DNA barcoding of fungal specimens using long-read high-throughput sequencing

Running title: High-throughput sequencing fungal specimens

Kadri Runnel[1,2*], Kessy Abarenkov[2,3], Ovidiu Copoț[1], Vladimir Mikryukov[1], Urmas Kõljalg[1,3], Irja Saar[1], Leho Tedersoo[2,4]

[1] Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

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

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

[4] College of Science, King Saud University, Riyadh, Saudi Arabia

Abstract

Molecular methods are increasingly used to identify species that lack conspicuous macro- or micromorphological characters. Taxonomic and ecological research teams barcode large numbers of collected voucher specimens annually. In this study we assessed the efficiency of long-read high throughput sequencing (HTS) as opposed to the traditionally used Sanger method for taxonomic identification of multiple vouchered fungal specimens, and providing reference information about intra-individual allele polymorphism. We developed a workflow based on a test-set of 423 fungal specimens (representing 205 species), PacBio HTS method, and ribosomal rRNA operon internal transcribed spacer (ITS) and 28S rRNA gene (LSU) markers. PacBio HTS had a higher success rate than Sanger sequencing at a comparable cost. Species identification based on PacBio reads was usually straightforward, because the dominant operational taxonomic unit (OTU) typically represented the targeted organism. Unlike the Sanger method, PacBio HTS enabled detecting widespread allele polymorphism within the ITS marker in the studied specimens. We conclude that multiplex DNA barcoding of the fungal ITS and LSU markers using a PacBio HTS is a useful tool for taxonomic identification of large amounts of collected voucher specimens at competitive price. Furthermore, PacBio HTS accurately recovers various alleles, which can provide crucial information for species delimitation and population-level studies.

Keywords

allele polymorphism, intragenomic diversity, species identification, long-read high-throughput sequencing

Introduction

A large proportion of living organisms can reliably be identified to species only based on molecular methods. Collected and vouchered specimens constitute a permanent record of biological diversity and form the basis of taxonomy as they provide material for description of new species and reference for taxonomic identification. The modern species identification relies increasingly on informative marker genes – termed as DNA barcodes (Hebert et al., 2003) – and sometimes on the entire genomes of organisms (Coissac et al., 2016; Misas et al., 2020). Because the DNA of non-living specimens degrades rapidly (Taylor & Swann, 1994), it is feasible to retrieve the molecular information in a reasonable time frame (usually within a few years following collection) to prevent loss of valuable genetic information. Large research teams of ecologists and taxonomists typically collect hundreds to tens of thousands of specimens annually, which necessitates efficient bulk analysis of large amounts of specimens in terms of cost, time, and labor (Hebert et al., 2018; Srivathsan et al., 2018).

DNA barcoding for taxonomic identification is traditionally performed using the Sanger sequencing method. Typically, one or several marker gene fragments are amplified and sequenced, allowing to produce up to 1000 base pair high-quality reads in a single pass, depending on the purity of DNA. The Sanger sequencing technology relies on chain termination signal averaged across all amplicons, and produces a consensus read of sequences of several marker gene alleles from the target specimen (Sanger et al., 1977). A well-known limitation of this approach is that the consensus read blurs the evolutionary signal among alleles. Additional technical shortfalls include low quality in the beginning of the sequence, disruption of reads in the case of length polymorphism in alleles, generation of ambiguous signal in the case of nucleotide substitutions, and failure to produce a readable

63 sequence if both the target and co-occurring organisms (e.g. contaminants) are amplified or
64 when the amplicon is of low quantity or purity (Hyde et al., 2013).
65
66 To overcome the issues with Sanger sequencing, high-throughput sequencing (HTS) methods
67 can alternatively be used for DNA barcoding (Coissac et al., 2016, Bohmann et al., 2020).
68 These methods include HTS-based analysis of single or multiple marker genes, genome
69 skimming (i.e., low-coverage genome sequencing to retrieve long contigs of marker genes),
70 and whole-genome sequencing. Genome-based methods are useful for obtaining full-length
71 marker genes for accurate identification and phylogenetic analyses, but they have a low
72 capacity to phase alleles differing by a few substitutions or indels >500 bases apart (Coissac
73 et al., 2016; Tedersoo et al., 2016). In addition, such methods are relatively costly, because
74 each sample requires preparation of a specific library and assembly from hundreds of
75 thousands to tens of millions of reads. Furthermore, no more than one or two marker genes
76 are still broadly used for DNA barcoding, except multi-gene phylogenetic analyses. Short-
77 read HTS platforms such as Illumina, Ion Torrent and DNBseq provide high-quality reads for
78 DNA fragments <550 base pairs, which may be insufficient for reliable identification and
79 phylogenetic analyses. In comparison, long-read and synthetic long-read sequencing methods
80 offer great promise for analysis of DNA markers up to ca. 3500 base pairs (Hebert et al.,
81 2018; Callahan et al., 2021; Karst et al., 2021). Despite high raw error rate, these methods are
82 highly accurate when calculating consensus (built-in option for PacBio and synthetic long-
83 reads). Long-read methods also enable to phase haplotypes, which is of great relevance in
84 population-level research (Tedersoo et al., 2021). Although these methods are relatively
85 costly, hundreds to thousands of samples can be multiplexed for a single run, bringing the
86 overall costs comparable to Sanger sequencing and one to three orders of magnitude less
87 compared with genome-based approaches (Hebert et al., 2018; Srivathsan et al., 2018).

The main purpose of this study was to assess the efficiency of long-read HTS for taxonomic identification of large sets of vouchered fungal specimens and providing reference information about intra-individual allele polymorphism. The latter is important to avoid describing artefactual "shadow taxa" (i.e., based on sequencing artefacts, rare alleles and pseudogenes; Thines et al., 2018; Porter & Hajibabaei, 2021) and inflating HTS-based biodiversity estimates. Here we developed a workflow based on PacBio multiplex DNA barcoding method and a test set of hundreds of fungal amplicons using the ribosomal rRNA operon internal transcribed spacer (ITS) and 28S rRNA gene (LSU) markers. These are the two most commonly used DNA barcodes for taxonomic identification of fungi and many protist groups (Pawlowski et al., 2012; Schoch et al., 2012). We demonstrate that the benefits of this multiplex DNA barcoding method include better recovery of low-quality samples and higher read quality compared with Sanger sequencing, as well as retrieval of multiple alleles differing by a single or more base pairs at a comparable cost.

Materials and methods

Molecular analyses

To test the relative performance of Sanger and PacBio sequencing on fruiting body samples, we compiled 423 vouchered specimens (polyporoid, resupinate, and agaricoid fruiting body types) belonging to 205 species from the fungarium of Natural History Museum of Tartu University (acronym TUF; Appendix 1). Most fruiting body samples were collected between 2015 and 2020 (Table S1) for different ecological and taxonomic studies.

The DNA of specimens was extracted from roughly 0.1-1 mg dried material using ammonium sulphate lysis buffer (Anslan & Tedersoo, 2015) that provides sufficient DNA quality from fruiting body material (Tedersoo et al., 2016). For all samples, our DNA barcoding approach targeted the ITS region, and for one batch of samples we also analyzed the LSU. To cover the entire ITS region, we chose the primers ITS1catta (5'-ACCWGCGGARGGATCATTA-3') and ITS4ngsUni (5'-CCTSCSCTTANTDATATGC-3') for PCR. The ITS1catta primer has a high affinity to Dikarya and it avoids amplification of the common intron in the 3' end of the 18S region (Tedersoo & Anslan, 2019). Both primers were tagged with one of the 115, 12-base, sample-specific indices (Tedersoo & Anslan, 2019). PCR was carried out in two replicates in the following thermocycling conditions: an initial 15 min at 95 °C, followed by 30 cycles of 95 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min, and a final cycle of 10 min at 72 °C. PCR products from replicate samples were pooled and their relative quantity was estimated by running 5 µl DNA on 1% agarose gel for 25 min. DNA samples with no visible bands were re-amplified with 35 cycles. To retrieve the LSU, we also amplified the DNA from a subset of 75 samples using the untagged primer LROR (5´-ACCCGCTGAACTTAAGC-3´) and tagged primer LR5 (5´-TCCTGAGGGAAACTTCG-3´) and the above-described options. In total, our test set yielded 497 fungal amplicons.

The PCR products were checked on a 1% agarose gel and the relative strength of the band was recorded at the scale of 0 (no band) to 5 (very strong band). It was also recorded whether there were a single or multiple bands on the gel. Altogether 20 µl of the PCR products were subjected to purification using Exo-SAP enzymatic treatment (Tedersoo et al., 2006) and shipped for Sanger sequencing in Macrogen, Inc., the Netherlands, using a single-pass with the untagged ITS4ngsUni primer (or LROR primer for LSU). Of the remaining PCR

138 products, between 1 and 10 µl of amplicon were taken based on the strength of the band on
139 the gel (categories 0 and 1, 10 µl; category 5, 1 µl), and pooled into five sequencing libraries
140 for the ITS region and one library for the LSU region. The amplicon pools were purified
141 using FavorPrep™ Gel/PCR Purification Kit (Favorgen-Biotech Corp., Austria), following
142 the manufacturer's instructions, and subjected to SMRTbell library preparation and PacBio
143 Sequel II sequencing on a single 8M SMRT cell in the University of Oslo, Norway.
144
145
146 **Sequence data workflow**
147 The raw data were obtained in ab1 format for Sanger sequences and fastq format for PacBio
148 sequences. Sanger sequences were inspected for quality using Sequencher v. 5.4.6 software.
149 Sanger sequences were manually checked and trimmed to comprise only the full ITS region
150 or LSU, or a shorter fragment in case of quality issues. PacBio reads were demultiplexed with
151 Lima v.2.4.0 (https://lima.how/; PacBio, 2021) and quality-filtered following Tedersoo et al.,
152 (2021). Sequences were trimmed to remove primers using cutadapt v.3.5 (Martin, 2011) and
153 ITS region was extracted using ITSx v.1.1.3 (Bengtsson-Palme et al., 2013). ITS and LSU
154 sequence data were processed separately using seqkit v.2.1.0 (Shen et al., 2016) and
155 VSEARCH v.2.18.0 (Rognes et al., 2016). Reads were grouped into 100% sequence-
156 similarity OTUs. Putative index-switches (also known as tag-jumps) were removed from the
157 OTU table based on the UNCROSS2 score (Edgar, 2018).
158
159 The representative reads of each PacBio OTU and all Sanger sequences were identified
160 against UNITE v.9.4b (Nilsson et al., 2018) and SILVA v.138.1 (Quast et al., 2013),
161 respectively, using the BLASTn algorithm and 10 best database hits. Based on morphological
162 identification of specimens, the PacBio OTUs and Sanger sequences were flagged as

potentially matching or potentially mismatching to the target taxa. The latter category suggested sequencing of naturally associated fungi, airborne or laboratory contaminants. We did not attempt to distinguish among these groups of unexpected taxa. For a comparison with Sanger sequencing, we only considered non-singleton OTUs with at least 1% relative abundance as putatively true alleles, while others were considered low-quality reads. For each sequenced specimen, we recorded the following properties: the overlap of PacBio OTUs with Sanger sequences from the same specimen (binary), the number of all non-singleton >1%-abundance PacBio OTUs and the number of potentially matching PacBio OTUs.

To calculate pairwise distances within specimens and species, sequences for each sample were aligned using mafft v7.487 (Katoh et al., 2002). Using the dist.seqs command of mothur v1.46.1 (Schloss et al., 2009), we calculated the mean and maximum uncorrected pairwise distances of polymorphic alleles within each sample sequenced for the ITS region. We also recorded corresponding intraspecific distances for 11 species that were successfully sequenced in >5 samples. The samples sequenced for LSU were excluded, because these yielded less data and only a small proportion of samples sequenced for LSU had allele polymorphisms. In allelic comparisons, each pairwise distance represents the percent of mismatches (including indels), where gaps of any length were penalized once and end gaps were ignored.

We also estimated the reasons of sequencing failure. Sanger sequencing was considered failed when the read (1) represented a contaminant; (2) had an unreadable chromatogram; (3) had >5 ambiguous bases; or (4) had >50 bases were missing from either end (due to low-quality 5’ or 3’ end or sequence disruption by length polymorphism of alleles). The PacBio

sequencing was considered unsuccessful if the sample (1) yielded no sequence reads; (2) did not contain a correct fungus; or (3) the relative abundance of correct fungus was <10%.

Statistical analyses

We ran three sets of binomial regression models (logarithmic link function) using the glm function in the R base package. To test whether PacBio sequencing success is more probable for some particular cases of Sanger sequencing failure, we focused on samples that could not be successfully sequenced using Sanger method, and ran a model where PacBio success was a binary dependent variable and Sanger failure reason the sole categorical explanatory variable (four levels, see above). To test if the sequencing success across all samples depended on the relative strength of the PCR band or sample age (years since sampling), we ran models where either Sanger or PacBio success was a binary dependent variable and, respectively, the relative strength of PCR band or sample age was a sole continuous explanatory variable.

Results

For our samples, the total sequencing costs (incl. library preparation) were ca. 10% less for PacBio than Sanger method. For PacBio, 387 out of 497 samples (77.9%) were successfully sequenced compared with 275 samples when using Sanger (55.3%). Altogether 122 (25%) of all samples could be successfully sequenced with PacBio but not with Sanger, whereas the opposite happened to 15 (3.0%) of samples (Fig 1A). In general, there was a higher probability of successful PacBio outcome in cases where Sanger sequencing failed because of yielding only a partial readable sequence or a sequence containing some low-quality regions

(Table 1). The probability for both Sanger and PacBio success increased with the relative strength of the PCR band (Table 1, Fig 2A). The probability of PacBio sequencing success but not Sanger sequencing success significantly decreased with sample age (Table 1, Fig 2B). In 98% of the samples successfully sequenced with PacBio, the true target species was represented by the most abundant OTU.

PacBio sequencing revealed ITS allele polymorphisms in 249 (75%) of samples that were successfully sequenced. These samples contained on average 5.1 (range 2 to16) polymorphic alleles with >1% relative abundance. Each allele generally differed from all others by only a few base pairs, yielding an average intraindividual distance of 0.44% (range, 0.15% to 2.88%) across all samples with allele polymorphisms. The average maximum distance across all samples with allele polymorphisms was 0.64% (range, 0.15% to 2.88%) (Fig 3B). The relative abundance of the most abundant allele and the total number of sequences in the sample were weakly correlated (Spearman r = 0.13, Fig 3A).

We also checked among-individual, intraspecific variation for 11 species that were successfully sequenced in >5 samples (Table 2). Those species had in average 2.5 polymorphic alleles per sequenced sample (min 0.29, max 5.17). The average and maximum intraspecific distance remained below 2% in all cases. The polypore *Sidera vulgaris* displayed the highest average and maximum intraspecific differences.

Discussion

1. Sequencing success

We demonstrated that sequencing of nearly 500 fungal amplicons using long-read HTS sequencing (PacBio) had a higher success rate compared to the traditional Sanger sequencing at comparable cost. Identifying the target species from PacBio samples was usually straightforward, as these were represented by the dominant OTU. However, contaminants were common and sometimes prevailed, especially in relatively old specimens. This indicates that taxonomic knowledge remains important when interpreting sequencing results (see Vu et al., 2018). We observed a rapid decay in PacBio sequencing success rate in >5-year-old specimens. Earlier studies have noticed a similar trend in Sanger sequencing (Larsson & Jacobsson, 2004). In the case of Sanger, we used single-end sequencing (as most taxonomic studies), additional sequencing of the complementary strand would have increased the success rate in samples with partial sequences, and low quality regions (Hyde et al., 2013).

There were two typical situations where the Sanger sequencing failed, but PacBio proved successful: length polymorphism of reads and field-contaminated samples. The length polymorphism in alleles caused disruption of Sanger reads but yielded no issues in PacBio. For example, in the boreal polypore species *Sidera vulgaris,* Sanger sequencing recovered a full-length read in 7% of the sequenced specimens, whereas PacBio was successful in 97% of specimens. The field-contamination was common in collections of the resupinate tropical *Tomentella* spp. and polypore *Rhodonia placenta*. In *R. placenta*, Sanger sequencing and PacBio sequencing were successful in 27% and 91% out of 11 specimens analyzed, respectively. It is worth noticing that we also adopted relatively stringent criteria for successful sequencing. If fruit body characteristics allowed restricting the species identification to a few options only, many of the “sequencing failures” were still informative for confirming the final identification.

264

265 **2. Allele polymorphism in studied samples**

266 Our workflow pointed to a widespread allele polymorphism within the ITS marker in the
267 studied fungal collections. We consider most of the common variants as "true" alleles, since
268 PacBio circular consensus sequencing yields high sequencing accuracy (Karst et al., 2021;
269 Tedersoo et al., 2021) and we only addressed OTUs represented by >1 read and at least 1%
270 total abundance, hence avoiding random PCR and sequencing errors (see also Ganley &
271 Kobayashi, 2007). Our interpretation contrasts to Lindner et al., (2013), who ascribed
272 intraindividual variation in a majority of 100 sampled fungal species to PCR and sequencing
273 errors that were an order of magnitude more common in the now-obsolete 454
274 pyrosequencing technology (Lindner et al., 2013). However, the alleles typically differed by a
275 single or few positions, and the maximum intragenomic and intraspecific differences
276 remained <2% (except in three samples). Metabarcoding studies typically use a 97%, 98% or
277 98.5% ITS sequence similarity threshold for species-level separation and identification of
278 taxa. Therefore, the observed differences typically remain within this threshold, especially
279 when compared to the closest read (single-linkage clustering) or centroid (greedy clustering)
280 based methods. However, when combined with geographic distance (population divergence),
281 inappropriate clustering methods and accumulating sequencing errors, intraspecific
282 differences may indeed account for artefactual, elevated richness in ecological studies.

283

284 Our results suggest that the exact sequence variant (ESV) based approaches (Callahan et al.,
285 2017) are not optimal for species-level metabarcoding analyses of fungal diversity (see also
286 Estensmo et al., 2021; Tedersoo et al., 2022) and perhaps eukaryotes in general (Antich et al.,
287 2021; Porter & Hajibabaei, 2021), by potentially retrieving artefactual taxa. In conclusion,
288 multiplex DNA barcoding of the fungal ITS marker using a PacBio third-generation HTS

289 protocol is a useful tool for taxonomic assessment of large sets of vouchered fungal
290 specimens. Besides costs comparable to Sanger sequencing, PacBio HTS provides more
291 complete and accurate recovery of various alleles, which can potentially be accounted for in
292 bordering the molecular species or species hypotheses (Kõljalg et al., 2013) and used in
293 population-level studies (Byrne et al., 2017).
294
295

296 Acknowledgements

297 The study was funded by the State Forest Management Centre (project "Enhancing the
298 conservation performance of protected forest fragments" to K.R.), the Estonian Research
299 Council (grants PRG632 and PRG1170), and the European Regional Development Fund
300 (Centre of Excellence EcolChange). Rasmus Puusepp performed the laboratory work.
301
302

303 References

304
305 Anslan, S., & Tedersoo, L. (2015). Performance of cytochrome c oxidase subunit I (COI), ribosomal
306 DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of
307 Collembola. *European Journal of Soil Biology*, 69, 1–7.
308
309 Antich, A., Palacin, C., Wangensteen, O.,S., & Turon, X. (2021). To denoise or to cluster, that is not
310 the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC*
311 *Bioinformatics*, 22, 1–24.
312
313 Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P.,
314 Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A., Jumpponen, A., Unterseher, M.,
315 Kristiansson, E., Abarenkov, K., Bertrand, Y.J.K., Sanli, K., Eriksson, K.M., Vik, U., Veldre, V., &
316 Nilsson, R.H. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal
317 ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods*
318 *in Ecology and Evolution*, 4, 914–919. https://doi.org/10.1111/2041-210X.12073
319
320 Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M.T.P. (2020). Beyond DNA barcoding: The
321 unrealized potential of genome skim data in sample identification. *Molecular Ecology*, 29, 2521–
322 2534.
323

Byrne, A. Q., Rothstein, A. P., Poorten, T. J., Erens, J., Settles, M. L., & Rosenblum, E. B. (2017). Unlocking the story in the swab: A new genotyping assay for the amphibian chytrid fungus *Batrachochytrium dendrobatidis*. *Molecular Ecology Resources*, 17, 1283–1292.

Callahan, B.J., Grinevich, D., Thakur, S., Balamotis, M.A., & Yehezkel, T.B. (2021). Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome*, 9, 1–13.

Callahan, B.J., McMurdie, P.J., & Holmes, S.P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11, 2639–2643.

Coissac, E., Hollingsworth, P., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology*, 25, 1423–1428.

Edgar, R.C. (2018). UNCROSS2: identification of cross-talk in 16S rRNA OTU tables, *BioRxiv*, p.400762. https://doi.org/10.1101/400762

Estensmo, E.L., Maurice, S., Morgado, L., Martin-Sanchez, P.M., Skrede, I., & Kauserud, H. (2021). The influence of intraspecific sequence variation during DNA metabarcoding: a case study of eleven fungal species. *Molecular Ecology Resources,* 21, 1141–1148.

Ganley, A.R., & Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome research,* 17, 184–191.

Hebert, P.D.N., Cywinska, A., Ball, S.L., & deWaard, J.R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London*, 270, 313–321.

Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., deWaard, J.R., Ivanova, N.V., Janzen, D.H., Hallwachs, W., Naik, S., Sones, J.E., & Zakharov, E.V. (2018). A sequel to Sanger: amplicon sequencing that scales. *BMC Genomics*, 19, 1–14.

Hyde, K.D., Udayanga, D., Manamgoda, D.S., Tedersoo, L., Larsson, E., Abarenkov, K., Bertrand, Y., Oxelman, B., Hartmann, M., Kauserud, H., Ryberg, M., Kristiansson, E., & Nilsson, R.H. (2013). Incorporating molecular data in fungal systematics: a guide for aspiring researchers. *Current Research in Environmental & Applied Mycology*, 3, 1–32.

Karst, S.M., Ziels, R.M., Kirkegaard, R.H., Sørensen, E.A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with nanopore or PacBio sequencing. *Nature Methods*, 18, 165–169.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30, 3059–3066. doi:10.1093/nar/gkf436

Kõljalg, U., Nilsson, R.H., Abarenkov, K., Tedersoo, L., Taylor, A.F., Bahram, M., Bates, S.T., Bruns, T.D., Bengtsson-Palme, J., Callaghan, T.M., & Douglas, B. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, 22, 5271–5277.

Larsson, E., & Jacobsson, S. (2004). Controversy over *Hygrophorus cossus* settled using ITS sequence data from 200 year-old type material. *Mycological Research*, 108, 781–786.

Lindner, D.L., Carlsen, T., Henrik Nilsson, R., Davey, M., Schumacher, T., & Kauserud, H. (2013). Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution*, 3, 1751–1764.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. (2011) *EMBnet. Journal,* 17, 10–12.

Misas, E., Gómez, O., Botero, V., Muñoz, J., Teixeira, M., Gallo, J., Clay, O., & McEwen, J. (2020). Updates and comparative analysis of the mitochondrial genomes of *Paracoccidioides* spp. using Oxford Nanopore MinION sequencing. *Frontiers in Microbiology*, 11, 1751.

Nilsson, R.H., Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F.O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2018). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, 47:D1, D259–D264.

Pawlowski, J., Audic, S., & Adl, S. (2012). CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology* 10, e1001419.

Porter, T.M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics*, 22, 1–20.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41:D1, D590–D596. https://doi.org/10.1093/nar/gks1219

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4, e2584. https://doi.org/10.7717/peerj.2584

Sanger, F., Nicklen, S., & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA*, 74:5463–5467

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS One*, 11(10): e0163962. https://doi.org/10.1371/journal.pone.0163962

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., Crous, P.W., Miller, A.N., Wingfield, M.J., Aime, M.C., An, K.D., Bai, F.Y., Barreto, R.W., Begerow, D., Bergeron, M.J., Blackwell, M., Boekhout, T., Bogale, M., Boonyuen, N., Burgaz, A.R., Buyck, B., Cai, L., Cai, Q., Cardinali, G., Chaverri, P., Coppins, B.J., Crespo, A., Cubas, P., Cummings, C., Damm, U., de Beer, Z.W., de Hoog, G.S., Del-Prado, R., Dentinger, B., Dieguez-Uribeondo, J., Divakar, P.K., Douglas, B., Duenas, M., Duong, T.A., Eberhardt, U., Edwards, J.E., Elshahed, M.S., Fliegerova, K., Furtado, M., Garcia, M.A., Ge, Z.W., Griffith, G.W., Griffiths, K., Groenewald, J.Z., Groenewald, M., Grube, M., Gryzenhout, M., Guo, L.D., Hagen, F., Hambleton, S., Hamelin, R.C., Hansen, K., Harrold, P., Heller, G., Herrera, G., Hirayama, K., Hirooka, Y., Ho, H.M., Hoffmann, K., Hofstetter, V., Hognabba, F., Hollingsworth, P.M., Hong, S.B., Hosaka, K., Houbraken, J., Hughes, K., Huhtinen, S., Hyde, K.D., James, T., Johnson, E.M., Johnson, J.E., Johnston, P.R., Jones, E.B., Kelly, L.J., Kirk, P.M., Knapp, D.G., Kõljalg, U., Kovacs, G.M., Kurtzman, C.P., Landvik, S., Leavitt, S.D., Liggenstoffer, A.S., Liimatainen, K., Lombard. L., Luangsa-Ard, J.J., Lumbsch, H.T., Maganti, H., Maharachchikumbura, S.S., Martin, M.P., May, T.W., McTaggart, A.R., Methven, A.S., Meyer, W., Moncalvo. J.M., Mongkolsamrit, S., Nagy, L.G., Nilsson, R.H., Niskanen, T., Nyilasi, I., Okada, G., Okane, I., Olariaga, I., Otte, J., Papp, T., Park, D., Petkovits, T., Pino-Bodas, R., Quaedvlieg, W., Raja, H.A., Redecker, D., Rintoul, T., Ruibal, C., Sarmiento-Ramirez, J.M., Schmitt, I., Schussler, A., Shearer, C., Sotome, K., Stefani, F.O., Stenroos, S., Stielow, B., Stockinger, H., Suetrong, S., Suh, S.O., Sung, G.H., Suzuki, M., Tanaka, K., Tedersoo, L., Telleria, M.T., Tretter, E., Untereiner, W.A., Urbina, H., Vagvolgyi, C., Vialle, A., Vu, T.D., Walther, G., Wang, Q.M., Wang, Y., Weir, B.S., Weiss, M.,

White, M.M., Xu, J., Yahr, R., Yang, Z.L., Yurkov, A., Zamora, J.C., Zhang, N., Zhuang, W.Y., & Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences USA*, 109, 6241–6246.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., & Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75, 7537–7541.

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W., Bertrand, D., Ng, A., Boey, E., Koh, J., Nagarajan, N., & Meier, R. (2018). A MinIONTM-based pipeline for fast and costeffective DNA barcoding. *Molecular Ecology Resources*, 18, 1035–1049.

Taylor, J.W., & Swann, E.C. (1994). DNA from herbarium specimens. In B., Herrmann, & S., Hummel (Eds.), *Ancient DNA* (pp. 166–181). Springer.

Thines, M., Crous, P.W., Aime, M.C., Aoki, T., Cai, L., Hyde, K.D., Miller, A.N., Zhang, N., & Stadler, M. (2018). Ten reasons why a sequence-based nomenclature is not useful for fungi anytime soon. *IMA Fungus*, 9, 177–183.

Tedersoo, L., Albertsen, M., Anslan, S., & Callahan, B. (2021). Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and Environmental Microbiology*, 87, e00626-21.

Tedersoo, L., & Anslan, S. (2019). Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. *Environmental Microbiology Reports*, 11, 659−668.

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, H., Kennedy, P., Yang, T., Anslan, S., & Mikryukov, V. (2021). Best practices in metabarcoding of fungi: from experimental design to results. *Molecular Ecology*, pending revision.

Tedersoo, L., Liiv, I., Kivistik, P.A., Anslan, S., Kõljalg, U., & Bahram, M. (2016). Genomics and metagenomics technologies to recover ribosomal DNA and single-copy genes from old fruitbody and ectomycorrhiza specimens. *MycoKeys*, 13, 1-20.

Tedersoo, L., Suvi, T., Larsson, E., & Kõljalg, U. (2006). Diversity and community structure of ectomycorrhizal fungi in a wooded meadow. *Mycological Research*, 110, 734–748. doi:10.1016/j.mycres.2006.04.00

Vu, D., Groenewald, M., De Vries, M., Gehrmann, T., Stielow, B., Eberhardt, U., Al-Hatmi, A., Groenewald, J.Z., Cardinali, G., Houbraken, J., & Boekhout, T. (2018). Large-scale generation and analysis of filamentous fungal DNA barcodes boosts coverage for kingdom fungi and reveals thresholds for fungal species and higher taxon delimitation. *Studies in Mycology*, 91, 23–36.

Data accessibility and benefit-sharing

Data accessibility: After the paper is published, the unique haplotype data will be made available through the PlutoF web platform.

Benefits Generated: Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author contributions

L.T and K.R conceived the research idea and designed the study, K.R and U.K collected data, K.R, L.T, I.S, V.M and O.C analyzed data, K.R and L.T wrote the paper, all authors discussed the results and commented on the manuscript.

Tables and figures

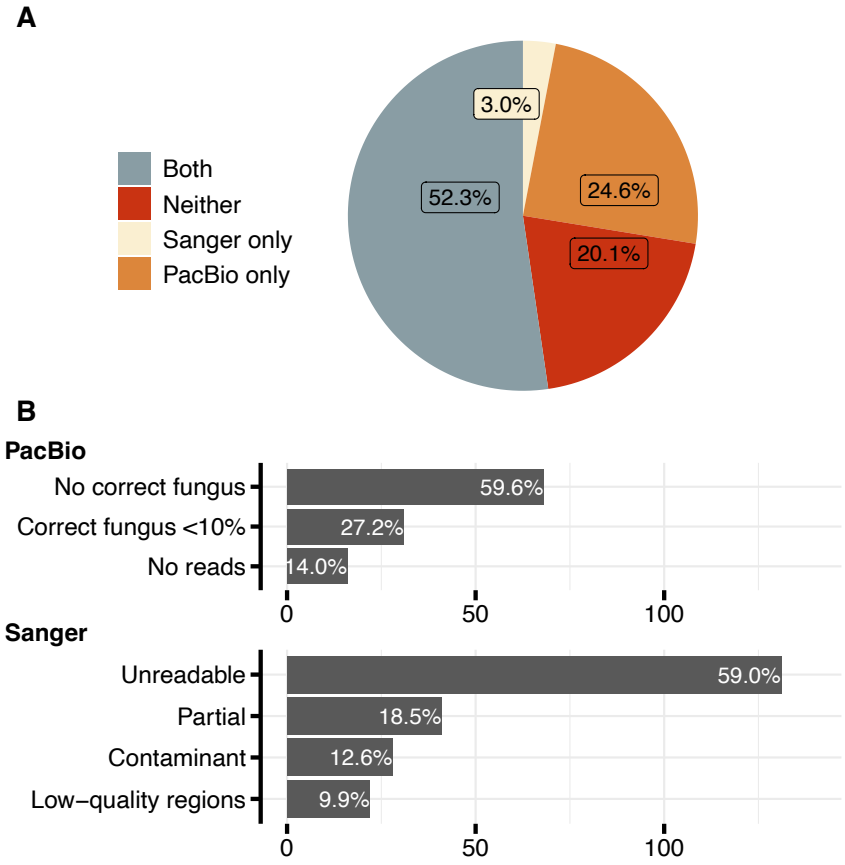

Fig 1. The sequencing success of 497 fungal amplicons with Sanger and PacBio methods (A), and the reasons for failed sequencing (B).

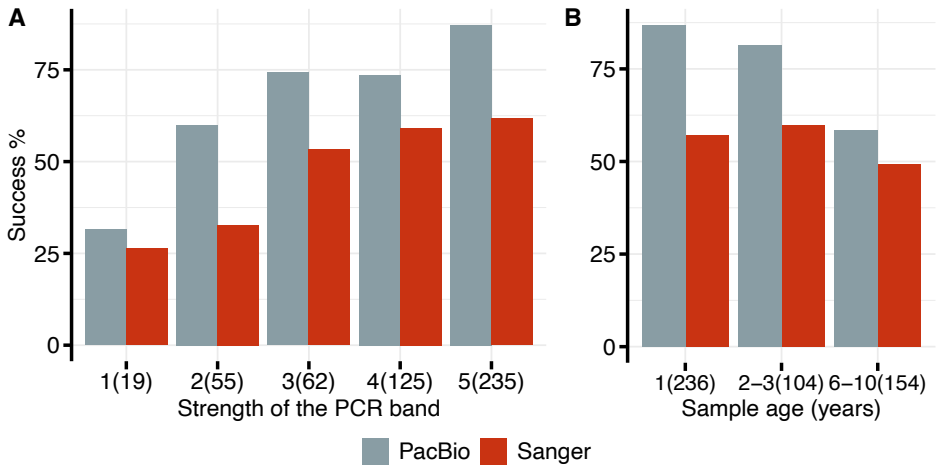

Fig 2. Sequencing success with PacBio and Sanger sequencing at different strengths of the PCR band (A) and sample ages (B). No. of samples in parentheses. There was a single sample with no PCR band.

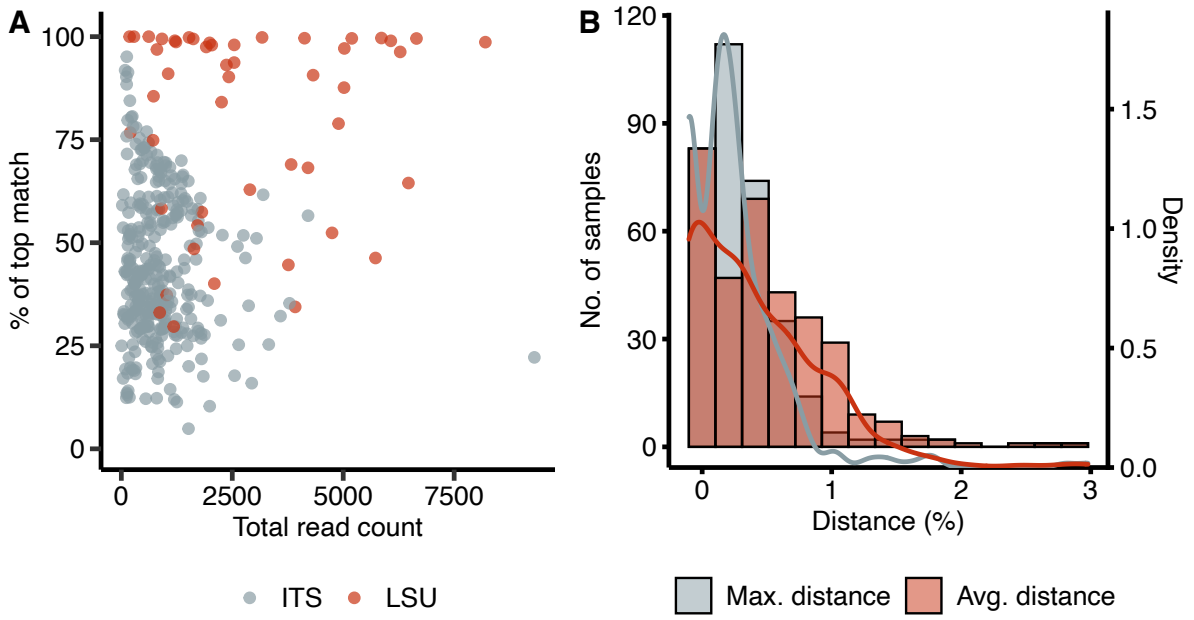

Fig 3. The major characteristics of successful PacBio samples. A: The relationships between total no. of sequences and relative abundance of the top (most abundant) allele. B: Count and density (probability density) of average and maximum distances between alleles in samples successfully sequenced for the ITS region. Distance is zero for samples with no allele polymorphism.

Table 1. Effects in the binomial regression models. "Sanger failure" is a categorical factor with four levels, where "partial sequence" is a reference group. p-values: * < 0.05; ** < 0.01; *** < 0.001.

	Estimate	Std error	z	P	
PacBio success vs. Sanger failure reasons					
(Intercept)	2.54	0.60	4.23	<0.001	***
Sanger failure: unreadable	-2.68	0.63	-4.28	<0.001	***
Sanger failure: contaminant	-4.33	0.81	-5.37	<0.001	***
Sanger failure: low q. regions	-0.69	0.86	-0.80	0.422	
Sanger success vs. PCR band strength					
(Intercept)	-1.19	0.33	-3.61	<0.001	***
PCR band strength	0.35	0.08	4.45	<0.001	***
Sanger success vs. sample age					
(Intercept)	0.43	0.15	2.88	0.004	**
Sample age	-0.07	0.04	-1.82	0.069	.
PacBio success vs. PCR band strength					
(Intercept)	-0.79	0.34	-2.34	0.019	*

PCR band strength	0.52	0.09	6.02	<0.001	***
PacBio success vs. sample age					
(Intercept)	2.14	0.20	10.56	<0.001	***
Sample age	-0.29	0.05	-6.21	<0.001	***

Table 2. Number of polymorphic alleles, and intraspecific distances in 11 studied species.

	No. of samples	No. of polymorphic alleles	Min distance (%)	Max distance (%)	Average distance (%)
Antrodia piceata	8	6	0.171	0.512	0.262
Antrodia serialis	6	19	0.181	1.093	0.565
Physisporinus vitreus L4	7	9	0.183	0.730	0.386
Postia tephroleuca	6	12	0.185	0.750	0.431
Rhodonia placenta	10	24	0.172	1.382	0.717
Sidera vulgaris	33	140	0.181	1.828	0.911
Skeletocutis nemoralis	10	35	0.167	1.003	0.510
Skeletocutis semipileata	14	61	0.166	1.815	0.720
Skeletocutis stellae	6	31	0.174	1.394	0.681
Tomentella sp. 15	6	2	0.173	0.173	0.173
Tomentella sp. 23	7	2	0.175	0.175	0.175