Discovering the unseen: a performance comparison of taxonomic classification methods under unknown DNA barcodes

Johanna Orsholm[1], Alessandro Zito[2,3], Panu Somervuo[4], Mariana P Braga[1], Nicolas Chazot[1], Tomas Roslin[1,4], and Brendan Furneaux[5]

[1]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden
[2]Department of Biostatistics, Harvard University, Boston, MA, U.S.A.
[3]Department of Data Sciences, Dana-Farber Cancer institute, Boston, MA, U.S.A.
[4]Faculty of Biological and Environmental Sciences, University of Helsinki, Finland
[5]Department of Biological and Environmental Sciences, University of Jyväskylä, Jyväskylä, Finland

Abstract

1. DNA barcoding and metabarcoding have emerged as cost-efficient, standardized methods for characterizing local biodiversity. Based on the sequencing of a small targeted gene fragment, it is theoretically possible to identify a wide diversity of taxa by comparing them with reference sequence databases. However, a key challenge for accurate taxonomic classification is the incompleteness of such databases, leading to most query sequences lacking species-level matches.
2. Where species-level matches are missing, it may be possible to classify query sequences to a higher taxonomic level, such as genus or family, based on the similarity of related reference taxa. The challenge then lies in confidently recognizing whether the sequence belongs to an unobserved (here, "novel") taxon on a given taxonomic level.
3. In this study, we evaluated the performance and utility of several methods for taxonomic classification. Methods were assessed based on the classification accuracy of both observed and novel taxa, training time, space requirements, and run time. We did this for two cases: the COI barcode for arthropods, and the ITS barcode for fungi, with the latter representing an instance with substantially greater variation within classes. To test classification of novel taxa, we used well-curated datasets with partially distinct taxonomic distribution. Novel taxa were present at multiple taxonomic levels, including genera, families, and orders. We further assessed the effect on performance when shifting from full-length barcodes to shorter sequences as generated through metabarcoding in the testing dataset.
4. This study sheds light on the strengths and limitations of different classification algorithms across varied ecological contexts and provides valuable guidance for researchers in selecting suitable algorithms for DNA barcoding and metabarcoding applications. In particular, it demonstrates the supreme performance of phylogenetic placement methods such as EPA-ng for classification of COI barcodes, and composition-based classifiers such as SINTAX, RDP, and IDTAXA for ITS.

Keywords: benchmarking, DNA barcoding, species identification, taxonomy

1 Introduction

During recent decades, DNA barcoding has emerged as a versatile method to characterize biological samples (Hebert et al., 2003). DNA barcoding relies on short, standardized gene segments to classify unknown specimens within a Linnaean taxonomy using previously assembled DNA reference databases. Such a procedure is especially effective when combined with high-throughput sequencing technologies, which enable both rapid parallel sequencing of a large number of individual specimens (Shokralla et al., 2014) and analysis of mixed samples in bulk, also known as DNA metabarcoding (Riaz et al., 2011). As a result, sequencing pipelines have become indispensable tools for large-scale biodiversity analysis and monitoring, enabling the identification of species across different taxa and allowing researchers to rapidly assess species diversity in various ecosystems.

Early work confirming the utility of DNA barcodes for taxonomic classification of unknown specimens has typically focused on cases where there are close matches for the query sequences in the reference database; see Hebert et al. (2003); Schoch et al. (2012); Janzen et al. (2005); Deshpande et al. (2016). However, these ideal conditions are rare: many taxonomic branches from highly diverse groups are either undescribed or currently lack reference sequences, as evident from a striking difference between current estimates of global species richness and the actual sizes of the annotated libraries. For example, insect species diversity is estimated at 2.6-7.2 million (Stork, 2018), while well-curated reference databases account for roughly 1.2 million species-level sequence clusters with only ≈283 thousand annotated with formal species names (statistics downloaded on 2025-08-15 from `https://www.boldsystems.org/`). As for fungi, estimates oscillate between 1-12 million species, but databases present ≈170 thousand clusters with only ≈53 thousand formally annotated (Ratnasingham et al., 2024; Abarenkov et al., 2025, statistics downloaded on 2025-08-15 from `https://unite.ut.ee/`). Additionally, sequencing efforts have shown a clear geographic bias towards Europe and North America (Leandro et al., 2024; Khomich et al., 2018), leaving many of the most diverse regions of the world underrepresented.

Due to the limited coverage of diverse taxonomic groups and geographical regions, classifying specimens from taxa lacking reference sequences is a central challenge of DNA barcoding. We refer to such a category as the *novel taxa*, in contrast to *observed taxa* that are represented in the reference database. In this context, novel taxa thus include both known species lacking reference sequences and currently undescribed species. In a realistic scenario, a classification algorithm must accomplish two tasks: (1) it must predict if a query sequence belongs to an observed species and, if not, (2) it must correctly classify the higher taxonomic ranks of novel taxa up to the last observed level; for instance, the family of novel genera, and the genus of novel species. The aim of this work is to test the ability of current taxonomic classifiers to satisfy both requirements.

When designing a taxonomic classification algorithm, one must first handle the biological heterogeneity of the commonly employed barcode regions. In principle, a genetic marker is suitable for taxonomic classification if the variation between species is larger than the variation within species (Somervuo et al., 2017; Hebert et al., 2004). These properties are seen in the mitochondrial cytochrome c oxidase subunit 1 (COI; Hebert et al., 2003) in animals, and in the nuclear ribosomal internal transcribed spacer (ITS; Schoch et al., 2012) in fungi. The traditional COI barcode region, known as the Folmer region, is usually 658 base pairs (bp) long (Folmer et al., 1994), and allows for relatively simple sequence alignments thanks to the strong functional constraints on its length and amino acid sequence (Pentinsaari et al., 2016). In contrast, the ITS barcode consists of two spacers—ITS1 and ITS2—surrounding the non-coding 5.8S rRNA, each varying widely in both base arrangement and

sequence length (Schoch et al., 2012). As a consequence, it is practically impossible to produce reliable multiple sequence alignments across broad taxonomic groups (Lindahl et al., 2013). In fungi, ITS varies between 400 and 800 bp (Manter and Vivanco, 2007), but in some cases it may be up to 1600 bp in length (Feibelman et al., 1994). Hence, classifying aligned or non-aligned queries often requires distinct modeling assumptions and pre-processing steps, adding an extra layer of difficulty.

Over the past two decades, a wealth of algorithms have been developed to classify query sequences using taxonomically labeled reference sequences. Here, we test the performance of such taxonomic classification methods for DNA barcodes, specifically focusing on the issue of incomplete reference libraries. In contrast to previous studies, we use real-world geographical and methodological divisions to generate partially overlapping reference and test sequence sets. Combining historical methods with more novel approaches, we focus on stand-alone programs using a variety of algorithms either designed or commonly used for taxonomic classification of barcodes. We specifically consider two distinct biological cases: the COI barcode as a protein-coding gene with conserved length, and ITS as a non-coding sequence of variable length. In the case of COI, we also consider amino acid translations of the barcode for algorithms that accept protein sequences. Additionally, we evaluate performance at classifying a subregion of the query sequences. This comparison is motivated by practical considerations: due to the limitations on sequence length posed by some high-throughput sequencing methods, a shorter subregion of the markers is often targeted for DNA metabarcoding, resulting in around 400 bp for COI (Elbrecht et al., 2019) and 300-400 bp for ITS (Winand et al., 2025). Hence, our goal is to mimic the shorter metabarcoding reads, which are associated with lower sequencing costs. Our metrics of comparison are classification accuracy, calibration of confidence scores, prediction coverage, and use of computational resources. *A priori*, we expect (i) a trade-off between accuracy and coverage, and (ii) higher accuracy at the expense of higher computational costs for algorithms incorporating detailed models of taxonomic structure in sequence data.

1.1 Overview of taxonomic classifiers

Taxonomic classification methods aim to label sequences using reference databases of DNA barcodes, often employing markedly different strategies such as similarity searches (Altschul et al., 1990; Camacho et al., 2009; Vu et al., 2022; Lanzén et al., 2012), likelihood-based approaches (Zito et al., 2023; Somervuo et al., 2016) and phylogenetic inference (Barbera et al., 2019). However, no classification algorithm is universally superior to the others *a priori*. Rather, each method may have been designed for specific purposes, either favoring speed of prediction, unbiased estimation of prediction probabilities, or the capacity to recognize novel taxa. Recently, Hleap et al. (2021) grouped taxonomic assignment algorithms into four categories: 1) similarity-based; 2) composition-based; 3) probabilistic; and 4) phylogenetic. We argue that recent advances in machine learning and deep learning methodologies have since led to the birth of a novel fifth category, that of 5) *neural network classifiers*. From each category, we selected one or several classifiers for testing, further described in the following sections and summarized in Table 1.

Similarity-based classifiers use pairwise alignment of query sequences to the reference sequences as the basis of classification. BLAST (Altschul et al., 1990; Camacho et al., 2009) is a commonly used tool for sequence similarity search. While it is not specifically designed for taxonomic classification, a widespread first approach is to simply take the taxonomy of the top BLAST hit; see Bonato et al. (2022); Vass et al. (2022) for examples.

A slightly more sophisticated approach uses *a priori* similarity thresholds to establish whether to accept the top hit classification at each taxonomic rank (Tedersoo et al., 2021). Here, a similarity threshold of 97% has been widely used for species-level identification across different taxa and barcode markers (Tedersoo et al., 2022; Porter and Hajibabaei, 2020). These thresholds can be further optimized for different taxa, as in dnabarcoder (Vu et al., 2022), or replaced by instead considering multiple hits simultaneously, assigning the last common ancestor (LCA) of the k nearest neighbors (KNN) from a BLAST search, as in CREST (Lanzén et al., 2012). Thresholding and KNN-consensus both limit over-classification, which occurs when sequences from novel taxa are assigned taxonomic labels at ranks where no true match exists. The KNN-consensus method can also help avoid over-classification in cases where closely related species have identical barcode sequences. However, neither method provides a numerical confidence score or distinguishes between uncertain classifications and novel taxa.

Composition-based classifiers avoid the need for computationally expensive pairwise alignments by decomposing both queries and reference sequences into k-mers, i.e., distinct overlapping subsequences of length k. A sequence can then be described by its k-mer profile – a fixed-length vector with one entry for each possible k-mer. The length of the vector is 4^k for nucleotide sequences and 20^k for amino acid sequences. From this category, we selected the RDP naïve Bayesian classifier (Wang et al., 2007), SINTAX (Edgar, 2016), and IDTAXA (Murali et al., 2018). All three models use bootstrap samples of the k-mer composition of the query sequences to produce numerical confidence scores for the classification at each taxonomic rank. This helps control over-classification in a more granular way than the similarity search methods, but does not directly address novel taxa.

Probabilistic classifiers use a statistical framework to estimate the probability that a query sequence belongs to a given taxon at each taxonomic rank. From this category, we selected BayesANT (Zito et al., 2023) and PROTAX (Somervuo et al., 2016). PROTAX can incorporate any covariates as predictors of taxonomy – sequence similarity being just one example – allowing users to tailor the model to better fit their specific needs. There are also existing implementations optimized for common taxonomic groups, such as animals (Roslin et al., 2022) and fungi (Abarenkov et al., 2018), which can be readily applied without further model customization. Both BayesANT and PROTAX explicitly distinguish between classification uncertainty and novel taxa. Both have variants that can operate either on unaligned or globally aligned sequences, and while the latter may be both faster and higher-performing, it is limited to sequences with low length polymorphism, such as the COI barcode for animals. Moreover, BayesANT and PROTAX provide confidence scores designed to be interpreted as well-calibrated probabilities. In addition, PROTAX can leverage information from a reference taxonomy that includes taxa lacking associated reference sequences.

Phylogenetic placement classifiers place query sequences individually into an existing phylogenetic tree generated from a set of reference sequences. In this framework, novel taxa are cleanly represented as placement on branches outside the clades which represent observed taxa at the rank under consideration. Construction of the reference tree is left to the user, with a wide range of algorithms available but differing in speed and accuracy. There are several opportunities to incorporate additional information beyond the reference sequences at the stage of building the reference tree, including constraining the topology to conform to the taxonomic hierarchy or to published phylogenies. Alternatively, the reference tree can be generated using longer sequences or additional genes beyond the region covered by the query sequences. From this category we selected EPA-ng (Barbera et al., 2019), a maximum-likelihood based phylogenetic placement algorithm that calculates the

likelihood of placement on each branch of the tree for each query sequence, according to a chosen evolutionary model. From these placements, Gappa (Czech et al., 2020), a helper tool for phylogenetic placement, can perform the mapping to taxonomic classification. Gappa can also report classification probabilities when they are provided by the phylogenetic placement algorithm. However, phylogenetic placement typically requires that the query and reference sequences are globally aligned, making it unsuitable for ITS fungal sequences.

Neural network classifiers are united by their use of a learned vector representation for barcode sequences. From this category, we selected MycoAI (Romeijn et al., 2024), which implements both a convolutional neural network (CNN) operating on sequences represented as real-valued k-mer profiles, and a bidirectional transformer (BERT) operating on byte-pair encoded sequences. Both architectures use hierarchical label smoothing during training and a multi-head output for the different taxonomic ranks in order to borrow information between taxonomic ranks and provide separate predictions and confidence values at each rank.

Previous comparisons of taxonomic classifiers have found that simple top-hit algorithms, such as BLAST, often achieve the highest classification accuracy of observed species (Edgar, 2018; Hleap et al., 2021), making them attractive choices for study systems with extensive reference libraries. However, classification accuracy typically declines with a decrease in the similarity between the query sequence and the top hit reference sequence (Edgar, 2018; Richardson et al., 2017), making it challenging to reliably classify novel taxa. As previously described, taxonomic classifiers should ideally make predictions down to the lowest observed rank of a query sequence. Thus, setting conservative confidence thresholds for predictions can help reduce misclassification and over-classification – but may simultaneously lead to high rates of under-classification, where too few taxonomic ranks are predicted. Classification of novel taxa is further complicated by the lack of a consistent relationship between sequence similarity and the lowest common rank across taxonomic branches (Edgar, 2018). The complexity of this task calls for well-calibrated estimations of prediction uncertainty (Somervuo et al., 2017), allowing researchers to make informed decisions about retaining classifications and potentially incorporating uncertainty in downstream analyses.

2 Materials and methods

2.1 Data description and pre-processing

In our evaluation, models were trained on a set of barcodes, called the *training set*, which had a partial taxonomic overlap with the sequences used to assess model performance, the *test set*. All sequences in both sets were associated with a species-level identification. While the taxonomic annotations could contain some errors, for the purposes of this evaluation we treat them as the ground truth. To reflect realistic differences between training and test sets, we used sequences from different countries for COI and sequences derived from different methodological approaches for ITS.

COI sequences: FinBOL and GBOL

For COI, we used DNA sequences from the Finnish Barcode of Life (FinBOL; Roslin et al., 2022) as the training set and from the German Barcode of Life (GBOL; German Barcode of Life Consortium, 2011) as the test set. FinBOL and GBOL aim to create comprehensive DNA barcode reference libraries for multicellular

organisms in Finland and Germany, respectively, using targeted sequencing of expert-identified specimens. Because they focus on separate regional faunas and, to some degree, reflect the research priorities of their respective taxonomic communities, the sets have only a partial taxonomic overlap.

For FinBOL, sequences and taxonomic annotations (including ranks class, order, family, subfamily, tribe, genus, and species) were retrieved from the BOLD (Ratnasingham and Hebert, 2007; Ratnasingham et al., 2024) data package 29-Mar-2024 (BOLD Systems, 2024) using the project code 'DS-FINPRO'. For GBOL, sequences were retrieved from the GBOL web portal (German Barcode of Life Consortium, 2011) using the filters "Sub-/Phylum: Arthropoda", "Collected: 11", and "With Barcode: 11" to retrieve COI barcodes from a maximum of eleven specimens per species within Arthropoda. To avoid discrepancies between the taxonomy of the training and test sets, the Process-ID corresponding to records in BOLD was used to retrieve taxonomic information for each sequence in the test set.

Secondary taxonomic ranks, here subfamily and tribe, are not assigned to all taxa in the arthropod taxonomy. To represent taxa without secondary ranks, we used *Incertae sedis* as placeholder taxa. In our evaluation of COI classifications, we treated these placeholder taxa as equivalent: for example, the genera *Acanthosoma* and *Elasmostethus* were classified in the subfamily *Acanthosomatinae*, but did not have a tribe classification. At the tribe rank, they were therefore both represented as *Acanthosomatinae_tribe_incertae_sedis*.

For both training and test sets, we retained only records which were identified to the species level, which did not belong to a "placeholder" taxon at the species level, which had been annotated as being a sequence of the COI-5P barcode region, and which were at least 600 bp in length. Test and training sequences were aligned to the Folmer barcode region (Folmer et al., 1994) and amino acid translations were generated using MACSE v2.07 (Ranwez et al., 2018) and the COI coding sequence from *Drosophila melanogaster* (NCBI accession NC_024511.2, positions 1474..3009) as a reference. Sequences with less than 600 bases in the nucleotide alignment after removal of gap columns were removed from both nucleotide and amino acid alignments, resulting in 35 603 training and 26 925 test sequences. The final COI test set included novel taxa at all ranks from class to species (Fig. 1).

We also created a "testshort" data set that mimics the shorter reads obtained from DNA metabarcoding, by replacing the first 240 positions of the nucleotide test alignment and the first 80 positions of the amino acid test alignment (i.e., up to and including the BF3 primer binding site; Elbrecht et al., 2019) with gaps ("-"). Trimmed but unaligned training, test, and testshort sequences were generated by removing all gap characters from the aligned sequences.

ITS sequences: Westerdijk culture collection and Unite reference sequences

For ITS, we used DNA sequences from the Westerdijk Fungal Biodiversity Institute culture collection of yeasts and fungi (Vu et al., 2019, 2016) as the training set. As the test set, we used curated "reference sequences", i.e., sequences that have been selected manually to represent species hypotheses, from the UNITE database (Abarenkov et al., 2023). The Westerdijk Institute maintains the largest culture collection of living fungi in the world, while the UNITE database is a community-curated database of eukaryotic ITS sequences generated from diverse substrates and methods. This difference in scope and methodological breadth results in a partial taxonomic overlap.

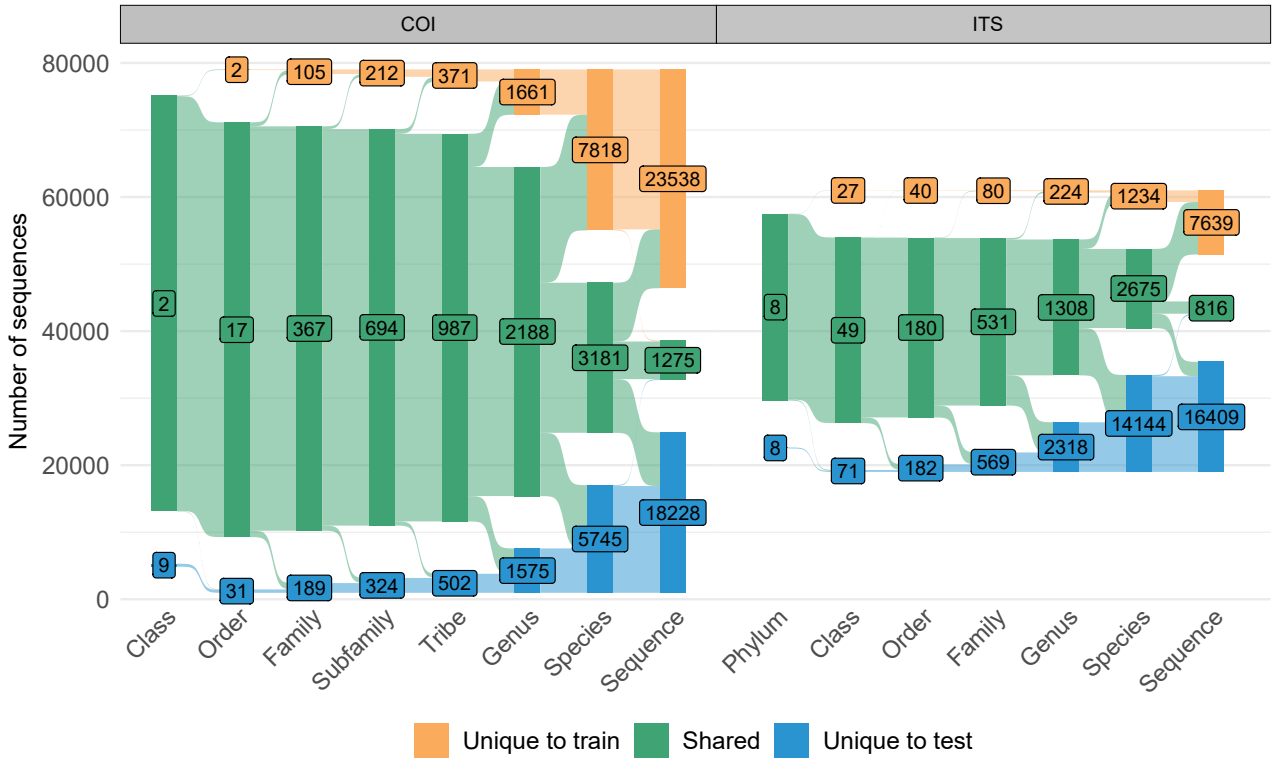

Figure 1: Taxonomic overlap between training set and test set for COI and ITS sequences. The x-axis displays the taxonomic ranks, with an additional column for unique DNA sequence variants. The height of columns and ribbons correlates to the number of sequences belonging to either shared taxa (green), taxa unique to the training set (yellow) or taxa unique to the test set (blue). Numbers show the count of unique taxa (or, for the last column, sequences) at each rank and group.

For the training set, sequences were retrieved from NCBI using PopSet 1102638064 and PopSet 1473241625 and matched with taxonomic annotations (including ranks kingdom, phylum, class, order, family, genus, and species) from the UNITE database (Abarenkov et al., 2023). For the test set, we downloaded the UNITE 10.0 general FASTA release from https://doi.org/10.15156/BIO/2959332 and selected sequences annotated as reference sequences ("RefS").

Many sequences from the Westerdijk culture collection also occur as reference sequences in UNITE. To remove this overlap between the training and test set, we removed duplicate sequences from the training set if multiple sequences were available for that species hypothesis. If the duplicate sequence was the single representative of a species hypothesis, we instead removed it from the test set.

For both training and test sets, we extracted the ITS region with ITSx (Bengtsson-Palme et al., 2013), which uses profile hidden Markov models (HMMs) to detect the flanking SSU and LSU regions. When the included fragments of SSU or LSU are too short, ITSx sometimes fails to detect them. For these cases, we used cutadapt (Martin, 2011) to trim sequences matching the end of SSU, and LSUx (Furneaux et al., 2021) to detect and remove the beginning of LSU.

Some fungal species do not have all primary taxonomic ranks in their classification, because their relationship to other fungi is unclear, and the Unite database uses *Incertae sedis* names as placeholder taxa to represent these. These taxa include the name of the closest named enclosing taxon, the rank they represent, and the words *Incertae sedis*. For instance, the genus *Ceratocladium* is classified in phylum *Ascomycota*, but not in any of the many described classes, orders, or families within *Ascomycota*, so its family is listed as "Ascomycota_fam_Incertae_sedis". Multiple taxa with the same placeholder cannot be assumed equivalent; for instance, genus *Septogloeum* is also listed at the family level as "Ascomycota_fam_Incertae_sedis", but this does not indicate that the genera *Ceratocladium* and *Septogloeum* belong to the same natural group at the family level. To disambiguate such cases, we appended the nearest named lower rank to each *Incertae sedis* label, in the example case creating the family-level placeholders "Ascomycota_fam_Incertae_sedis_Ceratocladium" and "Ascomycota_fam_Incertae_sedis_Septogloeum".

Finally, we created testshort sequences including only ITS2 by deleting all nucleotide bases up to the end of 5.8S, as detected by the Rfam covariance model in LSUx.

2.2 Classification algorithms

We trained each classification algorithm on the training set and then classified each sequence in the test set, using both the full length barcodes, the testshort data, and, for COI, the amino acid sequence. For some algorithms, there are many adjustable parameters for training and/or classification, but we used default parameter values when possible, or parameter values which we believe reflect standard usage. In particular, we did not tune parameters to optimize results on our particular data set, nor did we modify the algorithms from their most recent publicly available versions for the purpose of improving performance. For each sequence and classification algorithm, in cases where algorithms produced more than a single prediction, we selected the prediction with the highest probability or confidence score. We determined whether the prediction was correct by comparing it with the taxonomic annotation in the original test data set. For sequences belonging to novel taxa, predictions were considered correct if the taxon was predicted novel at the correct rank and all higher ranks were predicted correctly.

Some algorithms do not predict novelty but can refrain from making predictions for sequences with high classification uncertainty. Since a high proportion of missing predictions can inflate classification accuracy, we distinguish between results that allow missing predictions (MP) and those where predictions are enforced for all query sequences (FP). In the MP setting, we applied confidence thresholds for classification algorithms where recommended in the original publication. In the FP setting, we interpreted missing predictions as novel taxa.

Algorithms were executed on a compute node of a Linux high-performance computing cluster equipped with two Intel Xeon Gold 6230 CPUs for a total of 2x20 processing cores at 2.1 GHz. For each algorithm, the full training and classification task was run with allocations of 1, 4, 16, and 40 processor cores to test parallel scalability. In cases where the algorithms accept parameters controlling the number of threads, we provided them with the actual number of allocated cores. Algorithms which support GPU execution were additionally run in allocations with 1 CPU core and 1 Nvidia V100 GPU. We adjusted the memory and time allocations as needed for the algorithms to complete, with a maximum limit of 192 GiB of memory and 14 days of processing time. Elapsed time, maximum memory usage, and average processor usage of the training and classification steps were measured using the Gnu time utility. Disk space required by the trained model was also recorded. For algorithms implemented as R packages or Python packages without command line interface, we wrote minimal scripts in the respective languages, separately for training and prediction, and measured the CPU and memory usage of the entire R/Python process executing the script. Open-source algorithms implemented in C or C++ were locally compiled using GCC X.X.X with additional options `-O3 -march=native -mtune=native` to optimize the executable for the local CPU.

2.2.1 Sequence similarity methods

We tested two separate approaches using BLAST as a taxonomic classifier: (1) top hit classification, where we used the top hit without any cutoff, i.e. we accepted the taxonomy of the top hit as a prediction regardless of the sequence similarity, and (2) threshold classification, where we used predetermined similarity thresholds to accept or decline classifications at each taxonomic rank. To reflect common practice, we used for COI: class: 90%, order: 93%, family: 95%, subfamily: 96%, tribe: 96.5%, genus: 97%, and species: 98% (**?**), and for ITS: kingdom: 65%, phylum: 70%, class: 75%, order: 80%, family: 85%, genus: 90%, and species: 98% (Tedersoo et al., 2021). Both BLAST top hit and BLAST threshold can generate missing predictions when sequence similarity is low, but they are more prevalent for BLAST threshold.

2.2.2 K-mer composition methods

For the RDP naïve Bayesian classifier we considered version 2.14 (2023), downloaded from https://sourceforge.net/projects/rdp-classifier/files/rdp-classifier/ and run using OpenJDK version 21. We used the `train` command to generate the trained model parameters, and added a renamed version of `src/data/classifier/16srrna/rRNAClassifier.properties` from the downloaded source code. We then used the `classify` command with confidence threshold 0 (`-conf 0`) to output all classifications regardless of confidence limit for the FP setting. For the MP setting we removed classifications below the suggested confidence threshold of 0.8. For SINTAX, we used the implementation in VSEARCH version 2.28.1 (Rognes et al., 2016), which made several minor changes to the algorithm to prevent incorrectly high confidence in the presence

of ties. The model was initialized by creating an index using the `vsearch -makeudb-usearch` command, and predictions were made with `vsearch -sintax`, with the `-sintax-random` option. As for RDP, we removed classifications with confidence lower than 0.8 in the MP setting. For IDTAXA, we installed DECIPHER 2.26.0 using renv 1.0.4 in R 4.2.3. Models were trained using the `LearnTaxa()` command with default parameters. Predictions were made using the `IdTaxa()` command with arguments `type=''collapsed''` for condensed output, `strand=''top''` to assume the input data are properly oriented, and `threshold=1` (meaning 1%) to output even low-confidence predictions. Because IDTAXA operates hierarchically, avoiding classification at lower ranks in low-confidence cases, it is likely that the low threshold resulted in somewhat slower performance than typical usage. For the MP setting we removed classifications below the suggested confidence threshold of 0.6 for IDTAXA.

2.2.3 Probabilistic models with novelty

For taxonomic classifications with PROTAX, we used separate implementations for COI and ITS barcodes. For COI, we used the model formulation and codebase for aligned sequences (ProtaxA), as in FinPROTAX (Roslin et al., 2022). Source code and scripts were downloaded from https://github.com/psomervuo/protaxA (commit dff9be3, dated 24 Aug, 2023). Input taxonomies were generated using a custom `GAWK` script, based on the training labels for the base case, or on the full Arthropod taxonomy from BOLD snapshot 29-Mar-2024 (Ratnasingham et al., 2024; BOLD Systems, 2024) for the augmented case. We weighted the "unknown" taxa such that, at each taxonomic rank, the total weight for "unknown" taxa was equal to 0.05 of the total weight (Abarenkov et al., 2018). Models were trained using the training scripts provided in the "readme.txt" file in ProtaxA, with modifications for the number of taxonomic levels. Classification was performed using the compiled "classify_v2" executable, implementing fast SIMD distance calculations. For ITS, we used the version available at https://github.com/psomervuo/PROTAX, which uses USEARCH (Edgar, 2010) to calculate pairwise distances. Models were trained using modified versions of the scripts in "step1.txt" and "step2.txt" scripts, and predictions were made using the script in "step4.txt". Interpreted scripts for both versions of Protax were run using Perl 5.26.3 and R 4.2.3. MCMC traces were evaluated and considered sufficiently converged after the default sampling, so no additional sampling was performed.

For BayesANT, we used the implementation available at https://github.com/alessandrozito/BayesANT. For the aligned COI sequences, we trained the model using the default multinomial kernel (option `typeseq = "aligned", type_location = "single"`) paired with the maximum likelihood estimation of the Pitman-Yor parameters controlling novelty at each taxonomic level. For ITS, we considered a 5-mer decomposition of the fungal sequences using the option `typeseq = "not_aligned"`.

2.2.4 Phylogenetic methods

Phylogenetic placement algorithms typically require that the query and reference sequences are globally aligned, making it unsuitable for use with "unalignable" barcode regions such as ITS. Therefore, we only tested this class of models for COI. We classified the test data using EPA-ng with two different reference trees; one built with full taxonomic constraints (hereafter the *taxonomically constrained tree*), and one built with

constraints based on previously established phylogenies resolving relationships between families (hereafter the *phylogenetically constrained tree*).

For the taxonomically constrained tree, a multifurcating taxonomic tree was constructed from the taxonomic annotations of the training sequences using the command `prepare taxonomy-tree` in Gappa version 0.8.5. This tree was then used as a constraint for maximum likelihood tree search on the aligned training set using IQTREE version 2.4.0 (Minh et al., 2020) with the GTR+G4 (generalized time reversible, with 4-category approximation to gamma-distributed rates) substitution model for the nucleotide alignment, and the mART+G4 (Arthropod mitochondiral substitution matrix, with 4-category approximation to gamma-distributed rates) for the amino acid alignment, using mode `-fast` in both cases for fast inference.

For the phylogenetically constrained tree, we retrieved phylogenies resolving family relationships within the 20 arthropod orders included in the training data, along with phylogenies spanning Chelicerata and Insecta, to use as backbone trees. All trees were rooted using outgroups from the original publications. Order and family classification of all genera in each tree were retrieved from BOLD (Ratnasingham et al., 2024; BOLD Systems, 2024) when present, otherwise from NCBI taxonomy (Schoch et al., 2020). These trees were then grafted to form a single combined tree, starting by grafting the Insecta tree at the LCA of insects in the Chelicerata tree, and then continuing by grafting each order at the LCA of that order within the combined tree. Outgroup taxa were removed from each order-level tree before it was grafted into the combined tree. The combined tree was then pruned to only contain families present in FinBOL, and to only contain one branch per family. In cases where a family was not monophyletic in the reference tree (27 of 380 families), the branch for that family was placed as child of the LCA of that family in the original tree . To generate a constraint tree for IQ-TREE, the training sequences belonging to each family were added as a polytomy on the branch for that family. 93 families in the training set were not represented in the backbone tree, so sequences belonging to these families were excluded from the initial alignment. From the subset alignment, a first reference tree was inferred with IQ-TREE X.X.X using mode `-fast`, substitution model GTR+G4, and the backbone tree to constrain the topology. To include the 93 families not represented in the initial backbone tree, the sequences belonging to these families were phylogenetically placed in the initial reference tree, using EPA-ng version 0.3.8. To find the most likely placement for each family, we calculated the product of likelihoods across all branches and placed sequences belonging to that family, identifying the maximum value to determine the optimal position. For three families (Baetidae, Boreidae, and Diprionidae), the most likely placement was within a different order than the family taxonomically belongs to, or forming a sister clade to a different order. These families were placed as children to the respective LCA of the correct order, i.e., as sister clade to all other sequences in that order. To create the final constraint tree, the 93 families were added at the position of their most likely placement in the first reference tree, and each family, including those in the initial reference tree, was again collapsed to a polytomy. To create the final reference tree, we ran IQ-TREE X.X.X with the full training set alignment, GTR+G substitution model, and the combined backbone tree as a topological constraint using mode `-fast`.

For both reference trees, we then phylogenetically placed all query sequences using EPA-ng version 0.3.8 and assigned taxonomic labels with Gappa version 0.8.5 (`examine assign`).

Table 1: Features of selected taxonomic classification algorithms. Amino: All selected algorithms operate on nucleotide sequences; some can additionally be applied to amino acid sequences. Stopping: Does the algorithm have any mechanism to avoid over-classifying sequences? Novelty: Does the algorithm distinguish between low classification certainty for existing categories and prediction of new categories which were not present in the reference? Confidence: Does the algorithm produce a numerical score on a fixed scale (i.e., 0-1 or 0-100) for the confidence of its classifications at each rank? Outside info: Does the algorithm allow additional data beyond the sequences and taxonomic labels of the training data? If so, what? Implementation language: What language(s) is/are used for the implementation? Does it depend on any other software? Interface: How do you use it?

Algorithm	Data type	Amino	Stopping	Novelty	Confidence	Outside info	Implementation language	Interface
BLAST top-hit	similarity	✓	✘	✘	✘		C++	cmd, web interface
BLAST threshold	similarity	✓	✓	✘	✘	Threshholds	C++, awk script	cmd
Crest4 LCA	similarity	✓	✓	✘	✘		Python + BLAST	Python pkg., cmd
dnabarcoder	similarity	✘	✓	✘	✘		Python + BLAST	cmd
RDP	k-mer profile	✘	✓	✘	✓		Java	cmd
Sintax	k-mer profile	✘	✓	✘	✓		C++	cmd
IDTAXA	k-mer profile	✓	✓	✘	✓		C, R	R pkg.
ProtaxA	similarity from alignment	✘	✓	✓	✓	Full taxonomy	C, Perl, R (training)	cmd, R script
PROTAX	similarity from alignment	✘	✓	✓	✓	Full taxonomy	Perl + USEARCH, R (training)	cmd, R script
EPA-ng	alignment	✓	✓	✓	✓	Backbone tree	C++ (EPA-ng; Gappa; IQ-TREE2)	cmd
BayesANT	alignment or k-mer profile	✓	✓	✓	✓		R	R pkg.
MycoAI-CNN	k-mer profile	✘	✓	✘	✓		Python (PyTorch)	Python mod., cmd
MycoAI-BERT	BPE	✘	✓	✘	✓		Python (PyTorch)	Python mod., cmd

2.2.5 Neural network methods

We installed MycoAI v0.0.5 from pypi (package name mycoai-its), which is the first version with the option to output confidence values along with classifications. We installed the package in a Python 3.9.7 virtual environment using pip 20.2.4. We used PyTorch 1.12, the minimum version compatible with MycoAI. Models were trained using `mycoai-train` from the command line. For the CNN model, the option `-base_arch_type CNN` was given. Test data were classified using `mycoai-classify` with option `-confidence` to output confidence scores.

2.3 Performance metrics

To explore the differences in algorithm performances on query sequences belonging to either observed or novel taxa, we split sequences into four different taxon sets. First, *observed species* included all test sequences that belonged to species represented in the training data, and *novel species* included all test sequences from species absent from the training data. Since species was the lowest taxonomic rank we considered, all higher taxonomic ranks of sequences within the *observed species* set were necessarily observed as well, while *novel species* included sequences from taxa that were novel at higher ranks, such as novel orders or families. Second, to explore separately how well each taxonomic rank was predicted, we split sequences into *observed taxa* and *novel taxa*. *Observed taxa* included, for each rank, all test sequences belonging to an observed taxon on that specific rank. I.e., at order level, it included all test sequences belonging to orders represented in the training data, regardless of whether that sequence belonged to an observed taxon at lower taxonomic ranks or not. Similarly, *novel taxa* included all test sequences at each rank that belonged to a taxon which was novel on that specific rank. For example, each sequence in *novel taxa* at family level belonged to observed orders but novel families. Consequently, in both observed taxa and novel taxa, the number of sequences was unique at each rank of the set.

To assess algorithm performance, we calculated classification accuracy so that correct novelty prediction were counted as true positives. Thus, accuracy was here defined as

$$\text{Acc} = \frac{TP}{N}$$

where TP was the number of correct predictions, including novel sequences correctly predicted as novel, and N was the number of sequences in the test set. We further calculated *over-classification rate*, where taxonomic names are assigned to novel taxa, *under-classification rate*, where observed taxa are predicted as novel, and *misclassification rate*, where incorrect taxonomic names are assigned to observed taxa, using the definitions by Edgar (2018). We evaluated error rates under the MP setting, thus implementing confidence thresholds for algorithms which recommend them. For observed and novel taxa, we further calculated marginal and conditional recall. Recall is a performance metric used to measure of a model's ability to identify positive instances out of all true instances in the dataset (also known as sensitivity). Here, we defined *marginal recall* for each taxonomic rank as

$$\text{Recall}_{t,r} = \frac{TP_{t,r}}{N_{t,r}}, \qquad t \in \{\text{new}, \text{obs}\},$$

where $TP_{\text{new},r}$, $TP_{\text{obs},r}$ were the number of correct taxonomic predictions of sequences belonging to the groups of novel and observed taxa, respectively, at rank r, and $N_{\text{new},r}$, $N_{\text{obs},r}$ were the number of sequences in the test set belonging to the groups of novel and observed taxa at rank r, respectively. Note that $TP_{\text{new},r}+TP_{\text{obs},r} = TP$ and $N_{\text{new},r} + N_{\text{obs},r} = N$ at any rank. For conditional recall, we considered only cases where the predictions for the higher ranks were correct. That is,

$$\text{Conditional_Recall}_{t,r} = \frac{TP_{t,r}}{TP_{t,r-1}} \qquad t \in \{\text{new}, \text{obs}\},$$

where $TP_{t,r-1}$ is the number of sequences correctly predicted at rank $r-1$ for group t.

Under the MP setting, we calculated prediction coverage, i.e., the proportion of sequences for which an algorithm assigned a taxonomic label, across all possible confidence thresholds. For algorithms which did not produce confidence estimates, prediction coverage was calculated as a single value.

Finally, we evaluated how well-calibrated confidence estimates were, by plotting the cumulative probability of the predictions against the cumulative proportion of cases where the prediction was correct. For well calibrated algorithms, probability and proportion of correct predictions are equivalent, so that for cases where the prediction probability is 90%, the predictions are correct in 90% of the cases. In the plot, this would display as the correlation line following the identity line. For algorithms missing confidence estimates, we assigned 100% confidence to all predictions.

3 Results

3.1 Classification accuracy and recall

When classifying observed species under the FP setting, most algorithms exceeded 93% and 78% accuracy at the genus and species level, respectively, for both COI and ITS sequences (Fig. 2). For novel species, all

algorithms exhibited a lower classification accuracy than for observed species, ranging between 9-58% accuracy at the genus level for COI, and 13-40% for ITS. Models which explicitly distinguished between classification uncertainty and novel taxon predictions, i.e., EPA-ng (COI only), BayesANT, and PROTAX, classified novel species with accuracies between 12-44% for COI, and 0-30% for ITS. For COI, EPA-ng consistently achieved the highest classification accuracy across ranks, a result that held for both the taxonomically and the phylogenetically constrained reference tree. For ITS, composition-based classifiers exhibited the highest prediction accuracy for novel species, with nearly identical accuracies for IDTAXA, RDP, and SINTAX (Fig. 2).

For the more difficult task of classifying observed taxa, where some may belong to a novel taxon at a lower rank, all algorithms showed lower marginal recall than accuracy for classifying observed species at the corresponding rank (Fig. 2). This suggests that higher-level classifications were more error-prone when the sequence belonged to a novel taxon at a lower rank, compared to when a species-level match was available. Compared to observed species, the difference between algorithms was substantially larger for classification of observed taxa. In general, similarity-based classifiers exhibited the lowest marginal recall for observed taxa for both COI and ITS, with the exception of BLAST top hit, which was consistently among the top performing algorithms across all ranks for COI, and at ranks below order for ITS. For novel taxa, only EPA-ng, BayesANT, and PROTAX explicitly predicted taxonomic novelty. Some similarity-based algorithms can abstain from making predictions, which we interpreted as novelty in the FP setting. Consequently, as expected, all composition-based and neural network classifiers exhibited zero marginal recall for novel taxa. Among the algorithms which explicitly predicted novelty, EPA-ng and PROTAX exhibited the highest marginal recall for COI and ITS, respectively (Fig. 2). BayesANT achieved markedly better performance for COI sequences than ITS sequences, where it failed to recover any novel taxa. For COI, BayesANT exhibited the highest conditional recall across ranks from family to species, suggesting it successfully predicted novelty in most cases where the higher ranks where predicted correctly (Fig. **??**). For COI, EPA-ng, BayesANT, and PROTAX all over-estimated the number of novel taxa at family level and below. For ITS, PROTAX over-estimated and BayesANT under-estimates the number of novel across all ranks.

When evaluated on the testshort dataset, designed to mimic shorter metabarcoding reads, most algorithms classified observed species with accuracies similar to full-length reads, with the exceptions of MycoAI for both markers and DNABarcoder and RDP for ITS. Notably, DNABarcoder assigned almost all ITS testshort sequences from observed species to a novel class, yielding accuracy close to zero. For novel species, classification accuracy declined on short reads compared to full-length reads, particularly for ITS at higher taxonomic ranks.

Translating COI to amino acid sequences yielded lower classification accuracies of observed species. However, BayesANT and Crest4 exhibited higher accuracies compared to nucleotide sequences for novel species, especially at higher ranks.

3.2 Error rates

Most models showed very low misclassification rates (<5% for COI and <13% for ITS; Fig. 3). Higher error rates were observed for MycoAI and BLAST top hit across both markers, and for BayesANT on ITS. In addition, at the ITS species level, RDP and SINTAX showed elevated error rates, misclassifying 20% and 14% of observed species, respectively. Overall, over-classification and under-classification showed opposing patterns, highlighting the trade-off between the two. Even under the MP setting, there is no recommended confidence

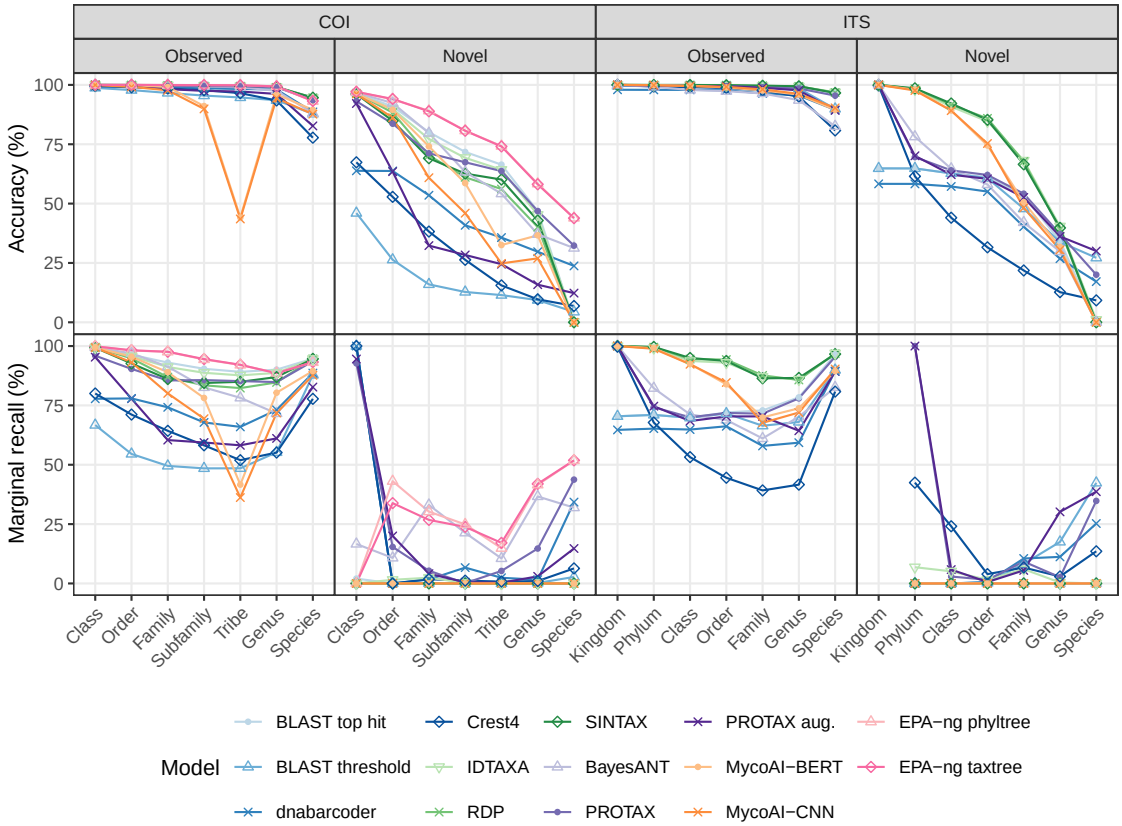

Figure 2: Accuracy and marginal recall of classification algorithms across taxonomic ranks under the FP setting, i.e. where predictions were enforced for all query sequences. The top row shows accuracy, i.e. the proportion of all taxonomic classifications that were correct for observed (right) and novel (left) species. The bottom row shows marginal recall, i.e. the proportion of correct predictions relative to the total number of sequences belonging to that respective taxon set and rank for observed (right) and novel (left) taxa. Each algorithm is shown with a different shape and color, with hue indicating algorithm group: sequence similarity (blue), composition (green), probabilistic (purple), phylogenetic (pink), and neural network (orange). Error rates are shown under the MP setting, i.e. allowing missing predictions and implementing confidence thresholds for algorithms which recommend them.

threshold or stopping rule for MycoAI or BLAST top hit, thus resulting in 100% over-classification for MycoAI for both markers, and nearly 100% over-classification for BLAST top hit and COI. For ITS, BLAST top hit had a higher proportion of missing predictions, resulting in lower over-classification rates. Among the algorithms that included a stopping rule, EPA-ng had the highest over-classification rate for COI. In contrast, BayesANT achieved both a low over-classification rate and a relatively low under-classification rate for COI. For ITS however, BayesANT over-classified most novel taxa. RDP and SINTAX had overall low over-classification rates for COI, except at class level, where they over-classified 73% and 98% of novel taxa, respectively.

3.3 Calibration of prediction probabilities

For algorithms that produced prediction probability estimates, probabilities were well calibrated for observed species, i.e., the correlation between prediction probability and prediction accuracy followed the identity line (Fig. 4). For novel species, MycoAI produced well-calibrated probability estimates for both COI and ITS predictions. The remaining algorithms exhibited a positive correlation between prediction probability and

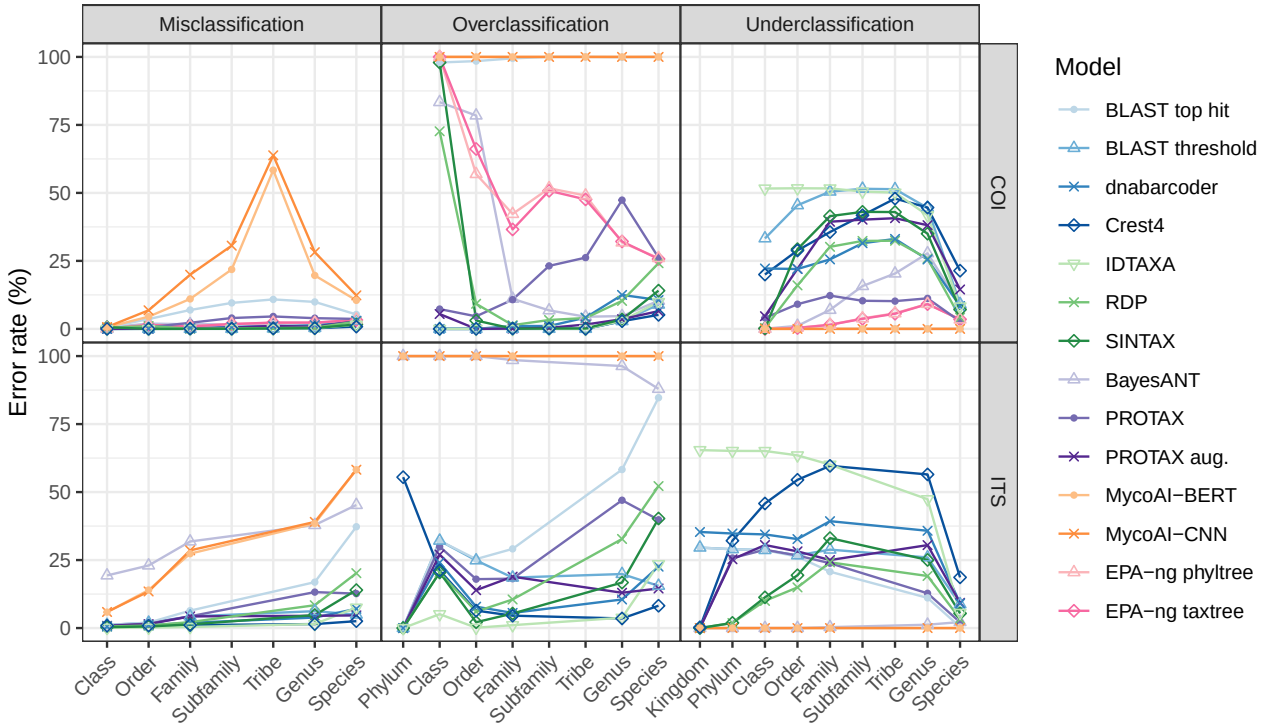

Figure 3: Rates of mis-, over-, and under-classification of COI and ITS sequences across classification algorithms and taxonomic ranks. Each algorithm is shown with a different shape and color, with hue representing the algorithm group: sequence similarity (blue), composition (green), probabilistic (purple), phylogenetic (pink), and neural network (orange).

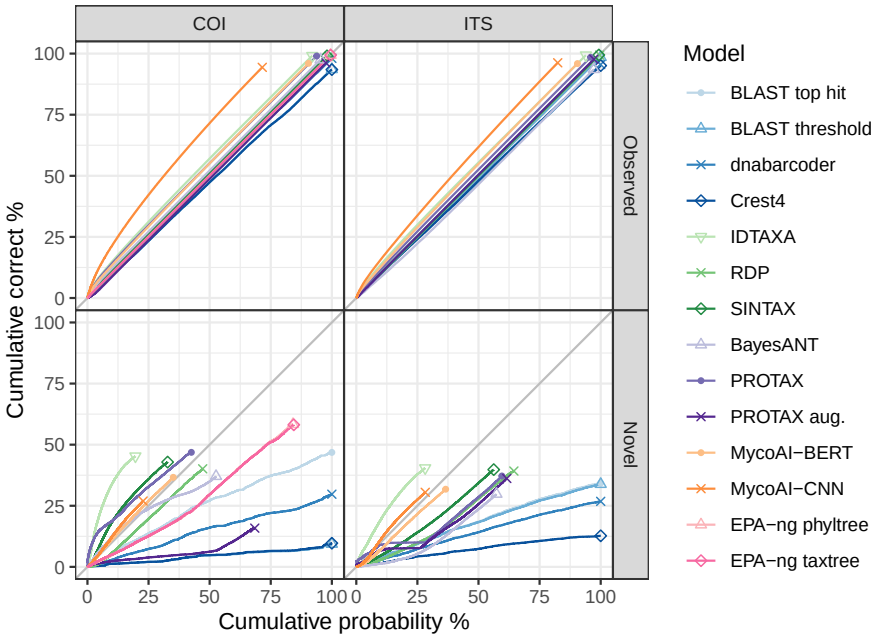

Figure 4: Calibration of different classification algorithms at genus-level taxonomic predictions. The x-axis shows the cumulative probability of the prediction with the highest probability or confidence score and the y-axis shows the cumulative proportion of correct predictions. Models are well calibrated if the line follows the identity line, here shown in gray. The y coordinate of the point at the end of the line displays the average accuracy. Calibration is shown for COI and ITS sequences, and for observed and novel species. Each algorithm is shown with a different shape and color, with hue indicating algorithm group: sequence similarity (blue), composition (green), probabilistic (purple), phylogenetic (pink), and neural network (orange). Calibration results are under the FP setting, i.e. where predictions are enforced for all query sequences.

accuracy as expected, but exact probability estimates were slightly too high (over-confident) or too low (under-confident). In general, algorithms were more over-confident for ITS than for COI, especially at lower ranks.

3.4 Classification coverage

There was a clear trade-off between accuracy and prediction coverage, where increasing confidence thresholds resulted in lower coverage but higher classification accuracy (Fig. 5). The exact relationship between prediction coverage and accuracy varied between algorithms; some achieved moderate accuracies at high coverage while others required aggressive filtering to achieve the same accuracy. Differences were especially pronounced for COI at speciel-level classifications, where EPA-ng preserved high accuracy even at 100% prediction coverage. Algorithms without confidence estimates, such as BLAST top hit, appeared as fixed points in this trade-off space, classifying relatively few sequences but often with high correctness. Importantly, there was no consistent threshold that yielded a high degree of correct classifications across algorithms or genetic markers. The optimal operating point instead depends on whether accuracy or recall is more important for the application.

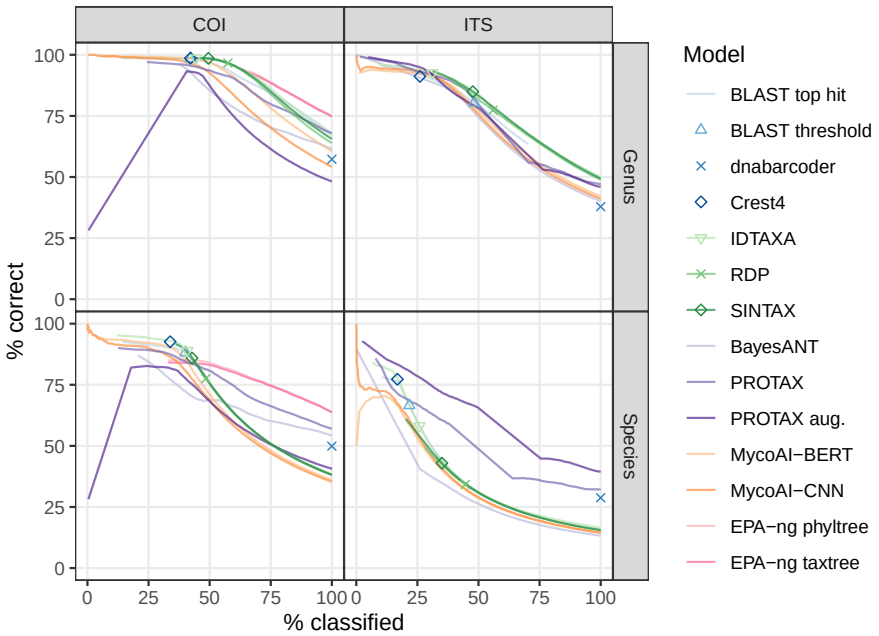

Figure 5: Classification accuracy as a function of prediction coverage for different algorithms. The x-axis shows the proportion of sequences classified and the y-axis shows the proportion of correct predictions. Continuous lines indicate algorithms with probability or confidence estimates, where varying the threshold yields a trade-off between coverage and accuracy. Single points represent algorithms without confidence estimates, corresponding to a fixed trade-off. Results are shown for COI and ITS sequences, and for genus- and species-level classifications. Each algorithm is represented by a different shape and color, with hue indicating algorithm group: sequence similarity (blue), composition (green), probabilistic (purple), phylogenetic (pink), and neural network (orange). Results are under the MP setting, i.e. allowing missing predictions.

3.5 Computational resources

Table 2: Computational resource use of different classification algorithms. Initialization/training and classification phases are shown separately for COI and ITS sequences. Reported metrics include wall-clock time, average CPU usage, peak memory consumption, and model size on disk. Results are provided for both single-core and multi-core CPU runs, as well as GPU runs where available.

		CPU /	Initialization/Training			Classification			
Marker	Algorithm	GPU	Time	CPU (%)	Mem (kB)	Time	CPU (%)	Mem (kB)	Disk (kB)
COI	bayesant	1	4:51	80%	1 538 500	1:33:10	98%	932 416	155 264
COI	bayesant	40	3:21	83%	1 539 312	3:05	3774%	917 572	155 264
COI	blast	1	0:02	49%	12 300	2:32:59	96%	45 768	13 424
COI	blast	40	0:02	30%	12 300	29:42	447%	145 068	13 428
COI	crest4	1	0:11	25%	56 740	2:50:34	98%	4 127 256	48 168
COI	crest4	40	0:11	58%	59 280	39:04	377%	4 128 224	48 168
COI	dnabarcoder	1	3d 18:30:54	97%	12 649 552	3:05:12	99%	1 626 492	480
COI	dnabarcoder	40	1:58:50	257%	50 514 432	38:14	438%	1 480 548	548
COI	epang-phyltree	1	10:46:02	99%	4 046 260	35:03	98%	36 606 576	15 452
COI	epang-phyltree	40	12:31:11	1237%	12 965 012	2:33	2073%	36 677 120	15 408
COI	epang-taxtree	1	15:26:02	99%	4 039 548	52:32	98%	36 601 288	13 436
COI	epang-taxtree	40	13:51:31	1150%	12 865 560	2:33	2046%	36 641 216	13 968
COI	idtaxa	1	13:31	95%	2 673 012	2:16:29	99%	1 015 012	50 688
COI	idtaxa	40	11:46	95%	2 687 336	12:13	1671%	1 023 304	50 688
COI	mycoai_bert	1	5d 22:03:58	99%	11 146 996	39:13	98%	2 274 284	425 776
COI	mycoai_bert	gpu	44:41	97%	3 151 880	0:41	97%	2 389 708	409 736
COI	mycoai_cnn	1	4:54:59	98%	1 116 584	1:06	50%	1 756 568	189 844
COI	mycoai_cnn	gpu	10:04	88%	3 761 492	0:35	91%	3 166 936	190 404
COI	protax-a	1	5:04:50	98%	428 900	1:56	96%	106 144	381 892
COI	rdp_nbc	1	0:38	82%	3 855 612	17:06	98%	1 112 828	122 840
COI	sintax	1	0:08	98%	78 672	5:47	99%	123 348	112 564
COI	sintax	40	0:07	79%	78 572	1:13	3856%	127 784	112 564
ITS	bayesant	1	0:36	68%	884 044	2:00:18	99%	1 181 204	30 064
ITS	bayesant	40	0:36	51%	894 472	3:52	3691%	1 372 000	30 064
ITS	blast	1	0:02	18%	12 164	30:18	99%	39 144	3 688
ITS	blast	40	0:02	11%	12 132	1:45	1807%	53 860	3 688
ITS	crest4	1	0:12	7%	41 936	39:21	99%	2 611 660	12 176
ITS	crest4	40	0:13	41%	41 288	6:54	559%	2 612 200	12 176
ITS	dnabarcoder	1	1d 01:29:45	98%	23 450 900	47:21	99%	360 412	132

ITS	dnabarcoder	40	15:06:21	102%	4 506 532	8:37	554%	322 716	132
ITS	idtaxa	1	7:38	96%	1 594 708	30:57	99%	780 868	12 296
ITS	idtaxa	40	6:00	88%	1 543 320	5:31	1591%	772 908	12 296
ITS	mycoai_bert	1	2d 03:07:01	98%	10 239 096	57:18	95%	2 247 300	308 012
ITS	mycoai_bert	gpu	16:55	70%	1 406 356	0:30	93%	927 632	292 680
ITS	mycoai_cnn	1	24:52	81%	687 324	0:19	92%	605 376	37 236
ITS	mycoai_cnn	gpu	7:43	47%	1 354 944	0:22	86%	816 624	37 552
ITS	protax	1	3:12:18	99%	373 084	8:34	99%	198 292	199 932
ITS	protax	40	2:27:59	101%	372 836	2:45	299%	198 264	199 896
ITS	rdp_nbc	1	0:16	68%	1 614 400	3:16	97%	720 780	35 876
ITS	sintax	1	0:02	77%	26 644	1:09	99%	33 036	27 776
ITS	sintax	40	0:02	69%	26 672	0:48	3865%	35 824	27 776

There was no clear relationship between classification performance and computational resource use (Table 2). Lightweight methods such as SINTAX achieved high accuracy with minimal runtimes and were the fastest overall when accounting for both training and classification, while resource-intensive algorithms such as MycoAI required substantially more time and memory without yielding superior accuracy. Model size on disk also varied widely, with MycoAI models occupying over 0.5–1.4GB compared to only a few hundred MB or less for composition-based classifiers like SINTAX and RDP, which makes the latter more practical for deployment in large-scale or resource-limited settings. Access to multiple CPU cores or GPUs reduced runtimes, especially during the classification phase, whereas fewer algorithms benefitted substantially from parallelization during initialization or training. Since training is typically a one-time cost, scalability during classification may be the more critical factor for real-world applications.

4 Discussion

DNA barcoding has proved a versatile tool for biodiversity research, and its applications are continuously expanding and developing following advances in sequencing technologies (Gostel and Kress, 2022; Van Klink et al., 2022; Niskanen et al., 2023). For many purposes, however, reliable taxonomic assignment of the resulting barcode sequences is paramount for downstream analyses. The primary factor determining classification accuracy is the completeness of the reference database, as emphasized in previous studies (Sickel et al., 2015; Taberlet et al., 2012; Edgar, 2018) and strongly supported by our result (e.g., Fig. 2). However, we have demonstrated that the sensitivity to incomplete reference data varies between taxonomic classifiers and that differences in performance between algorithms are more pronounced when species-level matches are lacking. Our results highlight that taxonomic classifiers vary in how they balance different types of classification errors, emphasizing the importance of careful selection of algorithms suitable to both the genetic marker targeted and the research question at hand.

4.1 High classification accuracy comes at the cost of prediction coverage

We found a general *trade-off* between classification accuracy and prediction coverage across all methods (Figure 5). This outcome is intuitive: algorithms that abstain from making predictions when uncertain have a lower risk of misclassification. When dealing with novel taxa, this corresponds to the balance between under- and over-classification. Under-classification increases accuracy at higher ranks by withholding predictions when uncertainty is high, while over-classification increases coverage but also increases the chance of errors at lower ranks. Because algorithms vary in how they navigate this trade-off, their suitability depends on the objectives of a given study. For instance, biodiversity monitoring or community surveys may benefit from a high prediction coverage at the expense of accuracy, as detecting rare taxa is crucial (Mouillot et al., 2013; Soliveres et al., 2016). Conversely, applications such as network analysis may prioritize accuracy, since misclassifications can create spurious links between taxa (Cuff et al., 2022). To support informed decision-making, we argue that classification algorithms should provide well-calibrated probabilities or confidence estimates.

Clearly reporting classification uncertainties not only enhances scientific reliability but also enables researchers to incorporate these probabilities into downstream analyses, thereby refining subsequent inferences. When reference databases are incomplete, algorithms should additionally distinguish between novelty and classification uncertainty, improving transparency and interpretability. Algorithms that rely on a single threshold to accept or reject classification must optimize thresholds to balance between different error types, often resulting in conservative cut-offs as observed here (e.g., Fig. 5). Treating confidence scores below such thresholds as indicative of novelty can thus inflate estimates of unseen taxa. For instance, both BLAST threshold and dnabarcoder largely over-assigned DNA sequences to novel branches (Fig. **??**), highlighting that many "under-classifications" reflect the stringency of the applied thresholds rather than true novelty. On the other hand, algorithms that explicitly account for the possibility of novel taxa and distinguish between novelty and classification uncertainty greatly enhance interpretability, enabling researchers to disentangle true biological signal from methodological conservatism and thereby draw more robust ecological and evolutionary inferences. Consequently, researchers should favor classification approaches that explicitly handle novelty when working in geographic regions or with taxonomic groups that are known to be underrepresented in reference databases.

4.2 Different algorithm classes perform best on different genetic markers

Genetic markers vary in their suitability for taxonomic classification of different taxonomic groups, owing to lineage-specific rates of evolution and variation in sequence composition (Hebert et al., 2003; Pentinsaari et al., 2016; Schoch et al., 2012). We evaluated classification algorithms on two widely used DNA barcodes chosen specifically for their contrasting characteristics: the COI barcode as a protein-coding gene with strong evolutionary constraints on the amino acid sequence and length, and ITS as a non-coding region of variable base composition and length.

While every method reached a high classification accuracy for observed species, we detected substantial differences in performance across genetic markers for novel taxa. For ITS, composition-based algorithms achieved the best results (e.g., Fig. 2, green colors). This is likely due to the high sequence variability and the resulting difficulty to establish position homology for distant species, which prevents reliable sequence alignments. Hence, k-mer-based approaches likely benefit from the strong signal provided by recurring short motifs, which are more

indicative of relatedness in a highly variable marker like ITS. Instead, the evolutionary constraints in COI may produce short sequence matches by pure chance. In this regard, the best performance for COI was achieved by phylogenetic placement (e.g., Fig. 2, pink colors), which explicitly employs evolutionary models to infer taxonomic relationships. We did not evaluate this class of algorithms for ITS due to the difficulties in providing a global alignment. However, alternative phylogenetic placement algorithms relying on k-mer composition are available (Balaban et al., 2019), and could be an interesting future avenue to explore for taxonomic classification of ITS sequences. Overall, our result indicates that a better understanding of both sequence composition and evolutionary constraints can help researchers select the most appropriate algorithm for their specific application.

4.3 Classification performance is not determined by computational resource use

We did not observe a general improvement in classification performance at the expense of higher computational costs. Indeed, the best performing algorithms for ITS included SINTAX, which had the lowest combined training and classification time, whereas computationally intensive phylogenetic placements were particularly rewarding for COI classifications. To save computational resources, large COI datasets may therefore benefit from a two-step strategy: first applying a fast algorithm such as SINTAX with a high confidence threshold to classify observed species, and then using the more costly EPA-ng to classify only those sequences that could not be confidently assigned, as in Sundh et al. (2024). Because SINTAX confidence estimates for COI were consistently under-confident (Fig. 4), a high confidence threshold would yield very few mis- or over-classifications.

Although we assessed scalability in terms of available computational resources, we did not specifically examine how the models scale with the size of the reference database, which is an important practical consideration. Realistically, the size of the present databases might limit the use of resource-heavy algorithms. For phylogenetic placement, for example, it may be necessary to preselect representative taxa for inclusion in the reference tree to make it computationally tractable.

4.4 Limited impact of read length on classification performance

In many applications, shorter sub-regions of the DNA barcode are employed. Shorter reads often amplify more reliably than full-length barcodes (Meusnier et al., 2008) and are therefore preferable when working with degraded DNA. They can also be sequenced more cost-effectively on high-throughput platforms, making them attractive for large-scale barcoding studies. For most algorithms, we observed minimal differences in classification performance between short reads and full-length barcodes, suggesting that the regions targeted by the shorter reads captured most of the variation relevant for species classification. Previous studies have reported similar results for both COI (Yeo et al., 2020) and ITS (Badotti et al., 2017).

Notably, dnabarcoder failed to classify 99% of ITS short reads as belonging to the kingdom Fungi, resulting in near-zero accuracy, and MycoAI showed significantly reduced performance on short reads across genetic markers and taxon sets. However, we trained all classifiers on the full-length barcodes and did not retrain them prior to classification of the short reads. The observed performance drop may therefore reflect sensitivity to mismatches between training and query sequence length, rather than an inherent inability of these algorithms to classify short sequences.

4.5 Barcode-based classifications under taxonomic uncertainty

For successful taxonomic classification of novel taxa from DNA barcodes, the genetic variation of the barcode must reflect the relatedness of taxonomic groups. For observed species, a genetic marker is generally suitable for taxonomic classification if the interspecific variation exceeds the intraspecific variation. For novel taxa, however, this criterion must be fulfilled also at higher taxonomic levels, so that the barcode sequence composition is sufficient to determine the correct genus, family, or even higher rank without an exact species match. This task is further complicated by the absence of universal divergence thresholds that reliably indicate the lowest shared taxonomic rank between two taxa (Edgar, 2018). In other words, even with the same genetic marker, a 3% divergence may correspond to taxa that share either a genus or a family. Additionally, there are incongruencies between currently accepted Linnean taxonomic classifications and molecular phylogenies, and taxonomic relationships are continually revised as new phylogenetic evidence corrects previous classifications (e.g., Möckel et al., 2022; Chen et al., 2021; Johnston et al., 2024; Ji et al., 2023). Beyond inconsistencies between reference databases, which differ in how frequently they update their taxonomies, such incongruencies can undermine the consistency of barcode-based classification by violating the fundamental assumption that taxonomy accurately reflects evolutionary relatedness. These challenges highlight the appeal of methods that infer evolutionary relationships directly rather than relying solely on taxonomic labels. Phylogenetic placement is one such approach, integrating evolutionary signal into the classification process. However, its utility for taxonomic classification ultimately depends on how closely the reference taxonomy reflects true phylogenetic relationships.

Phylogenetic placement accuracy may also depend on the quality of the reference tree. Previous studies have shown that phylogenetic backbones can improve tree inference from DNA barcodes by supplementing their limited signal for deep evolutionary relationships (Talavera et al., 2022; Liu et al., 2019). To examine whether a more accurate phylogenetic representation of higher-level relationships translates into improved classification performance, we evaluated EPA-ng using two reference trees: one constrained by taxonomy alone and one incorporating phylogenetic relationships on the family level and above. Hope-inspiringly, we observed minimal differences in classification result between the two reference trees, suggesting that constraining the tree to match taxonomy is generally sufficient for accurate classifications.

4.6 Recommendations for algorithm selection

Our results clearly show that the choice of taxonomic classifier is less critical when reference libraries are complete, allowing researchers to prioritize computational efficiency. Nevertheless, even with comprehensive reference data, it remains important to choose an algorithm that provides confidence estimates to flag uncertain cases. For this purpose, our results highlight SINTAX as a fast, memory-efficient option that produces well-calibrated, albeit slightly conservative, classification confidence estimates.

When reference libraries are incomplete, our results suggest that phylogenetic placement methods (e.g., EPA-ng) are preferable for COI barcodes, whereas composition-based methods (e.g., RDP, SINTAX, and IDTAXA) perform best for ITS barcodes. While our findings are specific for COI and ITS, the observed trends might be indicative for other barcodes sharing similar characteristics, though this remains to be validated. Under incomplete reference libraries, we further highlight the importance of selecting algorithms which explicitly

distinguishes between classification uncertainty and novel taxa, such as EPA-ng, PROTAX, and BayesANT. However, for ITS, no algorithm consistently achieved reliable identification of novel taxa across ranks. Consequently, the interpretation of whether uncertain classifications indicate true novelty is left to the user, underscoring the importance of transparent criteria for assigning novelty and clear reporting of well-calibrated confidence estimates.

To reduce computational demands for large COI barcode datasets, we recommend a two-step strategy, as also suggested by Sundh et al. (2024). First, apply a fast and resource-efficient method such as SINTAX with a high confidence threshold to classify most observed species. Then, apply phylogenetic placement to the remaining uncertain cases to optimize classification. Alhough not evaluated here, the computational efficiency of phylogenetic placement may improve if the reference tree is built from a preselected set of representative sequences (i.e., one sequence per species) rather than the full set used in this study.

Finally, our evaluations considered off-the-shelf approaches to taxonomic classification that leveraged default settings. In principle, one can further improve performance by properly tuning model-specific parameters to the training data and the specific marker, noting that some algorithms may be more sensitive to parameter choices than others (Hleap et al., 2021). For example, we observed more overconfidence in ITS than COI, potentially reflecting a calibration more optimized for barcodes with lower base heterogeneity.

References

Abarenkov, K., R. H. Nilsson, K.-H. Larsson, A. S. Taylor, T. May, T. G. Frøslev, J. Pawlowska, B. Lindahl, K. Põldmaa, C. Truong, D. Vu, T. Hosoya, T. Niskanen, T. Piirmann, F. Ivanov, A. Zirk, M. Peterson, T. Cheeke, Y. Ishigami, A. Jansson, T. Jeppesen, E. Kristiansson, V. Mikryukov, J. Miller, R. Oono, F. Ossandon, J. Paupério, I. Saar, D. Schigel, A. Suija, L. Tedersoo, and U. Kõljalg (2023, 11). The unite database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Research 52*(D1), D791–D797.

Abarenkov, K., P. Somervuo, R. H. Nilsson, P. M. Kirk, T. Huotari, N. Abrego, and O. Ovaskainen (2018). Protax-fungi: A web-based tool for probabilistic taxonomic placement of fungal internal transcribed spacer sequences. *New Phytologist 220*(2), 517–525.

Abarenkov, K., A. Zirk, T. Piirmann, R. Pöhönen, F. Ivanov, R. H. Nilsson, and U. Kõljalg (2025). Unite general fasta release for fungi 2. Dataset.

German Barcode of Life Consortium (2011). Gbol webportal at https://www.bolgermany.de. [Dataset]. Version: 20170316. Data Publisher: Zoological Research Museum Koenig – Leibniz Institute for Animal Biodiversity. Downloaded 2023-11-22.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). Basic local alignment search tool. *Journal of Molecular Biology 215*(3), 403–410.

Badotti, F., F. S. de Oliveira, C. F. Garcia, A. B. M. Vaz, P. L. C. Fonseca, L. A. Nahum, G. Oliveira, and A. Góes-Neto (2017). Effectiveness of its and sub-regions as dna barcode markers for the identification of basidiomycota (fungi). *BMC microbiology 17*(1), 42.

Balaban, M., S. Sarmashghi, and S. Mirarab (2019, 09). Apples: Scalable distance-based phylogenetic placement with or without alignments. *Systematic Biology 69*(3), 566–578.

Barbera, P., A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, and A. Stamatakis (2019). EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic Biology 68*(2), 365–369.

Bengtsson-Palme, J., M. Ryberg, M. Hartmann, S. Branco, Z. Wang, A. Godhe, P. De Wit, M. Sánchez-García, I. Ebersberger, F. de Sousa, A. Amend, A. Jumpponen, M. Unterseher, E. Kristiansson, K. Abarenkov, Y. J. K. Bertrand, K. Sanli, K. M. Eriksson, U. Vik, V. Veldre, and R. H. Nilsson (2013). Improved software detection and extraction of its1 and its2 from ribosomal its sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution 4*(10), 914–919.

BOLD Systems (2024). Bold dna barcode reference library. https://doi.org/10.5883/DP-BOLD_Public.29-Mar-2024. Accessed via https://boldsystems.org on 2024-05-21.

Bonato, K. O., P. C. Silva, F. R. Carvalho, and L. R. Malabarba (2022). Trophic interactions of vampire catfishes (siluriformes: Vandelliinae) revealed by metabarcoding analysis of stomach contents. *Freshwater Biology 67*(3), 542–548.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (2009). BLAST+: Architecture and applications. *BMC bioinformatics 10*, 421.

Chen, H., Z. Lahey, E. J. Talamas, A. A. Valerio, O. A. Popovici, L. Musetti, H. Klompen, A. Polaszek, L. Masner, A. D. Austin, et al. (2021). An integrated phylogenetic reassessment of the parasitoid superfamily platygastroidea (hymenoptera: Proctotrupomorpha) results in a revised familial classification. *Systematic Entomology 46*(4), 1088–1113.

Cuff, J. P., F. M. Windsor, M. P. Tercel, J. J. Kitson, and D. M. Evans (2022). Overcoming the pitfalls of merging dietary metabarcoding into ecological networks. *Methods in Ecology and Evolution 13*(3), 545–559.

Czech, L., P. Barbera, and A. Stamatakis (2020). Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics 36*(10), 3263–3265.

Deshpande, V., Q. Wang, P. Greenfield, M. Charleston, A. Porras-Alfaro, C. R. Kuske, J. R. Cole, D. J. Midgley, and N. Tran-Dinh (2016). Fungal identification using a bayesian classifier and the warcup training set of internal transcribed spacer sequences. *Mycologia 108*(1), 1–5.

Edgar, R. C. (2010, October). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics 26*(19), 2460–2461.

Edgar, R. C. (2016). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161.

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ 6*, e4652.

Elbrecht, V., T. W. A. Braukmann, N. V. Ivanova, S. W. J. Prosser, M. Hajibabaei, M. Wright, E. V. Zakharov, P. D. N. Hebert, and D. Steinke (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ 7*, e7745.

Feibelman, T., P. Bayman, and W. G. Cibula (1994, June). Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. *Mycological Research 98*(6), 614–618.

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek (1994, October). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit i from diverse metazoan invertebrates. *Molecular marine biology and biotechnology 3*(5), 294—299.

Furneaux, B., M. Bahram, A. Rosling, N. S. Yorou, and M. Ryberg (2021). Long- and short-read metabarcoding technologies reveal similar spatiotemporal structures in fungal communities. *Molecular Ecology Resources 21*(6), 1833–1849.

Gostel, M. R. and W. J. Kress (2022). The expanding role of dna barcodes: Indispensable tools for ecology, evolution, and conservation. *Diversity 14*(3), 213.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences 270*(1512), 313–321.

Hebert, P. D. N., M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis (2004, 09). Identification of birds through dna barcodes. *PLOS Biology 2*(10).

Hleap, J. S., J. E. Littlefair, D. Steinke, P. D. N. Hebert, and M. E. Cristescu (2021). Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources 21*(7), 2190–2203.

Janzen, D. H., M. Hajibabaei, J. M. Burns, W. Hallwachs, E. Remigio, and P. D. Hebert (2005). Wedding biodiversity inventory of a large and complex lepidoptera fauna with dna barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences 360*(1462), 1835–1845.

Ji, X., Y.-F. Sun, D.-M. Wu, N. Gao, and B.-K. Cui (2023). An updated phylogenetic assessment and taxonomic revision of perenniporia sensu lato (polyporales, basidiomycota). *Journal of Fungi 9*(2), 173.

Johnston, N. P., T. Pape, M. Piwczyński, J. F. Wallman, B. M. Wiegmann, B. K. Cassel, K. Akbarzadeh, and K. Szpila (2024). Anchored phylogenomics and revised classification of the miltogramminae (diptera: Sarcophagidae). *Systematic Entomology 49*(1), 138–155.

Khomich, M., F. Cox, C. J. Andrew, T. Andersen, H. Kauserud, and M. L. Davey (2018, December). Coming up short: Identifying substrate and geographic biases in fungal sequence databases. *Fungal Ecology 36*, 75–80.

Lanzén, A., S. L. Jørgensen, D. H. Huson, M. Gorfer, S. H. Grindhaug, I. Jonassen, L. Øvreås, and T. Urich (2012). CREST – Classification Resources for Environmental Sequence Tags. *PLOS ONE 7*(11), e49334.

Leandro, C., P. Jay-Robert, and J. Pétillon (2024). edna for monitoring and conserving terrestrial arthropods: Insights from a systematic map and barcode repositories assessments. *Insect Conservation and Diversity 17*(4), 565–578.

Lindahl, B. D., R. H. Nilsson, L. Tedersoo, K. Abarenkov, T. Carlsen, R. Kjøller, U. Kõljalg, T. Pennanen, S. Rosendahl, J. Stenlid, et al. (2013). Fungal community analysis by high-throughput sequencing of amplified markers–a user's guide. *New Phytologist 199*(1), 288–299.

Liu, J., J. Liu, Y.-X. Shan, X.-J. Ge, and K. S. Burgess (2019, May). The use of DNA barcodes to estimate phylogenetic diversity in forest communities of southern China. *Ecology and Evolution 9*(9), 5372–5379.

Manter, D. K. and J. M. Vivanco (2007, October). Use of the ITS primers, ITS1F and ITS4, to characterize fungal abundance and diversity in mixed-template samples by qPCR and length heterogeneity analysis. *Journal of Microbiological Methods 71*(1), 7–14.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal 17*(1), 10–12.

Meusnier, I., G. A. Singer, J.-F. Landry, D. A. Hickey, P. D. Hebert, and M. Hajibabaei (2008). A universal dna mini-barcode for biodiversity analysis. *BMC genomics 9*(1), 214.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution 37*(5), 1530–1534.

Möckel, L., K. Meusemann, B. Misof, V. U. Schwartze, H. H. De Fine Licht, K. Voigt, B. Stielow, S. de Hoog, R. G. Beutel, and J. Buellesbach (2022). Phylogenetic revision and patterns of host specificity in the fungal subphylum entomophthoromycotina. *Microorganisms 10*(2), 256.

Mouillot, D., D. R. Bellwood, C. Baraloto, J. Chave, R. Galzin, M. Harmelin-Vivien, M. Kulbicki, S. Lavergne, S. Lavorel, N. Mouquet, et al. (2013). Rare species support vulnerable functions in high-diversity ecosystems. *PLoS biology 11*(5), e1001569.

Murali, A., A. Bhargava, and E. S. Wright (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome 6*(1), 140.

Niskanen, T., R. Lücking, A. Dahlberg, E. Gaya, L. M. Suz, V. Mikryukov, K. Liimatainen, I. Druzhinina, J. R. Westrip, G. M. Mueller, et al. (2023). Pushing the frontiers of biodiversity research: Unveiling the global diversity, distribution, and conservation of fungi. *Annual review of Environment and resources 48*(1), 149–176.

Pentinsaari, M., H. Salmela, M. Mutanen, and T. Roslin (2016). Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports 6*(1), 35275.

Porter, T. M. and M. Hajibabaei (2020). Putting coi metabarcoding in context: The utility of exact sequence variants (esvs) in biodiversity analysis. *Frontiers in Ecology and Evolution 8*, 248.

Ranwez, V., E. J. P. Douzery, C. Cambon, N. Chantret, and F. Delsuc (2018). MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Molecular Biology and Evolution 35*(10), 2582–2584.

Ratnasingham, S. and P. D. N. Hebert (2007). Bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes 7*(3), 355–364.

Ratnasingham, S., C. Wei, D. Chan, J. Agda, J. Agda, L. Ballesteros-Mejia, H. A. Boutou, Z. M. El Bastami, E. Ma, R. Manjunath, D. Rea, C. Ho, A. Telfer, J. McKeowan, M. Rahulan, C. Steinke, J. Dorsheimer, M. Milton, and P. D. N. Hebert (2024). *BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data*, pp. 403–441. New York, NY: Springer US.

Riaz, T., W. Shehzad, A. Viari, F. Pompanon, P. Taberlet, and E. Coissac (2011, November). ecoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research 39*(21), e145.

Richardson, R. T., J. Bengtsson-Palme, and R. M. Johnson (2017). Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. *Molecular Ecology Resources 17*(4).

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ 4*, e2584.

Romeijn, L., A. Bernatavicius, and D. Vu (2024). MycoAI: Fast and accurate taxonomic classification for fungal ITS sequences. *Molecular Ecology Resources n/a*(n/a), e14006.

Roslin, T., P. Somervuo, M. Pentinsaari, P. D. N. Hebert, J. Agda, P. Ahlroth, P. Anttonen, J. Aspi, G. Blagoev, S. Blanco, D. Chan, T. Clayhills, J. deWaard, S. deWaard, T. Elliot, R. Elo, S. Haapala, E. Helve, J. Ilmonen, P. Hirvonen, C. Ho, J. Itämies, V. Ivanov, J. Jakovlev, A. Juslén, R. Jussila, J. Kahanpää, L. Kaila, Jari-PekkaKaitila, A. Kakko, I. Kakko, A. Karhu, S. Karjalainen, J. Kjaerandsen, J. Koskinen, E. M. Laasonen, L. Laasonen, E. Laine, P. Lampila, V. Levesque-Beaudin, L. Lu, M. Lähteenaro, P. Majuri, S. Malmberg, R. Manjunath, P. Martikainen, J. Mattila, J. McKeown, P. Metsälä, M. Miklasevskaja, M. Miller, R. Miskie, A. Muinonen, Veli-MattiMukkala, S. Naik, N. Nikolova, K. Nupponen, O. Ovaskainen, I. Österblad, L. Paasivirta, T. Pajunen, P. Parkko, J. Paukkunen, R. Penttinen, K. Perez, J. Pohjoismäki, S. Prosser, M. Raekunnas, M. Rahulan, M. Rannisto, S. Ratnasingham, P. Raukko, A. Rinne, T. Rintala, S. Miranda Romo, J. Salmela, J. Salokannel, R. Savolainen, L. Schulman, P. Sihvonen, D. Soliman, J. Sones, C. Steinke, G. Ståhls, J. Tabell, M. Tiusanen, G. Várkonyi, E. J. Vesterinen, E. Viitanen, V. Vikberg, M. Viitasaari, J. Vilen, C. Warne, C. Wei, K. Winqvist, E. Zakharov, and M. Mutanen (2022). A molecular-based identification resource for the arthropods of Finland. *Molecular Ecology Resources 22*(2), 803–822.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. (2020). Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database 2020*, baaa062.

Schoch, C. L., K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, and F. B. Consortium (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences 109*(16), 6241–6246.

Shokralla, S., J. F. Gibson, H. Nikbakht, D. H. Janzen, W. Hallwachs, and M. Hajibabaei (2014, February). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, n/a–n/a.

Sickel, W., M. J. Ankenbrand, G. Grimmer, A. Holzschuh, S. Härtel, J. Lanzen, I. Steffan-Dewenter, and A. Keller (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC ecology 15*(1), 20.

Soliveres, S., P. Manning, D. Prati, M. M. Gossner, F. Alt, H. Arndt, V. Baumgartner, J. Binkenstein, K. Birkhofer, S. Blaser, et al. (2016). Locally rare species influence grassland ecosystem multifunctionality. *Philosophical Transactions of the Royal Society B: Biological Sciences 371*(1694), 20150269.

Somervuo, P., S. Koskela, J. Pennanen, R. Henrik Nilsson, and O. Ovaskainen (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics 32*(19), 2920–2927.

Somervuo, P., D. W. Yu, C. C. Xu, Y. Ji, J. Hultman, H. Wirta, and O. Ovaskainen (2017, April). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution 8*(4), 398–407.

Stork, N. E. (2018, January). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology 63*(1), 31–45.

Sundh, J., E. Granqvist, E. Iwaszkiewicz-Eggebrecht, L. Manoharan, L. J. A. van Dijk, R. Goodsell, N. N. Godeiro, B. C. Bellini, P. Łukasik, A. Miraldo, T. Roslin, A. J. M. Tack, A. F. Andersson, and F. Ronquist (2024). Happ: High-accuracy pipeline for processing deep metabarcoding data. *bioRxiv*.

Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev (2012). Towards next-generation biodiversity assessment using dna metabarcoding. *Molecular ecology 21*(8), 2045–2050.

Talavera, G., V. Lukhtanov, N. E. Pierce, and R. Vila (2022, February). DNA Barcodes Combined with Multilocus Data of Representative Taxa Can Generate Reliable Higher-Level Phylogenies. *Systematic Biology 71*(2), 382–395.

Tedersoo, L., M. Bahram, L. Zinger, R. H. Nilsson, P. G. Kennedy, T. Yang, S. Anslan, and V. Mikryukov (2022). Best practices in metabarcoding of fungi: From experimental design to results. *Molecular Ecology 31*(10), 2769–2795.

Tedersoo, L., V. Mikryukov, S. Anslan, M. Bahram, A. N. Khalid, A. Corrales, A. Agan, A.-M. Vasco-Palacios, A. Saitta, A. Antonelli, A. C. Rinaldi, A. Verbeken, B. P. Sulistyo, B. Tamgnoue, B. Furneaux, C. D. Ritter, C. Nyamukondiwa, C. Sharp, C. Marín, D. Q. Dai, D. Gohar, D. Sharmah, E. M. Biersma, E. K. Cameron, E. De Crop, E. Otsing, E. A. Davydov, F. E. Albornoz, F. Q. Brearley, F. Buegger, G. Gates,

G. Zahn, G. Bonito, I. Hiiesalu, I. Hiiesalu, I. Zettur, I. C. Barrio, J. Pärn, J. Heilmann-Clausen, J. Ankuda, J. Y. Kupagme, J. Sarapuu, J. G. Maciá-Vicente, J. D. Fovo, J. Geml, J. M. Alatalo, J. Alvarez-Manjarrez, J. Monkai, K. Põldmaa, K. Runnel, K. Adamson, K. A. Bråthen, K. Pritsch, K. I. Tchan, K. Armolaitis, K. D. Hyde, K. K. Newsham, K. Panksep, L. A. Adebola, L. J. Lamit, M. Saba, M. E. da Silva Cáceres, M. Tuomi, M. Gryzenhout, M. Bauters, M. Bálint, N. Wijayawardene, N. Hagh-Doust, N. S. Yorou, O. Kurina, P. E. Mortimer, P. Meidl, R. H. Nilsson, R. Puusepp, R. Casique-Valdés, R. Drenkhan, R. Garibay-Orijel, R. Godoy, S. Alfarraj, S. Rahimlou, S. Põlme, S. V. Dudov, S. Mundra, T. Ahmed, T. Netherway, T. W. Henkel, T. Roslin, V. E. Fedosov, V. G. Onipchenko, W. A. E. Yasanthika, Y. W. Lim, M. Piepenbring, D. Klavina, U. Kõljalg, and K. Abarenkov (2021, November). The Global Soil Mycobiome consortium dataset for boosting fungal diversity research. *Fungal Diversity 111*(1), 573–588.

Van Klink, R., T. August, Y. Bas, P. Bodesheim, A. Bonn, F. Fossøy, T. T. Høye, E. Jongejans, M. H. Menz, A. Miraldo, et al. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in ecology & evolution 37*(10), 872–885.

Vass, M., K. Eriksson, U. Carlsson-Graner, J. Wikner, and A. Andersson (2022, 10). Co-occurrences enhance our understanding of aquatic fungal metacommunity assembly and reveal potential host–parasite interactions. *FEMS Microbiology Ecology 98*(11).

Vu, D., M. Groenewald, M. De Vries, T. Gehrmann, B. Stielow, U. Eberhardt, A. Al-Hatmi, J. Z. Groenewald, G. Cardinali, J. Houbraken, T. Boekhout, P. W. Crous, V. Robert, and G. J. M. Verkley (2019). Large-scale generation and analysis of filamentous fungal dna barcodes boosts coverage for kingdom fungi and reveals thresholds for fungal species and higher taxon delimitation. *Studies in mycology 92*(1), 135–154.

Vu, D., M. Groenewald, S. Szöke, G. Cardinali, U. Eberhardt, B. Stielow, M. De Vries, G. Verkleij, P. Crous, T. Boekhout, et al. (2016). Dna barcoding analysis of more than 9 000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in mycology 85*(1), 91–105.

Vu, D., R. H. Nilsson, and G. J. M. Verkley (2022). Dnabarcoder: An open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. *Molecular Ecology Resources 22*(7), 2793–2809.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol. 73*(16), 5261–5267.

Winand, R., E. D'hooge, A. Van Uffelen, B. Bogaerts, J. Van Braekel, S. Hoffman, N. H. C. J. Roosens, P. Becker, S. C. J. De Keersmaecker, and K. Vanneste (2025). Investigating fungal diversity through metabarcoding for environmental samples: assessment of its1 and its2 illumina sequencing using multiple defined mock communities with different classification methods and reference databases. *BMC Genomics 26*(1), 729.

Yeo, D., A. Srivathsan, and R. Meier (2020, 02). Longer is not always better: Optimizing barcode length for large-scale species discovery and identification. *Systematic Biology 69*(5), 999–1015.

Zito, A., T. Rigon, and D. B. Dunson (2023). Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. *Methods in Ecology and Evolution 14*(2), 529–542.