

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

[View the peer-reviewed version](https://peerj.com/articles/5126) (peerj.com/articles/5126), which is the preferred citable publication unless you specifically need to cite this preprint.

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>

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

3 Rodney T. Richardson^{1*}, Johan Bengtsson-Palme^{2,3}, Mary M. Gardiner⁴, Reed M. Johnson¹

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

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

8 ³ Center for Antibiotic Resistance research (CARE) at University of Gothenburg, Box 440, SE-
9 40530, Gothenburg, Sweden

10 ⁴ Department of Entomology, The Ohio State University, 2021 Coffey Road, Columbus, Ohio
11 43210 USA.

12 Corresponding Author:

13 Rodney Richardson

14 Email address: richardson.827@osu.edu

15 Abstract

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

38 Introduction

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

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

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

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

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

84 **Methods**

85 *Data collection and curation*

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

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

105 this process, we split the original input sequences on the basis of length prior to running the
106 database builder for amplicon extraction, creating four files with sequences of 250-500 bp, 501-
107 600 bp, 601-2500bp and whole mitochondrion genomes. Following sequence extraction, the
108 database builder tool aligns trimmed sequences using MAFFT (Kato and Standley 2013) and
109 from this alignment the conservation of each residue in the sequence is determined. The most
110 conserved regions are selected for building HMMs using the HMMER package (Eddy 2011).
111 Input sequences that cover most of the barcoding region and are taxonomically annotated are
112 used to build a BLAST (Altschul et al. 1997) database for sequence classification. Finally, the
113 sequences in the BLAST database are aligned using MAFFT, and the intra- and inter-taxonomic
114 sequence identities are calculated to derive meaningful sequence identity cutoffs at each
115 taxonomic level. This entire process is described in more detail in the Metaxa2 2.2 manual
116 (<http://microbiology.se/software/metaxa2/>) and in Bengtsson-Palme et al. (2018).

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

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

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

143 *Classifier performance evaluation*

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

150 sequence dataset. Evaluating classification performance on these short sequences provides a test
151 of the classifiers robustness to sequence length variation and enables estimation of the potential
152 for classifying sequences from short, high-throughput technology, such as 100 cycle single-end
153 Illumina HiSeq sequencing. We then performed the following analyses on both the full-length
154 (157 bp) and half-length (80 bp) test case sequences, separately.

155 To characterize the relationship between the Metaxa2 reliability score, an estimate of
156 classification confidence, and the probability of classification error, we used the COI trained
157 classifier to classify the testing datasets, requiring the software to classify to the family rank
158 regardless of the reliability score of the assignment. After comparing the known taxonomic
159 identity of each reference test case to the Metaxa2 predicted taxonomic identity, we regressed
160 5,000 randomly chosen binary classification outcomes, ‘1’ representing an incorrect
161 classification and ‘0’ representing a correct classification, against the Metaxa2 reliability score
162 using local polynomial logistic regression in R (v3.3.1; R Core Team 2014).

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

173 Results

174 Following curation and extraction, we obtained 199,206 reference amplicon sequences
175 belonging to 51,416 arthropod species. Over 90 percent of the references were between 142 and
176 149 bp in length, with a minimum reference sequence length of 94 bp. For the final database
177 creation and classifier training procedure, many reference amplicons were shorter than the 157
178 bp region of interest due to the incompleteness of some reference sequences and the trimming of
179 taxonomically uninformative ends during Metaxa2 training. Prior to this step, 82 percent of the
180 sequences were between 150 and 157 bp in length following the original extraction and these
181 longer sequences can be found at https://github.com/RTRichar/Zeale_COI_Database. The
182 taxonomic representativeness of the database across different arthropod classes and insect orders,
183 including the number of families, genera and species in each, are presented in Tables 2 and 3.

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

195 Regressing classification outcome against the Metaxa2 reliability score yielded a similar
196 best fit model for both the 80 bp and full length test sequence datasets (Figure 2). For both
197 regressions, the probability of sequence mis-assignment was below 10 percent for reliability
198 scores above 70. For the remainder of our evaluations, we chose a reliability score of 75, which
199 corresponded to family-level error probabilities of approximately 5.4 percent and 3.8 percent for
200 80 bp and full length sequences, respectively.

201 In evaluating the performance of our classification database when used with Metaxa2, we
202 found generally low rates of proportional error and false discovery and variable sensitivity,
203 though these estimates varied by taxonomic rank (Figure 2A). Interestingly, Metaxa2 displayed
204 low variance in classification sensitivity when classifying 80 bp sequences relative to the full
205 length sequences of 147 bp in median length. Overall, sensitivity was greatest at the order level,
206 the lowest resolution rank tested, wherein 90 and 96 percent of sequences were assigned for 80
207 bp and full length sequences, respectively. Sensitivity was decreased at higher resolution ranks
208 with only 20 to 22 percent of sequences being classified to species. Conversely, proportional
209 error varied more strongly by sequence length and was greatest at lower taxonomic levels. At the
210 species level, 1.97 percent of 80 bp sequences were misclassified, compared to only 1.13 percent
211 for full length sequences. At the order level, the percent of sequences misclassified was 0.59 and
212 0.65 for 80 bp and full length sequences, respectively. Considering both the number of sequences
213 assigned and mis-assigned, the classification false discovery rate was similarly highest at the
214 species level, with 8.82 and 5.66 percent of assignments being incorrect for 80 bp and full length
215 sequences, respectively. False discovery rates decreased to 0.66 and 0.68 percent of assignments
216 being incorrect at the order level for 80 bp and full length sequences, respectively.

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

223 **Discussion**

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

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

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

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

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

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

277 As expected, analyzing cases of overclassification in our data revealed that sequences
278 from taxa lacking representation in the database are far more likely to be misclassified relative to
279 sequences from well-represented taxa. This is supported by an approximately 10 percent
280 probability of genus level misclassification for sequences from unrepresented genera relative to a
281 1 to 2 percent probability, depending on sequence length, for all sequence test cases. While this
282 level of overclassification is not desirable, it is considerably lower than genus level
283 overclassification estimates for the RDP classifier, which range from 21.3 percent to 67.8 percent
284 depending on the database, locus analyzed and cross-validation approach used (Edgar 2016;
285 Richardson et al. 2017). Further, the observed degree of genus level overclassification using

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

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

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

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

317 **Acknowledgments**

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

320 **Funding**

321 This work was supported by a Project *Apis m.* - Costco Honey Bee Biology Fellowship to RTR,
322 state and federal appropriations to the Ohio Agricultural Research and Development Center
323 (USDA-NIFA Projects OHO01277 and OHO01355-MRF). JBP was supported by the Swedish
324 Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS; grant
325 2016-00768).

326 **Grant Disclosures**

- 327 • FORMAS grant 2016-00768
- 328 • USDA-NIFA Projects OHO01277
- 329 • USDA-NIFA Projects OHO01355-MRF

330 **Author contributions**

- 331 • RTR, RMJ and MMG conceived the work

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

335 **Data Availability**

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

338 **Supplemental Information**

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

341 **References**

- 342 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped
343 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic
344 Acids Research*, 25, 3389-3402.
- 345 Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification
346 programs. *BMC Bioinformatics*, 13, 1-13.
- 347 Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH
348 (2015) Metaxa2: Improved identification and taxonomic classification of small and large
349 subunit rRNA in metagenomic data. *Molecular Ecology Resources*, 15, 1403-1414.
- 350 Bengtsson-Palme J, Richardson RT, Meola M, Wurzbacher C, Tremblay ED, Thorell K, Kanger
351 K, Ericksson M, Bilodeau GJ, Johnson RM, Hartmann M, Nilsson H (2018) Taxonomic
352 identification from metagenomic and metabarcoding data using any genetic marker.
353 *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) UTAH 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

- 398 (COI) DNA barcode sequences using a naive Bayesian classifier. *Molecular Ecology*
399 *Resources* 14, 929-942.
- 400 R Core Team (2014) R: A language and environment for statistical computing. Vienna, Austria.
401 <http://www.R-project.org/>
- 402 Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzler R, Martin JW, Cunningham CW
403 (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-
404 coding sequences. *Nature*, 463, 1079-1083.
- 405 Richardson RT, Bengtsson-Palme J, Johnson RM (2017) Evaluating and optimizing the
406 performance of software commonly used for the taxonomic classification of DNA
407 metabarcoding sequence data. *Molecular Ecology Resources* 17, 760-769.
- 408 Richardson RT, Lin C-H, Quijia JQ, Riusech NS, Goodell K, Johnson RM (2015a) Rank-based
409 characterization of pollen assemblages collected by honey bees using a multi-locus
410 metabarcoding approach. *Applications in Plant Sciences*, 3, 1500043.
- 411 Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K et al (2013) Database
412 resources of the National Center for Biotechnology Information. *Nucleic Acids Research*,
413 41, D8-D20.
- 414 Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, Steffan-Dewenter I,
415 Keller A (2015) Increased efficiency in identifying mixed pollen samples by meta-
416 barcoding with a dual-indexing approach. *BMC Ecology*, 15, 1-9.
- 417 Sierwald P (2017) MilliBase. Accessed at <http://www.millibase.org> on 2017-04-12
- 418 Somervuo P, Koskela S, Pennanen J, Nilsson RH, Ovaskainen O (2016) Unbiased probabilistic
419 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 Dichelesthidae	Family Dichelesthidae	
Genus	Genus Phallocryptus	Family Thamnocephalidae	
Class	Order Symphyla	Class Myriapoda	
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 Prionodiptomus	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

453 **Figures**

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

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

464 Figure 3: Proportional accuracy, sensitivity and false discovery rate for classification of all test
 465 reference sequences, conducted on both the full-length and half-length sequences (A).
 466 Proportional genus-level overclassification rate, family-level error and family-level sensitivity
 467 for test cases belonging to genera not represented in the Metaxa2 COI reference training data
 468 (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

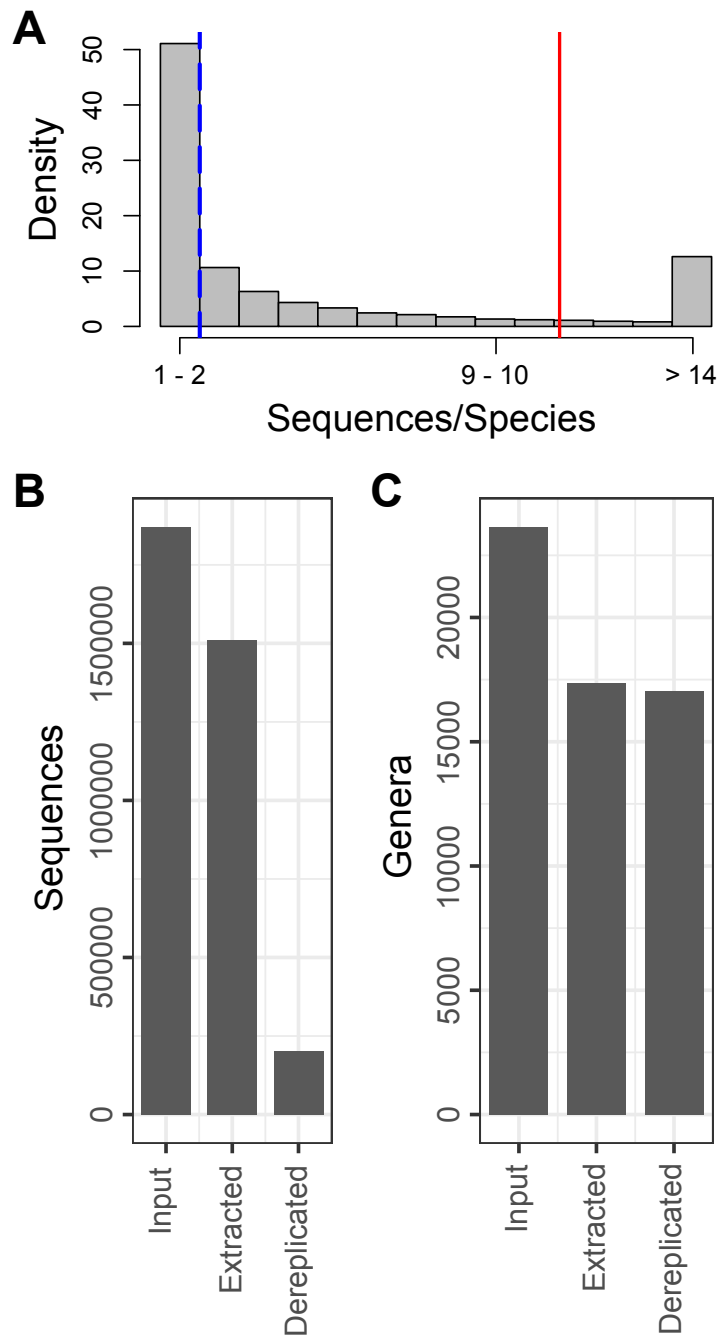


Figure 2(on next page)

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

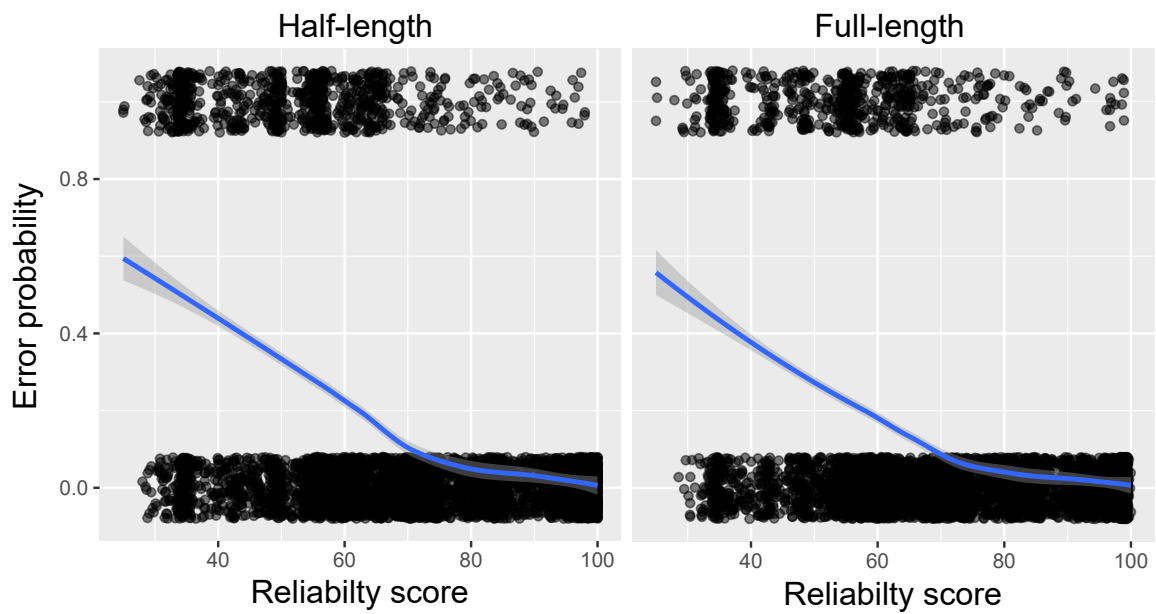


Figure 3(on next page)

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