**Inclusion of database outgroups improves accuracy of fungal meta-amplicon taxonomic assignments**

*Clayton Rawson and Geoffrey Zahn\**

Biology Department, Utah Valley University, Orem, UT, USA

800 W University Parkway, SB243  
Orem, UT, 84058, USA

gzahn@uvu.edu

**Inclusion of database outgroups improves accuracy of meta-amplicon taxonomic assignments**

ABSTRACT

Keywords: meta-amplicon; taxonomic assignment; databases

INTRODUCTION:

Mycobiome studies commonly use high-throughput sequencing of either the ITS1 or ITS2 region amplified from environmental DNA. These 'barcode' regions are useful for identifying fungal species, but are hypervariable and thus unsuitable for alignment and phylogenetic placement. Thus, fungal meta-amplicon studies are limited by the databases available to match unknown amplicons against. Popular fungal sequence databases such as UNITE and Maarjam, have been developed and maintained by the mycological community to serve as improvements over uncurated databases such as the NCBI non-redundant nucleotide database, etc.

Common fungal primers for the ITS1 and ITS2 regions can co-amplify other taxa from the environment, such as metazoans. Despite what *in silico* tests might suggest, real-world PCR conditions can allow more permissive binding to non-target templates than might be expected.

There is a danger of using a curated database containing only fungi for assigning taxonomy to unknown amplicons: non-fungal reads can be assigned to "Unknown Fungus" if there are not sufficient outgroups present in the database. Taxonomic assignment methods such as the Ribosomal Database Project Classifier, Least Common Ancestor, or simply using the top BLAST hit will all attempt to find the closest match to a query sequence in a given database. If a reasonable match cannot be made at a given taxonomic level, the assignment algorithms find the most inclusive level at which taxonomy can be assigned. This leads to assignments such as k\_\_Fungi; p\_\_Ascomycota; c\_\_Sordariomycetes; o\_\_Xylariales; f\_\_Xylariaceae; g\_\_NA; s\_\_NA. In this example, taxonomy could only be assigned unambiguously up to the family level. Amplicon sequences might also be assigned as simply k\_\_Fungi. In some cases these amplicons might truly be from fungal taxa that are just not well-represented in a given curated fungal database, but metazoan amplicons, for example, can also be assigned to this "Unknown Fungus" taxonomy. If metazoan outgroup sequences are present in a database, these unknown metazoan amplicons should theoretically be more similar to those outgroups and be assigned to k\_\_Metazoa. Being able to detect and remove non-fungal amplicons is important for accurate measures of alpha- and beta-diversity

Here, we compare taxonomic assignments from 15 published fungal meta-amplicon studies (Table 1) using both the UNITE (only fungal sequences) and UNITE+Euk (fungal and eukaryotic outgroup sequences) databases to test the effects of including outgroups in taxonomic assignments. We examine agreement between assignments at each taxonomic level and discuss potential implications for ecological inferrences. Our results suggest that many fungal meta-amplicon studies may be unwittingly including non-fungi in their estimates of diversity, and that the inclusion of outgroups in taxonomic databases is essential for accurate ecological inferrences of fungal communities.

Alpha-diversity and taxonomic identity are influenced by database choice in ITS amplicon studies

**Introduction**

Despite how important the mycobiome is, it remains understudied. Meta-amplicon methods, which use the ITS (Internal Transcribed Spacer) barcode region of the eukaryotic ribosome, have allowed for the unlocking of large droves of data that have allowed for the characterization of fungal microbiota.1 Taxonomic annotations, specific taxonomic sequences used to identify an organism in a database, are a critical component for accuracy and reproducibility in mycobiome studies.1 ITS DNA barcode sequences are matched against a database of known fungi to identify the species present.1,2

Historically, the UNITE database has been the main database used for identification; however, the UNITE database is largely based on soil fungi and is lacking in outgroups (reference groups used to determine evolutionary relationships between monophyletic organisms).14 A newer database known as UNITE+EUK differs from UNITE in that it incorporates non-fungal eukaryotic sequences.14,15 These serve as outgroups, which may help boost the accuracy of the database by allowing for non-fungal sequences from samples to be assigned to their accurate groups, rather than being erroneously assigned to the nearest fungus in the UNITE database. The goal of this research is to compare published mycobiome studies that use meta-amplicon methods to both databases to see if there is a difference in taxonomic identification. We hypothesize that UNITE+EUK will be better at identifying fungi at the family and genus level than UNITE, while both will be more similar at the phylum level.

**MATERIALS AND METHODS:**

**Overview**

**We selected 15 fungal meta-amplicon studies with publicly available raw ITS amplicons. Raw data for each study was downloaded and processed into ASV tables. All ASV tables were combined into a single table and taxonomy was assigned to ASVs using both a fungus-only database (UNITE Fungi 2022-10-16; https://doi.org/10.15156/BIO/2483911) and a fungus+eukaryotes database (UNITE All 2022-10-16; https://doi.org/10.15156/BIO/2483913). Taxonomic assignments were made using both the RDP Classifier within DADA2, and by selecting the top BLAST hit from a standalone blastn call against both databases. All analyses were performed in R (version 4.2.0).**

**Data aquisition**

**We searched the literature for studies that reported fungal mycobiome data from either ITS1 or ITS2 amplicons. Studies were selected for inclusion based on availability of raw data in the Sequence Read Archive (SRA) and came from varied habitats (Table 1). Sequences were downloaded directly from the SRA using the fasterq-dump module from the sra-toolkit (version 2.10.9; SRA Toolkit development team, https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Metadata from each study was downloaded directly from SRA. Scripts for downloading data are included in Supporting Info.**

**Data processing and taxonomic assignment**

**Raw data from each study was processed as follows: First, the ITS regions were extracted using itsxpress (Bengtsson-Palme et al., 2013)**. Next, extracted forward reads were subjected to quality filtration using the *dada2* R package (Callahan et al., 2016) where sequences with uncalled bases were removed and sequences were truncated when quality scores dropped below a phred score of 25. Quality filtered reads were denoised with the DADA2 algorithm and processed into ASV tables for each study. Each study was combined with metadata from the SRA and placed into a phyloseq object using the *phyloseq* R package (McMurdie & Holmes, 2013). All phyloseq objects were combined into a single object and taxonomy was assigned using the assign\_taxonomy() function from *dada2* using both the UNITE and UNITE\_All databases. ASV sequencess were also exported for assignment against those two databases using the blastn script from BLAST (version 2,13,0) (Altschul et al., 1990). The top BLAST hit, based on e-value and coverage, was selected as the "BLAST taxonomy." These various taxonomic assignments were added to phyloseq objects for all downstream analyses.

**Analyses**

**Taxonomic assignments were compared at each taxonomic level to examine agreement between the UNITE\_Fungi and UNITE\_All databases. Agreement was indicated as exact taxonomic match at a given level for each ASV. After agreement values were determined, we removed any taxa not assigned to Kingdom Fungi for estimating fungal diversity.**

**All ASVs identified as non-fungal by both the BLAST and RDP Classifier methods against the UNITE\_All database were investigated under the corresponding ASV assignments by the UNITE\_Fungi database. All analysis code can be found in the GitHub repository associated with this manuscript (https://github.com/gzahn/Fungal\_Database\_Comparison).**

**RESULTS**

**Out of 40,232 unique ASVs recovered from all the studies, 16,106 were determined to be of non-fungal origin based on RDP Classifier and BLAST assignments against the UNITE\_All database. Of these, taxonomic assignment against the UNITE\_Fungi database falsely identified 16,075 of them as Kingdom Fungi (See Supporting Info). The majority of non-fungal ASVs that were falsely identified as "Unknown Fungus" belonged to either unknown or metazoan kingdoms (Figure 3). Since none of the studies in question described any methods for removing non-fungi, these ASVs presumably remained in the studies in question, contributing to "fungal" diversity measures. This effect can be seen in principle in Fig 5, where fungal alpha diversity measures were often artificially inflated by the accidental inclusion of these non-fungal ASVs.**

**Taxonomic agreement between the two databases varied between studies (Figure 1), and between major habitats (Figure 2). ASVs from aerial and soil habitats showed fewer falsely assigned fungi while ASVs from plant and animal hosts had propotionally more falsely assigned fungi. Limited replication reduced our statistical power so that these patterns are not statistically significant, however.**

**Alpha diversity (richness) was affected by these falsely assigned "fungal" ASVs. Comparing the fungal richness between database methods showed artificial inflation of "fungal" diversity when ASVs were assigned taxonomy using the UNITE\_Fungi database. The vast majority of these assignments were simply "*Fungi sp.*" so it is unlikely to have affected any lineage-specific conclusions reached in the associated studies. However, it seems plausible that these and other studies that use the UNITE\_Fungi database without outgroups will show inflated alpha diversity measures. Some ecosystems may be more affected by this issue than others. We found some indication that studies focused on soil are less affected by false positive fungi than studies focused on plant or animal mycobiota, but these were not statistaically significant differences.**

**DISCUSSION**

**Alpha diversity measures are a common tool for comparing samples within and between studies. If a goal of a study is to examine fungal diversity, care must be taken to remove non-fungal sequences. This depends on reliable taxonomic assignments of those sequences. By assigning taxonomy with a database that *only* includes fungi, the chances of non-fungal sequences being assigned as some unspecified "*Fungus sp*" are dramatically increased. In this analysis of 15 published studies, we found that 99.8% of all non-fungal sequences were assigned thus when using the UNITE\_Fungi database.**

**The ability of common fungal primers to co-amplify non-target templates has been discussed before** (Bellemain et al., 2010; Ihrmark et al., 2012; Tedersoo & Anslan, 2019). While *in silico* tests may help to quantify the danger of this occurring for specific primer sets and specific conditions, even the most careful PCR protocol is not a homogenous reaction. Temperature variations and other conditions can lead to non-specific binding. For studies of the fungal mycobiome where host or environmental DNA is present, there will always be a risk that other eukaryotic taxa are amplified.

Our results highlight that this has occurred and gone undetected in the published literature, but that it is not an unsurmountable problem. It can be mitigated by simply including outgroups in the databse when taxonomy is assigned and then by removing taxa not unambiguously assigned to the fungal kingdom. We strongly suggest that future studies of this nature use the UNITE\_All database for assigning taxonomy.

The raw ITS sequences from fifteen published studies involving plant, soil, animal, aquatic, and air mycobiomes were compiled. Studies were targeted that had open-source sequence read data available at the National Center for Biotechnology Information (NCBI). Data from each study, along with accession numbers and primers used in each study, was collected. Each study was processed using custom R scripts that facilitated data processing, which resulted in data files containing only the ITSI region for each sample and study. The extracted ITS1 reads were then used to create amplicon sequence variant (ASV) using the DADA2 algorithm.7 ASVs make it possible to distinguish sequence variation by as little as a single nucleotide change. The RDP Classifier was employed in the pipeline with DADA2, exact sequence variants, and taxonomic assignments to process the data.6 The R package phyloseq was used to analyze the data and to create graphs.6 Analysis was done as a simple true or false agreement. If the same answer came out for both UNITE + UNITE+EUK, then R would give a true. If the answers were not the same, then a false would be given. This was done for all the studies. Bar plots were outputted from R, and the agreement between the databases was put as a proportional agreement based on the number of samples, agreement, and taxonomic level. Complete methods and R scripts can be found at https://github.com/gzahn/Clayton\_SRA.

Comparisons were done by looking at each study and how the taxonomic identification varied between UNITE and UNITE+EUK. The taxonomic ranks analyzed included phylum, class, order, family, genus, and species. Alpha diversity, the species diversity or species richness on a local scale, was also analyzed for each study. Measuring and testing of alpha diversity was done by vegan.8 Results from all the studies were tabulated to show the overall agreement between UNITE and UNITE+EUK.

**RESULTS:**

The fifteen articles were tabulated and sorted by host (Table 1). Results from the studies were varied depending on the study. Nine of the studies had high agreements across the board with over 50% agreement (Figure 1). Three studies had agreements between 30-50%. The three remaining studies have little to no agreement across the board. Among the 15 studies, kingdom had the highest agreement after which depending on the study agreement either stayed the same or went down from kingdom to phylum, phylum to class, etc.

Of the seven environments, coastal had the highest agreement, with temperate, and tropical following close behind though the agreement fell as the taxa got closer to genus and species (Figure 2). Freshwater, desert, and laboratory had fairly constant agreement, but on the lower end near or below 50%. Marine had by far the lowest with less 25% agreement.

Among the five hosts, aerobiota had the highest agreement among the group. Kingdom, phylum, class, and order being near 100% with family and genus not far behind (Figure 3). Species was the lowest in aerobiota with around 50%. The animal group was the most consistent with all taxonomic groups being near 50%. The aquatic group was similar to animal with the exception of a little more agreement at the phylum level and less than half the agreement found in animal on the species level. Soil was similar to aerobiota just with 10-15% agreement less in all taxonomic levels. Plant had the worst agreement of all hosts. Kingdom and phylum had around 25% agreement with all other taxonomic groups being less than 25%.

Disagreements among the 11 true kingdoms/clades were measured. The kingdom metazoa had the highest level of disagreements coming close to 50%. An unidentified group also had a high level of disagreement near 40%. All others of the 10 major groups had well less than 10% disagreement. The species rich infrakingdom of rhizaria was the highest next to metazoa and the unidentified groups with around 5% disagreement. All other groups had little to no disagreement.



Table 1: Table shows the 15 studies used. The table includes the organism the sample was taken from, the habitat, number of taxa, Shannon diversity, and SRA Accession number for each sample.

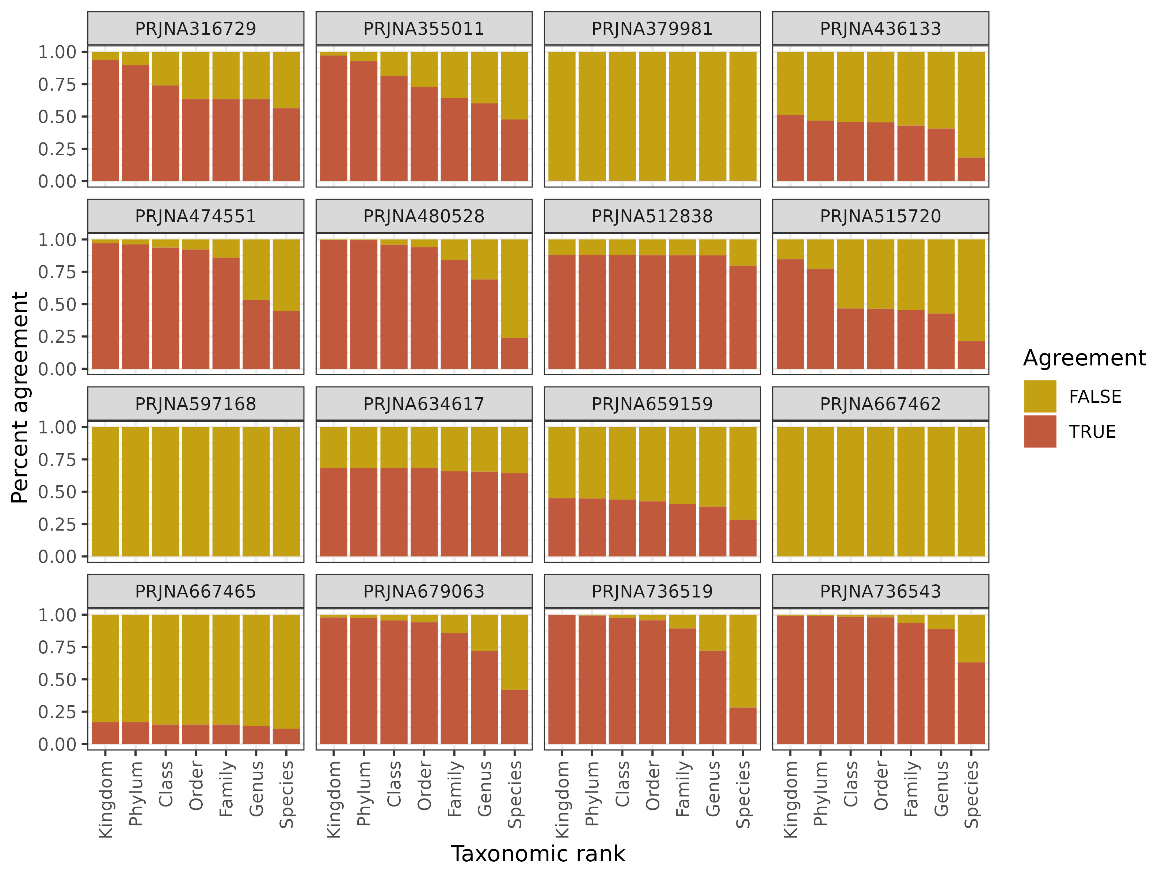


Figure 1: Figure shows the proportional agreement between databases UNITE and UNITE+EUK for each of the 15 studies. True or false agreement is shown for each of the studies on the taxonomic levels of kingdom, phylum, class, order, family, genus, and species.

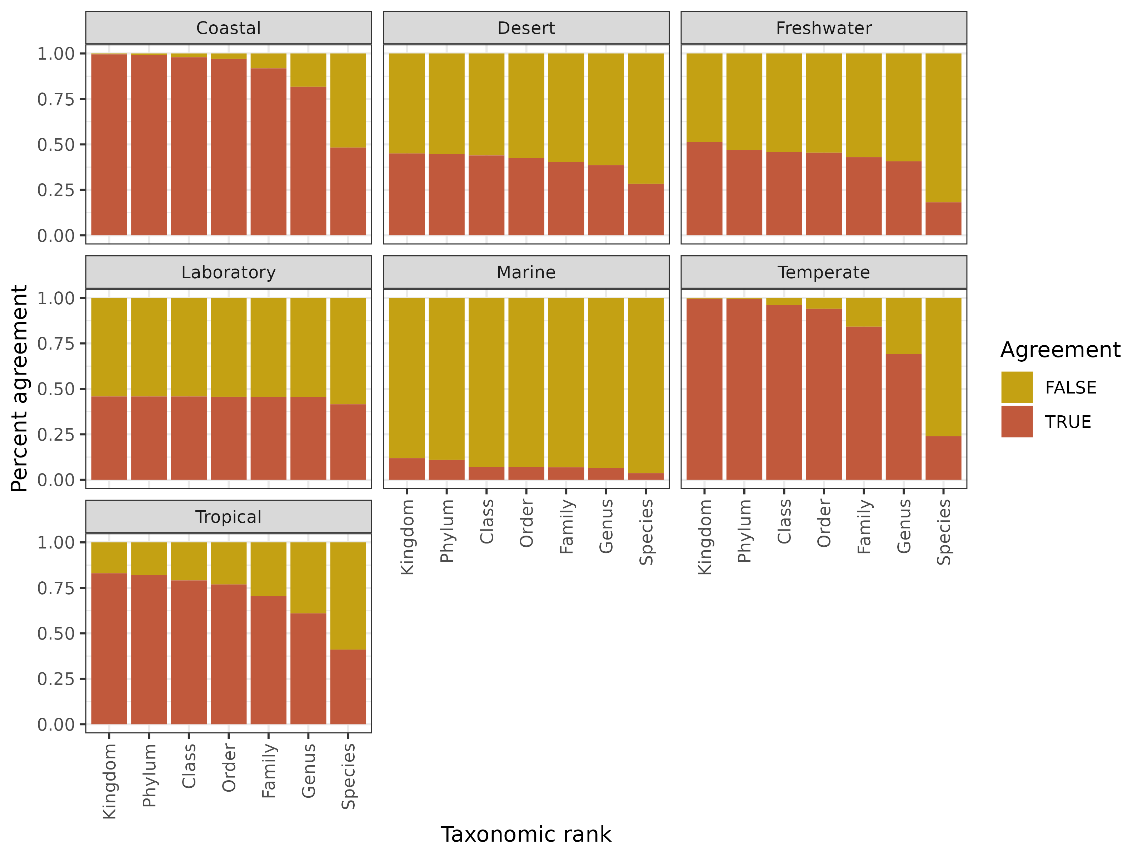
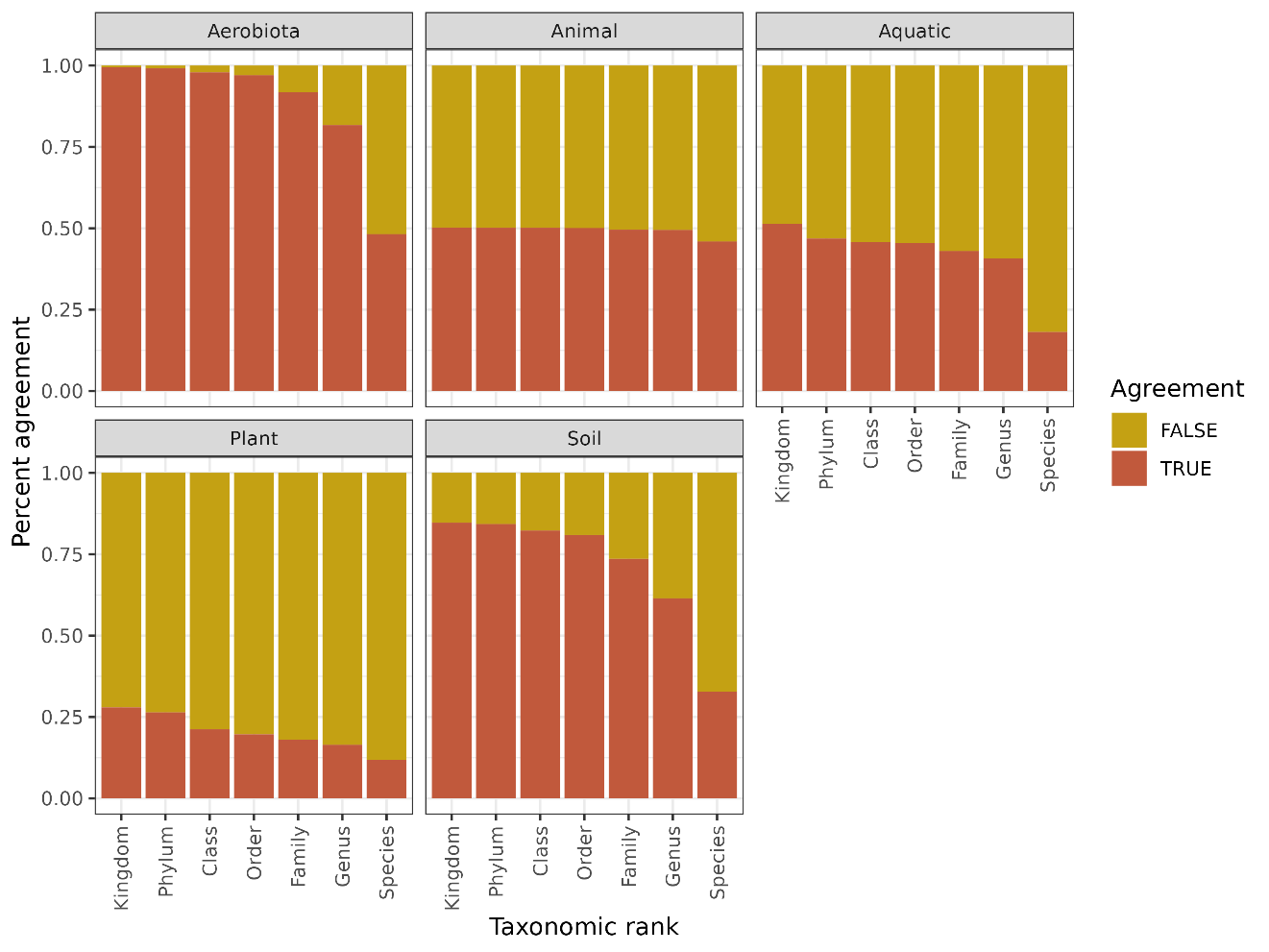
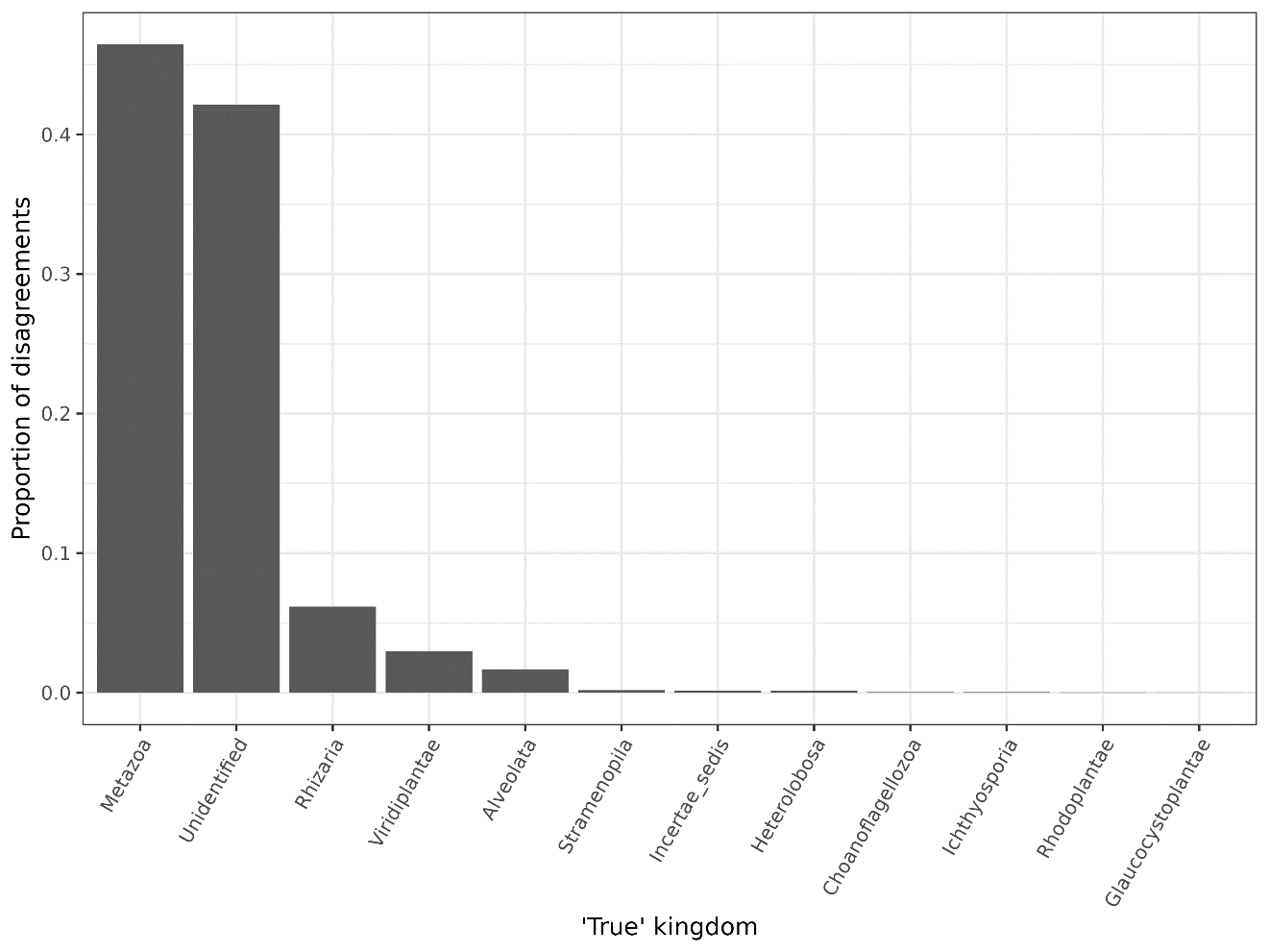
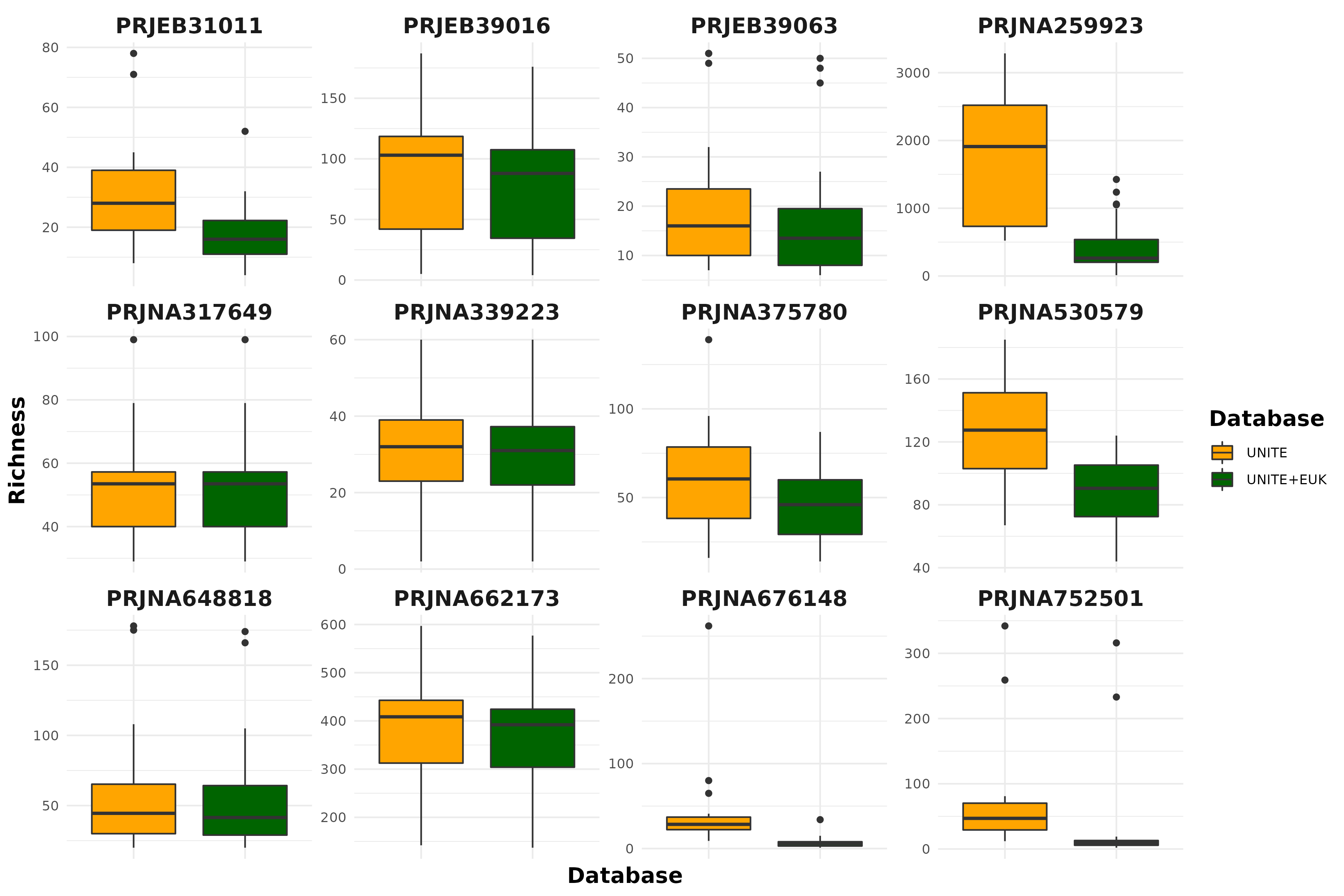


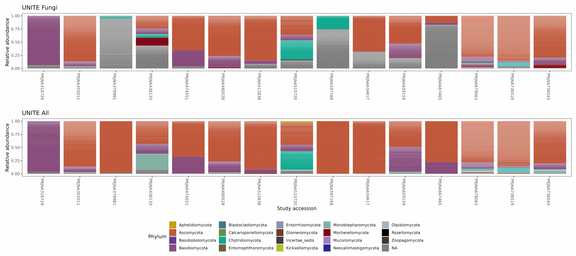
Figure 2: Figure shows the proportional agreement between databases UNITE and UNITE+EUK for the seven different habitats found in the studies. True or false agreement is shown for each of the studies on the taxonomic levels of kingdom, phylum, class, order, family, genus, and species.

Figure 3: Figure shows the proportional agreement between databases UNITE and UNITE+EUK for the five different hosts found in the studies. True or false agreement is shown for each of the studies on the taxonomic levels of kingdom, phylum, class, order, family, genus, and species.

 Figure 4: Figure shows the amount of disagreements between the twelve most common taxa found in the studies.

****

**Figure 5: Alpha diversity of each study after removing non-fungi; Falsely assigned fungi can inflate alpha diversity measures.**

**Figure 6 - Phylum relative abundance after removing non-fungi from each taxonomic assignment method. The bulk of "Fungus sp." that were assigned using UNITE\_Fungi had no phylum-level assignment (Gray fill in this plot).**

**DISCUSSION:**

**Our strong recommendation is to use a database with non-fungal outgroups such as the UNITE+Eukaryotes database (Abarenkov, et al, 2022).**

**Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; Nilsson, R. Henrik; Kõljalg, Urmas (2022): UNITE general FASTA release for eukaryotes. Version 16.10.2022. UNITE Community.** [**https://doi.org/10.15156/BIO/2483913**](https://doi.org/10.15156/BIO/2483913)

**DATA AVAILABILITY STATEMENT:**

**All raw sequenced used in this study come from previously published work and are publicly available on the Sequence Read Archive under their listed accessions (Table 1). All analysis code is publicly available as a GitHub repository release on Zenodo (Citation).**

**References**

1. Yang AM, Inamine T, Hochrath K, et al. Intestinal fungi contribute to development of alcoholic liver disease. *J Clin Invest*. 2017;127(7):2829-2841. doi:10.1172/JCI90562

2. Sitarik AR, Arora M, Austin C, et al. Fetal and early postnatal lead exposure measured in teeth associates with infant gut microbiota. *Environment International*. 2020;144:106062. doi:10.1016/j.envint.2020.106062

3. Sciavilla P, Strati F, Di Paola M, et al. Gut microbiota profiles and characterization of cultivable fungal isolates in IBS patients. *Appl Microbiol Biotechnol*. 2021;105(8):3277-3288. doi:10.1007/s00253-021-11264-4

4. Shah S, Locca A, Dorsett Y, et al. Alterations of the gut mycobiome in patients with MS. *EBioMedicine*. 2021;71:103557. doi:10.1016/j.ebiom.2021.103557

5. Pérez-Brocal V, Magne F, Ruiz-Ruiz S, et al. Optimized DNA extraction and purification method for characterization of bacterial and fungal communities in lung tissue samples. *Sci Rep*. 2020;10(1):17377. doi:10.1038/s41598-020-74137-2

6. Perera M, Al-hebshi NN, Perera I, et al. A dysbiotic mycobiome dominated by Candida albicans is identified within oral squamous-cell carcinomas. *Journal of Oral Microbiology*. 2017;9(1):1385369. doi:10.1080/20002297.2017.1385369

7. Cho O, Sugita T. Comprehensive Analysis of the Fungal Microbiota of the Human External Auditory Canal Using Pyrosequencing. *Med Mycol J*. 2017;58(1):E1-E4. doi:10.3314/mmj.16-00023

8. Khadija B, Imran M, Faryal R. Keystone salivary mycobiome in postpartum period in health and disease conditions. *J Mycol Med*. 2021;31(1):101101. doi:10.1016/j.mycmed.2020.101101

9. Jayasudha R, Das T, Chakravarthy SK, et al. Gut mycobiomes are altered in people with type 2 Diabetes Mellitus and Diabetic Retinopathy. *PLOS ONE*. 2020;15(12):e0243077. doi:10.1371/journal.pone.0243077

10. Luan C, Xie L, Yang X, et al. Dysbiosis of Fungal Microbiota in the Intestinal Mucosa of Patients with Colorectal Adenomas. *Sci Rep*. 2015;5(1):7980. doi:10.1038/srep07980

11. Gu Y, Wang Y, Wang P, et al. Study on the Diversity of Fungal and Bacterial Communities in Continuous Cropping Fields of Chinese Chives (*Allium tuberosum*). *BioMed Research International*. 2020;2020:e3589758. doi:10.1155/2020/3589758

12. Guarro J, Gené J, Stchigel AM. Developments in Fungal Taxonomy. *Clinical Microbiology Reviews*. 1999;12(3):454-500. doi:10.1128/CMR.12.3.454

13. Seed PC. The human mycobiome. *Cold Spring Harb Perspect Med*. 2014;5(5):a019810. doi:10.1101/cshperspect.a019810

14. Auchtung TA, Fofanova TY, Stewart CJ, et al. Investigating Colonization of the Healthy Adult Gastrointestinal Tract by Fungi. *mSphere*. 2018;3(2):e00092-18. doi:10.1128/mSphere.00092-18

15. Oksanen J, Blanchet FG, Kindt R, et al. Vegan: Community Ecology Package. *R Package Version 22-1*. 2015;2:1-2.

16. DADA2: High-resolution sample inference from Illumina amplicon data | Nature Methods. Accessed November 19, 2022. https://www.nature.com/articles/nmeth.3869

17. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*. 2013;8(4):e61217. doi:10.1371/journal.pone.0061217

18. Stevenson PC, Bidartondo MI, Blackhall-Miles R, et al. The state of the world’s urban ecosystems: What can we learn from trees, fungi, and bees? *PLANTS, PEOPLE, PLANET*. 2020;2(5):482-498. doi:10.1002/ppp3.10143

19. Wu B, Hussain M, Zhang W, Stadler M, Liu X, Xiang M. Current insights into fungal species diversity and perspective on naming the environmental DNA sequences of fungi. *Mycology*. 2019;10(3):127-140. doi:10.1080/21501203.2019.1614106

20. Crous PW, Carris LM, Giraldo A, et al. The Genera of Fungi - fixing the application of the type species of generic names - G 2: Allantophomopsis, Latorua, Macrodiplodiopsis, Macrohilum, Milospium, Protostegia, Pyricularia, Robillarda, Rotula, Septoriella, Torula, and Wojnowicia. *IMA Fungus*. 2015;6(1):163-198. doi:10.5598/imafungus.2015.06.01.11

21. Blackwell M. The Fungi: 1, 2, 3 … 5.1 million species? *American Journal of Botany*. 2011;98(3):426-438. doi:10.3732/ajb.1000298

22. Hibbett DS, Binder M, Bischoff JF, et al. A higher-level phylogenetic classification of the Fungi. *Mycological Research*. 2007;111(5):509-547. doi:10.1016/j.mycres.2007.03.004