

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

3 **Tavish G. Eenjes^{1†}, Yiheng Hu^{1,7†}, Laszlo Irinyi^{2,3,4}, Minh Thuy Vi**
4 **Hoang^{2,3,4}, Leon M. Smith^{1,8}, Celeste C. Linde¹, Wieland Meyer^{2,3,4,5},**
5 **Eric A. Stone^{1,6}, John P. Rathjen¹, Benjamin Mashford^{6*}, Benjamin**
6 **Schwessinger^{1*}**

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

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

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

13 ⁴ Westmead Institute for Medical Research, Westmead, NSW Australia.

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

15 ⁶ ANU-CSIRO Centre for Genomics, Metabolomics and Bioinformatics, Canberra, ACT,
16 Australia

17 ⁷ Current address: Department of Microbial Interactions, IMIT/ZMBP, University of
18 Tübingen, Tübingen, Germany

19 ⁸ Diversity Arrays Technology, Canberra, ACT, Australia

20 [†] These authors contributed equally to this work.

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

24

25 **ABSTRACT**

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

34

35 **KEYWORDS**

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

37

38 **BACKGROUND**

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

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

53 With the advent and increasing use of long-read sequencing, such as that enabled by the
54 nanopore sequencing technology of the MinION from Oxford Nanopore Technologies
55 (ONT), some of the limitations of short-reads can be bypassed [7]. With long-reads, an
56 extended ITS region can be sequenced including both ITS1 and ITS2 in addition to the minor
57 variable regions of the 18S and 28S rRNA subunits using one set of ‘universal’ primers [8-
58 11]. Here, we focus on the region amplified by the NS3 and LR6 primers [12], spanning close
59 to 2.9 kbp in size. We refer to this amplicon hereafter as the fungal ribosomal DNA region.
60 Nanopore sequencing introduces a relatively high read error of around 10% at the time of
61 conducting our study [13]. These make individual reads less suited for species identification
62 using DNA metabarcodes combined with currently existing sequence alignment and k-mer
63 based methods because the genetic distance of the variable regions between closely related
64 species are often lower than the per read error rate [14]. In addition, the entries in most fungal
65 DNA barcode databases, such as NCBI and Unite, are relatively short with a median
66 sequence length of 580 bp and 540 bp [15], respectively. This limits the analysis capacity of
67 long-reads which completely entail both ITS sequences and include minor variable regions in
68 both 18S and 28S rRNA.

69 In our current study we address these shortcomings and assess the applicability of novel
70 sequence analysis methods for metabarcodes using the fungal kingdom as a test case. The
71 fungal kingdom is diverse, with an estimated 1.5-5 million species globally, performing
72 important ecosystem functions [16]. At the same time fungi can have adverse effects on
73 human and animal health and agriculture. An estimated 300 million people suffer from
74 fungal-related diseases each year [17], which often have a high mortality rate and limited
75 treatment options, resulting in the deaths of over 1.5 million people annually [18]. Similarly,
76 fungi can cause large-scale biodiversity loss [19, 20] as demonstrated by the near extinction
77 of many amphibian taxa by the globally devastating fungal pathogen *Batrachochytrium*
78 *dendrobatidis* [21] and the local extinction of several myrtaceae tree species by the rust
79 fungus *Austropuccinia psidii* [22]. Fungal pathogens also cause an estimated loss of about
80 \$200 billion dollars in global food production annually [23]. The importance of fungi
81 warrants the development of improved sequence-based detection methods for fungi as
82 illustrated in our proof-of-concept study.

83 We explored machine learning classifiers as an alternative method for assigning individual
84 error-prone sequence long-reads to taxa, because machine learning techniques are ideally
85 suited to identify deterministic spatial relationships between features for classification [24].
86 For example, it might be that specific DNA bases have a unique spatial relationship within
87 the fungal ribosomal DNA region that is deterministic for a given fungal species. These
88 relationships are difficult to capture with currently available (local) alignment or k-mer based
89 methods when combined with error-prone sequence long-reads, especially when these
90 features (DNA bases) are not located in close proximity in the primary DNA sequence. There
91 exist many machine learning methods for identifying patterns across a variety of data types
92 [25-27]. Convolutional neural networks (CNNs) are one type of machine learning methods
93 that are especially suited for identifying the deterministic spatial relationships in DNA

94 sequence, as they are capable of learning from both small-scale and higher order
95 discretionary features, including important spatial relationships between said features [24, 28,
96 29]. So, we applied a CNN approach to metabarcoding based fungal species identification
97 using a uniquely labelled sequencing dataset of the 2.9 kbp fungal ribosomal DNA region
98 from 44 individually sequenced fungal species. We compared our machine learning approach
99 to three commonly used analysis approaches including alignment and k-mer based methods
100 on different in house and publicly available databases. Our machine learning approach faired
101 especially well when identifying closely related species. Furthermore, we show that the
102 training of a limited set of general and specific machine learning taxa classifiers provides a
103 reasonable approach to targeted species identification from a complex sample of unknown
104 composition.

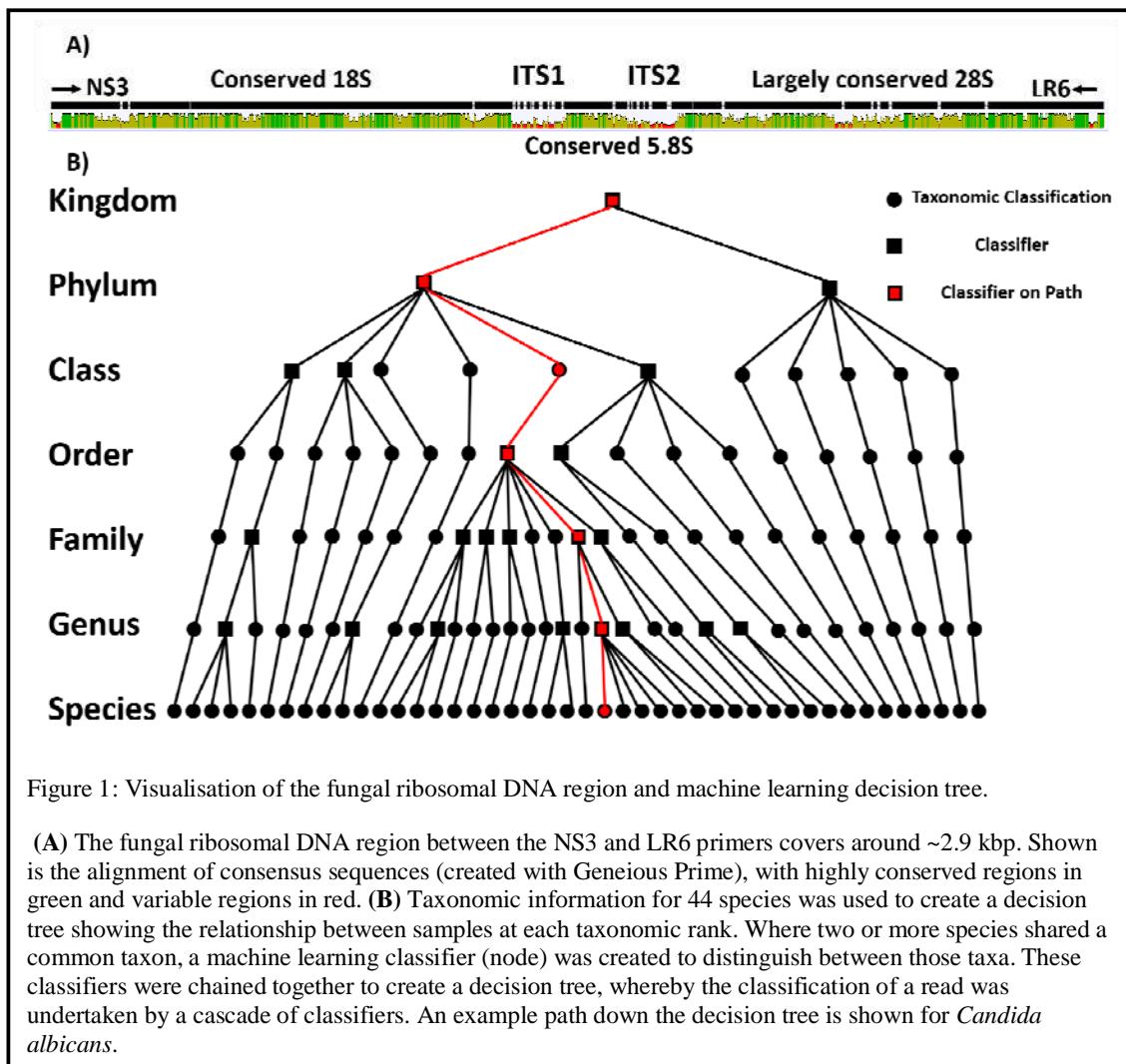
105

106 **RESULTS**

107 **Design of a decision tree for machine learning classifiers for taxonomic assignment of 108 fungal species**

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

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



132

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

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

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

157 **Comparison of methods for species classification of fungal pathogens**

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

167 approach because the gold standard and machine learning approaches are directly derived
168 from our sequencing dataset. To compare the machine learning approach with methods where
169 the sequencing data was not used to create the classification database in some way, we
170 applied *minimap2* to a large publicly-available database of fungal ITS sequences from NCBI
171 [31, 32] (NCBI alignment), and applied *Kraken2*, a k-mer-based algorithm designed for use
172 with metagenomic DNA sequences, to the same NCBI database (Kraken2).

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

182 Our machine learning decision tree approach maintained a consistently high recall rate across
183 all taxonomic ranks, with a mean species level recall rate of $93.0 \pm 2.8\%$. Notably, it
184 performed very well for closely related taxa, including the cryptic species *Candida*
185 *metapsilosis* and *Candida orthopsilosis* and another closely related species *Candida albicans*.
186 The two cryptic *Candida* species (*C. metapsilosis* and *C. orthopsilosis*) had a very high
187 consensus sequence similarity, with a genetic distance of 2.74% (97.26% identity) in our
188 fungal ribosomal DNA region target region representing the genetically least distinct species
189 pair. Our machine learning approach did achieve species level recall rates of 90.1% and
190 89.1% for *C. metapsilosis* and *C. parapsilosis*, respectively, even with per read error rates of
191 about 10%. This highlights the strength of our approach.

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

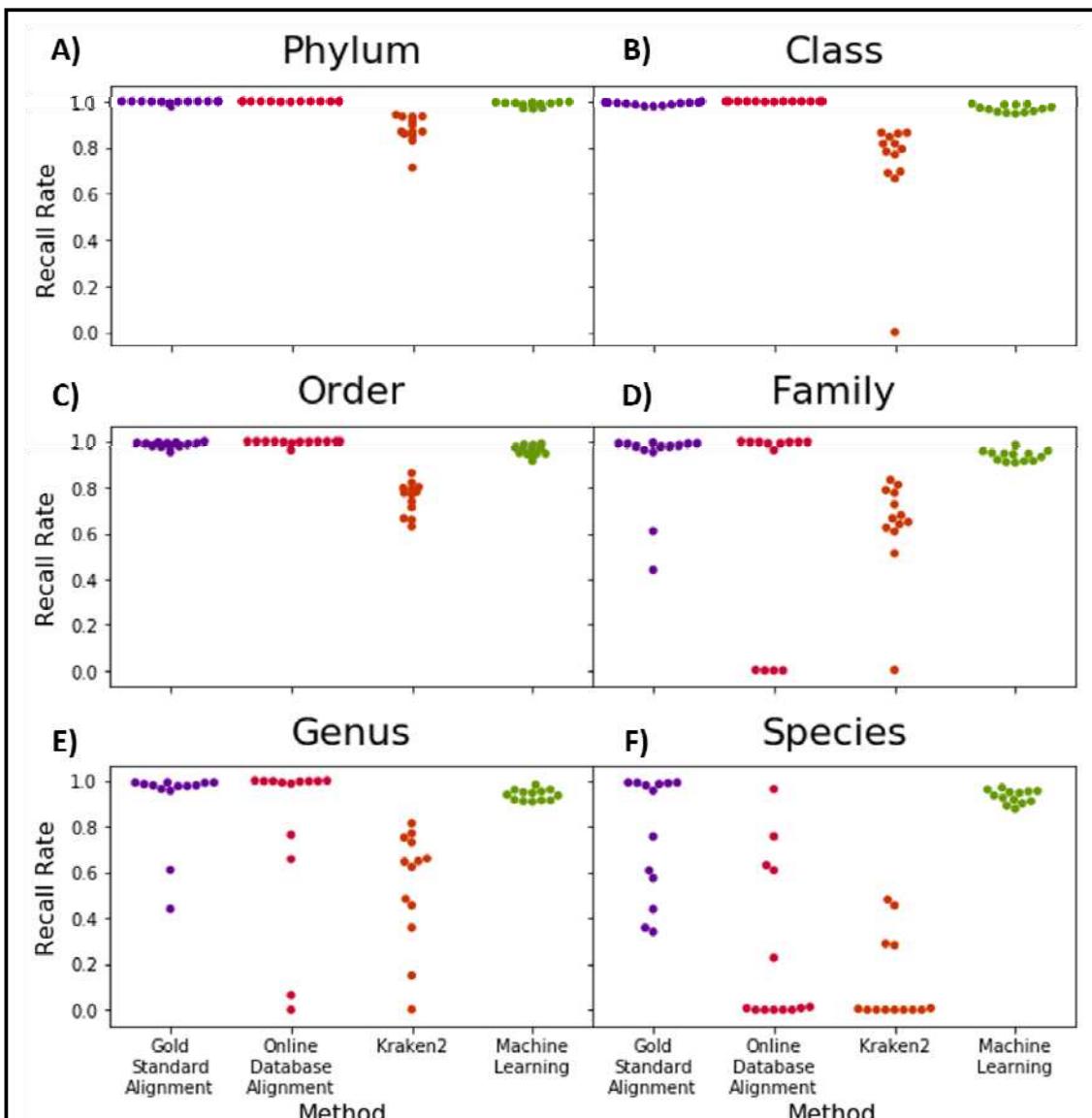


Figure 2: Machine learning based species identification performs especially well for closely related species

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.

201

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

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

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

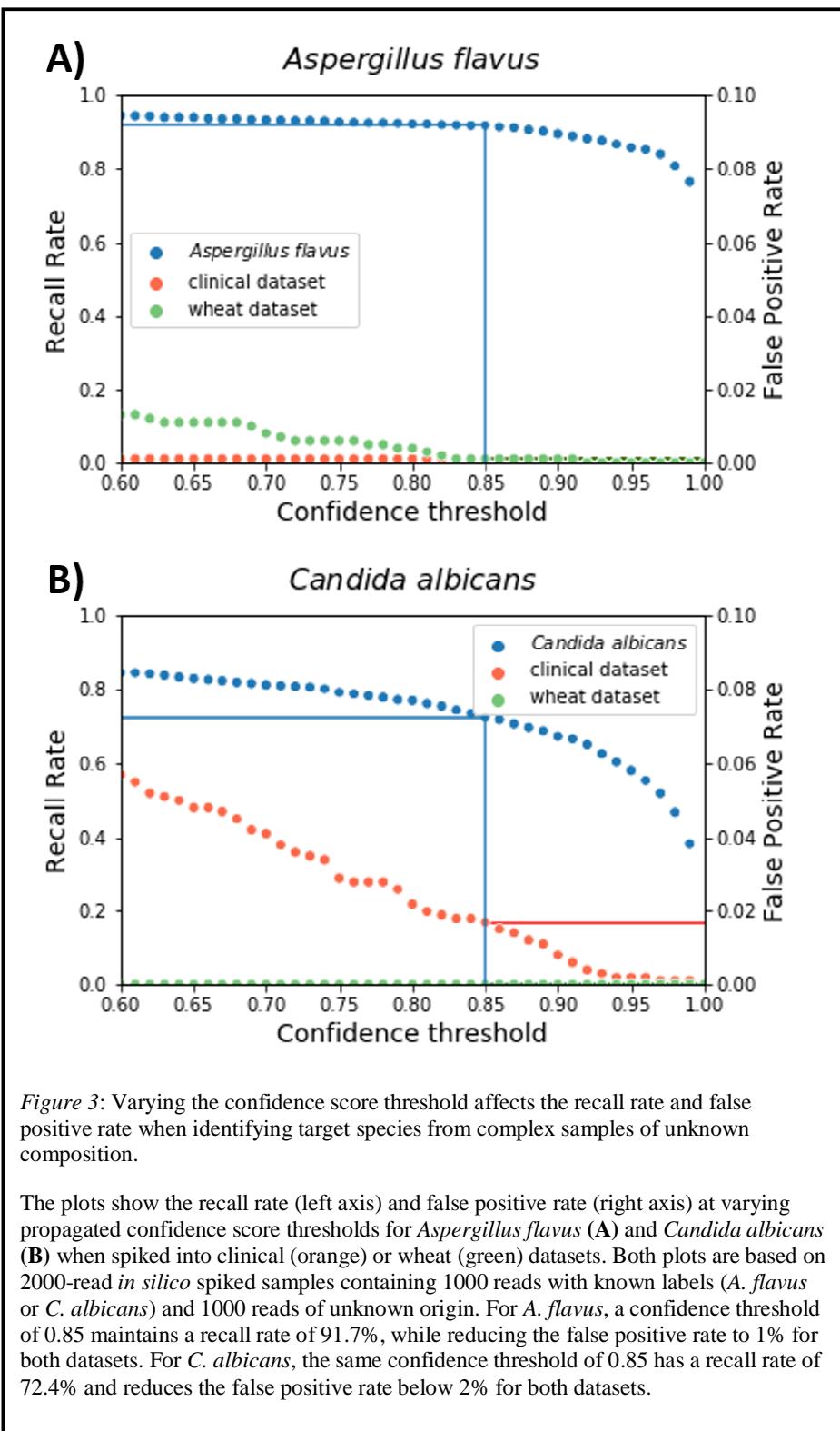
213 A key feature of a species classification tool is its ability to identify a known target species
214 from a complex sample of unknown composition. This is especially important when
215 attempting to identify the presence of a target species, such as a specific pathogen, from a
216 metagenomic sample.

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

228 We first plotted the propagated confidence score of the species level classification for all
229 reads in each *in silico* spiked dataset to better understand the behaviour of our machine
230 learning decision tree on samples containing reads of unknown origin (Supplemental Figure
231 S1). This clearly shows that the propagated confidence scores for reads of unknown origin
232 are far lower than reads of species the classifiers were trained on. We then assessed the recall
233 and false positive rate of the *in silico* spiked datasets at different confidence scores thresholds
234 (Figure 3). Increasing the thresholds reduced the recall and false positive rate in both cases.
235 For *A. flavus*, the recall rate remained above 90% until the confidence threshold reached 0.9,
236 and the false positive rate was consistently low across both the clinical and wheat datasets
237 with reads of unknown origin. A confidence threshold of 0.85 resulted in a high recall rate of
238 0.917, while maintaining a low false positive rate of just one percent. For *C. albicans*, not
239 using a confidence threshold at all resulted in a recall rate of 87.7% and false positive rate of
240 11.7%. However, by using a confidence threshold of 0.85, the recall rate was only decreased
241 to 72.4% while reducing the false positive rate to only 1.7% in the clinical dataset. We
242 recommend this confidence score threshold of 0.85 as suitable for retaining a high recall rate
243 while achieving a low false positive rate, even for a member of a difficult-to-distinguish
244 genus like *Candida*.

245

246



249 **DISCUSSION**

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

257 Our machine learning approach is comparable to – albeit slightly outperformed by - the gold
258 standard alignment approach across all taxonomic ranks for most of the species tested.

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

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

273 error prone long-reads with alignment (NCBI alignment) and k-mer (Kraken2) based
274 approaches using the NCBI ITS RefSeq Targeted Loci database performed relatively poorly
275 especially at lower taxonomic ranks. Clearly, the discrepancy between read and database
276 sequence lengths (~2900 bp vs ~580 bp) negatively impacted the alignment success.
277 Interestingly, the Kraken2 approach underperformed compared to the alignment-based
278 approach in our current study. This is consistent with previous work with long-read MinION
279 nanopore data, where Kraken2 classification success never exceeded that for BLAST, another
280 alignment-based classification program, when using the default 35 bp k-mers [38]. It is likely
281 using a smaller k-mer length would improve classification accuracy for long-read nanopore
282 sequencing due to the high read error, which impacts perfect matches for 35 bp k-mers.
283 Another common issue when using public databases for species identification was that many
284 species were not included in the NCBI database or present with different taxonomic labels,
285 which resulted in some family and species level recall rates being zero. Changing
286 nomenclature over time can be an issue when using these online databases when trying to
287 identify a species or detect the presence of a known, named species, as the nomenclature is
288 not always updated, leading to outdated or uncorrected taxonomic information persisting in
289 databases [39, 40].
290 We also tested if our machine learning approach can accurately identify specific target
291 species in complex samples of unknown composition without having classifiers for all fungal
292 species present in the sample. We were able to show that by only training a limited set of
293 classifiers we can detect target species with relatively low false positive and high recall rates
294 in *in silico* spiked datasets with known ground truth of the spiked reads only. By adjusting the
295 confidence score one can decide how much false positive and false negatives one is willing to
296 tolerate. We found a threshold of 0.85 on the propagated confidence score at the species level
297 classification was sufficient to reduce the false positive rate while maintaining high recall

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

317

318 CONCLUSIONS

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

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

330

331 MATERIALS AND METHODS

332 Fungal pathogen sample collection, DNA extraction and ITS amplification

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

335 We used three different DNA extraction methods for all the species in the mock
336 communities. The methods for each species are listed in the Supplemental Table T1.
337 Collectively, we used two commercially available kits: The Qiagen DNeasy Plant Mini Kit
338 (cat. no. 69106) for most of the plant pathogenic fungi, and the Quick-DNA Fungal/Bacterial
339 Miniprep Kit (cat. no. D6005, Zymo Research) for some of the human pathogenic fungi
340 following the manufacturer's protocol. We used a phenol chloroform-based DNA extraction
341 method for some other human pathogenic fungi modified from Ferrer et al [49]. Briefly, 100
342 mg of leaf tissue was homogenized, and cells were lysed using cetyl trimethylammonium
343 bromide (CTAB, Sigma-Aldrich) buffer (added RNase T1, Thermo Fisher, 1,000 units per
344 1750 µl), followed by a phenol/chloroform/isoamyl alcohol (25:24:1, Sigma-Aldrich)
345 extraction to remove protein and lipids. The DNA was precipitated with 700 µl of

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

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

363 **Library preparation and DNA sequencing using the MinION**

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

370 ethanol, the pellet was allowed to dry for 2 mins and the DNA was eluted in 51 μ l nuclease
371 free water and quantified using NanoDrop[®] (Thermo Fisher Scientific, USA) and Promega
372 Quantus[™] Fluorometer (cat. No. E6150, Promega, USA) follow the manufacturer's
373 instructions. All DNA samples showed a with absorbance ratio A260/A280 > 1.8 and
374 A260/A230 > 2.0 from the NanoDrop[®]. DNA was end-repaired using NEBNext Ultra II End-
375 Repair/ dA-tailing Module (cat. No. E7546, New England Biolabs (NEB), USA) by adding 7
376 μ l Ultra II End-Prep buffer, 3 μ l Ultra II End-Prep enzyme mix. The mixture was incubated at
377 20°C for 10 mins and 65°C for 10 mins. A 1x volume (60 μ l) Agencourt AMPure XP clean-
378 up was performed, and the DNA was eluted in 31 μ l nuclease free water. Barcoding reaction
379 was performed by adding 2 μ l of each native barcode and 20 μ l NEB Blunt/TA Master Mix
380 (cat. No. M0367, New England Biolabs (NEB), USA) into 18 μ l DNA, mixing gently and
381 incubating at room temperature for 10 mins. A 1x volume (40 μ l) Agencourt AMPure XP
382 clean-up was then performed, and the DNA was eluted in 15 μ l nuclease free water. Ligation
383 was then performed by adding 20 μ l Barcode Adapter Mix (EXP-NBD103 Native Barcoding
384 Expansion Kit, ONT, UK), 20 μ l NEBNext Quick Ligation Reaction Buffer, and Quick T4
385 DNA Ligase (cat. No. E6056, New England Biolabs (NEB), USA) to the 50 μ l pooled
386 equimolar barcoded DNA, mixing gently and incubating at room temperature for 10 mins.
387 The adapter-ligated DNA was cleaned-up by adding a 0.4x volume (40 μ l) of Agencourt
388 AMPure XP beads, incubating for 5 mins at room temperature and resuspending the pellet
389 twice in 140 μ l ABB provided in the SQK-LSK108 kit. The purified-ligated DNA was
390 resuspended by adding 15 μ l ELB provided in the SQK-LSK108 (ONT, UK) kit and
391 resuspending the beads. The beads were pelleted again, and the supernatant transferred to a
392 new 0.5 ml DNA LoBind tube (cat. No. 0030122348, Eppendorf, Germany).
393 In total, four independent sequencing reactions were performed on a MinION flow cell (R9.4,
394 ONT) connected to a MK1B device (ONT) operated by the MinKNOW software (version

395 2.0.2): 11 species for each flowcell. Each flow cell was primed with 1 ml of priming buffer
396 comprising 480 µl Running Buffer Fuel Mix (RBF, ONT) and 520 µl nuclease free water. 12
397 µl of amplicon library was added to a loading mix including 35 µl RBF, 25.5 µl Library
398 Loading beads (ONT library loading bead kit EXP-LLB001, batch number EB01.10.0012)
399 and 2.5 µl water with a final volume of 75 µl and then added to the flow cell via the SpotON
400 sample port. The “NC_48Hr_sequencing_FLOMIN106_SQK-LSK108” protocol was
401 executed through MinKNOW after loading the library and run for 48 h. Raw fast5 files were
402 processed using Albacore 2.3.1 software (ONT) for basecalling, barcode de-multiplexing and
403 quality filtering (Phred quality (Q) score of > 7) as per the manufacturer's recommendations.

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

406 **Processing and manipulation of fungal pathogen reads**

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

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

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

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

426 **Augmenting read datasets**

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

434 **Generating consensus sequences for each species**

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

440 **Determining the relationships between samples**

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

449 **Creation of asset of neural network classifiers to distinguish between samples**

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

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

465 but an equal length was required to avoid sequence length being a distinguishing factor in the
466 classifier, all sequences were padded out to a length of 5,000 bp. The padding used a value of
467 4 to avoid the padding data from affecting the identification of key features for classification.

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

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

478 **Evaluation of the machine learning classifiers**

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

486
$$Recall\ Rate = \frac{True\ Positives}{True\ Positives+False\ Negatives}$$
 [1]

487 **Chaining machine learning classifiers into a decision tree**

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

496 **Alternative methods for fungal pathogen read classification**

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

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

512 sequences from NCBI
513 (<ftp://ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Fungi/fungi.ITS.fna.gz>, downloaded Feb
514 2021). *Minimap2* was applied to each of these databases using the map-ont flag. As the
515 alignment tool can return multiple hits if alignment is good enough, only the best hit was
516 taken for each read.

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

522 **Identifying a key species from a complex sample using machine learning**

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

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

538
$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \quad [2]$$

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

548

549 **DECLARATIONS**

550 **Ethics Approval**

551 Not applicable.

552 **Consent for Publication**

553 Not applicable.

554 **Availability of data and materials**

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

557 generated and/or analysed during the current study are available in SRA under BioProject

558 PRJNA725648, available at <http://www.ncbi.nlm.nih.gov/bioproject/725648>.

559 **Competing interests**

560 The authors declare that they have no competing interests.

561 **Funding**

562 This work was supported by an NHMRC grant (#GNT1121936) to WM.). BS

563 is supported by an ARC Future Fellowship (FT1801000024).

564 **Authors' contributions**

565 TGE, YH, BM, and BS designed the experiments and performed the analysis. YH extracted

566 fungal DNA and performed all sequencing reactions. LL, MTV, LMS, CCL, and WM

567 provided fungal material and/or DNA. ES and JR provided feedback on experimental design

568 and data analysis. WM, ES, JR, and BS provided funding for the project. TGE and BS wrote

569 the manuscript. All authors commented on the manuscript and approved submission.

570 **Acknowledgements**

571 We thank Eduardo Eyras, Jen Taylor, and Peter Solomon for suggestions on improving the

572 machine learning models and determining statistics for identifying a species from a complex

573 sample. We thank Peter Solomon and David Jones for providing fungal samples. We thank

574 Andrew Milgate for providing infected wheat leaf material.

575 This work was supported by computational resources provided by the Australian Government

576 through the National Computational Infrastructure (NCI) under the ANU Merit Allocation

577 Scheme.

578 **REFERENCES**

579 1. Seifert KA: **Progress towards DNA barcoding of fungi.** 2009, **9**:83-89.

580 2. Valentini A, Pompanon F, Taberlet P: **DNA barcoding for ecologists.** *Trends in*
581 *Ecology & Evolution* 2009, **24**:110-117.

582 3. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W:
583 **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA**
584 **barcode marker for Fungi.** 2012, **109**:6241-6246.

585 4. Xu J: **Fungal DNA barcoding.** *Genome* 2016, **59**:913-932.

586 5. Badotti F, de Oliveira FS, Garcia CF, Vaz ABM, Fonseca PLC, Nahum LA, Oliveira
587 G, Góes-Neto A: **Effectiveness of ITS and sub-regions as DNA barcode markers**
588 **for the identification of Basidiomycota (Fungi).** *BMC Microbiology* 2017, **17**:42.

589 6. James TY, Stajich JE, Hittinger CT, Rokas A: **Toward a Fully Resolved Fungal**
590 **Tree of Life.** 2020, **74**:291-313.

591 7. Mafune KK, Godfrey BJ, Vogt DJ, Vogt KA: **A rapid approach to profiling diverse**
592 **fungal communities using the MinION™ nanopore sequencer.** 2020, **68**:72-78.

593 8. Tedersoo L, Anslan S: **Towards PacBio-based pan-eukaryote metabarcoding**
594 **using full-length ITS sequences.** 2019, **11**:659-668.

595 9. Vilgalys R, Hester M: **Rapid genetic identification and mapping of enzymatically**
596 **amplified ribosomal DNA from several Cryptococcus species.** *Journal of*
597 *bacteriology* 1990, **172**:4238-4246.

598 10. White TJ, Bruns T, Lee S, Taylor J: **38 - AMPLIFICATION AND DIRECT**
599 **SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES FOR**
600 **PHYLOGENETICS.** In *PCR Protocols*. Edited by Innis MA, Gelfand DH, Sninsky
601 JJ, White TJ. San Diego: Academic Press; 1990: 315-322

602 11. GARDES M, BRUNS TD: **ITS primers with enhanced specificity for**
603 **basidiomycetes - application to the identification of mycorrhizae and rusts.** 1993,
604 **2:113-118.**

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

608 13. Loit K, Adamson K, Bahram M, Puusepp R, Anslan S, Kiiker R, Drenkhan R,
609 **Tedersoo L: Relative Performance of MinION (Oxford Nanopore Technologies)**
610 **versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in**
611 **Identification of Agricultural and Forest Fungal Pathogens.** 2019, **85:e01368-**
612 **01319.**

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

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

619 16. Choi J, Kim S-H: **A genome Tree of Life for the Fungi kingdom.** *Proceedings of*
620 *the National Academy of Sciences of the United States of America* 2017, **114:9391-**
621 **9396.**

622 17. Rodrigues ML, Nosanchuk JD: **Fungal diseases as neglected pathogens: A wake-up**
623 **call to public health officials.** *PLOS Neglected Tropical Diseases* 2020,
624 **14:e0007964.**

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

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

629 20. Powell D, Jones A, Kent N, Kaur P, Bar I, Schwessinger B, Frère CH: **Genome**
630 **Sequence of the Fungus Nannizziopsis barbatae, an Emerging Reptile Pathogen.**
631 2021, **10**:e01213-01220.

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

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

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

641 24. Arratia A, Sepúlveda E: **Convolutional Neural Networks, Image Recognition and**
642 **Financial Time Series Forecasting.** In; *Cham.* Springer International Publishing;
643 2020: 60-69.

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

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

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

650 28. Flagel L, Brandvain Y, Schrider DR: **The Unreasonable Effectiveness of**
651 **Convolutional Neural Networks in Population Genetic Inference.** *Mol Biol Evol*
652 2019, **36**:220-238.

653 29. Ji S, Xu W, Yang M, Yu K: **3D Convolutional Neural Networks for Human Action**
654 **Recognition.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2013,
655 **35**:221-231.

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

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

662 32. Robbertse B, Strope PK, Chaverri P, Gazis R, Ciufo S, Domrachev M, Schoch CL:
663 **Improving taxonomic accuracy for fungi in public sequence databases: applying**
664 **'one name one species' in well-defined genera with Trichoderma/Hypocrea as a**
665 **test case.** *Database : the journal of biological databases and curation* 2017,
666 **2017**:bax072.

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

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

675 35. D'Andreano S, Cuscó A, Francino O: **Rapid and real-time identification of fungi**
676 **up to species level with long amplicon nanopore sequencing from clinical**
677 **samples.** *Biology Methods and Protocols* 2020, **6**.

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

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

683 38. Pearman WS, Freed NE, Silander OK: **Testing the advantages and disadvantages**
684 **of short- and long- read eukaryotic metagenomics using simulated reads.** *BMC*
685 *bioinformatics* 2020, **21**:220-220.

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

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

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

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

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

699 44. Hu Y, Wilson S, Schwessinger B, Rathjen JP: **Blurred lines: integrating emerging**
700 **technologies to advance plant biosecurity.** *Current Opinion in Plant Biology* 2020,
701 **56:**127-134.

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

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

709 47. Meyer W, Irinyi L, Hoang MTV, Robert V, Garcia-Hermoso D, Desnos-Ollivier M,
710 Yurayart C, Tsang C-C, Lee C-Y, Woo PCY, et al: **Database establishment for the**
711 **secondary fungal DNA barcode translational elongation factor 1 α (TEF1 α).** 2019,
712 **62:**160-169.

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

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

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

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

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

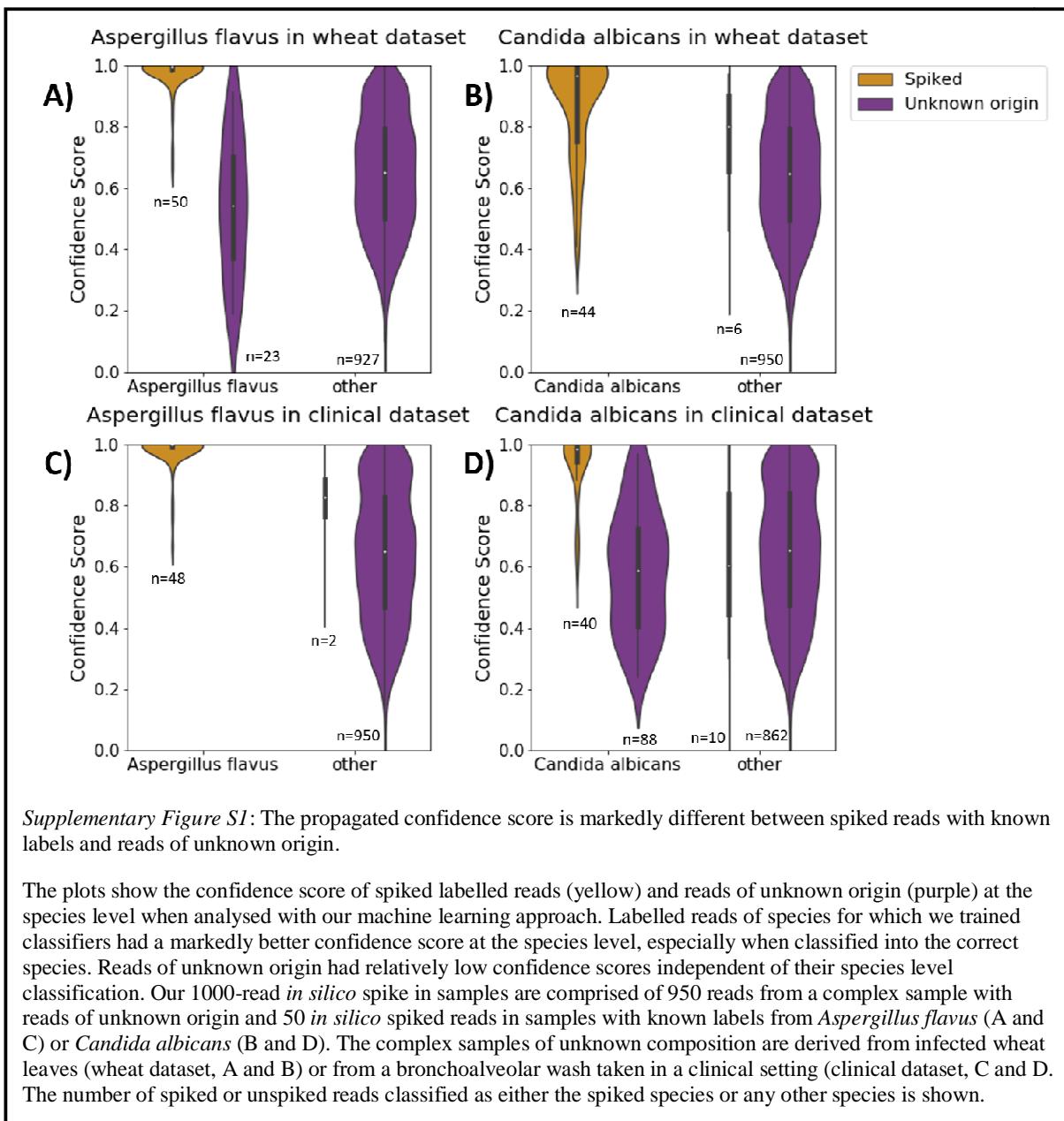
728 53. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high**
729 **throughput.** *Nucleic acids research* 2004, **32:**1792-1797.

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

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

734

735

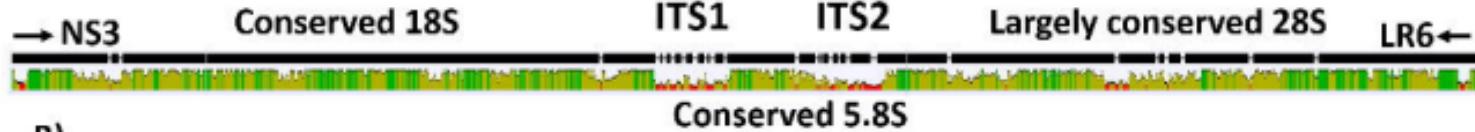


Supplementary Table T1

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

A)



B)

Kingdom

Phylum

Class

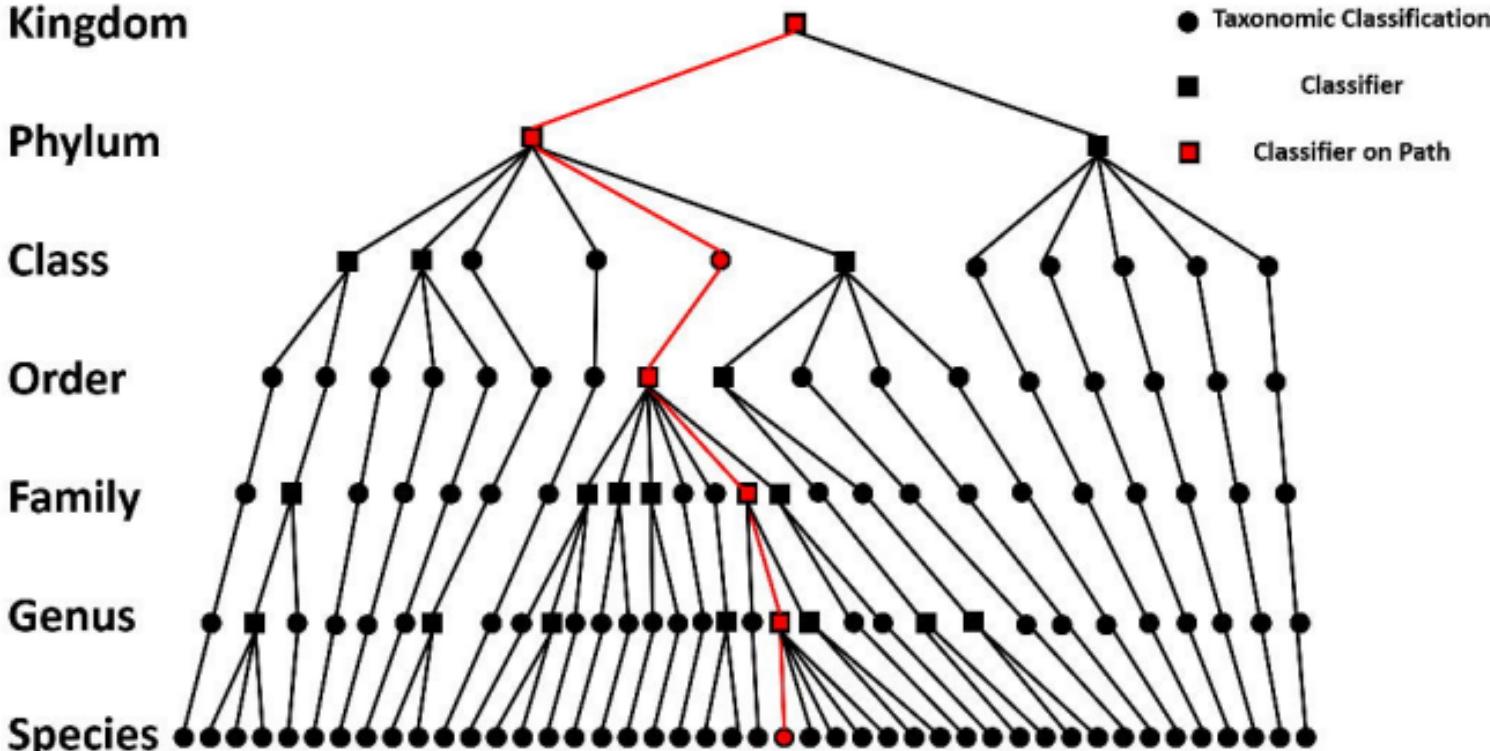
Order

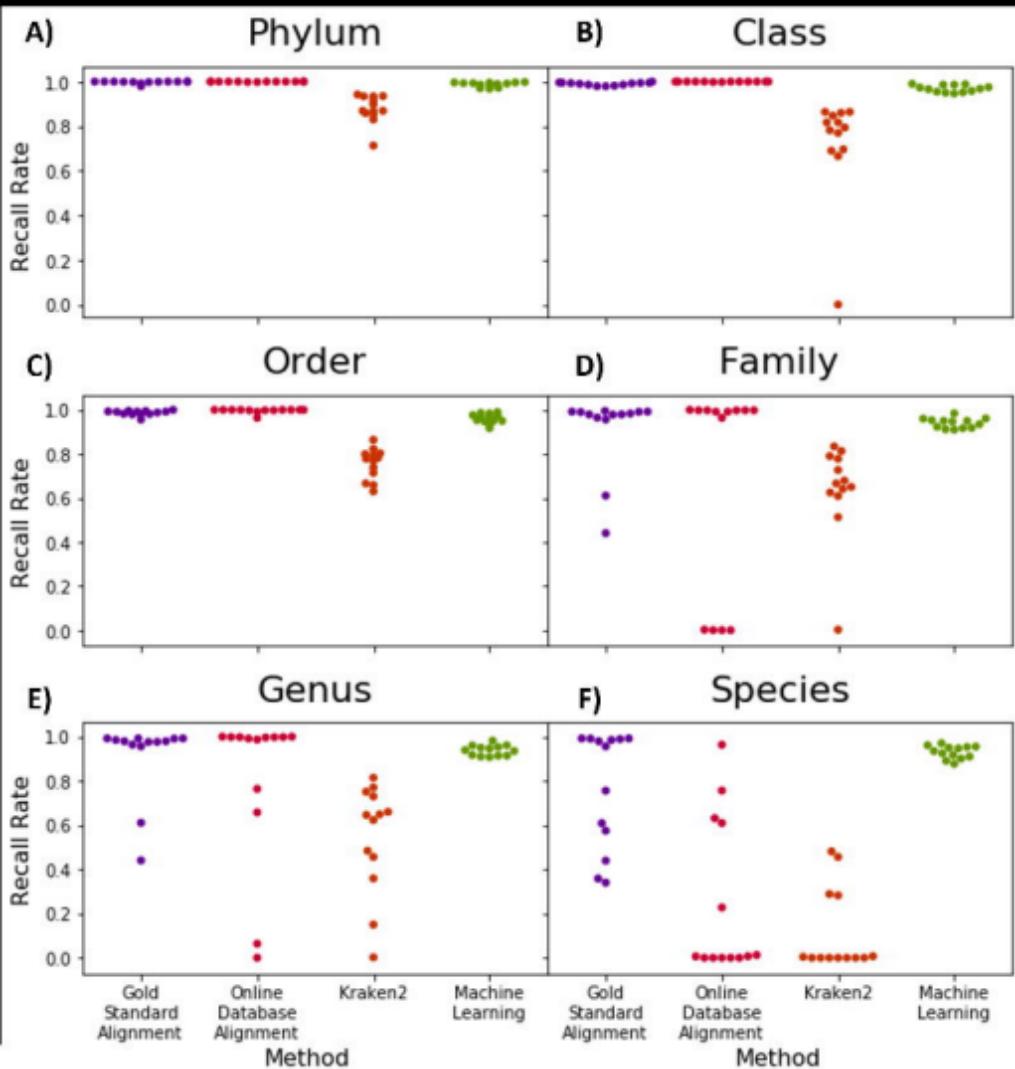
Family

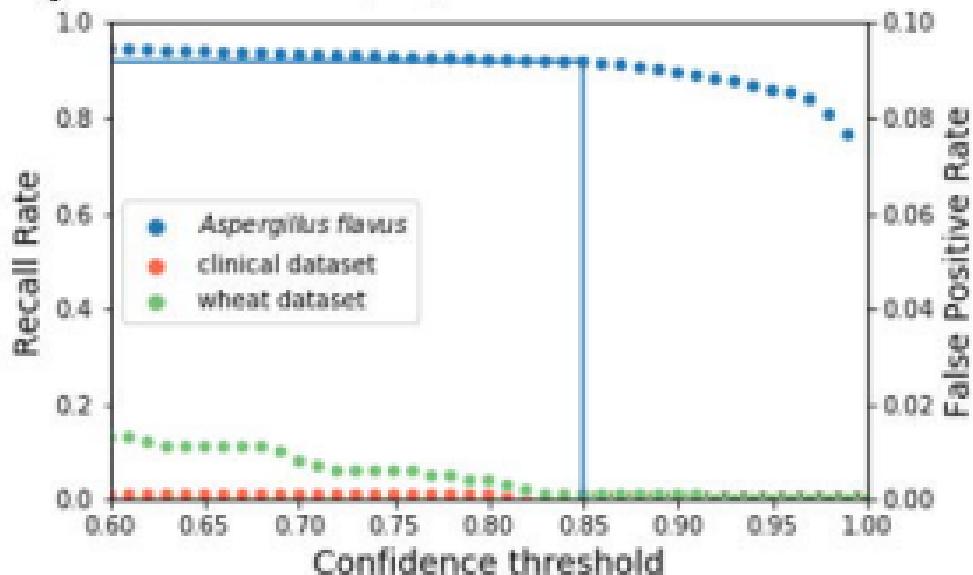
Genus

Species

- Taxonomic Classification
- Classifier
- Classifier on Path





A)*Aspergillus flavus***B)***Candida albicans*