communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-026-09820-7

Multiple hypervariable markers improve mycobiome classification in metatranscriptome and metagenome data

Haihua Wang[1,2], Steven H. Wu[3] ✉, Kaile Zhang[1,2], Ko-Hsuan Chen[4], Rytas Vilgalys[5] & Hui-Ling Liao[1,2] ✉

Profiling the taxonomic and functional composition of mycobiome using metagenomic and metatranscriptomic sequencing is advancing our understanding of fungal functions in ecosystems. However, the sensitivity and accuracy of mycobiome classification using genome- or core protein-based approaches, is limited by the availability of reference genomes and the resolution of sequence databases. To address this, we propose the MicroFisher, a novel tool to identify taxonomically useful reads from metagenomic or metatranscriptomic data, enabling taxonomic identification of community members based on multiple hypervariable markers. We applied MicroFisher to profile the simulated fungal communities to assess the performance of the developed tool, and found higher performance in fungal prediction and abundance estimation compared to existing tools. In addition, we also used metagenomes from forest soil and metatranscriptomes of root eukaryotic microbes to test our method and found that MicroFisher provided more accurate profiling of environmental microbiomes compared to other classification tools. MicroFisher leverages high-resolution hypervariable marker gene databases and weighted integration algorithms to deliver more accurate fungal community classification compared to existing state-of-the-art tools. Additionally, it enables the detection of rare taxa, which is challenging with other available tools. Thus, MicroFisher serves as a novel pipeline for classification of fungal communities from metagenomes and metatranscriptomes.

Eukaryotic microbiota that comprise a major component of microbial communities in terrestrial and marine ecosystems perform crucial roles in regulating biogeochemical processes. Therefore, it is critical to study the community structure and functional composition of fungi to better understand their ecosystem functions[1]. Next-generation sequencing approaches, including metagenomics and metatranscriptomics, provide revolutionary ways to identify critical functions of mycobiome in environments and within a host organism[2–5]. In addition to functional profiling, metagenomics and metatranscriptomics offer opportunities to characterize mycobiome community composition and diversity at multi-scales[6]. However, the classification of fungal taxonomy through metagenomic and metatranscriptomic sequencing remains challenging due to the limitations of current sequencing platforms and the lack of comprehensive databases[7].

Current methods for taxonomic profiling using metagenomic and metatranscriptomic sequences mainly rely on the read-mapping approach[8–10] and the read-assembly approach[11–13]. The read- (or k-mer) mapping approach, as used by tools such as Eukdetect[14], Kaiju[9], Kraken2[8], and Metaphlan3[10], is a method for profiling microbial taxa via mapping reads to pangenomic references, followed by taxa prediction. However, this approach relies on the availability of genome references, which limits the identification of many taxa in environmental samples. The limited availability of fungal reference genomes significantly constrains the taxonomic resolution of fungal profiling in metagenomic and metatranscriptomic analyses[15–17]. By contrast, the read-assembly approach allows for the identification of microbial community composition via performing de-novo assembly and binning to reconstruct genomes and identify taxonomic composition of bacteria, archaea, and viruses[18,19]. However, it is challenging to generate accurate de novo assemblies for eukaryotes from metagenomic and metatranscriptomic datasets using this approach, since eukaryotic genmones are typically larger and contain highly conserved and repetitive regions, which complicate assembly and increase the likelihood of false positive assignments that are indistinguishable from true identifications[20].

[1]North Florida Research and Education Center, University of Florida, Quincy, FL, USA. [2]Department of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, USA. [3]Departement of Agronomy, National Taiwan University, Taipei, Taiwan. [4]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. [5]Department of Biology, Duke University, Durham, NC, USA. ✉e-mail: stevenwu@ntu.edu.tw; sunny.liao@ufl.edu

The presence of multiple chromosomes in a fungal genome can lead to inaccurate predictions when using a de novo assembly classification approach, potentially causing different chromosomes from the same fungal genome to be misclassified as different species[21]. The limitations of these two methods impede comprehensive understanding of metagenomic-level community composition and function in complex ecosystems, particularly for eukaryotes. Thus, it is crucial to develop a non-genome, hypervariable marker-based classifier to improve the accuracy and precision of fungal taxonomic classification in sequencing datasets.

Although pangenome-[8,10] and core protein domain-based[22] profiling pipelines have been widely used for the taxonomic classification of metagenomic and metatranscriptomic sequencing, a number of marker regions (such as ITS, LSU, SSU, COX1, RPB1, RPB2, β-tubulin, MCM7, TEF1-α, γ-actin, ATP6 and CaM genes) have been recognized to provide higher resolution taxonomic composition and more reliable prediction of microbial communities[23]. This is because these marker genes provide taxa-specific fingerprints for classification[24]. For instance, there is a long history of identifying fungal taxa based on the marker genes derived from ribosomal RNA (rRNA) regions, and genes such as elongation factor 1α[25] and beta-tubulin[26] have also been used to identify certain taxonomic groups. Over the years, researchers have developed a variety of analytical tools and curated many databases that contain a massive number of marker sequences for fungal taxonomic classification, such as UNITE[27], SILVA[28], and RDP[29]. However, due to highly conserved regions present in those marker sequences[5], both read-assembly and read-mapping approaches can result in substantial false-postive assignments in fungal classification from metagenomic and metatranscriptomic datasets. Although most of the ribosomal markers are eliminated during library preparation, the residual ribosomal sequences that remain can still provide valuable taxonomic information. For example, based on previous studies, ~22.05–31.98% of ribosomal RNA may still remain in metatranscriptomic datasets prepared using the RiboZero kit[30].

We present a computational workflow (MicroFisher) that can produce accurate and sensitive taxonomic profiles of fungal communities from metagenomes and metatranscriptomes. MicroFisher uses databases of hypervariable internal transcribed spacer (ITS) and Large-Subunit (LSU) rRNA markers, which provide species-specific fingerprints for fungal identification, and maps intact reads directly to reference markers. Abundances of taxa are estimated from multiple markers, using a "weight-based abundance integration algorithm". Overall, this study provides an initial framework for accurately profiling the fungal community from sequencing datasets by extracting hypervariable and frequently used markers. In the long run, the MicroFisher protocol is capable of incorporating newly developed marker sequences that are specialized for certain fungal groups. MicroFisher has the potential toclassify not only fungal but also bacterial communities. While our current database focuses exclusively on fungi, we intend to expand the tool to include bacterial groups in future work.

Methods

Generation of hypervariable marker databases

In this study, we selected four sets of hypervariable marker sequences, including LSU rRNA D1 and D2 domains as well as ITS1 and ITS2, which are highly variable and popular marker sequences for fungal identification, specifically focusing on highly polymorphic sub-regions (Fig. 1A), and comprise a large number of reference sequences available in open-access databases.

Pre-filtration of fungal hypervariable marker sequences (Fig. 1A, Steps 1–4)

We developed a four-step process to pre-filter marker sequences from fungal nucleotide sequences originally obtained from International Nucleotide Sequence Database Collaboration (INSDC)[31]. Specifically, a total of 16,301,484 fungal nucleotide sequences, including genomic DNA, mRNA, and rRNA, were collected using the esearch command with taxonomy ID 4751 (sequence accession numbers are listed in Supplementary Data Sheet 1; Step 1). In Step 2, fungal LSU rRNA sequences were predicted by a hidden Markov model using RNAmmer[32] and HMMER2[33] with default settings, and then Metaxa2[34] with default parameters was used to remove non-LSU rRNA sequences. In Step 3, a custom pipeline was developed to extract the marker sequences. The universal LSU primer sequences (Supplementary Table S1) were applied to extract the D1 and D2 sequences (Supplementary Table S1) using SeqKit[35]. For ITS sequences, the pipeline first identified ITS sequences based on universal ITS primer sequences (Supplementary Table S1), then ITSxpress[36] was applied to predict and extract fungal ITS1 and ITS2 sequences. Chimeric sequences were subsequently detected and removed from the resulting ITS1 and ITS2 database using UCHIME[37] in high-confidence mode. The extracted ITS and LSU rRNA D1 and D2 sequences served as input for the next step after removing sequences with unexpected length (sequence length >1000 bp) (Step 4).

Detection of hypervariable marker region (Fig. 1A, Steps 5–8)

We applied Entrez-direct[38] and Taxonkit[39] to refine the taxonomic identities of sequences collected from Step 4 against the NCBI Taxonomy database (Step 5). For each marker, sequences assigned to the same order were grouped and aligned using ClustalW[40] with default parameters (Step 6). Alignment at the order level was applied to facilitate clear characterization of hypervariable regions. Our in-house tests suggested that multiple sequence alignment results at the order level retain the high sequence variety of fungal fingerprint for species-specific identification. The multiple sequence alignments were used to identify hypervariable markers. Regions with low consensus (i.e., containing more gaps) were designated as candidate regions for "hypervariable marker" (Step 7) (Supplementary Fig. S1). Each base site was quantified among these candidate regions using the bit value that was calculated based upon the equation described by Bembom and Ivanek (2020)[41]. Using seqLogo, regions with low bit values between conserved flanking regions withhigh bit values were identified as "hypervariable markers" to serve as marker fingerprints for fungal classification[41–43] (Supplementary Fig. S2). The conserved flanking regions of each "hypervariable marker" were then truncated and the remaining subset of hypervariable sequences was clustered to remove duplicate sequences with similarity >99.99% using Vsearch package[44] (Step 8). The obtained hypervariable sequences were pooled to generate a database for each marker, resulting in ITS1, ITS2, LSU D1, and LSU D2 databases. The average length of each marker is 188.8 bp for ITS1, 186.8 bp for ITS2, 172.8 bp for D1, and 189.4 bp for D2 (Supplementary Table S2).

In total, we retrieved 405,672 ITS1 sequences, 294,737 ITS2 sequences, 81,704 LSU D1 sequences, and 121,580 LSU D2 sequences (Supplementary Table S2). A total of 903,693 hypervariable marker sequences span 10 phyla, 62 classes, 234 orders, 873 families, 6545 genera, and 104,072 species. Approximately 19,714 fungal species were shared among all four databases (Supplementary Fig. S3). Since the process of creating these hypervariable marker databases is computationally intensive, we provide the latest reference databases and updated information for users to download for direct use in the MicroFisher pipeline.

MicroFisher pipeline: classification, integration, and optimization

Alignment: align the qualified reads of metagenomics and metatranscriptomics to marker databases. After quality trimming, metagenomic and metatranscriptomic reads generated from environmental samples (e.g., soils, marine, plants, and animals) were used as the input files for the MicroFisher pipeline. Firstly, MicroFisher classifier searches these reads against the marker databases using Centrifuge[45] with default settings, except for the minimum hit (MiniHit) length. In MicroFisher, Centrifuge[45] serves as the microbial classification engine, which identifies candidate taxa and estimates their abundances, using an indexing scheme based on the Burrows–Wheeler transform (BWT) and the Ferragina–Manzini (FM) index. This step generates one search report for each marker database. In this step, the MiniHit length is one of the key

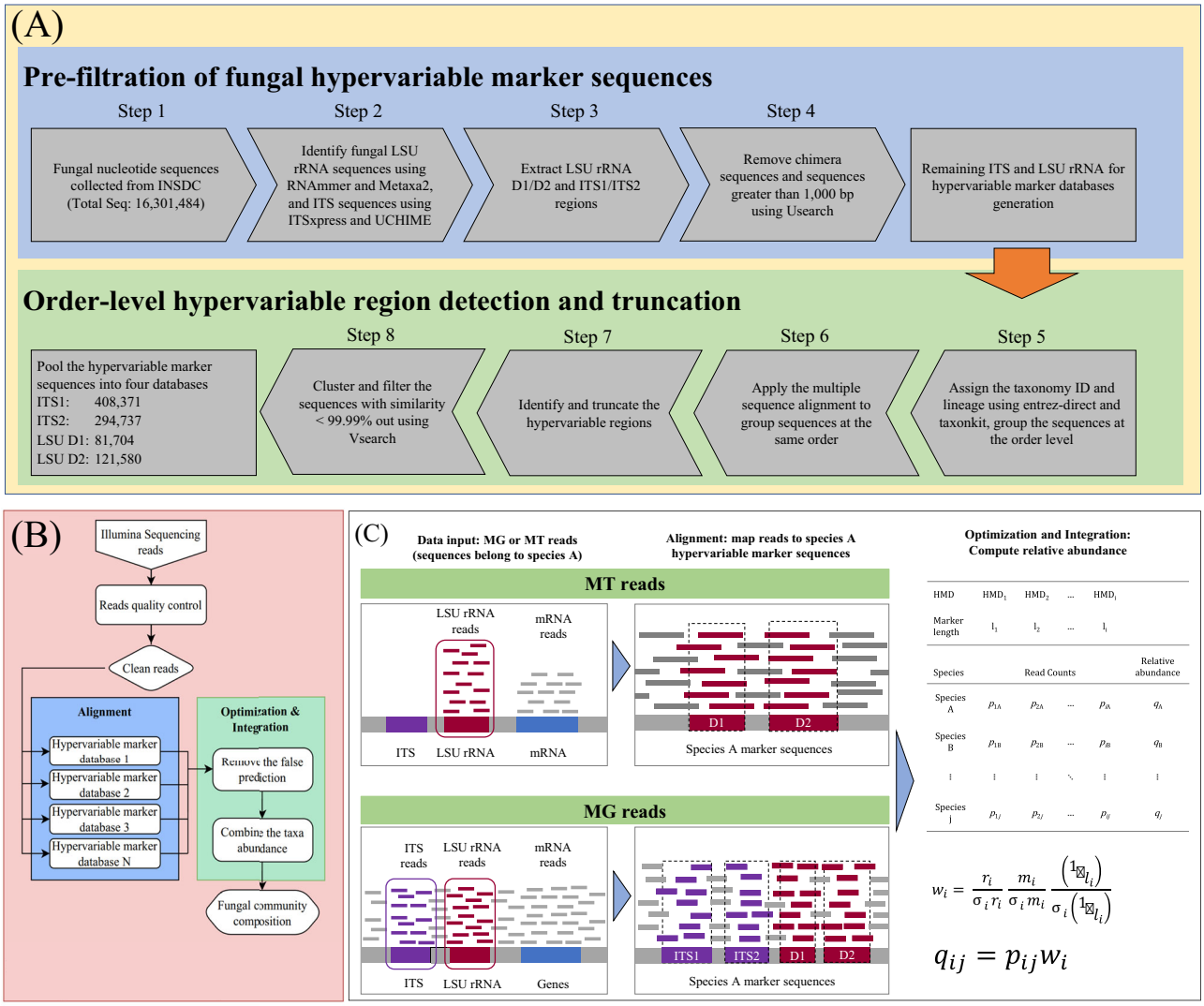

Fig. 1 | Overview of hypervariable marker database generation and the principle of MicroFisher classification. A Computational flowchart for the generation of hypervariable marker databases. **B**, **C** MicroFisher pipeline for fungal identification and abundance estimation in metagenomic and metatranscriptomic analysis (1) Alignment: aligning the short reads in metagenomes and metatranscriptomes to multiple hypervariable marker databases; (2) Optimization and integration: use a weight-based algorithm to remove false positives and optimize the estimated abundance of each taxon from multiple hypervariable databases. The final output results were returned fungal community composition across taxonomic levels. q_{ij} is the weighted proportion for an identified taxon j against marker database i, p_{ij} is the original abundance of identified taxon j estimated by Centrifuge against marker database i, and w_i is the weighting associated with the abundance report from a particular marker database i. r_i is the total number of reads obtained in the Centrifuge classification aligned to marker database i, m_i is the minimum length parameter (MiniHit) used in the Centrifuge classification for the reads aligned to marker database i, and l_i is the mean length of hypervariable marker sequences in database i.

parameters in Centrifuge to determine the minimum alignment length required for reads to match a marker sequence in the hypervariable marker databases. Different from standard Centrifuge classifcation, which only applies a single marker database at a time, MicroFisher integrates search reports from Centrifuge using multiple hypervariable marker databases for microbial classification. Therefore, the MiniHit length is critical for the classification sensitivity and accuracy of MicroFisher.

Integration and optimization: integrate abundance reports using weight-based algorithms. The MicroFisher pipeline generates multiple abundance reports obtained from the reads number aligned to different marker databases. Subsequently, weight-based algorithms are applied to calculate the integrated abundance of each taxon and create a "final abundance table". By default, MicroFisher identifies fungal taxa and calculates their final abundances only when they are detected through alignment with two or more marker databases to enhance the precision of taxonomic classification. Then, weight-based algorithms are applied to improve the accuracy of abundance estimation of detected taxa (Fig. 1C).

Three factors are incorporated in the algorithms to weight the abundance reports: 1) the total number of reads aligned to each marker database (e.g., databases with more aligned reads receive higher weights), 2) the MiniHit length, and 3) the inverse weighting of the average length of hypervariable marker sequences in each marker database (i.e., databases with longer average length are assigned lower weights; Supplementary Table S2). Incorporating these factors in the algorithms improves both the accuracy and precision of fungal taxonomic identification.

$$q_{ij} = p_{ij} w_i \quad (1)$$

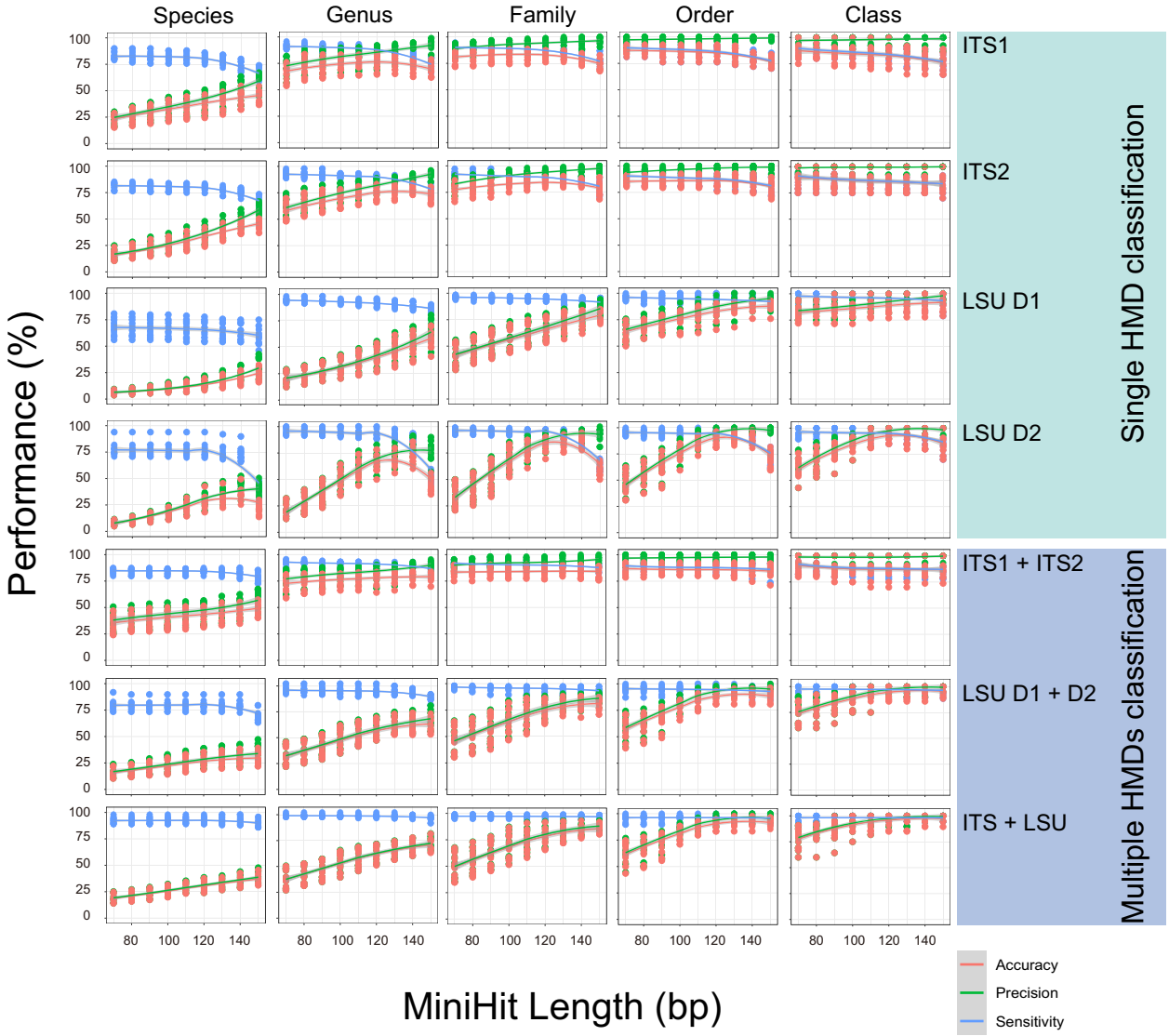

Fig. 2 | Performance of MicroFisher's fungal classification. Fungal classification and profiling were performed on these simulated communities using MicroFisher (multiple hypervariable marker databases applied) and Centrifuge (single hypervariable marker database applied). The simulated communities (151 bp) were simulated from three levels of fungal composition complexity (50, 100, and 200 fungal species), with five simulated sequencing datasets for each level. Accuracy (in red), precision (in green) and sensitivity (in blue) were calculated at each taxonomic level across a range of Minimum Hit Length (MiniHit Length) values from 70 to 150 bp.

where q_{ij} is the weighted adjusted abundance for an identified taxon j against marker databases i, p_{ij} is the original abundance of identified taxon j estimated by Centrifuge against marker database i, and w_i is the weighting associated with the abundance report from a particular marker database i, which is subsequently used to merge reports from multiple marker databases. w_i can be calculated by the following formula:

$$w_i = \frac{r_i}{\sum_i r_i} \frac{m_i}{\sum_i m_i} \frac{(1/l_i)}{\sum_i (1/l_i)} \quad (2)$$

Where r_i is the total number of reads obtained in the Centrifuge classification aligned to marker database i, m_i is the MiniHit length used in the Centrifuge classification for the reads aligned to marker database i, and l_i is the mean length of hypervariable marker sequences in marker database i.

The value of these three factors (number of reads, MiniHit length, average length of marker databases) can be configured by users to achieve alternative weighting schemes if needed. By setting all parameters to the same value for a dataset aligned to different marker databases, MicroFisher applies equal weighting across all databases.

Using MicroFisher to profile fungal communities with simulated and real datasets

Fungal community profiling for simulated sequencing dataset. To optimize the parameters of MicroFisher, a small-scale simulated dataset composed exclusively of sequences covering the ITS and LSU rRNA regions was generated to facilitate an accurate performance evaluation. We evaluated MicroFisher performance by testing its ability to profile the fungal community from a simulated sequencing dataset with known ground truth. This simulated sequencing data was generated using the InSilicoSeq sequencing simulator with default kernel density estimation error mode[46], simulated under NovaSeq error model with read length 151 bp. Fungal rRNA sequences sourced from the NCBI RefSeq database were used as references. We simulated three fungal taxa pools by randomly selecting 50, 100, and 200 fungal species with their sequences covering both ITS and LSU rRNA regions (Supplementary Table S3). Five replicates were conducted at each pool level, each with a different randomly selected fungal composition, resulting in a total of 15 simulated communities with 5 million (M) noisy reads in each dataset, which reflect rRNA and ITS reads with substitution, insertion and deletion errors in

real metagenomic and metatranscriptomic datasets. Reads of the selected fungal species in these simulated communities were distributed exponentially to represent the gradient of dominant toward rare fungal taxa. These 15 simulated community datasets then served as data input for fungal classification. The four hypervariable marker databases generated in section (a), including ITS1, ITS2, LSU D1, and LSU D2, were used individually for classification. Meanwhile, the MicroFisher package was applied with the default algorithm described in section (b) using three marker database combinations, including combination of ITS1 and ITS2 databases (ITS only: ITS1 + ITS2), combination of LSU D1 and LSU D2 databases (LSU only: LSU D1 + D2), and a combination of four databases (ITS + LSU). We evaluated the effects of the MiniHit length parameter (between 70 and 150 bp) on prediction of the fungal community profile. Classification accuracy was assessed by comparing taxanomic composition and abundance estimated through Centrifuge and MicroFisher againstthe original simulated fungal species pools.

In addition, simulated metagenomic datasets were generated to evaluate the performance of MicroFisher compared to existing tools, including Eukdetect, Metaphlan, Kraken2, and Kaiju, in fungal community prediction. Fungal genomes downloaded from the NCBI database and their representative ITS and LSU rRNA sequences obtained from NCBI RefSeq database were used as references for metagenomic simulation by InSilicoSeq sequencing simulator with the same parameters described above. To enrich the proportion of reads simulated from ITS and LSU rRNA genes, these representative sequences were reproduced 1000 times and merged into the genome reference file. Using these fungal genome files, simulated metagenomic sequencing datasets (500 M reads) were generated with 20 and 50 randomly selected fungal species followed an exponentially distribution mode. Then, the composition and abundance of fungal communities in these simulated metagenomic sequencing datasets were predicted using MicroFisher and existing tools.

Three metrics were applied to evaluate the performance of MicroFisher in fungal classification, including accuracy, precision, and sensitivity (Fig. 2). True positive (TP), false positive (FP), and false negative (FN) of fungal classification across different taxonomic levels were identified. These metrics were calculated followed the formulas described by Lu and Salzberg (2020)[5]. Briefly, sensitivity was defined as TP/(TP + FN); accuracy was calculated by TP/(TP + FP + FN), and precision was calculated by TP/(TP + FP) (Definitions are provided Supplementary Table S5). These metrics were plotted to illustrate the effect of the database-combination algorithm, MiniHit length, and community complexity on MicroFisher's prediction performance (Fig. 2 and Supplementary Fig. S4). Furthermore, histogram plots of predicted abundance were applied to compare the abundance distributions of true and false predictions (Figs. 3A and 4). The results showed that MicroFisher accurately classified fungal genera with abundances higher than 0.1%; thus, genus-level classification profiles of abundant fungi (abundance >0.5%) from simulated metagenomic datasets using MicroFisher were used for comparison with existing tools. Furthermore, the F1 score was calculated to compare the overall classification performance of MicroFisher to other existing tools at the genus level [$F1\ score = 2 * (\text{precision} * \text{sensitivity})/(\text{precision} + \text{sensitivity})$]. We also calculated the absolute error of abundance and root mean square (r.m.s.) error of MicroFisher predictions compared to the actual taxon abundances (Fig. 3B and Supplementary Fig. S5). The absolute error for a fungal taxon was defined as the difference between estimated abundance and the actual abundance. The r.m.s error was used to measure the error rate of taxon abundance estimates (formula listed in Supplementary Table S5). Finally, Pearson correlation analysis was performed to assess the relationship between estimated and actual abundances of taxa (Fig. 3C and Supplementary Figs. S6–S8).

Fungal community profiling from environmental metagenomic and metatranscriptomic data. We further classified fungal composition and estimated taxa abundance in real metagenomic and metatranscriptomic datasets (Supplementary Table S4) using both MicroFisher and existing tools to compare their performance in fungal taxonomic profiling. The sequencing datasets selected for this study included pine forest soil metagenomes, their fungal amplicon sequences, and pine root metatranscriptomes. Soils were sampled from *Pinus* forests located in Jelonan and Belanglo, Australia and in Cambria, Monterey, and Swanton, United States. Total soil DNA was extracted using the DNA PowerSoil kit (MoBio, Carlsbad, California, USA) following the manufacturer's instructions. Shotgun metagenomic sequencing of the soil DNA samples was performed at the DOE Joint Genome Institute (JGI), United States. Moreover, the collected soils were used to grow *Pinus radiata* seedlings for three months and pine roots were harvested for total RNA extraction using the CTAB/LiCl method following the protocol described in Liao et al.[47]. Poly-A enriched cDNA library construction method targeting eukaryotic mRNA was applied to construct RNAseq libraries for pine root metatranscriptomic sequencing. Sequencing was conducted using an Illumina NovaSeq 6000 system (Illumina Inc., San Diego, CA, USA) at JGI. Fungal classification in these sequencing datasets was performed by MicroFisher. To compare performance of MicroFisher with other alignment-based taxonomic classification techniques, we also ran taxonomic classification in metagenomic and metatranscriptomic datasets using other classification tools, including Eukdetect, Kaiju, Kraken2, and Metaphlan3. Default settings were applied for taxonomic classification. Amplicon sequencing was conducted for the same soil DNA extracts that were used to generate metagenomic data as a standard result in order to compare the fungal classification performance of MicroFisher against the existing tools. A two-step PCR approach[48] targeting ITS1 and ITS2 regions (Primers: ITS1F/ITS4) was used to generate the amplicon library, which was sequenced on an Illumina MiSeq platform (v3 300 bp, 13 Gb sequencing capacity) at the Duke Center for Genomic and Computational Biology, United States. Soil fungal ITS amplicon sequences were processed using QIIME2[49]. Community composition profiles derived from soil metagenomes, root metatranscriptomes, and amplicon sequencing datasets were compared and visualized using ggplot2[50] package in R.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Database and pipeline overview

We developed the MicroFisher package, a comprehensive bioinformatics tool for analyzing fungal community composition from metagenomic and metatranscriptomic sequencing datasets using multiple hypervariable marker databases. Our approach integrates 104,072 fungal species spanning 6545 genera (Fig. 1A, Supplementary Table S2), utilizing four curated hypervariable markers from ITS1, ITS2, LSU D1, and LSU D2 with sequence lengths ranging from 120 to 350 bp (Fig. 1A). Through applying hypervariable sequence markers, the compiled databases provide high-resolution references for fungal classification from metagenomic and metatranscriptomic datasets. The MicroFisher pipeline employs multiple hypervariable marker databases (including ITS1, ITS2, LSU D1, and LSU D2) and weight-based integration algorithms to accurately classify taxonomic composition of fungal communities from metagenomics and metatranscriptomic sequences (Fig. 1B, C). The first step of the MicroFisher pipeline is to map reads to several marker databases and generate corresponding abundance reports using Centrifuge. Then, weight-based algorithms are further employed to optimize the classified fungal composition and integrate these reports into a final abundance table for detected taxa. The weight-based integration algorithms incorporate three key factors (the total number of mapped reads, the MiniHit length, and average sequence length of the mapped data). The weight-based integration algorithms effectively reduce mismatches (false positives), and the combination of multiple marker databases lowers the probability of missed taxa (false negatives). This innovative approach provides a robust, multiple

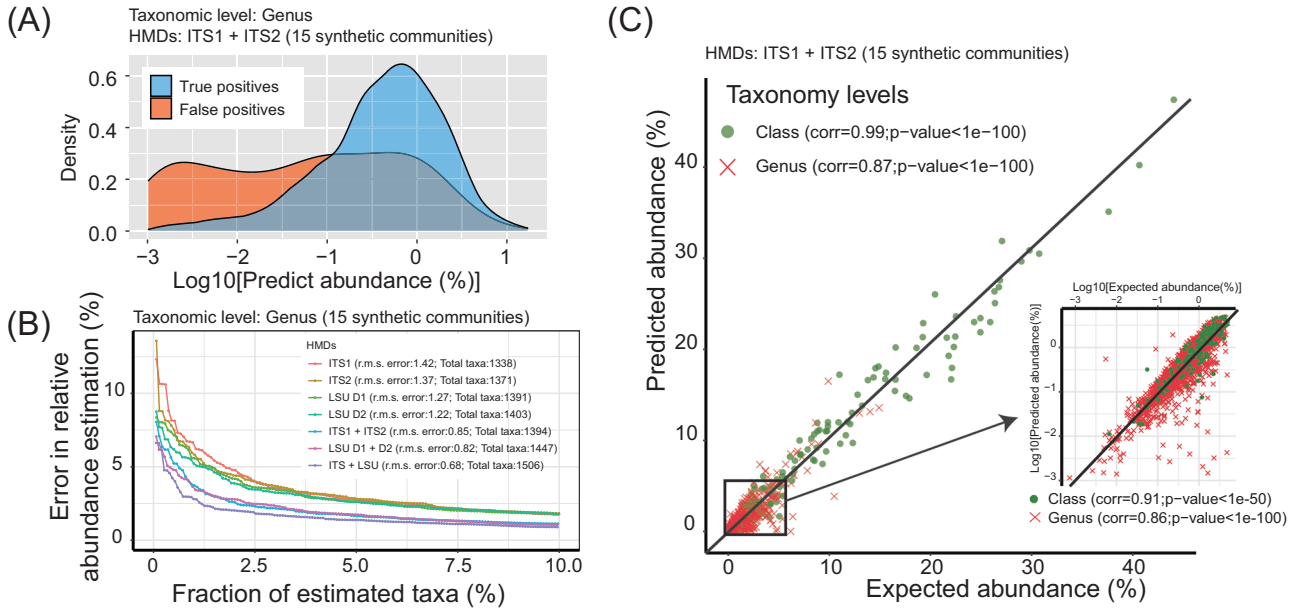

Fig. 3 | Evaluation of accuracy of abundance estimated from 15 simulated sequencing datasets using MicroFisher. Fungal classification and abundance estimation were performed using a MiniHit length of 120 bp in the MicroFisher pipeline. **A** Density plot showing the distribution of fungal abundances (at the genus level) for true positives and false positives. **B** Absolute error and root mean square (r.m.s) error estimated when predicting abundances of fungal taxa at the genus level using single marker databases (ITS1, ITS2, LSU D1, and LSU D2; profiled by Centrifuge with a MiniHitlength of 120 bp) and multiple marker databases (ITS1 + ITS2, LSU D1 + D2, and ITS + LSU; profiled by MicroFisher with default settings). Taxa were ordered by the value of error and exhibited with a fraction of the estimated taxa. Root mean square (r.m.s.) error was calculated for each hypervariable marker combination, and the total number of true positive predictions detected by different combinations from 15 simulated communities was indicated as "Total taxa: numbers". **C** Correlations between true and inferred abundances for taxa estimated from these 15 simulated sequencing datasets. Green and red dots denote taxa at the class and genus level, respectively. To explore the correlations for taxa with abundances less than 5%, abundance values were $\log_{10}$-transformed before performing the Pearson correlation analysis. Pearson correlation coefficients and *P*-values for each taxonomic level were calculated. MicroFisher classification and abundance evaluation were conducted using ITS1 + ITS2 hypervariable markers with default parameters (MiniHit length = 120 bp, weighted and filter mode).

hypervariable marker framework for comprehensive and accurate fungal taxonomic prediction from metagenomic and metatranscriptomic datasets, overcoming key limitations of traditional classification methods.

Evaluate and optimize MicroFisher performance in fungal community classification

We applied the MicroFisher pipeline to a small-scale simulated sequencing dataset generated using fungal ITS and LSU rRNA genes to evaluate and optimize its performance in fungal classification and abundance estimation. Overall, most expected species were identified from the simulated sequencing dataset (~92%) using MicroFisher. The results showed that the combination of hypervariable marker databases and the MiniHit length of mapped reads primarily determined MicroFisher's performance. For example, fungal community classification improved when multiple marker databases were used instead of a single database (Fig. 2). By applying a MiniHit length of 120 bp for genus-level detection, the average sensitivity, accuracy, and precision using multiple marker databases were 95%, 66%, 69%, respectively, compared with 90%, 64%, 69% when a single marker database was used. Specifically, classification using LSU D1 + D2 markers resulted in 95% sensitivity, 59% precision, and 57% accuracy, while applying LSU D1 database only achieved91% sensitivity, 42% precision, and 40% accuracy. In addition, we found that prediction accuracy and precision of MicroFisher increased with increasing MiniHit length (Fig. 2). Compared with MiniHit lengths of 80 bp and 100 bp, a MiniHit length of 120 bp or greater yielded better accuracy, precision, and sensitivity. The average precision of taxonomic predictions was 36%, 69%, and 83% at the species, genus, and family levels at a MiniHit length of 120 bp, respectively. However, reducing the MiniHit length to 80 bp reduced precision to 27%, 53%, and 66% across these taxonomic levels (Fig. 2). At a MiniHit length of 120 bp, 86%, 95%, and 94% of expected fungal taxa on average were recovered at the species, genus, and family levels, respectively, reflecting high sensitivity of fungal prediction using MicroFisher. However, increasing the MiniHit length to 150 bp reduced recovery to 79%, 91%, and 92% at the same taxonomic levels. Based on these results, aMiniHit length of 120 bp was then set as the default parameter in the MicroFisher pipeline. More importantly, genus- and higher-level classifications were recommended for downstream analyses because of their higher precision (>60%).

We further evaluated the accuracy of abundance estimates generated by MicroFisher (Fig. 3). After implementing the weight-based integration algorithms, predicted microbial abundances demonstrated high concordance withthe actual abundances, particularly when utilizing a comprehensive multiple marker database combination for taxonomic mapping (Figs. 3A and 4). The consistency between predicted and true positive abundance profiles suggests robust performance of the integration methodology in accurately reconstructing microbial community composition. In addition, most dominant (relative abundance >1%) and intermediate (relative abundance between 0.01% and 1%) fungal taxa in predicted results were accurately (82.9%) assigned across the genus (Fig. 3A) and higher taxonomic levels (Fig. 4) when multiple marker combinations were applied. Notably, for fungal genera with relative abundance greater than 0.1%, the true positive rate (91.8%) was much higher than the false positive rate (8.2%) (Fig. 3A), with similar patterns observed at higher taxonomic levels (Fig. 4), indicating that MicroFisher can precisely predict dominant community composition. Simultaneously, removing fungal taxa with relative abundance below 0.1% could reduce the false positive rate from 14.11%, 38.56%, and 34.97% to 8.21%, 22.45%, and 18.41% when using ITS1 + ITS2, LSU D1 + D2, and ITS + LSU at the genus level, respectively. This suggests that abundance-based filtering could be a practical way to filter out misclassified taxa manually. However, we also found that low-abundance taxa had a higher probability of false predictions at the species level using MicroFisher (Fig. 4). This finding indicates that hypervariable markers are limited in their ability to resolve species-level identification with high accuracy. The

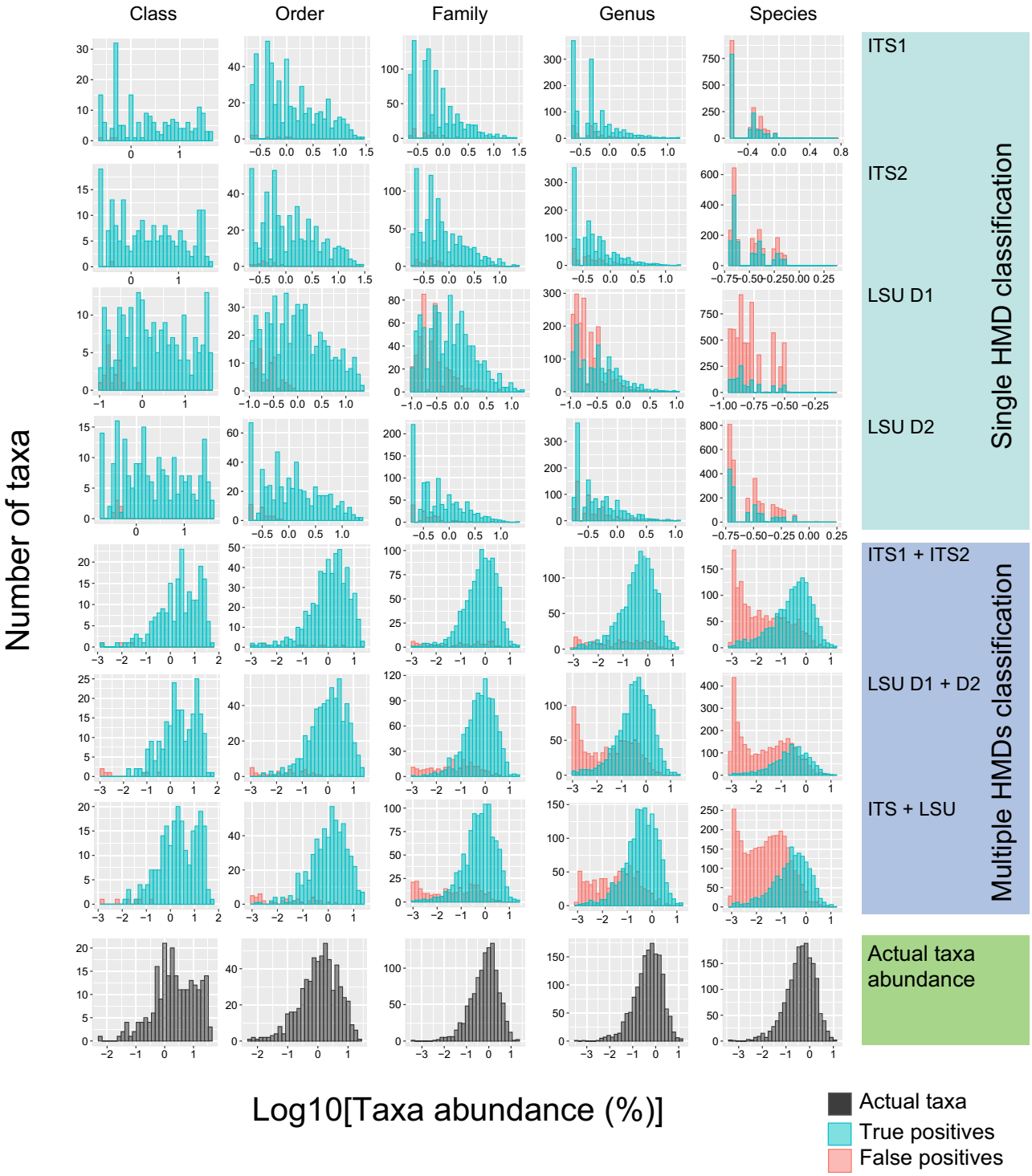

Fig. 4 | Histogram plots comparing patterns of fungal abundances across taxonomic levels between predicted results and actual data. Single hypervariable markers (ITS1, ITS2, LSU D1, and LSU D2; profiled by Centrifuge) and multiple (ITS1 + ITS2, LSU D1 + D2, and ITS + LSU; profiled by MicroFisher with default settings) were applied to compare the similarity of taxon abundance patterns. A minimum hit (MiniHit) length of 120 bp was applied for Centrifuge and MicroFisher.

absolute error identifying the percentage difference between measured and actual abundance of each taxon was also calculated to evaluate accuracy of abundance predictions (Fig. 3B, Supplementary Fig. S5). Overall, the abundances of fungal taxa estimated by MicroFisher using multiple markers had higher accuracy compared to single markers at the genus level (Fig. 3B) and across other taxonomic levels (Supplementary Fig. S5). At the genus level, only 1.2–2.7% of detected taxa had abundance errors higher than 2.5% when multiple marker databases were applied, compared with 5.4%–6.9% when any of the four single markers was used alone. The results suggest that this package can estimate taxon abundances of the fungal community with low error rates across different taxonomic levels.

Similar patterns were found while evaluating accuracy of predicted taxon abundances with r.m.s. errors. For example, r.m.s. errors of 0.85, 0.82, and 0.68 were identified using ITS1 + ITS2, LSU D1 + D2, and ITS + LSU, respectively, whereas higher r.m.s. errors of 1.42, 1.37, 1.27, and 1.22 were observed using ITS1, ITS2, LSU D1, and LSU D2 (Fig. 3B; MiniHit length of

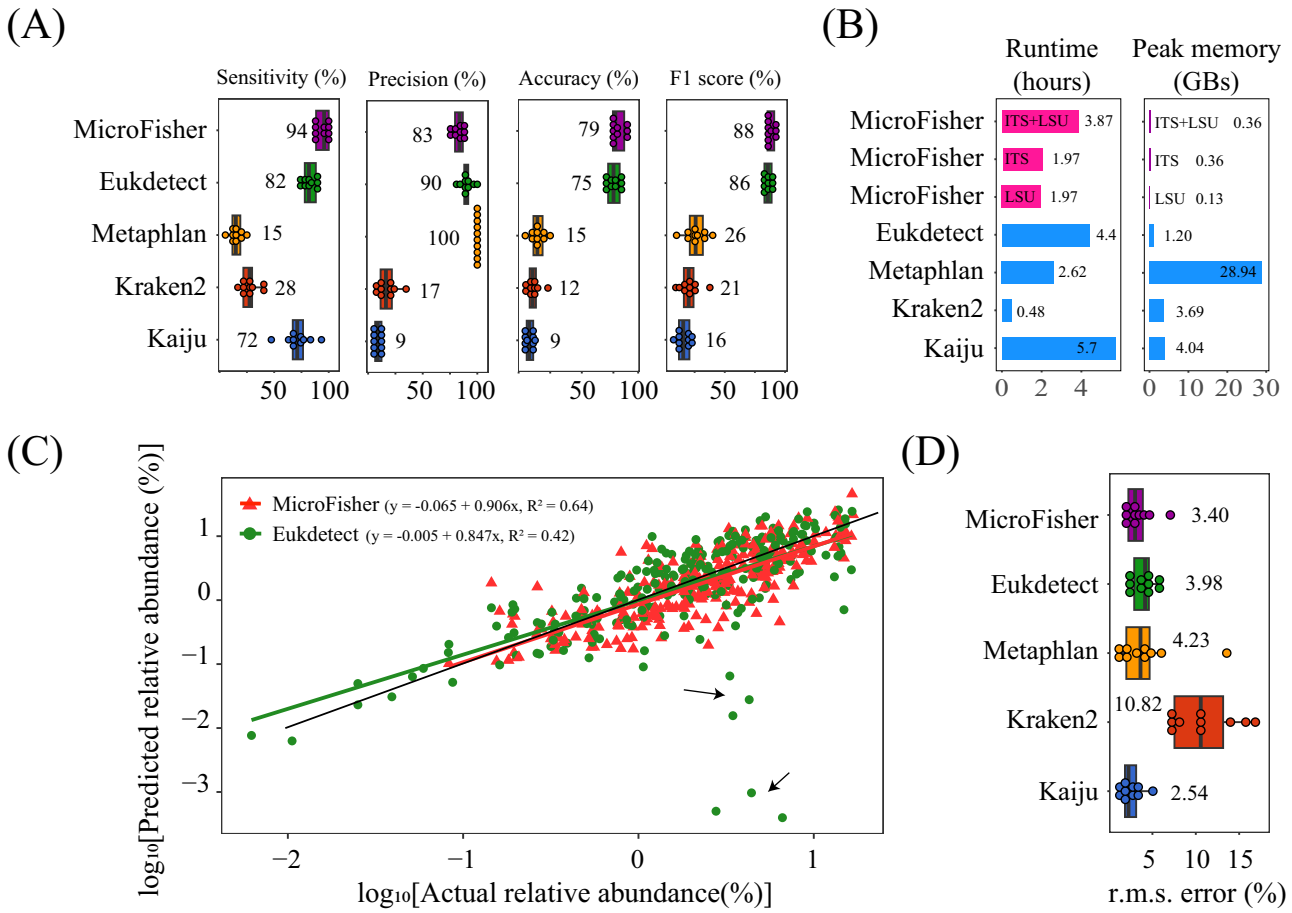

Fig. 5 | Comparison of MicroFisher in mycobiome classification with existing tools. Simulated metagenomic datasets were generated using fungal genome sequences and representative ITS and LSU rRNA sequences. Bioinformatic tools, including MicroFisher, Eukdetect, Metaphlan, Kraken2, and Kaiju, were applied to predict composition and taxon abundances in the simulated metagenomic datasets with default settings. After rare taxa (relative abundance <0.1%) classified by MicroFisher were removed, all comparisons were conducted at genus level. **A** Identification accuracy of MicroFisher and existing tools in mycobiome classification at the genus level. **B** Computational performance comparison of MicroFisher and other methods in profiling fungal community. Runtimes reflect multithreaded execution on 8 CPU cores for 500 million reads. **C** Correlations between true and inferred abundances of genera estimated from simulated metagenomic datasets. Green and red dots denote genera classified by MicroFisher and Eukdetect, respectively. **D** Root mean square (r.m.s) error of the abundances of fungal genera estimated by MicroFisher and existing tools.

120 bp). Moreover, high correlations between actual and inferred abundances at the genus (Pearson r = 0.87, *P*-value < 1e−100) and class (Pearson r = 0.99, *P*-value < 1e−100) levels were observed when ITS1 + ITS2 markers were applied in MicroFisher (Fig. 3C), with similar patterns observed for LSU D1 + D2 and ITS + LSU (Supplementary Fig. S6). High correlations between actual and inferred abundances were also consistently detected across all taxonomic levels when multiple marker databases were applied (Supplementary Figs. S7, S8). In addition, taxa with real and inferred log-transformed abundances less than 5% showed high correlations (genus: r = 0.86, *P*-value < 1e−100; class: r = 0.91, *P*-value < 1e−50; Fig. 3C, subfigure), suggesting that MicroFisher can accurately predict the abundances of taxa with abundances <5%. Overall, the application of multiple hypervariable markers combined with weight-based algorithms allows MicroFisher to sensitively and accurately classify dominant and intermediate fungal taxa from metagenomic datasets.

Comparative performance of MicroFisher and existing tools for fungal classification using simulated metagenomic sequencing datasets

Evaluation of classification performance using light-size simulated sequencing datasets comprised exclusively of ITS and LSU rRNA sequences showed that MicroFisher can accurately predict the occurrence and abundance of fungal genera with abundances higher than 0.1% (Fig. 4). Thus, we filtered out fungal taxa that were classified as rare species (abundance <0.1%) in the classification results obtained from the simulated metagenomic datasets by MicroFisher. Classification performance was evaluated by sensitivity, precision, accuracy, and F1 score, and we found that MicroFisher exhibited significantly higher identification performance than the other tools at the genus level (Fig. 5A). In detail, MicroFisher had the highest prediction sensitivity (94%) and accuracy (79%), followed by Eukdetect (82% and 75%, respectively), Metaphlan (15% and 15%, respectively), Kraken2 (28% and 12%, respectively) and Kaiju (72% and 9%, respectively). All genera predicted by Metaphlan (100% precision) were correctly identified, while 90% and 83% of genera predicted by Eukdetect and MicroFisher, respectively, were correct. In contrast, only a small fraction of identified genera were correctly predicted using Kraken2 (17%) and Kaiju (9%). Meanwhile, the overall classification performance evaluated by F1 score showed that MicroFisher (88%) and Eukdetect (86%) provided highly reliable fungal community predictions, compared to Metaphlan (26%), Kraken2 (21%), and Kaiju (16%). Among the 25 expected fungal genera with relative abundances below 0.1%, MicroFisher correctly recovered 10 genera (10% accuracy), whereas Eukdetect, Kaiju, Kraken, and Metaphlan correctly predicted 7 (30% accuracy), 10 (0.6% accuracy), 1 (0.9% accuracy), and 1 (4% accuracy) genera, respectively. Furthermore, fungal classification using MicroFisher required fewer computation resources compared to the other tools, with lower memory usage (0.36 GBs

of the peak memory usage) and shorter running time (3.87 hours per 500 M reads using 8 CPU cores) (Fig. 5B).

Finally, we compared the accuracy of MicroFisher and Eukdetect in abundance estimation using simulated metagenomic datasets (Fig. 5C). The results exhibited that most genera predicted by MicroFisher and Eukdetect had accurate abundance estimates, although the abundances of some genera were underestimated by Eukdetect (arrowheads). However, MicroFisher (3.4%) had a lower r.m.s. error was than Eukdetect (3.98%). Similar patterns were also observed for Metaphlan (4.23%) and Kraken2 (10.82%), suggesting that MicroFisher provides better performance in fungal taxon abundance estimation than these tools (Fig. 5D). Despite yielding the lowest r.m.s. error (2.54%), Kaiju showed limited effectiveness in fungal genus prediction.

MicroFisher accurately profiles fungal communities in real environmental sequencing datasets

We performed fungal community analyses on both soil metagenomic and root eukaryotic metatranscriptomic datasets using MicroFisher and compared the results with four existing tools (Eukdetect, Kaiju, Kraken2, and Metaphlan3). A combination of four hypervariable marker databases (ITS1, ITS2, LSU D1, LSU D2) was used for MicroFisher workflow at MiniHit lengths of 80, 120, and 150 bp, respectively. DNA amplicon sequencing data generated from the same soil DNA samples were also included for comparison. Based on metabarcoding results, MicroFisher outperformed other classification tools in fungal community profiling from metagenomic (Fig. 6 and Supplementary Fig. S9) and metatranscriptomic (Supplementary Figs. S9, S11) datasets.

For alpha-diversity estimation of the metagenomic data, on average of 50 genera were detected using amplicon sequencing across the soil samples, and 18, 31, and 56 genera were detected from soil metagenomes using MicroFisher at MiniHit length of 150, 120, and 80 bp, respectively, while only 17 genera were predicted using Eukdetect (Fig. 6A). However, MetaPhlan3 exhibited limited capabiity for fungal detection from the metagenomic dataset, with no fungal taxa detected, and analyses using Kaiju and Kraken2 revealed no statistically significant differences (P-value > 0.05) in fungal alpha diversity across the examined samples. This observation underscores the methodological challenges in comprehensively characterizing fungal community composition using current taxonomic profiling approaches. When comparing genera detected from the metagenomic data using these selected classification tools with those identified by amplicon sequencing, Eukdetect and MicroFisher predicted more shared genera in comparison with Kraken2 and Kaiju (Fig. 6B). Also, MicroFisher recovered the highest number of shared genera (118 genera at MiniHit length of 120 bp) with the amplicon sequencing results (Fig. 6B). These results suggest that MicroFisher is more sensitive in fungal detection than the other examined classification tools. In addition, 177 and 185 genera were exclusively detected by amplicon sequencing and MicroFisher (MiniHit length of 120 bp), respectively. This discrepancy may reflect differences in sequencing technique, database application, and mapping algorithms utilized by metagenomics and amplicon sequencing.

Similar patterns in the abundances of dominant fungal taxa were observed when MicroFisher was applied with MiniHit lengths of 120 bp and 150 bp, whereas profiles generated witha MiniHit length of 80 bp showed distinct differences (Fig. 6C). This suggests that a MiniHit length of 120 bp or longer provides more accurate outcomes in profiling fungal community abundance. Interestingly, the abundances of identified dominant fungal taxa was different among amplicon sequencing, MicroFisher, and Eukdetect. For example, relatively higher abundances of *Sailozyma*, *Russula*, and *Inocybe* were identified using DNA amplicon sequencing in many samples, while MicroFisher more frequently detected higher abundances of *Russula*, *Suillus*, and *Amanita*. In contrast, relatively higher abundances of *Oidiodendron*, *Saitozyma*, and *Hyaloscypha* were consistently detected across samples using Eukdetect. Overall, dominant taxa were distributed more evenly among samples in the MicroFisher results compared to amplicon-sequencing and Eukdetect profiles. These patterns indicate that, among the computational tools evaluated, MicroFisher provided abundance estimates from metagenoic data that are more consistent with amplicon-based fungal community profiles. However, more studies are needed to determine whetheramplicon sequencing or MicroFisher yields more accurate estimates of fungal abundance from environmental DNA. By contrast, dominant fungal taxa identified by amplicon sequencing were largely not detected by Kaiju and Kraken2, which failed to differentiate fungal community composition across soil samples (Fig. 6C). These findings reveal significant limitations in current protein- and genome-based metagenomic classification approaches for fungal taxonomic profiling. This could be due to the lack of robust genome databases and the functional redundancy of protein databases used in these classification tools[8,9,51]. We also compared fungal beta diversity classified using MicroFisher and the other classification tools. Principal Coordinates Analysis (PCoA) and Bray–Curtis distance analysis indicated that fungal communities classified by MicroFisher at a MiniHit length of 120 bp and Eukdetect are more similar to the amplicon-sequencing results (P-value > 0.05), suggesting that MicroFisher may accurately predict fungal composition and abundance at the genus level (Fig. 6D, E) and at higher taxonomic levels (Supplementary Fig. S9D, E). This could also indicate that the marker gene prediction tool (Eukdetect) and ITS and LSU-based prediction tools (amplicon sequencing and MicroFisher) may generate more similar outcomes in fungal community prediction compared with genome- and protein-based prediction tools. Furthermore, our findings demonstrated that MicroFisher also effectively performed for fungal taxonomic classification across multiple taxonomic levels in the metatranscriptomic dataset (Supporting Information A.1), indicating the robustness of this computational approach for microbial community characterization.

Overall, MicroFisher detected 306 clades (including 86 genera from 12 classes) in soil metagenomes and 142 clades (including 35 genera from 10 classes) in root metatranscriptomes with abundances greater than 0.1%. MicroFisher also identified commonly occurring root-associated symbiotic fungal taxa, including *Russula*, *Amanita*, *Inocybe*, *Suillus*, *Microbotryum*, *Wilcoxina*, *Sebacina*, *Trichoderma*, and *Tomentella*, in the pine forest system (Fig. 5F and Supplementary Fig. S10B), indicating that our approach can successfully predict expected fungal taxa in certain small-scale studies. Together, these results show that MicroFisher outperforms other whole-genome-based methods for fungal prediction, as its hypervariable marker databases are not limited by databse size or dependence on specific targeted regions/genes. Our results pave the way for the application of multiple hypervariable markers to profile the composition and abundance of microbiomes from high-throughput sequencing datasets.

Discussion

With the development of high-throughput sequencing approaches (such as metagenomics and metatranscriptomics) that provides a better understanding of relationships between microbial community composition and ecological function, the impacts of fungal diversity and composition on ecosystem functioning and services have been well recognized. However, accurately inferring microbial composition and abundance remains a significant challenge, as genome-reliant microbial classification often results in a high number of false-positive and missed predictions, especially when profiling fungal communities from short-reads sequencing datasets.

To address this issue, we developed MicroFisher, which employs curated multiple hypervariable markers to classify fungal community composition and estimate taxon abundance from metagenomic and metatranscriptomic sequencing data. Applying hypervariable marker databases is likely to have important implications for improving the accuracy of fungal classification and strengthening the biological interpretation of ecological patterns. The hypervariable marker databases comprise highly species-discriminative sequences that enable more precise fungal taxonomic identificaiton from sequencing data. When coupled with the MicroFisher pipeline, which compiles mapping-based classification and weight-based integration algorithms, multiple hypervariable markers provide high sensitivity on both low- and high-complexity fungal communities. We also demonstrate that MicroFisher achieves substantially higher accuracy and

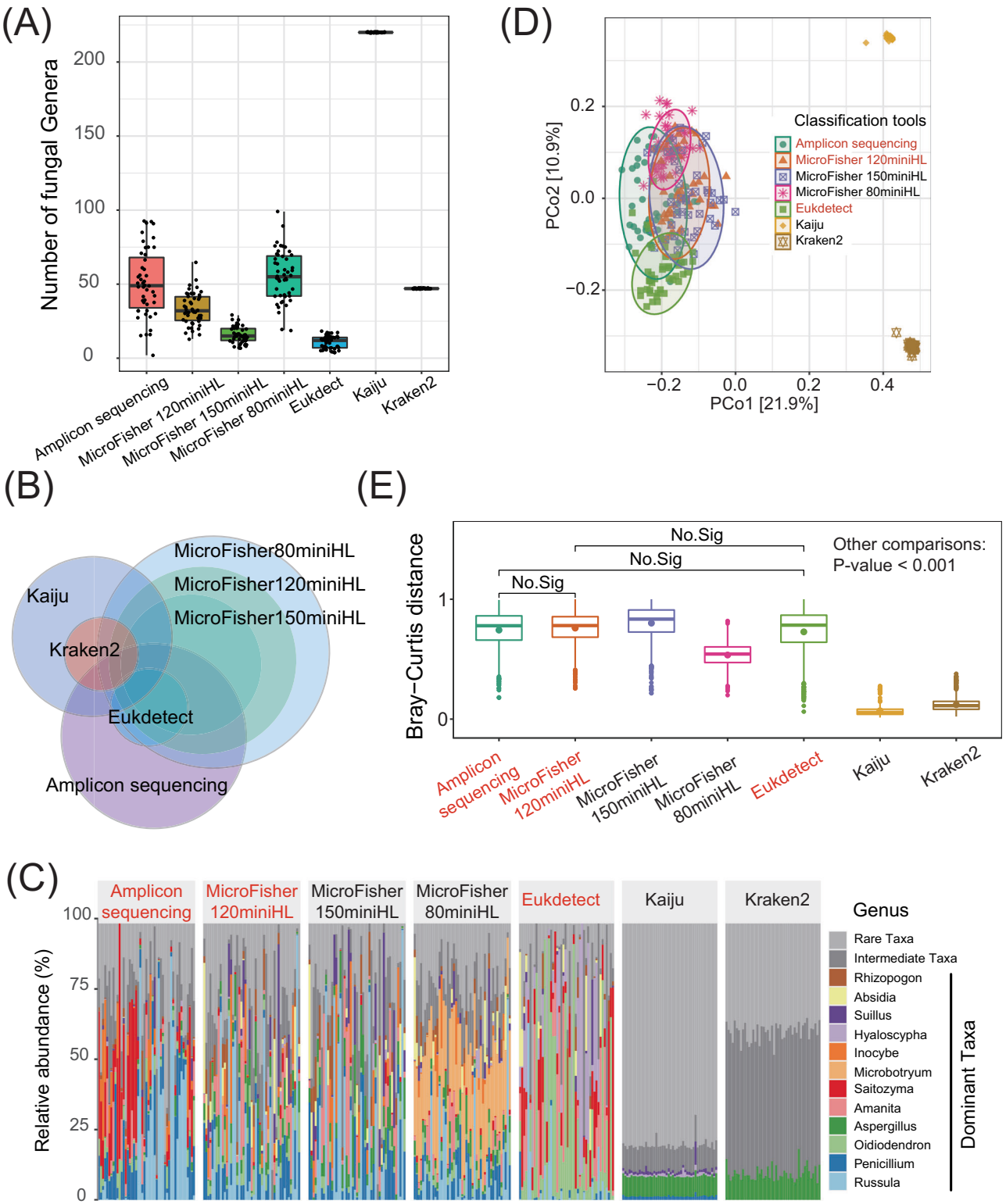

Fig. 6 | Analysis of soil metagenomic data using MicroFisher and other fungal classification tools. DNA amplicon sequencing data generated from the same soil samples were included in this comparison, indicating "amplicon sequencing". Fungal profiling from amplicon sequencing was performed using QIIME2, and the metagenomic dataset was analyzed using MicroFisher with minimum hit lengths of 80, 120, and 150 bp, as well as Eukdetect, Kaiju, Kraken2, and Metaphlan3 (no fungal taxa detected) with default parameters. Only genus-level taxonomy was performed for this comparison. The analyses include: **A** Richness of soil fungal communities; **B** Venn plot illustrating the overlap of fungal taxa classified by different methods; **C** Dominant taxa (relative abundance >1%) obtained from different pipelines. **D** Principal Coordinates Analysis (PCoA) and **E** Bray–Curtis dissimilarity analysis of soil fungal community composition inferred by different classification tools. MiniHL: minimum hit length; Dominant Taxa: taxa with the average relative abundance greater than 1%; Intermediate Taxa: taxa with the average relative abundance between 0.01% and 1%; Rare Taxa: taxa with the average relative abundance lower than 0.01%. MicroFisher classification and abundance evaluation were conducted using ITS + LSU hypervariable markers with default parameters (MiniHit length of 120 bp, weighted and filter mode).

precision in fungal classification at the genus and higher taxonomic levels in the simulated datasets. Moreover, compared to existing tools, MicroFisher provides more reliable profiles of fungal communities from real metagenomic and metatranscriptomic datasets, specifically for abundant taxa (relative abundance > 0.1%). For rare taxa with relative abundances below 0.1%, both MicroFisher and the other tested tools exhibited poor predictive performance. As to novel taxa that potentially present in metagenomic and metatranscriptomic sequencing datasets, lower minimum hit lengths may increase the likelihood of mismatches to known fungal species, which can be considered false positives. Although MicroFisher cannot directly identify such mismatches, several strategies can effectively reduce them, including requiring detection by at least two hypervariable databases, increasing the minimum hit length, and removing rare taxa with relative abundance below 0.1%. Accordingly, optimizing MicroFisher by applying a minimum hit length of 120 bp and filtering out rare taxa (relative abundance < 0.1%) improved performance, achieving a better balance between sensitivity and accuracy in fungal community profiling.

Additionally, MicroFisher demonstrates strong performance in the abundance estimation of fungal communities. Our results show that MicroFisher outperformed alternative whole-genome-based methods for fungal prediction (Fig. 5), as the number of sequences and coverage of taxonomy in hypervariable marker databases are unrestricted by limitations in database size and by reliance on specific genomic regions or genes. However, the ITS and LSU regions have inherent limitations when it comes to identifying taxa at the species level for most organisms[52]. Currently, MicroFisher only utilizes ITS and LSU databases for its pipeline; nevertheless, it is capable of incorporating additional marker sequences in the future. The implementation of this extension will depend on the availability of robust databases for other informative gene regions, such as elongation factor genes, beta tubulins, and other potential markers. Additionally, as metagenomic technologies continue to generate longer reads, MicroFisher may enable more reliable species-level taxonomic identification. At the current stage, even with short-read sequencing data (151 bp), we demonstrated that MicroFisher provides more accurate classificaiton results at the genus level and higher taxonomic ranks compared to other classification platforms, such as Eukdetect, Kaiju, Metaphlan, and Kraken2 (Figs. 5 and 6).

While MicroFisher offers clear advantages for profiling fungal communities, several methodological challenges remain and warrant careful consideration. First, generation of hypervariable markers relies on the INSDC database, where misidentified sequences and/or outdated taxonomic information may introduce false predictions. Although we implemented stringent quality-control and purification steps to refine taxonomy before generating our database, it is impossible to entirely eliminate inaccurately sequenced or incorrectly annotated database entries. Furthermore, fungal genomes naturally contain multiple rDNA copies that vary significantly among species, and this biological complexity can compromise accurate estimation of fungal abundance. Consequently, even though our abundance results may closely mirror metabarcoding data, these similarities cannot be interpreted unequivocally as accurate measures of species representation due to intrinsic variations in rDNA copy numbers. Amplicon sequencing is inherently limited by primer bias introduced during library construction, restricting its potential for methodological improvements. However, MicroFisher stands to benefit from advancements in short-read sequencing methods, including improvements in read accuracy, depth, and throughput on platforms, such as Illumina NovaSeq. In addition, MicroFisher's default settings require fungal taxa to be detected in at least two marker databases to enhance classification accuracy and reduce false-positive predictions. However, this strategy introduces a trade-off between precision and coverage, because fungal taxa represented in only a single marker database will not be reported. The default settings may result in missed detections of fungal taxa in environmental samples, particularly low-abundance taxa that can only be identified using a single hypervariable marker database. Conversely, using a single database to enhance coverage may compromise precision, potentially increasing erroneous predictions of fungal community composition from metagenomic and metatranscriptomic datasets. Despite these, MicroFisher offers a significant advantage over protein- or genome-based identification tools, because the latter are restricted to fungal species represented in existing genome or protein databases, which typically include only a small fraction of total fungal diversity. MicroFisher is capable of predicting a much broader range of fungal species, including those that are unculturable. However, the capability of MicroFisher to classify fungal composition in long-read sequencing datasets, such as those generated by PacBio and Oxford Nanopore sequencing, warrants further investigation to assess its applicability.

In conclusion, we developed MicroFisher, an accurate and sensitive tool for fungal classification from metagenomic and metatranscriptomic datasets. By integrating high-resolution hypervariable marker gene databases with weight-based integration algorithms, MicroFisher provides more reliable fungal community classification than existing state-of-the-art tools. MicroFisher also enables the detection of rare taxa that are difficult to identify using other available tools. This highlights the advancement of predicting fungal communities using multiple hypervariable markers. Our results demonstrate the effectiveness of MicroFisher for fungal classification and its potential for advancing the field of taxonomic profiling from metagenomes and metatranscriptomes. We use the name "MicroFisher" for this bioinformatic tool to reflect its broader potential, as the pipeline is designed to accommodate classification of other microorganisms in the future. While our current database focuses exclusively on fungi, we intend to expand the framework to include bacterial groups in future work.

Data availability

The MicroFisher Hypervariable Marker Databases have been deposited in Figshare, the DOI for these databases is https://doi.org/10.6084/m9.figshare.19679595.v3[53]. The light-size simulated sequencing datasets using fungal ITS and LSU rRNA Sequences are available in Figshare with the https://doi.org/10.6084/m9.figshare.20001542.v1[54]. The simulated metagenomic dataset are accessible in NCBI Sequence Read Archive [SRA] under BioProject PRJNA1321326. The pine forest soil metagenomic sequencing data and pine root metatranscriptomic sequencing data can be accessed through Joint Genome Institute (JGI) under Award https://doi.org/10.46936/10.25585/60001406. The soil fungal ITS amplicon sequencing datasets are available in NCBI Sequence Read Archive [SRA] under BioProject PRJNA1263842. The full tables of results for the fungal community classification and performance evaluation of simulated datasets using MicroFisher and other existing methods are available at Supplementary Data Sheet 2-7, which is available at https://doi.org/10.6084/m9.figshare.31400742.

Code availability

MicroFisher is written in Python and is available for download at GitHub repository with the URL: https://github.com/NFREC-Liao-Lab/MicroFisher[55], under the folder "microfisher". The manual for helping users to quickly start the fungal community classification using MicroFisher can be found with the URL: https://github.com/NFREC-Liao-Lab/MicroFisher[55]. The codes for generation of hypervariable databases, scripts for performance evaluation and statistical analysis are available at https://github.com/NFREC-Liao-Lab/MicroFisher[55], under the folder "database_generation" and "manuscript", respectively.

Received: 7 July 2025; Accepted: 25 February 2026;
Published online: 08 April 2026

References

1. Jansson, J. K. & Hofmockel, K. S. Soil microbiomes and climate change. *Nat. Rev. Microbiol.* **18**, 35–46 (2020).
2. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

3. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
4. Bork, P. et al. Tara Oceans. Tara Oceans studies plankton at planetary scale. *Science* **348**, 873 (2015).
5. Lu, J. & Salzberg, S. L. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* **8**, 124 (2020).
6. Antunes, L. P. et al. Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Sci. Rep.* **6.1**, 38915 (2016).
7. Avershina, E., Qureshi, A. I., Winther-Larsen, H. C. & Rounge, T. B. Challenges in capturing the mycobiome from shotgun metagenome data: lack of software and databases. *Microbiome* **13**, 66 (2025).
8. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
9. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).
10. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
11. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40**, e155 (2012).
12. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
13. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
14. Lind, A. L. & Pollard, K. S. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* **9**, 58 (2021).
15. Donovan, P. D., Gonzalez, G., Higgins, D. G., Butler, G. & Ito, K. Identification of fungi in shotgun metagenomics datasets. *PLoS One* **13**, e0192898 (2018).
16. Soverini, M. et al. HumanMycobiomeScan: a new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. *BMC Genomics* **20**, 496 (2019).
17. Marcelino, V. R., Holmes, E. C. & Sorrell, T. C. The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics* **21**, 184 (2020).
18. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol* **2**, 1533–1542 (2017).
19. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
20. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, 1125–1136 (2019).
21. Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K.-H. Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evol. Bioinform. Online* **4**, 193–201 (2008).
22. Tovo, A., Menzel, P., Krogh, A., Cosentino Lagomarsino, M. & Suweis, S. Taxonomic classification method for metagenomics based on core protein families with Core-Kaiju. *Nucleic Acids Res.* **48**, e93 (2020).
23. Tekpinar, A. D. & Kalmer, A. Utility of various molecular markers in fungal identification and phylogeny. *Nova Hedwig.* **109**, 187–224 (2019).
24. Wurzbacher, C. et al. Introducing ribosomal tandem repeat barcoding for fungi. *Mol. Ecol. Resour.* **19**, 118–127 (2019).
25. Meyer, W. et al. Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α) 1. *Genome* **62**, 160–169 (2019).
26. Gao, J. et al. A Rapid colorimetric assay for field-portable authentication of Amanita sepiacea using the beta-tubulin gene sequence. *J. Food Compost. Anal.* **105**, 104169 (2022).
27. Nilsson, R. H. et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* **47**, D259–D264 (2019).
28. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
29. Cole, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2014).
30. Deng, Z.-L., Münch, P. C., Mreches, R. & McHardy, A. C. Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.* **50**, e60 (2022).
31. Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
32. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
33. Eddy, S. R. Hidden markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365 (1996).
34. Bengtsson-Palme, J. et al. METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* **15**, 1403–1414 (2015).
35. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
36. Rivers, A. R., Weber, K. C., Gardner, T. G., Liu, S. & Armstrong, S. D. ITSxpress: software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis. *F1000Res* **7**, 1418 (2018).
37. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
38. Kans, J. Entrez direct: E-utilities on the UNIX command line. In *Entrez Programming Utilities Help [Internet]* (National Center for Biotechnology Information (US), 2021).
39. Shen, W. & Ren, H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics* **48**, 844–850 (2021).
40. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
41. Bembom, O. & Ivanek, R. seqLogo: Sequence logos for DNA sequence alignments, https://github.com/ivanek/seqLogo/blob/HEAD/vignettes/seqLogo.Rmd (2020).
42. Gorodkin, J., Heyer, L. J., Brunak, S. & Storomo, G. D. Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics* **13**, 583–586 (1997).
43. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
44. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
45. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
46. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J. & Bongcam-Rudloff, E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521–522 (2019).
47. Liao, H.-L. et al. Metatranscriptomic analysis of ectomycorrhizal roots reveals genes associated with Piloderma-Pinus symbiosis: improved methodologies for assessing gene expression in situ. *Environ. Microbiol.* **16**, 3730–3742 (2014).
48. Chen, K.-H., Longley, R., Bonito, G. & Liao, H.-L. A two-step PCR protocol enabling flexible primer choice and high sequencing yield for illumina MiSeq meta-barcoding. *Agronomy* **11**, 1274 (2021).

49. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
50. Gómez-Rubio, V. ggplot2-elegant graphics for data analysis. *J. Stat. Softw.* **77**, 1–3 (2017).
51. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
52. Kiss, L. Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for Fungi. *Proc. Natl. Acad. Sci.* **109**, E1811 (2012).
53. Wang, H. MicroFisher_DBs [Dataset]. *Figshare*, https://doi.org/10.6084/m9.figshare.19679595.v3 (2022).
54. Wang, H. Mock Fungal Community - metagenomic sequencing datasets [Dataset]. *Figshare*, https://doi.org/10.6084/m9.figshare.20001542.v1 (2022).
55. Wang, H. & Wu, S. HaihuaWang-hub/MicroFisher: MicroFisher.v1 [Software]. *Zenodo*, https://doi.org/10.5281/zenodo.18316826 (2026).

Acknowledgements

This work was supported by several fundings to H.-L.L., including DE-SC0020403, DE-SC0012704, NSF IOS-PBI (2029168), USDA-NIFA (2019–67013-29107, 2022-51106-37926), USDA-NIFA Research Capacity Fund McIntire-Stennis accession no. 1026825 and Hatch accession no. 7001162. This research was also supported by the USDA-NIFA (2024-67019-42344) to K.Z. and H-L.L. A portion of this research was performed on a project award JGI CSP (https://doi.org/10.46936/10.25585/60001406) under JGI's community science program, and used resources at the DOE Joint Genome Institute (https://ror.org/04xm1d337), which are DOE Office of Science User Facility. The facility is sponsored by the Biological and Environmental Research program and operated under Contract Nos. DE-AC02-05CH11231 (JGI).

Author contributions

H.L. Liao and R. Vilgalys coordinated the project; H. Wang constructed the hypervariable marker databases and developed the MicroFisher pipeline; H. Wang and S. Wu developed the MicroFisher package; H. Wang, K. Zhang and S. Wu performed the data analysis; K.H. Chen, R. Vilgalys, and H.L. Liao conducted the raw data generation for DNA amplicon, metagenomic and metatranscriptomic data; H. Wang, K. Zhang and H.L. Liao wrote the manuscript; all authors contribute to manuscript edit.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-026-09820-7.

Correspondence and requests for materials should be addressed to Steven H. Wu or Hui-Ling Liao.

Peer review information *Communications Biology* thanks Meganathan P. Ramakodi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Yu-Wei Wu and Tobias Goris. A peer review file is available.