 2020; Nilsson et al., 2022; Penev et al., 2017). Fortunately, this topic is receiving increasing attention. Journals that are solely devoted to the publishing of datasets, which can then be cited and the appropriate credit afforded to the data producers, e.g. the Biodiversity Data Journal and Scientific Data (Penev et al., 2017) are emerging. Additionally, national funding bodies have started to advocate for a more open science that considers other forms of publications, including published data sets, for grant application assessments which will help entice authors to publish their data timeously. Journal editors and reviewers also have a critical role to play by ensuring strictly enforced journal policies regarding both raw data and metadata standards and sharing which meets international standards. Continued efforts by all stakeholders will be essential to meeting the FAIR data principles, so that knowledge generation and innovation are not hindered. Thus, the time is ripe to utilise this vast collection of fungal occurrence data to bring fungal biogeography to the forefront of biogeographical research so that we may advance our knowledge of the kingdom Fungi, reflecting the vital role of its members in ecosystems from the poles to the tropics.

Declaration of competing interest

We know of no conflicts of interest that are associated with this potential publication. Additionally, we declare that there has been no significant financial support for conducting this research that could have influenced its overall outcome. As the corresponding author, I confirm that the entire manuscript with all supplementary materials has been read and approved for submission by all authors named above.

Acknowledgements

The authors would like to thank the reviewers for their helpful insights and suggestions for the manuscript. Additionally, the authors would also like to thank the Tree Protection Co-operative Program (TPCP), the center of excellence in plant health and biotechnology (CPHB) and the South African National Research Foundation (Grant numbers 98889 and 116333) for funding this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fbr.2023.100329.

REFERENCES

Abarenkov, K., Adams, R.I., Irinyi, L., Agan, A., Ambrosio, E., Antonelli, A., Bahram, M., Bengtsson-Palme, J., Bok, G., Cangren, P., Coimbra, V., Coleine, C., Gustafsson, C., He, J., Hofmann, T., Kristiansson, E., Larsson, E., Larsson, T., Liu, Y., Martinsson, S., Meyer, W., Panova, M., Pombubpa, N., Ritter, C., Ryberg, M., Svantesson, S., Scharn, R., Svensson, O., Topel, M., Unterseher, M., Visagie, C., Wurzbacher, C., Taylor, A.F.S., Kõljalg, U., Schriml, L., Nilsson, R.H., 2016. Annotating public fungal ITS sequences from the built environment according to the MIxS-Built Environment standard – a report from a May 23-24 , 2016 workshop (Gothenburg, Sweden). MycoKeys 16, 1–15. https://doi.org/10.3897/mycokeys.16.10000.

Abarenkov, K., Kristiansson, E., Ryberg, M., Nogal-Prata, S., Gomez-Martinez, D., Stuer-Patowsky, K., Jansson, T., Põlme, S., Ghobad-Nejhad, M., Corcoll, N., Scharn, R., Sanchez-Garcia, M., Khomich, M., Wurzbacher, C., Nilsson, R.H., 2022. The curse of the uncultured fungus. MycoKeys 86, 177–194.

Ackerly, D.D., Loarie, S.R., Cornwell, W.K., Weiss, S.B., Hamilton, H., Branciforte, R., Kraft, N.J.B., 2010. The geography of climate change: implications for conservation biogeography. Divers. Distrib. 16, 476–487. https://doi.org/10.1111/j.1472-4642.2010.00654.x.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A., Knight, R., 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2e00191-16. https://doi.org/10.1128/msystems.00191-16.

Andrew, C., Heegaard, E., Kirk, P.M., Assler, C.B., Heilmann-clausen, J., Krisai-greilhuber, I., Kuyper, T.W., Senn-irlet, B., Untgen, U.B., Diez, J., Egli, S., Gange, A.C., Halvorsen, R., Rustøen, F., Boddy, L., Høiland, K., En, J.N., 2017. Big data integration: pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. Fungal Biol. Rev. 31, 88–98. https://doi.org/10.1016/j.fbr.2017.01.001.

Arita, M., Karsch-Mizrachi, I., Cochrane, G., 2021. The international nucleotide sequence database collaboration. Nucleic Acids Res. 41, D121–D124. https://doi.org/10.1093/nar/gkaa967.

Arnold, A.E., Lutzoni, F., 2007. Diversity and host range of foliar fungal endophytes: are tropical leaves biodiversity hotspots? Ecology 88, 541–549. https://doi.org/10.1890/05-1459.

Bahram, M., Põlme, S., Kõljalg, U., Zarre, S., Tedersoo, L., 2012. Regional and local patterns of ectomycorrhizal fungal diversity and community structure along an altitudinal gradient in the Hyrcanian forests of northern Iran. New Phytol. 193, 465–473. https://doi.org/10.1111/j.1469-8137.2011.03927.x.

Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., Bengtsson-Palme, J., Anslan, S., Coelho, L.P., Harend, H., Huerta-Cepas, J., Medema, M.H., Maltz, M.R., Mundra, S., Olsson, P.A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., Tedersoo, L., Bork, P., 2018. Structure and function of the global topsoil microbiome. Nature 560, 233–237. https://doi.org/10.1038/S41586-018-0386-6.

Barrett, T., 2013. BioSample. In: The NCBI Handbook [Internet]. National Center for Biotechnology Information (US), Bethesda (MD), pp. 1–6.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K., Resenchuk, S., Tatusova, T., Yaschenko, E., Ostell, J., 2012. BioProject and BioSample databases at NCBI: facilitating capture and organisation of metadata. Nucleic Acids Res. 40, 57–63. https://doi.org/10.1093/nar/gkr1163.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., Abebe, E., 2005. Defining operational taxonomic units using DNA barcode data. Philos. Trans. R. Soc. B Biol. Sci. 360, 1935–1943. https://doi.org/10.1098/rstb.2005.1725.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K. Bin, Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

Brown, E.D., Williams, B.K., 2019. The potential for citizen science to produce reliable and useful information in ecology. Conserv. Biol. 33, 561–569. https://doi.org/10.1111/cobi.13223.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869.

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11. https://doi.org/10.1038/ismej.2017.119, 2693–2643.

Ceballos-Escalera, A., Richards, J., Arias, M.B., Inward, D.J.G., Vogler, A.P., 2022. Metabarcoding of insect-associated fungal communities: a comparison of internal transcribed spacer (ITS) and large-subunit (LSU) rRNA markers. MycoKeys 88, 1–33. https://doi.org/10.3897/mycokeys.88.77106.

Chalmers, N., Henderson, P., 1998. Raising the relevance to outside needs. Nature 394, 118. https://doi.org/10.1038/28019.

Chandler, M., See, L., Copas, K., Bonde, A.M.Z., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A., Turak, E., 2017. Contribution of citizen science towards international biodiversity monitoring. Biol. Conserv. 213, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004.

Chiarello, M., McCauley, M., Villéger, S., Jackson, C.R., 2022. Ranking the biases: the choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. PLoS One 17e0264443. https://doi.org/10.1371/journal.pone.0264443.

Clark, K., Pruitt, K., Tatusova, T., Mizrachi, I., 2013. BioProject. In: The NCBI Handbook [Internet]. National Center for Biotechnology Information (US), Bethesda (MD), pp. 1–6.

Cline, L.C., Song, Z., Al-Ghalith, G.A., Knights, D., Kennedy, P.G., 2017. Moving beyond de novo clustering in fungal community ecology. New Phytol. 216, 629–634. https://doi.org/10.1111/nph.14752.

Compant, S., Samad, A., Faist, H., Sessitsch, A., 2019. A review on the plant microbiome: ecology, functions, and emerging trends in microbial application. J. Adv. Res. 19, 29–37. https://doi.org/10.1016/j.jare.2019.03.004.

Cowan, D.A., Lebre, P.H., Amon, C., Becker, R.W., Boga, H.I., Boulangé, A., Chiyaka, T.L., Coetzee, T., de Jager, P.C., Dikinya, O., Eckardt, F., Greve, M., Harris, M.A., Hopkins, D.W., Houngnandan, H.B., Houngnandan, P., Jordaan, K., Kaimoyo, E., Kambura, A.K., Kamgan-Nkuekam, G., Makhalanyane, T.P., Maggs-Kölling, G., Marais, E., Mondlane, H., Nghalipo, E., Olivier, B.W., Ortiz, M., Pertierra, L.R., Ramond, J.B., Seely, M., Sithole-Niang, I., Valverde, A., Varliero, G., Vikram, S., Wall, D.H., Zeze, A., 2022. Biogeographical survey of soil microbiomes across sub-Saharan Africa: structure, drivers, and predicted climate-driven changes. Microbiome 10, 131. https://doi.org/10.1186/s40168-022-01297-w.

Lang, D.T.CRAN Team, 2019. XML: Tools for Parsing and Generating XML within R and S-Plus.

Daru, B.H., Park, D.S., Primack, R.B., Willis, C.G., Barrington, D.S., Whitfeld, T.J.S., Seidler, T.G., Sweeney, P.W., Foster, D.R., Ellison, A.M., Davis, C.C., 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytol. 217, 939–955. https://doi.org/10.1111/nph.14855.

Davison, J., Moora, M., Öpik, M., Adholeya, A., Ainsaar, L., Bâ, A., Burla, S., Diedhiou, a G., Hiiesalu, I., Jairus, T., Johnson, N.C., Kane, A., Koorem, K., Kochar, M., Ndiaye, C., Pärtel, M., Reier, Ü., Saks, Ü., Singh, R., Vasar, M., Zobel, M., 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 127, 970–973. https://doi.org/10.1126/science.aab1161, 80.

De Kort, H., Prunier, J.G., Ducatez, S., Honnay, O., Baguette, M., Stevens, V.M., Blanchet, S., 2021. Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. Nat. Commun. 12. https://doi.org/10.1038/s41467-021-20958-2.

Dobson, A.D.M., Milner-Gulland, E.J., Aebischer, N.J., Beale, C.M., Brozovic, R., Coals, P., Critchlow, R., Dancer, A., Greve, M., Hinsley, A., Ibbett, H., Johnston, A., Kuiper, T., Le Comber, S., Mahood, S.P., Moore, J.F., Nilsen, E.B., Pocock, M.J.O., Quinn, A., Travers, H., Wilfred, P., Wright, J., Keane, A., 2020. Making messy data work for conservation. One Earth 2, 455–465. https://doi.org/10.1016/j.oneear.2020.04.012.

Durkin, L., Jansson, T., Sanchez, M., Khomich, M., Ryberg, M., Kristiansson, E., Nilsson, R.H., 2020. When mycologists describe new species, not all relevant information is provided (clearly enough). MycoKeys 72, 109–128. https://doi.org/10.3897/mycokeys.72.56691.

Elith, J., Leathwick, J., 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Divers. Distrib. 13, 265–275. https://doi.org/10.1111/j.1472-4642.2007.00340.x.

Ellis, C.J., Coppins, B.J., Dawson, T.P., Seaward, M.R.D., 2007. Response of British lichens to climate change scenarios: trends and uncertainties in the projected impact for contrasting biogeographic groups. Biol. Conserv. 140, 217–235. https://doi.org/10.1016/J.BIOCON.2007.08.016.

Genomic Standards Consortium, 2022. GSC Defined Terms [WWW Document]. URL. http://www.gensc.org/pages/standards/all-terms.html. accessed September.28.22.

GO FAIR, 2022. FAIR Principles [WWW Document]. URL. https://www.go-fair.org/fair-principles/. accessed September.28.22.

Gonçalves, R.S., Musen, M.A., 2019. Analysis: the variable quality of metadata about biological samples used in biomedical experiments. Sci. Data 6, 1–15. https://doi.org/10.1038/sdata.2019.21.

Greve, M., Lykke, A.M., Fagg, C.W., Bogaert, J., Friis, I., Marchant, R., Marshall, A.R., Ndayishimiye, J., Sandel, B.S., Sandom, C., Schmidt, M., Timberlake, J.R., Wieringa, J.J., Zizka, G., Svenning, J.C., 2012. Continental-scale variability in browser diversity is a major driver of diversity patterns in acacias across Africa. J. Ecol. 100, 1093–1104. https://doi.org/10.1111/j.1365-2745.2012.01994.x.

Greve, M., Lykke, A.M., Fagg, C.W., Gereau, R.E., Lewis, G.P., Marchant, R., Marshall, A.R., Ndayishimiye, J., Bogaert, J., Svenning, J.C., 2016. Realising the potential of herbarium records for conservation biology. South Afr. J. Bot. 105, 317–323. https://doi.org/10.1016/j.sajb.2016.03.017.

Harris, M.A., Kemler, M., Slippers, B., Jamison-Daniels, S.-L., Witfeld, F., Botha, M., Begerow, D., Brachmann, A., Greve, M., 2023. Deterministic processes have limited impacts on foliar fungal endophyte communities along a savanna-forest successional gradient. Fungal Ecol. 64, 101249. https://doi.org/10.1016/j.funeco.2023.101249.

Hawksworth, D.L., 2001. The magnitude of fungal diversity: the 1.5 million. Mycol. Res. 105, 1422–1432. https://doi.org/10.1017/S0953756201004725.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: the history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Heberling, J.M., Isaac, B.L., 2017. Herbarium specimens as exaptations: new uses for old collections. Am. J. Bot. 104, 963–965. https://doi.org/10.3732/ajb.1700125.

Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Rodney Brister, J., O'Sullivan, C., 2022. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 50, D387–D390. https://doi.org/10.1093/nar/gkab1053.

Klich, M.A., 2002. Biogeography of Aspergillus species in soil and litter. Mycologia 94, 20–27. https://doi.org/10.1080/15572536.2003.11833245.

Klie, A., Tsui, B.Y., Mollah, S., Skola, D., Dow, M., Hsu, C.N., Carter, H., 2021. Increasing metadata coverage of SRA BioSample entries using deep learning-based named entity recognition. Database 2021, 1–11. https://doi.org/10.1093/database/baab021.

Korpelainen, H., Pietiläinen, M., Huotari, T., 2016. Effective detection of indoor fungi by metabarcoding. Ann. Microbiol. 66, 495–498. https://doi.org/10.1007/s13213-015-1118-x.

Lavoie, C., 2013. Biological collections in an ever changing world: herbaria as tools for biogeographical and environmental studies. Perspect. Plant Ecol. Evol. Systemat. 15, 68–76. https://doi.org/10.1016/j.ppees.2012.10.002.

Linnaeus, C. von, 1758. Systema Naturae Per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis, tenth ed. Holmiae (Salvius), Stockholm.

Liu, Y., Fu, B., Wang, S., Zhao, W., 2018. Global ecological regionalization: from biogeography to ecosystem services. Curr. Opin. Environ. Sustain. 33, 1–8. https://doi.org/10.1016/j.cosust.2018.02.002.

Lomolino, M., Riddle, B., Whittaker, R., Brown, J.H., 2010. Biogeography, sixth ed. Oxford University Press, Sunderland.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., Chilquillo, E., Rønsted, N., Antonelli, A., 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecol. Biogeogr. 24, 973–984. https://doi.org/10.1111/geb.12326.

Meineke, E.K., Davies, T.J., Daru, B.H., Davis, C.C., 2018. Biological collections for understanding biodiversity in the Anthropocene. Philos. Trans. R. Soc. B 374, 20170386. https://doi.org/10.1098/rstb.2017.0386.

Miralles, A., Bruy, T., Wolcott, K., Scherz, M.D., Begerow, D., Beszteri, B., Bonkowski, M., Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F.O., Hawlitschek, O., Kostadinov, I., Nattkemper, T.W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., Wilke, T., Renner, S.S., Vences, M., 2020. Repositories for taxonomic data : where we are and what is missing. Syst. Biol. 69, 1231–1253. https://doi.org/10.1093/sysbio/syaa026.

Mueller, G.M., Schmit, J.P., Leacock, P.R., Buyck, B., Cifuentes, J., Desjardin, D.E., Halling, R.E., Hjortstam, K., Iturriaga, T., Karlhenrik, L., Lodge, D.J., May, T.W., Minter, D., Rajchenberg, M., Redhead, S.A., Ryvarden, L., Trappe, J.M., Watling, R., Wu, Q., 2007. Global diversity and distribution of macrofungi. Biodivers. Conserv. 16, 37–48. https://doi.org/10.1007/s10531-006-9108-8.

Nemergut, D.R., Schmidt, S.K., Fukami, T., O'Neill, S.P., Bilinski, T.M., Stanish, L.F., Knelman, J.E., Darcy, J.L., Lynch, R.C., Wickey, P., Ferrenberg, S., 2013. Patterns and processes of microbial community assembly. Microbiol. Mol. Biol. Rev. 77, 342–356. https://doi.org/10.1128/MMBR.00051-12.

Nilsson, R.H., Larsson, K., Taylor, A.F.S., Bengtsson-palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Oliver, F., Tedersoo, L., Saar, I., Urmas, K., 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res. 47, 259–264. https://doi.org/10.1093/nar/gky1022.

Nilsson, R.H., Andersson, A.F., Bissett, A., Finstad, A.G., Fossøy, F., Grosjean, M., Hope, M., Jeppesen, T.S., Kõljalg, U., Lundin, D., 2022. Introducing guidelines for publishing DNA-derived occurrence data through biodiversity data platforms. Metabarcoding and Metagenomics 6, 239–244. https://doi.org/10.3897/mbmg.6.84960.

Peay, K.G., Kennedy, P.G., Talbot, J.M., 2016. Dimensions of biodiversity in the Earth mycobiome. Nat. Rev. Microbiol. 14, 434–447. https://doi.org/10.1038/nrmicro.2016.59.

Penev, L., Mietchen, D., Chavan, V.S., Hagedorn, G., Smith, V.S., Shotton, D., Tuama, E.O., Senderov, V., Georgiev, T., Stoev, P., Groom, Q.J., Remsen, D., Edmunds, S.C., 2017. Strategies and guidelines for scholarly publishing of biodiversity data. Res. Ideas Outcomes 3e12431. https://doi.org/10.3897/rio.3.e12431.

Pocock, M.J.O., Chapman, D.S., Sheppard, L.J., Roy, H.E., 2014. Choosing and Using Citizen Science: a Guide to when and How to Use Sitizen Science to Monitor Biodiversity and Environment. Centre for Ecology & Hydrology, Wallingford, Oxfordshire.

Pyke, G.H., Ehrlich, P.R., 2010. Biological collections and ecological/environmental research: a review, some observations and

a look to the future. Biol. Rev. 85, 247–266. https://doi.org/10.1111/j.1469-185X.2009.00098.x.

R Core Team, 2019. R: A Language and Environment for Statistical Computing.

Ramirez, K.S., Knight, C.G., de Hollander, M., Brearley, F.Q., Constantinides, B., Cotton, A., Creer, S., Crowther, T.W., Davison, J., Delgado-Baquerizo, M., Dorrepaal, E., Elliott, D.R., Fox, G., Griffiths, R.I., Hale, C., Hartman, K., Houlden, A., Jones, D.L., Krab, E.J., Maestre, F.T., McGuire, K.L., Monteux, S., Orr, C.H., van der Putten, W.H., Roberts, I.S., Robinson, D.A., Rocca, J.D., Rowntree, J., Schlaeppi, K., Shepherd, M., Singh, B.K., Straathof, A.L., Bhatnagar, J.M., Thion, C., van der Heijden, M.G.A., de Vries, F.T., 2018. Detecting macroecological patterns in bacterial communities across independent studies of global soils. Nat. Microbiol. 3, 189–196. https://doi.org/10.1038/s41564-017-0062-x.

Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N.V., Finegold, C., Garrett, K.A., Gilligan, C.A., Jones, C.M., Martin, M.D., MacDonald, G.K., Neenan, P., Records, A., Schmale, D.G., Tateosian, L., Wei, Q., 2021. The persistent threat of emerging plant disease pandemics to global food security. Proc. Natl. Acad. Sci. U.S.A. 118, 1–9. https://doi.org/10.1073/pnas.2022239118.

Romeiras, M.M., Figueira, R., Duarte, M.C., Beja, P., Darbyshire, I., 2014. Documenting biogeographical patterns of African timber species using herbarium records: a conservation perspective based on native trees from Angola. PLoS One 9e103403. https://doi.org/10.1371/journal.pone.0103403.

Runnel, K., Abarenkov, K., Copoț, O., Mikryukov, V., Kõljalg, U., Saar, I., Tedersoo, L., 2022. DNA barcoding of fungal specimens using PacBio long-read high-throughput sequencing. Mol. Ecol. Resour. 22, 2871–2879. https://doi.org/10.1111/1755-0998.13663.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Consortium, F.B., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. USA 109, 6241–6246. https://doi.org/10.1073/pnas.1117018109.

Steidinger, B.S., Crowther, T.W., Liang, J., Van Nuland, M.E., Werner, G.D.A., Reich, P.B., Nabuurs, G.J., de-Miguel, S., Zhou, M., Picard, N., Herault, B., Zhao, X., Zhang, C., Routh, D., Peay, K.G., Abegg, M., Adou Yao, C.Y., Alberti, G., Almeyda Zambrano, A., Alvarez-Davila, E., Alvarez-Loayza, P., Alves, L.F., Ammer, C., Antón-Fernández, C., Araujo-Murakami, A., Arroyo, L., Avitabile, V., Aymard, G., Baker, T., Bałazy, R., Banki, O., Barroso, J., Bastian, M., Bastin, J.-F., Birigazzi, L., Birnbaum, P., Bitariho, R., Boeckx, P., Bongers, F., Bouriaud, O., Brancalion, P.H.S., Brandl, S., Brearley, F.Q., Brienen, R., Broadbent, E., Bruelheide, H., Bussotti, F., Cazzolla Gatti, R., Cesar, R., Cesljar, G., Chazdon, R., Chen, H.Y.H., Chisholm, C., Cienciala, E., Clark, C.J., Clark, D., Colletta, G., Condit, R., Coomes, D., Cornejo Valverde, F., Corral-Rivas, J.J., Crim, P., Cumming, J., Dayanandan, S., de Gasper, A.L., Decuyper, M., Derroire, G., DeVries, B., Djordjevic, I., Iêda, A., Dourdain, A., Obiang, N.L.E., Enquist, B., Eyre, T., Fandohan, A.B., Fayle, T.M., Feldpausch, T.R., Finér, L., Fischer, M., Fletcher, C., Fridman, J., Frizzera, L., Gamarra, J.G.P., Gianelle, D., Glick, H.B., Harris, D., Hector, A., Hemp, A., Hengeveld, G., Herbohn, J., Herold, M., Hillers, A., Honorio Coronado, E.N., Huber, M., Hui, C., Cho, H., Ibanez, T., Jung, I., Imai, N., Jagodzinski, A.M., Jaroszewicz, B., Johannsen, V., Joly, C.A., Jucker, T., Karminov, V., Kartawinata, K., Kearsley, E., Kenfack, D., Kennard, D., Kepfer-Rojas, S., Keppel, G., Khan, M.L., Killeen, T., Kim, H.S., Kitayama, K., Köhl, M., Korjus, H., Kraxner, F., Laarmann, D., Lang, M., Lewis, S., Lu, H., Lukina, N., Maitner, B., Malhi, Y., Marcon, E., Marimon, B.S., Marimon-Junior, B.H., Marshall, A.R., Martin, E., Martynenko, O., Meave, J.A., Melo-Cruz, O., Mendoza, C., Merow, C., Monteagudo Mendoza, A., Moreno, V., Mukul, S.A., Mundhenk, P., Nava-Miranda, M.G., Neill, D., Neldner, V., Nevenic, R., Ngugi, M., Niklaus, P., Oleksyn, J., Ontikov, P., Ortiz-Malavasi, E., Pan, Y., Paquette, A., Parada-Gutierrez, A., Parfenova, E., Park, M., Parren, M., Parthasarathy, N., Peri, P.L., Pfautsch, S., Phillips, O., Piedade, M.T., Piotto, D., Pitman, N.C.A., Polo, I., Poorter, L., Poulsen, A.D., Poulsen, J.R., Pretzsch, H., Ramirez Arevalo, F., Restrepo-Correa, Z., Rodeghiero, M., Rolim, S., Roopsind, A., Rovero, F., Rutishauser, E., Saikia, P., Saner, P., Schall, P., Schelhaas, M.-J., Schepaschenko, D., Scherer-Lorenzen, M., Schmid, B., Schöngart, J., Searle, E., Seben, V., Serra-Diaz, J.M., Salas-Eljatib, C., Sheil, D., Shvidenko, A., Silva-Espejo, J., Silveira, M., Singh, J., Sist, P., Slik, F., Sonké, B., Souza, A.F., Stereńczak, K., Svenning, J.-C., Svoboda, M., Targhetta, N., Tchebakova, N., Steege, H. ter, Thomas, R., Tikhonova, E., Umunay, P., Usoltsev, V., Valladares, F., van der Plas, F., Van Do, T., Vasquez Martinez, R., Verbeeck, H., Viana, H., Vieira, S., von Gadow, K., Wang, H.-F., Watson, J., Westerlund, B., Wiser, S., Wittmann, F., Wortel, V., Zagt, R., Zawila-Niedzwiecki, T., Zhu, Z.-X., Zo-Bi, I.C., consortium, G., 2019. Climatic controls of decomposition drive the global biogeography of forest-tree symbioses. Nature 569, 404–408. https://doi.org/10.1038/s41586-019-1128-0.

Sun, X., Hu, Y.-H., Wang, J., Fang, C., Li, J., Han, M., Wei, X., Zheng, H., Luo, X., Jia, Y., Gong, M., Xiao, L., Song, Z., 2021. Efficient and stable metabarcoding sequencing data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses. Gigabyte 1–15. https://doi.org/10.46471/gigabyte.16.

Talbot, J.M., Bruns, T.D., Taylor, J.W., Smith, D.P., Branco, S., Glassman, S.I., Erlandson, S., Vilgalys, R., Liao, H.-L., Smith, M.E., Peay, K.G., 2014. Endemism and functional convergence across the North American soil mycobiome. Proc. Natl. Acad. Sci. USA 111, 6341–6346. https://doi.org/10.1073/pnas.1402584111.

Taylor, J.W., Jacobson, D.J., Kroken, S., Kasuga, T., Geiser, D.M., Hibbett, D.S., Fisher, M.C., 2000. Phylogenetic species recognition and species concepts in fungi. Fungal Genet. Biol. 31, 21–32. https://doi.org/10.1006/fgbi.2000.1228.

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., Smith, M.E., Sharp, C., Saluveer, E., Saitta, A., Rosas, M., Riit, T., Ratkowsky, D., Pritsch, K., Põldmaa, K., Piepenbring, M., Phosri, C., Peterson, M., Parts, K., Pärtel, K., Otsing, E., Nouhra, E., Njouonkou, A.L., Nilsson, R.H., Morgado, L.N., Mayor, J., May, T.W., Majuakim, L., Lodge, D.J., Lee, S.S., Larsson, K.-H., Kohout, P., Hosaka, K., Hiiesalu, I., Henkel, T.W., Harend, H., Guo, L., Greslebin, A., Grelet, G., Geml, J., Gates, G., Dunstan, W., Dunk, C., Drenkhan, R., Dearnaley, J., Kesel, A. De, Dang, T., Chen, X., Buegger, F., Brearley, F.Q., Bonito, G., Anslan, S., Abell, S., Abarenkov, K., 2014. Global diversity and geography of soil fungi. Science 346, 1256688. https://doi.org/10.1126/SCIENCE.1256688, 80.

Tedersoo, L., Ramirez, K.S., Nilsson, R.H., Kaljuvee, A., Kõljalg, U., Abarenkov, K., 2015. Standardizing metadata and taxonomic identification in metabarcoding studies. GigaScience 4, 1–4. https://doi.org/10.1186/s13742-015-0074-5.

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R.H., Kennedy, P.G., Yang, T., Anslan, S., Mikryukov, V., 2022a. Best practices in metabarcoding of fungi: from experimental design to results. Mol. Ecol. 31, 2769–2795. https://doi.org/10.1111/mec.16460.

Tedersoo, L., Mikryukov, V., Zizka, A., Bahram, M., Doust, N.H.-, Anslan, S., Prylutskyi, O., Maestre, F.T., Pärn, J., Öpik, M., Moora, M., Zobel, M., Espenberg, M., Mander, Ü., Nasir, A., Adriana, K., Aída, A.A., Albornoz, F.E., Brearley, F.Q., Buegger, F., Zahn, G., Bonito, G., Hiiesalu, I., Barrio, I.C., 2022b. Global patterns in endemicity and vulnerability of soil fungi.

Global Change Biol. 28, 6696–6710. https://doi.org/10.1111/gcb.16398.

U'Ren, J.M., Lutzoni, F., Miadlikowska, J., Zimmerman, N.B., Carbone, I., May, G., Arnold, A.E., 2019. Host availability drives distributions of fungal endophytes in the imperilled boreal realm. Nat. Ecol. Evol. 3, 1430–1437. https://doi.org/10.1038/s41559-019-0975-2.

van Wilgen, B.W., Measey, J., Richardson, D.M., Wilson, J.R., Zengeya, T.A., 2021. Biological Invasions in South Africa, Invading Nature: Springer Series in Invasion Ecology. Springer Nature, Switzerland AG, Cham. https://doi.org/10.1080/0035919x.2021.1969797.

Větrovský, T., Kohout, P., Kopecký, M., Machac, A., Man, M., Bahnmann, B.D., Brabcová, V., Choi, J., Meszárošová, L., Human, Z.R., Lepinay, C., Lladó, S., López-Mondéjar, R., Martinović, T., Mašínová, T., Morais, D., Navrátilová, D., Odriozola, I., Štursová, M., Švec, K., Tláskal, V., Urbanová, M., Wan, J., Žifčáková, L., Howe, A., Ladau, J., Peay, K.G., Storch, D., Wild, J., Baldrian, P., 2019. A meta-analysis of global fungal distribution reveals climate-driven patterns. Nat. Commun. 10, 1–9. https://doi.org/10.1038/s41467-019-13164-8.

Větrovský, T., Morais, D., Kohout, P., Lepinay, C., Algora, C., Alò, F.D., Human, Z.R., Jomura, M., Kolařík, M., Mašínová, T., Meszárošová, L., Michalčíková, L., Mic, T., Mundra, S., Na, D., Odriozola, I., Piché-choquette, S., Štursová, M., Švec, K., Tláskal, V., Urbanová, M., Vlk, L., Voříšková, J., Žifčáková, L., Baldrian, P., 2020. GlobalFungi, a global database of fungal occurrences from metabarcoding studies. Sci. Data 7, 1–14. https://doi.org/10.1038/s41597-020-0567-7.

Wang, Z., Lachmann, A., Ma'ayan, A., 2019. Mining data and metadata from the gene expression omnibus. Biophys. Rev. 11, 103–110. https://doi.org/10.1007/s12551-018-0490-8.

White III, R.A., Callister, S.J., Moore, R.J., Baker, E.S., Jansson, J.K., 2016. The past, the present and the future of microbiome analyses. Nat. Protoc. 11, 2049–2053. https://doi.org/10.1038/nprot.2016.148.

Wilkinson, M.D., Dumontier, M., Aalbersberg, IjJ., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. Comment: the FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 1–9. https://doi.org/10.1038/sdata.2016.18.

Wüest, R.O., Zimmermann, N.E., Zurell, D., Alexander, J.M., Fritz, S.A., Hof, C., Kreft, H., Normand, S., Cabral, J.S., Szekely, E., Thuiller, W., Wikelski, M., Karger, D.N., 2020. Macroecology in the age of Big Data – where to go from here? J. Biogeogr. 47, 1–12. https://doi.org/10.1111/jbi.13633.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B.W., Blaser, M.J., Bonazzi, V., Booth, T., Bork, P., Bushman, F.D., Buttigieg, P.L., Chain, P.S.G., Charlson, E., Costello, E.K., Huot-Creasy, H., Dawyndt, P., Desantis, T., Fierer, N., Fuhrman, J.A., Gallery, R.E., Gevers, D., Gibbs, R.A., Gil, I.S., Gonzalez, A., Gordon, J.I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A.L., Kelley, S.T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C.L., Legg, T., Ley, R.E., Lozupone, C.A., Ludwig, W., Lyons, D., Maguire, E., Methé, B.A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K.E., Nemergut, D., Neufeld, J.D., Newbold, L.K., Oliver, A.E., Pace, N.R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D.A., Assunta-Sansone, S., Schloss, P.D., Schriml, L., Sinha, R., Smith, M.I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J.M., Ward, D.V., Weinstock, G.M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J.R., Yatsunenko, T., Glöckner, F.O., 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415–420. https://doi.org/10.1038/nbt.1823.

Zani, D., Crowther, T.W., Mo, L., Renner, S.S., Zohner, C.M., 2020. Increased growing-season productivity drives earlier autumn leaf senescence in temperate trees. Science 370, 1066–1071. https://doi.org/10.1126/science.abg1438, 80.