Linked machine learning classifiers improve species classification of fungi when using error-prone long-reads on extended metabarcodes

Tavish G. Eenjes[1†], Yiheng Hu[1,7†], Laszlo Irinyi[2,3,4], Minh Thuy Vi Hoang[2,3,4], Leon M. Smith[1,8], Celeste C. Linde[1], Wieland Meyer[2,3,4,5], Eric A. Stone[1,6], John P. Rathjen[1], Benjamin Mashford[6*], Benjamin Schwessinger[1*]

[1] Research School of Biology, Australian National University, Canberra, ACT, Australia

[2] Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, The University of Sydney, Sydney, NSW, Australia.

[3] Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Sydney, NSW, Australia.

[4] Westmead Institute for Medical Research, Westmead, NSW Australia.

[5] Westmead Hospital (Research and Education Network), Westmead, NSW, Australia.

[6] ANU-CSIRO Centre for Genomics, Metabolomics and Bioinformatics, Canberra, ACT, Australia

[7] Current address: Department of Microbial Interactions, IMIT/ZMBP, University of Tübingen, Tübingen, Germany

[8] Diversity Arrays Technology, Canberra, ACT, Australia

[†] These authors contributed equally to this work.

* Correspondence should be addressed to
Benjamin Mashford (benjamin.mashford@anu.edu.au)
and Benjamin Schwessinger (benjamin.schwessinger@anu.edu.au)

ABSTRACT

The increased usage of long-read sequencing for metabarcoding has not been matched with public databases suited for error-prone long-reads. We address this gap and present a proof-of-concept study for classifying fungal species using linked machine learning classifiers. We demonstrate its capability for accurate classification using labelled and unlabelled fungal sequencing datasets. We show the advantage of our approach for closely related species over current alignment and k-mer methods and suggest a confidence threshold of 0.85 to maximise accurate target species identification from complex samples of unknown composition. We suggest future use of this approach in medicine, agriculture, and biosecurity.

KEYWORDS

Machine learning, fungi, species classification, long-read sequencing, metabarcodes

BACKGROUND

DNA sequencing is increasingly becoming an important part of identifying and classifying fungal species, particularly through DNA barcoding. To date this process involves the use of short, variable regions of DNA that differ between species and are surrounded by highly conserved regions which are suitable targets for ‘universal’ primers enabling PCR amplification over a large variety of fungal taxa [1, 2]. The internal transcribed spacer (ITS)

region, is used as the primary DNA barcode region for fungal diversity studies [3]. This regions contains the two variable components, ITS1 and ITS2, which are on average 550-600 bp long [4]. The ITS1 and ITS2 are separated by the conserved 5.8S rRNA gene and is flanked by the conserved 18S and 28S rRNA genes. Although these regions offer a targetable region for identifying fungal species, they have some limitations that affect the ability to accurately classify fungi especially at lower taxonomic ranks [4, 5]. The length of the complete ITS1/2 region prevents short-read sequencing platforms to use both in combination for taxonomic classification. Furthermore, the limited selection of 'universal' primers in the region can subject taxonomic studies to primer biases [6].

With the advent and increasing use of long-read sequencing, such as that enabled by the nanopore sequencing technology of the MinION from Oxford Nanopore Technologies (ONT), some of the limitations of short-reads can be bypassed [7]. With long-reads, an extended ITS region can be sequenced including both ITS1 and ITS2 in addition to the minor variable regions of the 18S and 28S rRNA subunits using one set of 'universal' primers [8-11]. Here, we focus on the region amplified by the NS3 and LR6 primers [12], spanning close to 2.9 kbp in size. We refer to this amplicon hereafter as the fungal ribosomal DNA region. Nanopore sequencing introduces a relatively high read error of around 10% at the time of conducting our study [13]. These make individual reads less suited for species identification using DNA metabarcodes combined with currently existing sequence alignment and k-mer based methods because the genetic distance of the variable regions between closely related species are often lower than the per read error rate [14]. In addition, the entries in most fungal DNA barcode databases, such as NCBI and Unite, are relatively short with a median sequence length of 580 bp and 540 bp [15], respectively. This limits the analysis capacity of long-reads which completely entail both ITS sequences and include minor variable regions in both 18S and 28S rRNA.

In our current study we address these shortcomings and assess the applicability of novel sequence analysis methods for metabarcodes using the fungal kingdom as a test case. The fungal kingdom is diverse, with an estimated 1.5-5 million species globally, performing important ecosystem functions [16]. At the same time fungi can have adverse effects on human and animal health and agriculture. An estimated 300 million people suffer from fungal-related diseases each year [17], which often have a high mortality rate and limited treatment options, resulting in the deaths of over 1.5 million people annually [18]. Similarly, fungi can cause large-scale biodiversity loss [19, 20] as demonstrated by the near extinction of many amphibian taxa by the globally devastating fungal pathogen *Batrachochytrium dendrobatidis* [21] and the local extinction of several myrtaceae tree species by the rust fungus *Austropuccinia psidii* [22]. Fungal pathogens also cause an estimated loss of about $200 billion dollars in global food production annually [23]. The importance of fungi warrants the development of improved sequence-based detection methods for fungi as illustrated in our proof-of-concept study.

We explored machine learning classifiers as an alternative method for assigning individual error-prone sequence long-reads to taxa, because machine learning techniques are ideally suited to identify deterministic spatial relationships between features for classification [24]. For example, it might be that specific DNA bases have a unique spatial relationship within the fungal ribosomal DNA region that is deterministic for a given fungal species. These relationships are difficult to capture with currently available (local) alignment or k-mer based methods when combined with error-prone sequence long-reads, especially when these features (DNA bases) are not located in close proximity in the primary DNA sequence. There exist many machine learning methods for identifying patterns across a variety of data types [25-27]. Convolutional neural networks (CNNs) are one type of machine learning methods that are especially suited for identifying the deterministic spatial relationships in DNA

sequence, as they are capable of learning from both small-scale and higher order discretionary features, including important spatial relationships between said features [24, 28, 29]. So, we applied a CNN approach to metabarcoding based fungal species identification using a uniquely labelled sequencing dataset of the 2.9 kbp fungal ribosomal DNA region from 44 individually sequenced fungal species. We compared our machine learning approach to three commonly used analysis approaches including alignment and k-mer based methods on different in house and publicly available databases. Our machine learning approach faired especially well when identifying closely related species. Furthermore, we show that the training of a limited set of general and specific machine learning taxa classifiers provides a reasonable approach to targeted species identification from a complex sample of unknown composition.

RESULTS

Design of a decision tree for machine learning classifiers for taxonomic assignment of fungal species

Here we explored the application of machine learning on individual nanopore reads for fungal taxonomic classification. We sequenced the fungal ribosomal DNA region of 44 fungal species individually to generate a labelled real-life dataset for which the ground truth is known for each individual read. This makes our dataset uniquely suited for our supervised machine learning approach and for benchmarking studies when comparing this to commonly used classification approaches. Our fungal species dataset included 39 ascomycetes species spanning 19 families and 27 genera in addition to five basidiomycetes. We performed several quality-control steps on all reads in each sample. We first filtered reads based on homology against a custom-curated database of the fungal ribosomal DNA region, to remove any partial

reads or reads from other areas of the fungal genome with partial primer binding. We then filtered reads by length, removing short or very long-reads that were not within a 90% confidence interval around the mean read length for the fungal ribosomal DNA region for each species (see Supplemental Table T1). The *Galactomyces geotrichum* sample had too few reads for further processing, hence we complimented those with simulated reads using NanoSim [30]. This resulted in an average of 54,832 ± 35,537 reads available across all species. We took a subsample of these quality-controlled reads and split them into a training set and a test set, containing 85% and 15% of the subsampled reads respectively, to be used for training the machine learning classifiers and assessing the performance of the newly generated machine learning classifiers, respectively. We implemented a decision tree to be able to classify individual reads at each taxonomic rank from phylum to species (Figure 1). The taxonomic information for the 44 available individually sequenced species was used to create the cladogram for this decision tree. We generated one machine learning classifier for each node in our decision tree (Figure 1).

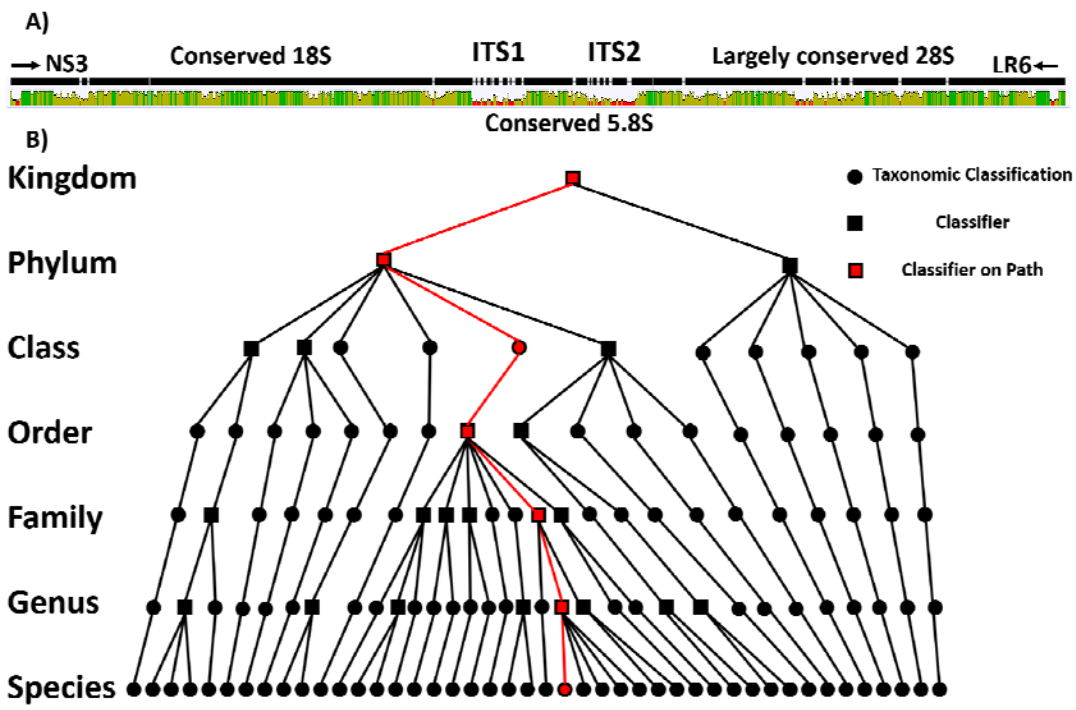


Figure 1: Visualisation of the fungal ribosomal DNA region and machine learning decision tree.

(A) The fungal ribosomal DNA region between the NS3 and LR6 primers covers around ~2.9 kbp. Shown is the alignment of consensus sequences (created with Geneious Prime), with highly conserved regions in green and variable regions in red. **(B)** Taxonomic information for 44 species was used to create a decision tree showing the relationship between samples at each taxonomic rank. Where two or more species shared a common taxon, a machine learning classifier (node) was created to distinguish between those taxa. These classifiers were chained together to create a decision tree, whereby the classification of a read was undertaken by a cascade of classifiers. An example path down the decision tree is shown for *Candida albicans*.

For training each of these classifiers, a balanced dataset was used, such that each possible outcome of the machine learning classifier had an equal number of reads. These individual classifiers had a mean recall rate of 97.9 ± 1.1% for correctly classifying reads using the test read dataset. The lowest recall rate belonged to the species-level classifier that distinguished between *Candida* species, with a recall rate of 94.4%.

To fully classify a read, we used the cladogram as a decision tree to link individual machine learning classifiers at each taxonomic rank. This allowed us to chain classifiers together to classify a read at each taxonomic rank, moving through the tree from phylum to species assignments. The outcome of a classifier at one taxonomic rank was used to decide the path

along the tree, and thus this decision defined which classifier was appropriate for use at the next lower taxonomic rank (Figure 1). We refer to a classifier by the taxonomic rank that it outputs. For example, a species-level classifier takes reads from a specific genus and outputs a species, while a class-level classifier takes reads from a specific phylum and outputs a decision on the taxonomic class of the read. The recall rate of the individual classifiers at different taxonomic ranks can affect the final species-level recall rate for each individual read as it moves through the decision tree. This means that the final species-level recall rate is equal to or worse than the individual species-level classifier's recall rate. Another limitation of our approach was that not every path through the decision tree had a node at each taxonomic rank, because of the taxonomic composition of our 44 individually sequenced species. For example, the basidiomycete species *Puccinia striiformis* f. sp. *tritici* has only two classifiers, at the phylum level and the class level. The latter decides the class classification which collapses with the species classification because *Puccinia striiformis* f. sp. *tritici* is the only species in the class Pucciniomycetes in our sequencing dataset. In total we trained 22 classifiers to distinguish our 44 fungal species.

Comparison of methods for species classification of fungal pathogens

We compared the machine learning decision tree to two other more standard methods for read classification to determine the effectiveness of this technique. We assessed the ability of the other methods at classifying reads across multiple taxonomic ranks because the tiered nature of the decision tree offers the potential to gleam taxonomic information from a read, even when it cannot be confidently classified at the species level. We used two additional classification techniques. We first applied *mimimap2*, a pairwise alignment-based method designed to be used with long-reads, against a gold-standard custom-curated database generated from the consensus sequences of all 44 species present in the decision tree (gold standard alignment). This is the most appropriate comparison for our machine learning

approach because the gold standard and machine learning approaches are directly derived from our sequencing dataset. To compare the machine learning approach with methods where the sequencing data was not used to create the classification database in some way, we applied *minimap2* to a large publicly-available database of fungal ITS sequences from NCBI [31, 32] (NCBI alignment), and applied *Kraken2*, a k-mer-based algorithm designed for use with metagenomic DNA sequences, to the same NCBI database (Kraken2).

To compare these methods, an *in silico* mock community was generated from our labelled sequencing data for which we know the ground truth classification for each sequencing read. This mock community contained 13 species from the original 44 species used to generate the original machine learning decision tree. Species were selected to focus on species for whom multiple machine learning classifiers would be required, in particular those species from populous genera. Although all species from this mock community were present in the gold standard database, the NCBI database was missing some genera and species. All of these missing or unclassified taxonomies were recorded as having a recall rate of zero percent, artificially decreasing the quality at lower taxonomic ranks.

Our machine learning decision tree approach maintained a consistently high recall rate across all taxonomic ranks, with a mean species level recall rate of 93.0 ± 2.8%. Notably, it performed very well for closely related taxa, including the cryptic species *Candida metapsilosis* and *Candida orthopsilosis* and another closely related species *Candida albicans*. The two cryptic *Candida* species (*C. metapsilosis* and *C. orthopsilosis*) had a very high consensus sequence similarity, with a genetic distance of 2.74% (97.26% identity) in our fungal ribosomal DNA region target region representing the genetically least distinct species pair. Our machine learning approach did achieve species level recall rates of 90.1% and 89.1% for *C. metapsilosis* and *C. parapsilosis*, respectively, even with per read error rates of about 10%. This highlights the strength of our approach.

The gold standard alignment approach also performed very well when compared to the machine learning approach across all taxonomic ranks (Figure 2). The majority of the species were classified with recall rates in excess of 95%. Yet this approach significantly underperformed when trying to differentiate taxa with low genetic distance such as those from the *Candida* genus. As with the machine learning approach, the three *Candida* species were classified with the lowest recall rate at the species level, with *C. albicans*, *C. metapsilosis* and *C. parapsilosis* being classified with recall rates of 35.8%, 34.0% and 57.5% respectively. These difficulties are also reflected in the overall mean species level recall rate of 76.6 ± 25.5%, which is much lower than our machine learning approach.

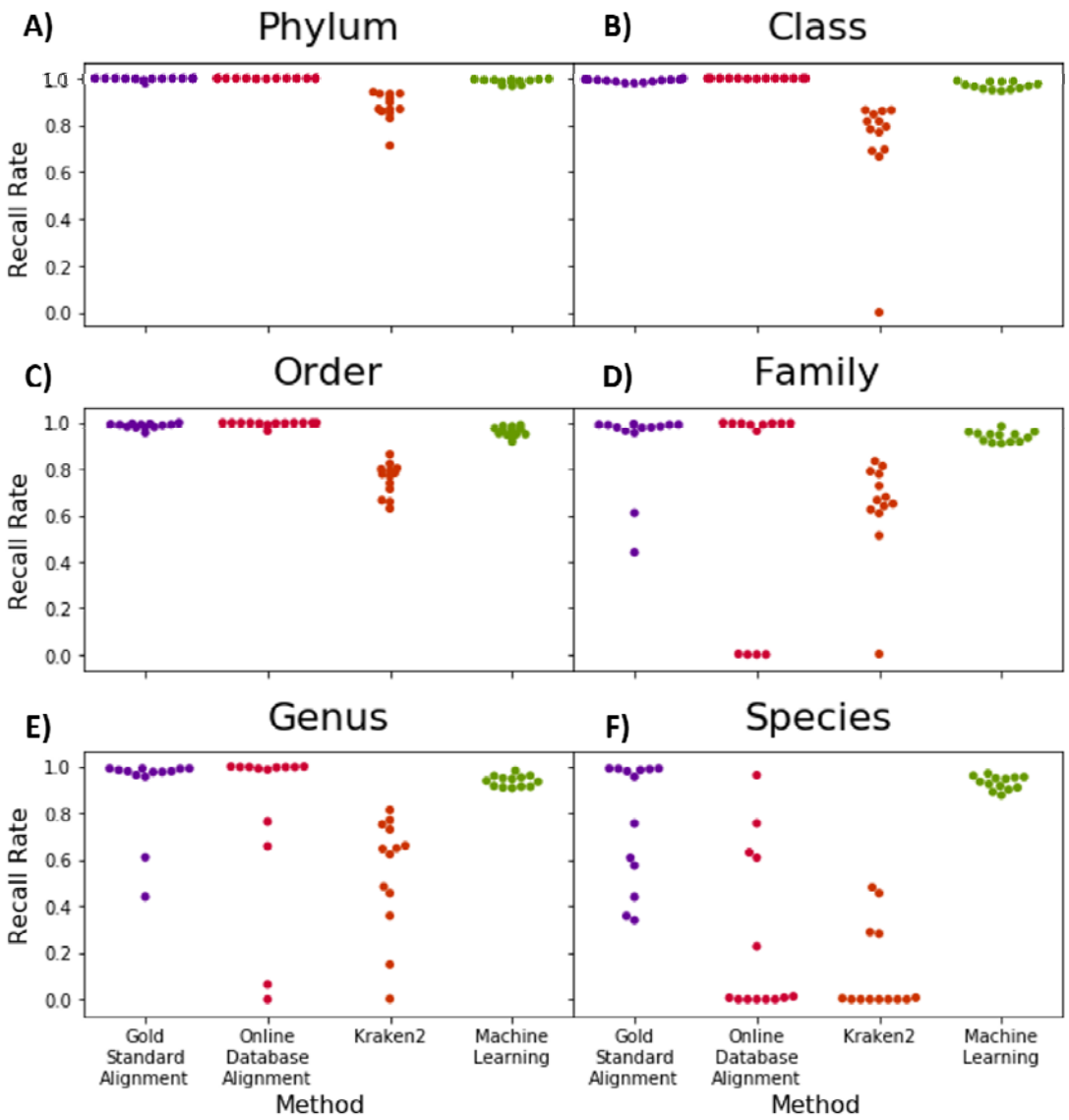


Figure 2: Machine learning based species identification performs especially well for closely related species

Recall rates of the alignment-based *minimap2* technique, k-mer-based *Kraken2* method and the machine learning decision tree across different taxonomic ranks **(A-F)**. The *minimap2* technique, as applied to the gold standard database, was successful across most taxonomic ranks, but lower recall rates were recorded for closely related species at the species level **(F)**. Both the *minimap2* and *Kraken2* methods were applied to the NCBI database, and while the *minimap2* NCBI alignment was more accurate across most taxonomic ranks, both showed comparable recall at the species level. The machine learning decision tree approach provided the greatest classification power for closely related species, despite lower recall rates for some distantly-related species than the gold standard alignment method.

Next, we assessed our dataset with alignment and k-mer based analysis approaches when using the publicly available NCBI database. Overall, NCBI alignment with *minimap2* performed similarly well at higher taxonomic ranks. However, inconsistent or missing

naming conventions at the family level and missing or alternate species labels, meant that the overall recall rate was low at the species level, although the vast majority of the samples were classified with a high recall rate at the genus level. This low species level recall rate is an artefact created from the choice of database, which is reflected in the similarly poor species level recall rates of the *Kraken2* method. Overall, the k-mer based *Kraken2* was less accurate than all other methods tested across all taxonomic ranks.

Identifying target species from a complex sample of unknown composition using the machine learning decision tree

A key feature of a species classification tool is its ability to identify a known target species from a complex sample of unknown composition. This is especially important when attempting to identify the presence of a target species, such as a specific pathogen, from a metagenomic sample.

We generated two additional sequencing datasets of truly unknown composition to test the capability of our machine learning decision tree to identify a given target species. These datasets were generated with the same PCR and sequencing protocols as for the individual 44 training species focusing on the fungal ribosomal DNA region. The first dataset was derived from fungi-infected wheat leaves (wheat dataset) [33] and the second was derived from bronchoalveolar wash in a clinical setting (clinical dataset) [34]. To each of these sequencing datasets of unknown composition, we spiked *in silico* a known number of reads with known labels as test case. We choose *Aspergillus flavus*, a crop pathogen, and *Candida albicans*, a human pathogen. We then tested recall and false positive rate of our machine learning classifiers using our *in silico* spiked reads, assuming that the original datasets of unknown composition did not contain any reads of either species.

We first plotted the propagated confidence score of the species level classification for all reads in each *in silico* spiked dataset to better understand the behaviour of our machine learning decision tree on samples containing reads of unknown origin (Supplemental Figure S1). This clearly shows that the propagated confidence scores for reads of unknown origin are far lower than reads of species the classifiers were trained on. We than assessed the recall and false positive rate of the *in silico* spiked datasets at different confidence scores thresholds (Figure 3). Increasing the thresholds reduced the recall and false positive rate in both cases. For *A. flavus*, the recall rate remained above 90% until the confidence threshold reached 0.9, and the false positive rate was consistently low across both the clinical and wheat datasets with reads of unknown origin. A confidence threshold of 0.85 resulted in a high recall rate of 0.917, while maintaining a low false positive rate of just one percent. For *C. albicans*, not using a confidence threshold at all resulted in a recall rate of 87.7% and false positive rate of 11.7%. However, by using a confidence threshold of 0.85, the recall rate was only decreased to 72.4% while reducing the false positive rate to only 1.7% in the clinical dataset. We recommend this confidence score threshold of 0.85 as suitable for retaining a high recall rate while achieving a low false positive rate, even for a member of a difficult-to-distinguish genus like *Candida*.

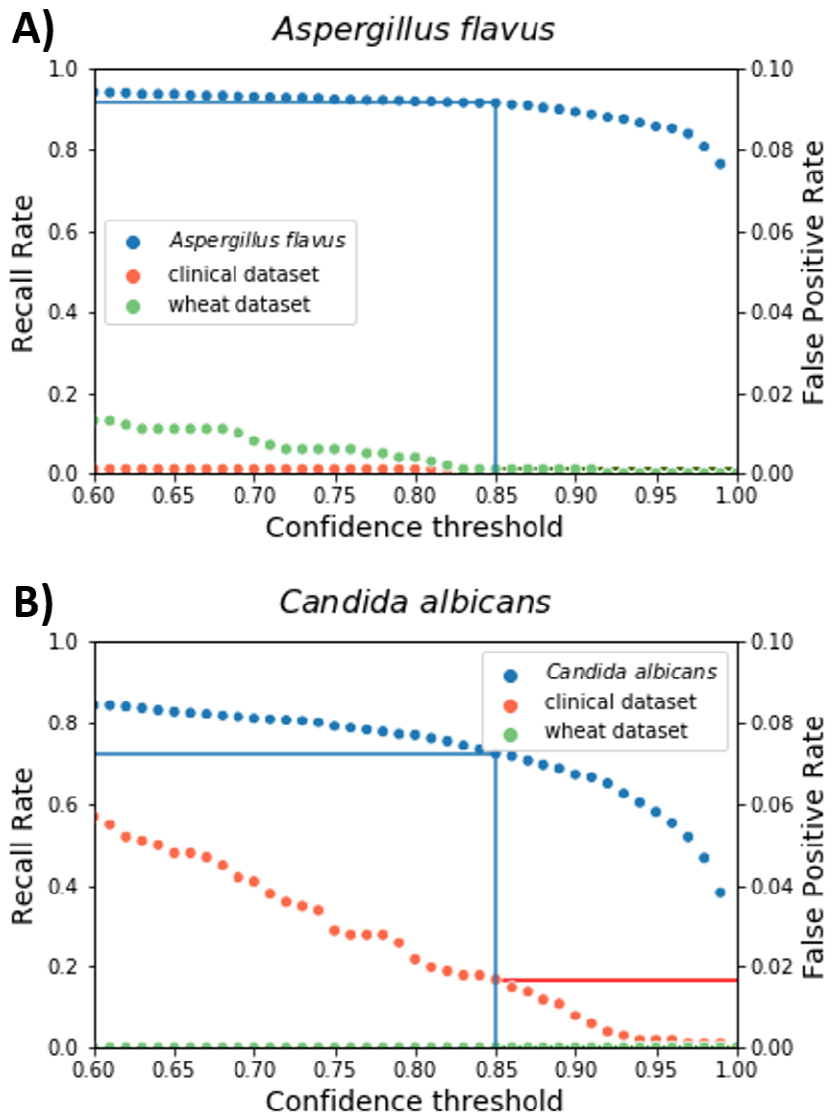


Figure 3: Varying the confidence score threshold affects the recall rate and false positive rate when identifying target species from complex samples of unknown composition.

The plots show the recall rate (left axis) and false positive rate (right axis) at varying propagated confidence score thresholds for *Aspergillus flavus* **(A)** and *Candida albicans* **(B)** when spiked into clinical (orange) or wheat (green) datasets. Both plots are based on 2000-read *in silico* spiked samples containing 1000 reads with known labels (*A. flavus* or *C. albicans*) and 1000 reads of unknown origin. For *A. flavus*, a confidence threshold of 0.85 maintains a recall rate of 91.7%, while reducing the false positive rate to 1% for both datasets. For *C. albicans*, the same confidence threshold of 0.85 has a recall rate of 72.4% and reduces the false positive rate below 2% for both datasets.

DISCUSSION

Nanopore sequencing offers portable, real-time sequencing using long-reads that can cover extended metabarcodes that are poised to include more sequence information suitable for species classification than more classic Illumina short-read sequencing [35]. Yet currently, metabarcode datasets in publicly available databases are limited in barcode length and often do not cover these extended regions. This can cause difficulties when using error-prone nanopore long-reads to classify reads at the species level using these databases [36]. Here, we implement a novel machine learning approach for species level classification.

Our machine learning approach is comparable to – albeit slightly outperformed by - the gold standard alignment approach across all taxonomic ranks for most of the species tested. However, the gold standard alignment approach has a very poor performance at the species level for very closely related species within the same genus. This is indicative of the problems of alignment-based classification methods for fungi, especially given the relatively high error rate of the nanopore long-reads [37]. Hence, it is at the species level where the greatest potential for improvement using machine learning lays. For example, some closely related species were highly misclassified with a recall rate lower than 50% using the *minimap2* alignment against the gold standard database. The same species were classified with recall rates equal to or greater than 90% using our machine learning decision tree. This is remarkable given the per read error rate of 10% for nanopore reads is much larger than the genetic distance of 2.74% that we observed between some closely related taxa.

These initial comparisons are based on idealised databases directly derived from our sequencing dataset for which sequencing read length and database entry length are equivalent. Hence, we expected these analyses to outperform other approaches relying on public databases with short reference sequences. This was indeed the case as analysing our

error prone long-reads with alignment (NCBI alignment) and k-mer (Kraken2) based approaches using the NCBI ITS RefSeq Targeted Loci database performed relatively poorly especially at lower taxonomic ranks. Clearly, the discrepancy between read and database sequence lengths (~2900 bp vs ~580 bp) negatively impacted the alignment success. Interestingly, the Kraken2 approach underperformed compared to the alignment-based approach in our current study. This is consistent with previous work with long-read MinION nanopore data, where Kraken2 classification success never exceeded that for BLAST, another alignment-based classification program, when using the default 35 bp k-mers [38]. It is likely using a smaller k-mer length would improve classification accuracy for long-read nanopore sequencing due to the high read error, which impacts perfect matches for 35 bp k-mers. Another common issue when using public databases for species identification was that many species were not included in the NCBI database or present with different taxonomic labels, which resulted in some family and species level recall rates being zero. Changing nomenclature over time can be an issue when using these online databases when trying to identify a species or detect the presence of a known, named species, as the nomenclature is not always updated, leading to outdated or uncorrected taxonomic information persisting in databases [39, 40].

We also tested if our machine learning approach can accurately identify specific target species in complex samples of unknown composition without having classifiers for all fungal species present in the sample. We were able to show that by only training a limited set of classifiers we can detect target species with relatively low false positive and high recall rates in *in silico* spiked datasets with known ground truth of the spiked reads only. By adjusting the confidence score one can decide how much false positive and false negatives one is willing to tolerate. We found a threshold of 0.85 on the propagated confidence score at the species level classification was sufficient to reduce the false positive rate while maintaining high recall

rates. To ensure a target species is identifiable, the species-level classifier in the machine learning decision tree must include other species closely related to the target species. If no closely related species is present, the likelihood of false positive hits increase as closely related taxa may be identified as false positives with high confidence scores even in the absence of the target species. As such, the more fungal species within a genus the machine learning decision tree classifiers are trained on, the higher the resolution of species-level identification. This is especially important when a genus contains both pathogenic and non-pathogenic species. In this way, our approach might be particularly applicable to targeted diagnostic tasks in specific settings, such as detecting fungal pathogens in agriculture [41] and medicine [42], or screening imports for specific invasive pathogen species in aid of border biosecurity [43, 44]. Here, the species used to train the classifiers are flexible and can be changed to suit the user's need. For example, additional species from a specific taxon could be added for increased resolution within that taxa. Furthermore, the principles behind the application of machine learning to the fungal ribosomal DNA region can be expanded to other barcoding regions for other organisms, such as *cytochrome c oxidase I* [45] or *elongation factor 1 alpha* [46, 47]. Recent work on improving barcoding cost-effectiveness and scalability with the MinION nanopore sequencer offers promise for expanding to more species using barcoding across multiple regions to improve the species-level resolution and overall classification accuracy [48].

CONCLUSIONS

Online databases for metabarcoding often contain only short sequences, and hence are traditionally useful for identifying taxa using high accuracy short-reads. As such, identifying species from error prone long-read sequencing data, such as that produced by ONT nanopore

sequencing, can be inaccurate when using these databases. We provide a tangible solution for species identification by applying a novel neural network-based machine learning approach with a proof-of-concept study using extended fungal ribosomal DNA barcodes on fungi. Our machine learning approach can identify target species with high accuracy from complex samples of unknown origin making it applicable to pathogen identification in biosecurity, agriculture, and clinical settings. Our approach performs especially well on closely related species where it provides an advantage in accuracy over current alignment-based or k-mer-based classification methods.

MATERIALS AND METHODS

Fungal pathogen sample collection, DNA extraction and ITS amplification

We collected different fungal tissue differently for DNA extractions. The tissue collection processes for each fungal species are summarized in Supplemental Table T1.

We used three different DNA extraction methods for all the species in the mock communities. The methods for each species are listed in the Supplemental Table T1. Collectively, we used two commercially available kits: The Qiagen DNeasy Plant Mini Kit (cat. no. 69106) for most of the plant pathogenic fungi, and the Quick-DNA Fungal/Bacterial Miniprep Kit (cat. no. D6005, Zymo Research) for some of the human pathogenic fungi following the manufacturer's protocol. We used a phenol chloroform-based DNA extraction method for some other human pathogenic fungi modified from Ferrer et al [49]. Briefly, 100 mg of leaf tissue was homogenized, and cells were lysed using cetyl trimethylammonium bromide (CTAB, Sigma-Aldrich) buffer (added RNAse T1, Thermo Fisher, 1,000 units per 1750 μl), followed by a phenol/chloroform/isoamyl alcohol (25:24:1, Sigma-Aldrich) extraction to remove protein and lipids. The DNA was precipitated with 700 μl of

isopropanol, washed with 1 ml of 70% ethanol, dried for 5 min at room temperature, and resuspended in 50 μl of TE buffer containing 10 mM Tris and 1 mM EDTA at pH 8. For the human clinical sample and the field infected wheat sample, we directly used the DNA described in the original article [33, 34] for PCR amplification. Quality and average size of genomic DNA was visualized by gel electrophoresis with a 1% agarose gel for 1 h at 100 volts. DNA was quantified by NanoDrop and Qubit (Life Technologies) according to the manufacturer's protocol.

We used the NS3 (GCAAGTCTGGTGCCAGCAGCC) and LR6 (CGCCAGTTCTGCTTACC) primers [12] to generate the fungal ribosomal DNA fragment of all samples, and the EF1-983F (GCYCCYGGHCAYCGTGAYTTYAT) and EF1-2218R (ATGACACCRACRGCRACRGTYTG) primers [12] were used to sequence a secondary region, the fungal elongation factor 1 alpha region, although this region was not used for assessing the machine learning method. We used the New England Biolabs Q5 High-Fidelity DNA polymerase (NEB #M0515) for the PCR reaction following the manufacturer's protocol. Around 10 – 30 nanograms of DNA were used in each PCR reaction. After PCR, DNA was purified with one volume of Agencourt AMPure XP beads (cat. No. A63881, Beckman Coulter) according to the manufacturer's protocol and stored at 4°C.

Library preparation and DNA sequencing using the MinION

DNA sequencing libraries were prepared using Ligation Sequencing 1D SQK-LSK108 and Native Barcoding Expansion (PCR-free) EXP-NBD103 Kits from ONT, as adapted by Hu and Schwessinger [50] which was adapted from the manufacturer's instructions with the omission of DNA fragmentation and DNA repair. DNA was first cleaned up using a 1x volume of Agencourt AMPure XP beads (cat. No. A63881, Beckman Coulter), incubated at room temperature with gentle mixing for 5 mins, washed twice with 200 μl fresh 70%

ethanol, the pellet was allowed to dry for 2 mins and the DNA was eluted in 51 µl nuclease free water and quantified using NanoDrop® (Thermo Fisher Scientific, USA) and Promega Quantus™ Fluorometer (cat. No. E6150, Promega, USA) follow the manufacturer's instructions. All DNA samples showed a with absorbance ratio A260/A280 > 1.8 and A260/A230 > 2.0 from the NanoDrop®. DNA was end-repaired using NEBNext Ultra II End-Repair/ dA-tailing Module (cat. No. E7546, New England Biolabs (NEB), USA) by adding 7 µl Ultra II End-Prep buffer, 3 µl Ultra II End-Prep enzyme mix. The mixture was incubated at 20°C for 10 mins and 65°C for 10 mins. A 1x volume (60 µl) Agencourt AMPure XP clean-up was performed, and the DNA was eluted in 31 µl nuclease free water. Barcoding reaction was performed by adding 2 µl of each native barcode and 20 µl NEB Blunt/TA Master Mix (cat. No. M0367, New England Biolabs (NEB), USA) into 18 µl DNA, mixing gently and incubating at room temperature for 10 mins. A 1x volume (40 µl) Agencourt AMPure XP clean-up was then performed, and the DNA was eluted in 15 µl nuclease free water. Ligation was then performed by adding 20 µl Barcode Adapter Mix (EXP-NBD103 Native Barcoding Expansion Kit, ONT, UK), 20 µl NEBNext Quick Ligation Reaction Buffer, and Quick T4 DNA Ligase (cat. No. E6056, New England Biolabs (NEB), USA) to the 50 µl pooled equimolar barcoded DNA, mixing gently and incubating at room temperature for 10 mins. The adapter-ligated DNA was cleaned-up by adding a 0.4x volume (40 µl) of Agencourt AMPure XP beads, incubating for 5 mins at room temperature and resuspending the pellet twice in 140 µl ABB provided in the SQK-LSK108 kit. The purified-ligated DNA was resuspended by adding 15 µl ELB provided in the SQK-LSK108 (ONT, UK) kit and resuspending the beads. The beads were pelleted again, and the supernatant transferred to a new 0.5 ml DNA LoBind tube (cat. No. 0030122348, Eppendorf, Germany).

In total, four independent sequencing reactions were performed on a MinION flow cell (R9.4, ONT) connected to a MK1B device (ONT) operated by the MinKNOW software (version

2.0.2): 11 species for each flowcell. Each flow cell was primed with 1 ml of priming buffer comprising 480 µl Running Buffer Fuel Mix (RBF, ONT) and 520 µl nuclease free water. 12 µl of amplicon library was added to a loading mix including 35 µl RBF, 25.5 µl Library Loading beads (ONT library loading bead kit EXP-LLB001, batch number EB01.10.0012) and 2.5 µl water with a final volume of 75 µl and then added to the flow cell via the SpotON sample port. The "NC_48Hr_sequencing_FLOMIN106_SQK-LSK108" protocol was executed through MinKNOW after loading the library and run for 48 h. Raw fast5 files were processed using Albacore 2.3.1 software (ONT) for basecalling, barcode de-multiplexing and quality filtering (Phred quality (Q) score of > 7) as per the manufacturer's recommendations.

Raw unfiltered fastq files were uploaded into NCBI Short Reads Archive under BioProject PRJNA725648.

Processing and manipulation of fungal pathogen reads

All reads from one species were held in a fastq file with reads of varying quality, that included sequences from both the fungal ribosomal DNA and the elongation factor 1 alpha regions of the fungal genome. Data was thus required to be processed so downstream use dealt only with fungal ribosomal DNA reads of the expected size range. A two-step data filtration method was applied for this purpose.

To select reads of a similar general structure to the ITS region, reads were first mapped to an in-house database of fungal ribosomal DNA regions. This homology-based filter assumes the structure of the fungal ribosomal DNA region will be similar between species due to shared ancestry, which has been repeatedly shown to be true [51]. The in-house database used here was curated from 28 ITS sequences from the NCBI Nucleotide database, from a range of genera across the fungal kingdom. This process mapped reads using *minimap2* (version 2.17),

using the map-ont flag. Reads that failed to map to any of the sequences in the in-house database were discarded.

Reads that successfully mapped were then filtered for read length. The expected read length for the fungal ribosomal DNA region varied by species, from 2600-3200 bp on average. As the mean length and spread of successfully filtered reads differed between samples, a 90% confidence interval cut-off around the mean read length was applied. This interval was sufficient to exclude those remaining short or very long reads, that may have resulted from incomplete or partial homology filtering, or errors in the sequencing or basecalling processes.

Augmenting read datasets

To ensure all samples had at least 15,000 reads for use in the design of the machine learning classifiers downstream, some reads were simulated based on the consensus sequence and error profile of the existing reads where the total number of filtered reads did not exceed the required number of reads. NanoSim (v2.0.0) [30] was used for one species, *Galactomyces geotrichum*, to generate an additional 8,782 simulated cDNA reads. These reads were generated using an identical error profile and length spread to the pre-existing non-simulated fungal pathogen reads.

Generating consensus sequences for each species

The consensus sequence, an aggregate sequence formed from the comparison of multiple sequences that represents the 'true' sequence, was generated using 200 randomly subsampled filtered reads for each sample. Primer sequences were removed using *Mothur* v1.44.11 [52], an alignment file was generated using *muscle* v3.8.1551 [53] and the consensus sequence was generated from this file using *EMBOSS cons* v6.6.0.0 [54].

Determining the relationships between samples

Prior to using the processed read data to train machine learning classifiers, the taxonomic relationships between the samples were needed to inform the samples present in each machine learning classifier at each taxonomic rank. Using the taxonomic information available for each sample in MycoBank and the results of a BLAST search with the generated consensus sequences, a cladogram was designed to show the relationships between samples at each of the major taxonomic ranks. A machine learning classifier would be required at each point where two or more samples split on the cladogram (a node) to distinguish between samples for each read.

Creation of asset of neural network classifiers to distinguish between samples

A convolutional neural network (CNN) was chosen as the most appropriate type of machine learning classifier due to its ability to use the spatial relationships between data features in the reads, such as the distance between ITS and other variable groups, as a factor in assigning a label to a read. CNNs are capable of learning from both minor variation and higher-order features, which is of particular importance given the high read error of nanopore reads.

CNNs work best when there is a balanced number of items in each classification class. As such, for each multiclass node on the cladogram, an equal number of reads were subsampled from each group of samples that would be represented in the node. So, for machine learning classifiers distinguishing between species, each species present contributes an equal number of reads, while at the kingdom level, each phylum contributes an equal number of reads, with said reads being distributed equally amongst all species belonging to that phylum. The number of reads subsampled was based on the largest number of reads available for each sample, with a maximum of 35,000 reads due to computational processing limitations. For each read subsampled, the nucleotide sequence was converted to a numeric sequence, where A, C, G, and T became 0, 1, 2, and 3, respectively. As not all sequences were of equal length,

but an equal length was required to avoid sequence length being a distinguishing factor in the classifier, all sequences were padded out to a length of 5,000 bp. The padding used a value of 4 to avoid the padding data from affecting the identification of key features for classification.

Each read was assigned a label representing the output class it would belong to in the one-hot format. Labelled reads were then separated into a training set and a test set. The training set contained 85% of the reads, and was used to train the machine learning classifiers, while the test set contained the remaining 15% of labelled reads and was used to test the efficacy of said classifiers on similar data that the classifier had not previously encountered. The neural network was created using the Sequential classifier of the *Keras* framework for neural networks [55], containing five layers of neurons.

Specific details for the design of the machine learning classifiers and the required software packages for machine learning and other analyses can be found at **https://github.com/teenjes/fungal_ML.**

Evaluation of the machine learning classifiers

The test set was used to assess the accuracy of the various machine learning classifiers. As the test set data was labelled, the expected outcome for each read was known, and could be compared to the output of the machine learning classifier. The accuracy, or classification rate, of these classifiers was the proportion of reads in the test set for whom the prediction of the machine learning classifier, as determined by the highest confidence score, matched the expected outcome. This is equivalent to the recall rate [1], where matches to the expected outcome were true positives and matches outside this outcome were false negatives.

$$Recall\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad [1]$$

Chaining machine learning classifiers into a decision tree

When seeking to identify members of a specific taxon in a community, where the members are not immediately obvious from the species name, it is useful to have samples classified at each taxonomic rank. A singular classifier would require excessive computational power to do this. As such, we chained the machine learning classifiers together into a decision tree based on the cladogram of the species present in our sample. The most confident outcome of the machine learning classifier at one taxonomic rank would be used to decide the path along the decision tree. This path could either lead into another machine learning classifier, if the path diverged again, or lead all the way down to the species level with the same confidence.

Alternative methods for fungal pathogen read classification

For comparison to the machine learning classifier, two different commonly used methods for fungal pathogen metabarcode classification: an alignment-based method in *minimap2*; and a k-mer-based method in *Kraken2*. To compare these methods, we generated an *in silico* mock community from our labelled sequencing data for which we know the ground truth classification for each sequencing read. This mock community contained 13 species from the original 44 species used to generate the original machine learning decision tree, randomly subsampling 1000 reads from those not previously used for training the machine learning classifiers. Species were selected to focus on species for whom multiple machine learning classifiers would be required, in particular species with populous genera.

For this *minimap2*-based alignment method, two separate databases were used for identification. Firstly, a gold standard database was created in-house to represent the best-case scenario for identification, when all the species present in a sample are also present in the database. This contained the labelled consensus sequences of all 44 species present in the machine learning decision tree, using the consensus sequences already generated from 200 randomly selected filtered reads. The second was a publicly available database of fungal ITS

sequences from NCBI (ftp://ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Fungi/fungi.ITS.fna.gz, downloaded Feb 2021). *Minimap2* was applied to each of these databases using the map-ont flag. As the alignment tool can return multiple hits if alignment is good enough, only the best hit was taken for each read.

We used Kraken2 (v2.0.8) to assign the NCBI taxonomic ID for the same 1000 reads of each species as used in the machine learning decision tree. We generated a Kraken2 NCBI ITS database with the same fasta file downloaded from above. We used the Kraken2-build command with the --add-to-library and --build flag. We used the Python pandas module to modify the Kraken2 output file and the numpy module to calculate the accuracy.

Identifying a key species from a complex sample using machine learning

To assess the suitability of machine learning for this problem, we utilised the two complex datasets sampled from fungi-infected sources of unknown compositions: the field infected wheat dataset [33] and the human clinical dataset [34], to create *in silico* mock communities. To create these initial mock communities, we used 950 reads randomly subsampled from these datasets, and spiked in 50 reads from one of two target species with known ground truth: *Aspergillus flavus*, a crop pathogen; and *Candida albicans*, an opportunistic human pathogen and common member of the human microbiome. This created a total of four 1000-read synthetic communities, two of which paired a target species and dataset from the same source (*A. flavus* with the wheat dataset and *C. albicans* with the clinical dataset) and two communities where the target species would not be expected to be present in the complex dataset unless it had been spiked in. We used the propagated confidence scores for assessing the recall rate for these spiked datasets, where the confidence score at each taxonomic rank was multiplied to give a final overall confidence at the species level.

We then created an additional four *in silico* mock communities to assess the change in recall rate and false positive rate [2] as a confidence threshold was applied.

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives} \quad [2]$$

Each mock community was created by randomly subsampling 1000 reads from one of *A. flavus* or *C. albicans* samples with known ground truth and adding an additional 1000 randomly subsampled reads from one of the wheat or clinical datasets containing reads of unknown origin. In total, this resulted in four 2000-read *in silico* mock communities. We assumed the datasets with reads of unknown origin did not contain any reads for the target species tested, placing an upper bound on the false positive rate and a lower bound on the true positive rate. Any positive identifications of the target species *A. flavus* or *C. albicans* with a propagated confidence score below the confidence threshold were instead classified as negative identifications.

DECLARATIONS

Ethics Approval

Not applicable.

Consent for Publication

Not applicable.

Availability of data and materials

The code generated and used for machine learning during the current study is available in the fungal_ML repository, available at https://github.com/teenjes/fungal_ML. The datasets

generated and/or analysed during the current study are available in SRA under BioProject PRJNA725648, available at http://www.ncbi.nlm.nih.gov/bioproject/725648.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by an NHMRC grant (#GNT1121936) to WM.). BS is supported by an ARC Future Fellowship (FT1801000024).

Authors' contributions

TGE, BM, and BS designed the experiments and performed the analysis. YH extracted fungal DNA and performed all sequencing reactions. LL, MTV, LMS, CCL, and WM provided fungal material and/or DNA. ES and JR provided feedback on experimental design and data analysis. WM, ES, JR, and BS provided funding for the project. TGE and BS wrote the manuscript. All authors commented on the manuscript and approved submission.

Acknowledgements

We thank Eduardo Eyras, Jen Taylor, and Peter Solomon for suggestions on improving the machine learning models and determining statistics for identifying a species from a complex sample. We thank Peter Solomon and David Jones for providing fungal samples. We thank Andrew Milgate for providing infected wheat leaf material.

This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation Scheme.

REFERENCES

1. Seifert KA: **Progress towards DNA barcoding of fungi.** 2009, **9:**83-89.
2. Valentini A, Pompanon F, Taberlet P: **DNA barcoding for ecologists.** *Trends in Ecology & Evolution* 2009, **24:**110-117.
3. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W: **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.** 2012, **109:**6241-6246.
4. Xu J: **Fungal DNA barcoding.** *Genome* 2016, **59:**913-932.
5. Badotti F, de Oliveira FS, Garcia CF, Vaz ABM, Fonseca PLC, Nahum LA, Oliveira G, Góes-Neto A: **Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi).** *BMC Microbiology* 2017, **17:**42.
6. James TY, Stajich JE, Hittinger CT, Rokas A: **Toward a Fully Resolved Fungal Tree of Life.** 2020, **74:**291-313.
7. Mafune KK, Godfrey BJ, Vogt DJ, Vogt KA: **A rapid approach to profiling diverse fungal communities using the MinION™ nanopore sequencer.** 2020, **68:**72-78.
8. Tedersoo L, Anslan S: **Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences.** 2019, **11:**659-668.
9. Vilgalys R, Hester M: **Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several Cryptococcus species.** *Journal of bacteriology* 1990, **172:**4238-4246.
10. White TJ, Bruns T, Lee S, Taylor J: **38 - AMPLIFICATION AND DIRECT SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES FOR PHYLOGENETICS.** In *PCR Protocols.* Edited by Innis MA, Gelfand DH, Sninsky JJ, White TJ. San Diego: Academic Press; 1990: 315-322

11. GARDES M, BRUNS TD: **ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts.** 1993, **2:**113-118.

12. Raja HA, Miller AN, Pearce CJ, Oberlies NH: **Fungal Identification Using Molecular Tools: A Primer for the Natural Products Research Community.** *Journal of Natural Products* 2017, **80:**756-770.

13. Loit K, Adamson K, Bahram M, Puusepp R, Anslan S, Kiiker R, Drenkhan R, Tedersoo L: **Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens.** 2019, **85:**e01368-01319.

14. Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H: **Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification.** *Evolutionary bioinformatics online* 2008, **4:**193-201.

15. UNITE Community: **UNITE QIIME release for Fungi. Version 18.11.2018.** UNITE Community; 2019.

16. Choi J, Kim S-H: **A genome Tree of Life for the Fungi kingdom.** *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114:**9391-9396.

17. Rodrigues ML, Nosanchuk JD: **Fungal diseases as neglected pathogens: A wake-up call to public health officials.** *PLOS Neglected Tropical Diseases* 2020, **14:**e0007964.

18. Brown GD, Denning DW, Gow NAR, Levitz SM, Netea MG, White TC: **Hidden Killers: Human Fungal Infections.** 2012, **4:**165rv113-165rv113.

19. Almeida F, Rodrigues ML, Coelho C: **The Still Underestimated Problem of Fungal Diseases Worldwide.** *Frontiers in microbiology* 2019, **10:**214-214.

20. Powell D, Jones A, Kent N, Kaur P, Bar I, Schwessinger B, Frère CH: **Genome Sequence of the Fungus Nannizziopsis barbatae, an Emerging Reptile Pathogen.** 2021, **10:**e01213-01220.

21. Lips KR: **Overview of chytrid emergence and impacts on amphibians.** *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2016, **371:**20150465.

22. Fensham RJ, Carnegie AJ, Laffineur B, Makinson RO, Pegg GS, Wills J: **Imminent Extinction of Australian Myrtaceae by Fungal Disease.** *Trends in Ecology & Evolution* 2020, **35:**554-557.

23. Birren B, Fink G, Lander EJC, MA: Whitehead Institute Center for Genome Research: **Fungal Genome Initiative: white paper developed by the fungal research community.** 2002.

24. Arratia A, Sepúlveda E: **Convolutional Neural Networks, Image Recognition and Financial Time Series Forecasting.** In*; Cham.* Springer International Publishing; 2020: 60-69.

25. Tarca AL, Carey VJ, Chen X-w, Romero R, Drăghici S: **Machine learning and its applications to biology.** *PLoS computational biology* 2007, **3:**e116-e116.

26. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics.** *Nature reviews Genetics* 2015, **16:**321-332.

27. Boža V, Brejová B, Vinař T: **DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads.** *PLOS ONE* 2017, **12:**e0178751.

28. Flagel L, Brandvain Y, Schrider DR: **The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference.** *Mol Biol Evol* 2019, **36:**220-238.

29. Ji S, Xu W, Yang M, Yu K: **3D Convolutional Neural Networks for Human Action Recognition.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2013, **35:**221-231.

30. Yang C, Chu J, Warren RL, Birol I: **NanoSim: nanopore sequence read simulator based on statistical characterization.** *Gigascience* 2017, **6:**1-6.

31. Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, Irinyi L, Meyer W, Nilsson RH, Hughes K, Miller AN, et al: **Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi.** *Database : the journal of biological databases and curation* 2014, **2014:**bau061.

32. Robbertse B, Strope PK, Chaverri P, Gazis R, Ciufo S, Domrachev M, Schoch CL: **Improving taxonomic accuracy for fungi in public sequence databases: applying 'one name one species' in well-defined genera with Trichoderma/Hypocrea as a test case.** *Database : the journal of biological databases and curation* 2017, **2017:**bax072.

33. Hu Y, Green GS, Milgate AW, Stone EA, Rathjen JP, Schwessinger B: **Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer.** 2019, **3:**92-101.

34. Irinyi L, Hu Y, Hoang MTV, Pasic L, Halliday C, Jayawardena M, Basu I, McKinney W, Morris AJ, Rathjen J, et al: **Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics.** *Medical Mycology* 2019, **58:**650-660.

35. 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.** *Biology Methods and Protocols* 2020, **6**.

36. Heeger F, Bourne EC, Baschien C, Yurkov A, Bunk B, Spröer C, Overmann J, Mazzoni CJ, Monaghan MT: **Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments.** 2018, **18:**1500-1514.

37. Hofstetter V, Buyck B, Eyssartier G, Schnee S, Gindro K: **The unbearable lightness of sequenced-based identification.** *Fungal Diversity* 2019, **96:**243-284.

38. 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:**220-220.

39. Sanford RA, Lloyd KG, Konstantinidis KT, Löffler FE: **Microbial Taxonomy Run Amok.** *Trends in Microbiology* 2021, **29:**394-404.

40. Lücking R, Aime MC, Robbertse B, Miller AN, Aoki T, Ariyawansa HA, Cardinali G, Crous PW, Druzhinina IS, Geiser DM, et al: **Fungal taxonomy and sequence-based nomenclature.** *Nature Microbiology* 2021, **6:**540-548.

41. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A: **A primer on deep learning in genomics.** *Nature Genetics* 2019, **51:**12-18.

42. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, et al: **Opportunities and obstacles for deep learning in biology and medicine.** 2018, **15:**20170387.

43. Bulman SR, McDougal RL, Hill K, Lear G: **Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity.** *Australasian Plant Pathology* 2018, **47:**467-474.

44. Hu Y, Wilson S, Schwessinger B, Rathjen JP: **Blurred lines: integrating emerging technologies to advance plant biosecurity.** *Current Opinion in Plant Biology* 2020, **56:**127-134.

45. Rodrigues MS, Morelli KA, Jansen AM: **Cytochrome c oxidase subunit 1 gene as a DNA barcode for discriminating Trypanosoma cruzi DTUs and closely related species.** *Parasites & Vectors* 2017, **10:**488.

46. Stielow JB, Lévesque CA, Seifert KA, Meyer W, Iriny L, Smits D, Renfurm R, Verkley GJM, Groenewald M, Chaduli D, et al: **One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes.** *Persoonia* 2015, **35:**242-263.

47. Meyer W, Irinyi L, Hoang MTV, Robert V, Garcia-Hermoso D, Desnos-Ollivier M, Yurayart C, Tsang C-C, Lee C-Y, Woo PCY, et al: **Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α).** 2019, **62:**160-169.

48. Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, Yeo D, Meier R: **MinION barcodes: biodiversity discovery and identification by everyone, for everyone.** 2021**:**2021.2003.2009.434692.

49. Ferrer C, Colom F, Frasés S, Mulet E, Abad JL, Alió JL: **Detection and identification of fungal pathogens by PCR and by ITS2 and 5.8S ribosomal DNA typing in ocular infections.** *Journal of clinical microbiology* 2001, **39:**2873-2879.

50. **Amplicon sequencing using MinION optimized from 1D native barcoding genomic DNA**

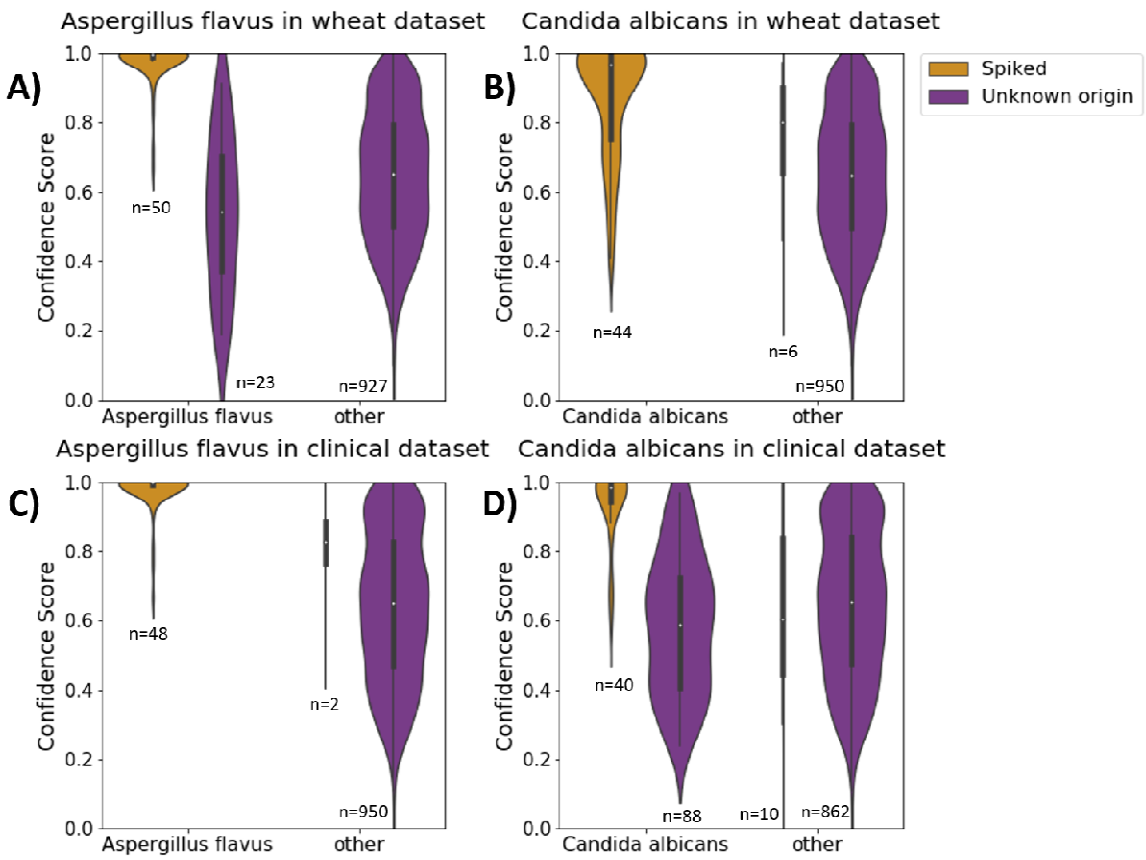
51. Narutaki S, Takatori K, Nishimura H, Terashima H, Sasaki T: **Identification of fungi based on the nucleotide sequence homology of their internal transcribed spacer 1 (ITS1) region.** *PDA J Pharm Sci Technol* 2002, **56:**90-98.

52. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al: **Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.** 2009, **75:**7537-7541.

53. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic acids research* 2004, **32:**1792-1797.

54. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R: **The EMBL-EBI search and sequence analysis tools APIs in 2019.** *Nucleic acids research* 2019, **47:**W636-W641.

55. Chollet F: **Keras.** GitHub; 2015.



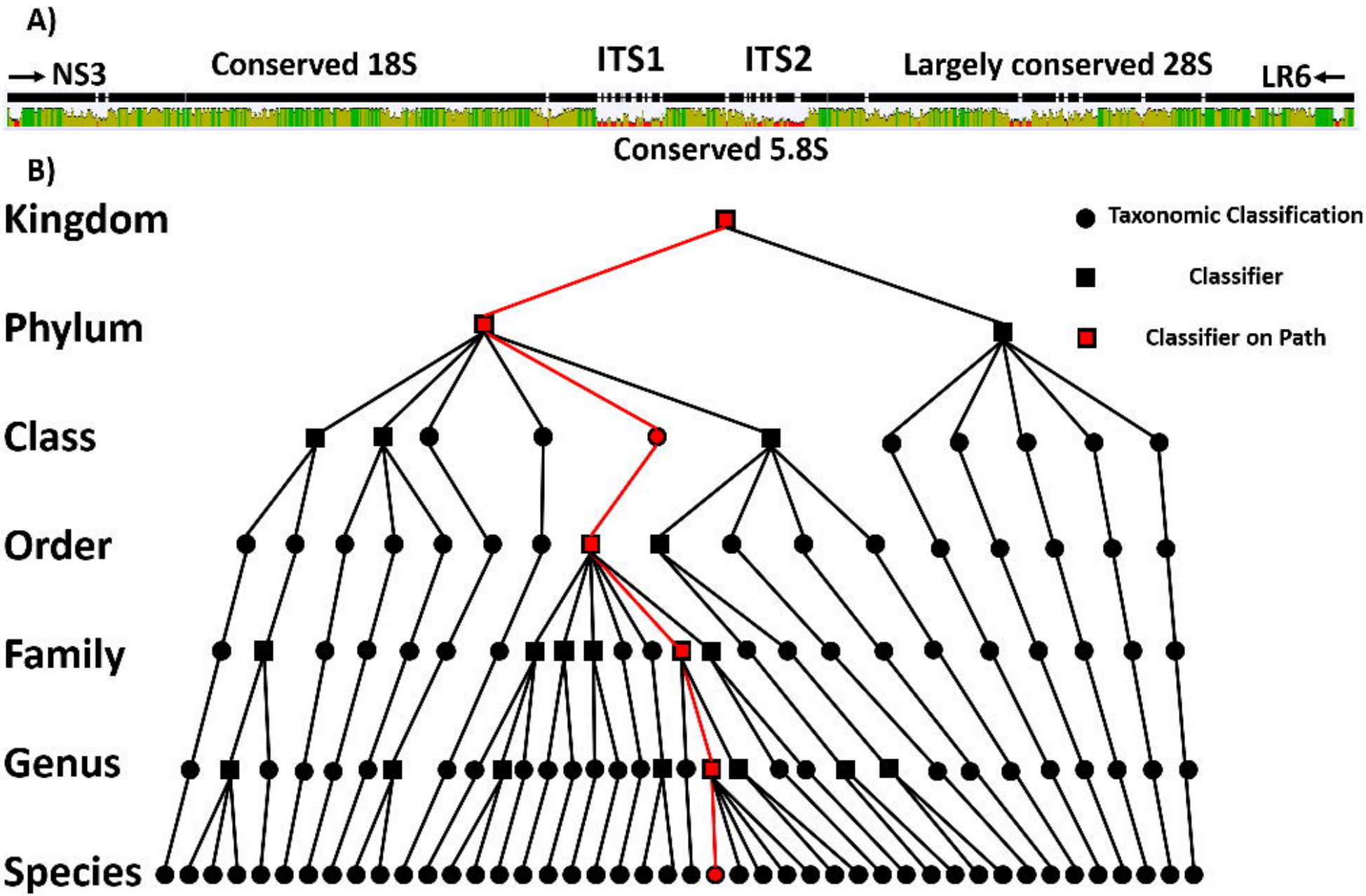
Supplementary Figure S1: The propagated confidence score is markedly different between spiked reads with known labels and reads of unknown origin.

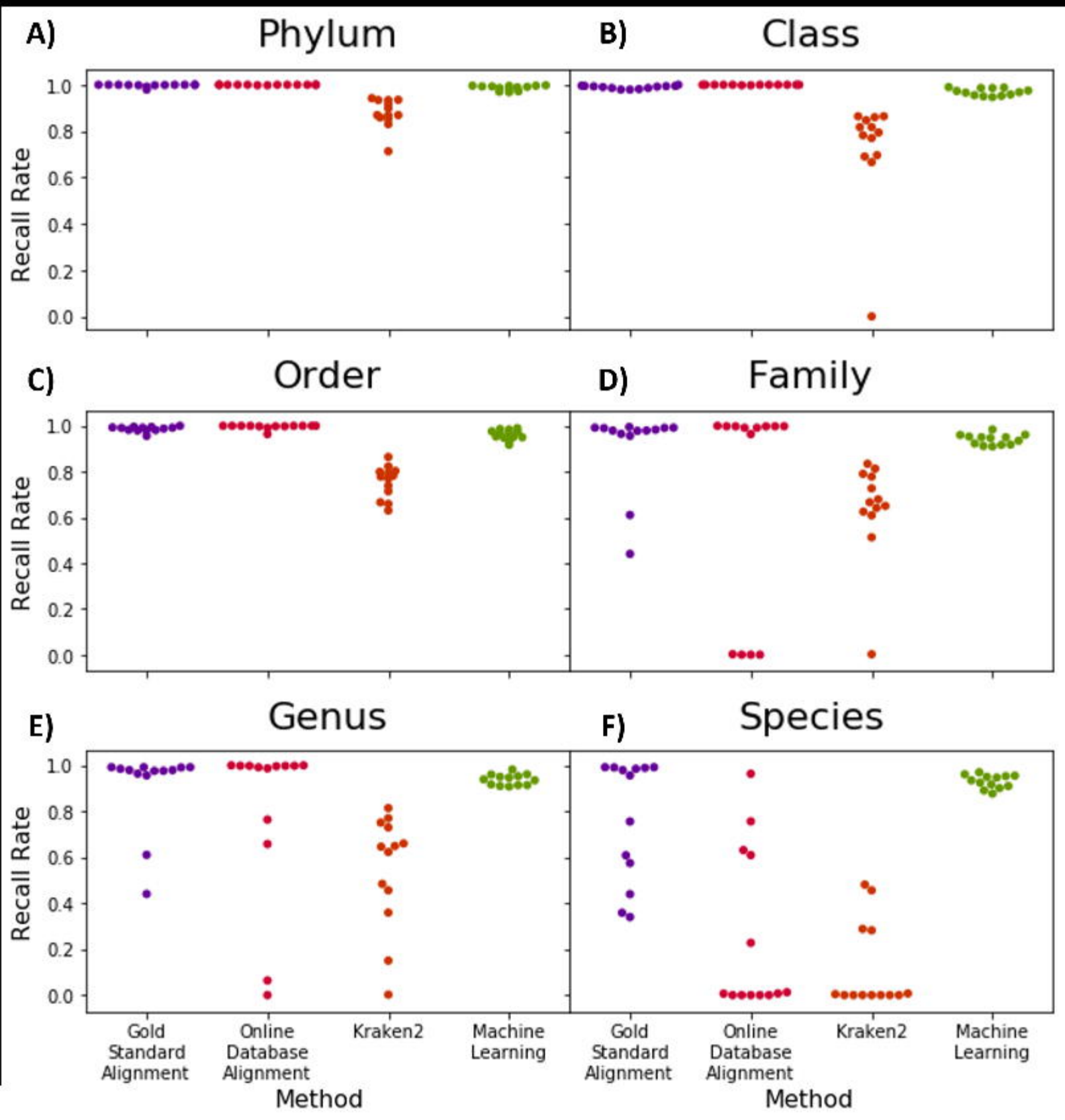
The plots show the confidence score of spiked labelled reads (yellow) and reads of unknown origin (purple) at the species level when analysed with our machine learning approach. Labelled reads of species for which we trained classifiers had a markedly better confidence score at the species level, especially when classified into the correct species. Reads of unknown origin had relatively low confidence scores independent of their species level classification. Our 1000-read *in silico* spike in samples are comprised of 950 reads from a complex sample with reads of unknown origin and 50 *in silico* spiked reads in samples with known labels from *Aspergillus flavus* (A and C) or *Candida albicans* (B and D). The complex samples of unknown composition are derived from infected wheat leaves (wheat dataset, A and B) or from a bronchoalveolar wash taken in a clinical setting (clinical dataset, C and D. The number of spiked or unspiked reads classified as either the spiked species or any other species is shown.

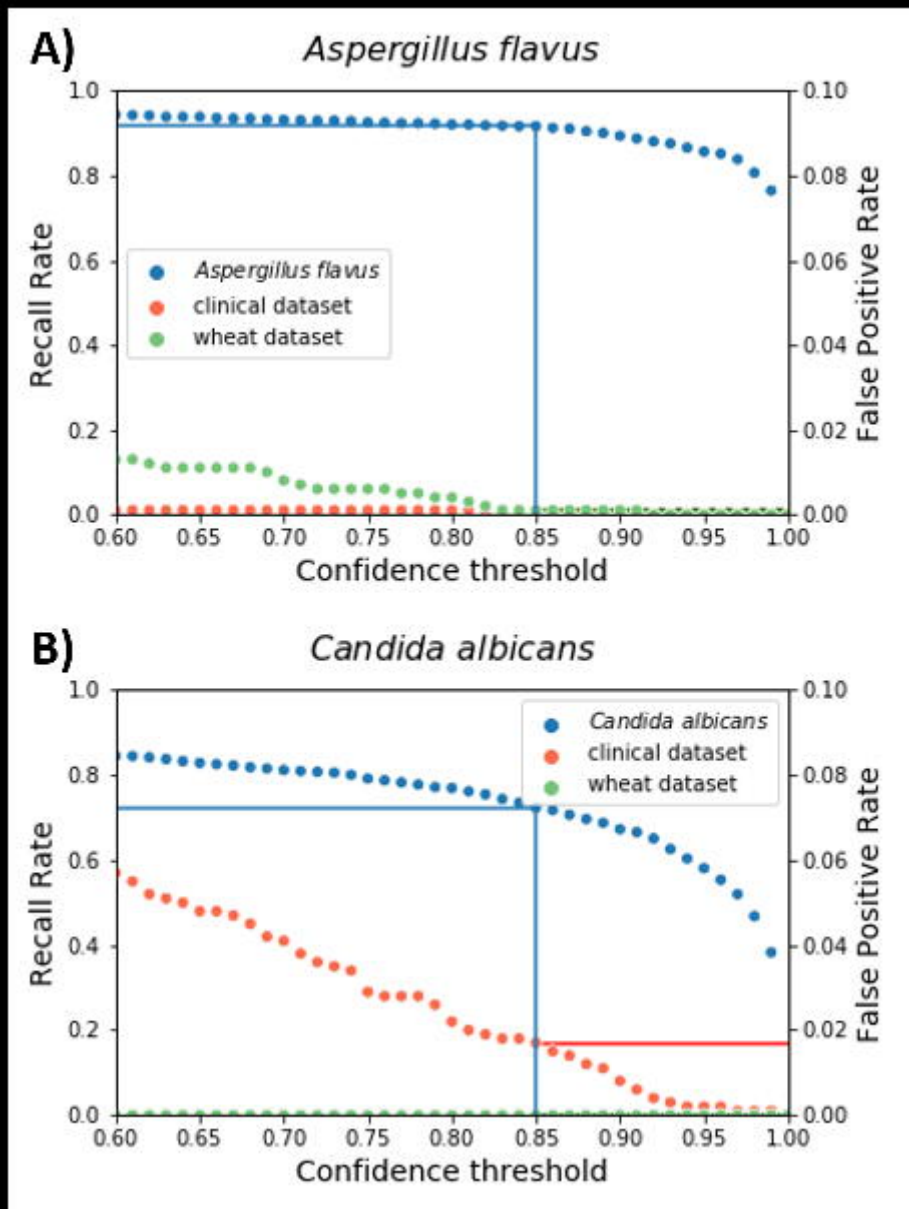
Supplementary Table T1

Genus label	Species label	Strain/Isolate	Sample Collection	DNA Extraction Method	# Raw Reads	# Homology Filtered	# Length Filtered
Aspergillus	*flavus*	WM03.230	Tissue from SDA plate [a]	Phenol chloroform	125014	54061	51340
Aspergillus	*niger*	WM06.98	Tissue from SDA plate [a]	Zymo kit	171615	65065	59406
Aspergillus	sp.	CCL015	Tissue from PDA plate	Qiagen kit	249468	42899	39988
Blastobotrys	*proliferans*	WM07.12	Tissue from SDA plate [a]	Phenol chloroform	133835	39631	37315
Candida	*albicans*	WM229	Tissue from SDA plate [a]	Zymo kit	91031	26114	25597
Candida	*metapsilosis*	WM01.56	Tissue from SDA plate [a]	Zymo kit	127633	40257	37426
Candida	*orthopsilosis*	WM03.414	Tissue from SDA plate [a]	Zymo kit	104214	32490	29765
Candida	*parapsilosis*	WM02.200	Tissue from SDA plate [a]	Zymo kit	135720	45958	42338
Candida	sp.	WM28	Tissue from SDA plate [a]	Zymo kit	109905	38085	35120
Cladophialophora	sp.	CLM599	Tissue from PDA plate [b]	Qiagen kit	115477	28063	25917
Clavispora	*lusitaniae*	WM18	Tissue from SDA plate [a]	Zymo kit	352768	141856	131936
Cortinarius	*globuliformis*	CM4	Fruiting tissue	Qiagen kit	347423	128993	117090
Cryptococcus	*zero*	CCL040	Tissue from PDA plate [b]	Qiagen kit	167818	42373	39235
Debaryomyces	sp.	WM03.458	Tissue from SDA plate [a]	Phenol chloroform	174974	35837	33499
Diaporthe	*foeniculina*	CCL060	Tissue from PDA plate [b]	Qiagen kit	206161	42329	39836
Diaporthe	sp.		Tissue from PDA plate [b]	Qiagen kit	198500	29833	27941
Discula	*quercina*	CCL067	Tissue from PDA plate [b]	Qiagen kit	172601	32847	30504
Discula	*quercina*	CCL068	Tissue from PDA plate [b]	Qiagen kit	188353	33438	31996
Dothiorella	*vidmadera*		Tissue from PDA plate [b]	Qiagen kit	204777	47318	44257
Entoleuca	sp.	CCL052	Tissue from PDA plate [b]	Qiagen kit	155158	33941	31356
Fusarium	*oxysporum*	Race3	Tissue from PDA plate [b]	Qiagen kit	382450	131411	123742
Galactomyces	*geotrichum*	WM17.23	Tissue from SDA plate [a]	Phenol chloroform	152933	8485	7805
Kluyveromyces	*marxianus*	WM13	Tissue from SDA plate [a]	Zymo kit	115282	31150	28382
Kluyveromyces	sp.	WM04.172	Tissue from SDA plate [a]	Zymo kit	370154	165113	152736
Kodamaea	*ohmeri*	WM10.200	Tissue from SDA plate [a]	Phenol chloroform	111257	38931	36478
Meyerozyma	*guilliermondii*	WM02.361	Tissue from SDA plate [a]	Phenol chloroform	211853	20333	18944
Penicillium	*chrysogenum*	WM06.341	Tissue from SDA plate [a]	Zymo kit	192173	78105	72307
Pichia	*kudriavzevii*	WM03.103	Tissue from SDA plate [a]	Zymo kit	122601	35604	33244
Pichia	*membranifaciens*	WM324	Tissue from SDA plate [a]	Zymo kit	104844	29540	26937
Puccinia	*striiformis-tritici*	104E	Fungal spores	Phenol chloroform	272465	122080	113337
Pyrenophora	*tritici-repentis*	Ptr8814	Tissue from PDA plate [b]	Qiagen kit	260896	97584	90015
Quambalaria	*cyanescens*	CCL055	Tissue from PDA plate [b]	Qiagen kit	205404	49780	46171
Rhodotorula	*mucilaginosa*	WM09.204	Tissue from SDA plate [a]	Zymo kit	318405	127801	117801
Saccharomyces	*cerevisiae*	YH2Gold	Tissue from YPD media [c]	Qiagen kit	96837	33025	30260
Scedosporium	*boydii*	WM09.122	Tissue from SDA plate [a]	Zymo kit	331947	102481	93723
Tapesia	*yallundae*	CCL029	Tissue from PDB	Qiagen kit	223186	59651	55589
Tapesia	*yallundae*	CCL031	Tissue from PDB	Qiagen kit	213143	52944	49481
Tuber	*brumale*		Fruiting tissue	Qiagen kit	275035	80614	74232
Wickerhamomyces	*anomalus*	WM03.505	Tissue from SDA plate [a]	Phenol chloroform	193187	45720	42589
Yamadazyma	*mexicana*	WM805	Tissue from SDA plate [a]	Phenol chloroform	179240	45093	42369
Yamadazyma	*scolyti*	WM06.835	Tissue from SDA plate [a]	Phenol chloroform	136650	37159	34841
Yarrowia	*lipolytica*	WM599	Tissue from SDA plate [a]	Phenol chloroform	141238	35950	33873
Zygoascus	*hellenicus*	WM02.460	Tissue from SDA plate [a]	Phenol chloroform	229073	36666	34002
Zymoseptoria	*tritici*	WA332	Tissue from PDA plate [b]	Qiagen kit	413127	143363	133089

Sample labels, collection methods, DNA extraction methods and read counts before and after two-step data filtering. a) Sabourand dextrose agar (SDA); b) Potato dextrose agar (PDA); c) Yeast extract peptone dextrose (YPD)


A)
NS3
Conserved 18S
ITS1
ITS2
Largely conserved 28S
LR6
Conserved 5.8S
B)
Kingdom
Phylum
Class
Order
Family
Genus
Species
Taxonomic Classification
Classifier
Classifier on Path


A)
Phylum
B)
Class
C)
Order
D)
Family
E)
Genus
F)
Species
Recall Rate
1.0
0.8
0.6
0.4
0.2
0.0
Gold Standard Alignment
Online Database Alignment
Kraken2
Machine Learning
Method


A)
Aspergillus flavus
Recall Rate
False Positive Rate
Confidence threshold
Aspergillus flavus
clinical dataset
wheat dataset
B)
Candida albicans
Recall Rate
False Positive Rate
Confidence threshold
Candida albicans
clinical dataset
wheat dataset