Microbiome

RESEARCH Open Access

Challenges in capturing the mycobiome from shotgun metagenome data: lack of software and databases

Ekaterina Avershina[1], Arfa Irej Qureshi[2], Hanne C. Winther-Larsen[2] and Trine B. Rounge[1,2,3*]

Abstract

Background The mycobiome, representing the fungal component of microbial communities, is increasingly acknowledged as an integral part of the gut microbiome. However, research in this area remains relatively limited. The characterization of mycobiome taxa from metagenomic data is heavily reliant on the quality of the software and databases. In this study, we evaluated the feasibility of mycobiome profiling using existing bioinformatics tools on simulated fungal metagenomic data.

Results We identified seven tools claiming to perform taxonomic assignment of fungal shotgun metagenomic sequences. One of these was outdated and required substantial modifications of the code to be functional and was thus excluded. To evaluate the accuracy of identification and relative abundance of the remaining tools (Kraken2, MetaPhlAn4, EukDetect, FunOMIC, MiCoP, and HumanMycobiomeScan), we constructed 18 mock communities of varying species richness and abundance levels. The mock communities comprised up to 165 fungal species belonging to the phyla Ascomycota and Basidiomycota, commonly found in gut microbiomes. Of the tools, FunOMIC and HumanMycobiomeScan needed source code modifications to run. Notably, only one species, *Candida orthopsilosis*, was consistently identified by all tools across all communities where it was included. Increasing community richness improved precision of Kraken2 and the relative abundance accuracy of all tools on species, genus, and family levels. MetaPhlAn4 accurately identified all genera present in the communities and FunOMIC identified most species. The top three tools for overall accuracy in both identification and relative abundance estimation were EukDetect, MiCoP, and FunOMIC, respectively. Adding 90% and 99% bacterial background did not significantly impact these tools' performance. Among the whole genome reference tools (Kraken2, HMS, and MiCoP), MiCoP exhibited the highest accuracy when the same reference database was used.

Conclusion Our survey of mycobiome-specific software revealed a very limited selection of such tools and their poor robustness due to error-prone software, along with a significant lack of comprehensive databases enabling characterization of the mycobiome. None of the implemented tools fully agreed on the mock community profiles. FunOMIC recognized most of the species, but EukDetect and MiCoP provided predictions that were closest to the correct compositions. The bacterial background did not impact these tools' performance.

Keywords Mycobiome, Microbiome, Shotgun metagenome sequencing, Software, Databases, Fungi, Genome

*Correspondence:
Trine B. Rounge
t.b.rounge@farmasi.uio.no
Full list of author information is available at the end of the article

Introduction

Traditionally, research in microbiome studies has predominantly focused on bacteria and often overshadowed the contribution of Fungi. Although the mycobiome, the fungal fraction of the microbiome, only makes up less than 1% of the microorganisms present in the gut, it plays an important role in maintaining host homeostasis and influencing both physiological and pathophysiological processes [1]. *Ascomycota* is the most prevalent phylum in the gut of healthy individuals, with *Basidiomycota* following closely behind [2, 3]. Gut fungal dysbiosis, i.e., the altered composition of fungal communities [4] including a loss of symbionts, growth of opportunists, and disturbed fungal diversity [1], has been linked to various health conditions. These include irritable bowel syndrome (IBS) [5], Crohn's disease (CD) [6], autism spectrum disorder (ASD) [7], obesity [8], and colorectal cancer (CRC) with polyps exhibiting a reduced overall fungal diversity compared to adjacent tissue [9–11].

Fungi exhibit extensive diversity in both their physical characteristics and functions. Although they are found everywhere, the taxonomic classification of Fungi remains largely uncharted. Out of the estimated 2.2–3.8 million fungal species existing on Earth, only a small fraction (estimated 4%) has been formally identified [12], likely due to factors including phenotypic diversity, genetic variability, and the challenge of cultivating many species [13].

By sequencing the genomes in an untargeted manner, shotgun metagenomics can be utilized to study both the taxonomic makeup and potential functions of all microbiome constituents at once, including not only bacterial but also viral and fungal community [14]. However, in October 2024, NCBI PubMed search produced only 54 original research papers published from 2014 to 2024 for the search query "(metagenome analysis) AND (human gut mycobiome)" (Supplementary Table 1). Majority of these papers ($n=31$) utilized marker-based amplicon sequencing [15, 16]. When whole genome sequencing was used for fungal communities, data analysis often implied an alignment to a custom reference database (f.ex. using *bowtie2*) followed by relative abundance estimation as a separate step [17, 18]. This contrasts with automated bacteriome profiling tools that perform both identification and relative abundance estimation. Collectively, this indicates that there are significant challenges in characterizing the mycobiome, likely stemming from the scarcity of specialized mycobiome bioinformatics tools and databases.

Here, we have addressed the feasibility of automated mycobiome characterization based on shotgun metagenome data from microbial communities using existing bioinformatic tools. Mycobiome identification and classification using amplicon sequencing is not part of this work. We used simulated shotgun metagenome data from defined fungal communities of various richness and relative abundances. We have assessed Kraken2 [19] and MetaPhlAn4 [20]—two widely used taxonomy classification tools for microbiome data across domains; and MiCoP [21], FunOMIC [22], EukDetect [23], and HumanMycobiomeScan [24] designed specifically for eukaryotic/fungal metagenome data classification, whereas FindFungi [13] failed to be implemented.

Materials and methods

Reads simulation, mock community metagenome data generation, and its classification by Kraken2, MetaPhlAn4, MiCoP, and FunOMIC were wrapped into the Snakemake v8.11.1 workflow. The EukDetect and the HumanMycobiomeScan (MycobiomeScan v2.0; hereafter referred to as HMS) classifications were performed independently using the original pipeline/script. All other analyses were performed in Python v3.12 using Spyder IDE v5.5.4 unless stated otherwise. The work was executed on a high-performance computing (HPC) cluster comprising 256 CPUs and a memory capacity of 1 TB RAM.

Fungal assembly collection source

To compile a set of fungal genomes, we searched NCBI records for species belonging to Agaricostilbomycetes, Microbotryomycetes, Malasseziomycetes, Spiculogloeomycetes, Wallemiomycetes, Saccharomycetes, Xylonomycetes, and Classiculomycetes classes and *Cryptococcaceae* and *Sordariaceae* families within Ascomycota and Basidiomycota phyla which include unicellular Fungi likely to be part of microbiomes [25], using the NCBI *datasets* tool v16.17.3. We then filtered the list, retaining only records that were linked to a scientific publication. The list was expanded by including parasitic Fungi species belonging to other classes or phyla (Microsporidia, Mucoromycota), before downloading all NCBI RefSeq deposited genomes for these records using the same tool (database compilation was performed on May 21, 2024). We included only NCBI RefSeq curated assemblies due to their manual curation, and thus likely higher annotation and completeness levels [26]. The number of fungal core genes per genome was assessed using the *profile* module within the Universal Fungal Core Genes (UFCG) tool v1.0.5 and the related database [27]. The core gene profiles were then used to infer phylogeny using the *tree* module within the same software (based on the MAFFT alignment and JTT model with IQ-TREE).

Mock communities generation

To get a range of community richness and abundance, we constructed 3 mock communities with 10 genomes, 3 communities with 50 genomes, 2 communities with 100 genomes, and 1 community with 165 genomes. To avoid creating a community that comprised only rare species, five commonly and widely known fungal species (*Saccharomyces cerevisiae, Debaryomyces hansenii, Malassezia restricta, Candida albicans, Kluyveromyces lactis*) were included in all generated communities. All other species were randomly selected from the list of species with available NCBI RefSeq curated assemblies. For each genome, we generated 1 M paired-end Illumina HiSeq 2500 reads using the ART read simulator v2.5.8 [28] (read length 150 bp; mean fragment size 300 ± 50 bp; quality range 30–40), which were then used for the fungal mock community metagenomic dataset construction.

We further used two approaches for the relative abundance profile generation—equal reads (ER) and equal coverage (EC). In the ER communities, each species was represented by 100,000 reads per genome regardless of the genome size. For the EC communities, each species was represented by n reads, where $n = \left(genome\ size \times 2\right)/read\ length$. Therefore, the complete simulated dataset comprised 18 mock communities of 4 richness levels and 2 relative abundance modes.

Additionally, we created mock communities with 90% and 99% bacterial backgrounds. We randomly selected 90 bacterial genomes from the HumGut database [29] and generated 100,000 reads/genome following the same approach as for the fungal genomes. All bacterial reads were combined into one file and added to the three 10-species fungal mock communities. For the 90% bacterial background, 100,000 reads/fungal genome were used, whereas for the 99% bacterial background community, each fungal genome was downsampled to 9091 reads/genome. We selected EukDetect, FunOMIC, and MiCoP—three tools with the highest accuracy of taxonomy detection and abundance estimation, for assessment of the impact of bacterial background on their performance.

Fungal classification and relative abundance estimation

The following tools and related databases were used: Kraken2 v2.1.3: the PlusPF database (January 12, 2024; https://benlangmead.github.io/aws-indexes/k2) with NCBI RefSeq records of Fungi and protozoa added to the Standard Kraken database; MetaPhlAn v 4.0.6: CHOCOPhlAnSGB_202307 (July 2023) comprising clade-specific markers from microbial genomes; EukDetect: EukDetect database v2 (April 23, 2022) comprising markers from eukaryotic organisms; HMS: Human_associated_fungi (April 8, 2021), FunOMIC: FunOMIC-T.v1 (July 2022) and MiCoP: MiCoP-fungi (based on NCBI RefSeq; February 2018). To ensure common taxonomy delineation, the databases were normalized to the same nomenclature using NCBI taxonomy identifier (December 17, 2024).

To test the effect of the database, we have downloaded fungal NCBI RefSeq records (using *kraken2-build –download library fungi*) and constructed Kraken2, HMS, and MiCoP databases, since these were the tools that use whole genome mapping and are not based on marker genes.

The accuracy of fungal species prediction by tested tools was assessed by precision (i.e., what proportion of the identified fungal species was included into the mock community), recall (i.e., what proportion of fungal species included into the community was identified), and the F1 score [_F1score = 2*(Precision*Recall)/(Precision + Recall)_] on species, genus, and family taxonomic levels. With regards to the relative abundance predictions, Kraken2, HMS, and MiCoP estimates were compared to the read-based relative abundance, whereas marker-based EukDetect, FunOMIC, and MetaPhlAn4 estimates were compared to the coverage-based relative abundance. Root mean squared error was used for the assessment of the relative abundance estimation accuracy. Significance was tested using the Kruskal–Wallis test, p-values were corrected for multiple testing using Benjamini–Hochberg FDR correction, and $p \leq 0.05$ were considered significant.

For Kraken2, the predicted relative abundance was estimated using Bracken v2.9; for EukDetect and FunOMIC, based on all classified reads; and for MetaPhlAn4, MiCoP, and HMS, no estimation of relative abundance was performed since it is reported by default.

Results

Scarcity of mycobiome-specific tools and fungal data in databases

To test the different tools published for mycobiome metagenomic analysis, we first searched for available fungal genomes. There were over 9000 records in the NCBI taxonomy database for species belonging to the query classes and families. Of these records, 1622 were linked to a scientific publication, and 170 of them had a link to NCBI RefSeq. Most species with NCBI RefSeq genomes available belonged to Saccharomycetes class (Fig. 1a). The genome size ranged between 2.5 Mbp (*Encephalitozoon cuniculi*) and 75.3 Mbp (*Blastomyces gilchristii*), median = 18.4 Mbp (IQR = 18.7 Mbp; Fig. 1b). Generally, genomes were complete, with a median proportion of universal fungal core genes = 91.8% (IQR = 6.5%; Fig. 1c), although

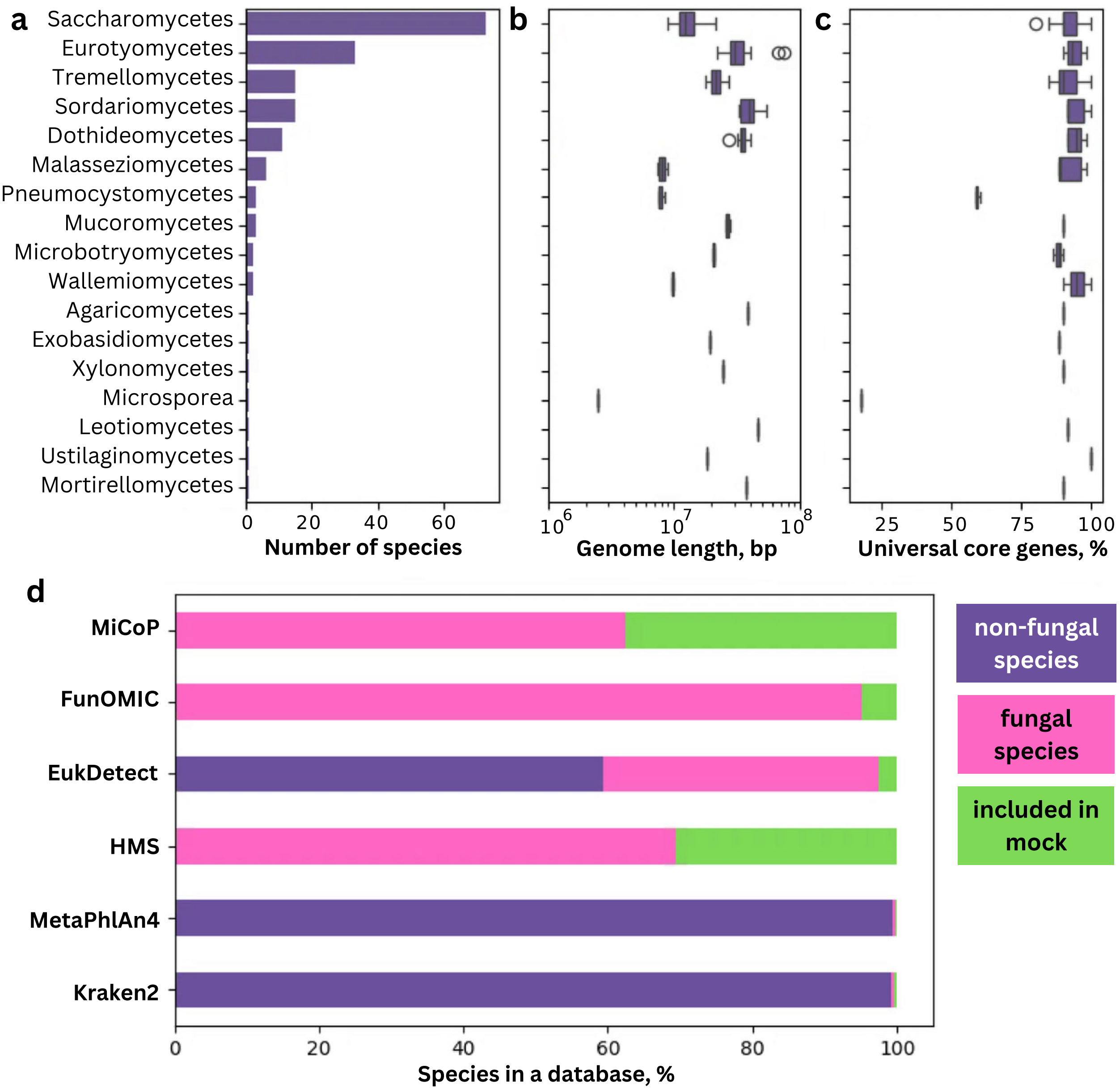

Fig. 1 Description of fungal species and databases used in the analysis. **a**–**c** Fungal species with NCBI RefSeq genomes included into the mock communities. **a** Number of species belonging to each class. **b** Genome size, whiskers indicate interquartile range (IQR). **c** Genome completeness by proportion of universal fungal core genes. **d** Proportion of fungal species in databases used by MiCoP, FunOMIC, EukDetect, HumanMycobiomeScan v2.0 (HMS), MethaPhlAn4, and Kraken2. Purple: all non-fungal species; pink: fungal species not included into the mock community; green: fungal species included into the mock community. For the MetaPhlAn4 database, only Ascomycota and Basidiomycota are included into the fungal group

a correlation between the size of the genome and the number of detected core genes was observed (Pearson $R^2 = 0.07$, $p = 0.0007$). The full list of fungal species, their taxonomy, NCBI RefSeq accession number, genome size, number of detected core genes, number of contigs, as well as species inclusion into each mock community is provided in Supplementary Table 2.

Then we searched for tools for taxonomic classification of the mycobiome [30]. Three approaches were utilized: (a) search for the published mycobiome metagenome data analysis tools; (b) summarizing tools that were used for shotgun metagenome data analysis in mycobiome research from 2014 (earliest paper with fungal shotgun metagenome data analysis referenced in NCBI PubMed was published in 2017; Supplementary Table 1); and (c) search for publicly available mycobiome tools on GitHub. Seven tools— Kraken2 [19] (number of research articles utilizing the tool for the mycobiome data analysis, $n=8$), MetaPhlAn4 [20] ($n=2$), FindFungi [13] ($n=0$), HMS [24] ($n=3$), EukDetect [23] ($n=0$), MiCoP [21] ($n=1$), and FunOMIC [22] ($n=1$)—were identified for this purpose. FindFungi, published in 2018, showed significant stability issues with numerous error messages, the setup process involved downloading multiple software packages and databases, modifying permissions and scripts, and substantial storage space requirements (at least 512 GB), rendering the implementation process cumbersome and not user-friendly (Supplementary Table 3). Implementation of FunOMIC, published in 2022, also proved challenging. The tool needed several script modifications, and its associated bacterial index database was likely corrupted (Supplementary Text 1). We have implemented another bacterial index database based on the HumGut database of human gut prokaryotes [29] in order to run the tool. Since the study focuses on taxonomic assignment of Fungi, we did not attempt to adapt the functional part of the FunOMIC pipeline which was also not running without modifications. Similarly, the HMS required changes in the main *MScan.sh* and related *bmtagger.sh* scripts; several dependencies were not listed in the manual (f.ex. *git, java* and *stringr, ggplot2 R* libraries), and *Normalising_table.txt* file was also missing from the latest release and had to be downloaded from the v1 release (Supplementary Text 2). EukDetect and MiCoP were the only Fungi-targeting tools that were successfully installed and implemented with no issues.

The fungal domain was poorly represented in the microbial databases (Fig. 1d). Only 1% (2146/217 432) of the Kraken2 PlusPF database records (v January 12, 2024) were tagged as "Fungi." These records comprised 98 fungal species, and only 49 of them matched the mock communities' species. All species listed in the database had primary names in the NCBI Taxonomy database (Supplementary Table 4), and when the preferred name was provided, the species name was updated accordingly. Similarly, only 1.3% of the species-level genome bins (SGBs) in the MetaPhlAn4 database (v July, 2023; 489/36 822 SGBs) were tagged as "Eukaryota"; 229 of them belonged to Ascomycota or Basidiomycota (although many SGBs had multiple taxonomic assignments on a species level), and only 50 mock communities' species were represented in the database. On the other hand, 65.8% of the EukDetect database (v April 23, 2022; 343 618/521 824 records) were tagged as "Fungi" and comprised 2006 fungal species, 128 of which were represented in the mock communities. However, 239 of these species names could not be identified by the NCBI Taxonomy Identifier (Supplementary Table 5), and many of these entries were labelled as "genus sp" or "genus SPcollapse." The Human_Associated_Fungi database (v April 8, 2021; 559,697 records) from HMS comprised 137 fungal species (100% Fungi), 42 of which were present in the mock communities. The MiCoP database (v February, 2018) comprised 72,668 records from 247 species, 99 of which were present in the mock communities. The FunOMIC-associated FunOMIC-T database (v1, July 2022; 1.6 M marker genes from 3060 fungal species) had the highest coverage of mock communities with 154 species represented. Pairwise species overlap between the databases is provided in Supplementary Table 5. Additionally, we used a Kraken2 RefSeq fungal database for the comparison of the whole genome based taxonomy tools (Kraken2, HMS and MiCoP). The database comprised 2259 records from 126 fungal species.

EukDetect exhibits highest detection and abundance estimation accuracy

In the ER communities (0.1 M reads/genome), genome coverage ranged between 0.2×and 6.0x (median = 0.8x, IQR = 1.2x). In the EC communities (2×genome coverage), the number of reads ranged between 0.03 M and 1 M reads per genome (median = 0.25, IQR = 0.25 M reads per genome). Even genome coverage of each genome did not significantly impact classification accuracy with Kraken2, MetaPhlAn4, MiCoP, or EukDetect, at any tested taxonomy level and community richness (FDR corrected p-value = 1 for all; Supplementary Table 6). However, seven species with genome size > 10 Mbp were detected in the EC, but not ER, communities by EukDetect (Supplementary Table 7). FunOMIC exhibited higher accuracy of genus level classification with the ER mode (F1 score (IQR) = 82.6% (3.6%) vs 74.0% (4.2%) for ER and EC modes respectively, FDR corrected p-value = 0.03). The HMS successfully analyzed all ER communities but failed during the 100 and 165 species EC communities runs when aggregating results with *mergescript.R* script, even with considerable memory available. Same as with other tools, there was no difference between ER and EC for 10 and 50 species communities using HMS.

The results reported below are provided for ER and EC communities combined. Overall, EukDetect exhibited significantly higher identification accuracy than other tools both at species, genus, and family levels (Fig. 2a). HMS had the steepest accuracy increase between all taxonomy levels (Fig. 2a). Kraken2 exhibited positive Pearson correlation between the community richness and detection accuracy; MetaPhlAn4, MiCoP, and EukDetect detection accuracy had a negative correlation to the community richness, whereas HMS and FunOMIC accuracy did not significantly correlate to the community richness

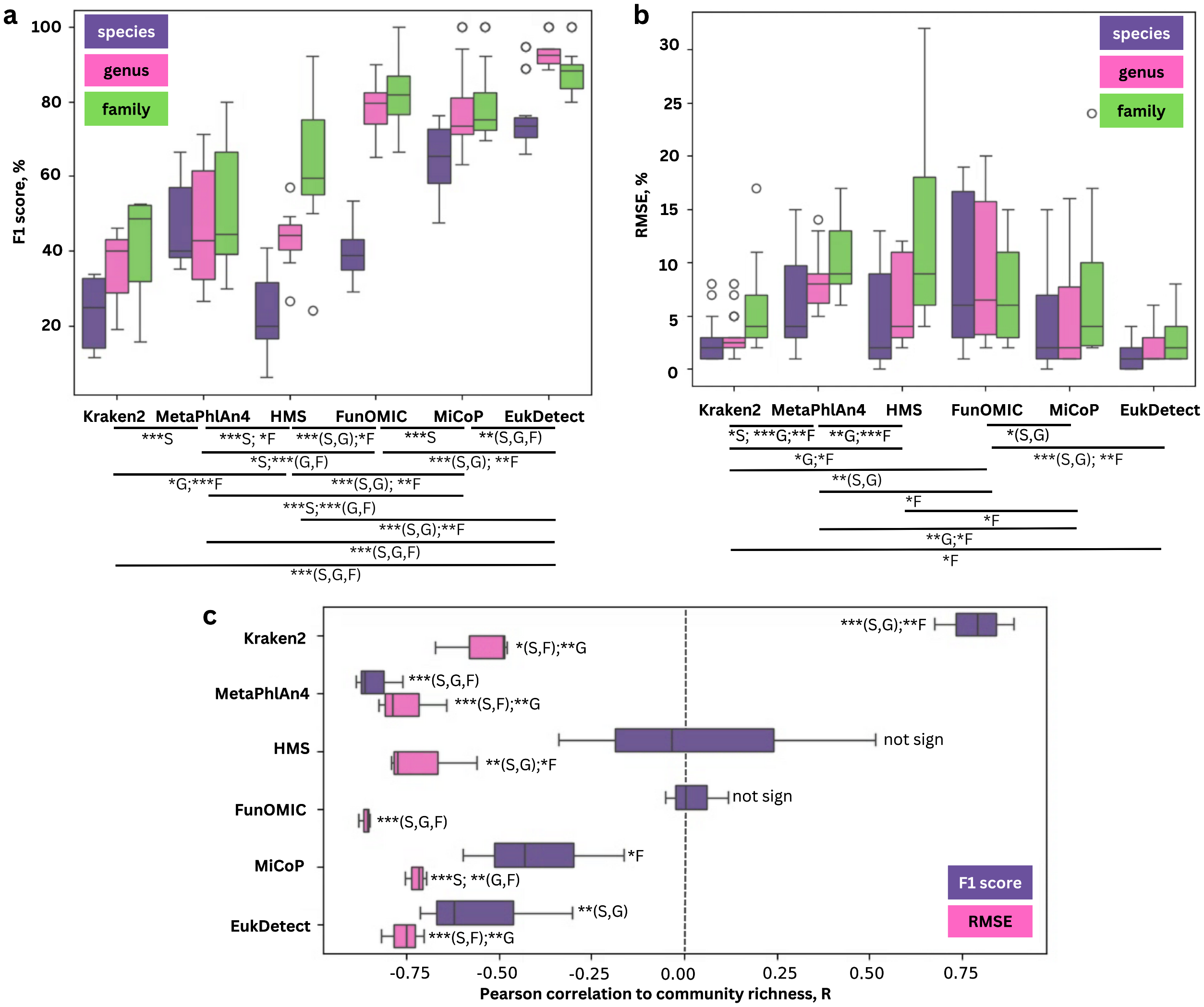

Fig. 2 Accuracy of Fungi characterization. **a** Identification accuracy at the species (purple), genus (green), and family (dark red) taxonomy levels. **b** Root mean square error of relative abundance prediction at the species (purple), genus (green), and family (dark red) taxonomy levels. **c** Pearson correlation between the identification accuracy (purple) or relative abundance prediction error (pink) and the community richness. Significance at species (S), genus (G), and family (F) levels is provided as *0.01 < FDRp ≤ 0.05, **0.001 < FDRp ≤ 0.01, ***FDRp ≤ 0.001

(Fig. 2c). At the genus level, all genera predicted by MetaPhlAn4 were correctly identified, whereas EukDetect precision at the genus level ranged between 90 and 100%, MiCoP between 80 and 90%; FunOMIC at most reached 80% precision; and Kraken2 and HMS remained below 80% precision (Supplementary Table 8). At the family level, although EukDetect had the highest accuracy, MetaPhlAn4 and HMS exhibited precision similar to that of EukDetect (median > 80%; Supplementary Table 8), but significantly lower recall, whereas FunOMIC had the highest recall with the lowest precision (Supplementary Table 8).

Unlike identification accuracy which tended to increase on higher taxonomy levels, relative abundance estimates on a family level were the least accurate, especially in the case of HMS (Fig. 2b). HMS, FunOMIC, and MetPhlAn4 exhibited the highest median abundance error, whereas EukDetect and Kraken2 had similarly low RMSE with a median = 5% (Fig. 2b). In all cases, RMSE negatively correlated to community richness (Fig. 2c).

FunOMIC covers highest phylogenetic diversity and has highest recovery of multiple closely related species

Kraken2, MetaPhlAn4, and EukDetect use databases comprising both fungal and non-fungal records. Kraken2 classified 0.01–2.5% reads as bacterial, 0–0.02% as viral, and 0.01–5.4% as human in all mock communities.

MetaPhlAn4 and EukDetect classifications, on the other hand, belonged solely to the Fungi domain. Out of 170 species used in mock communities, MetaPhlAn4, Kraken2, and HMS recognized 41, 46, and 42 species respectively, whereas MiCoP, EukDetect, and FunOMIC successfully identified 98, 111, and 132. Saccharomycetes and Eurotiomycetes were the most recognized fungal classes; *Candida orthopsilosis* was the only species recognized in all 3 mock communities, and it was included by all tools (Fig. 3a). For Eurotiomycetes, MetaPhlAn4 and Kraken2 captured mostly *Aspergillus* and *Penicillium* genera, whereas EukDetect, HMS, MiCoP, and FunOMIC recovered species across the class. Kraken2, MiCoP, and FunOMIC were the most robust with regard to the detection of species in communities of various richness—all species were either detected in all communities or missed in all communities, whereas other tools sporadically detected a given species in one community but missed it in another.

Cumulatively, there were 86 occurrences when a genus was represented by more than one species in a mock community (Supplementary Table 9). FunOMIC, MiCoP, and EukDetect exhibited the highest identification rate of at least one species within a genus (Fig. 3b). FunOMIC and MiCoP identified all species within the genus more frequently than other tools, and FunOMIC was the only tool that identified all genera although sometimes with a wrong species assignment. With Kraken2, all missing genera were missed due to their absence in the PlusPF database, whereas other tools also failed to identify genera present in the respective databases. MetaPhlAn4, for example, additionally failed to identify *Ogataeae*, MiCoP missed *Fusarium*, HMS missed *Blastomyces* and *Trichophyton*, and EukDetect failed to identify *Kazachstania* (now reclassified to *Huiozyma, Arxiozyma,* and *Maudiozyma*). Equal genome coverage enabled more frequent identification of all representative species within a genus for EukDetect (10 cases vs 6 in the ER communities) but not the other tools.

With regard to five species included in all mock communities (*D. hansenii*, *C. albicans*, *S. cerevisiae*, *K. lactis*, *M. restricta*), FunOMIC and Kraken2 exhibited the best performance having identified all species in all communities. *Kluyveromyces lactis* remained unclassified at species or genus level by MetaPhlAn4 (*Kluyveromyces* genus was absent in the database) and was misclassified to other *Kluyveromyces* by HMS. *Malassezia restricta* was misidentified to other species by MiCoP due to its absence in the database. *S. cerevisiae* was missed by HMS due to the absence of *Saccharomyces* genus from the associated database. There were also instances when *C. albicans, S. cerevisiae, or D. hansenii* were missed despite having been present in the reference database. All these misidentifications occurred in the > 10 species communities, where there were other representatives within the genus.

MiCoP performs best among whole genome-based tools when used with the same database

Using the same RefSeq fungal database (*kraken2-build –download library fungi*; retrieved October 10, 2024) with whole genome-based tools (Kraken2, MiCoP, and HMS), MiCoP exhibited highest accuracy on species level, whereas on the genus and family level, its accuracy was similar to that of HMS (Supplementary Fig. 1). Kraken2, on the other hand, exhibited significantly lower accuracy compared to the other tools (Supplementary Fig. 1).

Bacterial background did not impact identification accuracy

Adding bacteria background (90% and 99%) did not significantly impact the accuracy of any tool. However, MiCoP exhibited a somewhat reduced F1 score in the presence of 99% bacterial reads, and FunOMIC F1 score and RMSE were slightly improved with the lower representation of fungal genomes (Supplementary Fig. 2).

Discussion

Our major findings highlighted a shortage of user-friendly and bug-free software for fungal identification using shotgun metagenomic data. The most accurate tool for taxonomic identification and abundance estimation was EukDetect, although it missed 59 out of 170 species. FunOMIC, on the other hand, missed only 38 species but exhibited a high false positive detection rate. The challenge with the taxonomic identification of fungal species is largely a result of insufficient representation of their genomes in databases.

A challenge in analyzing metagenome sequencing data is the lack of benchmarking and maintenance of the bioinformatic tools [31]. This is even more problematic in

(See figure on next page.)

Fig. 3 **a** Universal core genes phylogenetic tree of the fungal dataset. Node color represents fungal classes; classes with only one representative are combined in "other." Outer circles depict detection of the species by Kraken2 (purple), MetaPhlAn4 (pink), EukDetect (green), HMS (beige), FunOMIC (blue), and MiCoP (grey) in the ER mock communities. Size of squares represent detection of species in all mock communities (largest), in at least one community (middle size), and in none of the communities (smallest). **b** Number of cases when all species (purple), at least one species (pink), only genus (green), or no genus (beige) were identified given that the genus was represented by several species in a community

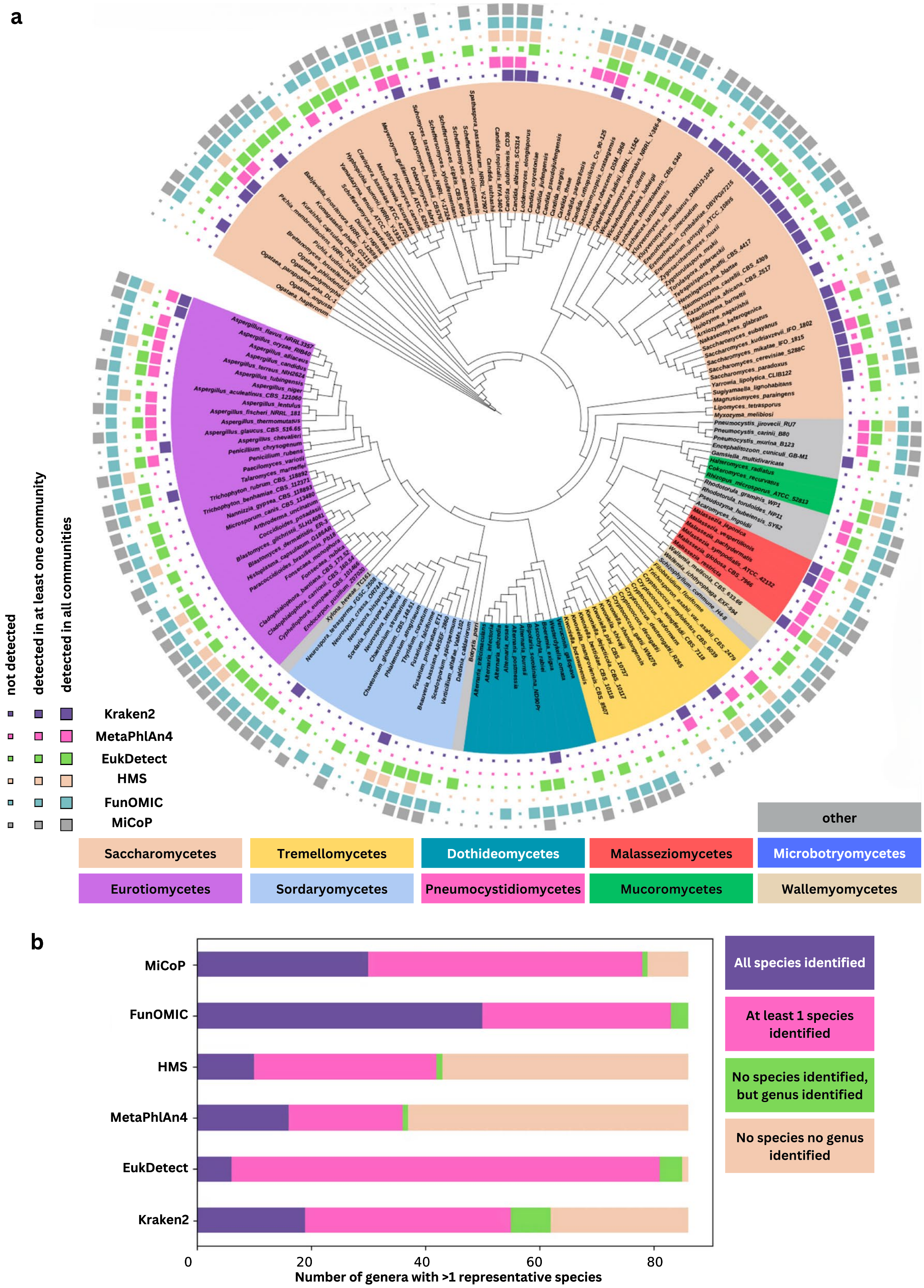

Fig. 3 (See legend on previous page.)

developing fields, such as the mycobiome, where the gold standards have not been established yet. Three out of seven tools tested here could not be implemented without substantial modification of their codes. FindFungi failed in our setup, but it was recently tested by Usyk et al., with the conclusion that it produced poor fungal reads detection [32]. FunOMIC was recently employed to demonstrate higher species profiling accuracy of shotgun sequencing compared to ITS by Xie et al. although they used an updated version that is not publicly available [33]. We managed to implement the previous version of FunOMIC with several modifications to the code and by implementing another index database for bacterial reads filtering. The HMS also required modifications. These issues make the use of Fungi-specific tools cumbersome, if not impossible, for researchers with no or little bioinformatic background, and their algorithms/code require optimization and benchmarking. Metagenomic sequencing data allows not only taxonomic but also functional characterization, commonly employed for bacteria. The FunOMIC is the only tool claiming to perform a functional annotation analysis; however, it failed implementation and we did not attempt to make it operational. Therefore, it is crucial to develop robust and accurate mycobiome metagenome data analysis tools that include functional annotation.

Kraken and MetaPhlAn are first and foremost focused on characterizing bacteria and were later expanded to capture other kingdoms [34, 35]. Both tools are well-established, easily implemented, and able to detect Fungi species deposited in databases. Even though both tools exhibited lower accuracy than Fungi targeting tools, Kraken2 more frequently recovered all representative species within a genus, whereas MetaPhlAn4 made zero false positive genus predictions. However, Kraken2 PlusPF and MetaPhlAn4 CHOCOPhlAnSGB databases mostly comprise information on bacterial genomes, and only ~ 1% of records belonged to Fungi. This general scarcity of representation in databases mirrors the lack of high-quality whole fungal genomes. In this work, only 10% of Fungi records that we searched through, had whole genome data deposited in the NCBI RefSeq database, and most of them belonged to Saccharomycetes and Eurotiomycetes classes. Whole genome sequencing enables strain-level resolution, which is successfully exploited in bacterial studies [20]. With Fungi, however, due to the scarcity of databases, most of the species remained identified only as belonging to a genus or family. To accomplish better resolution, it is imperative to sequence a larger diversity of fungal genomes. Recently, Yan et al. have developed the cultivated gut fungi catalog which encompasses 760 fungal genomes belonging to 206 species [36]. All Fungi were isolated from stool samples of healthy Chinese individuals and such efforts should be expanded to other geographical regions. With sequencing, fungal taxonomy delineation is rapidly evolving as more genomes become available. It is thus important to ensure correct species names, especially when utilizing databases that were constructed before recent updates.

We anticipated that equal genome coverage might enhance identification accuracy, particularly with marker-based tools. However, aside from several cases of species identification with EukDetect (not significantly), no overall improvement in Fungi detection was observed and FunOMIC's genus accuracy was reduced. Moreover, HMS failed to run with equal coverage mock communities of > 100 species due to memory limitations in underlying R scripts.

All tools tested here exhibited higher accuracy of genus or family identification compared to species, albeit with higher relative abundance error. FunOMIC exhibited the highest recall among the tested tools and had the most comprehensive fungal database compared to other tools, but its precision was low resulting in reduced detection accuracy. EukDetect, on the other hand, had the highest accuracy of taxonomy identification and relative abundance estimation although fewer mock community species were represented in its associated database. MetaPhlAn4 ranked next to last among tested tools based on recall, but all genera that were predicted by MetaPhlAn4 were indeed present in the communities. Unexpectedly, Kraken2 exhibited a strong correlation between the accuracy of taxa identification and the underlying community richness—the larger the community was, the less false positive predictions it made. Relative abundance estimates depended on the community richness, with higher RMSE in smaller communities, and higher RMSE for genus and family vs species, for all tested tools.

Interestingly, only *C. orthopsilosis* was persistently recovered with the tested tools. When a genus had several species representatives, only one species per genus tended to be recovered although both were present in the reference database. True microbiome communities are commonly complex with several species per genus, and this may thus lead not only to misidentifications but also to drawing potentially wrong associations between a species and a host.

The difference in tool performance likely reflects differences in the associated databases. To compare tool performance with regard to the algorithm itself, we implemented the same fungal RefSeq database to Kraken2, HMS, and MiCoP. We selected these tools because they perform whole genome comparisons and are thus more database flexible than marker gene-based EukDetect, FunOMIC, and MetaPhlAn4. In this setting, MiCoP performed best, whereas Kraken2 exhibited the lowest accuracy of fungal detection. True microbiome data often comprise down to < 1% of the fungal data, whereas in this work, we tested a best-case scenario of only fungal communities. Therefore, we have additionally constructed fungal communities with 90% and 99% of the bacterial background and assessed EukDetect, FunOMIC, and MiCoP—three tools with highest accuracy. None of the tools were significantly affected by the background, even with 9091 reads per fungal genome representation.

To enhance fungal characterization, efforts must be focused on both advancing database development and refining algorithms. Relying solely on algorithms is insufficient to encapsulate fungal diversity without comprehensive knowledge of fungal genomes. Furthermore, sensitive algorithms are crucial for capturing minuscule fungal part of complex metagenomic data and for predicting their functional potential.

The strength of this software survey was the comprehensive analysis of the tools with mock communities of various richness and abundance based on high-quality genomes. The persistent inclusion of five species in all communities, random occurrence of several species per genus in a community, and species nomenclature normalization enabled the assessment of identification reproducibility in various contexts. We constructed most mock communities based solely on fungal genomes to have controlled experiments and tested bacterial background contribution with the most promising tools. This does not reveal how viral and human background data can interfere with the fungal profiling.

Conclusion

Overall, we have detected a limited selection of software and approaches for mycobiome shotgun metagenome data analysis. Out of seven tools we identified, one failed an implementation in our hands, and two required alterations to their source code to be implemented. No established strategies for functional annotation of mycobiome data are available to date. Nevertheless, FunOMIC had the most extensive fungal database, but EukDetect exhibited the highest performance score regarding taxonomic profiling. To unlock the broad capabilities of shotgun metagenomics in mycobiome research, it is of utmost importance to develop Fungi-tailored bioinformatics tools, improve the reference databases and conduct meticulous benchmarking studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-025-02048-3.

Additional file 1: Supplementary Table 1. PubMed search results using "metagenomics & human gut mycobiome" keywords, publications 2014–2024; excluding reviews and systematic reviews. Date of search - 07.10.2024. Supplementary Table 2. NCBI RefSeq genomes used in the study and their metadata. Supplementary Table 3. List of mycobiome-tailored and microbiome tools tested. Supplementary Table 4. Correspondence of species names in databases to NCBI Taxonomy (December 2024). Supplementary Table 5. Pairwise intersect between the species deposited in each of the databases. Species names were normalized to the NCBI Taxonomy in December 2024. Supplementary Table 6. Identification accuracy between EC and ER modes. Supplementary Table 7. Species detected in equal coverage but not in equal reads communities. Supplementary Table 8. Precision and recall on species, genus and family levels. Supplementary Table 9. List of all cases when a genus was represented by several species and their identification by all tools in equal reads community type. Supplementary Table 10. Detection of mock community members by mycobiome tools and their inclusion into related databases.

Additional file 2: Supplementary Figure 1. Identification accuracy (A) and RMSE (B) by Kraken2, HMS and MiCoP using the same fungal RefSeq reference database on a species (purple), genus (pink) and family (green) levels. Supplementary Figure 2. Identification accuracy (A, B) and RMSE (C, D) by FunOMIC (blue), MiCoP (grey) and EukDetect (green) on a species (square), genus (x) and family (circle) levels in mock communities with 10 fungal species and bacterial background at 90 % (A, C) and 99 % (B, D) level (y-axis) vs no bacterial background (x-axis). Supplementary Text 1. Script and database modifications required for FunOMIC. Supplementary Text 2. Script modifications and unlisted dependencies required for MycobiomeScan 2.0 (HMS).

Acknowledgements

We would also like to thank the CRCbiome group, including Einar Birkeland, Maja Sigerseth Jacobsen, Paula Istvan, Paula Berstad, Ane S Kværner, and Milan Stosic for their valuable feedback and discussions. This document has benefited from the assistance of ChatGPT, an AI language model developed by OpenAI. The AI was used to help enhance the clarity, coherence, and grammar of the text, as well as to provide support with scripting and coding tasks.

Authors' contributions

EA, TBR and HCWL contributed to the study conception and design. AIQ and EA participated in the data collection; EA and AIQ performed the data analysis and produced the first draft. All authors interpreted the results; have read and approved the final manuscript.

Funding

This project would not have been possible without funding from The South Eastern Norway Regional Health Authority projects 2022067 (TBR).

Data availability

All scripts, selected genomes, seeds to recreate the mock communities, and related files are available at https://github.com/Rounge-lab/mock_mycobiome.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details
[1]Department of Tumor Biology, Oslo University Hospital, Oslo, Norway. [2]Department of Pharmacy, Section for Pharmacology and Pharmaceutical Biosciences, University of Oslo, Oslo, Norway. [3]Department for Research, Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway.

Received: 28 August 2024 Accepted: 28 January 2025
Published online: 07 March 2025

References

1. Zhang F, Aschenbrenner D, Yoo JY, Zuo T. The gut mycobiome in health, disease, and clinical applications in association with the gut bacterial microbiome assembly. Lancet Microbe. 2022;3(12):e969–83. https://doi.org/10.1016/S2666-5247(22)00203-8.
2. Hoffmann C, et al. Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. PLoS ONE. 2013;8(6):e66019. https://doi.org/10.1371/journal.pone.0066019.
3. Hallen-Adams HE, Suhr MJ. Fungi in the healthy human gastrointestinal tract. Virulence. 2017;8(3):352–8. https://doi.org/10.1080/21505594.2016.1247140.
4. Iliev ID, Leonardi I. Fungal dysbiosis: immunity and interactions at mucosal barriers. Nat Rev Immunol. 2017;17(10):635–46. https://doi.org/10.1038/nri.2017.55.
5. Y. Gu, G. Zhou, X. Qin, S. Huang, B. Wang, and H. Cao, "The potential role of gut mycobiome in irritable bowel syndrome," *Front. Microbiol.*, vol. 10, Aug. 2019, https://doi.org/10.3389/fmicb.2019.01894.
6. Liguori G, et al. Fungal dysbiosis in mucosa-associated microbiota of Crohn's disease patients. J Crohns Colitis. 2016;10(3):296–305. https://doi.org/10.1093/ecco-jcc/jjv209.
7. Strati F, et al. New evidences on the altered gut microbiota in autism spectrum disorders. Microbiome. 2017;5(1):24. https://doi.org/10.1186/s40168-017-0242-1.
8. M. Mar Rodríguez *et al.*, "Obesity changes the human gut mycobiome," *Sci. Rep.*, vol. 5, p. 14600, Oct. 2015, https://doi.org/10.1038/srep14600.
9. Coker OO, et al. Enteric fungal microbiota dysbiosis and ecological alterations in colorectal cancer. Gut. 2019;68(4):654–62. https://doi.org/10.1136/gutjnl-2018-317178.
10. Luan C, et al. Dysbiosis of fungal microbiota in the intestinal mucosa of patients with colorectal adenomas. Sci Rep. 2015;5:7980. https://doi.org/10.1038/srep07980.
11. Gao R, et al. Dysbiosis signature of mycobiota in colon polyp and colorectal cancer. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2017;36(12):2457–68. https://doi.org/10.1007/s10096-017-3085-6.
12. Y. Hu *et al.*, "Inferring species compositions of complex fungal communities from long- and short-read sequence data," *mBio*, vol. 13, no. 2, pp. e02444–21, Apr. 2022, https://doi.org/10.1128/mbio.02444-21.
13. Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K. Identification of fungi in shotgun metagenomics datasets. PLoS ONE. 2018;13(2):e0192898. https://doi.org/10.1371/journal.pone.0192898.
14. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44. https://doi.org/10.1038/nbt.3935.
15. Seelbinder B, et al. Candida expansion in the gut of lung cancer patients associates with an ecological signature that supports growth under dysbiotic conditions. Nat Commun. 2023;14(1):2673. https://doi.org/10.1038/s41467-023-38058-8.
16. Marfil-Sánchez A, et al. Gut microbiome functionality might be associated with exercise tolerance and recurrence of resected early-stage lung cancer patients. PLoS ONE. 2021;16(11):e0259898. https://doi.org/10.1371/journal.pone.0259898.
17. Xing Y, et al. Multikingdom characterization of gut microbiota in patients with rheumatoid arthritis and rheumatoid arthritis-associated interstitial lung disease. J Med Virol. 2024;96(7):e29781. https://doi.org/10.1002/jmv.29781.
18. Xiao Z, et al. Characterizations of gut bacteriome, mycobiome, and virome of healthy individuals living in sea-level and high-altitude areas. Int Microbiol Off J Span Soc Microbiol. 2024. https://doi.org/10.1007/s10123-024-00531-9.
19. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. https://doi.org/10.1186/s13059-019-1891-0.
20. Blanco-Míguez A, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023;41(11):1633–44. https://doi.org/10.1038/s41587-023-01688-w.
21. LaPierre N, et al. MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. BMC Genomics. 2019;20(5):423. https://doi.org/10.1186/s12864-019-5699-9.
22. Xie Z, Manichanh C. FunOMIC: pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling. Comput Struct Biotechnol J. 2022;20:3685–94. https://doi.org/10.1016/j.csbj.2022.07.010.
23. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. Microbiome. 2021;9(1):58. https://doi.org/10.1186/s40168-021-01015-y.
24. Soverini M, Turroni S, Biagi E, Brigidi P, Candela M, Rampelli S. HumanMycobiomeScan: a new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. BMC Genomics. 2019;20(1):496. https://doi.org/10.1186/s12864-019-5883-y.
25. Rintarhat P, et al. Assessment of DNA extraction methods for human gut mycobiome analysis. R Soc Open Sci. 2023;11(1):231129. https://doi.org/10.1098/rsos.231129.
26. T. Goldfarb *et al.*, "NCBI RefSeq: reference sequence standards through 25 years of curation and annotation," *Nucleic Acids Res.*, p. gkae1038, Nov. 2024, https://doi.org/10.1093/nar/gkae1038.
27. Kim D, Gilchrist CLM, Chun J, Steinegger M. UFCG: database of universal fungal core genes and pipeline for genome-wide phylogenetic analysis of fungi. Nucleic Acids Res. 2023;51(D1):D777–84. https://doi.org/10.1093/nar/gkac894.
28. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4. https://doi.org/10.1093/bioinformatics/btr708.
29. Hiseni P, Rudi K, Wilson RC, Hegge FT, Snipen L. HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. Microbiome. 2021;9(1):165. https://doi.org/10.1186/s40168-021-01114-w.
30. A. I. Qureshi, "Exploring the landscape of bioinformatic tools for fungal identification in shotgun metagenomic sequencing data: for potential applications in colorectal cancer biomarker discovery," Master thesis, 2023. Accessed: Aug. 26, 2024. [Online]. Available: https://www.duo.uio.no/handle/10852/104186
31. Sun Z, et al. Challenges in benchmarking metagenomic profilers. Nat Methods. 2021;18(6):618–26. https://doi.org/10.1038/s41592-021-01141-3.
32. Usyk M, et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. Cell Rep Methods. 2023;3(1):100391. https://doi.org/10.1016/j.crmeth.2022.100391.
33. Xie Z, Canalda-Baltrons A, d'Enfert C, Manichanh C. Shotgun metagenomics reveals interkingdom association between intestinal bacteria and fungi involving competition for nutrients. Microbiome. 2023;11(1):275. https://doi.org/10.1186/s40168-023-01693-w.
34. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46.
35. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique cladespecific marker genes. Nat Methods. 2012;9(8):811–4. https://doi.org/10.1038/nmeth.2066.
36. Yan Q, et al. A genomic compendium of cultivated human gut fungi characterizes the gut mycobiome and its relevance to common diseases. Cell. 2024;187(12):2969-2989.e24. https://doi.org/10.1016/j.cell.2024.04.043.

Publisher's Note