A reference cytochrome c oxidase subunit I database curated for hierarchical classification of arthropod metabarcoding data

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

[1] Department of Entomology, Ohio State University, Columbus, Ohio, United States

[2] Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[3] Center for Antibiotic Resistance research (CARe), University of Gothenburg, Gothenburg, Sweden

[4] Department of Entomology, Ohio State University, Wooster, Ohio, United States

Corresponding Author: Rodney T Richardson
Email address: richardson.827@osu.edu

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, the latter of 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 performed a cross-validation analysis to characterize 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.

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. Gardiner[4], Reed M. Johnson[1]

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, the latter of which provided the only source of representation for

23 Zoraptera, Callipodida and Holothyrida. The database contains extracted sequences of the target

24 amplicon 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 performed a cross-validation analysis to

32 characterize the relationship between the Metaxa2 reliability score, an estimate of classification

33 confidence, and classification error probability. We used this analysis to select a reliability score

34 threshold which minimized error. We then estimated classification sensitivity, false discovery

35 rate and overclassification, the propensity to classify sequences from taxa not represented in the

36 reference database. Our work will help researchers design and evaluate classification databases

37 and conduct metabarcoding on arthropods and alternate taxa.

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 is 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 top 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 using the

89   search term 'Arthropoda cytochrome oxidase subunit I'. To supplement this collection, all

90   arthropod whole mitochondrion genomes were downloaded from NCBI Nucleotide on March 3rd,

91   2017 using the search term 'Arthropoda mitochondrion genome'. For metagenetic analysis, the

92   inclusion of close outgroup sequences is useful for estimating the sequence space boundaries

93   between arthropods and alternate phyla. The Perl script provided in Sickel et al. (2015) was then

94   used along with the NCBI Taxonomy module (NCBI Resource Coordinators 2018) to retrieve

95   the taxonomic identity of each sequence across each of the major Linnaean ranks, from kingdom

96   to species.

97   After obtaining the available sequences and rank annotations, we created an intermediate

98   database to obtain extracted barcode amplicons of interest from the reference data using the

99   Metaxa2 database builder tool (v1.0, beta 4; http://microbiology.se/software/metaxa2/). This tool

100   creates the hidden Markov models (HMMs) and BLAST reference databases underpinning the

101   Metaxa2 classification procedures. Prior to extraction, we randomly selected a reference

102   sequence, trimmed it to the exact 157 bp barcode amplicon of interest and designated it as the

103   archetypical reference during database building using the '-r' argument. The section of the

104   arthropod COI gene we trimmed this sequence to is the amplicon product of the commonly used

105 primers of Zeale et al. (2011). The reference sequence is used in the database builder tool to

106 define the range of the barcoding region of interest, and the software then trims the remainder of

107 the input sequences to this region using the Metaxa2 extractor (Bengtsson-Palme et al. 2015). To

108 increase the accuracy of multiple sequence alignment during this process, we split the original

109 input sequences on the basis of length prior to running the database builder for amplicon

110 extraction, creating four files with sequences of 250-500 bp, 501-600 bp, 601-2500bp and whole

111 mitochondrion genomes. Following sequence extraction, the database builder tool aligns

112 trimmed sequences using MAFFT (Katoh and Standley 2013) and from this alignment the

113 conservation of each residue in the sequence is determined. The most conserved regions are

114 selected for building HMMs using the HMMER package (Eddy 2011). Input sequences that

115 cover most of the barcoding region and are taxonomically annotated are used to build a BLAST

116 (Altschul et al. 1997) database for sequence classification. Finally, the sequences in the BLAST

117 database are aligned using MAFFT, and the intra- and inter-taxonomic sequence identities are

118 calculated to derive meaningful sequence identity cutoffs at each taxonomic level. This entire

119 process is described in more detail in the Metaxa2 2.2 manual

120 (http://microbiology.se/software/metaxa2/) and in Bengtsson-Palme et al. (2018).

121   After extraction, sequences were then curated by removal of duplicate sequences using

122 the Java code provided with the RDP classifier (v2.11; Wang et al. 2007), which removes

123 identical sequences or any sequence contained within another sequence. At this point, we

124 conducted extensive curation of the available lineage data for the reference sequence database.

125 For references lacking complete annotation at midpoints within the Linnaean lineage, we used

126 Perl regular expression-based substitution to complete the annotation according to established

127 taxonomic authorities, including MilliBase (Sierwald 2017), the Integrated Taxonomic

128   Information System (http://www.itis.gov) and the phylogenomic analysis of Regier et al. (2010).

129   Table 1 shows the substitutions made. Further, we removed ranks containing annotations

130   reflective of open nomenclature, such as sp., cf. and Incertae sedis, as well as ranks annotated as

131   'undef.' Lastly, we removed entries containing more than two consecutive uncalled base pairs.

132   Upon analyzing the representativeness of this initial database across arthropod classes

133   and insect orders, we found that amplicon sequences from two insect orders, Strepsiptera and

134   Embioptera, were not present in the curated database, likely due to their poor sequence similarity

135   to the reference sequence used to designate the amplicon barcode region of interest. To add

136   Strepsiptera and Embioptera COI amplicons, all NCBI COI sequences belonging to these orders

137   were downloaded on October 10th, 2017, curated and added to the Metaxa2 COI database. To

138   improve recovery of amplicons from these insect orders during curation, a representative

139   sequence from both Embioptera and Strepsiptera, representing the 157 bp COI amplicon of

140   interest, was used when building the Metaxa2 database. This retrospective addition of sequences

141   belonging to Strepsiptera and Embioptera contributed 102 and 3 non-redundant reference

142   sequences to the database, respectively. After this final sequence addition step, a Metaxa2

143   database was built to include all curated sequences and this database is available through the

144   Metaxa2 software package (http://microbiology.se/software/metaxa2/).

145   To assess the degree to which our amplicon sequence extraction, dereplication and

146   curation procedures worked, we took inventory of the number of sequences per species in the

147   initial input data as well as the number of sequences and genera present in the data at three points

148   during curation: 1) in the initial input data, 2) following Metaxa2 database builder-based

149   amplicon sequence extraction and 3) in the final database following dereplication and taxonomic

150   curation.

151   *Classifier performance evaluation*

152   For performance evaluations, the methods used were highly similar to those of

153   Richardson et al. (2017). For three repeated samplings, we randomly selected 10 percent of the

154   curated sequences to obtain testing data, using the remaining 90 percent of sequences to train the

155   Metaxa2 classifier for performance evaluations. To assess the effect of sequence length on

156   classifier performance, we used a Python script to crop the test case sequences to 80 bp in length,

157   approximately half the median length of the original reference sequence dataset. Evaluating

158   classification performance on these short sequences provides a test of the classifiers robustness

159   to sequence length variation and enables estimation of the potential for classifying sequences

160   from short, high-throughput technology, such as 100 cycle single-end Illumina HiSeq

161   sequencing. We then performed the following analyses on both the full-length (157 bp) and half-

162   length (80 bp) test case sequences, separately.

163   To characterize the relationship between the Metaxa2 reliability score, an estimate of

164   classification confidence, and the probability of classification error, we used the COI trained

165   classifier to classify the testing datasets, requiring the software to classify to the family rank

166   regardless of the reliability score of the assignment. After comparing the known taxonomic

167   identity of each reference test case to the Metaxa2 predicted taxonomic identity, we regressed

168   5,000 randomly chosen binary classification outcomes, '1' representing an incorrect

169   classification and '0' representing a correct classification, against the Metaxa2 reliability score

170   using local polynomial logistic regression in R (v3.3.1; R Core Team 2014) with the span set to a

171   value of 0.5.

172   For each of the three testing and training datasets, we classified the testing sequences

173   using Metaxa2 with a reliability score threshold (-R) of 68. With the resulting classifications, we

174   compared the known taxonomic identity of each reference test case to its Metaxa2 classification,

175   from kingdom to species, to assess the proportion of true positive, true negative, false negative

176   and false positive predictions. We also calculated false discovery rate, as measured in errors per

177   assignment for each rank.

178   To assess the rate of taxonomic overclassification at the genus and species levels, we

179   searched the testing dataset for sequence cases belonging to arthropod genera and species not

180   represented in the training database. For each of the two ranks, we then determined the

181   proportion of these sequences which were classified. Since the actual identity of such a sequence

182   case is not represented in the training data, any such classification represents a particular type of

183   misclassification known as an overclassification or overprediction. Lastly, for these sequence

184   cases, we looked at how the classifier performed at the preceding rank (e.g. for the species-level

185   cases, we analyzed classifier performance at the genus-level). Such analysis provides a measure

186   of how well the software is able to perform at the next higher rank for these worst-case-scenario

187   input sequences.

188   For each order in the testing data, we estimated the family, genus and species-level

189   proportion of sequences assigned and false discovery rate to estimate the degree of variance in

190   performance across major arthropod lineages. For this analysis, the false discovery rate was

191   again defined as the number of errors per assignment. After conducting this analysis, we limited

192   our interpretation of the results to orders with at least 100 tests sequence cases at all of the ranks

193   analyzed, family, genus and species. A Python script which takes the testing sequence

194   taxonomies, training sequence taxonomies and Metaxa2 predicted taxonomies as input and

195   provides the summaries of classification performance described above is provided with the

196   GitHub repository associated with this work, which is detailed in the Supplemental Information

197   section.

198   **Results**

199   Following curation and extraction, we obtained 199,206 reference amplicon sequences

200   belonging to 51,416 arthropod species. Over 90 percent of the references were between 142 and

201   149 bp in length, with a minimum reference sequence length of 94 bp. For the final database

202   creation and classifier training procedure, many reference amplicons were shorter than the 157

203   bp region of interest due to the incompleteness of some reference sequences and the trimming of

204   taxonomically uninformative ends during Metaxa2 training. Prior to this step, 82 percent of the

205   sequences were between 150 and 157 bp in length following the original extraction and these

206   longer sequences can be found at the GitHub repository associated with this work. The

207   taxonomic representativeness of the database across different arthropod classes and insect orders,

208   including the number of families, genera and species in each, are presented in Tables 2 and 3.

209   Analyzing the number of sequences per species in the input reference sequence data, we

210   observed a heavily right-skewed distribution, with a median of 2 and a mean of 11.1 sequences

211   per species (Figure 1A). Further, 32.0 percent of species were represented by 5 or more

212   sequences and 40 species, including *Bemisia tabaci* and *Delia platura*, were represented by

213   between 1,000 and 9,736 entries. After conducting amplicon sequence extraction using the

214   Metaxa2 database builder tool, we were able to extract the COI region of interest from 80.8

215   percent of the input sequences, which corresponded to 73.4 percent of the genus-level diversity

216   found in the original input data. Following sequence dereplication, removal of sequences with

217   three or more ambiguous base calls and taxonomic lineage curation, our final database contained

218   approximately 13 percent of the input extracted sequences, which represented 98.2 percent of the

219   genus-level richness of the input extracted reference amplicon sequences (Figure 1B and 1C)

220   Regressing classification outcome against the Metaxa2 reliability score yielded a similar

221   best fit model for both the 80 bp and full length test sequence datasets (Figure 2). For both

222   regressions, the probability of sequence mis-assignment was below 10 percent for reliability

223   scores above 70. Thus, for our evaluations, we chose a reliability score of 68, which

224   corresponded to family-level error probabilities of approximately 11.3 percent and 9.5 percent

225   for 80 bp and full length sequences, respectively.

226   In evaluating the performance of our classification database when analyzed with Metaxa2

227   at a reliability score cutoff of 68, we found consistently low proportions of false positives across

228   all ranks, though the proportion of true positive, true negative and false negatives varied more

229   considerably from kingdom to species (Figure 3A). Further, while the Metaxa2 false discovery

230   rate increased with higher resolution ranks (Figure 3B), it was generally low, never exceeding 5

231   percent at the genus level. Interestingly, Metaxa2 displayed low variance in the proportion of

232   sequences assigned when classifying 80 bp sequences relative to full length sequences of 147 bp

233   in median length. Overall, the proportion of sequences assigned was greater than 90 percent

234   through the order level for both full length and half length sequences. Beyond the order level,

235   this statistic decreased to 53 and 56 percent at the species level for half length and full length

236   sequences, respectively. Conversely, the proportion of false positives varied more strongly by

237   sequence length and was greatest at higher-resolution taxonomic levels. At the species level, 1.97

238   percent of 80 bp sequences were misclassified, compared to only 1.13 percent for full length

239   sequences. At the order level, the percent of sequences misclassified was 0.59 and 0.65 for 80 bp

240   and full length sequences, respectively. As measured in errors per assignment, the classification

241 false discovery rate was similarly highest at the species level, with 7.3 and 6.3 percent of

242 assignments being incorrect for 80 bp and full length sequences, respectively. False discovery

243 rates decreased to 1.2 and 0.7 percent of assignments being incorrect at the order level for 80 bp

244 and full length sequences, respectively.

245 During our evaluation of taxonomic overclassifiction, we found between 3,141 and 3,202

246 sequence test cases belonging to species not represented in the corresponding training data and

247 between 612 and 630 sequence test cases belonging to genera not represented in the

248 corresponding training data across the three iterations of training and testing data. At the species

249 level, the proportion of these cases which were overclassified was roughly equal for full length

250 and half length sequences, with 5.4 and 5.2 percent being overclassified (Figure 4A). With

251 respect to genus level classification performance on these species overclassification test cases,

252 approximately 31 and 37 percent of test cases were classified correctly as true positives or true

253 negatives for half length and full length sequences. Genus level false positive proportions for

254 these sequence cases were 6.2 and 5.1 percent (Figure 4B). For the genus level overclassification

255 cases, the difference in overclassification rates by sequence length was slightly larger, with

256 approximately 8.9 and 10.9 percent of full length and half length test cases being overclassified

257 (Figure 4C). Family level performance on genus level overclassification cases was slightly lower

258 with the proportion of true positive and true negative identifications summing to 25 and 27

259 percent for half length and full length sequences, with corresponding false positive proportions

260 of 8.8 and 8.0

261 Evaluating classification performance for each order resulted in unsurprising outcomes

262 across taxonomic ranks, from family to species. Generally, the proportion of sequences assigned

263 decreased and false discovery rate increased with increasing taxonomic resolution. Between

264 orders, there was noteworthy variation in performance. For example, the proportion of sequences

265 assigned was highest among trichopteran sequences and lowest among lepidopteran sequences,

266 with approximately 94 and 50 percent of sequences being assigned to the family rank for each

267 order, respectively. Further, while the proportion of sequences assigned was similarly lowest for

268 lepidopteran sequences at the genus and species levels, wherein 39 and 21 percent of sequences

269 were assigned for each respective rank, the highest proportion of sequences assigned at these

270 ranks was not observed with trichopteran sequences. Instead, genus and species level proportion

271 assigned was highest among Sessilia sequences, with 86 and 78 percent of sequences being

272 assigned to each respective rank.

273 **Discussion**

274   While species-specific PCR and immunohistochemistry-based methods have been useful

275 in documenting arthropod food webs (Stuart and Greenstone, 1990; Symondson 2002; Weber et

276 al. 2006; Blubaugh et al. 2016), the narrow species-by-species nature of such approaches has

277 limited their utility for answering large-scale or open-ended ecological questions. With the

278 increasing availability of high-throughput sequencing, arthropod metabarcoding will continue to

279 become more broadly applicable to scientific questions spanning a diversity of research areas.

280 The development of improved methods for drawing maximal inferences from sequence data is an

281 important area for further methodological research. In creating a highly curated COI reference

282 amplicon sequence database and evaluating its performance when used with the Metaxa2

283 taxonomic classifier, we have developed a new method to aid researchers in the analysis of

284 arthropod metabarcoding data.

285   Though predictions vary greatly, researchers have estimated the species richness of

286 arthropods to be between 2.5 to 3.7 million (Hamilton et al. 2010). Further, according to the

287 literature review of Porter et al. (2014), 72,618 insect genera have been described to date. Thus

288 the 51,416 species and 17,039 genera represented in our database account for only a small

289 fraction of arthropod biodiversity. The limited representativeness of currently available, high

290 quality reference sequence amplicons for the COI region highlights the need for continued

291 efforts to catalogue arthropod biodiversity with molecular techniques. Despite this current

292 limitation, the combination of molecular gut content analysis with high-throughput sequencing is

293 a promising path toward investigating arthropod trophic ecology and biodiversity monitoring

294 with greater sensitivity and accuracy relative to alternate approaches.

295   The results of our inventory of sequences per species and genus level richness at various

296 stages in the database curation process revealed that our amplicon extraction procedure was

297 highly sensitive, extracting and trimming approximately 81 percent of the input sequences down

298 to the 157 bp region of interest. Further, approximately 87 percent of these extracted sequences

299 represented sequence redundancies and were removed during dereplication. As mentioned

300 previously, the trimming of sequence residues adjacent to the barcode of interest and removal of

301 redundant sequences not only makes computational analysis less resource intensive, it can also

302 improve classification performance. For k-mer style classifiers, extraneous sequence residues can

303 bias model selection during classifier training, while abundant sequence duplicates can result in

304 an overwhelming number of identical top hit alignments for alignment-based classifiers.

305   Overall, the best fit local regression models summarizing the relationship between the

306 Metaxa2 reliability score and the probability of classification error were useful in that the

307 likelihood of misclassification was always less than what would be expected based on the

308 reliability score. For example, a reliability score of 90 corresponded to only a 3.3 percent

309 probability of family-level misclassification for full length sequences. We selected a reliability

310 score of 68 for subsequent analysis as this provided a balanced trade-off between sensitivity and

311 accuracy. Using this reliability score we observed minimal false positive rates and overall

312 proportions of misclassification when comparing our results to those of similar studies (Porter et

313 al. 2014; Bengtsson-Palme et al. 2015; Edgar 2016; Richardson et al. 2017). Given that the

314 family level probability of error was only 9.5 percent at a reliability score of 68, a lower

315 reliability score threshold may be justifiable for certain research situations. However, further

316 testing should be conducted to ensure that the relationship between reliability score and

317 classification confidence is similar across taxonomic ranks and between different DNA barcodes.

318   With respect to sensitivity using a reliability score of 68, our results were highly

319 dependent upon the rank being analyzed, with sensitivities, as measured by the total proportion

320 of sequences classified, above 60 percent only being achieved at the family and order ranks. To

321 some degree, these sensitivity estimates reflect the large degree of database incompleteness at the

322 genus and species ranks, wherein approximately 3.6 and 25.5 percent of unclassified sequences

323 were true negatives. However, to our knowledge, no other studies have reported classification

324 sensitivity data for this COI amplicon locus. This makes it difficult to ascertain if Metaxa2 –

325 with an average of approximately 44.7 percent of false negative assignments at the genus and

326 species ranks – exhibits relatively low sensitivity or if this locus is limited in discriminatory

327 power. The short length of the 157 bp COI amplicon region relative to other barcoding regions

328 such as the ITS and 18S rRNA regions (Hugerth et al. 2014; Wang et al. 2015) could be a cause

329 of such limited discriminatory power.

330   As expected, analyzing cases of overclassification in our data revealed that sequences

331 from taxa lacking representation in the database are far more likely to be misclassified relative to

332 sequences from well-represented taxa. This is supported by an approximately 10 percent

333 probability of genus level overclassification for sequences from unrepresented genera relative to

334 a 1 to 2 percent probability, depending on sequence length, for all sequence test cases.

335 Interestingly, the genus level overclassification rate was approximately double that observed at

336 the species level. This seems counterintuitive but is expected in light of the discussion put forth

337 by Edgar (2018), wherein the percent identity difference from the closest reference sequence

338 match is considered one of the most important predictors of classification performance. For a

339 species-level overclassification case, the closest reference sequence match in the corresponding

340 database is a sequence from a congeneric species in the best case scenario. Since the closest

341 reference sequence match for a genus level overclassification case is a sequence from a

342 confamilial species at best, the average percent identity difference from the closest reference is

343 greater for a genus level overclassification case than for a species level overclassification case.

344 Thus, for overclassification cases in particular, higher levels of error are expected at more

345 inclusive taxonomic ranks.

346   While the genus level overclassification estimates we observed are not desirable, they are

347 considerably lower than similar estimates for the RDP classifier, which range from 21.3 percent

348 to 67.8 percent depending on the database, locus analyzed and cross-validation approach used

349 (Edgar 2016; Richardson et al. 2017). Further, the observed degree of genus level

350 overclassification using Metaxa2 with our COI database was similar to or less than that of the

351 recently developed SINTAX classifier (Edgar 2016, Edgar 2018). Interestingly, the recent

352 analysis of Edgar (2018), resulted in a similar estimate of the Metaxa2 genus level

353 overclassification rate while revealing a considerably higher overclassification rate for the

354 majority of other classifiers tested. Though this analysis found Metaxa2 to be relatively less

355 sensitive than other classifiers, a weakness of the work was that it only tested Metaxa2 using the

356   default reliability score of 80, while testing multiple confidence thresholds for alternate

357   classifiers, such as SINTAX and RDP. Given that our analysis has revealed the Metaxa2 default

358   reliability score to be too conservative – at least for the COI locus – such results are difficult to

359   interpret. In general, such comparisons across studies should be approached with caution as

360   multiple factors complicate the interpretation of classification performance, such as locus

361   discriminatory power, database completeness and the choice of evaluation metrics used.

362   Ultimately, direct comparisons of classification methods using standardized loci and databases

363   are needed to more rigorously compare performance.

364   With respect to Metaxa2 classification of full length relative to half length amplicon

365   sequences, we observed surprisingly small differences in performance. Consistently, the

366   proportion of misclassified sequences was greater for half length sequences. When considering

367   error and sensitivity together, the false discovery rate or errors per assignment for full length

368   sequences was consistently less than or equal to that achieved during the classification of half

369   length sequences. Lastly, when considering the relationship between the Metaxa2 reliability

370   score and the probability of classification error at the family level, we noted highly similar local

371   polynomial regression models of error probability for both full length and half length sequences.

372   **Conclusions**

373   Here, we assembled a highly curated database of arthropod COI reference amplicon

374   sequences, trained a recently developed hierarchical DNA sequence classifier using the database

375   and conducted extensive *in silico* performance evaluations on the resulting classification

376   pipeline. Overall, we found a high degree of sequence redundancy within the initial, uncurated

377   dataset, highlighting the importance of effective sequence dereplication during the creation of

378   databases designed for metabarcoding analysis. Further, the limited representativeness of the

379 database with respect to arthropod biodiversity indicates that additional sequencing effort is

380 needed to further improve the performance of arthropod metabarcoding techniques. Though the

381 performance evaluations presented in this work were conducted on a large corpus of available

382 biological data, the results are not necessarily directly transferable to all experimental settings.

383 For example, variations in sequence error profiles and taxonomic distributions among datasets

384 are potential confounding factors. Despite these limitations, this work provides researchers with

385 a new resource for arthropod COI sequence analysis and novel data for gauging the strengths and

386 limitations of different approaches to arthropod metabarcoding.

387 **Acknowledgments**

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

389 Center.

390 **Data Availability**

391 The entire set of curated sequences associated with this work can be downloaded from the

392 associated GitHub repository: https://github.com/RTRichar/Zeale_COI_Database. This

393 repository has additionally been archived using Zenodo (DOI: 10.5281/zenodo.1256869)

394 **References**

395 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped

396   BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic

397   Acids Research, 25, 3389-3402.

398 Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification

399   programs. BMC Bioinformatics, 13, 1-13.

400 Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH

401   (2015) Metaxa2: Improved identification and taxonomic classification of small and large

402   subunit rRNA in metagenomic data. Molecular Ecology Resources, 15, 1403-1414.

403 Bengtsson-Palme J, Richardson RT, Meola M, Wurzbacher C, Tremblay ED, Thorell K, Kanger

404   K, Eriksson M, Bilodeau GJ, Johnson RM, Hartmann M, Nilsson RH (2018) Taxonomic

405   identification from metagenomic and metabarcoding data using any genetic marker.

406   bioRxiv, 253377, doi: https://doi.org/10.1101/253377

407 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. Nucleic

408   Acids Research, 33, D34-D38.

409 Blubaugh CK, Hagler JR, Machtley SA, Kaplan I (2016) Cover crops increase foraging activity

410   of omnivorous predators in seed patches and facilitate weed biological control.

411   Agriculture, Ecosystems and Environment, 231, 264-270.

412 Corse E, Costedoat C, Chappaz R, Pech N, Martin J-F, Gilles A (2010) A PCR-based method for

413   diet analysis in freshwater organisms using 18S rDNA barcoding on faeces. Molecular

414   Ecology Resources, 10, 96-108.

415 Eddy SR (2011) Accelerated profile HMM searches. PLoS Computational Biology, 7, e1002195.

416 Edgar RC (2015) UTAX algorithm. Available at:

417   http://www.drive5.com/usearch/manual/utax_algo.html (accessed 6 January 2016).

418 Edgar RC (2016) SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS

419   sequences. Biorxiv. https://doi.org/10.1101/074161.

420 Edgar RC (2018) Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences.

421   PeerJ, 6, e4652.

422 Elbrecht V and Leese F (2017) Validation and development of COI metabarcoding primers for
423   freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science, 5, 11.
424 Guardiola M, Uriz MJ, Taberlet P, Coissac E, Wangensteen OW, Turon X (2015) Deep-sea,
425   deep-sequencing: Metabarcoding extracellular DNA from sediments of marine canyons.
426   PloS One, 10, e0139633.
427 Hamilton AJ, Basset Y, Benke KK, Grimbacher PS, Miller SE, Novotny V, Samuelson A, Stork
428   NE, Weiblen GD, Yen JDL (2010) Quantifying uncertainty in estimation of tropical
429   arthropod species richness. The American Naturalist, 175, 90-95.
430 Hugerth LW, Muller EEL, Hu YOO, Lebrun LAM, Roume H, Lundin D, Wilmes P, Andersson
431   AF (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic
432   diversity in microbial consortia. PloS ONE, 9, e95567.
433 Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of
434   environmental sequences using MEGAN 4. Genome Research, 21, 1552-1560.
435 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
436   improvements in performance and usability. Molecular Biology and Evolution, 30, 772-
437   780.
438 Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M (2009) 5.8S-28S rRNA
439   interaction and HMM-based ITS2 annotation. Gene, 430, 50-57.
440 Mollot G, Duyck P-F, Lefeuvre P, Lescourret F, Martin J-F, Piry S, Canard A, Tixier P (2014)
441   Cover cropping alters the diet of arthropods in a banana plantation: A metabarcoding
442   approach. PloS ONE, 9, e93740.
443 NCBI Resource Coordinators (2018) Database Resources of the National Center for
444   Biotechnology Information. Nucleic Acids Research, 46, D8–D13.

445 Ohio Supercomputer Center (1987) Citation. Columbus, Ohio, USA.

446   http://osc.edu/ark:/19495/f5s1ph73.

447 Peabody MA, Van Rossum T, Lo R, Brinkman FSL (2015) Evaluation of shotgun metagenomics

448   sequence classification methods using in silico and in vitro simulated communities. BMC

449   Bioinformatics, 16, 363.

450 Porter TM, Golding GB (2012) Factors that affect large subunit ribosomal DNA amplicon

451   sequencing studies of fungal communities: Classification method, primer choice, and

452   error. PloS One, 7, e35749.

453 Porter TM, Gibson JF, Shokralla S, Baird DJ, Golding GB, Hajibabaei M (2014) Rapid and

454   accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1

455   (COI) DNA barcode sequences using a naive Bayesian classifier. Molecular Ecology

456   Resources 14, 929-942.

457 R Core Team (2014) R: A language and environment for statistical computing. Vienna, Austria.

458   http://www.R-project.org/

459 Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW

460   (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-

461   coding sequences. Nature, 463, 1079-1083.

462 Richardson RT, Bengtsson-Palme J, Johnson RM (2017) Evaluating and optimizing the

463   performance of software commonly used for the taxonomic classification of DNA

464   metabarcoding sequence data. Molecular Ecology Resources 17, 760-769.

465 Richardson RT, Lin C-H, Quijia JQ, Riusech NS, Goodell K, Johnson RM (2015a) Rank-based

466   characterization of pollen assemblages collected by honey bees using a multi-locus

467   metabarcoding approach. Applications in Plant Sciences, 3, 1500043.

468 Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, Steffan-Dewenter I,

469   Keller A (2015) Increased efficiency in identifying mixed pollen samples by meta-

470   barcoding with a dual-indexing approach. BMC Ecology, 15, 1-9.

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

472 Somervuo P, Koskela S, Pennanen J, Nilsson RH, Ovaskainen O (2016) Unbiased probabilistic

473   taxonomic classification for DNA barcoding. Bioinformatics, 32, 2920-2927.

474 Somervuo P, Yu DW, Xu CCY, Ji Y, Hultman J, Wirta H, Ovaskainen O (2017) Quantifying

475   uncertainty of taxonomic placement in DNA barcoding and metabarcoding. Methods in

476   Ecology and Evolution, 8, 398-407.

477 Stuart MK, Greenstone MH (1990) Beyond ELISA: A rapid, sensitive, specific immunodot assay

478   for identification of predator stomach contents. Annals of the Entomological Society of

479   America, 83, 1101-1107.

480 Symondson WOC (2002) Molecular identification of prey in predator diets. Molecular Ecology,

481   11, 627-641.

482 Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce

483   C, Nevard T, Potts SG, Zhou X, Yu DW (2015) High-throughput monitoring of wild bee

484   diversity and abundance via mitogenomics. Methods in Ecology and Evolution, 6, 1034-

485   1043.

486 Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment

487   of rRNA sequences into the new bacterial taxonomy. Applied and Environmental

488   Microbiology, 73, 5261-5267.

489   Wang X-C, Liu C, Huang L, Bengtsson-Palme J, Chen H, Zhang J-H, Cai D, Li J-Q (2015)

490   ITS1: A DNA barcode better than ITS2 in eukaryotes? Molecular Ecology Resources, 15,

491   573-586.

492   Weber DC, Rowley DL, Greenstone MH, Athanas MM (2006) Prey preference and host

493   suitability of the predatory and parasitoid carabid beetle, *Lebia grandis*, for several

494   species of *Leptinotarsa* beetles. Journal of Insect Science, 6, 1-14.

495   Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup:

496   Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring.

497   Methods in Ecology and Evolution, 3, 613-623.

498   Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G (2011) Taxon-specific PCR for DNA

499   barcoding arthropod prey in bat faeces. Molecular Ecology Resources, 11, 236-244.

500   **Tables**

501   Table 1: Summary of taxonomic annotations made for references which had undefined ranks at

502   midpoints in their respective taxonomic lineages.

503   Table 2: Summary of taxonomic representation across all arthropod classes and associated sister

504   groups. Numbers may include sub and super groupings.

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

506   Numbers may include sub and super groupings.

507   **Figures**

508   Figure 1: A percent density histogram of the number of sequences per species (A) shows the

509   distribution of redundancy within the NCBI Nucleotide entries used. The dashed blue line and

510   solid red line indicate the median and mean number of sequences per species, respectively.

511   Inventories of the number of sequences (A) and genera (B) input into the curation process,

512 following Metaxa2 extraction and following dereplication of redundant sequences and curation

513 of taxonomic lineages.

514 Figure 2: A logistic regression analysis of case-by-case classification accuracy, '1' indicating a

515 false-positive identification and '0' indicating a true-positive identification, regressed against

516 classification reliability score for half length (A) and full length (B) test sequence cases. A best

517 fit local polynomial regression line (solid blue with 95 percent confidence interval) was used to

518 estimate the relationship between reliability score and the probability of mis-classification.

519 Dashed red lines illustrate the hypothetically ideal 1 to 1 relationship between error probability

520 and the Metaxa2 reliability score, an estimate of classification confidence. Solid black lines

521 highlight the 10 percent error probability.

522 Figure 3: Mean proportion and standard error of true positives (TP), true negatives (TN), false

523 negatives (FN) and false positives (FP) for the classification of all testing sequences, conducted

524 on both the full-length and half-length sequences (A). Mean and standard error of the false

525 discovery rate for both full length and half length sequences, as measured in errors per

526 assignment, during classification of all testing sequences (B).

527 Figure 4: Proportional species level overclassification rate (A) and genus level classification

528 performance (B) for test sequence cases from species not represented in the corresponding

529 training data. Proportional genus level overclassification rate (C), and family level classification

530 performance (D) for test sequence cases from genera not represented in the corresponding

531 training data.

532 Figure 5: Mean and standard error of the proportion of sequences assigned and false discovery

533 rate, as measured in errors per assignment, measured at the family, genus and species levels for

534 sequences belonging to each order. These results represent classification performance by

535   arthropod order for the full length sequences only and orders with fewer than 100 test sequence

536   cases at any of the ranks analyzed were excluded from analysis.

Taxonomic midpoint annotation corrections.

Summary of taxonomic annotations made for references which had undefined ranks at 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	
Genus	Species Tanypodinae spp.	Genus Tanypodinae	
Genus	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	

Summary of database completeness with respect to all arthropods.

Summary of taxonomic representation across all arthropod classes and associated sister groups. Numbers may include sub and super groupings.

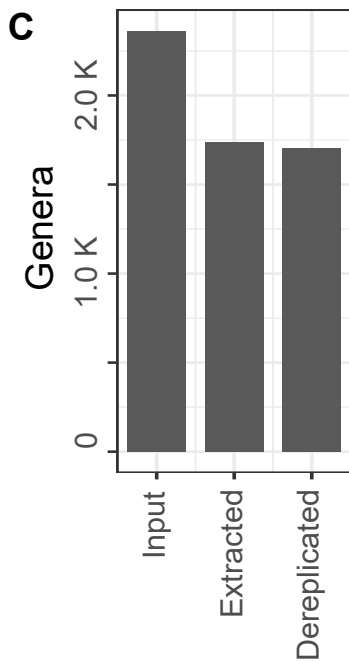
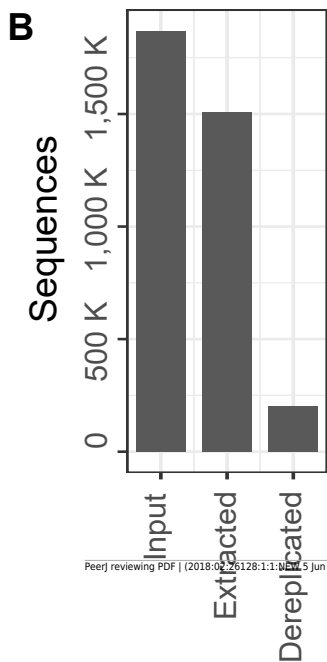
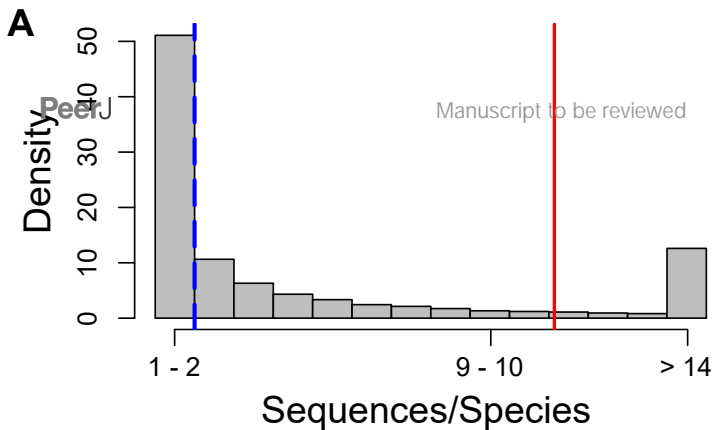
Class	Number of Orders	Number of Families	Number of Genera	Number of Species
Heterotardigrada	1	2	2	1
Eutardigrada	1	3	12	20
Onychophora	1	2	17	42
Pycnogonida	1	10	27	89
Cephalopoda	1	1	1	1
Merostomata	1	1	3	4
Arachnida	17	226	740	1804
Myriapoda	2	4	7	9
Chilopoda	5	16	53	172
Diplopoda	11	33	95	181
Ostracoda	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**

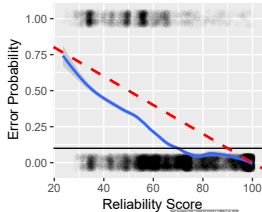
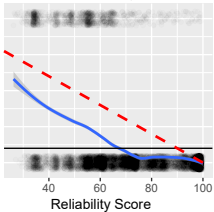
Summary of database completeness with respect to insects.

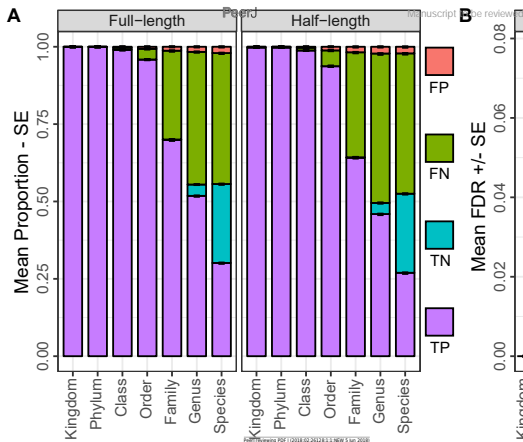
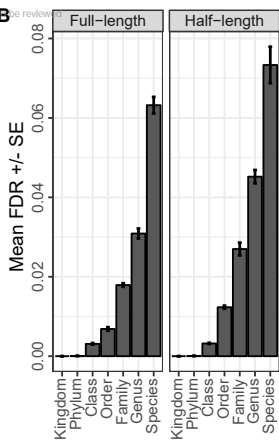
Summary of insect taxa included in the arthropod COI database following curation. Numbers may include sub and super groupings.

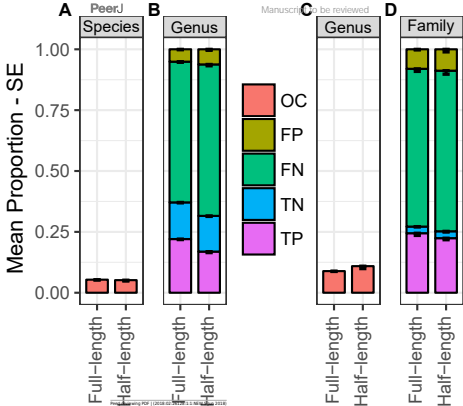
Order	Number of Families	Number of Genera	Number of Species
Archaeognatha	2	15	18
Zygentoma	3	3	3
Odonata	34	243	488
Ephemeroptera	26	108	378
Zoraptera	1	1	1
Dermaptera	4	6	7
Plecoptera	15	84	199
Orthoptera	30	299	603
Mantophasmatodea	1	1	1
Grylloblattodea	1	1	0
Embioptera	1	2	2
Phasmatodea	6	20	28
Mantodea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

Data inventory throughout 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.

Estimating reliability score accuracy.

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 for half length (A) and full length (B) test sequence cases. A best fit local polynomial regression line (solid blue with 95 percent confidence interval) was used to estimate the relationship between reliability score and the probability of mis-classification. Dashed red lines illustrate the hypothetically ideal 1 to 1 relationship between error probability and the Metaxa2 reliability score, an estimate of classification confidence. Solid black lines highlight the 10 percent error probability.

Classification performance accross entire testing corpus.

Mean proportion and standard error of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP) for the classification of all testing sequences, conducted on both the full-length and half-length sequences (A). Mean and standard error of the false discovery rate for both full length and half length sequences, as measured in errors per assignment, during classification of all testing sequences (B).

Overclassification analysis.

Proportional species level overclassification rate (A) and genus level classification performance (B) for test sequence cases from species not represented in the corresponding training data. Proportional genus level overclassification rate (C), and family level classification performance (D) for test sequence cases from genera not represented in the corresponding training data.

Classification performance by order.

Mean and standard error of the proportion of sequences assigned and false discovery rate, as measured in errors per assignment, measured at the family, genus and species levels for sequences belonging to each order. These results represent classification performance by arthropod order for the full length sequences only and orders with fewer than 100 test sequence cases at any of the ranks analyzed were excluded from analysis.