Check for updates

RESEARCH ARTICLE

REVISED Nanopore sequencing of DNA barcodes succeeds in unveilling the diversity of fungal mock communities

[version 2; peer review: 4 approved with reservations]

Cristiano Pedroso-Roussado[1], Fergus Guppy[2], Lucas Bowler[1], Joao Inacio[1]

[1]School of Applied Sciences, University of Brighton, Brighton, BN2 4HQ, UK
[2]Institute of Life and Earth Sciences, School of Energy, Geoscidence, Infrastructure and Society, Heriot-Watt University, Edinburgh, EH14 4AS, UK

v2 **First published:** 08 Mar 2023, **3**:45
https://doi.org/10.12688/openreseurope.15485.1
Latest published: 17 Jul 2023, **3**:45
https://doi.org/10.12688/openreseurope.15485.2

Abstract

Background: The fungal component of the gut microbiome has been underrepresented in most gut microbiome studies. While metabarcoding approaches have been used to assess the diversity and role of the gut mycobiome, differences in experimental design and downstream analyses may induce bias and differential outcomes. This study assessed the capacity of nanopore sequencing to retrieve the microbial profile and relative abundance of a mock fungal community comprised of Candida glabrata, Meyerozyma guilliermondii, Pichia kudriavzevii, Clavispora lusitaniae and Candida parapsilosis.
Methods: The approach was implemented using the MinION system and involved the analysis of sequencing libraries made from equimolar mixtures of the PCR-amplified nuclear ribosomal internal transcribed spacer (ITS) of the target species.
Results: Nanopore sequencing successfully retrieved the composition of the fungal mock community in terms of the different taxa present. However, the approach was unable to correctly assess the expected relative abundances of each species in the same community, showing some yet undetermined bias that may be related to the size of the respective target DNA fragments. Clavispora lusitaniae was consistently overrepresented in the mixtures, while C. glabrata was underrepresented. The remaining three species showed relative abundances more aligned with the expected values of an equimolar mixture.
Conclusions: Although not yielding the expected results for the relative abundances, the values obtained from independent sequencing runs were similar for all species, suggesting a good reliability but questionable accuracy in this sequencing approach.

Keywords

Nanopore sequencing, fungal mock community

Open Peer Review

Approval Status ? ? ? ?

	1	2	3	4
version 2 (revision) 17 Jul 2023	? view		? view	? view
version 1 08 Mar 2023	✗ view	? view		

1. **Alfonso Benítez-Páez**, Principe Felipe Research Center (CIPF), Valencia, Spain
2. **R. Henrik Nilsson**, Goteborgs universitet, Gothenburg, Sweden
3. **Christoph Ammer-Herrmenau**, University medical center Goettingen, Goettingen, Germany
4. **Laurence Delhaes**, Univ-Bordeaux, Bordeaux, France

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Genetics and Genomics gateway.



This article is included in the Horizon 2020 gateway.



This article is included in the Microbiology gateway.



This article is included in the Cell, Molecular and Structural Biology gateway.



This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

Corresponding author: Cristiano Pedroso-Roussado (cristiano.roussado@outlook.pt)

Author roles: Pedroso-Roussado C: Conceptualization, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Guppy F**: Conceptualization, Methodology, Supervision, Validation, Writing – Review & Editing; **Bowler L**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Inacio J**: Conceptualization, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 801604. The funder had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Copyright: © 2023 Pedroso-Roussado C *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Pedroso-Roussado C, Guppy F, Bowler L and Inacio J. **Nanopore sequencing of DNA barcodes succeeds in unveilling the diversity of fungal mock communities [version 2; peer review: 4 approved with reservations]** Open Research Europe 2023, **3**:45 https://doi.org/10.12688/openreseurope.15485.2

First published: 08 Mar 2023, **3**:45 https://doi.org/10.12688/openreseurope.15485.1

REVISED **Amendments from Version 1**

We accepted all the comments from the reviewers and inserted revisions on the text. We did not perform the analysis on the coverage rate of the primers nor any other method to assess the rRNA copy number of the initial fungal mock community and we fully understand that this constitutes ground to reject out manuscript entirely: however we would find it surprising. Focusing on the changes we did: starting with the grammar and updated references with the correct links. Then, more details on methods regarding the origin of the fungal strains, and the inclusion in the conclusion section of a brief comment on the recent papers from Hu *et al.*, 2022 and Conti *et al.*, 2023. In respect to the novelty of this work we might disagree with reviewers since it is still relevant and innovative to develop easy experimental and analytical protocols to study fungal communities using nanopore sequencing. We agree that other studies were better equipped to answer such questions, but, in our view that does not decrease the quality of our work. We believe that our work represents a useful reading for small labs that want to start working with nanopore sequencing for studying fungal isolates and possess little or no bioinformatics knowledge. We also believe that we address the copy number effect and we do not overstate the robustness of nanopore sequencing specifically in regard to the relative abundances' results. Nevertheless, the taxonomic results are trustworthy with the BugSeq pipeline. We also want to thank Dr Sam Chorlton from the BugSeq Bioinformatics Inc. for addressing some of the reviewers comments and making the case for the robustness of their algorithm despite the conflict of interest demonstrated.

Any further responses from the reviewers can be found at the end of the article

Plain summary

The fungi present in the human gut have not been widely studied. Most laboratory sequencing techniques and data analyses protocols still impose limitations on the study of the fungal composition of the human gut. In this study, we assessed the capacity of the nanopore sequencing methodology to correctly classify a fungal mock community of known composition and abundance. The artificial community was composed of *Candida glabrata*, *Meyerozyma guilliermondii*, *Pichia kudriavzevii*, *Clavispora lusitania*, and *Candida parapsilosis*. The approach was implemented using the MinION system and involved the previous amplification of a target genomic region called internal transciber spacer from the five species. These amplified regions were mixed in equal concentrations and then the nanopore sequencing protocol was performed. Nanopore sequencing successfully retrieved the composition of the fungal mock community. However, the approach was unable to correctly assess the expected relative abundances of each species, showing some yet undetermined bias that may be related to the size of the respective target DNA fragments. *C. lusitaniae* was consistently overrepresented in the mixtures, while *C. glabrata* was underrepresented. The remaining three species showed relative abundances more aligned with the expected values. Although not yielding the expected results for the relative abundances, the values obtained from two sequencing runs were similar for all species, suggesting a good reliability but questionable accuracy in this sequencing approach.

Introduction

The global fungal diversity is underestimated (Garcia-Solache & Casadevall, 2010; Pal, 2017; Taylor *et al.*, 2001; Tedersoo *et al.*, 2014). The gut mycobiome is shaped from birth, and ongoing analysis from this early stage provides a window for study of its development and maturation. After the early development years, the gut mycobiome stabilizes, although possessing a lower alpha-diversity in comparison with the bacterial microbiota (Nash *et al.*, 2017; Raimondi *et al.*, 2019). The most abundant and observed phyla at this stage are *Ascomycota*, *Basidiomycota*, and *Mucoromycota*, the last one much less abundant than the rest (Nash *et al.*, 2017; Raimondi *et al.*, 2019; Wu *et al.*, 2020). Even though more stable, the mature gut mycobiome is still characterized by relative instability with both inter- and intra-individual differences occurring over time (Nash *et al.*, 2017; Raimondi *et al.*, 2019). Transient gut colonization can potentially explain the fungal composition instability, because some fungi reach the gut as a result of environmental exposure and via dietary intake (Nash *et al.*, 2017; Raimondi *et al.*, 2019). Despite recent advances in mycobiome studies, associations with the onset of some human diseases remains to be further explored. As per the bacterial microbiome, the fungal microbiome's variability between individuals presents further challenges in identifying patterns between mycobiome and remaining host physiology (Hallen-Adams & Suhr, 2017). For example, it is not clear how dynamic the mycobiome is in relation with ageing. Prolonged exposure to a processed diet and high isolation periods in pathogen-free environments has been shown to reduce fungal gut diversity (Mims *et al.*, 2021). While metabarcoding approaches have been used to assess the diversity and role of the gut mycobiome, differences in experimental design and downstream analyses may induce bias and differential outcomes (Nilsson *et al.*, 2019a). Nanopore sequencing is a recently developed third-generation sequencing technology allowing the development of fast, cost-effective, and portable approaches to study complex ecosystems, such as the gut microbiome, and it represents a promising tool to further unveil the fungal component of these microbial communities (Wang *et al.*, 2021). Thus, it is necessary to further investigate the impact of environmental and pharmacological agents on the whole gut microbiome, specifically the mycobiome component, allowing a more detailed comprehension about its composition, functionality and relationship with the host (Gutierrez & Arrieta, 2021).

Methods

Fungal growth conditions

A fungal mock community was prepared with five strains, representing species *Meyerozyma guilliermondii*, *Candida glabrata*, *Pichia kudriavzevii*, *Candida parapsilosis*, and *Clavispora lusitaniae*, that are usually found as members of the human gut mycobiota. These strains were obtained from a collection of historical clinical fungi isolates stored at -80° located at the School of Applied Sciences, University of Brighton

(United Kingdom) originally obtained from the Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa (Portugal). All stock isolates were grown in 5 mL of Sabouraud dextrose broth in universal flasks for 2 days at 37°C under shaking (270 rpm). After growth was observed, isolates were plated and streaked twice to achieve purification in Potato Dextrose Agar and incubated at 30°C for 18–24h until DNA extraction.

Fungal DNA extraction and amplicons generation

Fungal isolates identification confirmation. For DNA extraction, a custom protocol was performed. From fresh culture plates (typically 18 to 24 h), a small amount of yeast cells was collected and transferred to a 1.5 mL microtube containing 500 µL sterile water. The microtube was vortexed at 10,000× g for 5 min and the supernatant discarded. Then, 300 µL of lysis buffer (100mM Tris-HCl [pH 8.0], 1 mM EDTA, 2% Triton X-100) were added to the pellet and mixed with the equivalent of 100 µL glass beads (ø = 0.5 mm). The mixture was briefly vortexed and incubated at 95°C for 10 min before cell' disruption in the Savant Bio 101 FastPrep FP120 (Savant Instruments, India) (maximum speed = 6.5; 40 sec). A final centrifugation step was performed for 5 min at 10,000× g. The supernatant containing the extracted DNA was stored at -20°C. Polymerase chain reaction (PCR) was performed to amplify the ITS region with diluted genomic DNA (1:10) from each isolate, with the following primer sequences: ITS1F – CTTGGTCATTTAGAGGAAGTAA; ITS4 – TCCTCCGCTTATTGATATGC (Op De Beeck *et al.*, 2014). The PCR mixture (25 µL total volume) contained between 1 and 4 ng of DNA template (or 6 µL of unquantified initial DNA), 1X Standard *Taq* Reaction Buffer, 0.2 mM of dNTPs, 0.4 µM of ITS1F and ITS4, and 0.625 U of *Taq* DNA Polymerase (New England Biolabs, United Kingdom). The PCR thermal profile consisted of an initial denaturation of 1 min at 95 °C, followed by 35 cycles of 30 s at 95°C, 40 s at 50°C, 60 s at 68°C, and a final step of 5 min at 68°C. The PCR products were visualized in an agarose gel 1.1% (w/v) running for 1 hour at 220 V (ThermoFisher, United States of America). After positive band check, the PCR products were purified with GeneJET purification kit according to the manufacturer's protocol (Thermofisher, United States of America) and sequenced by Sanger sequencing using the same primers (Eurofins, Germany).

Fungal mock community amplicons generation and library preparation. To achieve the best quality gDNA, purified isolates' DNA extraction was performed with MasterPure Yeast DNA Purification Kit, according to the manufacturer's protocol (Lucigen, United States of America). Extracted DNA was stored at -20° C until further use. The PCR mixture and thermal profile performed was described above. PCR products were visualized in an agarose gel 1.1% (w/v). Then, PCR products were purified with GeneJET purification kit, according to the manufacturer's protocol and stored in the fridge at 4°C until further use. To build the mock community, equimolar concentrations of the purified PCR products originating from the five strains were pooled together for library preparation, *i.e.*, 40 fmol per strain, accordingly to ONT protocols.

Nanopore sequencing approaches

Library of ITS amplicons from a fungal mock community. For the nanopore sequencing approach, optimization using a fungal mock community, the Ligation Sequencing Kit (SQK-LSK 109, ONT, United Kingdom) was applied. A mixture comprising ITS amplicons previously generated by PCR were used as template for DNA library preparation. Two independent DNA sequencing libraries were prepared for separate sequencing runs. The pooled ITS amplicon concentration was adjusted to reach 1 µg in a volume of 47 µL, with 1 µL of DNA CS (sequencing positive control containing DNA obtained from lambda phage), 3.5 µL of NEBNext FFPE DNA Repair Buffer (New England Biolabs, United Kingdom), 2 µL of NEBNext FFPE DNA Repair Mix (New England Biolabs, United Kingdom), 3.5 µL Ultra II End-prep reaction buffer, and 3 µL of Ultra-prep enzyme mix (New England Biolabs, UK), completed with molecular biology-grade water. The mixture was incubated in a thermal cycler at 20°C for 5 min and 65 °C for 5 min. A purification step using 1× Agencourt AMPure XP beads was performed. Briefly, the mixture was incubated at room temperature with gentle mixing for 5 min, washed twice with 200 µL fresh 70% ethanol, with pellet allowed to dry at room temperature for 1 min before DNA being eluted in 61 µL nuclease-free water, and incubated at room temperature for less than 5 min. A 1 µL aliquot was quantified by fluorometry using Qubit 4 Fluorometer. For the adapter ligation step, the total 60 µL of end-prepped DNA from the previous step were added in a mix containing 25 µL of ligation buffer LNB (NEBNext Quick Ligation Module, New England Biolabs, United Kingdom), 10 µL NEBNext Quick T4 DNA Ligase (New England Biolabs, United Kingdom), and 5 µL of Adapter Mix (ONT, United Kingdom), and incubated at room temperature for 10 min. A purification step using 0.4× Agencourt AMPure XP beads was performed. Briefly, the mixture was incubated at room temperature with gentle mixing for 5 min, washed twice with 250 µL of short fragment buffer (ONT, United Kingdom) (to allow the recovery of fragments of all sizes), pellet allowed to dry at room temperature for 1 min before DNA being resuspended in 15 µL of elution buffer (ONT, United Kingdom), and incubated at room temperature for 10 min or more. The mixture was retained and coined as 'DNA sequencing library'. A 1 µL aliquot was quantified by fluorometry using Qubit 4 Fluorometer and quality measured with DeNovix DS-11 Spectrophotometer to ensure no contaminants were present before the library loading step. If the DNA sequencing library did not show $A_{260/280}$ between 1.8–2.0 and $A_{260/230}$ between 2.0–2.2, it was discarded, and the protocol repeated.

MinION sequencing and basecalling. Nanopore sequencing was performed using MinION MK1B sequencer (ONT, United Kingdom) and flowcells R9.4.1 (FLOW-MIN106D, ONT, United Kingdom) using MinKNOW software v.21.02.1 (ONT, United Kingdom). After the MinION Flow Cell was setup accordingly to manufacturer's instructions, the pre-sequencing mix was prepared by adding 12 µL of the DNA sequencing library containing at least 50 fmol, 37.5 µL of sequencing buffer (ONT, United Kingdom), and 25.5 µL of Loading Beads (ONT, United Kingdom). The flowcell was primed with

a mixture of 30 µL Flush Tether introduced directly into the flush buffer tube (Flow Cell Priming Kit, EXP-FLP002), according to manufacturer's protocol (ONT, United Kingdom). Five minutes after the priming of the flowcell, the DNA nanopore sequencing library was mixed by pipetting and loaded in a dropwise fashion using the SpotON port, according to manufacturer's protocol. Once the DNA sequencing library was loaded, the 24 h sequencing run was initiated using MinKNOW software v.21.02.1. Basecalling was performed with *guppy* v3.2.6 (Oxford Nanopore Technologies) from the command line using the *fast basecalling* mode, with the quality score filter of 8, and adapters' trimming mode *on*. Quality controls were performed using the *nanoplot* tool (De Coster *et al.*, 2018) with the length filters parameters of maximum length at 1300 and minimum at 300.

Taxonomical classification

What's in my Pot and BugSeq. WIMP (version 2021.11.26) was used as a first try for the classification assignments in the fungal mock community nanopore sequencing profile experiments. WIMP is a workflow available in the EPI2ME Desktop Agent (ONT, United Kingdom), for the taxonomic classification of basecalled sequences (reads) generated by nanopore sequencing. Firstly, WIMP filtered FASTQ files with a quality Qscore below a minimum threshold (default is 7). The sequencing read quality was also measured by the Qscore. For reads that pass the quality threshold, Centrifuge software (Kim *et al.*, 2016) was used to assign a taxon to each read, and thus a score was attributed to every assignment. Additionally, BugSeq online software (Fan *et al.*, 2021) was also employed. The BugSeq algorithm performs *minimap2* alignments using a nucleotide database, followed by Bayesian reassignment and Lowest-Common Ancestor identification (Fan *et al.*, 2021).

Results

Fungal mock community nanopore sequencing

Nanopore sequencing quality control assessment. The mock community was generated by amplifying the ITS regions of five fungal species - *C. glabrata*, *C. parapsilosis*, *C. lusitaniae*, *M. guilliermondii*, and *P. kudriavzevii*. Two independent DNA sequencing libraries were prepared and sequenced in separate experiments. One first classification attempt was tried with WIMP that produced an inaccurate taxonomical assignment (Figure 1).

Therefore, BugSeq's pipeline may be more curated than WIMP's, since it includes a post-classification step to remove and quality control the assignments which is absent in WIMP pipeline (Fan *et al.*, 2021; Marti, 2019). Therefore, BugSeq pipeline was employed for nanopore sequenced reads data analyses (Table 1).

Both average Qscore and *recentrifuge* scores revealed a high basecalling error-rate (~6–9%) and a low alignment score (56.9–58.0, with 57–58% of the sequenced reads matched to one entry in the reference databases). Indeed, nanopore sequencing could retrieve the correct assignments of the mock community in both sequencing runs (Table 2). BugSeq could

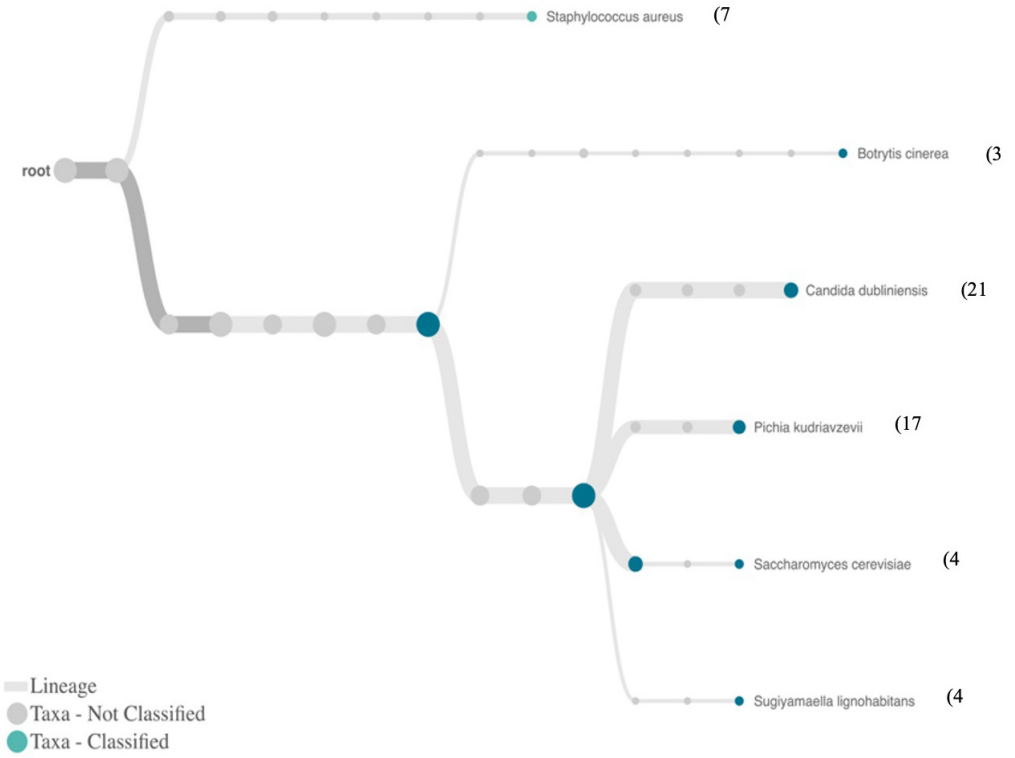

Figure 1. Heat tree displaying the community structure of the Mock community A. A total of 1,314,572 reads were analysed with WIMP, from which 1,084,471 were classified. The average quality score was 9.43 and the average sequence length was 550 bases. Only taxa with relative abundances above 3% are showed (in parenthesis are depicted their relative abundances).

Table 1. Nanopore sequencing performance overview of two independent sequencing runs from the same fungal mock community targeting ITS amplicons.

Sample ID	Quality control[1]		Alignment parameters			
	No. reads	Average Qscore	No. analysed reads	*recentrifuge* score average	Average read length (b)	No. taxa detected
Run A	1,314,571	12.03	1,308,263	56.9	512	146
Run B	4,879,747	10.65	4,655,632	58	543	255

[1] – Quality control was obtained through *multiqc* tool and alignment's quality through *recentrifuged* tool in BugSeq pipeline.

Table 2. Nanopore sequencing performance overview of two independent runs (termed 'Run A' and 'Run B') from the same fungal mock community targeting ITS amplicons. Taxonomic assignments were retrieved through recentrifuge tool in BugSeq pipeline.

Taxa ID	ITS fragment size (b)	No. reads		Relative abundance (%)		Normalized average *recentrifuge* score	
		Run A	Run B	Run A	Run B	Run A	Run B
Clavispora lusitaniae ATCC 42720	552	607,695	1,842,887	46	40	53.5	56.7
Candida parapsilosis	690	237,533	532,047	18	11	62.4	65.3
Meyerozyma gulliermondii ATCC 6260	775	207,528	773,358	16	17	60	63.1
Pichia kudriavzevii	679	179,138	859,268	14	18	60.6	62.6
Candida glabrata	1050	54,718	376,756	4	8	61.3	64

attribute the expected taxa identifiers even when the *recentrifuge* assignment score barely surpassed the 50% of matching. However, the presence of ITS sequences in the references databases is sufficiently represented in the way that a 50% match is considered sufficient to assign the correct taxon using BugSeq bioinformatic pipeline. Still, the relative abundances of these species in the fungal mock community did not yield the expected value (Table 2). *C. lusitaniae* was consistently overrepresented in the mixtures, showing relative abundances of 40% to 47% of the assigned reads instead of the expected 20%, while *C. glabrata* was underrepresented, with only 4% to 8% of the assigned reads.

The two independent fungal mock community nanopore sequencing runs showed high precision. Despite a difference of around 3.5 million reads between the two sequencing runs, the average Qscore (12.03 *vs* 10.65), *recentrifuge* alignment score (56.9 *vs* 58), and average read length (512 *vs* 543) were similar. The five species were detected and were the most abundant taxa observed between the assigned taxa. The overall performance here described revealed an expected nanopore sequencing accuracy, with relatively high error-rate and a long list of 'taxonomic noise' (146 total taxa found in Run A, and 255 in Run B). Indeed, a higher number of taxa detected was observed in Run B, the sequencing run where the most sequenced reads were generated. In the Run A, 1% of the total reads were not assigned, and less than 1% was assigned to bacteria (0.2%), synthetic constructs (0.2%), viruses (0.2%), and archaea (less than 0.01% assigned to *Sulfolobus acidocaldarius*). In the Run B, 5% of the total reads were not assigned, and less than 1% was assigned to bacteria (0.6%), synthetic constructs (0.1%), viruses (less than 0.01%), and archaea (less than 0.01% assigned to *Euryarchaeota*). However, the average classification score and the high error-rate observed were not sufficient to hamper the fungal mock community taxonomic assignments since the original fungi mock community's members were correctly assigned (Table 2).

Discussion

A fungal mock community comprising *C. lusitaniae*, *C. glabrata*, *C. parapsilosis*, *M. guilliermondii*, and *P. kudriavzevii* was used to assess the precision and accuracy of targeted nanopore sequencing protocols based on the use of the fungal ITS genomic taxonomical marker. The goal was to check the capacity of nanopore sequencing to correctly assign the community composition while detecting the correct abundances of each species. The incongruent output retrieved by WIMP strange and cannot be related to the increased representation of closely related sequences in the database. However, WIMP's developers stated that by allowing increased specificity of assignments

at the family and genus level it also might blur the line between lower levels of taxonomic classification, such as genus and species (ONT technical team, personal communication). Despite less curated entries in the reference databases, the bioinformatic pipeline behind WIMP might not be the best choice to analyse our type of data. Although the advantage of *centrifuge* is that it is optimized for solving metagenomic classification problems and it is capable of accurate identification of reads even when databases containing multiple highly similar references genomes are used (ONT technical team, personal communication), it has limitations compared with other algorithms such as *minimap2* (Ammer-Herrmenau *et al.*, 2021). Thus, it can be hypothesized that a different approach based on BLAST/MEGAN can surpass these limitations (Bağcı *et al.*, 2019).

WIMP attributes a score after determining the most reliable place in the taxonomy tree for those reads (Kim *et al.*, 2016). On the other hand, the BugSeq pipeline assigns taxonomical classifications based on the *recentrifuge* algorithm instead of *centrifuge* performed by WIMP, and it relies on a three-step pipeline node to assign taxonomical identifiers to every read (Fan *et al.*, 2021). The first step is the reads mapping using *minimap2*, the second is the reassignment of mapped reads based on a Bayesian statistical framework using *pathoscope* (Francis *et al.*, 2013). This reassignment step optimizes the multiple strains coverage, mediating the absence of target taxa in the reference database, surpassing the need of multiple alignments, extensive homology searches, or genome assembly steps (Francis *et al.*, 2013). The last step in the BugSeq pipeline is the calculation of the Lowest-Common Ancestor as input for *recentrifuge* (Marti, 2019), which facilitates vigorous contamination removal (such as crossovers) and quality comparative analysis of multiple samples, giving a confidence level score for every result that propagates to further downstream analyses.

Recently, D'Andreano *et al.* (2020) were successful in developing a nanopore sequencing approach to assess the fungal composition from clinical samples. The authors targeted a 3.5 kb region (V3, 18S-ITS1-5.8S-ITS2-28S D2) and a 6 kb region (V1, 18S-ITS1-5.8S-ITS2-28S D12) for amplicons generation, performed taxonomical assignments using WIMP, and compared the performance with a known fungal mock community – ZymoBIOMICS (D'Andreano *et al.*, 2020). Besides using longer amplicons than the approach tested here, the authors also removed barcodes and adapters using *porechop* (which is unsupported since October 2018). However, it is unlikely that the inclusion of barcodes and adapters to downstream analyses could explain such significant discrepancies found in the fungal mock community analysed in this report, because *C. parapsilosis* (Table 2, run A), *M. guilliermondii* (Table 2, run B), and *P. kudriavzevii* (Table 2, run 2) showed relative abundances as expected (Table 2). Moreover, there was a difference of 1–7 percentage points between the two independent runs attributed to the same taxa, which was more probably explained by chance or other factors, such as base-calling error-rate, or *centrifuge* indexation, than to the presence of barcodes and adapters. In another study, Hu *et al.* (2021) evaluated a nanopore sequencing approach to study a fungal mock community comprised of 43 species. The authors generated amplicons by amplifying the ITS1 regions of the rRNA gene with the universal fungal primers ITS1F and ITS2 and performed adapters and barcode trimming with *qcat* after a barcoded nanopore sequencing run (Hu *et al.*, 2021). The authors compared classification and community analysis pipelines for nanopore sequencing data in amplicon and shotgun approaches from fungal mock communities comprised of 43 species (Hu *et al.*, 2021). Their approach consisted of a generation of a custom reference database comprised solely with the species from the mock community, *i.e.*, all the genomes of the 43 species in the mock community were downloaded from the NCBI (Hu *et al.*, 2021). Then, the authors used *kraken2* to perform a search to identify potential contaminated regions in the concatenated FASTA and masked those regions using *bedtools* (Quinlan & Hall, 2010). Low complexity regions were also masked using *dustmasker* from the BLAST+ package (Camacho *et al.*, 2009). Lastly, they used *makeblastdb* to construct the mock community custom database (Hu *et al.*, 2021). In regard to the amplicon-based approach, the authors used *kraken2* as the k-mer based-algorithm and *minimap2* as the alignment-based algorithm. In the *minimap2* analysis, the accessions of the best hits were extracted from the output files and their taxonomic correspondence were searched in the NCBI taxonomic map using python *pandas* module (Hu *et al.*, 2021). Finally, the authors merged information from different output files and performed *ete3* module again to assign taxonomic information to each read (Hu *et al.*, 2021). The authors concluded that combining BLAST with the fungal specific RefSeq fungal database was the best approach to obtain the most precise classifications for the fungal mock community data sets (Hu *et al.*, 2021). However Unite is a consistently better tool for metabarcoding (Nilsson *et al.*, 2019b) The classification can also be improved if cutoffs on query coverage were applied, including flow-on effects on downstream community composition analysis sourced from complementary shotgun/metagenomics data sets (Hu *et al.*, 2021). Such curated nanopore sequencing approach decreased the chance of misclassification and mismatches in the classification. These results cannot be explained by PCR amplification biases since the mock community was built with equimolar concentrations of each species ITS amplicons after PCR (Hu *et al.*, 2021). In detail, PCR amplification biases derive mainly from inconsistent amplification of the genomic barcoding region that differs in every species, caused by copy number variations and different primer binding specificities (Nearing *et al.*, 2021; Silverman *et al.*, 2021). Additionally, in fungi, variations in barcode genomic regions lengths are the most probable cause to explain biases in recording fungal community compositions, with longer amplicons being underrepresented (Castaño *et al.*, 2020). Thus, these results give evidence that we can treat nanopore sequence data with confidence despite the high error-rate. Since nanopore sequencing is a recent technology, there is a lack of experimental and analytical standards for particular approaches like the one here tested and so it is challenging to define

what could be accepted or not as a correct approach (Leggett & Clark, 2017). Additionally, recent studies have observed that the extra length of nanopore reads overcome the accuracy limitations (Pearman *et al.*, 2020). Still, it is expected that ONT or other researchers develop better sequencing chemistry and bioinformatic tools to improve these challenges, as it is already expected by the super-accurate basecalling model (announced as <2% error-rate).

Taking the limitations of WIMP into consideration, the fungal mock community was also analyzed with the BugSeq pipeline. It was observed that a BugSeq's *recentrifuge* classification score as low as 54 could still perform the correct taxonomical assignments and reasonably differentiate the different species' relative abundances. Therefore, it is absolutely crucial to carefully evaluate the bioinformatic pipeline choice, since they can profoundly impact the taxonomic and microbial abundances analyses in the samples under study. In conclusion, these preliminary results suggest that either upstream experimental steps, downstream bioinformatics analyses, or both, can introduce biases that are confounding the fungal relative abundances from biological samples.

In conclusion, these results confirm that nanopore sequencing is a very sensitive approach, and the protocol execution must be performed with care and optimized to the specific biological sample and scientific question under investigation. Impairments may occur either upstream during sample collection, DNA extraction, PCR amplicons generation (among others), or downstream, during or post-sequencing, such as basecalling errrors, and further issues in the bioinformatic pipeline performed. These conclusions are supported by recent papers that performed similar but more complete approaches (Conti *et al.*, 2023; Hu *et al.*, 2022). Their results were improved in terms of reducing the taxonomic noise and better sequence-based classification by applying custom-based analysis pipelines on fungal mock communities.

Data availability

Underlying data

The dataset analysed during the current study is available in the NCBI Sequence Read Archive, under the BioProject accession number PRJNA914708.

Author contributions

Conceptualization, C.P-R., F.G., L.B., and J.I.; Formal Analysis C.P-R; Investigation, C.P-R.; Methodology, C.P-R, F.G., L.B., and J.I.; Supervision and Validation, F.G., L.B., and J.I.; Writing – Original Draft, C.P-R.; Writing – Review & Editing, C.P-R., F.G., L.B., and J.I.

References

Ammer-Herrmenau C, Pfisterer N, van den Berg T, *et al.*: **Comprehensive wet-bench and bioinformatics workflow for complex microbiota using Oxford Nanopore technologies.** *mSystems.* 2021; **6**(4): e0075021.
PubMed Abstract | Publisher Full Text | Free Full Text

Bağcı C, Beier S, Górska A, *et al.*: **Introduction to the analysis of environmental sequences: metagenomics with MEGAN.** *Evolutionary Genomics.* Humana, New York, NY, 2019; **1910**: 591–604.
Publisher Full Text

Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: architecture and applications.** *BMC Bioinformatics.* 2009; **10**(1): 421.
PubMed Abstract | Publisher Full Text | Free Full Text

Castaño C, Berlin A, Durling MB, *et al.*: **Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities.** *New Phytol.* 2020; **228**(3).
PubMed Abstract | Publisher Full Text

Conti A, Pierantoni DC, Robert V, *et al.*: **MinION Sequencing of Yeast Mock Communities To Assess the Effect of Databases and ITS-LSU Markers on the Reliability of Metabarcoding Analysis.** *Microbiol Spectr.* 2023; **11**(1): e0105222.
PubMed Abstract | Publisher Full Text | Free Full Text

D'Andreano S, Cuscó A, Francino O: **Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples.** *Biol Methods Protoc.* 2020; **6**(1): bpaa026.
PubMed Abstract | Publisher Full Text | Free Full Text

De Coster W, D'Hert S, Schultz DT, *et al.*: **NanoPack: visualizing and processing long-read sequencing data.** *Bioinformatics.* 2018; **34**(15): 2666–2669.
PubMed Abstract | Publisher Full Text | Free Full Text

Fan J, Huang S, Chorlton SD: **BugSeq: a highly accurate cloud platform for long-read metagenomic analyses.** *BMC Bioinformatics.* 2021; **22**(1): 160.
PubMed Abstract | Publisher Full Text | Free Full Text

Francis OE, Bendall M, Manimaran S, *et al.*: ***Pathoscope*: species identification and strain attribution with unassembled sequencing data.** *Genome Res.* 2013; **23**(10): 1721–1729.
PubMed Abstract | Publisher Full Text | Free Full Text

Garcia-Solache MA, Casadevall A: **Global warming will bring new fungal diseases for mammals.** *mBio.* 2010; **1**(1): e00061–10.
PubMed Abstract | Publisher Full Text | Free Full Text

Gutierrez MW, Arrieta MC: **The intestinal mycobiome as a determinant of host immune and metabolic health.** *Curr Opin Microbiol.* 2021; **62**: 8–13.
PubMed Abstract | Publisher Full Text

Hallen-Adams HE, Suhr MJ: **Fungi in the healthy human gastrointestinal tract.** *Virulence.* 2017; **8**(3): 352–358.
PubMed Abstract | Publisher Full Text | Free Full Text

Hu Y, Irinyi L, Hoang MTV, *et al.*: **Inferring species compositions of complex fungal communities from long- and short-read sequence data.** *bioRxiv.* 2021.
Publisher Full Text

Hu Y, Irinyi L, Hoang MTV, *et al.*: **Inferring species compositions of complex fungal communities from long- and short-read sequence data.** *mBio.* 2022; **13**(2): e0244421.
PubMed Abstract | Publisher Full Text | Free Full Text

Kim D, Song L, Breitwieser FP, *et al.*: **Centrifuge: rapid and sensitive classification of metagenomic sequences.** *Genome Res.* 2016; **26**(12): 1721–1729.
PubMed Abstract | Publisher Full Text | Free Full Text

Leggett RM, Clark MD: **A world of opportunities with nanopore sequencing.** *J Exp Bot.* 2017; **68**(20): 5419–5429.
PubMed Abstract | Publisher Full Text

Martí JM: **Recentrifuge: Robust comparative analysis and contamination removal for metagenomics.** *PLoS Comput Biol.* 2019; **15**(4): e1006967.
PubMed Abstract | Publisher Full Text | Free Full Text

Mims TS, Abdallah QA, Stewart JD, *et al.*: **The gut mycobiome of healthy mice is shaped by the environment and correlates with metabolic outcomes in response to diet.** *Commun Biol.* 2021; **4**(1): 281.
PubMed Abstract | Publisher Full Text | Free Full Text

Nash AK, Auchtung TA, Wong MC, *et al.*: **The gut mycobiome of the Human Microbiome Project healthy cohort.** *Microbiome.* 2017; **5**(1): 153.
PubMed Abstract | Publisher Full Text | Free Full Text

Nearing JT, Comeau AM, Langille MGI: **Identifying biases and their potential solutions in human microbiome studies.** *Microbiome.* 2021; **9**(1): 113.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Nilsson RH, Anslan S, Bahram M, *et al.*: **Mycobiome diversity: high-throughput sequencing and identification of fungi.** *Nat Rev Microbiol.* 2019a; **17**(2): 95–109.
PubMed Abstract | **Publisher Full Text**

Nilsson RH, Larsson KH, Taylor AFS, *et al.*: **The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications.** *Nucleic Acids Res.* 2019b; **47**(D1): D259–D264.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Op De Beeck M, Lievens B, Busschaert P, *et al.*: **Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies.** *PLoS One.* 2014; **9**(6): e97629.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pal M: **Morbidity and mortality due to fungal infections.** *J Appl Microbiol Biochem.* 2017; **1**(1): 2.
Publisher Full Text

Pearman WS, Freed NE, Silander OK: **Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads.** *BMC Bioinformatics.* 2020; **21**(1): 220.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Raimondi S, Amaretti A, Gozzoli C, *et al.*: **Longitudinal survey of fungi in the human gut: ITS profiling, phenotyping, and colonization.** *Front Microbiol.* 2019; **10**: 1575.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Sayers EW, Bolton EE, Brister JR, *et al.*: **Database resources of the National Center for Biotechnology Information in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D29–D38.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Silverman JD, Bloom RJ, Jiang S, *et al.*: **Measuring and mitigating PCR bias in microbiota datasets.** *PLoS Comput Biol.* 2021; **17**(7): e1009113.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Taylor LH, Latham SM, Woolhouse MEJ: **Risk factors for human disease emergence.** *Philos Trans R Soc Lond B Biol Sci.* 2001; **356**(1411): 983–989.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Tedersoo L, Bahram M, Põlme S, *et al.*: **Fungal biogeography. Global diversity and geography of soil fungi.** *Science.* 2014; **346**(6213): 1256688.
PubMed Abstract | **Publisher Full Text**

Wang Y, Zhao Y, Bollas A, *et al.*: **Nanopore sequencing technology, bioinformatics and applications.** *Nat Biotechnol.* 2021; **39**(11): 1348–1365.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wu L, Zeng T, Deligios M, *et al.*: **Age-related variation of bacterial and fungal communities in different body habitats across the young, elderly, and centenarians in Sardinia.** *mSphere.* 2020; **5**(1): e00558–19.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ? ? ? ?

Version 2

Reviewer Report 09 September 2024

https://doi.org/10.21956/openreseurope.17642.r34873

© **2024 Delhaes L.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Laurence Delhaes
INSERM, Centre de Recherche Cardio-thoracique de Bordeaux & CHU de Bordeaux, Laboratoire de Parasitologie, Mycologie, Service d'exploration fonctionnelle respiratoire, Univ-Bordeaux, Bordeaux, France

The article "Nanopore sequencing of DNA barcodes succeeds in unveilling the diversity of fungal mock communities [version 2; peer review: 2 approved with reservations]" is well written, easy to read with an exhaustive bibliography.

In this paper, using Nanopore technology (ONT) (especially MiniONT) to analyze fungal community (mycobiome) is documented trough a mock community analysis.

- A positive point of this work is the bioinformatic aspects, highly deciphered which clearly highlight several pitfalls such as having a critical view on the bioinformatic pipeline, not using WIMP as the unique solution provided by ONT, but keeping in mind that WIMP might not be the best choice to analyze mycobiota data.

- While the work is well conducted, it would have been of major interest to build a more complex mock community, composed of other fungi than yeasts (such as mucorales or other molds: *Aspergillus* for example).

- Finally, the added value of the whole data appears somewhat limited, and should be added by the mycobiota analysis of human gut samples for example.

Sincerely yours

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Mycobiome, NGS

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 20 September 2023

https://doi.org/10.21956/openreseurope.17642.r34875

© 2023 Ammer-Herrmenau C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Christoph Ammer-Herrmenau
Gastroenterology, Gastrointestinal Oncology and Endocrinology, University medical center Goettingen, Goettingen, Lower Saxony, Germany

Pedroso-Roussado *et al.* performed ONT marker gene sequencing with five fungi species. The authors highlight that the ONT build-in classification pipeline (WIMP) is by far not sufficient for mycobiome analysis and that other classification tools are needed. They demonstrate that ONT sequencing is able to detect fungi species. However, the relative abundances are differing vastly between two species. I kindly ask the authors to discuss and/ or address following points.

Major:

1. ITS-Amplicons have a high risk of chimerics[1]. It might be of high interest to identify chimerics and remove them before classification.

2. The unsatisfying results regarding relative abundances might be also explained by the low qscore that was set for the basecalling. Can the authors provide additional results with higher qscore (e.g. > 12)? That might also reduce the high false positive rate of classified taxa. It might be also worth to implement the super-accuracy basecalling algorithm for this experiment.

Minor:

1. Please specify, which strains of these five fungi were used.

2. Can you please provide information about the library, which is behind WIMP? Are all five species in this database which centrifuge uses for classification? It might be that the results are skewed by the fact that not all species were in this database. Are that complete and incomplete genomes?

3. Our group (Ammer-Herrmenau *et al.* 2021[2]) used the combination of centrifuge annotation followed by minimap2 alignment control of centrifuge annotated taxids. In your discussion, it sounds that only minimap2 was used.

4. In the discussion part, the authors mention that they kept barcodes and adapter sequences at their reads for classification. That is a discrepancy to the method section, where they stated that they chose adapter trimming on flag for guppy basecalling.

5. Please discuss that all five taxa are taxonomical highly related. Most are former or still species from the genus *Candida* and all belong to the same family. It might be that *Clavispora lusitaniae* and *Canidida glabrata* are too similar according their ITS genes. Please expand information about this.

6. Please provide information about average read length per species.

References

1. Nilsson RH, Tedersoo L, Ryberg M, Kristiansson E, et al.: A Comprehensive, Automatically Updated Fungal ITS Sequence Dataset for Reference-Based Chimera Control in Environmental Sequencing Efforts.*Microbes Environ*. 2015; **30** (2): 145-50 PubMed Abstract | Publisher Full Text
2. Ammer-Herrmenau C, Pfisterer N, van den Berg T, Gavrilova I, et al.: Comprehensive Wet-Bench and Bioinformatics Workflow for Complex Microbiota Using Oxford Nanopore Technologies. *mSystems*. 2021; **6** (4): e0075021 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbiome research based on ONT sequencing, pancreatology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 24 July 2023

https://doi.org/10.21956/openreseurope.17642.r33762

© 2023 Benítez-Páez A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Alfonso Benítez-Páez
Host-Microbe Interactions in Metabolic Health Laboratory, Principe Felipe Research Center (CIPF), Valencia, Spain

Thanks to the authors for addressing some of the critical points raised by this peer.

Notwithstanding, I think this investigation still has essential gaps to explaining discrepancies in the number of taxa retrieved (noise or uncontrolled contamination) from the taxonomy tool employed and the uneven relative abundance obtained in the mock community despite the load equimolar proportions of all amplicon samples produced. The authors argue this approach is good enough for non-specialists in bioinformatics analyses. Still, I think user-friendly applications cannot compromise quality and global performance in favour of being manageable.

I do not demerit the effort of the authors to present a worthwhile work, but it has significant limitations compared to current research trends in this field.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human microbiome, nanopore-based sequencing, computational biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 06 April 2023

https://doi.org/10.21956/openreseurope.16741.r31093

© 2023 Nilsson R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

R. Henrik Nilsson
[1] Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, Goteborgs universitet, Gothenburg, Västra Götaland County, Sweden
[2] Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, Goteborgs universitet, Gothenburg, Västra Götaland County, Sweden

I read Pedrouso-Roussado *et al.* with interest but also curiosity. Since several papers of this kind have been published already, I was eager to find out what the authors were bringing to the table.

The lack of line numbers is unfortunate and makes it difficult to provide feedback.

Page 1.

- Title: You would expect the title of a study in the natural sciences to say something about the field of application of the study, something about the methods used, and something about the results. The authors tick off the two first boxes but not the third. “to unveil” is a quite uninformative verb. “succeeds in unveiling” would be more informative, if the authors indeed feel that this is what happened.

- Abstract and elsewhere: The authors say “next-generation sequencing (NGS)”, but I’d say “metabarcoding” would be more accurate. See Taberlet *et al.,* (2012)[1].

- "internal transcribed spacer genomic regions" > "nuclear ribosomal internal transcribed spacer (ITS) region" [also on P3].

Page 3.

- I wouldn't necessarily call nanopore sequencing "novel" since it has been around for a number of years by now.
- "*Clavispora Lusitania*" > "*Clavispora lusitania*".
- In the Methods section, the authors should bring up from where the cultures were obtained. I see in Table 2 that two of them have ATCC strain IDs. The others don't. Please provide this information in the interest of scientific reproducibility.
- "cells' disruption" > "cell disruption".
- "ITSF1" – I imagine that the authors intended "ITS1F" here. It's a most unfortunate primer choice for metabarcoding, and it has been strongly recommended against for this purpose by the mycological community for more than 10 years (e.g., Bellemain *et al.,* (2010)[2] ; Ihrmark *et al.,* (2012)[3]; Tedersoo *et al.,* (2022)[4].
- "...The broadly used fungus-specific forward primer ITS1F is particularly problematic because of several critical mismatches in certain groups of moulds and putative animal pathogens..."). The authors were lucky to escape unscathed (did they?) here.

Page 4.

- "amplicons' concentration" > "amplicon concentration" ?
- Is "Short Fragment Buffer" a proper noun or a capitonym? Because if not, I'd go "short fragment buffer". Same with "Elution Buffer", "Sequencing Buffer", "Loading Buffer", and "Flush Buffer".
- "guppy standalone tool" > "guppy". But in the interest of scientific reproducibility, please provide the version number of guppy, nanoplot, and all other software tools used in the making of the study.
- "BugSeq algorithm" > "The BugSeq algorithm".

Page 5.

- "originated" > "produced"?
- "barely surpass" > "barely surpasses" or "barely surpassed" ?
- Sounds awkward to me that "the presence" was "satisfactorily represented". Please rewrite this sentence for clarity.
- Figure 1 legend – should it be "Heat tree" rather than "Heat trees"?
- Table 1 and 2. I would argue that the length of a sequence read is measured in bases rather than base-pairs (bp).

Page 6.

- The authors seem to get some non-targeted sequences and taxa. I wouldn't rely on a

machine to classify these. Instead, I would have taken them on a manual BLAST run in GenBank to learn what they are.

- To see a “was then” in the first sentence of the Discussion is somewhat confusing. What was done before that, the reader wonders?

- What do the authors mean by ”The incongruent output retrieved by WIMP is related to the increased representation of closely related sequences in the database…"? After all, densely populated reference databases are normally a source of scientific resolution rather than “incongruent output”?

- “taxonomical” > “taxonomic”, I’d say. Here and elsewhere in the manuscript.

- “BugSeq pipeline” > “The BugSeq pipeline”

- What, exactly, is a “homology search” in this context? Nearly every time a researcher writes “homology search”, what they should be writing is a “sequence similarity search”. Precious few tools can assess homology, and BLAST is not one of them. But sequence similarity, yes, BLAST is good at assessing that.

Page 7.

- “amplicons’ generation” > “amplicon generation”.

- The authors variously use “et al.”, variously “and colleagues”. Is a difference intended? If so, please be more clear on what that difference is and what it is supposed to tell the reader.

- “barcodes’ trimming” > “barcode trimming”.

- “consisted in” > “consisted of” ?

- “Hu *et al.* 2021” is not a good reference for “NCBI”. This one is though: Sayers *et al.,* (2023)[5].

- RefSeq – by which I imagine the authors mean this – is a great resource for type-derived fungal sequences. I use it myself frequently – almost daily, in fact. It comprises several thousand species of fungi. Then again, some 150,000 species of fungi have been described, which means that RefSeq represents <10% of the described species of fungi. It should not be recommended as a taxonomic reference database for metabarcoding. Unite does a better job at that.

- What does it mean that there is “not enough…standards”? When is there enough standards, and how do you know we have reached the point where there are enough standards?

- The second to last paragraph (on biases) comes across as superficial and somewhat out of sync with recent developments in the field (e.g., Tedersoo *et al.,* (2022)[4]).

  Also, surely, ITS copy number must come into play here. I paste, “This potential arises because the ribosomal tandem array occurs at high copy number, which in fungi can range from approximately 45 to 200 copies per genome and span several chromosomes (Maleszka and Clark-Walker 1990; Ganley and Kobayashi 2007).” from Lindner *et al.,* (2013)[6], for

instance.

So when the authors write that they did not fully see the “expected” abundances, I’d argue that it was unknown already from the start what those “expected” abundances were. The targeted species were all likely to have distinct number of ITS copies per cell. Thus, the authors cannot really say much about the extent to which the approach used reflected abundances properly. Amend, Seifert & Bruns (2010)[7] struggled with the same topic, but using pyrosequencing instead.

- Which, finally, brings me to the question of scientific novelty. Given that there are other similar papers out there (e.g., Conti et al., (2023)[8] and Hu et al., (2022)[9]), it’s not clear to me what the authors are bringing to the table. Please position the present findings in relation to previous, similar studies.

- I suppose I also struggled a bit to understand why nanopore sequencing was used here to begin with. After all, the authors sequence a tiny fragment (<1000 bases) by nanopore standards. The authors are bringing a machine gun to a knife fight, and I’m interested to know more on why.

References

1. Taberlet P, Coissac E, Pompanon F, Brochmann C, et al.: Towards next-generation biodiversity assessment using DNA metabarcoding.*Mol Ecol*. 2012; **21** (8): 2045-50 PubMed Abstract | Publisher Full Text
2. Bellemain E, Carlsen T, Brochmann C, Coissac E, et al.: ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases.*BMC Microbiol*. 2010; **10**: 189 PubMed Abstract | Publisher Full Text
3. Ihrmark K, Bödeker IT, Cruz-Martinez K, Friberg H, et al.: New primers to amplify the fungal ITS2 region--evaluation by 454-sequencing of artificial and natural communities.*FEMS Microbiol Ecol*. 2012; **82** (3): 666-77 PubMed Abstract | Publisher Full Text
4. Tedersoo L, Bahram M, Zinger L, Nilsson RH, et al.: Best practices in metabarcoding of fungi: From experimental design to results.*Mol Ecol*. 2022; **31** (10): 2769-2795 PubMed Abstract | Publisher Full Text
5. Sayers EW, Bolton EE, Brister JR, Canese K, et al.: Database resources of the National Center for Biotechnology Information in 2023.*Nucleic Acids Res*. 2023; **51** (D1): D29-D38 PubMed Abstract | Publisher Full Text
6. Lindner DL, Carlsen T, Henrik Nilsson R, Davey M, et al.: Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi.*Ecol Evol*. 2013; **3** (6): 1751-64 PubMed Abstract | Publisher Full Text
7. Amend AS, Seifert KA, Bruns TD: Quantifying microbial communities with 454 pyrosequencing: does read abundance count?. *Mol Ecol*. 2010; **19** (24): 5555-65 PubMed Abstract | Publisher Full Text
8. Conti A, Casagrande Pierantoni D, Robert V, Corte L, et al.: MinION Sequencing of Yeast Mock Communities To Assess the Effect of Databases and ITS-LSU Markers on the Reliability of Metabarcoding Analysis.*Microbiol Spectr*. 2023; **11** (1): e0105222 PubMed Abstract | Publisher Full Text
9. Hu Y, Irinyi L, Hoang MTV, Eenjes T, et al.: Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data.*mBio*. 2022; **13** (2): e0244421 PubMed

Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Mycology, taxonomical bioinformatics, data mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Jul 2023

Cristiano M. Pedroso-Roussado

Dear Dr Nillsson,

First of all thank you for your time to review our manuscript. We are taking all your comments into consideration for the uploading of a new version. Here I will take the chance to just add a note on your last two comments regarding the positioning to two other papers, namely Conti *et al*., 2023, and Hu *et al*., (2022). At the time of performing and writing of the first version of this manuscript these papers were not out so we couldn't have discussed them. This answer also partially answer your question regarding the novelty; the other explanation for bringing a machine gun to a knife fight, we disagree since that for lower equipped labs perhaps it may seem a better management decision to choose for nanopore sequencing rather other technologies - if they want to control the entire process and not rely on third-party analyses. So, our motivation was to walk the path of the democratization of the sequencing methods.

Competing Interests: No competing interests were disclosed.

Reviewer Report 23 March 2023

https://doi.org/10.21956/openreseurope.16741.r30985

© 2023 Benítez-Páez A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Alfonso Benítez-Páez
[1] Host-Microbe Interactions in Metabolic Health Laboratory, Principe Felipe Research Center (CIPF), Valencia, Spain
[2] Host-Microbe Interactions in Metabolic Health Laboratory, Principe Felipe Research Center (CIPF), Valencia, Spain

Pedrosso-Roussado and coworkers present a study assessing a fungal mock community via a long-amplicon sequencing approach. In particular, the authors aimed to evaluate the capacity of nanopore-based sequencing and related analysis methods to delineate fungal communities, which are often neglected in human microbiota assessments. I found this work interesting, but at the same time, it is not new, and several similar studies previously published pursued the same authors' aim with more promising results. The major concerns raised by this peer reviewer after reading this material carefully are stated as follows:

1) I'm wondering why the authors selected the ITS1 and ITS2 regions only to analyze despite the fact that they used a nanopore-base sequencing method with no amplicon size limitation like conventional SBS-based (sequencing-by-synthesis, Illumina). The amplification of larger segments of the ribosome RNA encoding operon leads to a better determination of the species composition in complex communities just because including a wide array of hypervariable regions to evaluate. The evaluation of a short region of less than one kbp in length is, in my opinion, a waste of the sequencing capacity the authors have with their nanopore-sequencing platform.

Furthermore, the current direction and evolution of this particular assay imply the study of the whole fungal ribosome RNA operon, which has been demonstrated to be more informative in delineating environmental- or host-associated mycobiota[1-3]. Consequently, the authors should better argue the use of this sequencing configuration and highlight what advances they bring up compared to previous investigations.

2) The results described by the authors included that their method configuration could efficiently reconstitute the occurrence of species in the fungal mock community assessed. Nevertheless, the sequencing protocol and data analysis failed to retrieve the theoretical abundance of every species in the mock community. The fungal ribosome RNA encoding operons are arranged in tandem repeats manner[1], which implies that copy number variation could be an additional discriminate variable to pay attention to. I've revised the material and found no reference to this matter. Suppose the species present in the mock community have a different ribosome RNA operons copy number than assumed theoretically. In that case, this could be the basis for the discrepancies in the relative abundance retrieved after analysis of the sequencing output. Is the genome of such species sequenced? If not (even if the genome is drafted and prone to misassembly by the repeated structure of operons), a qPCR (absolute quantification) of target

markers against a single copy housekeeping gene could help solve this inconsistency. A digital-droplet PCR can help in this issue as well. The coverage values of primers used should also be described in M&M section[4].

3) The authors employed the manufacturer solution to asses in real-time the composition of the mock community. Moreover, they also used an alternative method developed by Fan and coworkers[5]. Unfortunately, to my best knowledge, none of the strategies used are aligned with the quasi-standards reached in this particular field involving comprehensive databases. For instance, the central repository for this specific case would be the UNITE database[6], which compiles a vast amount of genomic information from eukaryote ITS regions. Alternatively, the database created for nanopore sequencing-based applications called FRODO could also be helpful [2], given this was concretely developed to adapt long-read protocols and improve the characterization of fungal communities. In both cases, using advanced and high-performance mapping strategies such as "minimap2" with a specific configuration for ONT reads will drastically improve the study of fungi species through the molecular protocols described in this work towards exploring more complex and real communities.

References

1. Wurzbacher C, Larsson E, Bengtsson-Palme J, Van den Wyngaert S, et al.: Introducing ribosomal tandem repeat barcoding for fungi.*Mol Ecol Resour*. 2019; **19** (1): 118-127 PubMed Abstract | Publisher Full Text
2. Lu J, Zhang X, Zhang X, Wang L, et al.: Nanopore sequencing of full rRNA operon improves resolution in mycobiome analysis and reveals high diversity in both human gut and environments. *Mol Ecol*. 2022. PubMed Abstract | Publisher Full Text
3. D'Andreano S, Cuscó A, Francino O: Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples.*Biol Methods Protoc*. 2021; **6** (1): bpaa026 PubMed Abstract | Publisher Full Text
4. Banos S, Lentendu G, Kopf A, Wubet T, et al.: A comprehensive fungi-specific 18S rRNA gene sequence primer toolkit suited for diverse research issues and sequencing platforms.*BMC Microbiol*. 2018; **18** (1): 190 PubMed Abstract | Publisher Full Text
5. Fan J, Huang S, Chorlton SD: BugSeq: a highly accurate cloud platform for long-read metagenomic analyses.*BMC Bioinformatics*. 2021; **22** (1): 160 PubMed Abstract | Publisher Full Text
6. Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, et al.: The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications.*Nucleic Acids Res*. 2019; **47** (D1): D259-D264 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human microbiome, nanopore-based sequencing, computational biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 27 Mar 2023

Cristiano M. Pedroso-Roussado

The authors come in this way to write a response to Alfonso Benitez-Paez comments, which we truly appreciate.

First of all the study presented is not a long-amplicon sequencing approach. Despite using a sequencing approach based on nanopore sequencing, the intent of this study was to evaluate its preciseness and accuracy to sequence short-reads, as stated in the abstract.

Second, we agree that other studies are published in the same space but their fragmentation motivated us to publish our work. Specifically, there is no standard to either upstream or downstream protocols to nanopore sequencing. Moreover, the novelty of this work relates to its results wwhere the data analysis recommended by Nanoporetech showed an unexpected discrepant taxonomical classification (Figure 1). We also agree that other studies have more promising results, but we disagree that the judgement of the results should be a parameter to evaluate whatever results. In this way, our results are new, and we believe they must be considered as that.

Next, we comment Alfonso's comment by order of appearance:

1) The goal of our work was to assess the capacity of nanopore sequencing to analyse short-reads. Since this technology has been enunciated as a promise to revolutionize the clinical practice, as a faster and easier approach, our study wanted to develop a faster and easier protocol to study fungal species. Therefore, we don't see the point of compare our study to and it was not aimed for. Still, we agree that the way forward is to analyse longer sequences.

2) We agree with the fact that nanopore sequencing approach we tested failed to retrieve the correct theoretical abundance of every species in the mock community. We also agree with this comment and the next work should take the Alfonso's advice into consideration.

3) In accordance with our response in 1), our goal was to deliver a fast and easier protocol based on nanopore sequencing. This implies the application of standalone solutions to data analyses. Therefore, we assumed that more complex bioinformatic tools were not practical for the context we developed this study for - a faster and easier protocol to study fungal species using nanopore sequencing. As commented by Alfonso, the quasi-standards exist, but they do not represent a satisfactory pipeline. Since our study aimed the study if ITS sequences, the standards for such short reads are not enough to be stated as 'standards', and this justified the discloser of our results, in our view. Still, we deep appreciate the suggestions about the use of more comprehensive databases, which we agree.

Final comment - In our view, we still believe that our study deserves to be accepted since both points raised in 1) and 3) do not directly address the aim of our study. However, we agree that the assessment of the copy number variations can explain the discrepancies found in the fungal mock community species' abundance, and it was not considered as a variable during the work.

Finally, we want to thank Alfonso Benitez-Paez for the time and care of the comments made.

Competing Interests: No competing interests were disclosed.

Reader Comment 25 Apr 2023

Sam Chorlton

Hi Dr. Benítez-Páez,

Thank you for your comments specifically around the bioinformatic methods. I am PI of the BugSeq method[1] cited in your review and also cited in this article, and therefore wanted to specifically address some points around bioinformatic methods. I also have a competing interest as detailed at the bottom.

First, I wholeheartedly agree that following best practices are paramount to achieving top accuracy and transparent results.

With regards to the BugSeq method, as published in our original article[1], BugSeq performs minimap2 alignment of reads using the “map-ont” preset against a reference database, exactly as you suggest in your comment. We also adjust the analysis based on data characteristics; for example, we detect full length 16S nanopore data based on read length and content, and process this data with a dedicated full length 16S pipeline, which has been shown to perform superiorly to alternative methods.[2]

Regarding the reference database, BugSeq offers several reference databases to our users for selection upon data submission, including a curated high quality metagenomic database and the NCBI nt database. The NCBI nt database includes a large compendium of fungal genomes and ITS sequences, and it can be freely downloaded from NCBI.

From my understanding, the cited FRODO database[3] was generated by extracting operons with full 18S and 28S rDNA sequences from public genomes including NCBI, JGI, FungiDB, Ensembl Fungi and Broad Institute. Therefore, although the NCBI nt database is probably overkill for an ITS analysis, due to differences in database update frequency (FRODO has not been updated since 2021), stricter inclusion of sequences in FRODO (NCBI could include shorter but otherwise unrepresented sequences) and other factors (NCBI gets data from some of the additional sources, eg. JGI), these databases potentially contain comparable ITS content. While the UNITE database is the standard in the field, lack of taxonomy database dumps similar to NCBI taxdumps precludes easy manipulation of sequence classifications; for example, BugSeq performs lowest common ancestor identification for equivalently good alignments to improve confidence in classification, as detailed our previously cited manuscript.[1]

Furthermore, as you allude to in your comment, databases such as UNITE, which often only contain the short ITS region and not surrounding regions, may no longer be suitable for the longer amplicons enabled with third-generation sequencers. While not directly evaluated with ITS, a recent preprint suggests that using metagenomic databases may surprisingly perform superiorly for amplicon sequencing data as they contain the variation of each copy of an operon within a single genome, whereas databases such as UNITE may collapse these slight variations.[4] Supporting this finding are equivalent performance of NCBI nt and UNITE in separate work.[5] Future work on our platform may incorporate the UNITE or other databases as explicit analysis options to meet the needs of the scientific community.

I hope these points alleviate some of your concerns for the BugSeq method evaluated in this manuscript and demonstrate an alignment of BugSeq's methods with quasi-best practices to achieve optimal accuracy (as demonstrated in this independent manuscript) in an easy and user-friendly manner (as highlighted in a response to your review from the study authors).

Please do not hesitate to reach out if you have any other questions/comments on the methods.

Best, Sam Chorlton MD D(ABMM)

1. Fan, J., Huang, S. & Chorlton, S. D. BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. BMC Bioinformatics 22, 160 (2021).

2. Jung, A. & Chorlton, S. D. BugSeq 16S: NanoCLUST with Improved Consensus Sequence Classification. bioRxiv 2021.03.16.434153 (2021) doi:10.1101/2021.03.16.434153.

3. Lu, J. et al. Nanopore sequencing of full rRNA operon improves resolution in mycobiome analysis and reveals high diversity in both human gut and environments. Molecular Ecology.

4. Faits, T., Odom-Mabey, A. R., Castro-Nallar, E., Crandall, K. A. & Johnson, W. E. Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data. 2022.07.27.501757 Preprint at https://doi.org/10.1101/2022.07.27.501757 (2022).

5. Rolling, T., Zhai, B., Frame, J., Hohl, T. M. & Taur, Y. Customization of a DADA2-based pipeline for fungal internal transcribed spacer 1 (ITS1) amplicon data sets. JCI Insight 7, (2022).

Competing Interests: SDC is a shareholder and employee of BugSeq Bioinformatics Inc.