Fungal Diversity (2021) 111:573–588
https://doi.org/10.1007/s13225-021-00493-7

The Global Soil Mycobiome consortium dataset for boosting fungal diversity research

Leho Tedersoo[1] · Vladimir Mikryukov[1,2] · Sten Anslan[1,2] · Mohammad Bahram[3] · Abdul Nasir Khalid[4] · Adriana Corrales[5] · Ahto Agan[1] · Aída-M. Vasco-Palacios[6] · Alessandro Saitta[7] · Alexandre Antonelli[8] · Andrea C. Rinaldi[9] · Annemieke Verbeken[10] · Bobby P. Sulistyo[11] · Boris Tamgnoue[12] · Brendan Furneaux[13] · Camila Duarte Ritter[14] · Casper Nyamukondiwa[15] · Cathy Sharp[16] · César Marín[17] · D. Q. Dai[18] · Daniyal Gohar[1] · Dipon Sharmah[19] · Elisabeth Machteld Biersma[20,21] · Erin K. Cameron[22] · Eske De Crop[10] · Eveli Otsing[1] · Evgeny A. Davydov[23] · Felipe E. Albornoz[24] · Francis Q. Brearley[25] · Franz Buegger[26] · Genevieve Gates[27] · Geoffrey Zahn[28] · Gregory Bonito[29] · Indrek Hiiesalu[1,2] · Inga Hiiesalu[1,2] · Irma Zettur[1] · Isabel C. Barrio[30] · Jaan Pärn[2] · Jacob Heilmann-Clausen[31] · Jelena Ankuda[32] · John Y. Kupagme[1] · Joosep Sarapuu[2] · Jose G. Maciá-Vicente[33] · Joseph Djeugap Fovo[12] · József Geml[34] · Juha M. Alatalo[35] · Julieta Alvarez-Manjarrez[36] · Jutamart Monkai[37] · Kadri Põldmaa[1,2] · Kadri Runnel[1,2] · Kalev Adamson[38] · Kari A. Bråthen[39] · Karin Pritsch[26] · Kassim I. Tchan[40] · Kęstutis Armolaitis[32] · Kevin D. Hyde[37] · Kevin K. Newsham[20] · Kristel Panksep[41] · Lateef A. Adebola[42] · Louis J. Lamit[43,44] · Malka Saba[45] · Marcela E. da Silva Cáceres[46] · Maria Tuomi[39] · Marieka Gryzenhout[47] · Marijn Bauters[48] · Miklós Bálint[49] · Nalin Wijayawardene[50] · Niloufar Hagh-Doust[1,2] · Nourou S. Yorou[51] · Olavi Kurina[52] · Peter E. Mortimer[53] · Peter Meidl[13] · R. Henrik Nilsson[54] · Rasmus Puusepp[1] · Rebeca Casique-Valdés[55] · Rein Drenkhan[38] · Roberto Garibay-Orijel[56] · Roberto Godoy[57] · Saleh Alfarraj[58] · Saleh Rahimlou[1] · Sergei Põlme[1] · Sergey V. Dudov[59] · Sunil Mundra[60] · Talaat Ahmed[35] · Tarquin Netherway[3] · Terry W. Henkel[61] · Tomas Roslin[3] · Vladimir E. Fedosov[59,62] · Vladimir G. Onipchenko[59] · W. A. Erandi Yasanthika[37] · Young Woon Lim[63] · Meike Piepenbring[64] · Darta Klavina[65] · Urmas Kõljalg[1,66] · Kessy Abarenkov[1,66]

Received: 26 May 2021 / Accepted: 21 October 2021 / Published online: 30 November 2021

Abstract

Fungi are highly important biotic components of terrestrial ecosystems, but we still have a very limited understanding about their diversity and distribution. This data article releases a global soil fungal dataset of the Global Soil Mycobiome consortium (GSMc) to boost further research in fungal diversity, biogeography and macroecology. The dataset comprises 722,682 fungal operational taxonomic units (OTUs) derived from PacBio sequencing of full-length ITS and 18S-V9 variable regions from 3200 plots in 108 countries on all continents. The plots are supplied with geographical and edaphic metadata. The OTUs are taxonomically and functionally assigned to guilds and other functional groups. The entire dataset has been corrected by excluding chimeras, index-switch artefacts and potential contamination. The dataset is more inclusive in terms of geographical breadth and phylogenetic diversity of fungi than previously published data. The GSMc dataset is available over the PlutoF repository.

Keywords Soil fungi · Global dataset · PacBio sequencing · Fungal richness

Introduction

Soil microorganisms such as bacteria, archaea, fungi and protists play integral roles in terrestrial and aquatic ecosystem functioning. In particular, fungi act as the main decomposers of organic material and regulators of the abundance of other organisms as, e.g., mutualists, pathogens or producers of antibiotics (Bahram et al. 2018). In terrestrial

Handling Editor: Jian-Kui Liu.

✉ Leho Tedersoo
leho.tedersoo@ut.ee

Extended author information available on the last page of the article

ecosystems, mycorrhizal fungi colonize the roots of vascular plants and provide water and mineral nutrients to their hosts (Smith and Read 2008). Lichenized fungi associated with green algae and/or cyanobacteria form an important component of the soil biocrusts in drylands and cold deserts, developing a suitable habitat for other soil biota (Asplund and Wardle 2017). Fungal species and higher taxonomic groups differ greatly in their ecological functions (Põlme et al. 2020; Zanne et al. 2020); therefore, knowledge about fungal taxonomy and functional grouping is essential for understanding their potential activity.

Despite the great ecological importance of various fungal guilds, there are major knowledge gaps about the global distribution of fungal taxa. This is mostly due to the poor sample coverage of tropical countries and certain difficult-to-access regions of large countries such as the polar regions of Russian Federation, Canada and Antarctica. The advent of high-throughput sequencing (HTS) has enabled researchers to explore the global distribution of fungi or specific fungal phyla based on datasets of a few hundred unevenly distributed sampling sites (Tedersoo et al. 2014; Davison et al. 2015, 2021; Maestre et al. 2015; Egidi et al. 2019). Vetrovsky et al. (2020) compiled HTS data from > 100 individual HTS-based studies (including soil and other terrestrial substrates) in the GlobalFungi (GF) database. Based on their analyses, the authors proposed that temperature-related variables are the key drivers of soil fungal diversity (Vetrovsky et al. 2019) and estimated global fungal richness to be around 6 million species (Baldrian et al. 2021). However, GF has the following disadvantages: (1) the major geographical gaps remain; (2) it comprises short-read DNA sequence data from ITS1 and ITS2 subregions; (3) the included studies employ different sampling and analytical procedures, compromising their comparability; (4) some data stem from older studies using 454 sequencing (e.g., Tedersoo et al. 2014) and are of poor quality by modern standards; (5) although automatically checked for chimeras and arranged into sequence variants (SVs; Callahan et al. 2016), the data are principally raw and unchecked for other artefacts such as index switches and contamination, which may greatly affect diversity analyses.

Here, we publish and describe a dataset of the Global Soil Mycobiome consortium (GSMc), which offers the following benefits over previous public datasets and databases: (1) the geographical breadth and number of individual samples (> 125,000) are by far the greatest to date; (2) the molecular barcode covers the entire ITS region and the V9 variable region of the 18S rRNA gene, which taken together provide a much greater species-level taxonomic resolution as well as phylum-level and kingdom-level resolution compared to the ITS1 and ITS2 subregions alone (Tedersoo et al. 2021); (3) the samples have been collected and processed following the same protocol; (4) molecular analyses including PCR, library preparation and HTS have been performed in single central laboratories (Mycology and Microbiology Center, University of Tartu; and Norwegian Sequencing Centre, University of Oslo); (5) the data are carefully quality-filtered by checking for potential chimeras and contaminants manually for each of the 61 sequencing runs. Furthermore, the dataset is equipped with up-to-date taxonomic and functional annotations and metadata, allowing the possibility of re-interpretation by the users. The dataset is freely available to all researchers in a principally ready-to-use form (i.e., requires formatting of non-numerical values and missing data according to specific programs). These data are released to facilitate incorporation of soil fungi into macroecological analyses and boost understanding of the diversity and distribution of these microbes, topics that have lagged far behind similar studies in macroorganisms (Xu et al. 2020; Guerra et al. 2021).

Methods

Sampling and sample pre-processing

The GSMc was formed in 2015 to increase the geographical and habitat type coverage of a previous global soil fungal survey (Tedersoo et al. 2014). We invited the participants of the latter study and other colleagues to participate in GSMc by collecting and pre-processing fresh samples according to relevant national legislations. Around 10% of the initially contacted researchers provided material.

The GSMc participants were supplied with a detailed protocol for conducting sampling in various vegetation types (Tedersoo et al. 2014; Item S1) and securing necessary permissions. In brief, 40 soil cores (5 cm diam. to 5 cm depth) were collected from a 50 m × 50 m square plot or 56 m diam. circular plot (2500 m^2). The individual cores were collected in pairs (2–3 m apart) on the opposite sides of a randomly selected tree (in forests) or randomly selected locations (in non-forested ecosystems) at least 8 m away from other sampling locations and evenly covering the plot. Roughly an equal volume (approximately one quarter of the total volume) from each soil core were pooled, followed by mixing and air-drying within 24 h. The dried samples were placed into ZipLock plastic bags and homogenised by vigorous rubbing of the plastic bags, after which approximately 30–50 g of the finest material was transferred into a new bag (through a hole cut into the bottom of the ZipLock bag). This material was then either subjected to DNA extraction in the contributor's laboratory (using the methods described below) or was shipped to the University of Tartu along with silica gel. In total, sampling covered 3251 plots in 110 countries from all continents (Fig. 1; Table S1).

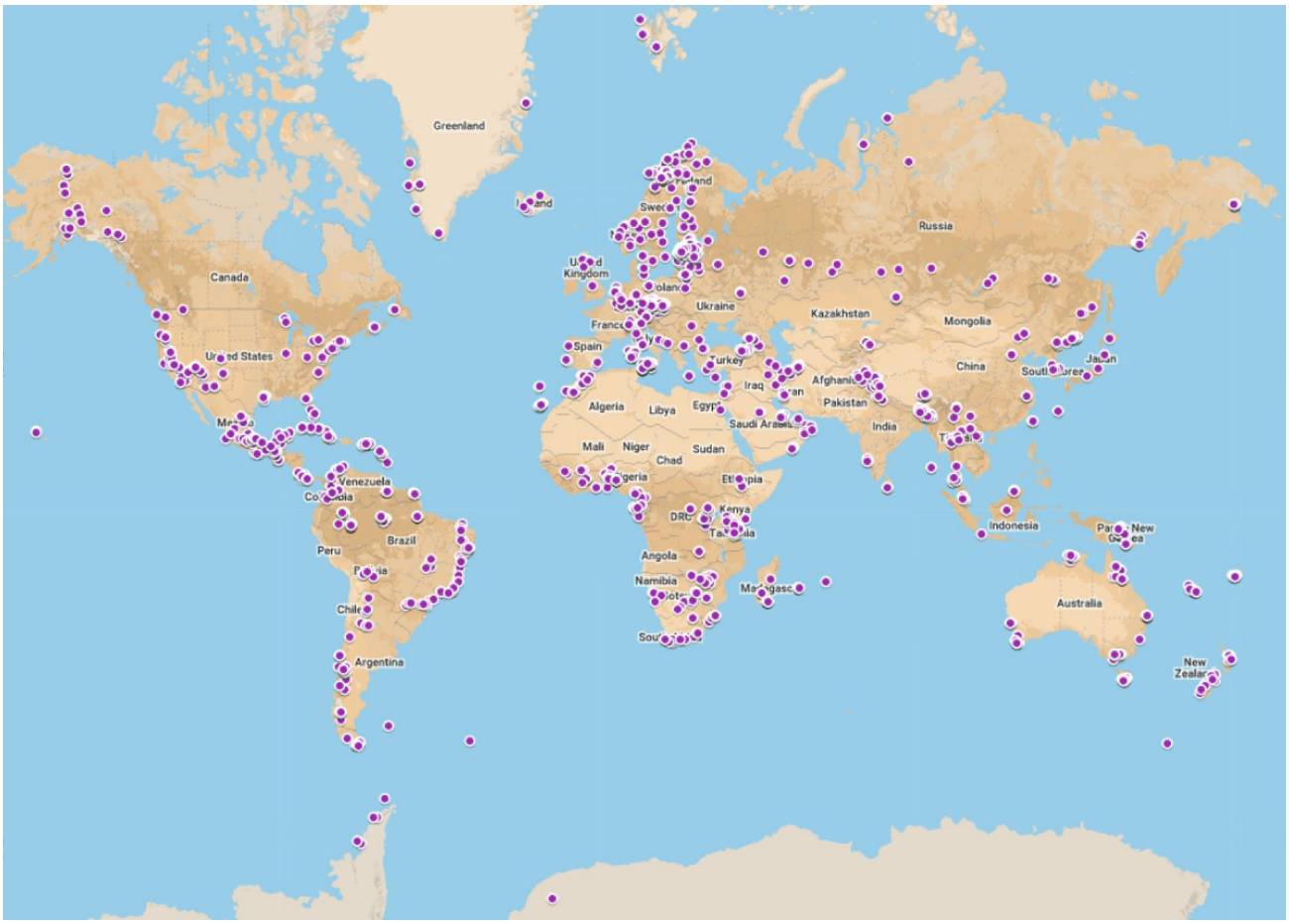


Fig. 1 Geographical distribution of 3200 sampling plots (purple dots)

Molecular analyses

DNA was extracted from 2.0 g of homogenised dry soil using the PowerMax Soil DNA Isolation kit (Qiagen, Carlsbad, CA, United States) following the manufacturer's instructions. The DNA extracts were further purified using the FavorPrepTM Genomic DNA Clean-Up kit (Favorgen, Vienna, Austria). PCR reactions were performed using the universal eukaryote primers ITS9mun and ITS4ngsuni (Tedersoo and Lindahl 2016; Tedersoo and Anslan 2019). These primers amplify nearly all known eukaryotes and all fungi excluding the Microsporidea (mismatches in ITS4ngsUni) and with potentially minor primer bias against Tulasnellaceae and Archaeorhizomycetes (one central mismatch; Tedersoo and Anslan 2019). The latter groups were, nonetheless, represented by hundreds of operational taxonomic units (OTUs) in our dataset. Both forward and reverse primers were equipped with the same 12-base index out of 115 combinations to minimise the risk of index switching (Table S2).

For amplification, the PCR mixture comprised 5 µl of 5 × HOT FIREPol Blend Master Mix (Solis Biodyne, Tartu, Estonia), 0.5 µl of each forward and reverse primer (20 mM), 1 µl of DNA extract and 18 µl ddH_2O. Thermal cycling included an initial denaturation at 95 °C for 15 min; 25–30 cycles of denaturation for 30 s at 95 °C, annealing for 30 s at 57 °C, elongation for 1 min at 72 °C; final elongation at 72 °C for 10 min; and storage at 4 °C. The duplicate PCR products were pooled and the presence of a 600–800 bp DNA band was checked on a 1% agarose gel. Samples yielding no visible PCR product were reamplified using 28 or 30 cycles (Tedersoo et al. 2020a). DNA concentrations were measured for a small subset of the amplicons using Qubit 3.0 (Thermo Fisher Scientific, Chicago, USA). Based on the correlations of Qubit measurements and amplicon band strength on a gel, we varied the quantity of amplicons (1–10 µl) for library preparation.

The pooled amplicons were shipped to the Norwegian Sequencing Centre at the University of Oslo for library preparation and sequencing. PacBio SMRTbell libraries were

prepared following the manufacturer's instructions (Pacific Biosciences, Palo Alto, USA) and sequenced on a Sequel II instrument using Sequel II Binding kit 2.1, sequencing chemistry 2.0, loading by diffusion, movie time of 15 h and pre-extension time of 20 min. The samples producing < 2000 reads were re-amplified and re-sequenced. In total, sequencing was performed on 61 SMRT cells. In some of these libraries and SMRT cells, other non-GSMc samples were included, which does not allow raw statistics for the GSMc samples to be reported.

Bioinformatics

Circular consensus sequences were generated using SMRT Tools v.9.0.0.92188 (PacBio) with default settings: minimum number of passes = 3 and minimum accuracy = 0.99. The FastQ-formatted output files were demultiplexed into samples based on the information of 12 bp primer index sequences using the software LIMA v.2.0.0 (PacBio) with the '–min-score 93' option to improve the precision of index identification. Sequence processing was performed using seqkit v.0.16.0 (Shen et al. 2016).

Across all samples, we compiled unique sequences and subjected these to ITS region extraction using the software ITSxpress v.1.8.0 (Rivers et al. 2018). All reads possessing > 1 ambiguous nucleotide or > 2 expected errors (Edgar and Flyvbjerg 2015) were removed prior to clustering.

Within each sample, chimeras were checked against the updated UNITE 9.1 beta dataset containing 920,399 reads (available at https://doi.org/10.15156/BIO/1444285). In this reference, taxonomy was re-checked by experts using ca. 80 person hours, taxonomically unidentified reads were removed, taxonomically identified reads from INSDc and UNITE (Nilsson et al. 2019) were added, taxonomically identified long reads (Tedersoo et al. 2020b; unpublished data) were added. Reference-based chimeras were detected using VSEARCH v.2.17.0 (Rognes et al. 2016), and removed when found. De novo chimeras were marked as such, but these were given the lowest priority for clustering. A manual examination indicated that ca. 80% of the global, putatively de novo chimeras were, in fact, false positives, and so they were retained.

For clustering, we used the open reference method by including all Sanger-derived ITS sequences in the UNITE database, including INSDc (update from 11.02.2021). These UNITE-INSDc sequences were subjected to similar quality filtering and ITS region extraction procedures as described above. Clustering was performed using a 98% sequence similarity threshold with VSEARCH. Our initial analysis suggested that sequence order was critical for formation of high-quality OTUs. To prevent formation of clusters with potentially low-quality reads acting as seeds, we used the VSEARCH options '–cluster_smallmem –usersort' and prioritized sequences as follows: (1) trimmed, high-quality UNITE-INSDc and GSMc sequences; (2) untrimmed, potentially partial sequences; and (3) GSMc sequences marked as putatively de novo chimeras. The initial analysis revealed that a careful selection of representative sequences is necessary to prevent low-quality sequences acting as references, which would compromise the OTUs thus derived. Therefore, we used the following order of preference for representative sequences: (1) highest similarity to centroid; (2) recognized as full-length ITS; and (3) recognized as non-chimeric. The alternative, widely used amplicon sequence variant (ASV) approach (Callahan et al. 2016) may not be suitable for full-length ITS sequences because of random PCR errors and the presence of multiple, usually highly similar, copies of the ITS region in eukaryote genomes (Lindner et al. 2013). Typically, the ASV approach tends to eliminate taxa that are both rare and phylogenetically unique (Joos et al. 2020). Furthermore, the ASV approach eliminated 21% more sequences compared to the OTU approach (V. Mikryukov, unpublished data). To produce a sample-by-OTU table, sequences from the dereplicated samples were mapped to OTUs using VSEARCH at 98% similarity with the options '–id 0.98 –iddef 2 –strand both'.

We removed all OTUs < 250 bases in length, which corresponds to the shortest full-length ITS sequences of Saccharomycetes and alveolates (Microsporidea may exhibit shorter ITS sequences, but these were either absent or had been removed during various filtering steps). Fungal OTUs displaying the ITS region < 350 bases were all screened for the presence of the highly conserved 5.8S rRNA motif "CGA TGAAG". OTUs without this motif were removed (except the phylum-level clade GS01 that has mutations in this motif) as partial reads (Tedersoo et al. 2017). A small proportion of OTUs was represented by partial ITS sequences (at least one of the ITS subregions and 5.8S rRNA genes present). Notably, some other OTUs were represented by ITS sequences that also contained a part of the flanking gene.

Taxonomic assignment of OTUs was performed using BLAST + 2.11.0 (Camacho et al. 2009) by running MegaBLAST queries of representative OTU sequences against the updated UNITE 9.1 beta reference dataset. These taxonomic assignments were checked against the 10 best MegaBLAST hits. Accordingly, we set the following taxon-specific thresholds: kingdom, $e_{max} = e^{-50}$; phylum, $e_{max} = e^{-55}$ to e^{-80}; class, $e_{max} = e^{-70}$ to e^{-100}; order, $e_{max} = e^{-80}$ to e^{-120}; genus, sequence similarity to the best match > 85–95% (Table S3). In general, lower thresholds were set for phyla comprising unicellular fungi and groups with divergent ITS sequences; higher thresholds were set for selected ascomycete groups displaying both low ITS sequence divergence and vigorous splitting of classes into orders and genera. At the kingdom level, OTUs best matching to Fungi or unspecified kingdoms at $e < e^{-50}$ were re-studied by custom BLAST + searches

(Tedersoo et al. 2020a) against a smaller reference dataset (https://doi.org/10.15156/BIO/1444347) that was populated with a single representative sequence per 1.5% SH and representative sequences of fungi and other eukaryotes (OTUs with e-values ranging from e^{-60} to e^{-100} to the best-hitting reference) from this dataset. If the five best matches were fungi, the corresponding OTUs were included in the fungal kingdom. For OTUs not classified to any kingdom or fungal phylum based on the above criteria, we performed an additional MegaBLAST analysis using the 18S V9 subregion as a target using both the UNITE 9.1 beta reference dataset and the SILVA 138.1 18S dataset (Quast et al. 2013). Based on the 10 best hits, we established similar e-value thresholds for the V9 matches (Table S3).

We used the taxonomic ranks genera, orders and phyla to present our findings, because these are typically the most well-defined and robust levels in fungal systematics. We acknowledge that the 98% sequence similarity criterion used for OTU delimitation represents an overall compromise across taxonomic groups, resulting in splitting species of non-Dikarya and lumping of certain species-rich groups of Ascomycota orders with slowly evolving ITS region (e.g., certain families in Helotiales, Hypocreales, Eurotiales and Sordariales of the Ascomycota, but also Hymenogastraceae and Cortinariaceae of the Basidiomycota; Visagie et al. 2014; Garnica et al. 2016). Using current bioinformatics developments, it is very laborious to apply different clustering thresholds to various fungal groups. Furthermore, in most cases, this information is poorly known. Taxonomy and classification of fungi follows Tedersoo et al. (2018a) and Wijayawardene et al. (2020). The taxonomical results were visualized as a Krona chart using KronaTools 2.8 (Ondov et al. 2011).

The OTUs marked as chimeric based on the de novo method and those represented by a single sequence (singletons) were studied in greater detail. Typically, singletons or OTUs represented by < 5 or < 10 reads are removed from metabarcoding studies, but this may result in the loss of up to two thirds of the biologically relevant information (Balint et al. 2016). A randomly selected set of 1000 sequences of putative chimeras was subjected to manual BLASTn queries against INSDc to search for patterns in the combination of best match coverage and sequence similarity indicative of chimeric origin (Nilsson et al. 2012). The analyses revealed that de novo chimeras were restricted to sequences with global abundance of up to 10 reads and frequency in up to three samples. Chimeras between the same dominant species may develop several times independently; if the chimera break point is located anywhere in the 5.8S region, the reads are similar enough to cluster into the same OTU. Commonly, putatively chimeric OTUs had unexpectedly long ITS sequences. We therefore also tested whether only the reference sequence was chimeric by selecting another reference sequence (the length of which was as close as possible to the median length of the sequences in that OTU and had the greatest abundance) for a new MegaBLAST search. In around half of the cases, this secondary reference was of regular length and displayed no evidence of a chimeric nature. This indicates that in the case of non-singleton OTUs, chimerism is commonly related only to the reference sequence. Hence, when selecting representative sequences, the longest reads should be avoided. We also observed that true chimeras typically have partial matches to the reference sequence, restricted to either side (but not the central part) of the read, with chimeric breakpoints commonly situated in the subregions. Such OTUs with non-central disruptions in alignments were removed when they exhibited a partial match at > 93% sequence similarity. We removed all singletons that were indicated as putatively chimeric or that had any ambiguous nucleotide.

Functional annotation of OTUs (trophic strategies and life forms) was performed at the level of genera for most fungal guilds using FungalTraits 1.3 (Põlme et al. 2020). We also used order-level annotation of certain life history traits (e.g., life form and arbuscular mycorrhizal fungi) when this was unequivocal for the entire order. Because many fungal genera comprise both ectomycorrhizal (EcM) and non-EcM taxa (e.g. *Hyaloscypha, Ramaria* and *Serendipita*), EcM fungi were additionally annotated at the level of sequence accessions based on information accumulated in UNITE. For EcM fungi, we generated taxon-specific e-value thresholds (Table S3) by utilizing information from 10 best blast hits, with additional guidance based on taxon occurrences in non-EcM habitats. Using functional annotations, we calculated the relative abundance of moulds (Umbelopsidales, Mortierellales, Mucorales, Aspergillaceae, Trichocomaceae and *Trichoderma*) to evaluate the relative quality of samples. High proportions of sequences from moulds are suggestive of sample degradation (Tedersoo et al. 2020a). We also calculated the relative abundance of all EcM fungi and of the /suillus-rhizopogon lineage and Mesophelliaceae (/hysterangium lineage) that are indicative of Pinaceae and Australian hosts (mainly Myrtaceae), respectively. The relative abundance of these groups and control samples were used as indicators of contamination, especially in plots where such EcM host plants are known to be absent. Samples with an estimated level of contamination of > 1% were removed from the dataset. Positive controls and single occurrences of contaminants were manually removed on a library by library case. To remove potential index switches, we removed all single occurrences and double occurrences in individual samples if the OTU total abundance exceeded 99 or 999, respectively, in the particular sequencing library (equivalent to 0.63% reads). Putatively uncontaminated replicates of the same samples in different libraries (4512 out of 4680) were pooled. However, samples with < 1000 reads (including

non-fungal eukaryotes) were excluded, because samples with low sequencing depth tended to accumulate a relatively higher proportion of artefacts. Altogether 51 (1.6%) of the initial 3,251 samples were removed. Subsequently, OTUs classified as non-fungi and artefactual were removed to limit the GSMc dataset to fungi.

Comparisons among databases

We compared the taxonomic richness and phylogenetic coverage of fungal phyla and orders of GSMc to the most recent versions of UNITE-INSDc (as of 11.02.2021) and GlobalFungi (GF; accessed 19.03.2021; soil data subset used in Baldrian et al. 2021). The UNITE-INSDc dataset was compared to GSMc for both full-length ITS and the ITS2 subregion. From GF, we requested only sequences corresponding to the ITS2 subset of OTUs assigned to the fungal kingdom (as published in Baldrian et al. 2021); therefore, we compared GSMc and GF only for the ITS2 subregion. To generate an ITS2-only version of the GSMc and UNITE + INSD datasets, we extracted the ITS2 subregion from all sequences that had passed the initial quality control for full-length ITS analyses. To provide a fair comparison to GF, we clustered the sequences from all datasets combined at 97% sequence similarity, ran BLAST + against our updated reference database, and considered the matches with e-values of $< e^{-50}$ as fungi (cf. Baldrian et al. 2021). We anticipate that slight differences between blast parameters and reference databases affect the e-value of the best match and identification decision.

Results

Taxonomic coverage

Demultiplexing of the 61 libraries yielded 30,043,967 sequences assigned to GSMc samples. Further quality filtering recovered 17,899,467 reads for clustering and 20,331,906 reads for mapping to OTUs. Clustering at the 98% sequence similarity threshold produced 1,251,637 OTUs including 709,791 singletons. Reference-based and de novo algorithms collectively suggested that 30,233 non-singleton (4.8%) and 59,855 singleton (10.5%) OTUs were putative chimeras. Based on additional checking and filtering (see Methods), 2.4% non-singleton and 10.2% singleton OTUs were regarded as chimeric, partial or of low quality, and were therefore removed. We acknowledge that multiple chimeric OTUs most likely remain in the dataset and many high-quality OTUs may be represented by sequences that are incomplete or contain flanking rRNA genes.

Of all 1,157,667 quality-filtered OTUs (18,782,650 reads), Fungi dominated (722,682 OTUs, 62.4%), followed by the kingdoms Alveolata (175,265, 15.1%), Metazoa (75,139, 6.5%), Rhizaria (34,376, 3.0%) and Viridiplantae (31,329, 2.7%) (Fig. 2a). Fungi also dominated in terms of sequence abundance (14,391,752 reads, 76.6%), followed by Alveolata (11.4%). The additional analysis of the V9 region enabled us to resolve roughly one half of the taxa that were not assigned to kingdoms and phyla based on the full-length ITS region alone. The 18S-V9 region was particularly useful for identification of certain taxonomic groups for which ITS reference data are scarce in public databases (e.g., Nucleariae, Amoebozoa and minor kingdoms of Excavata). Altogether 7.4% of the OTUs and 2.7% of the reads could not be assigned to any eukaryote kingdom. Based on the best matches of ITS and 18S-V9 regions, most of these unknowns probably represent Apicomplexa (Alveolata), invertebrates (Metazoa), Rozellomycota (Fungi) and Amoebozoa.

Of the GSMc fungal OTUs, 393,375 (54.4%) were singletons and 329,307 (45.6%) were represented by more than one sequence. Fungi were relatively more common among non-singletons (71.9%) than singletons (56.2%), suggesting that fungi are relatively better covered by our data compared to other eukaryote groups. Regarding the fungal singletons, 29,117 (7.4%) of the OTUs were also found in the other full-length ITS datasets analysed in parallel.

In the fungal kingdom, Ascomycota (330,054 OTUs, 45.7%) and Basidiomycota (204,667, 28.3%) dominated, followed by Rozellomycota (37,046, 5.1%), Glomeromycota (22,361, 3.1%) and Chytridiomycota (19,794, 2.7%) in terms of OTU richness (Fig. 2b). However, Basidiomycota (42.8%), Ascomycota (41.0%) and Mortierellomycota (6.2%) harboured the greatest number of reads. Overall, 7.1% of the fungal OTUs, representing 1.4% of the total number of reads, could not be assigned to any phylum. At the order level, Agaricales (59,385 OTUs, 8.2%), Helotiales (42,177, 5.8%) and Thelephorales (42,080, 5.8%) were the most OTU-rich groups (Figs. 3, 4). The genera *Tomentella* (34,202 OTUs, 4.7%), *Penicillium* (20,893, 2.9%) and *Russula* (15,784, 2.2%) comprised the greatest number of OTUs (Figs. 3, 4).

In terms of functional guilds, EcM fungi, soil saprotrophs, unspecified saprotrophs and unspecified pathotrophs were the most taxon-rich groups, represented by 124,616 (17.2%), 75,530 (10.5%), 49,521 (6.9%) and 43.758 (6.1%) OTUs, respectively. Notably, 38.6% of the OTUs could not be assigned to any functional guild. As mycorrhizal fungi are relatively well-known, these unassigned taxa are expected to mainly represent various saprotrophs and antagonists.

Plots

The GSMc plots are relatively evenly distributed globally, covering 108 countries (two countries had only failed samples) and all continents. There was a comparatively high

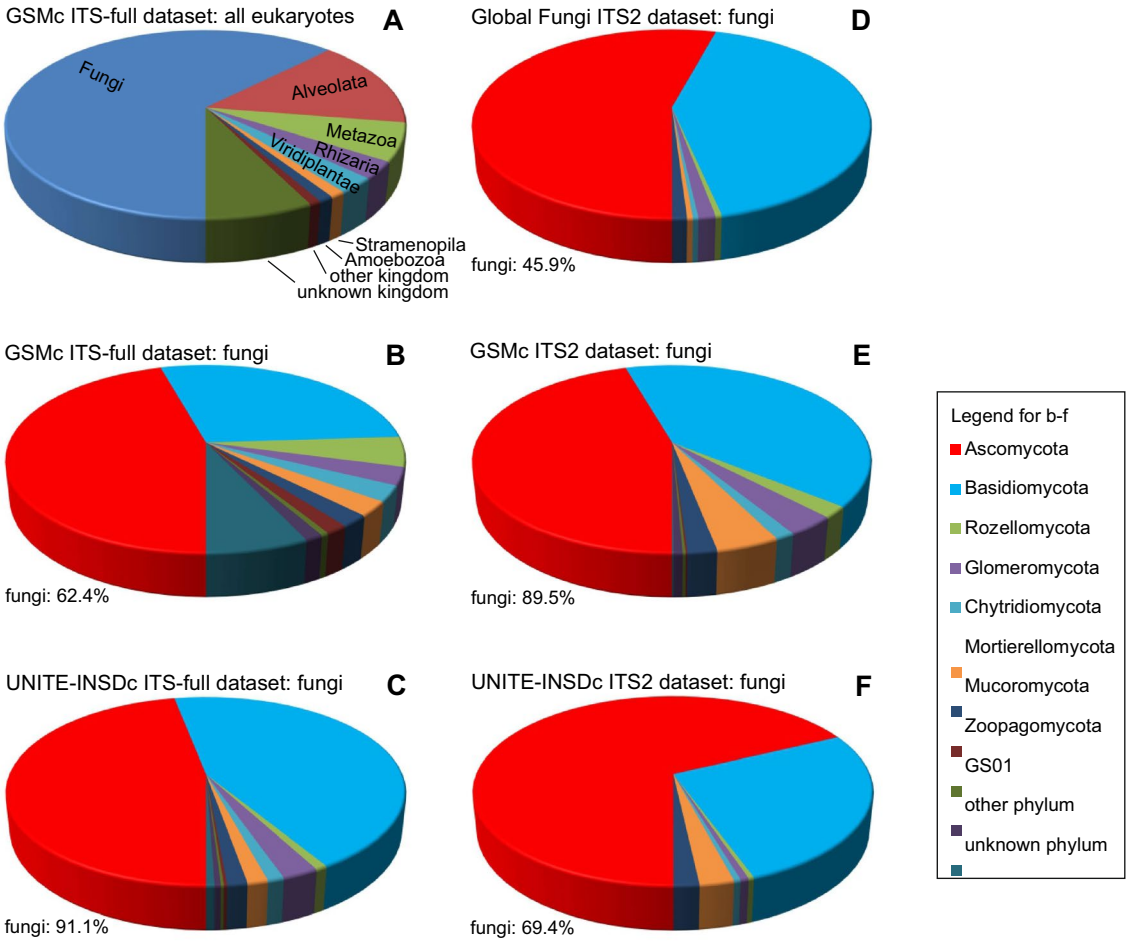


Fig. 2 Taxonomic distribution of fungi and fungal phyla in ITS-full (**a**–**c**) and ITS2 (**d**–**f**) datasets in the GSMc (**a**, **b**, **e**), UNITE-INSDc (**c**, **f**) and GlobalFungi (**d**) databases. The overall proportion (%) of fungal OTUs is indicated below each graph

coverage in Estonia and Latvia due to multiple case studies. We achieved the lowest sampling coverage in East Asia, Indo-Malaya, Pacific islands, Central Africa, the Canadian Arctic and Siberia.

Quality-screening revealed that 130 DNA sample and library combinations out of 4668 total combinations produced a low number of reads (< 1000 in total) and 38 were suspected of contamination from the positive control, other samples or an unknown source. DNA sequences from the uncontaminated plot and library combinations were pooled by plot, revealing information from 3200 plots. Combined data from several libraries comprised on average 3832 fungal reads (SD, 3075 reads) and 748 OTUs (SD, 401 OTUs) per plot. The most deeply sequenced samples harboured > 2000 fungal OTUs.

The Global Soil Mycobiome consortium dataset

The GSMc dataset of 3200 samples and 722,682 OTUs has a matrix fill (connectance, non-zero values) of 0.006%. Rarefaction analysis across all plots revealed a steady increase in OTU richness. Ugland's logarithmic extrapolator (Ugland et al. 2003) predicted the presence of 1.15 million and 1.53 million fungal OTUs at twofold greater sampling depth (6400 plots) and at the depth of 10,000 plots, respectively (Fig. 5). We refrain from extrapolating further and from estimating the global soil fungal richness because of steep OTU accumulation and uncertainties in the quality of singletons that play a key role in the accuracy of parametric and non-parametric estimators (Bunge et al. 2014; Balint et al. 2016).

The dataset is associated with abundant metadata. Around 99% of the plots are tagged with precise sampling dates and geographical coordinates; for most others, the approximate coordinates are deduced from maps if known at 0.1 degree precision or better. Based on the description of plots and remote sensing, we established the biome and type of land use. These custom biomes match the local vegetation cover rather than geographically delimited biomes (sensu Olson et al. 2001). In addition,

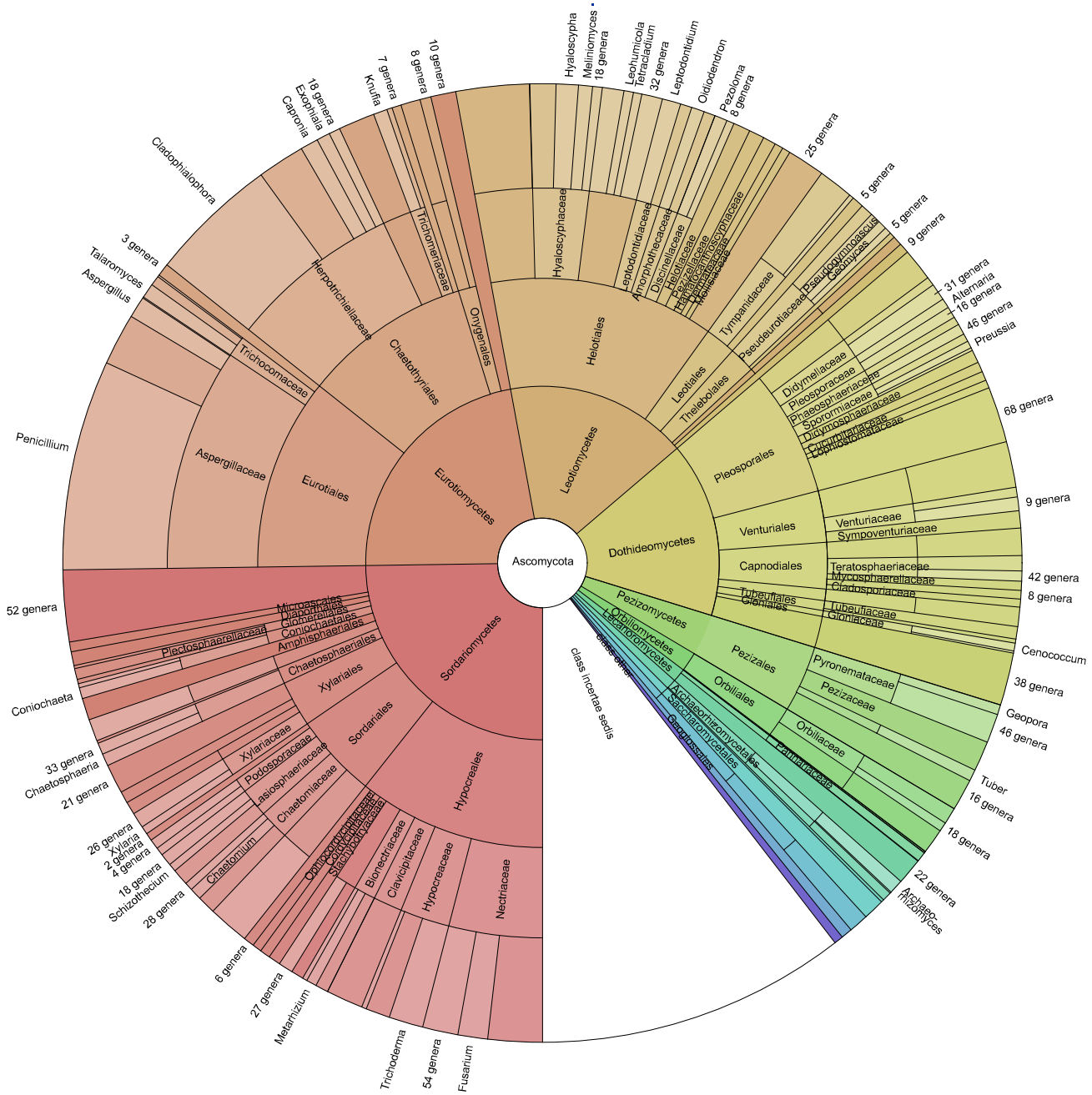


Fig. 3 Taxonomic profile of the Ascomycota component in the samples based on a Krona chart. The figure can be interactively expanded at https://doi.org/10.15156/BIO/1436941

we supplement data on soil total phosphorus and nitrogen content and $\delta^{15}N$ values for ^{15}N to ^{14}N isotopic ratio relative to a standard (Tedersoo et al. 2014). We provide information about whether the vegetation of the plot was considered native or non-native and the last reported burn. We also note differences to the original protocol (e.g., number of subsamples and plot area) and provide warning for samples that may have compromised quality (e.g., high mould abundance or potential contaminants).

Discussion

Taxonomy

The GSMc dataset comprises 722,682 fungal OTUs, exceeding the full-length ITS sequence data from the combined UNITE-INSDc dataset (820,138 fungal sequences classified into 125,363 OTUs) by six-fold. It also surpasses, by an order of magnitude, the Tedersoo et al. (2014) dataset that

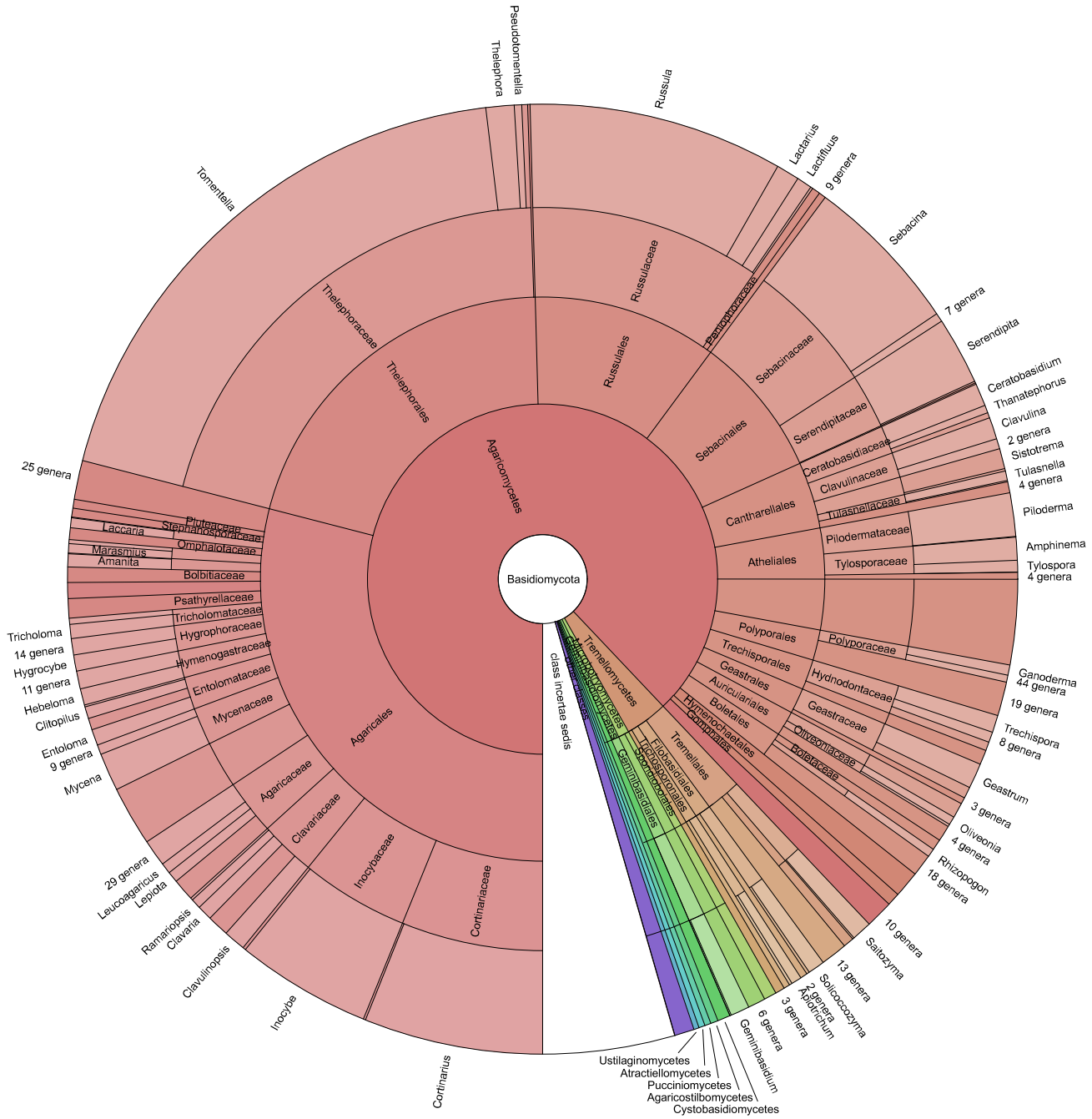


Fig. 4 Taxonomic profile of the Basidiomycota components in the samples based on a Krona chart. The figure can be interactively expanded at https://doi.org/10.15156/BIO/1436941

relied on the now-obsolete 454 sequencing technique of the ITS2 region (963,458 sequences classified into 80,486 OTUs including unfiltered singletons). The GSMc and UNITE-INSDc datasets share 47,310 fungal OTUs, corresponding to 6.5% of the GSMc OTUs and 37.7% of the UNITE-INSDc OTUs. The UNITE-INSDc dataset tends to lack a relatively higher proportion of non-Dikarya. Conversely, a majority of the OTUs missing from the GSMc dataset (but abundant in UNITE-INSDc) include various vertebrate pathogens and lichenized fungi, and also several plant pathogenic species of *Colletotrichum*, suggesting that soil sampling does not necessarily capture obligate biotrophs unrelated to the soil environment.

We compared the taxonomic profile of the GSMc full ITS data to the ITS2 data subset based on the clustering parameters and identification thresholds recommended for the

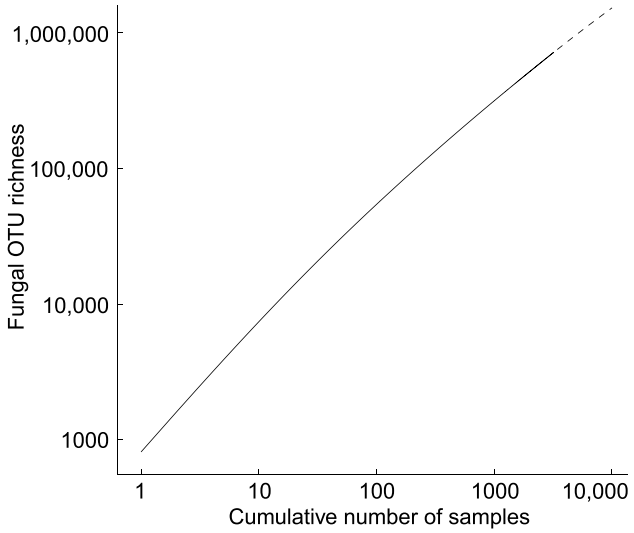


Fig. 5 Plot-based rarefaction and extrapolation (dashed line) curve of OTU accumulation with increasing spatiotemporal sampling depth

GlobalFungi (GF) dataset (Baldrian et al. 2021). The GSMc ITS2 subset comprised 435,192 fungal OTUs (12,909,562 reads), 39.8% less than the full ITS dataset. We attribute the lower fungal richness of the GSMc ITS2 subset to a much lower capacity of kingdom-level identification of distantly related OTUs, which is particularly evident in Rozellomycota and other non-Dikarya (Fig. 2b, e). For example, Zoopagomycota and the clade BCG2 (cf. Tedersoo et al. 2017) had > tenfold greater richness based on the full ITS region compared to the ITS2 subregion. Conversely, the greater OTU richness of UNITE-INSDc ITS2 sequences reflects a much higher proportion of reads passing the quality filtering and poor clustering of short reads (Tedersoo et al. 2018b).

Comparisons between the GSMc and GF datasets are hampered by the heterogeneous nature of the data in the latter (i.e., different sampling and molecular analysis protocols), as well as differences between the datasets in bioinformatics protocols. When reanalyzed following the options of Baldrian et al. (2021), the ITS2 subsets of GSMc, UNITE and GF (as in Baldrian et al. 2021) comprised 435,192 (12,909,562 reads), 102,563 (805,278) and 951,833 (193,411,059) fungal OTUs, respectively. The GSMc and GF dataset comprise 5537 (SD, 3686) and 20,832 (47,104) reads, and 1037 (556) and 716 (1274) OTUs, respectively.

Taxonomic comparison of fungal phyla among the GSMc, UNITE-INSDc and GF datasets revealed unexpected differences (Fig. 2d–f). The relative proportion and richness of many non-Dikarya lineages was much greater in the GSMc dataset compared to the other datasets. For example, the relative and absolute richness of the fungal phyla GS01, Zoopagomycota, Entomophthoromycota, Blastocladiomycota and Kickxellomycota were > tenfold greater in the GSMc dataset than in the GF dataset based on the ITS2 data. This can be explained by various sampling and analytical biases. First, species classified in these groups do not form fruiting bodies and mycorrhizas that are common isolation sources in UNITE-INSDc. Second, the classical primers used for fungal metabarcoding may have substantial biases against several groups of non-Dikarya. Third, and perhaps most importantly, ITS1 and ITS2 reads of many non-Dikarya are insufficient for reliable classification due to the lack of properly annotated reference sequences (Heeger et al. 2019; Tedersoo et al. 2020b).

As judged from the ITS2 subset, GSMc, UNITE and GF had a surprisingly low proportion of overlapping fungal OTUs. Interestingly, GSMc and UNITE tended to share a higher proportion of taxa compared with GF, despite the fact that both GSMc and GF data include samples from Tedersoo et al. (2014). At the level of genus, it was evident that many taxa such as *Laccaria* and *Ruhlandiella* display enormous OTU richness in GF (Fig. 6c–f), while other taxa, for example *Coemansia* and *Conidiobolus*, are absent (Fig. 6g–j). The ultra-high OTU richness of genera with relatively low known species richness may reflect uneven data quality in GF, as a large proportion of these records are derived from single-end Illumina reads from an individual study. Conversely, the absence of specific taxa can be attributable to PCR and primer biases as well as the length bias of Illumina library preparation (Sato et al. 2019). In the GSMc dataset, multiple genera were found to have full-length ITS regions of over 1000 bases (e.g., *Leccinum, Cantharellus, Balsamia, Piptocephalis, Spizellomyces p.parte* and *Entoloma p.parte*). These taxa may be very difficult to retrieve using short-read sequencing when considering the technical biases and their capacity to sequence up to at most some 590-base amplicons (including indexes and primers).

The Global Soil Mycobiome consortium dataset

The GSMc dataset includes 3200 composite samples (127,263 samples) from 3084 sites with unique geographical coordinates and 108 countries, surpassing our previous effort (Tedersoo et al. 2014) by an order of magnitude. The ITS2 data subset of GF (Baldrian et al. 2021) comprises 10,561 entries that represent nearly 95,000 samples and 3097 plots (63 countries) with a unique geocode, being roughly comparable in size. The GF dataset has relatively more dense sampling in Australia (data from Bissett et al. 2016) and China (multiple studies). Conversely, the GSMc sample coverage is relatively greater in Africa, South and Central America, North Europe, East Asia, Central Asia and the Pacific islands. Both datasets exhibit poor coverage of the Canadian Arctic and Indo-Malayan regions.

The GSMc dataset is ready-to-use for macroecological analyses after considering exclusion of samples of relatively

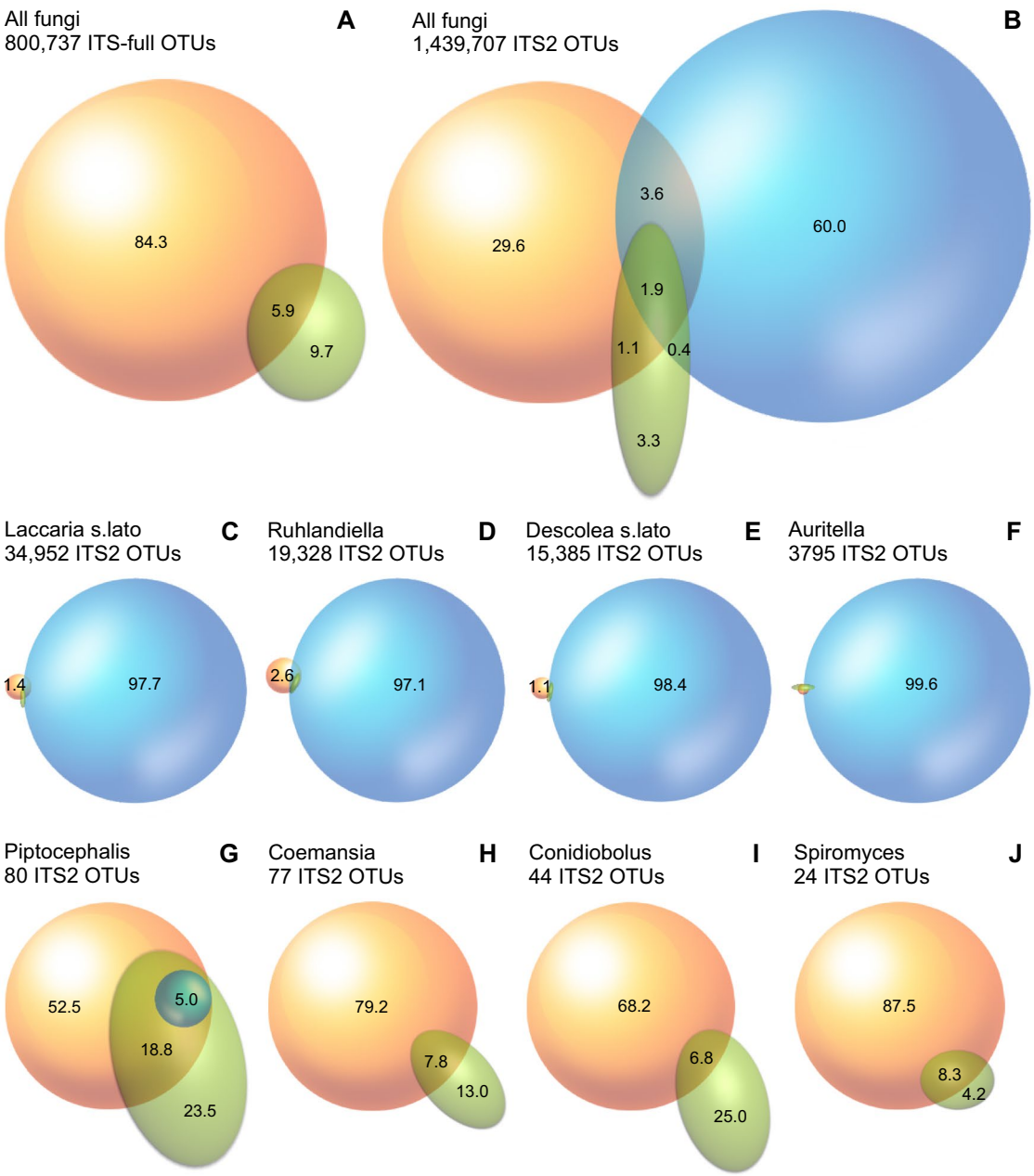


Fig. 6 Venn diagrams indicating unique and shared operational taxonomic units (OTUs) based on the ITS-full (**a**) and ITS2 (**b–j**) datasets of the GSMc (orange), UNITE-INSDc (green) and GlobalFungi (blue) databases. In **c–j**, the most strongly conflicting taxonomic groups are indicated. The relative abundance of unique and shared taxa is proportionally indicated in the area of Venn ellipses. Percent values < 0.2 are not indicated

low sequencing depth, high mould content and non-standard sample size (Table S1). Use of these data requires essentially no taxonomic or molecular ecological expertise. The GSMc data have been analysed with specifically optimised bioinformatics workflows accounting for the specific amplicon, indices and primers used. Taxonomic and functional annotations were provided by experts using specifically updated taxonomic and functional reference databases. The matrices have been manually curated to remove problematic samples and potentially contaminating OTUs. Conversely, the GF database is available upon reasonable request from its authors (although indicated as fully accessible over the web) and requires the user to sort specific bioinformatics data on a case-by-case basis. In addition, the taxonomic and functional

annotations require specific handling. In GF, problematic samples and index switch artefacts have not been removed, and these may impact downstream analyses. Furthermore, analyses using GF require specific accounting for multiple analytical variables such as volume and number of (sub)samples, DNA extraction methods, PCR primers and sequencing technology, many of which are poorly documented in the database.

Methodological considerations and limitations

In terms of methods, PacBio Sequel and Sequel II platforms provide sequence data of unprecedented quality to support full-length ITS or 18S-ITS marker gene analyses of fungi that exhibit much-improved taxonomic resolution (Tedersoo et al. 2021). Analysis of longer markers puts higher standards to both the initial DNA quality and bioinformatics quality-filtering due to higher rates of forming chimeras, incomplete reads and other artefacts. Our ongoing analyses with GSMc data indicate that a few samples with exceptionally high OTU richness (> 2000 OTUs) may harbour dominant species (e.g., *Tomentella* sp.) that encompass hundreds of OTUs, due potentially to the escape from concerted evolution. This phenomenon has been illustrated in analyses of fruiting bodies (Lindner et al. 2013) and mycorrhizal root tips (Tedersoo et al. 2010) and cannot be solved with current methods.

An anonymous referee pointed out that sampling top 5 cm of soil has limitations. We agree that the top 5 cm is not representative of the entire soil profile, because fungal diversity differs vertically (e.g. Lindahl et al. 2007). Nonetheless, our sampling captured both the organic soil horizon and top mineral soil (except bogs), where most of the microbial biomass and biodiversity is concentrated. Sampling to 10 or 20 cm depth was considered unfeasible on a global scale, because (1) deep sampling in rocky soils is virtually impossible; and (2) soil chemical properties of topsoil are better correlated with biodiversity, because the more dense deeper soil would contribute disproportionately more to soil pH and chemical composition compared with the less dense organic-rich horizons.

The main limitation of the GSMc dataset is its fixed nature, because it is difficult to integrate more data into it. However, further samples with full-length ITS sequences will be added in the course of ongoing research that uses the same sampling and analytical design or similar approaches. Since all taxonomic and functional annotation details are provided, users can easily check and improve the original annotations at any taxonomic and functional level or at the level of UNITE SHs. The numbers of fungal OTUs will slightly change with more efficient algorithms for clustering and quality filtering and improved reference databases.

Conclusions

Taken together, the GSMc dataset is the largest collection of soil fungal distribution data obtained following standardised procedures. This fully open dataset has been rigorously curated for taxonomy, functional traits and supplemented with original, plot-associated metadata. We hope that the GSMc dataset will boost our understanding of fungal biogeography and the role of fungi in macroecological processes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13225-021-00493-7.

Acknowledgements We Thank Liis Tiirmann for assistance during manuscript preparation and multiple students for assistance in sample and metadata collection. We also thank two anonymous referees for their constructive suggestions.

Funding The bulk of this work was supported by the Estonian Science Foundation (Grant Nos. PRG632, PSG136, MOBTP198, PUT1170), Norway-Baltic EEA financial mechanism (Grant No. EMP442), RSF19-14-00038, DSFP-2021 and Novo Nordisk Fonden (Silva Nova).

Data availability The GSMc OTU-by-sample dataset is available from the PlutoF data repository (https://doi.org/10.15156/BIO/2263453) in six files: information file, OTU matrix in spreadsheet format, sample metadata (equivalent to Table S1), taxonomic and functional description of OTUs, OTU sequences in FASTA format, as well as data in Biological Observation Matrix (BIOM) format. Representative sequences of identical sequences per sample will be available from the UNITE database.

Code availability The scripts used for the bioinformatic analysis are available at GitHub: https://github.com/Mycology-Microbiology-Center/GSMc

Declarations

Conflict of interest There are no conflicts of interest to declare related to this study.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors give their consent to publish this study in Fungal Diversity.

References

Asplund J, Wardle DA (2017) How lichens impact on terrestrial community and ecosystem properties. Biol Rev 92:1720–1738

Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Tedersoo L, Bork P (2018) Structure and function of the global topsoil microbiome. Nature 560:233–237

Baldrian P, Vetrovsky T, Lepinay C, Kohout P (2021) High-throughput sequencing view on the magnitude of global fungal diversity. Fung Divers

Balint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, O'Hara RB, Öpik M, Sogin ML, Unterseher M, Tedersoo L (2016) Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiol Rev 40:686–700

Bissett A, Fitzgerald A, Meintjes T, Mele PM, Reith F, Dennis PG, Breed MF, Brown B, Brown MV, Brugger J, Byrne M (2016) Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. GigaScience 18:21

Bunge J, Willis A, Walsh F (2014) Estimating the number of species in microbial diversity studies. Annu Rev Stat Appl 1:427–445

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP (2016) DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–584

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:421

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Ba A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T (2015) Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 349:970–973

Davison J, Moora M, Semchenko M, Adenan SB, Ahmed T, Akhmetzhanova AA, Alatalo JM, Al-Quraishy S, Andriyanova E, Anslan S, Bahram M (2021) Temperature and pH define the realised niche space of arbuscular mycorrhizal fungi. New Phytol 231:763–776

Edgar RC, Flyvbjerg H (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–3482

Egidi E, Delgado-Baquerizo M, Plett JM, Wang J, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK (2019) A few Ascomycota taxa dominate soil fungal communities worldwide. Nature Commun 10:2369

Garnica S, Schön ME, Abarenkov K, Riess K, Liimatainen K, Niskanen T, Dima B, Soop K, Frøslev TG, Jeppesen TS, Peintner U (2016) Determining threshold values for barcoding fungi: lessons from *Cortinarius* (Basidiomycota), a highly diverse and widespread ectomycorrhizal genus. FEMS Microbiol Ecol 92:fiw045

Guerra CA, Bardgett RD, Caon L, Crowther TW, Delgado-Baquerizo M, Montanarella L, Navarro LM, Orgiazzi A, Singh BK, Tedersoo L, Vargas-Rojas R (2021) Tracking, targeting and conserving soil biodiversity. Science 371:239–241

Heeger F, Wurzbacher C, Bourne EC, Mazzoni CJ, Monaghan MT (2019) Combining the 5.8S and ITS2 to improve classification of fungi. Methods Ecol Evol 10:1702–1711

Joos L, Beirinckx S, Haegeman A, Debode J, Vandecasteele B, Baeyen S, Goormachtig S, Clement L, De Tender C (2020) Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. BMC Genomics 21:733

Lindahl B, Ihrmark K, Boberg J, Trumbore SE, Högberg P, Stenlid J, Finlay RD (2007) Spatial separation of litter decomposition and mycorrhizal nutrient uptake in a boreal forest. New Phytol 173:611–620

Lindner DL, Carlsen T, Nilsson RH, Davey M, Schumacher T, Kauserud H (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. Ecol Evol 3:1751–1764

Maestre FT, Delgado-Baquerizo M, Jeffries TC, Eldridge DJ, Ochoa V, Gozalo B, Singh BK (2015) Increasing aridity reduces soil microbial diversity and abundance in global drylands. Proc Natl Acad Sci USA 112:15684–15689

Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JAA, Bergsten J, Porter TM, Jumpponen A, Vaishampayan P, Ovaskainen O, Hallenberg N, Bengtsson-Palme J, Eriksson KM, Larsson K-H, Larsson E, Kõljalg U (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4:37–63

Nilsson RH, Larsson KH, Taylor AF, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Kõljalg SI (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucl Acids Res 47:D259–D264

Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GWN, Underwood EC, Damico JA, Itoua I, Strand HE, Morrison JC, Loucks CJ, Allnutt TF, Ricketts TH, Kura Y, Lamoreux JF, Wettengel WW, Hedao P, Kassem KR (2001) Terrestrial ecoregions of the World: a new map of life on earth. Bioscience 51:933–938

Ondov BD, Bergman NH, Philippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinform 12:385

Põlme S, Abarenkov K, Nilsson RH, Lindahl BD, Clemmensen KE, Kauserud H, Nguyen N, Kjøller R, Bates ST, Baldrian P, Tedersoo L (2020) FungalTraits: a user-friendly traits database of fungi and fungus-like stramenopiles. Fung Divers 105:1–16

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl Acids Res 41:D590–D596

Rivers AR, Weber KC, Gardner TG, Liu S, Armstrong SD (2018) ITSxpress: software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis. F1000Research 7:1418

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584

Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, Hisatsune J, Sugai M, Takehiko I, Hayashi T (2019) Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res 26:391–398

Shen W, Le S, Li Y, Hu F (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE 11:e0163962

Smith SE, Read DJ (2008) Mycorrhizal symbiosis, 3rd edn. Academic Press, London

Tedersoo L, Anslan S (2019) Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. Environ Microbiol Rep 11:659–668

Tedersoo L, Lindahl B (2016) Fungal identification biases in microbiome projects. Environ Microbiol Rep 8:774–779

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytol 188:291–301

Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R, Villarreal-Ruiz L, Vasco-Palacios A, Quang Thu P, Suija A, Smith ME, Sharp C, Saluveer E, Saitta A, Ratkowsky D, Pritsch K, Riit T, Põldmaa K, Piepenbring M, Phosri C, Peterson M, Parts K, Pärtel K, Otsing E, Nouhra E, Njouonkou AL, Nilsson RH, Morgado LN, Mayor J, May TW, Kohout P, Hosaka K, Hiiesalu I, Henkel TW, Harend H, Guo L, Greslebin A, Grelet G, Geml J, Gates G, Dunstan W, Dunk C, Drenkhan R, Dearnaley J, De Kesel A, Dang T, Chen X, Buegger F, Brearley FQ, Bonito G, Anslan S, Abell S, Abarenkov K (2014) Global diversity and geography of soil fungi. Science 346:1078

Tedersoo L, Bahram M, Puusepp R, Nilsson RH, James TY (2017) Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome 5:42

Tedersoo L, Sánchez-Ramírez S, Kõljalg U, Bahram M, Döring M, Schigel D, May T, Ryberg M, Abarenkov K (2018a) High-level classification of the Fungi and a tool for evolutionary ecological analyses. Fung Divers 90:135–159

Tedersoo L, Tooming-Klunderud A, Anslan S (2018b) PacBio metabarcoding of fungi and other eukaryotes: biases and perspectives. New Phytol 217:1370–1385

Tedersoo L, Anslan S, Bahram M, Drenkhan R, Pritsch K, Buegger F, Padari A, Hagh-Doust N, Mikryukov V, Kõljalg U, Abarenkov K (2020a) Regional-scale in-depth analysis of soil fungal diversity reveals strong pH and plant species effects in Northern Europe. Front Microbiol 11:1953

Tedersoo L, Anslan S, Bahram M, Kõljalg U, Abarenkov K (2020b) Identifying the 'unidentified' fungi: a global-scale long-read third-generation sequencing approach. Fung Divers 103:273–293

Tedersoo L, Albertsen M, Anslan S, Callahan B (2021) Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. Appl Environ Microbiol 87:e00626-e721

Ugland KI, Gray JS, Ellingsen KE (2003) The species-accumulation curve and estimation of species richness. J Anim Ecol 72:888–897

Vetrovsky T, Kohout P, Kopecký M, Machac A, Man M, Bahnmann BD, Brabcová V, Choi J, Meszárošová L, Human ZR, Lepinay C, Baldrian P (2019) A meta-analysis of global fungal distribution reveals climate-driven patterns. Nat Commun 10:1–9

Vetrovsky T, Morais D, Kohout P, Lepinay C, Algora C, Awokunle Hollá S, Bahnmann BD, Bílohnědá K, Brabcová V, D'Alò F, Human ZR, Jomura M, Kolařík M, Kvasničková J, Lladó S, López-Mondéjar R, Martinović T, Mašínová T, Meszárošová L, Michalčíková L, Michalová T, Mundra S, Navrátilová 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 high-throughput-sequencing metabarcoding studies. Sci Data 7:228

Visagie CM, Houbraken J, Frisvad JC, Hong SB, Klaassen CH, Perrone G, Seifert KA, Varga J, Yaguchi T, Samson RA (2014) Identification and nomenclature of the genus *Penicillium*. Stud Mycol 78:343–371

Wijayawardene NN, Hyde KD, Al-Ani LK, Tedersoo L, Haelewaters D, Rajeshkumar KC, Zhao RL, Aptroot A, Leontyev D, Saxena RK, Tokarev YS (2020) Outline of Fungi and fungus-like taxa. Mycosphere 11:1060–1456

Xu X, Wang N, Lipson D, Sinsabaugh R, Schimel J, He L, Soudzilovskaia NA, Tedersoo L (2020) Microbial macroecology: In search of mechanisms governing microbial biogeographic patterns. Glob Ecol Biogeogr 29:1870–1886

Zanne AE, Abarenkov K, Afkhami ME, Aguilar-Trigueros CA, Bates S, Bhatnagar JM, Busby PE, Christian N, Cornwell W, Crowther TW, Moreno HF (2020) Fungal functional ecology: Bringing a trait-based approach to plant-associated fungi. Biol Rev 95:409–433

Authors and Affiliations

Leho Tedersoo[1] · Vladimir Mikryukov[1,2] · Sten Anslan[1,2] · Mohammad Bahram[3] · Abdul Nasir Khalid[4] · Adriana Corrales[5] · Ahto Agan[1] · Aída-M. Vasco-Palacios[6] · Alessandro Saitta[7] · Alexandre Antonelli[8] · Andrea C. Rinaldi[9] · Annemieke Verbeken[10] · Bobby P. Sulistyo[11] · Boris Tamgnoue[12] · Brendan Furneaux[13] · Camila Duarte Ritter[14] · Casper Nyamukondiwa[15] · Cathy Sharp[16] · César Marín[17] · D. Q. Dai[18] · Daniyal Gohar[1] · Dipon Sharmah[19] · Elisabeth Machteld Biersma[20,21] · Erin K. Cameron[22] · Eske De Crop[10] · Eveli Otsing[1] · Evgeny A. Davydov[23] · Felipe E. Albornoz[24] · Francis Q. Brearley[25] · Franz Buegger[26] · Genevieve Gates[27] · Geoffrey Zahn[28] · Gregory Bonito[29] · Indrek Hiiesalu[1,2] · Inga Hiiesalu[1,2] · Irma Zettur[1] · Isabel C. Barrio[30] · Jaan Pärn[2] · Jacob Heilmann-Clausen[31] · Jelena Ankuda[32] · John Y. Kupagme[1] · Joosep Sarapuu[2] · Jose G. Maciá-Vicente[33] · Joseph Djeugap Fovo[12] · József Geml[34] · Juha M. Alatalo[35] · Julieta Alvarez-Manjarrez[36] · Jutamart Monkai[37] · Kadri Põldmaa[1,2] · Kadri Runnel[1,2] · Kalev Adamson[38] · Kari A. Bråthen[39] · Karin Pritsch[26] · Kassim I. Tchan[40] · Kęstutis Armolaitis[32] · Kevin D. Hyde[37] · Kevin K. Newsham[20] · Kristel Panksep[41] · Lateef A. Adebola[42] · Louis J. Lamit[43,44] · Malka Saba[45] · Marcela E. da Silva Cáceres[46] · Maria Tuomi[39] · Marieka Gryzenhout[47] · Marijn Bauters[48] · Miklós Bálint[49] · Nalin Wijayawardene[50] · Niloufar Hagh-Doust[1,2] · Nourou S. Yorou[51] · Olavi Kurina[52] · Peter E. Mortimer[53] · Peter Meidl[13] · R. Henrik Nilsson[54] · Rasmus Puusepp[1] · Rebeca Casique-Valdés[55] · Rein Drenkhan[38] · Roberto Garibay-Orijel[56] · Roberto Godoy[57] · Saleh Alfarraj[58] · Saleh Rahimlou[1] · Sergei Põlme[1] · Sergey V. Dudov[59] · Sunil Mundra[60] · Talaat Ahmed[35] · Tarquin Netherway[3] · Terry W. Henkel[61] · Tomas Roslin[3] · Vladimir E. Fedosov[59,62] · Vladimir G. Onipchenko[59] · W. A. Erandi Yasanthika[37] · Young Woon Lim[63] · Meike Piepenbring[64] · Darta Klavina[65] · Urmas Kõljalg[1,66] · Kessy Abarenkov[1,66]

[1] Mycology and Microbiology Center, University of Tartu, Tartu, Estonia

[2] Institute of Ecology and Earth Sciences, Tartu, Estonia

[3] Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[4] Department of Botany, University of the Punjab, Quaid-e-Azam Campus, Lahore, Pakistan

[5] Centro de Investigaciones en Microbiología y Biotecnología-UR (CIMBIUR), Facultad de Ciencias Naturales, Universidad del Rosario, Bogotá, Colombia

[6] Grupo de Microbiología Ambiental y BioMicro, Escuela de Microbiología, Universidad de Antioquia UdeA, Medellín, Colombia

[7] Department of Agricultural, Food and Forest Sciences, University of Palermo, Palermo, Italy

[8] Royal Botanic Gardens, Surrey, UK

[9] Department of Biomedical Sciences, University of Cagliari, Cagliari, Italy

[10] Research Group Mycology, Department of Biology, Ghent University, Ghent, Belgium

[11] Department of Biomedicine, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

[12] Plant Pathology and Agricultural Zoology Research Unit, Department of Plant Protection, Faculty of Agronomy and Agricultural Sciences, University of Dschang, Dschang, Cameroon

[13] Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

[14] Eukaryotic Microbiology, University of Duisburg-Essen, Essen, Germany

[15] Department of Biological Sciences and Biotechnology, Botswana International University of Science and Technology, Palapye, Botswana

[16] Natural History Museum of Zimbabwe, Bulawayo, Zimbabwe

[17] Institute of Botany, Czech Academy of Sciences, Průhonice, Czech Republic

[18] Center for Yunnan Plateau Biological Resources Protection and Utilization, College of Biological Resource and Food Engineering, Qujing Normal University, Qujing, People's Republic of China

[19] Department of Botany, Jawaharlal Nehru Rajkeeya Mahavidyalaya, Pondicherry University, Port Blair, India

[20] British Antarctic Survey, NERC, Cambridge, UK

[21] Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

[22] Department of Environmental Science, Saint Mary's University, Halifax, NS, Canada

[23] Altai State University, Barnaul, Russia

[24] CSIRO Land and Water, Wembley, WA, Australia

[25] Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK

[26] Institute of Biochemical Plant Pathology, Helmholtz Zentrum München - Deutsches Forschungszentrum Für Gesundheit und Umwelt, Neuherberg, Germany

[27] Tasmanian Institute of Agriculture, Hobart, TAS, Australia

[28] Biology Department, Utah Valley University, Orem, UT, USA

[29] Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA

[30] Faculty of Environmental and Forest Sciences, Agricultural University of Iceland, Reykjavík, Iceland

[31] Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

[32] Department of Ecology, Institute of Forestry of Lithuanian Research Centre for Agriculture and Forestry (LAMMC), Kaunas distr., Lithuania

[33] Plant Ecology and Nature Conservation, Wageningen University & Research, Wageningen, The Netherlands

[34] Lendület Environmental Microbiome Research Group, Eszterházy Károly University, Eger, Hungary

[35] Environmental Science Center, Qatar University, Doha, Qatar

[36] Instituto de Geología, Universidad Nacional Autónoma de México, Coyoacán, Mexico City, Mexico

[37] Center of Excellence in Fungal Research, Mae Fah Luang University, Chiang Rai, Thailand

[38] Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Tartu, Estonia

[39] Department of Arctic and Marine Biology, UiT – Arctic University of Norway, Tromsø, Norway

[40] Research Unit Tropical Mycology and Plant-Soil Fungi Interactions, Faculty of Agronomy, University of Parakou, Parakou, Benin

[41] Chair of Hydrobiology and Fishery, Estonian University of Life Sciences, Tartu, Estonia

[42] Department of Plant Biology, University of Ilorin, Ilorin, Nigeria

[43] Department of Biology, Syracuse University, Syracuse, NY, USA

[44] Department of Environmental and Forest Biology, State University of New York, College of Environmental Science and Forestry, Syracuse, NY, USA

[45] Department of Plant Sciences, Quaid-I-Azam University, Islamabad, Pakistan

[46] Departamento de Biociências, Universidade Federal de Sergipe, Itabaiana, Sergipe, Brazil

[47] Department of Genetics, Faculty of Natural Sciences, University of the Free State, Bloemfontein, South Africa

[48] Department of Environment, Ghent University, Ghent, Belgium

[49] LOEWE Centre for Translational Biodiversity Genomics, Frankfurt am Main, Germany

[50] Center for Yunnan Plateau Biological Resources Protection and Utilization, College of Biological Resource and Food Engineering, Qujing Normal University, Qujing, Yunnan, People's Republic of China

[51] Faculty of Agronomy, University of Parakou, Parakou, Benin

[52] Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Tartu, Estonia

[53] CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

[54] Department of Biological and Environmental Sciences, Gothenburg Global Biodiversity Centre, University of Gothenburg, Göteborg, Sweden

[55] Instituto de Ciencias y Humanidades Lic. Salvador González Lobo, Universidad Autónoma de Coahuila, Saltillo, Mexico

[56] Instituto de Biología, Universidad Nacional Autónoma de México, Mexico, Mexico

[57] Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile, Valdivia, Chile

[58] Zoology Department, College of Science, King Saud University, Riyadh, Saudi Arabia

[59] Department of Ecology and Plant Geography, Lomonosov Moscow State University, Moscow, Russia

[60] Department of Biology, College of Science, United Arab Emirates University, Al Ain, Abu Dhabi, UAE

[61] Department of Biological Sciences, Humboldt State University, Arcata, CA, USA

[62] Botanical Garden-Institute FEB RAS, Vladivostok, Russia

[63] School of Biological Sciences and Institute of Microbiology, Seoul National University, Seoul, Korea

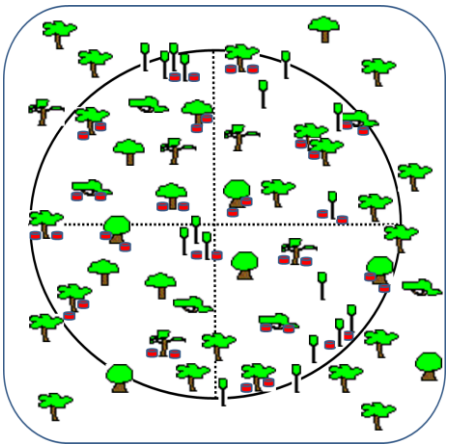
[64] Mycology Working Group, Goethe University Frankfurt am Main, Frankfurt am Main, Germany

[65] Latvian State Forest Research Institute Silava, Salaspils, Latvia

[66] Natural History Museum, University of Tartu, Tartu, Estonia

Soil collection protocol for the Global Soil Mycobiome consortium project

1. Select sampling sites of differing habitat, approximately 0.25-ha (50 x 50m or circular: 28m diam.) that is very little disturbed.
2. Record GPS coordinates & altitude for each site.
3. Record the relative basal area of all ECM host tree species and estimate their contribution to the basal area of all trees (in %). For arctic/grassland/desert sites, give a rough estimation about EcM plant cover (arctic EcM plants are Bistorta, Dryas, Salix, Betula, Kobresia – that's it). Typically, the type of root symbiosis explains much of the soil microbial community.
4. Within each site, select 20 random trees at least 8 m apart. For arctic/grassland/desert sites, select vegetated "spots" 8 m apart.
5. Below each tree, collect two soil samples, 5 cm diam. & 5 cm deep, from the opposite directions 1-1.5 m from the trunk, as follows (around "spots" in arctic/grassland/desert sites, but include rhizosphere!):
6. Remove loose litter (fallen leaves) from points to be sampled.
7. Hammer a hard 5-cm inner diameter PVC pipe 5-6 cm into the ground. In rocky sites, carve soil to a comparable depth and breadth with a knife
8. The sample should typically include both the organic and upper mineral layer.



9. From the core thus obtained, pick about ¼ of soil from all sides and place it in a clean plastic bag. The remaining part can be discarded.
10. Collect and pool all 40 soil cores into the same bag.
11. Remove coarse roots (>2 mm diam.) and stones.
12. Within 4 hours of collection, air dry the pooled samples in a dry place, (do not allow their temperature to rise above 40° C). If not possible, freeze-drying and deep-freezing are alternatives (i.e., dry later in more optimal conditions).
13. Make sure that samples are fully dried (no moisture is emitted when covered by plastic bag). After 24 hours, place dried samples into a strong large ZipLoc bag, seal, and rub vigorously between two hands for 3 min. to rupture soil into dust.
14. With the bag still closed, tip it so that one corner becomes empty.
15. Point that corner downwards and pinch off the bag to keep all soil above it.
16. Massage the bag gently, while opening your hands slightly, to allow about 30-50 of grams of the finest dust to fall into the low corner.
17. Pinch off the bag to prevent the coarse material from falling into the fine dust.
18. Place the corner with fine soil dust into another ZipLoc bag, cut off the corner, and allow the dust to empty into the new bag (the final sample).
19. Discard the corner, the first bag and its remaining soil (or keep it as a back-up).
20. Wrap ca 10-20 g Silica gel in a paper towel and add to the second bag with the collected fine dust.
21. Seal the bag and store gel-dried samples at room temperature for mailing.

Note: To avoid contaminating and cross-contaminating the samples, disposable gloves should be worn and changed each time when handling soil from a different site. If gloves are unavailable, hands should be well washed before beginning and each time soil from a different site is handled. Do secure all relevant permissions!

Please send all the samples as 'fungal samples', 'geological samples' or leave undeclared in the post office to:

Leho Tedersoo
University of Tartu
14A Ravila Street
50411 Tartu
Estonia

Standard protocol for DNA extraction

1. Distribute the finest soil material to one of the corners of a plastic bag. Cut off the corner and gently pour the soil dust on a clean piece of paper or foil. Weigh 2.0 +/- 0.05 grams of soil and deliver this to the DNA extraction tubes with grinding sand of MoBio PowerSoil kit (alternatively, put it to other tubes and later distribute it to the tubes provided by the kit).
2. Follow the DNA extraction protocol of the MoBio kit, except in step 4: vortex 5 min at max speed (instead of 10 min), followed by incubation at 60 °C for 10 min., and final vortex at max speed for 10 min. Continue following the original protocol.
3. Distribute the DNA extract into two tubes; keep one at -20 °C or -80 °C and a working solution at +4 °C.
4. To check the DNA amplificability, run PCR using any primer pair for a short product (400-800 bases) using 30 cycles; e.g. primers ITS1-ITS4. Primers yielding visible broad band at expected length is indicative of good quality. The lack of band and brownish colour indicates presence of inhibitors. Smear of short DNA fragments indicates degraded DNA.

If you chose to extract DNA, please send the DNA samples to the address above.