Received: 17 May 2021 | Revised: 12 July 2021 | Accepted: 13 July 2021

DOI: 10.1111/1755-0998.13473

MOLECULAR ECOLOGY RESOURCES WILEY

NEWS AND VIEWS

PERSPECTIVE

Just keep it simple? Benchmarking the accuracy of taxonomy assignment software in metabarcoding studies

Holly M. Bik

Department of Marine Sciences and Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA

Correspondence
Holly M. Bik, Department of Marine Sciences and Institute of Bioinformatics, University of Georgia, Athens, GA, USA.
Email: hbik@uga.edu

How do you put a name on an unknown piece of DNA? From microbes to mammals, high-throughput metabarcoding studies provide a more objective view of natural communities, overcoming many of the inherent limitations of traditional field surveys and microscopy-based observations (Deiner et al., 2017). Taxonomy assignment is one of the most critical aspects of any metabarcoding study, yet this important bioinformatics task is routinely overlooked. Biodiversity surveys and conservation efforts often depend on formal species inventories: the presence (or absence) of species, and the number of individuals reported across space and time. However, computational workflows applied in eukaryotic metabarcoding studies were originally developed for use with bacterial/archaeal data sets, where microbial researchers rely on one conserved locus (nuclear 16S rRNA) and have access to vast databases with good coverage across most prokaryotic lineages – a situation not mirrored in most multicellular taxa. In this issue of *Molecular Ecology Resources*, Hleap et al. (2021) carry out an extensive benchmarking exercise focused on taxonomy assignment strategies for eukaryotic metabarcoding studies utilizing the mitochondrial Cytochrome C oxidase I marker gene (COI). They assess the performance and accuracy of software tools representing diverse methodological approaches: from "simple" strategies based on sequence similarity and composition, to model-based phylogenetic and probabilistic classification tools. Contrary to popular assumptions, less complex approaches (BLAST and the QIIME2 feature classifier) consistently outperformed more sophisticated mathematical algorithms and were highly accurate for assigning taxonomy at higher levels (e.g. family). Lower-level assignments at the genus and species level still pose significant challenge for most existing algorithms, and sparse eukaryotic reference databases further limit software performance. This study illuminates current best practices for metabarcoding taxonomy assignments, and underscores the need for community-driven efforts to expand taxonomic and geographic representation in reference DNA barcode databases.

KEYWORDS
COI, fish, insects, metabarcoding, species identification, zooplankton

The authors take a rigorous and exhaustive approach towards evaluating these seven computational algorithms. First, they assembled a taxonomically diverse set of mock communities, focusing on three COI metabarcoding data sets obtained from zooplankton, fish, and insects (Figure 1). Mock communities were designed to be compositionally heterogeneous (including varying numbers of taxonomic groups and individuals per taxa), thus mimicking community DNA obtained from natural ecosystems. Next, Hleap et al. (2021) manually constructed a well-annotated reference database of COI barcodes, including realistic subsampling and gene alignments required for phylogenetic and probabilistic taxonomy classifiers. Finally, the performance of taxonomy assignment software was objectively compared using a suite of defined metrics, including scoring of false discovery rates (the proportion of false predictions), true positive rates (the proportion of true positive predictions), and the Matthews correlation coefficient (a metric comparing the observed versus predicted taxonomy classification), as well as other composite metrics. Software tools were also compared in terms of optimization runs (e.g., algorithm training), parameter choices, and the computational time and resources needed to complete metabarcoding taxonomy assignments.

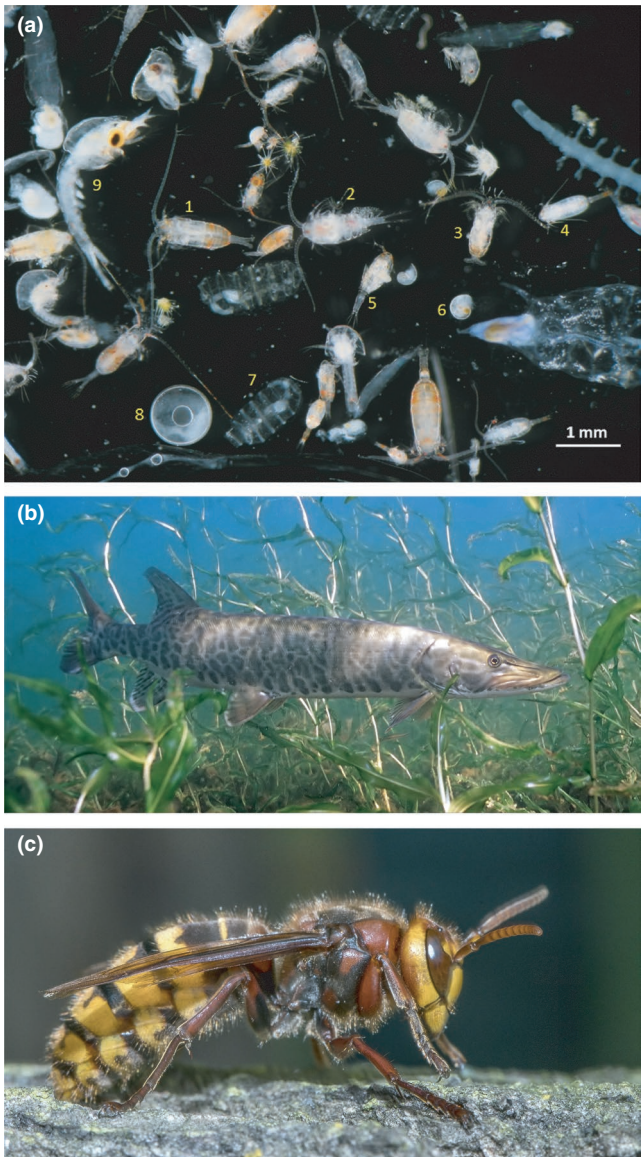

FIGURE 1 The article by Hleap et al. (2021) assesses metabarcoding taxonomy assignments using three taxonomically diverse mock communities spanning freshwater, marine, and terrestrial habitats: (a) Zooplankton, a type of pelagic marine community comprising a mixture of microbial metazoa taxa, fish larvae, and eggs; (b) Bony fish, and (c) Arthropods. All images are public domain or CC-BY licensed content obtained from Wikipedia. Image credit: Adriana Zingone and LTER-MC team (Zooplankton), Eric Engbretson (Muskellunge, *Esox masquinongy*), and Jerzy Strzelecki (European Hornet, *Vespa crabro*)

Notably, this study represents the first neutral benchmarking effort undertaken for a eukaryotic marker gene, leveraging the largest in vitro eukaryotic mock communities to date. Unlike other benchmarking studies, mock communities were manually curated in order to eliminate signals from environmental and extracellular DNA and ensure that sequencing data sets represented only the target taxa. These realistic mock communities also proved to be integral for parameter optimization, and the accuracy of most taxonomic assignment methods was notably improved after using a training data set where the community resembled the expected sample. Similarly, the authors cleverly chose to include "fake sequences" (shuffled nucleotides with the same composition as the original reads) into the data set so that taxonomy assignment strategies could be assessed using true negatives within each mock community. The methods implemented here have obvious relevance to other metabarcoding loci and taxonomic groups. Hleap et al. (2021) placed strong emphasis on reproducibility so that these benchmarking exercises can be easily extended to other data sets: all scripts, pipelines, and study outputs have been made freely available in an open source GitHub repository (https://github.com/jshleap/TA_pipes).

The results of Hleap et al. (2021) challenge the assumption that complex mathematical approaches return more accurate taxonomy assignments for metabarcoding amplicons. In fact, "simple" tools outperformed all other methods, with the QIIME2 feature classifier and BLAST top hit ranking first and second in accuracy of taxonomic classification, respectively. These two tools performed consistently well across all mock community data sets (accuracy did not decrease with different community compositions), and taxonomy assignments were particularly accurate at higher levels (e.g., family). In contrast, all other software tools showed variable performance and accuracy depending on the underlying data set. More complex machine learning methods performed especially poorly if the underlying reference barcode database was heterogeneous or had patchy coverage across taxonomic groups. It is important to note that even BLAST and QIIME2 demonstrated significant reductions in accuracy at the genus and species level. Sequence similarity/composition approaches will always return a "best hit" from the reference database. Reference barcode databases are truly massive for many taxa, and it is not difficult to locate a distantly related sequence given the conserved nature of many metabarcoding loci. Furthermore, even basic taxonomy assignment methods require customization and parameter optimization that are fine tuned for each metabarcoding data set.

BLAST approaches require an indexed reference database and the QIIME2 feature classifier involves a computationally intensive training step for reference sequences. Although public databases and pretrained reference sets exist for both of these tools, these precomputed resources often contain misannotations or glaring gaps in taxonomic coverage (Leray et al., 2018; Macheriotou et al., 2019).

Perhaps the most surprising result of the Hleap et al. (2021) study is the abysmal performance of the automated phylogenetic strategies. In theory, a phylogeny-aware algorithm could improve metabarcoding taxonomy assignments by leveraging the inherent evolutionary information contained within nucleotide or protein alignments for any given locus. However, automated phylogenetic approaches use heuristics that sacrifice accuracy in favor of speed, even assigning fake shuffled sequences (true negatives) to a wrong taxon. Other popular phylogenetic packages exist, such as Bayesian-based pplacer (Matsen et al., 2010) and the RAxML-based Evolutionary Placement Algorithm (Stamatakis, 2014), but these require significant downstream manual effort to view and interpret taxonomy assignments. Phylogenetic strategies thus offer substantial promise for improving the accuracy of lower-level taxonomic assignments (genus and species level), but the current software landscape requires labour-intensive curation efforts in order to achieve this goal. There is a clear need for improved phylogenetic workflows that can be more easily applied to metabarcoding studies.

The benchmarking results from Hleap et al. (2021) also convey an important warning about the interdependence of taxonomy assignment workflows and reference databases. Achieving adequate taxonomic coverage in reference databases is of paramount importance for all metabarcoding workflows, especially for studies of poorly sampled taxonomic groups and geographic regions – notably, only BLAST and the QIIME2 feature classifier were able to sufficiently cope with existing gaps in reference databases. Genetic heterogeneity presents another challenge, since current taxonomy assignment algorithms do not explicitly incorporate models accounting for intraspecific variability (e.g., mitochondrial haplotypes recovered in COI-based studies). This issue even impacts well-sampled vertebrate taxa, and Hleap et al. (2021) point to hybridization in bony fish species as a possible cause of low taxonomy assignment accuracy in probabilistic classification tools. Other studies have conveyed the importance of having local fauna represented in study-specific reference databases, demonstrating that the inclusion of endemic versus non-native taxa can potentially bias taxonomy assignments even in well-sampled vertebrate groups (Gold et al., 2021). Taking into account all these concerns, Hleap et al. (2021) recommend a multistep taxonomy assignment workflow, beginning with a narrow (focused) search across the expected sample diversity, and progressively widening the search scope until all metabarcoding reads have been accounted for.

For metabarcoding studies, it is clear that a "one size fits all" approach to taxonomy assignment is unlikely to produce the most biologically accurate results. Putting a name on an unknown piece of DNA will depend on the choice of genetic locus, the taxonomic groups best captured by that locus, and the availability and coverage of reference barcode databases. From a scientific perspective, computational power and resources are no longer the limiting factor – Hleap et al. (2021) note that most algorithms can rapidly assign taxonomy using similar memory requirements, and computationally intensive tools do not offer any advantage in terms of accuracy. Instead, the true bottleneck for metabarcoding studies appears to lie with reference databases themselves. For most taxonomic groups on earth, it is unlikely that existing DNA barcodes have adequately captured biodiversity across local, regional, and global scales (Porter & Hajibabaei, 2018). Even the most well-performing taxonomy assignment algorithms are at the mercy of public databases and training sets. There is a clear need to incentivize traditional barcoding alongside high-throughput approaches, helping to build stronger collections of reference sequences linked to formal species names and descriptions.

ORCID

Holly M. Bik https://orcid.org/0000-0002-4356-3837

REFERENCES

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.

Gold, Z., Curd, E., Goodwin, K., Choi, E., Frable, B., Thompson, A., Walker, H., Burton, R., Kacev, D., & Barber, P. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Authorea*. Preprints. https://doi.org/10.22541/au.161407483.33882798/v1

Hleap, J. S., Littlefair, J. E., Steinke, D., Hebert, P. D. N., & Cristescu, M. E. (2021). Assessment of current taxonomy assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources*, *21*, 2190–2203. https://doi.org/10.1111/1755-0998.13407

Leray, M., Ho, S.-L., Lin, I.-J., & Machida, R. J. (2018). MIDORI server: A webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics*, *34*, 3753–3754. https://doi.org/10.1093/bioinformatics/bty454

Macheriotou, L., Guilini, K., Bezerra, T. N., Tytgat, B., Nguyen, D. T., Phuong Nguyen, T. X., Noppe, F., Armenteros, M., Boufahja, F., Rigaux, A., Vanreusel, A., & Derycke, S. (2019). Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecology and Evolution*, *9*, 1211–1226. https://doi.org/10.1002/ece3.4814

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*, 538. https://doi.org/10.1186/1471-2105-11-538

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS One*, *13*(9), e0200177.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

How to cite this article: Bik, H. M. (2021). Just keep it simple? Benchmarking the accuracy of taxonomy assignment software in metabarcoding studies. *Molecular Ecology Resources*, 21, 2187–2189. https://doi.org/10.1111/1755-0998.13473