



**HAL**  
open science

## A measure of the DNA barcode gap for applied and basic research

Jarrett D Phillips, Cortland K Griswold, Robert G Young, Nicolas Hubert,  
Robert H Hanner

► **To cite this version:**

Jarrett D Phillips, Cortland K Griswold, Robert G Young, Nicolas Hubert, Robert H Hanner. A measure of the DNA barcode gap for applied and basic research. Robert DeSalle. DNA Barcoding: Methods and Protocols, 2744, Humana, pp.375-390, 2024, 978-1-0716-3583-4. 10.1007/978-1-0716-3581-0\_24. hal-04579301

**HAL Id: hal-04579301**

**<https://hal.science/hal-04579301>**

Submitted on 17 May 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 A measure of the DNA barcode gap for applied and basic research

2

3 Jarrett D. Phillips<sup>1,4‡\*</sup> (ORCID: 0000-0001-8390-386X),

4 Cortland K. Griswold<sup>4‡</sup> (ORCID: 0000-0003-2993-7043)

5 Robert G. Young<sup>4</sup> (ORCID: 0000-0002-6731-2506)

6 Nicolas Hubert<sup>2</sup> (ORCID: 0000-0001-9248-3377)

7 Robert H. Hanner<sup>3,4</sup> (ORCID: 0000-0003-0703-1646)

8

9 <sup>1</sup>*School of Computer Science, University of Guelph, Ontario, Canada N1G 2W1*

10 <sup>2</sup>*UMR ISEM (IRD, UM, CNRS), Université de Montpellier, Place Eugène Bataillon,*

11 *34095 Montpellier cedex 05, France*

12 <sup>3</sup>*Biodiversity Institute of Ontario, University of Guelph, Ontario, Canada N1G*

13 *2W1*

14 <sup>4</sup>*Department of Integrative Biology, University of Guelph, Ontario, Canada N1G*

15 *2W1*

16

17 ‡ Equal contribution

18

19 \***Correspondence:** Jarrett D. Phillips

20 **Email:** [jphill01@uoguelph.ca](mailto:jphill01@uoguelph.ca)

21

22 **Keywords:** bootstrapping, DNA barcoding, intraspecific genetic distance,

23 interspecific genetic distance, multispecies coalescent, nonparametrics,

24 speciation

25

26 **Running Title:** A method to detect DNA barcode gaps

27

28

29

30

31

32

33

34

35

36

37

38 **Abstract**

39  
40 DNA barcoding has largely established itself as a mainstay for rapid molecular  
41 taxonomic identification in both academic and applied research. The use of  
42 DNA barcoding as a molecular identification method depends on a “DNA  
43 barcode gap” — the separation between the maximum within-species  
44 difference and the minimum between-species difference. Previous work  
45 indicates the presence of a gap hinges on sampling effort for focal taxa and  
46 their close relatives. Furthermore, both theory and empirical work indicate a  
47 gap may not occur for related pairs of biological species. Here, we present a  
48 novel evaluation approach in the form of an easily calculated set of  
49 nonparametric metrics to quantify the extent of proportional overlap in inter-  
50 and intraspecific distributions of pairwise differences among target species  
51 and their conspecifics. The metrics are based on a simple count of the number  
52 of overlapping records for a species falling within the bounds of maximum  
53 intraspecific distance and minimum interspecific distance. Our approach takes  
54 advantage of the asymmetric directionality inherent in pairwise genetic  
55 distance distributions, which has not been previously done in the DNA  
56 barcoding literature. We apply the metrics to the predatory diving beetle  
57 genus *Agabus* as a case study because this group poses significant  
58 identification challenges due to its morphological uniformity despite both

59 relative sampling ease and well-established taxonomy. Results herein show  
60 that target species and their nearest neighbour species were found to be  
61 tightly clustered, and therefore difficult to distinguish. Such findings  
62 demonstrate that DNA barcoding can fail to fully resolve species in certain  
63 cases. Moving forward, we suggest the implementation of the proposed  
64 metrics be integrated into a common framework to be reported in any study  
65 that uses DNA barcoding for identification. In so doing, the importance of the  
66 DNA barcode gap and its components on the success of DNA-based  
67 identification using DNA barcodes can be better appreciated.

68  
69 **Introduction**  
70

71 The utility of DNA barcodes **(1, 2)** to provide species-level  
72 identifications to unknown specimens depends strongly on historical,  
73 geographical, and ecological processes shaping mitochondrial DNA (mtDNA)  
74 polymorphism **(3, 4)**. Current identification based on DNA sequences further  
75 depends on the comprehensiveness of genomic sequence libraries housed in  
76 databases such as the Barcode of Life Data Systems (BOLD)  
77 (<http://www.barcodinglife.org>; **(5)**) and GenBank  
78 (<https://www.ncbi.nlm.nih.gov/genbank>). In addition, the successful use of  
79 such data assumes sequencing, as well as species labelling, are error-free.

80 Despite access to over 12 million DNA barcodes across more than 340  
81 thousand multicellular species in BOLD as of February 2023, current  
82 specimen sampling efforts are likely not sufficient to enable accurate  
83 taxonomic assignment of many sequences, as a majority species are still  
84 missing, and oftentimes less than 10 individuals (typically comprising only  
85 singletons) have been sampled for a species **(6)**.

86       Having a broad representation of within-species genetic diversity is  
87 necessary when it comes to elucidating genetic species boundaries and to  
88 accurately estimate the DNA barcoding gap — the difference between  
89 maximum intraspecific and minimum (nearest neighbour) interspecific  
90 sequence variation **(7, 8)**. Empirical observation inferred from marker-  
91 specific variation suggests intraspecific distances do not usually exceed those  
92 found among species, thereby indicating DNA sequences aptly capture the  
93 evolutionary genetic boundaries of species **(1, 2)**. The existence of the DNA  
94 barcode gap has been invoked as the primary explanation for why DNA  
95 barcoding works well in practice; however, it has been demonstrated to fail in  
96 (population genetic) theory **(9, 10, 11)**. In fact, several studies have suggested  
97 that DNA barcoding gaps can be an artifacts of insufficient specimen sampling,  
98 particularly across biological space (see **(12)** for some taxon-specific  
99 examples). With greater spatial coverage, intraspecific distances increase as a

100 result of isolation-by-distance, whereas interspecific distances tend to shrink  
101 due to encountering more closely related species **(13)**. While much work has  
102 been devoted to assessing patterns of intraspecific variation using DNA  
103 barcodes and highlighting the importance of specimen sampling for reliable  
104 DNA barcode gap estimation **(12)**, the role of interspecific genetic diversity on  
105 the success of DNA-based taxon identification has attracted less attention.  
106 Early on in DNA barcoding, barcode gaps were calculated using the mean  
107 intraspecific and mean interspecific genetic distance **(7)**. However, this  
108 scheme was prone to exaggerating the width of gaps, leading to poor species  
109 delimitation and resolution of taxon boundaries. False positives in the form of  
110 taxonomic oversplitting (where the minimum interspecific distance falls  
111 below the maximum intraspecific distance), and false negatives manifesting as  
112 excessive species lumping (where the nearest neighbour distance lies above  
113 the maximum intraspecific distance), were common **(7, 12)**. Later arguments  
114 for the use of the maximum intraspecific distance and the minimum  
115 interspecific distance greatly reduced identification biases in threshold-based  
116 species delimitation, but did not eliminate them altogether **(8)**. Further,  
117 problems arise from the reporting of the DNA barcode gap, particularly in  
118 taxon reference library publications, where it is often treated as a binary  
119 (Yes/No) response, rather than as a continuous random variable with

120 measurable statistical error reflecting the overall extent of intraspecific and  
121 interspecific distribution overlap. Robust sampling is therefore necessary to  
122 reliably detect DNA barcode gaps when they actually exist **(12)**.

123         Several computational and statistical approaches, in both parametric  
124 (model-based) and nonparametric (data-driven) settings, have been proposed  
125 to address the proper use of DNA barcodes. Tools for specimen identification  
126 and species delimitation include both frequentist likelihood and Bayesian  
127 posterior estimation models fitted via Markov Chain Monte Carlo (MCMC), for  
128 instance **(14, 15)**. Taxon delineation methods like the Generalized Mixed Yule  
129 Coalescent (GMYC) **(16)**, Automatic Barcode Gap Discovery (ABGD) **(17)**,  
130 Poisson Tree Processes (PTP) **(18)**, the Barcode Index Number (BIN) system  
131 **(19)**, and Assemble Species by Automatic Partitioning (ASAP) **(20)** have seen  
132 extensive use as stand-alone tools for species delineation tasks (*e.g.*, the *splits*  
133 (Species Limits by Threshold Statistics) R package **(21)**, as well as various  
134 webservers), despite numerous computational, statistical, and  
135 interpretational challenges **(12)**. Often, the plethora of available methods for  
136 specimen identification and species delimitation tasks produces conflicting  
137 results when applied across the same dataset, making the selection of a single  
138 method daunting for end users.

139           Generally, specimen identification using DNA barcodes is routine  
140 whenever species boundaries delimited using standardized loci such as  
141 cytochrome *c* oxidase subunit I (COI) are concordant with those inferred from  
142 the examination of morphological characters. Despite DNA barcoding's strong  
143 presence and widespread use over the past two decades, what appears to still  
144 be missing is an automatic multi-use tool to assess genomic markers for  
145 downstream specimen identification. While consensus largely exists regarding  
146 marker choice for multicellular eukaryotic taxa like animals, plants, and fungi,  
147 the same cannot be said for unicellular and prokaryotic organisms such as  
148 protists and cyanobacteria, where a multimarker approach is likely required  
149 for accurate DNA barcode gap assessment (*e.g.*, **(22, 23)**). This makes the  
150 development of such a marker evaluation scheme imperative.

151           In this paper, we propose statistical metrics to evaluate the DNA  
152 barcode gap based on coalescent theory **(24)**, whose deep role in DNA-based  
153 species delimitation was previously noted by Hubert and Hanner **(25)** and  
154 applied using the multispecies coalescent (MSC) **(26, 27)**. The application of  
155 the MSC in a multilocus context to help resolve instances of mixed haplotype  
156 clusters arising from non-monophyly was also previously advocated for by  
157 Collins and Cruickshank **(28)**. They argued that multilocus MSC approaches  
158 provide stronger biological support for speciation events between sister taxon



159 pairs compared to simpler distance-based heuristics. Here, easily computed  
160 nonparametric statistics are proposed to measure the separation between  
161 intraspecific and interspecific sequence variation between pairs of species.  
162 They are relatively straightforward to implement and evaluate.

163

## 164 **DNA Barcode Gap Metrics**

165 Define the proposed DNA barcode gap metrics as follows:

$$166 \quad p_x = \frac{\# \{d_{ij} \geq \min d_{XY}\}}{\# \{d_{ij}\}} \quad (1),$$

$$167 \quad q_x = \frac{\# \{d_{XY} \leq \max d_{ij}\}}{\# \{d_{XY}\}} \quad (2).$$

168 where  $d_{ij}$  are intraspecific differences,  $d_{XY}$  are interspecific differences and #  
169 corresponds to the “Number of”. Note in traditional estimation of the DNA  
170 barcode gap, the distribution of interspecific distances *excludes* the target  
171 species. In contrast, the present approach deviates from this by including *all*  
172 pairwise interspecific distances across all species pairs. The justification for  
173 employing this scheme is that it more accurately accounts for all species’ and  
174 coalescent histories in the available sample of DNA sequences. Note,  $p_x$  and  $q_x$   
175 are calculated for each species ( $x$ ), such that  $\# \{d_{ij}\}$  is the number of pairwise  
176 differences for a particular species and  $\max d_{ij}$  is for a given species. See **Table**  
177 **1** for a complete set of definitions.

178 **[Fig. 1 near here]**

179 [Fig. 2 near here]

180 [Table 1 near here]

181 Equations (1) and (2) above express the degree of proportional overlap of  
182 intraspecific and interspecific genetic distance distributions. The metrics can  
183 be applied to either differentiate species at the genus level, or resolve  
184 separation of target taxa and their nearest neighbours (*i.e.*, sister species)  
185 within a genus of interest. For example,  $XY$  would cover either an entire genus,  
186 with each species being well sampled, or the nearest neighbour species. The  
187 quantity  $p_x$  measures the overlap of intraspecific differences with interspecific  
188 ones, whereas  $q_x$  measures the overlap of interspecific differences with  
189 intraspecific ones. To distinguish metrics for nearest neighbours, we add a  
190 prime symbol (') to  $p_x$  and  $q_x$ . The proposed metrics outlined here stand in  
191 sharp contrast to previous ones which only examine “average” effects at the  
192 “community” level. While we here apply the metrics at the species level, they  
193 are easily extended to the level of Operational Taxonomic Units (OTUs) and  
194 other similar species proxies. **Figures 1** and **2** motivate the statistics. In  
195 **Figure 1**, the quantity  $t_{AB}$  corresponds to the time species  $A$  and  $B$  formed.  $t_{XY,$   
196  $t_{\min}$  is the (unobserved) minimum divergence time for two mitochondrial genes  
197 from different species within the genus.  $t_{j, \max}$  is the (unobserved) maximum  
198 divergence time for two mitochondrial genes from individuals of the same

199 species across the entire genus. The Most Recent Common Ancestor (MRCA) of  
200 the sample is noted, but in principle, the maximum (unobserved) divergence  
201 time between species *A* and *B* ( $t_{AB, \max}$ ) may be further back in time. **Figure 2**  
202 relates times in **Figure 1** to the proposed statistics, and the distributions of  
203 intraspecific and interspecific distances.

204 Quantities  $p_x$  and  $q_x$  (as well as  $p_x'$  and  $q_x'$ ) are bounded between zero  
205 and one. Values near or equal to zero suggest evidence of species' DNA  
206 barcode gaps within the genus under consideration for a given sample and  
207 species identifications. Conversely, values near or equal to one indicate little  
208 to no evidence of species' DNA barcode gaps, and therefore lack of support for  
209 specimen assignment to species. Distances may be measured using, for  
210 example, *p*-distance, Jukes-Cantor (JC69) (**29**) or Kimura-Two-Parameter  
211 (K2P) models (**30**). More specifically,  $p_x$  and  $q_x$  are calculated through  
212 counting the number of specimen record distances lying within the bounds for  
213 intraspecific and interspecific comparisons and then dividing this sum by the  
214 total number of considered distances, respectively. It should be noted that  $p_x$   
215 and  $q_x$ , as well as  $p_x'$  and  $q_x'$ , are calculated herein using a hybrid approach  
216 based on both the mean and the smallest intraspecific distance (see next  
217 section). While the use of the minimum is more common than the mean in the  
218 DNA barcoding literature, it is susceptible to specimen assignment errors in

219 BOLD and GenBank. Nevertheless, while the minimum interspecific distance  
220 can more easily expose cases of DNA barcode sharing, in well sampled taxa  
221 identified using highly diagnosable gene markers, phenomena such as  
222 introgressive hybridization still do occasionally occur.

223         A logical next step is testing for statistical significance and assessing  
224 deviations from expectations for a particular hypothesis. The proposed  
225 metrics are associated with considerable uncertainty given low sample sizes  
226 for most taxa within DNA barcode libraries. Thus, calculation of standard  
227 errors (SEs), which can be estimated through nonparametric bootstrapping  
228 **(31)** for instance, provided sample sizes are sufficiently large, is necessary. In  
229 cases where sample sizes are small however, no amount of bootstrapping will  
230 help ameliorate uncertainty, and if applied naively, give a false assessment of  
231 certainty **(12)**, especially when the metrics are found to be equal to zero.  
232 Bootstrapping distances will not be informative in such cases. A more robust  
233 bootstrapping approach in the context of estimating the true degree of  
234 distribution overlap entails resampling alignment sites directly, like is done in  
235 phylogenetics, as opposed to distances themselves. Because a distance is  
236 already a summary over a set of sites, resampling nucleotide positions has the  
237 advantage of retaining more information given DNA barcode sequences are  
238 short in length. Thus, bootstrapping at the level of sites may be more

239 informative because, in principle, sites leading to no overlap may not be  
240 included and an overlap may arise. In fact, because genetic distances used in  
241 calculating the proposed metrics are not statistically independent and  
242 identically distributed (IID) **(12)**, estimator SEs will depend directly on the  
243 number of sampled species within the focal genus ( $K$ ), as well as indirectly on  
244 the number of sampled specimens per species ( $N$ ). Therefore, plots of the  
245 proposed metrics *versus* their estimated SEs could be informative. In this  
246 regard, considering two species, one well sampled, and the other poorly  
247 sampled, the lower values of the proposed metrics will likely produce larger  
248 SEs for the well-sampled species when its sample size is downsized to that of  
249 the poorly sampled species. Appropriate  $(1 - \alpha)100\%$  confidence intervals for  
250 the “true” DNA barcode gap metrics are also easily constructed, where  $\alpha$  is the  
251 desired significance level cutoff (*e.g.*,  $\alpha = 0.05$  for 95% confidence).

252

### 253 **Case Study: *Agabus* Diving Beetles (Coleoptera: Dytiscidae)**

254 We use the predaceous diving beetle genus *Agabus* (Coleoptera:  
255 Dytiscidae), as a test case, which was previously examined by Bergsten *et al.*  
256 **(13)** in the context of assessing the scale of geographic sampling on DNA  
257 barcoding. *Agabus* consists of 199 named Linnaean species according to the  
258 Global Biodiversity Information Facility (GBIF). DNA sequences for COI-5P,

259 COI-3P, and CYTB were automatically downloaded from GenBank and BOLD,  
260 aligned to reference sequences for *Agabus bipustulatus*, and cleaned on  
261 November 1, 2022 with the R package *MACER* (Molecular Acquisition,  
262 Cleaning and Evaluation in R) using default parameters **(32)**. Data processing  
263 included removing sequences with ambiguous and/or missing nucleotide data  
264 (which was handled using pairwise deletion), omitting sequences containing  
265 sequencing artifacts such as stop codons, and excluding genus- and species  
266 level outliers (through use of the  $1.5 \times \text{IQR}$  (interquartile range) rule to  
267 discard taxon records having excessively divergent distance values). Aside  
268 from being well-represented within BOLD and GenBank, gene markers  
269 investigated were specifically selected due to their high representativeness in  
270 *Agabus*, widespread use in the DNA barcoding literature, as well as centrality  
271 in studies of phylogenetics and molecular evolution. All statistics were  
272 computed in R **(33)** using  $p$ -distances and integrated within *MACER*.  
273 Nonparametric bootstrap 95% percentile confidence intervals for the  
274 population means of  $p_x$ ,  $q_x$ ,  $p_x'$ , and  $q_x'$  across all species for each marker were  
275 calculated with 10000 replications where possible using the *boot* R package  
276 **(34, 35)**. In computing  $p_x'$ , and  $q_x'$ , it was not uncommon for focal taxa to be  
277 equally distant to multiple nearest neighbours based on the minimum  
278 interspecific distance. In such cases, ties were broken by employing the

279 species having the smallest mean interspecific distance among all nearest  
280 neighbours for a given target species. Scatterplots displaying  $\log_{10}$ -  
281 transformed probabilities for all examined species were generated using the  
282 *ggplot2* R package (36). This transformation was selected to aid overall  
283 visualization of the difference between  $p_x$  and  $q_x$  versus  $p_x'$  and  $q_x'$  by reducing  
284 any skewness inherent in the data. In cases where corresponding  $(p_x, q_x)$  and  
285  $(p_x', q_x')$  pairs were equal to zero on the untransformed scale, resulting in  
286 infinite values on the  $\log_{10}$  scale, observations were replaced by a large  
287 negative number (here, -5) prior to plotting. Following this, the number of  
288 species displaying zero values was then indicated.

289

## 290 **Statistical Interpretation, Notes and Caveats**

291 Calculated values of the proposed statistics and other important  
292 quantities can be found in **Table 2**.

293 A defining characteristic of metrics (1) and (2) is their asymmetric  
294 *directionality* with respect to intraspecific and interspecific distribution  
295 overlap (as well as those for target species and their closely aligned relatives).  
296 This is akin to measuring the Kullback-Leibler (KL) divergence (37) between  
297 two probability distributions  $P$  and  $Q$ : the distance between  $P$  and  $Q$  is not  
298 necessarily equal to the distance between  $Q$  and  $P$ . When no DNA barcode gap

299 is observed for a species, the extent of overlap will be different depending on  
300 whether one is looking at intraspecific or interspecific distances. Note,  $p_x$  and  
301  $q_x$  are informative in comparison to  $p_x'$  and  $q_x'$  because computed probabilities  
302 reveal the extent to which the distribution of nearest neighbours differs from  
303 the genus as a whole.

304         The newly described statistics can be employed in several ways. Firstly,  
305 the proposed quantities can be utilized to assess gene marker efficacy for  
306 specimen identification: markers attaining lower values of the proposed  
307 metrics are preferable as DNA barcoding loci to those displaying higher  
308 values, provided selected markers reflect relevant species' histories. Such  
309 behaviour would occur in the case where gene markers possess low rates of  
310 substitution and display no polymorphism within species; however, it will not  
311 solve the problem of incomplete lineage sorting or introgression. Secondly,  
312 through coalescent simulations, the defined metrics can be used to assess  
313 whether obtained values are consistent with a population-level process  
314 having a particular combination of  $N_e$  values, mutation rate ( $\mu$ ), dispersal rates  
315 ( $m$ ), and divergence times ( $\tau$ ) as seen in **Figures 1, 2, and Table 1**, and in a  
316 similar vein to Yang and Rannala (**27**). This agrees well with the universal  
317 observation that molecular identifications are significantly hampered when  
318 both intraspecific and interspecific distances are below a predefined arbitrary



319 threshold (*e.g.*, the 2% rule **(1, 2)** or the 10× rule **(38)**), which may occur  
320 given variable rates of molecular evolution in both taxa and gene markers.  
321 However, cautious interpretation should be exercised as these patterns are  
322 unlikely to hold strongly in practice for highly speciose groups showing recent  
323 adaptive diversification at large spatial scales. In fact, evolutionary processes  
324 such as random genetic drift and purifying selection occurring within genes  
325 commonly employed in studies of speciation effectively cancel out traces of  
326 intraspecific variation over short timeframes. However, such a pattern is less  
327 likely to be a problem for neutral mitochondrial loci, but exceptions have been  
328 described (*e.g.*, **(39)**).

329         In comparing results for *Agabus* across sequenced markers for  
330 examined taxa, several findings are noteworthy.

331 [**Table 2** near here]

332 Overall, results point to COI-5P being the most suitable gene marker for  
333 species delimitation using DNA barcodes in this group, followed by COI-3P.  
334 CYTB was found to perform poorly as a DNA barcode for diving beetles,  
335 whereas species coverage, albeit still poor, was 21.6% (43/199) and 18.1%  
336 (36/199) for COI-5P and COI-3P, respectively. Two equally important factors  
337 complicate accurate estimation of DNA barcode gaps, and therefore optimal  
338 selection of genomic markers for unambiguous specimen identification: the

339 number of sampled specimens per species in a genus, and the number of  
340 species in the genus. While the COI-5P dataset comprised more species, that  
341 for COI-3P contained a larger number of specimens, suggesting these two  
342 crucial elements are difficult to balance in tandem. This can be seen through  
343 examining the 95% bootstrap percentile intervals for  $\bar{p}_x$ ,  $\bar{q}_x$ ,  $\bar{p}_x'$ , and  $\bar{q}_x'$ ,  
344 which are all quite wide in most cases, meaning there is considerable  
345 uncertainty as to their true population means, given poor sampling of taxon  
346 genetic diversity. Results overall point to considerable overlap between target  
347 species of interest and all other taxa within a given genus. For example,  
348 computed mean probabilities for all loci in the  $q_x$ ,  $p_x'$ , and  $q_x'$  directions  
349 differed by an order of magnitude (**Table 2**). This finding suggests that the  
350 proposed metrics have sufficient discriminatory power to detect DNA barcode  
351 gaps. In comparing intraspecific and interspecific differences, values of  $p_x$   
352 were found to all be equal to one (COI-5P; 43 species; COI-3P: 36 species;  
353 CYTB: two species), indicating complete overlap of the intraspecific  
354 distributions with interspecific distributions, and hence no evidence of DNA  
355 barcode gaps across all three assessed genes. A probability of one will always  
356 occur whenever a minimum interspecific distance of zero is found. In contrast,  
357 much less, but still significant, overlap was noted in the  $q_x$  direction, ranging  
358 from  $5.191 \times 10^{-5}$  (7 species: *A. bifarius*, *A. biguttatus*, *A. brunneus*,

359 *A. conspersus*, *A. discolor*, *A. pallens*, and *A. setulosus*) to 0.1759 (*A. thomsoni*)  
360 for COI-5P, 0.0030 (4 species: *A. alexandrae*, *A. binotatus*, *A. clypealis*, and *A.*  
361 *faldermanni*) to 0.3175 (*A. ambiguus*) for COI-3P, and 0.0100 (*A. nevadensis*) to  
362 1 (*A. bipustulatus*) for CYTB, leading to similar conclusions regarding the  
363 presence (or lack thereof) of DNA barcode gaps; however, this depended  
364 strongly on the locus being considered (**Table 2; Figure 3**). In comparing  
365 target taxa to their closest conspecifics for all assessed loci, values of  $p_x'$  for  
366 COI-5P ranged from zero (21 species: *A. ajax*, *A. anthracinus*, *A. audeni*,  
367 *A. bicolor*, *A. bifarius*, *A. biguttatus*, *A. brunneus*, *A. colymbus*, *A. conspersus*,  
368 *A. crassipes*, *A. didymus*, *A. discolor*, *A. erichsoni*, *A. infuscatus*, *A. labiatus*,  
369 *A. moestus*, *A. nebulosus*, *A. pallens*, *A. sturmii*, *A. uliginosus* and *A. undulatus*) to  
370 one (eight species: *A. clavicornis*, *A. clypealis*, *A. congener*, *A. inscriptus*,  
371 *A. lapponicus*, *A. serricornis*, *A. setulosus*, and *A. thomsoni*), whereas those for  
372  $q_x'$  ranged from zero (21 species, same as above) to 0.3922 (*A. thomsoni*).  
373 Similarly, values of  $p_x'$  for COI-3P ranged from zero (20 species: *A. affinis*,  
374 *A. alexandrae*, *A. amoenus*, *A. aubei*, *A. bipustulatus*, *A. cephalotes*, *A. clypealis*,  
375 *A. didymus*, *A. disintegrates*, *A. faldermanni*, *A. fulvaster*, *A. lapponicus*,  
376 *A. melanarius*, *A. pseudoclypealis*, *A. serricornis*, *A. sturmii*, *A. tristis*,  
377 *A. uliginosus*, *A. undulatus*, and *A. unguicularis*) to one (10 species: *A. binotatus*,  
378 *A. brunneus*, *A. congener*, *A. glacialis*, *A. guttatus*, *A. heydeni*, *A. labiatus*,

379 *A. ramblae*, *A. rufulus*, and *A. zimmermanni*) and those for  $q_x'$  spanned zero (20  
380 species, same as above) to 0.2230 (*A. ambiguus*). In the case of CYTB, all  
381 values of  $p_x'$  and  $q_x'$  were identical to those of  $p_x$  and  $q_x$  since only two species  
382 were sequenced. Based on these findings, 21/43 (48.8%) species, 20/36  
383 (55.5%) species, and 0/2 (0%) show evidence of a DNA barcode gap for COI-  
384 5P, COI-3P, and CYTB, respectively. It is interesting to note that while  
385 calculated means of  $q_x$ ,  $p_x'$  and  $q_x'$  point to COI-5P being superior to all other  
386 examined genetic loci in terms of identification performance, since they attain  
387 the smallest values (as would be expected in animal DNA barcoding studies),  
388 the above results point to COI-3P instead. This reinforces the tradeoff between  
389 balancing the number of specimens sampled for a given species, and the  
390 number of species included within a target genus, as well as the fact that DNA  
391 barcoding has been one-sided. Results for  $p_x'$  and  $q_x'$  compared to  $p_x$  and  $q_x$   
392 make intuitive sense since the genus-level distribution of pairwise differences  
393 encompasses a mixed distribution comprising closely- and more distantly-  
394 related species. Findings outline here suggest the above species may warrant  
395 further DNA barcode scrutiny, due in part to *Agabus*' difficult morphology and  
396 widespread Holarctic range hampering successful classification in certain  
397 cases (**13**). For instance, a large degree of overlap between intraspecific and  
398 interspecific distributions (as well as those for target species and their closest

399 neighbours) could be evidence for a recent evolutionary origin, species  
400 hybridization, or mitochondrial introgression. Thus, DNA barcoding will fail to  
401 discriminate species displaying such patterns.

402 A potentially useful tool to visualize values of  $p'$  and  $q'$  computed for all  
403 species' neighbours across all assessed genetic markers is the heatmap, where  
404 colour intensity is directly proportional to the magnitude of calculated  
405 probabilities. In this way, generated heatmaps can be employed as "look-up  
406 tables" to better gauge intraspecific and interspecific distribution overlap for  
407 specific pairs of species of interest (and not just the closest neighbour) to a  
408 given study. Preliminary investigation has shown this type of visualization to  
409 be revealing, particularly when species are not well separated based on plots  
410 like **Figure 3**; thus, they will be included in future work.

411 It seems only one similar methodology for DNA barcode marker  
412 selection has been proposed before: the probability of correct identification  
413 (PCI) (**40, 41**). However, this statistic has not gained wide traction in the DNA  
414 barcoding community, appearing to be employed only for Fungi (**42**), and is  
415 further based purely on statistical, and not coalescent, theory (**12**). The PCI  
416 employs jackknife resampling of a binomially distributed random variable to  
417 obtain estimates of standard error and confidence intervals of specimen  
418 identification success with and without the influence of PCR failure (**12, 41**).

419 An approach that has seen a much larger body of research as a species  
420 delimitation tool for both sexually- and asexually-reproducing organisms is  
421 the so-called  $K/\theta$  ratio **(43, 44, 45, 46, 47, 48, 49)**. It measures average  
422 pairwise sequence differences between clade pairs ( $K$ ) *versus* said differences  
423 found within a clade ( $\theta$ ). If  $K/\theta \geq 4$  (the so-called “4× rule”), lineages belong to  
424 distinct species with at least 95% probability based on Rosenberg’s index of  
425 reciprocal monophyly **(50)**. Unlike the PCI, it appears  $K/\theta$  has not seen  
426 adoption within the DNA barcoding community-at-large, despite showing  
427 promise for DNA barcode gap detection **(12, 47)**. It is stressed that the  
428 approach here is simpler than the abovementioned ones and also overcomes  
429 the need for order statistics in the case of Phillips *et al.* **(12)**, as our metrics  
430 are easily summarized as simple arithmetic means with well-defined sampling  
431 distributions in the limit as sample sizes increase. However, indices discussed  
432 here are not the only ones possible, and others should be explored. Problems  
433 may arise when multiple curve intersection points or when missing/zero  
434 genetic distance values are present. Hence, more sophisticated techniques  
435 such as data interpolation, kernel smoothing, local regression, numerical  
436 integration, or data transformation may be required. Statistical frameworks  
437 like mixture models may also be worth exploring. While our method  
438 demonstrates that DNA barcoding has been a one-sided argument, it must be

439 stressed that obtained estimates likely do not mirror ground truth, since  
440 calculated probabilities are based on small taxon sample sizes, making  
441 pairwise distance distributions difficult to estimate in practice. A recent  
442 article by De Sanctis *et al.* (51) presents a coalescent theory simulation  
443 framework to examine accuracy of taxonomic binning in query sequences to  
444 references found within curated genomic databases. This approach may  
445 additionally prove informative in introducing greater theoretical rigor into  
446 DNA barcoding for molecular species identification purposes.

## 447 **Conclusion**

449         Here, we characterize the DNA barcode gap using the multispecies  
450 coalescent through proposing a suite of easily computed and interpreted  
451 nonparametric estimators inspired by population genetics theory along with  
452 observed trends in taxon DNA sequence diversity. Application to the beetle  
453 genus *Agabus* demonstrates the promise of the proposed statistical metrics  
454 for DNA barcode locus selection. The present approach, while having a strong  
455 basis in coalescent theory, is ideally suited to addressing applied research  
456 questions pertaining to DNA-based specimen identification, such as those  
457 encountered in studies of seafood fraud and invasive pest management using  
458 techniques like environmental DNA (eDNA)-based targeted species detection  
459 and metabarcoding, in addition to characterizing interspecific sequence  
460

461 diversity. While our metrics are simple, incorporating other coalescent  
462 approaches is worthwhile, as is research integrating sample size estimation  
463 with tools like *HACSim* (Haplotype Accumulation Curve Simulator) **(52)**.  
464 These are left for future studies.

465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496



497 **Author Contributions**

498

499 J.D.P., C.K.G., and R.G.Y. derived and coded the DNA barcode gap coalescent  
500 metrics, analyzed and interpreted the data, generated figures, and wrote the  
501 manuscript. N.H. and R.H.H. provided insight on DNA barcoding and other  
502 applications of the present work. All authors commented on and approved the  
503 final version.

504

505 **Data Accessibility**

506

507 FASTA files and R code can be found on GitHub at  
508 <https://github.com/jphill01/DNA-Barcode-Gap-Coalescent>.

509

510 **Acknowledgements**

511

512 We sincerely thank Rob DeSalle for the invitation to contribute to this  
513 important volume. We acknowledge that the University of Guelph resides on  
514 the ancestral lands of the Attawandaron people and the treaty lands and  
515 territory of the Mississaugas of the Credit. We recognize the significance of the  
516 Dish with One Spoon Covenant to this land and offer our respect to our  
517 Anishinaabe, Haudenosaunee and Métis neighbours as we strive to strengthen  
518 our relationships with them.

519

520

521

522

523

524

525

526

527

528

529

530

531

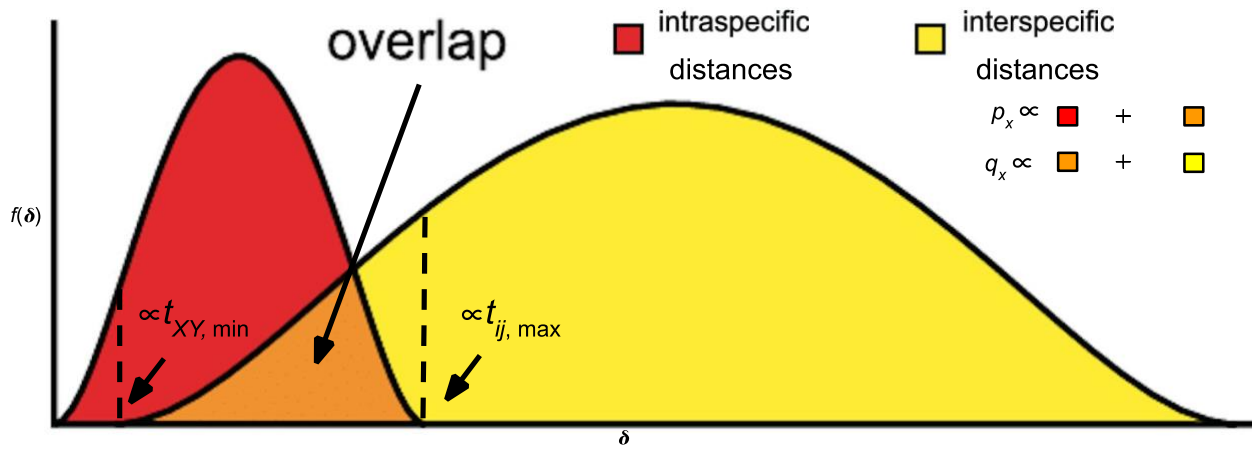
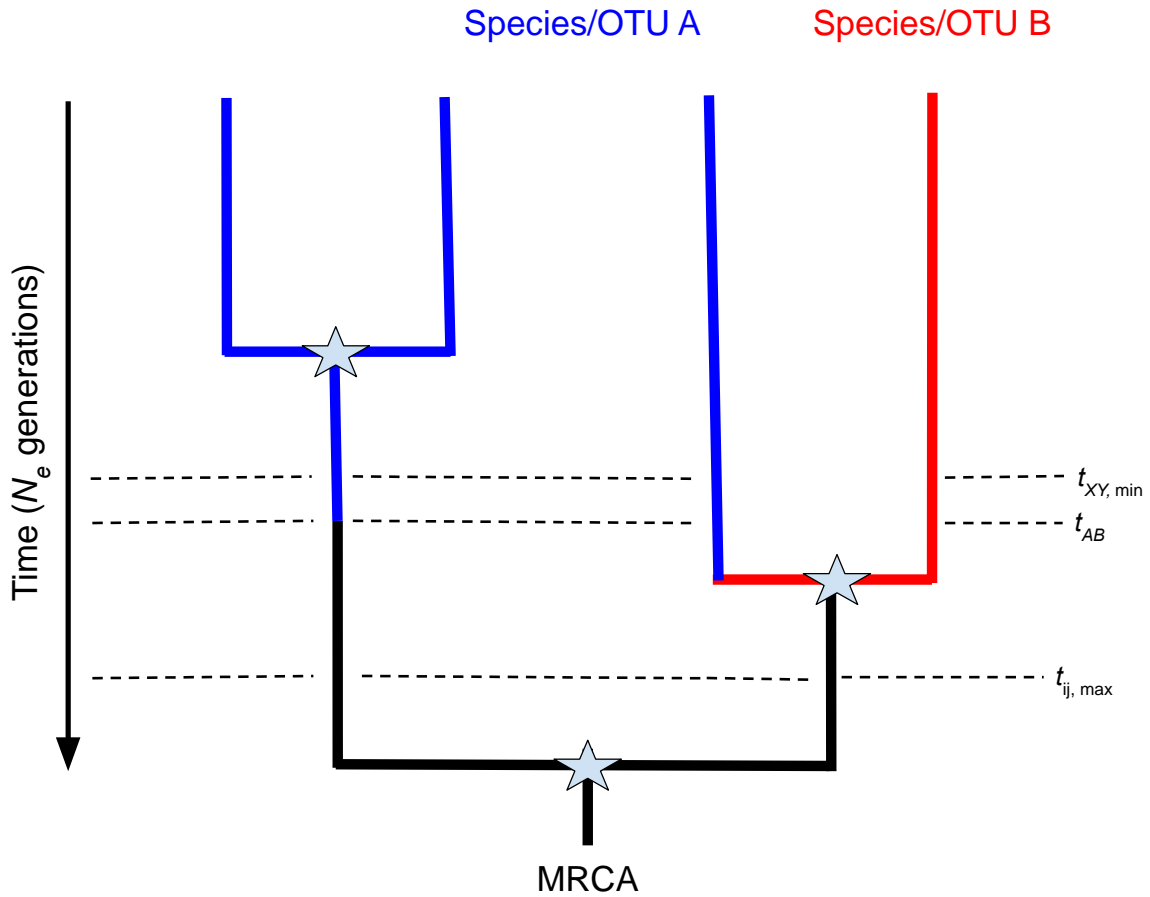
532

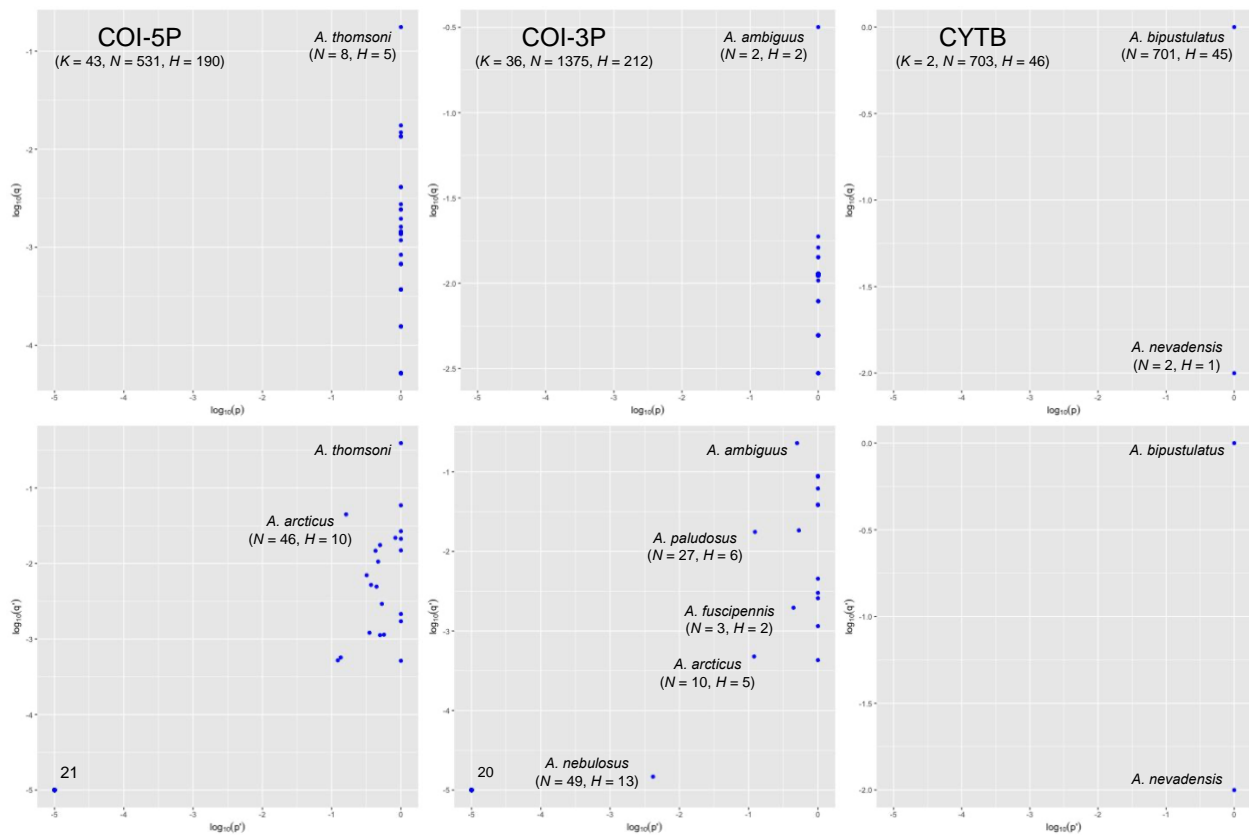
533

534

535

536





543

544

## 545 Figure Captions

546

547 **Figure 1.** Multispecies coalescent tree depicting a coalescence history for two  
 548 species/OTUs and four sampled individuals of a given genus backwards in  
 549 time to the Most Recent Common Ancestor (MRCA). A case of incomplete  
 550 lineage sorting is depicted, where species/OTU A is paraphyletic to  
 551 species/OTU B. Indicated divergence times are also shown.

552

553 **Figure 2.** Graphical depiction (modified from (7)) of the DNA barcode gap  
 554 metrics, for a single hypothetical species/OTU  $x$  with genetic distances  $\delta$ . Note  
 555 that distribution overlap (orange area) is implicitly accounted for in the  
 556 calculation of  $p_x$  and  $q_x$ . Similar visualizations can be generated for both  $p_x'$ , and  
 557  $q_x'$ .

558

559 **Figure 3.** Plots of  $p_x$  and  $q_x$  for all assessed *Agabus* species on the  $\log_{10}(x)$  scale  
 560 across three mitochondrial loci calculated using  $p$ -distance. **Top:** Intraspecific  
 561 vs. genus-level comparisons. **Bottom:** Target species vs. nearest neighbour  
 562 comparisons.

563

<b>Parameter</b>	<b>Definition</b>	<b>Range</b>
$K$	Number of identified species/OTUs within a genus of interest	$(0, \infty)$
$t_{XY, \min}$	Minimum divergence time within the genus or set of sister species	$(0, \infty)$
$t_{ij, \max}$	Maximum divergence time within the genus or set of sister species	$(0, \infty)$
$t_{AB}$	Divergence time between species/OTUs $A$ and $B$ within the genus	$(0, \infty)$
$\delta_x$	Pairwise genetic distances for all sampled specimens for species $x$ within the genus or set of sister species	$[0, 1]$
$f(\delta_x)$	Distribution of all pairwise genetic distances for all specimens of species $x$ within the genus or set of sister species	$[0, 1]$
$f(d_{ij, x})$	Distribution of intraspecific genetic distances for species $x$ within the genus or set of sister species	$[0, 1]$
$f(d_{XY})$	Distribution of interspecific genetic distances for the <b>entire</b> genus or set of sister species	$[0, 1]$
$p_x$	Proportional overlap of $f(d_{ij, x})$ with $f(d_{XY})$	$[0, 1]$
$q_x$	Proportional overlap of $f(d_{XY})$ with $f(d_{ij, x})$	$[0, 1]$
$p_{x'}$	Proportional overlap of $f(d_{ij, x'})$ with $f(d_{XY})$ for nearest neighbour species $x'$ within the genus	$[0, 1]$
$q_{x'}$	Proportional overlap of $f(d_{XY})$ with $f(d_{ij, x'})$ for nearest neighbour species $x'$ within the genus	$[0, 1]$

566

567

568

Marker	<i>L</i>	<i>N</i>	<i>H</i>	<i>K</i>	$\bar{p}$	$\bar{q}$ (95% CI)	$\bar{p}'$ (95% CI)	$\bar{q}'$ (95% CI)
COI-5P	600	531	190	43	1	0.0064 (0.0015, 0.0151)	0.3195 (0.2055, 0.4382)	0.0152 (0.0035, 0.0354)
COI-3P	600	1375	212	36	1	0.0180 (0.0085, 0.0360)	0.3256 (0.1868, 0.4745)	0.0165 (0.0051, 0.0325)
CYTB	342	703	46	2	1	0.5050 (0.0100, 1.0000)	1	0.5050 (0.0100, 1.0000)

569

### 570 Table Captions

571

572 **Table 1.** Coalescent DNA barcode gap model parameters, definitions and  
573 ranges. Ranges are indicated using interval notation, where parentheses and  
574 brackets signify open and closed intervals respectively.

575

576 **Table 2.** Calculated DNA barcoding gap statistics using *p*-distance for  
577 several sequenced mitochondrial loci in *Agabus* considering only species with  
578 at least two specimens. *L* is the alignment length in basepairs. *N* is the number  
579 of sequences used to compute the pairwise genetic distance matrix. *H* is the  
580 number of unique haplotypes.

581

582

### 583 References

584

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

588 2. Hebert, P.D.N., Ratnasingham, S., and deWaard, J.R. (2003b). Barcoding  
589 animal life: cytochrome *c* oxidase subunit 1 divergences among closely related  
590 species. *Proceedings of the Royal Society of London B (Suppl. 1)*, S96-S99.

591 3. Ballard, J.W.O. and Rand, D.M. (2005). The population biology of  
592 mitochondrial DNA and its phylogenetic implications. *Annual Reviews of*  
593 *Ecology, Evolution, and Systematics*, **36**: 621-642.

- 594  
595 4. Phillips, J.D., Gillis, D.J., and Hanner, R.H. (2019). Incomplete estimates of  
596 genetic diversity within species: Implications for DNA barcoding. *Ecology and*  
597 *Evolution*, **9**: 2996-3010.
- 598  
599 5. Ratnasingham S. and Hebert P.D.N. (2007) Bold: The barcode of life data  
600 system (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, **7**, 355–364.
- 601 6. Zhang, A.B., He, L.J., Crozier, R.H., Muster, C., and Zhu, C.-D. (2010).  
602 Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and*  
603 *Evolution*, **54**: 1035-1039.
- 604 7. Meyer C.P. and Paulay, G. (2005). DNA barcoding: Error rates based on  
605 comprehensive sampling. *PLOS Biology*, **3**: e422.
- 606 8. Meier, R., (2008). The use of mean instead of smallest interspecific distances  
607 exaggerates the size of the “barcoding gap” and leads to misidentification.  
608 *Systematic Biology*, **57**: 809-813.
- 609  
610 9. Dasmahapatra, K.K., Elias, M., Hill, R.I., Hoffman, J.I, and Mallet, J. (2010).  
611 Mitochondrial DNA barcoding detects some species that are real, and some  
612 that are not. *Molecular Ecology Resources*, **10**, 254-273.
- 613  
614 10. Hickerson, M.J., Meyer, C.P., and Moritz, C. (2006). DNA barcoding will  
615 often fail to discover new animal species in broad parameter space. *Systematic*  
616 *Biology*. **55**, 729–739. <sup>[1]</sup><sub>SEP</sub>
- 617 11. Stoeckle M.Y. and Thaler D.S. (2014). DNA barcoding works in practice but  
618 not in (neutral) theory. *PLOS ONE*, **9**, e100755.
- 619 12. Phillips, J.D., Gillis, D.J., and Hanner, R.H. (2022). Lack of statistical rigor in  
620 DNA barcoding likely invalidates the presence of a true species’ barcode gap.  
621 *Frontiers in Ecology and Evolution*, **10**: 859099.
- 622  
623 13. Bergsten, J., Bilton, D.T., Fujisawa, T., Elliott, M., Monaghan, M.T., Balke, M.,  
624 Hendrich, L., Geijer, J., Herrmann, J., Foster, G.N., Ribera, I., Nilsson, A.N.,  
625 Barraclough, T.G., and Vogler, A.P. (2012). The effect of geographical scale of  
626 sampling on DNA barcoding. *Systematic Biology*, **61**: 851-869.
- 627  
628 14. Matz, M. and Nielsen, R. (2005). A likelihood ratio test for species  
629 membership based on DNA sequence data. *Philosophical Transactions of the*  
630 *Royal Society of London B: Biological Sciences*, **360**: 1969-1974.

- 631  
632 15. Nielsen, R. and Matz, M. (2006). Statistical approaches for DNA barcoding.  
633 *Systematic Biology*, **55**: 162-169.  
634
- 635 16. Pons, j., Barraclough, T., Gomez-Zurita, J., Cardoso, A., Hazell, S., Kamoun, S.,  
636 Sumlin, W.D., and Vogler, A.P. (2006). Sequence-based species delimitation for  
637 the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**: 585-609.  
638
- 639 17. Puillandre, N., Lambert, A., Brouillet, S., and Achez, G. (2011). ABGD,  
640 Automatic Barcode Gap Discovery for species delimitation. *Molecular Ecology*,  
641 **21**: 1864-1877.  
642
- 643 18. Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, P. (2013). A general species  
644 delimitation method with applications to phylogenetic placements.  
645 *Bioinformatics*, **29**: 2869-2876.  
646
- 647 19. Ratnasingham, S. and Hebert, PDN (2013) A DNA-based registry for all  
648 animal species: The Barcode Index Number (BIN) system. *PLOS ONE*, **8**,  
649 e66213.
- 650 20. Puillandre, N. Brouillet, S. and Achez, G. (2021). ASAP: assemble species by  
651 automatic partitioning. *Molecular Ecology Resources*, **21**: 609-622.  
652
- 653 21. Ezard, T., Fujisawa, T., and Barraclough, T. (2017). splits: Species Limits by  
654 Threshold Statistics. R package version 1.0.  
655
- 656 22. Eckert, E.M., Fontaneto, D., Coci, M., and Callieri, C. (2015). Does a  
657 barcoding gap exist in prokaryotes? Evidences from species delimitation in  
658 cyanobacteria. *Life*, **5**: 50-64.  
659
- 660 23. Zimmerman, J., Jahn, R., and Gemeinholzer, B. (2011). Barcoding diatoms:  
661 evaluation of the V4 subregion on the 18S rRNA gene, including new primers  
662 and protocols. *Organisms Diversity and Evolution*, **11**: 1-20.  
663
- 664 24. Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their*  
665 *Applications*, **13**: 235-248.  
666
- 667 25. Hubert, N. and Hanner, R. (2015) DNA barcoding, species delineation and  
668 taxonomy: a historical perspective. *DNA Barcodes*, **3**: 44-58.

- 669 26. Rannala B and Yang Z. (2003). Bayes estimation of species divergence  
670 times and ancestral population sizes using DNA sequences from multiple loci.  
671 *Genetics*. **164**: 1645–1656.
- 672
- 673 27. Yang, Z. and Rannala, B. (2017). Bayesian species identification under the  
674 multispecies coalescent provides significant improvements to DNA barcoding  
675 analyses. *Molecular Ecology*, **26**: 3028-3036.
- 676
- 677 28. Collins, R.A. and Cruickshank. R.H. (2014). Known knowns, known  
678 unknowns, unknown unknowns and unknown knowns in DNA barcoding: a  
679 comment on Downton et al. *Systematic Biology*, **63**: 1005-1009.
- 680 29. Jukes, T.H. and Cantor, C.R. (1969) Evolution of Protein Molecules. In:  
681 Munro, H.N., Ed., Mammalian Protein Metabolism, Academic Press, New York,  
682 21-132.
- 683
- 684 30. Kimura, M. (1980). A simple method for estimating evolutionary rates of  
685 base substitutions through comparative studies of nucleotide sequences.  
686 *Journal of Molecular Evolution*, **16**: 111-120.
- 687 31. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The*  
688 *Annals of Statistics*, **7**: 1-26.
- 689
- 690 32. Young, RG, Gill, R, Gillis, D., and Hanner, RH (2021) Molecular Acquisition,  
691 Cleaning and Evaluation in R (MACER) – A tool to assemble molecular marker  
692 datasets from BOLD and GenBank. *Biodiversity Data Journal*, **9**, e71378.
- 693 33. R Core Team (2022). R: A language and environment for statistical  
694 computing. R Foundation for Statistical Computing, Vienna, Austria. URL  
695 <https://www.R-project.org/>.
- 696 34. Canty, A. and Ripley, B. (2021). boot: Bootstrap R (S-Plus) Functions. R  
697 package version 1.3-28.
- 698 35. Davison, A. C. and Hinkley, D. V. (1997). Bootstrap Methods and Their  
699 Applications. Cambridge University Press, Cambridge.
- 700
- 701 36. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag  
702 New York, 2016.



- 703 37. Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals*  
704 *of Mathematical Statistics*, 22: 49-86.  
705
- 706 38. Hebert, P.D.N., Stoeckle, M.Y., Zemplak, T.S., and Francis, C.M. (2004).  
707 Identification of Birds through DNA barcodes. *PLOS Biology*, 2: e312.
- 708 39. D'Ercole, J., Dapporto, L., Schmidt, B.C., Dincă, V., Talavera, G., Vila, R., and  
709 Hebert, P.D.N. (2022). Patterns of DNA barcode diversity in butterfly species  
710 (Lepidoptera) introduced to the Nearctic. *European Journal of Entomology*,  
711 119: 379-387.
- 712
- 713 40. Martin, M.P., Daniëls, P.P, Erickson, D., and Spouge, J.L. (2020). Figures of  
714 merit and statistics for detecting faulty species identification with DNA  
715 barcodes: a case study in *Ramaria* and related fungal genera. *PLOS ONE*, 15:  
716 e0237507  
717
- 718 41. Spouge, J. and Mariño-Ramirez, L. (2012). The practical evaluation of DNA  
719 barcode efficacy. In: *DNA Barcodes: Methods and Protocols*. W.J. Kress and D.L  
720 Erickson, eds. *Springer*.  
721
- 722 42. Suwannasai, N., Martin, M.P., Phosri, C., Sihanonth, P., Whalley, A.J.S., and  
723 Spouge, J.L. (2013). Fungi in Thailand: A case study of the efficacy of an ITS  
724 barcode for automatically identifying species within the *Annulohyphomyces* and  
725 *Hyphomyces* genera. *PLOS ONE*, 8: e54529.  
726
- 727 43. Birky, C.W.J, Wolf, C., Maughan, H., Herbertson, L., and Henry, E. (2005).  
728 Speciation and selection without sex. *Hydrobiologia* 546: 29–45.  
729
- 730 44. Birky C.W.J. and Barraclough T.G. (2009). Asexual Speciation. In: Schon, I.,  
731 Martens, K., van Dijk, P., eds. *Lost Sex*. New York: Springer. pp. 201–216.  
732
- 733 45. Birky, C.W.J., Adams, J., Gemmel, M., and Perry, J. (2010). Using population  
734 genetic theory and DNA sequences for species detection and identification in  
735 asexual organisms. *PLOS ONE*, 5: e10609.  
736
- 737 46. Birky, C.W.J., Ricci, C., Melone, G., and Fontaneto, D. (2011). Integrating  
738 DNA and morphological taxonomy to describe diversity in poorly studied  
739 microscopic animals: new species of the genus *Abrochtha* Bryce, 1910  
740 (Rotifera: Bdelloidea: Philodinavidae). *Zoological Journal of the Linnean*  
741 *Society*, 161: 723–734.

- 742  
743 47. Birky, C.W.J. (2013) Species detection and identification in sexual  
744 organisms using population genetic theory and DNA sequences. *PLOS ONE*, **8**:  
745 e52544.
- 746 48. Birky, C. W.J. and H. Maughan (2020). Evolutionary genetic species  
747 detected in prokaryotes by applying the  $K/\theta$  ratio to DNA sequences. *bioRxiv*.  
748 <https://www.biorxiv.org/content/10.1101/2020.04.27.062828v3.full>
- 749
- 750
- 751 49. Spöri, Y. Stoch, F., Dellicour, S., Birky, C.W.J., and Flot, J.-F. (2021). KoT: an  
752 automatic implementation of the  $K/\theta$  method for species delimitation. *bioRxiv*.  
753 <https://www.biorxiv.org/content/10.1101/2021.08.17.454531v2>
- 754 50. Rosenberg, N. (2007). Statistical tests for taxonomic distinctiveness from  
755 observation of monophyly. *Evolution*, **61**: 317-323.
- 756 51. De Sanctis, B., Money, D., Winther Pedersen, M., and Durbin, R. (2021). A  
757 theoretical analysis of taxonomic binning accuracy. *Molecular Ecology*  
758 *Resources*, **22**: 2208-2219.
- 759
- 760 52. Phillips, J.D., French, S.H., Hanner, R.H., and Gillis, D.J. (2020). HACSim: An  
761 R package to estimate intraspecific sample sizes for genetic diversity  
762 assessment using haplotype accumulation curves. *PeerJ Computer Science*, **6**:  
763 1-37.