A peer-reviewed version of this preprint was published in PeerJ on 26 June 2018.

Richardson RT, Bengtsson-Palme J, Gardiner MM, Johnson RM. 2018. A reference cytochrome c oxidase subunit I database curated for hierarchical classification of arthropod metabarcoding data. PeerJ 6:e5126 https://doi.org/10.7717/peerj.5126

A Reference Cytochrome C Oxidase Subunit I Database Curated for Hierarchical Classification of Arthropod Metabarcoding Data

Rodney T. Richardson[1*], Johan Bengtsson-Palme[2,3], Mary M. Gardiner[4], Reed M. Johnson[1]

[1] Department of Entomology, The Ohio State University–Ohio Agricultural Research and Development Center, 1680 Madison Ave., Wooster, Ohio 44691 USA.

[2] Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, SE-413 46, Gothenburg, Sweden

[3] Center for Antibiotic Resistance research (CARe) at University of Gothenburg, Box 440, SE-40530, Gothenburg, Sweden

[4] Department of Entomology, The Ohio State University, 2021 Coffey Road, Columbus, Ohio 43210 USA.

Corresponding Author:
Rodney Richardson

Email address: richardson.827@osu.edu

Abstract

Metabarcoding is a popular application which warrants continued methods optimization. To maximize barcoding inferences, hierarchy-based sequence classification methods are increasingly common. We present methods for the construction and curation of a database designed for hierarchical classification of a 157 bp barcoding region of the arthropod cytochrome c oxidase subunit I (COI) locus. We produced a comprehensive arthropod COI amplicon dataset including annotated arthropod COI sequences and COI sequences extracted from arthropod whole mitochondrion genomes, which provided the only source of representation for Zoraptera, Callipodida and Holothyrida. The database contains extracted sequences of the target amplicon from all major arthropod clades, including all insect orders, all arthropod classes and Onychophora, Tardigrada and Mollusca outgroups. During curation, we extracted the COI region of interest from approximately 81 percent of the input sequences, corresponding to 73 percent of the genus-level diversity found in the input data. Further, our analysis revealed a high degree of sequence redundancy within the NCBI nucleotide database, with a mean of approximately 11 sequence entries per species in the input data. The curated, low-redundancy database is included in the Metaxa2 sequence classification software (http://microbiology.se/software/metaxa2/). Using this database with the Metaxa2 classifier, we characterized the relationship between the Metaxa2 reliability score, an estimate of classification confidence, and classification error probability. We used this analysis to select a reliability score threshold which minimized error. We then estimated classification sensitivity, false discovery rate and overclassification, the propensity to classify sequences from taxa not represented in the reference database. Our work will help researchers design and evaluate classification databases and conduct metabarcoding on arthropods and alternate taxa.

Introduction

With the increasing availability of high-throughput DNA sequencing, scientists with a wide diversity of backgrounds and interests are increasingly utilizing this technology to achieve a variety of goals. One growing area of interest involves the use of metabarcoding, or amplicon sequencing, for biomonitoring, biodiversity assessment and community composition inference (Yu et al. 2012; Guardiola et al. 2015; Richardson et al. 2015). Using universal primers designed to amplify conserved genomic regions across a broad diversity of taxonomic groups of interest, researchers are afforded the opportunity to survey biological communities at previously unprecedented scales. While such advancements hold great promise for improving our knowledge of the biological world, they also represent new challenges to the scientific community.

Given that bioinformatic methods for taxonomic inference of metabarcoding sequence data are relatively new, the development, validation and refinement of appropriate analytical methods are ongoing. Relatively few studies have characterized the strengths and weaknesses of different bioinformatic sequence classification protocols (Porter et al. 2012; Bengtsson-Palme et al. 2015; Peabody et al. 2015; Somervuo et al. 2016; Richardson et al. 2017). Further, researchers continue to utilize a diversity of methods to draw taxonomic inferences from amplicon sequence data. Relative to alignment-based nearest-neighbor and lowest common ancestor-type classification approaches, methods involving hierarchical classification of DNA sequences are popular as they are often designed to estimate the probabilistic confidence of taxonomic inferences at each taxonomic rank. However, studies explicitly examining the accuracy of classification confidence estimates are rare (Somervuo et al. 2016).

When performing hierarchical classification, the construction, curation and uniform taxonomic annotation of the reference sequence database is an important methodological consideration. Database quality can affect classification performance in numerous ways. For example, artifacts within the taxonomic identifiers of a reference database can represent artificial diversity and the inclusion of sequence data adjacent to the exact barcoding locus of interest likely display sequence composition that is unrepresentative of the barcoding locus. Lastly, sequence redundancy within reference databases increases computational resource use and is particularly problematic for classification software programs that classify sequences based on a set number of alignments. In general, such database artifacts have the potential to bias model selection and confidence estimation both with k-mer style classifiers such as UTAX, SINTAX and the RDP Naïve Bayesian Classifier (Wang et al. 2007; Edgar 2015; Edgar 2016) and alignment based classification approaches such as Metaxa2 and Megan (Huson et al. 2011; Bengtsson-Palme et al. 2015). Thus, it is important to identify and manage reference sequence database artifacts during curation for optimal downstream classification performance.

The use of molecular barcoding and metabarcoding in arthropod community assessment and gut content analysis has gained popularity in recent years (Corse et al. 2010; Yu et al. 2012; Mollot et al. 2014; Elbrecht and Leese 2017). However, as with other non-microbial taxonomic groups of interest, few researchers have developed hierarchical DNA sequence classification techniques for arthropods (Porter et al. 2014; Tang et al. 2015; Somervuo et al. 2017). Here, we detail the construction, curation and evaluation of a database designed for hierarchical classification of amplicon sequences belonging to a 157 bp COI locus commonly used for arthropod metabarcoding (Zeale et al. 2011). This work will serve as both a resource for those

conducting experiments using arthropod metabarcoding and as a template for future work curating and evaluating hierarchical sequence classification databases.

Methods

Data collection and curation

To produce a comprehensive reference set, all COI annotated sequences from Arthropoda as well as three sister phyla, Mollusca, Onychophora and Tardigrada, between 250 and 2500 bp in length were downloaded from the NCBI Nucleotide repository on October 21st, 2016. To supplement this collection, all arthropod whole mitochondrion genomes were downloaded from NCBI Nucleotide on March 3rd, 2017. For metagenetic analysis, the inclusion of close outgroup sequences is useful for estimating the sequence space boundaries between arthropods and alternate phyla. The Perl script provided in Sickel et al. (2016) was then used along with the NCBI Taxonomy module (Sayers et al. 2013) to retrieve the taxonomic identity of each sequence across each of the major Linnaean ranks, from kingdom to species.

After obtaining the available sequences and rank annotations, we created an intermediate database to obtain extracted barcode amplicons of interest from the reference data using the Metaxa2 database builder tool (v1.0, beta 4; http://microbiology.se/software/metaxa2/). This tool creates the hidden Markov models (HMMs) and BLAST reference databases underpinning the Metaxa2 classification procedures. During extraction, we designated an archetypical reference sequence trimmed to the exact 157 bp barcode amplicon of interest from the COI gene. This section of the arthropod COI gene is the amplicon product of the commonly used primers of Zeale et al. (2011). The reference sequence is used in the database builder tool to define the range of the barcoding region of interest, and the software then trims the input sequences to this region using the Metaxa2 extractor (Bengtsson-Palme et al. 2015). To increase the accuracy of

this process, we split the original input sequences on the basis of length prior to running the database builder for amplicon extraction, creating four files with sequences of 250-500 bp, 501-600 bp, 601-2500bp and whole mitochondrion genomes. Following sequence extraction, the database builder tool aligns trimmed sequences using MAFFT (Katoh and Standley 2013) and from this alignment the conservation of each residue in the sequence is determined. The most conserved regions are selected for building HMMs using the HMMER package (Eddy 2011). Input sequences that cover most of the barcoding region and are taxonomically annotated are used to build a BLAST (Altschul et al. 1997) database for sequence classification. Finally, the sequences in the BLAST database are aligned using MAFFT, and the intra- and inter-taxonomic sequence identities are calculated to derive meaningful sequence identity cutoffs at each taxonomic level. This entire process is described in more detail in the Metaxa2 2.2 manual (http://microbiology.se/software/metaxa2/) and in Bengtsson-Palme et al. (2018).

After extraction, sequences were then curated by removal of duplicate sequences using the Java code provided with the RDP classifier (v2.11; Wang et al. 2007). At this point, we conducted extensive curation of the available lineage data for the reference sequence database. For references lacking complete annotation at midpoints within the Linnaean lineage, we used Perl regular expression-based substitution to complete the annotation according to established taxonomic authorities, including MilliBase (Sierwald 2017), the Integrated Taxonomic Information System (http://www.itis.gov) and the phylogenomic analysis of Regier et al. (2010). Table 1 shows the substitutions made. Further, we removed ranks containing annotations reflective of open nomenclature, such as sp., cf. and Incertae sedis, as well as ranks annotated as 'undef.' Lastly, we removed entries containing more than two consecutive uncalled base pairs.

Upon analyzing the representativeness of this initial database across arthropod classes and insect orders, we found that amplicon sequences from two insect orders, Strepsiptera and Embioptera, were not present in the curated database, likely due to their poor sequence similarity to the reference sequence used to designate the amplicon barcode region of interest. To add Strepsiptera and Embioptera COI amplicons, all NCBI COI sequences belonging to these orders were downloaded on October 10th, 2017, curated and added to the Metaxa2 COI database. To improve recovery of amplicons from these insect orders during curation, a representative sequence from both Embioptera and Strepsiptera, representing the 157 bp COI amplicon of interest, was used when building the Metaxa2 database. This retrospective addition of sequences belonging to Strepsiptera and Embioptera contributed 102 and 3 non-redundant reference sequences to the database, respectively.

To assess the degree to which our amplicon sequence extraction, dereplication and curation procedures worked, we took inventory of the number of sequences per species and genera present in the data at three points during curation: 1) in the initial input data, 2) following Metaxa2 database builder-based amplicon sequence extraction and 3) in the final database following dereplication and taxonomic curation.

Classifier performance evaluation

For performance evaluations, the methods used were highly similar to those of Richardson et al. (2017). With the final set of curated amplicon reference sequences, we randomly sampled 10 percent of the sequences to obtain testing data, using the remaining 90 percent of sequences to train the Metaxa2 classifier for performance evaluations. To assess the effect of sequence length on classifier performance, we used a Python script to crop the test case sequences to 80 bp in length, approximately half the median length of the original reference

sequence dataset. Evaluating classification performance on these short sequences provides a test of the classifiers robustness to sequence length variation and enables estimation of the potential for classifying sequences from short, high-throughput technology, such as 100 cycle single-end Illumina HiSeq sequencing. We then performed the following analyses on both the full-length (157 bp) and half-length (80 bp) test case sequences, separately.

To characterize the relationship between the Metaxa2 reliability score, an estimate of classification confidence, and the probability of classification error, we used the COI trained classifier to classify the testing datasets, requiring the software to classify to the family rank regardless of the reliability score of the assignment. After comparing the known taxonomic identity of each reference test case to the Metaxa2 predicted taxonomic identity, we regressed 5,000 randomly chosen binary classification outcomes, '1' representing an incorrect classification and '0' representing a correct classification, against the Metaxa2 reliability score using local polynomial logistic regression in R (v3.3.1; R Core Team 2014).

To assess classifier sensitivity and accuracy on all testing data, we analyzed both testing datasets using Metaxa2 with an e-value threshold (-E) of 1e-25 and a reliability score threshold (-R) of 75. With the resulting classifications, we compared the known taxonomic identity of each reference test case to its Metaxa2 classification, from species to order, to assess the proportion of sequences classified and the proportion of incorrect taxonomic assignments as well as the false discovery rate, or errors per assignment. Lastly, to assess the rate of taxonomic overclassification at the genus level, we searched the testing dataset for sequence cases belonging to arthropod genera not represented in the training database. We then determined the proportion of these sequences classified to the genus level, or overclassified. Further, we assessed family level error and sensitivity for these test cases.

Results

Following curation and extraction, we obtained 199,206 reference amplicon sequences belonging to 51,416 arthropod species. Over 90 percent of the references were between 142 and 149 bp in length, with a minimum reference sequence length of 94 bp. For the final database creation and classifier training procedure, many reference amplicons were shorter than the 157 bp region of interest due to the incompleteness of some reference sequences and the trimming of taxonomically uninformative ends during Metaxa2 training. Prior to this step, 82 percent of the sequences were between 150 and 157 bp in length following the original extraction and these longer sequences can be found at https://github.com/RTRichar/Zeale_COI_Database. The taxonomic representativeness of the database across different arthropod classes and insect orders, including the number of families, genera and species in each, are presented in Tables 2 and 3.

Analyzing the number of sequences per species in the input reference sequence data, we observed a heavily right-skewed distribution, with a median of 2 and a mean of 11.1 sequences per species (Figure 1A). Further, 32.0 percent of species were represented by 5 or more sequences and 40 species, including *Bemisia tabaci* and *Delia platura*, were represented by between 1,000 and 9,736 entries. After conducting amplicon sequence extraction using the Metaxa2 database builder tool, we were able to extract the COI region of interest from 80.8 percent of the input sequences, which corresponded to 73.4 percent of the genus-level diversity found in the original input data. Following sequence dereplication, removal of sequences with three or more ambiguous base calls and taxonomic lineage curation, our final database contained approximately 13 percent of the input extracted sequences, which represented 98.2 percent of the genus-level richness of the input extracted reference amplicon sequences (Figure 1B and 1C)

Regressing classification outcome against the Metaxa2 reliability score yielded a similar best fit model for both the 80 bp and full length test sequence datasets (Figure 2). For both regressions, the probability of sequence mis-assignment was below 10 percent for reliability scores above 70. For the remainder of our evaluations, we chose a reliability score of 75, which corresponded to family-level error probabilities of approximately 5.4 percent and 3.8 percent for 80 bp and full length sequences, respectively.

In evaluating the performance of our classification database when used with Metaxa2, we found generally low rates of proportional error and false discovery and variable sensitivity, though these estimates varied by taxonomic rank (Figure 2A). Interestingly, Metaxa2 displayed low variance in classification sensitivity when classifying 80 bp sequences relative to the full length sequences of 147 bp in median length. Overall, sensitivity was greatest at the order level, the lowest resolution rank tested, wherein 90 and 96 percent of sequences were assigned for 80 bp and full length sequences, respectively. Sensitivity was decreased at higher resolution ranks with only 20 to 22 percent of sequences being classified to species. Conversely, proportional error varied more strongly by sequence length and was greatest at lower taxonomic levels. At the species level, 1.97 percent of 80 bp sequences were misclassified, compared to only 1.13 percent for full length sequences. At the order level, the percent of sequences misclassified was 0.59 and 0.65 for 80 bp and full length sequences, respectively. Considering both the number of sequences assigned and mis-assigned, the classification false discovery rate was similarly highest at the species level, with 8.82 and 5.66 percent of assignments being incorrect for 80 bp and full length sequences, respectively. False discovery rates decreased to 0.66 and 0.68 percent of assignments being incorrect at the order level for 80 bp and full length sequences, respectively.

During our evaluation of taxonomic overclassification, we found 650 sequence test cases belonging to genera not represented in the reference training database. Of these cases, Metaxa2 overclassified 11.5 and 9.7 percent of 80 bp and full length sequences, respectively. At the family level, Metaxa2 misclassified 6.15 and 7.08 percent of 80 bp and full length test sequences, respectively. Lastly, family level sensitivity for these test cases was 19.4 and 28.6 percent for 80 bp and full length sequences, respectively (Figure 2B).

Discussion

While species-specific PCR and immunohistochemistry-based methods have been useful in documenting arthropod food webs (Stuart and Greenstone, 1990; Symondson 2002; Weber et al. 2006; Blubaugh et al. 2016), the narrow species-by-species nature of such approaches has limited their utility for answering large-scale or open-ended ecological questions. With the increasing availability of high-throughput sequencing, arthropod metabarcoding will continue to become more broadly applicable to scientific questions spanning a diversity of research areas. The development of improved methods for drawing maximal inferences from sequence data is an important area for further methodological development. In creating a highly curated COI reference amplicon sequence database and evaluating its performance when used with the Metaxa2 taxonomic classifier, we have developed a new method to aid researchers in the analysis of arthropod metabarcoding data.

Though predictions vary greatly, researchers have estimated the species richness of arthropods to be between 2.5 to 3.7 million (Hamilton et al. 2010). Further, according to the literature review of Porter et al. (2014), 72,618 insect genera have been described to date. Thus the 51,416 species and 17,039 genera represented in our database account for only a small fraction of arthropod biodiversity. The limited representativeness of currently available, high

quality reference sequence amplicons for the COI region highlights the need for continued efforts to catalogue arthropod biodiversity with molecular techniques. Despite this current limitation, the combination of molecular gut content analysis with high-throughput sequencing is a promising path toward investigating arthropod trophic ecology and biodiversity monitoring with greater sensitivity and accuracy relative to alternate approaches.

The results of our inventory of sequences per species and genus-level richness at various stages in the database curation process revealed that our amplicon extraction procedure was highly sensitive, extracting and trimming approximately 81 percent of the input sequences down to the 157 bp region of interest. Further, approximately 87 percent of these extracted sequences represented sequence redundancies and were removed during dereplication. As mentioned previously, the trimming of sequence residues adjacent to the barcode of interest and removal of redundant sequences not only makes computational analysis less resource intensive, it can also improve classification performance. For k-mer style classifiers, extraneous sequence residues can bias model selection during classifier training, while abundant sequence duplicates can result in an overwhelming number of identical top hit alignments for alignment-based classifiers. Overall, the best fit local regression models summarizing the relationship between the Metaxa2 reliability score and the probability of classification error were useful in that the probability of misclassification was always less than what would be expected based on the reliability score. For example, a reliability score of 90 corresponded to only a 3.7 percent chance of family-level misclassification for full length sequences. We selected a reliability score of 75 for subsequent analysis as this provided a balanced trade-off between sensitivity and accuracy. Using this reliability score we observed minimal false positive rates and overall proportions of misclassification when comparing our results to those of similar studies (Porter et al. 2014;

Bengtsson-Palme et al. 2015; Edgar 2016; Richardson et al. 2017). Given that the family-level probability of error was only 8.6 percent at a reliability score of 68, a lower reliability score threshold may be justifiable for certain research situations. However, further testing should be conducted to ensure that the relationship between reliability score and classification confidence is similar across taxonomic ranks and between different DNA barcoding loci.

With respect to sensitivity using a reliability score of 75, our results were highly dependent upon the rank being analyzed, with sensitivities above 60 percent only being achieved at the family and order ranks. These sensitivity estimates likely reflect a large degree of database incompleteness at the genus and species ranks. However, to our knowledge, no other studies have reported classification sensitivity data for this COI amplicon locus, making it difficult to ascertain if Metaxa2 is exhibiting low sensitivity or if this locus is limited in discriminatory power. The short length of the locus relative to other barcoding regions such as the 18S rRNA gene and the ITS regions (Hugerth et al. 2014; Wang et al. 2015) could be a cause of such limited discriminatory power.

As expected, analyzing cases of overclassification in our data revealed that sequences from taxa lacking representation in the database are far more likely to be misclassified relative to sequences from well-represented taxa. This is supported by an approximately 10 percent probability of genus level misclassification for sequences from unrepresented genera relative to a 1 to 2 percent probability, depending on sequence length, for all sequence test cases. While this level of overclassification is not desirable, it is considerably lower than genus level overclassification estimates for the RDP classifier, which range from 21.3 percent to 67.8 percent depending on the database, locus analyzed and cross-validation approach used (Edgar 2016; Richardson et al. 2017). Further, the observed degree of genus level overclassification using

Metaxa2 with our COI database was similar to or less than that of the recently developed SINTAX classifier (Edgar 2016). However, such comparisons should be approached with caution as multiple factors can affect classification performance, such as locus discriminatory power and database completeness. Ultimately, direct comparisons of classification methods using standardized loci and databases are needed to more rigorously compare performance.

With respect to Metaxa2 classification of full-length relative to half-length amplicon sequences, we observed surprisingly small differences in performance. In general, the proportion of misclassified sequences was greater for half-length sequences. However, this was not the case for all ranks. At the order level, a greater proportion of sequences were misclassified, likely an artifact of the greater proportion of sequences assigned to the order level. When considering error and sensitivity together, the false discovery rate for full-length sequences was consistently less than or equal to that achieved during the classification of half-length sequences. Further, with respect to the classification of test sequence cases belonging to genera not represented in the reference training database, proportional genus-level overclassification and family-level misclassification were similar for both full-length and half-length sequences, while family-level sensitivity was considerably greater for full-length sequences, 28.6 percent, relative to half-length sequences, 19.4 percent. Lastly, when considering the relationship between the Metaxa2 reliability score and the probability of classification error at the family-level, we noted highly similar local polynomial regression models of error probability for both full-length and half-length sequences.

In summary, we assembled a highly curated database of arthropod COI reference amplicon sequences, trained a recently developed hierarchical DNA sequence classifier using the database and conducted extensive performance evaluations on the resulting classification

pipeline. The limited representativeness of the database with respect to arthropod biodiversity indicates that additional sequencing effort is needed to further improve the performance of arthropod metabarcoding techniques. To evaluate classification performance, we characterized classification sensitivity, false discovery rate and classification confidence. Based on this evaluation, researchers will be better prepared to gauge the strengths and limitations of different approaches to arthropod metabarcoding. Further, the COI database produced as a part of this work will be useful for researchers searching for improved methods for arthropod COI sequence classification.

Acknowledgments

This work was supported by an allocation of computing time from the Ohio Supercomputer Center.

Funding

This work was supported by a Project *Apis m*. - Costco Honey Bee Biology Fellowship to RTR, state and federal appropriations to the Ohio Agricultural Research and Development Center (USDA-NIFA Projects OHO01277 and OHO01355-MRF). JBP was supported by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS; grant 2016-00768).

Grant Disclosures

- FORMAS grant 2016-00768
- USDA-NIFA Projects OHO01277
- USDA-NIFA Projects OHO01355-MRF

Author contributions

- RTR, RMJ and MMG conceived the work

- JBP adapted the database building software
- RTR created and evaluated the database and wrote the manuscript
- RTR, RMJ, JBP and MMG edited the manuscript

Data Availability

The database associated with this work can be downloaded from the associated GitHub repository: https://github.com/RTRichar/Zeale_COI_Database

Supplemental Information

The bioinformatics workflow and code associated with this work is disclosed at the associated GitHub repository: https://github.com/RTRichar/Zeale_COI_Database

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389-3402.

Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs. BMC Bioinformatics, 13, 1-13.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH (2015) Metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Molecular Ecology Resources, 15, 1403-1414.

Bengtsson-Palme J, Richardson RT, Meola M, Wurzbacher C, Tremblay ED, Thorell K, Kanger K, Ericksson M, Bilodeau GJ, Johnson RM, Hartmann M, Nilsson H (2018) Taxonomic identification from metagenomic and metabarcoding data using any genetic marker. bioRxiv, 253377, doi: https://doi.org/10.1101/253377

354 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. Nucleic
355 Acids Research, 33, D34-D38.
356 Blubaugh CK, Hagler JR, Machtley SA, Kaplan I (2016) Cover crops increase foraging activity
357 of omnivorous predators in seed patches and facilitate weed biological control.
358 Agriculture, Ecosystems and Environment, 231, 264-270.
359 Corse E, Costedoat C, Chappaz R, Pech N, Martin J-F, Gilles A (2010) A PCR-based method for
360 diet analysis in freshwater organisms using 18S rDNA barcoding on faeces. Molecular
361 Ecology Resources, 10, 96-108.
362 Eddy SR (2011) Accelerated profile HMM searches. PLoS Computational Biology, 7, e1002195.
363 Edgar RC (2015) UTAX algorithm. Available at:
364 http://www.drive5.com/usearch/manual/utax_algo.html (accessed 6 January 2016).
365 Edgar RC (2016) SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS
366 sequences. Biorxiv. https://doi.org/10.1101/074161.
367 Elbrecht V and Leese F (2017) Validation and development of COI metabarcoding primers for
368 freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science, 5, 11.
369 Guardiola M, Uriz MJ, Taberlet P, Coissac E, Wangensteen OW, Turon X (2015) Deep-sea,
370 deep-sequencing: Metabarcoding extracellular DNA from sediments of marine canyons.
371 PloS One, 10, e0139633.
372 Hamilton AJ, Basset Y, Benke KK, Grimbacher PS, Miller SE, Novotny V, Samuelson A, Stork
373 NE, Weiblen GD, Yen JDL (2010) Quantifying uncertainty in estimation of tropical
374 arthropod species richness. The American Naturalist, 175, 90-95.

375 Hugerth LW, Muller EEL, Hu YOO, Lebrun LAM, Roume H, Lundin D, Wilmes P, Andersson
376 AF (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic
377 diversity in microbial consortia. PloS ONE, 9, e95567.
378 Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of
379 environmental sequences using MEGAN 4. Genome Research, 21, 1552-1560.
380 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
381 improvements in performance and usability. Molecular Biology and Evolution, 30, 772-
382 780.
383 Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M (2009) 5.8S-28S rRNA
384 interaction and HMM-based ITS2 annotation. Gene, 430, 50-57.
385 Mollot G, Duyck P-F, Lefeuvre P, Lescourret F, Martin J-F, Piry S, et al. (2014) Cover cropping
386 alters the diet of arthropods in a banana plantation: A metabarcoding approach. PloS
387 ONE, 9, e93740.
388 Ohio Supercomputer Center (1987) Citation. Columbus, Ohio, USA.
389 http://osc.edu/ark:/19495/f5s1ph73.
390 Peabody MA, Van Rossum T, Lo R, Brinkman FSL (2015) Evaluation of shotgun metagenomics
391 sequence classification methods using in silico and in vitro simulated communities. BMC
392 Bioinformatics, 16, 363.
393 Porter TM, Golding GB (2012) Factors that affect large subunit ribosomal DNA amplicon
394 sequencing studies of fungal communities: Classification method, primer choice, and
395 error. PloS One, 7, e35749.
396 Porter TM, Gibson JF, Shokralla S, Baird DJ, Golding GB, Hajibabaei M (2014) Rapid and
397 accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1

(COI) DNA barcode sequences using a naive Bayesian classifier. Molecular Ecology Resources 14, 929-942.

R Core Team (2014) R: A language and environment for statistical computing. Vienna, Austria. http://www.R-project.org/

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature, 463, 1079-1083.

Richardson RT, Bengtsson-Palme J, Johnson RM (2017) Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources 17, 760-769.

Richardson RT, Lin C-H, Quijia JQ, Riusech NS, Goodell K, Johnson RM (2015a) Rank-based characterization of pollen assemblages collected by honey bees using a multi-locus metabarcoding approach. Applications in Plant Sciences, 3, 1500043.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K et al (2013) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 41, D8-D20.

Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, Steffan-Dewenter I, Keller A (2015) Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. BMC Ecology, 15, 1-9.

Sierwald P (2017) MilliBase. Accessed at http://www.millibase.org on 2017-04-12

Somervuo P, Koskela S, Pennanen J, Nilsson RH, Ovaskainen O (2016) Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics, 32, 2920-2927.

420 Somervuo P, Yu DW, Xu CCY, Ji Y, Hultman J, Wirta H, Ovaskainen O (2017) Quantifying
421 uncertainty of taxonomic placement in DNA barcoding and metabarcoding. Methods in
422 Ecology and Evolution, 8, 398-407.
423 Stuart MK, Greenstone MH (1990) Beyond ELISA: A rapid, sensitive, specific immunodot assay
424 for identification of predator stomach contents. Annals of the Entomological Society of
425 America, 83, 1101-1107.
426 Symondson WOC (2002) Molecular identification of prey in predator diets. Molecular Ecology,
427 11, 627-641.
428 Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce
429 C, Nevard T, Potts SG, Zhou X, Yu DW (2015) High-throughput monitoring of wild bee
430 diversity and abundance via mitogenomics. Methods in Ecology and Evolution, 6, 1034-
431 1043.
432 Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment
433 of rRNA sequences into the new bacterial taxonomy. Applied and Environmental
434 Microbiology, 73, 5261-5267.
435 Wang X-C, Liu C, Huang L, Bengtsson-Palme J, Chen H, Zhang J-H, Cai D, Li J-Q (2015)
436 ITS1: A DNA barcode better than ITS2 in eukaryotes? Molecular Ecology Resources, 15,
437 573-586.
438 Weber DC, Rowley DL, Greenstone MH, Athanas MM (2006) Prey preference and host
439 suitability of the predatory and parasitoid carabid beetle, *Lebia grandis*, for several
440 species of *Leptinotarsa* beetles. Journal of Insect Science, 6, 1-14.

441 Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup:
442 Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring.
443 Methods in Ecology and Evolution, 3, 613-623.
444 Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G (2011) Taxon-specific PCR for DNA
445 barcoding arthropod prey in bat faeces. Molecular Ecology Resources, 11, 236-244.

446 **Tables**

447 Table 1: Summary of taxonomic annotations made for references which had undefined ranks at
448 midpoints in their respective taxonomic lineages

Undefined Rank	Higher Resolution Assignment	Assignment Made	Authority Used
Order	Family Sphaerotheriidae	Order Sphaerotheriida	MilliBase
Order	Family Zephroniidae	Order Sphaerotheriida	
Order	Family Lepidotrichidae	Order Zygentoma	ITIS
Order	Family Lepismatidae	Order Zygentoma	
Order	Family Nicoletiidae	Order Zygentoma	
Class	Order Pauropoda	Class Myriapoda	
Genus	Genus Pseudocellus	Family Ricinididae	
Genus	Genus Chanbria	Family Eremobatidae	
Species	Species Tanypodinae spp.	Genus Tanypodinae	
Species	Species Ennominae spp.	Genus Ennominae	
Genus	Genus Dichelesthiidae	Family Dichelesthiidae	
Genus	Genus Phallocryptus	Family Thamnocephalidae	
Class	Order Symphyla	Class Myriapoda	Regier *et al.* 2010
Order	Family Peripatidae	Order Onychophora	
Class	Family Peripatidae	Class Onychophora	
Order	Family Peripatopsidae	Order Onychophora	
Class	Family Peripatopsidae	Class Onychophora	
Genus	Genus Lasionectes	Family Speleonectidae	WoRMS
Family	Family Speleonectidae	Order Nectiopoda	
Genus	Genus Prionodiaptomus	Family Diaptomidae	
Family	Family Diaptomidae	Order Calanoida	
Order	Order Calanoida	Class Maxillopoda	

449 Table 2: Summary of taxonomic representation across all arthropod classes and associated sister
450 groups. Numbers may include sub and super groupings.

Class	Number of Orders	Number of Families	Number of Genera	Number of Species
Heterotardigrada	1	2	2	1
Eutardigrada	1	3	12	20
Onychophora	1	2	17	42
Pycnogonida	1	10	27	89
Cephalopoda	1	1	1	1
Merostomata	1	1	3	4
Arachnida	17	226	740	1804
Myriapoda	2	4	7	9
Chilopoda	5	16	53	172
Diplopoda	11	33	95	181
Ostracoda	2	6	19	40
Branchiopoda	3	25	76	254
Malacostraca	13	256	969	2654
Maxillopoda	11	85	240	568
Cephalocarida	1	1	2	1
Remipedia	1	2	5	8
Protura	1	4	12	13
Diplura	1	5	7	11
Collembola	4	18	98	203
Insecta	29	789	14654	45341
Total	**107**	**1489**	**17039**	**51416**

451 Table 3: Summary of insect taxa included in the arthropod COI database following curation.

452 Numbers may include sub and super groupings.

Order	Number of Families	Number of Genera	Number of Species
Archaeognatha	2	15	18
Zygentoma	3	3	3
Odonata	34	243	488
Ephemeroptera	26	108	378
Zoraptera	1	1	1
Dermaptera	4	6	7
Plecoptera	15	84	199
Orthoptera	30	299	603
Mantophasmatodea	1	1	1
Grylloblattodea	1	1	0
Embioptera	1	2	2
Phasmatodea	6	20	28
Mantodea	11	74	76
Blattodea	8	82	95

Isoptera	6	91	186
Thysanoptera	3	14	30
Hemiptera	101	1245	2730
Psocoptera	12	17	19
Hymenoptera	74	1500	4418
Raphidioptera	2	9	11
Megaloptera	2	10	22
Neuroptera	14	65	153
Strepsiptera	3	10	36
Coleoptera	124	2287	7452
Trichoptera	43	320	1269
Lepidoptera	134	6554	21626
Siphonaptera	6	14	20
Mecoptera	4	5	10
Diptera	118	1574	5460
Total	**789**	**14654**	**45341**

Figures

Figure 1: A percent density histogram of the number of sequences per species (A) shows the distribution of redundancy within the NCBI Nucleotide entries used. The dashed blue line and solid red line indicate the median and mean number of sequences per species, respectively. Inventories of the number of sequences (A) and genera (B) input into the curation process, following Metaxa2 extraction and following dereplication of redundant sequences and curation of taxonomic lineages.

Figure 2: A logistic regression analysis of case-by-case classification accuracy, '1' indicating a false-positive identification and '0' indicating a true-positive identification, regressed against classification reliability score. A best fit local polynomial regression line was used to estimate the relationship between reliability score and the probability of mis-classification.

Figure 3: Proportional accuracy, sensitivity and false discovery rate for classification of all test reference sequences, conducted on both the full-length and half-length sequences (A). Proportional genus-level overclassification rate, family-level error and family-level sensitivity for test cases belonging to genera not represented in the Metaxa2 COI reference training data (B).

Figure 1(on next page)

Inventory of database size and diversity at various stages in the curation process

A percent density histogram of the number of sequences per species (A) shows the distribution of redundancy within the NCBI Nucleotide entries used. The dashed blue line and solid red line indicate the median and mean number of sequences per species, respectively. Inventories of the number of sequences (A) and genera (B) input into the curation process, following Metaxa2 extraction and following dereplication of redundant sequences and curation of taxonomic lineages.

Figure 1

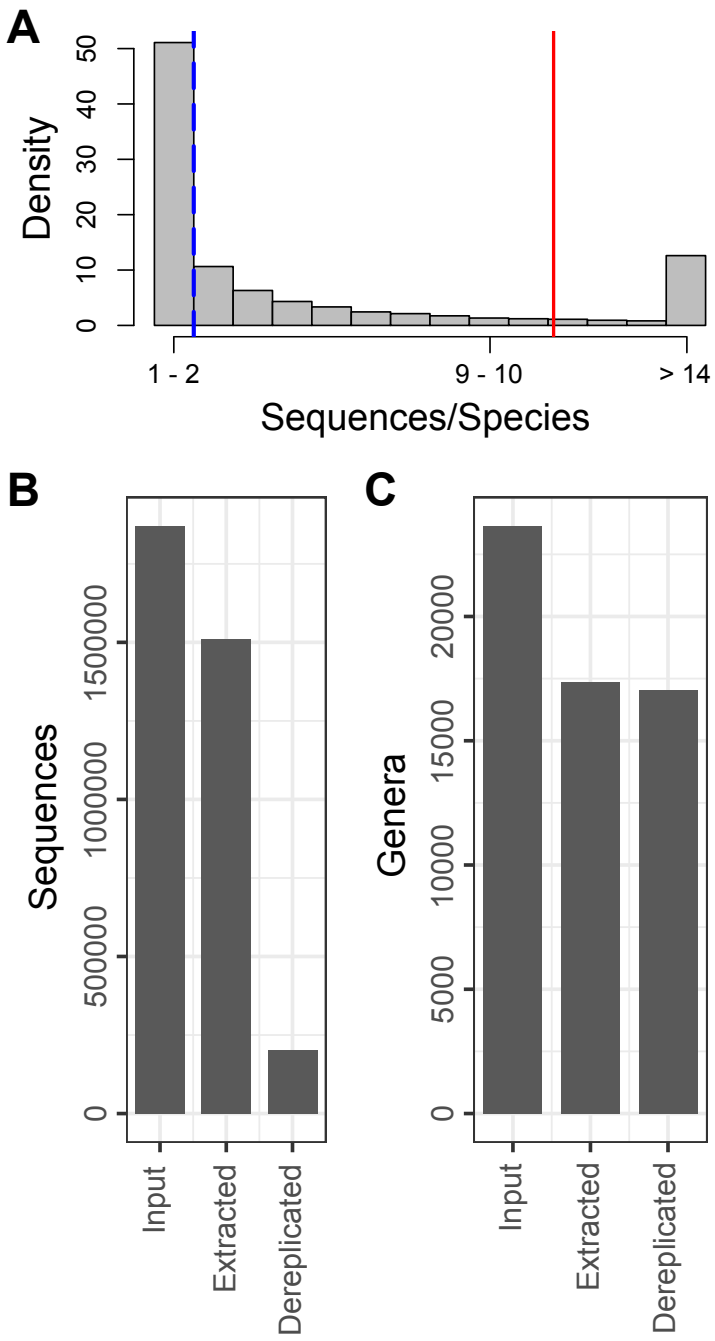

Figure 2(on next page)

Relationship between Metaxa2 reliability score and classification error probability

A logistic regression analysis of case-by-case classification accuracy, ‘1’ indicating a false-positive identification and ‘0’ indicating a true-positive identification, regressed against classification reliability score. A best fit local polynomial regression line was used to estimate the relationship between reliability score and the probability of mis-classification.

Figure 2

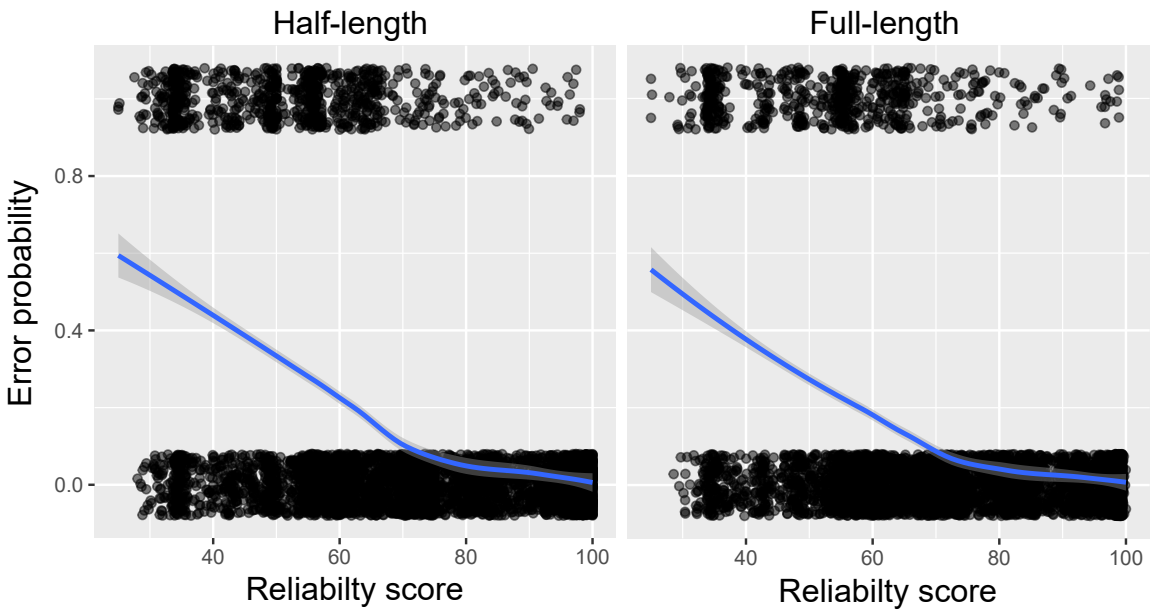

Figure 3(on next page)

Evaluation of classification performance

Proportional accuracy, sensitivity and false discovery rate for classification of all test reference sequences, conducted on both the full-length and half-length sequences (A). Proportional genus-level overclassification rate, family-level error and family-level sensitivity for test cases belonging to genera not represented in the Metaxa2 COI reference training data (B).

Figure 3