Longer is not always better: Optimizing barcode length for large-scale species discovery and identification

Running Title: mini-barcodes for species discovery

Darren Yeo[1], Amrita Srivathsan[1], Rudolf Meier[1*]

[1]Department of Biological Sciences, National University of Singapore, 14 Science 8 Drive 4, Singapore 117543

[*]Corresponding author: meier@nus.edu.sg

Abstract

New techniques for the species-level sorting of millions of specimens have to be developed in order to answer the question of how many species live on earth. These methods should be reliable, scalable, and cost-effective as well as largely insensitive to the low-quality genomic DNA commonly obtained from museum specimens. Mini-barcodes seem to satisfy these criteria, but it is unclear whether they are sufficiently informative for species-level sorting. This is here tested based on 20 datasets covering ca, 30,000 specimens of 5,500 species. All specimens were first sorted based on morphology before being barcoded with full-length *cox1* barcodes. Mini-barcodes of different lengths and positions were then obtained *in silico* from the full-length barcodes using nine published mini-barcode primers (length: 94 – 407-bp) and a sliding window approach (3 windows: 100-bp, 200-bp, 300-bp). Afterwards, we determined whether barcode length and/or position reduces congruence between morphospecies and molecular Operational Taxonomic Units (mOTUs) that were obtained using three different species delimitation techniques (ABGD, PTP, objective clustering). We also evaluate how useful the published mini-barcodes are for species identification with the “best close match” algorithm. We find that there is no significant difference in performance for both species delimitation and identification between full-length and mini-barcodes as long as they are of moderate length (>200-bp). Only very short mini-barcodes (<200-bp) perform poorly, especially when they are located near the 5’ end of the Folmer region. Overall, congruence between morphospecies and mOTUs is ca. 80% for barcodes that are >200-bp. The congruent mOTUs contain ca. 75% of the specimens and we estimate that most of the conflict is caused by ca. 10% of the specimens that should be targeted for re-examination in order to resolve conflict efficiently. Overall, barcode length (>200-bp)

and species delimitation methods have minor effects on congruence. Our study suggests that large-scale species discovery, identification and metabarcoding can utilize mini-barcodes without substantial loss of information compared to full-length barcodes. This is good news given that mini-barcodes can be obtained via cost-effective tagged amplicon sequencing using short-read sequencing platforms (Illumina: “NGS barcodes”).

Keywords: DNA barcoding, mini-barcodes, species discovery, metabarcoding

Introduction

The question of how many species live on earth has intrigued biologists for a very long time, but we are nowhere close to having a robust answer. We do know that fewer than 2 million of the estimated 10-100 million multicellular species have been described (Roskov et al. 2018) and that many are currently being extirpated by the “sixth mass extinction” (Ceballos et al. 2015; Sánchez-Bayo & Wyckhuys 2019) with potentially catastrophic consequences for the environment (Cafaro 2015). Monitoring, halting, and perhaps even reversing this process is hampered by the “taxonomic impediment”. This impediment is particularly severe for “invertebrates” that collectively contribute much of the animal biomass (e.g., arthropods, annelids, nematodes: Stork et al. 2015; Bar-On et al. 2018). Most biologists thus agree that there is a pressing need for accelerating species discovery and description. This is very likely to require the development of new molecular methods. We would argue that they should not only be accurate, but also (1) rapid, (2) cost-effective, and (3) and largely insensitive to DNA quality. These criteria are important because tackling the earth’s biodiversity would likely require the processing of >500 million specimens even under the very conservative assumption that there are only 10 million species and a new species is discovered with every 50 specimens processed. Cost-effectiveness is similarly important because millions of species are found in countries with only basic research facilities. Fortunately, many species are already represented in museum holdings, but such specimens often yield degraded DNA (Cooper 1994). Therefore, methods that require DNA of high-quality and quantity are not likely to be suitable for large-scale species discovery in invertebrates.

Conceptually, species discovery and description can be broken up into three steps. The first is obtaining specimens, the second, species-level sorting, and the

92 third, species identification or description. Fortunately, centuries of collecting have
93 generated many of the specimens that are needed for large-scale species discovery.
94 Indeed, for many invertebrate groups it is likely that the museum collections contain
95 more specimens of undescribed than described species; i.e., this unsorted collection
96 material represents vast and still underutilized source for species discovery (Lister &
97 Climate Change Research Group 2011; Kemp 2015; Yeates et al. 2016). The
98 second step in species discovery/description is species-level sorting. This step is in
99 dire need for acceleration. Traditionally, it starts with the sorting of unsorted material
100 into major taxa (e.g., order-level in insects). This task can still be accomplished by
101 parataxonomists but may in the future be taken over by machines utilizing neural
102 networks (Valan et al. 2019). In contrast, the subsequent species-level sorting is
103 usually time-limiting because the specimens for many invertebrate taxa have to be
104 prepared and studied by highly-skilled specialists (e.g., dissected; slide-mounted)
105 before the material can be sorted into putative species. This means that the
106 traditional techniques are neither rapid nor cost-effective for many invertebrate
107 groups. This impediment is likely to be responsible for why certain taxa that are
108 known to be abundant and species-rich are particularly poorly studied (Bickel 1999).

109 An alternative way to sort specimens to species-level would be with DNA
110 sequences. This approach is particularly promising for metazoan species because
111 most multicellular animal species can be distinguished based on cytochrome c
112 oxidase subunit I (*cox1*) barcode sequences (Hebert et al. 2003). However, such
113 sorting requires that each specimen is barcoded. This creates cost- and scalability
114 problems when the barcodes are obtained with Sanger sequencing (see Taylor and
115 Harris 2012). Such sequencing is currently still the standard in barcoding studies
116 because the animal barcode was defined as a 658-bp long fragment of *cox1*

117 (“Folmer region”: Folmer et al. 1994), although sequences >500-bp with <1%
118 ambiguous bases are also considered BOLD-compliant (BOLDsystems.org). The
119 658-bp barcode was optimized for ABI capillary sequencers but it has become a
120 burden because it is not suitable for most new sequencing technologies, while
121 Sanger sequencing remains expensive and only scalable when expensive liquid-
122 handling robots are used. This approach is hence unlikely to become widely
123 available in those countries that harbour most of the species diversity. Due to these
124 constraints, very few studies have utilized DNA barcodes to sort entire samples into
125 putative species (but see Fagan-Jeffries et al. 2018). Instead, most studies use a
126 mixed approach where species-level sorting is carried out based on morphology
127 before a select few specimens per morphospecies are barcoded (e.g., Riedel et al.
128 2010). This two-step process requires considerable amounts of skilled manpower
129 and time.

130 Scalability and cost-effectiveness are hallmark features of the new short-read
131 high throughput sequencing technologies. In addition, these technologies are
132 particularly suitable for sequencing the kind of degraded DNA that is typical for
133 museum specimens. Indeed, anchored hybrid enrichment (AHE) has already been
134 optimized for the use with old museum specimens (Bi et al. 2013; Guschanski et al.
135 2013; Blaimer et al. 2016) and is likely to play a major role for the integration of rare
136 species into taxonomic and systematic projects. It will be difficult, however, to apply
137 AHE to millions of specimens because it requires time-consuming and expensive
138 molecular protocols (e.g., specimen-specific libraries). Fortunately, for most species
139 it is likely that species-level sorting does not require a very large number of markers.
140 We would thus argue that the initial species-level sorting can be achieved using
141 barcodes that can be obtained via “tagged amplicon sequencing” on next-generation

142 sequencing platforms (“NGS barcodes”; Wang et al. 2018; Yeo et al. 2018). Until
143 recently, obtaining full-length NGS barcodes via tagged amplicon sequencing was
144 difficult because the reads of most next-generation-sequencing platforms were too
145 short for sequencing the full-length barcode. This has now changed with the arrival
146 of third-generation platforms (ONT: MinION: Srivathsan et al. 2018; PacBio: Sequel:
147 Hebert et al. 2018). These platforms, however, come with drawbacks; *viz* elevated
148 sequencing error rates and higher cost. These problems are likely to be overcome in
149 the future (e.g., Yang et al. 2018), but such solutions would not solve the main
150 challenge posed by full-length barcodes; i.e., the problems with reliably obtaining
151 amplicons from museum specimens with degraded DNA (e.g., Hajibabaei et al.
152 2006). We thus submit that one should optimize the barcode length based on
153 empirical evidence; it should be only as long as needed for accurate pre-sorting of
154 specimens into putative species.

155 Barcodes that are shorter than the full-length barcode are called mini-
156 barcodes. They are obtained with primers that amplify shorter subsets of the original
157 barcode region and have several advantages. Firstly, such amplicons are easier to
158 obtain when the DNA in the sample is degraded (Hajibabaei & McKenna 2012).
159 Secondly, mini-barcodes can be sequenced at low cost using tagged amplicon
160 sequencing on short-read sequencing platforms (e.g., Illumina). Thirdly, mini-
161 barcode primers are available for a large number of species-rich metazoan clades
162 (Hajibabaei et al. 2006; Meusnier et al. 2008; Hebert et al. 2013; Little 2014) as well
163 as for specific taxa such as fruit flies, catfish and sharks (Fan et al. 2009;
164 Bhattacharjee & Ghosh 2014; Fields et al. 2015). It is thus not surprising that short
165 barcodes are already the barcodes of choice when the template DNA is degraded.
166 This is often the case for museum specimens (Zuccon et al. 2012; Hebert et al.

167 2013) or for environmental DNA which is usually analysed via metabarcoding (e.g.:
168 processed food: Armani et al. 2015; Shokralla et al. 2015; water, soil, fecal matter:
169 Epp et al. 2012; Srivathsan et al. 2015; Lim et al. 2016). Mini-barcodes were initially
170 obtained via Sanger sequencing, but they can now be sequenced much more
171 efficiently via tagged amplicon sequencing on short-read platforms ("NGS
172 barcoding": Wang et al. 2018: sequencing cost < 4 cents). Wang et al. (2018) could
173 thus implement a "reverse workflow" based on sequencing all specimens without any
174 species-level pre-sorting based on morphology. Four-thousand specimens of ants
175 were barcoded with a 313-bp mini-barcode. The specimens were then pre-sorted
176 into 89 molecular operational taxonomic units (mOTUs) that were largely congruent
177 with morphospecies (86 species). However, it remained unclear whether full-length
178 DNA barcodes would have further improved congruence, whether the results from
179 this study can be generalized, and which mini-barcode is optimal for large-scale pre-
180 sorting of specimens into putative species.

181 The answers to these questions are elusive, because mini-barcodes remain
182 insufficiently tested despite their ubiquitous use in metabarcoding. Arguably, existing
183 tests suffer from lack of scale (the largest study includes 6695 barcodes for 1587
184 species: Meusnier et al. 2008) and taxonomic scope (usually only 1-2 family-level
185 taxa: e.g. Hajibabaei et al. 2006; Yu & You 2010). Furthermore, the tests yielded
186 conflicting results. Hajibabaei et al. (2006) found high congruence with the full-length
187 barcode when species are delimited based on mini-barcodes and Meusnier et al.
188 (2008) find similar BLAST identification rates for mini-barcodes and full-length
189 barcodes in their *in silico* tests. However, Yu & You (2010) conceded that mini-
190 barcodes may have worse accuracy despite having close structural concordance
191 with the full-length barcode. In addition, Sultana et al. (2018) concluded that the

192 ability to identify species is compromised when the barcodes are too short (<150-bp),
193 but it remained unclear at which length and in which position mini-barcodes stop
194 performing poorly. Furthermore, published tests of mini-barcodes compare their
195 performance to results obtained with full-length barcodes. All conflict is then implicitly
196 considered evidence for the failure of mini-barcodes. However, results obtained with
197 longer barcodes should not automatically be assumed to be accurate given that the
198 Folmer region varies in nucleotide variability (Roe & Sperling 2007). Lastly, the
199 existing tests of mini-barcodes do not include a sufficiently large number of different
200 mini-barcodes in order to be able to detect positional and lengths effects across the
201 658-bp barcode region.

202 Here, we address the lack of scale by including 20 studies covering 5500
203 species represented by ca. 30,000 barcodes. We furthermore test a large number of
204 different mini-barcodes by applying a sliding window approach to generate mini-
205 barcodes of different lengths (100, 200, 300-bp window length, 60-bp intervals) and
206 compare the results to the performance of nine mini-barcodes with published primers
207 (mini-barcode length: 94 – 407-bp). The taxonomic scope of our study is broad and
208 includes a wide variety of metazoans ranging from earthworms to butterflies and
209 birds. Lastly, we do not assume that mOTUs based on full-length barcodes are
210 automatically more accurate than those obtained with mini-barcodes. Instead, we
211 use morphology as an external criterion for assessing whether mOTUs obtained with
212 different-length barcodes have different levels of congruence with morphospecies.
213 Note that this does not imply that morphology is more suitable for species
214 delimitation than molecular data. Instead, we test whether shortening barcodes
215 influences congruence with morphology; i.e. morphology is treated as a constant
216 while testing whether barcode length and/or position influences the number of

217 morphospecies that are recovered. Given that morphology is a generally accepted
218 type of data that can be used for species delimitation, mini-barcodes that significantly
219 lower congruence with morphospecies are unlikely to be useful for accurate species-
220 level sorting.

221 We also compare the performance of different species delimitation methods.
222 There has been substantial interest in developing algorithms for mOTU estimation,
223 leading to the emergence of various species delimitation algorithms over the past
224 decade (e.g., objective clustering: Meier et al. 2006; BPP: Yang & Rannala 2010;
225 jmOTU: Jones et al. 2011; ABGD; Puillandre et al. 2012; BINs: Ratnasingham &
226 Hebert 2013; PTP: Zhang et al. 2013; etc.). For the purposes of this study, we
227 selected three algorithms as representatives of distance and tree-based methods:
228 objective clustering, Automatic Barcode Gap Discovery (ABGD) and Poisson Tree
229 Process (PTP). Objective clustering utilizes an *a priori* distance threshold to group
230 sequences into clusters, ABGD groups sequences into clusters based on an initial
231 prior and recursively uses incremental priors to find stable partitions, while PTP
232 utilizes the branch lengths on the input phylogeny to delimit species units. Arguably,
233 barcode data may not be appropriate for the application of PTP because a single
234 marker is not likely to yield reliable phylogenetic trees (including branch lengths), but
235 PTP has been frequently applied to barcode data in the literature (e.g. Ermakov et al.
236 2015; Han et al. 2016; Hollatz et al. 2016) and is thus included here. There are
237 numerous additional techniques for species delimitation, but most require multiple
238 markers and/or are even more reliant on accurately reconstructed phylogenetic
239 trees. They may thus not easily scale to millions of specimens and are excluded
240 here.

241 Lastly, we examine the utility of mini-barcodes for species identification. This
242 is the original use of DNA barcodes (Hebert et al. 2003). The widespread use of the
243 metazoan *cox1* Folmer region barcode greatly contributed to the usefulness of DNA
244 barcodes for this purpose, with the value of barcodes increasing as the databases
245 grow. While not as relevant to large-scale species discovery, species identification
246 with barcodes is particularly valuable for targeted species detection and community
247 characterization in metabarcoding studies (Ficetola et al. 2008; Lim et al. 2016;
248 Morinière et al. 2019). Most of these studies tend to involve poor quality genetic
249 material (i.e., gut content, faecal matter, environmental DNA) and hence it would be
250 valuable to know whether short markers can reliably assign species identities to an
251 unknown sequence. We here test the performance of the barcodes in our test
252 datasets using "best close match", which considers those sequences correctly
253 identified when they match conspecific barcodes within a predefined threshold
254 (Meier et al. 2006).

255

256 **Materials & Methods**

257 *Dataset selection*

258 We surveyed the barcoding literature in order to identify publications that cited
259 the original barcode paper by Hebert et al. (2003) and met the following criteria: 1)
260 have pre-identified specimens where the barcoded specimens were pre-
261 sorted/identified based on morphology and 2) the dataset had at least 500
262 specimens with *cox1* barcodes >656-bp. We identified 20 most recent datasets
263 dating backwards from 2017 (Table S1); all had >500 barcoded specimens even
264 after removing those that were not sorted to species level (e.g., only identified to

265 genus or higher) or had short sequences <657-bp (the full-length barcode is
266 technically 658-bp long, but a 1-bp concession was made to prevent the loss of too
267 much data). The barcode sequences were downloaded from BOLDSystems or NCBI
268 GenBank and aligned with MAFFT v7 (Katoh & Standley 2013) with a gap opening
269 penalty of 5.0.

270 Using a custom python script, we generated three sets of mini-barcodes along
271 a "sliding window". They were of 100-, 200- and 300-bp lengths. The first iteration
272 begins with the first base pair of the 658-bp barcode and the shifting windows jump
273 60-bp at each iteration, generating ten 100-bp windows, eight 200-bp windows and
274 six 300-bp windows. Additionally, we identified nine mini-barcodes with published
275 primers within the *cox1* Folmer region (Fig. 1 & Table S2). These mini-barcodes
276 have been repeatedly used in the literature and were used for a broad range of taxa.
277 The primers for the various mini-barcodes were aligned to the homologous regions
278 of each dataset with MAFFT v7 --addfragments (Katoh & Standley 2013) in order to
279 identify the precise position of the mini-barcodes within the full-length barcode. The
280 mini-barcode subsets from each barcode were then identified after alignment to full-
281 length barcodes to mark the start and end-point of the amplicon. Note that most of
282 the published primers are in the 5' prime region of the full-length barcode.

283 *Species delimitation*

284 The mini-barcodes and the full-length barcodes were clustered into putative
285 species using three species delimitation algorithms: objective clustering (Meier et al.
286 2006), ABGD (Puillandre et al. 2012) and PTP (Zhang et al. 2013). For objective
287 clustering, the mOTUs were clustered at 2 – 4% uncorrected p-distance thresholds
288 (Srivathsan & Meier 2012) using a python script which reimplements the objective

clustering of Meier et al. 2006 and allows for batch processing. The p-distance thresholds selected are the typical distance thresholds used for species delimitation in the literature (Meier et al. 2006; Ratnasingham & Hebert 2013; Meier et al. 2016). The same datasets were also clustered with ABGD (Puillandre et al. 2012) using the default range of priors and with uncorrected p-distances, but the minimum slope parameter (-X) was reduced in a stepwise manner (1.5, 1.0, 0.5, 0.1) if the algorithm could not find a partition. We then considered the ABGD clusters at priors P=0.001, P=0.01 and P=0.04. The priors (P) refer to the maximum intraspecific divergence and function similarly to p-distance thresholds at the first iteration, before being refined by recursive application of the ABGD algorithm. Lastly, in order to use PTP, the datasets were used to generate maximum likelihood (ML) trees in RAxML v.8 (Stamatakis 2014) via rapid bootstrapping (-f a) and the GTRCAT model. The best tree generated for each dataset was then used for species delimitation with PTP (Zhang et al. 2013) under default parameters.

Performance assessment

We assess the performance of mini-barcodes by using morphospecies as an external arbiter. Species-level congruence was quantified using match ratios between molecular and morphological groups (Ahrens et al. 2016). The ratio is defined as $\frac{2 \times N_{match}}{N_1 + N_2} \times 100$, where N_{match} is the number of clusters identical across both mOTU delimitation methods/thresholds (N_1 & N_2). Incongruence between morphospecies and mOTUs is usually caused by a few specimens that are assigned to the “incorrect” mOTUs. Conflict at the specimen-level can thus be quantified as the number of specimens that are in mOTUs that cause conflict with morphospecies.

312 In order to test whether barcode length is a significant predictor of congruence,
313 MANOVA tests were carried out in R (R Core Team 2017) with “match ratio”
314 (species-level congruence) as the response variable and “dataset” and “mini-
315 barcode” as categorical explanatory variables. We found that most of the variance in
316 our study was generated by the “dataset” variable ($P < 0.05$ in MANOVA tests).
317 Given that we were particularly interested in the effect of barcode length and
318 position, “dataset” was subsequently treated as a random effect “mini-barcode” as
319 the explanatory variable (categorical) in a linear mixed effects model (R package
320 *lme4*: Bates et al. 2015). The *emmeans* R package (Lenth 2018) was then used to
321 perform pairwise post-hoc Tukey tests between mini- and full-length barcodes so as
322 to assess whether either barcode was performing significantly better/worse. To
323 compare the differences in performance between objective clustering, ABGD and
324 PTP, ANOVA tests were performed in R. After which, pairwise Tukey tests were
325 used to determine which species delimitation method was responsible for significant
326 differences. Lastly, in order to explore the reasons for positional effects, the
327 proportion of conserved sites for each mini-barcode was obtained using MEGA6
328 (Tamura et al. 2013).

329 Match ratios indicate congruence at the species level, but it is also important
330 to determine how many specimens have been placed in congruent units. Species-
331 and specimen-level congruence are only identical when all mOTUs are represented
332 by the same number of specimens. However, specimen abundances are rarely equal
333 across species and hence match ratio is insufficient at characterizing congruence
334 between mOTUs and morphospecies. It is straightforward to determine the number
335 of congruent specimens as follows:

(1) Congruence Class I specimens: If $\boldsymbol{A} = \boldsymbol{B}$ then number of congruent specimens is $Nc_1 = |A| \text{ OR } |B|$.

Incongruence is caused by morphospecies that are split, lumped, or split and lumped in the mOTUs. This means that a single mis-sorted specimen placed into a large-sized mOTU leads to all specimens in two mOTUs to be considered "incongruent" according to the criterion outlined above. Yet, most specimens are congruent and full congruence could be restored by re-examining the mis-sorted specimen. It is therefore also desirable to determine the number of specimens that require re-examination or, conversely, the number of specimens that would be congruent if one were to remove a few incongruent specimens. This number of specimens can be estimated by counting congruent specimens as follows:

(2) Congruence Class II specimens: Specimens that are in split or lumped mOTUs relative to morphospecies. Here, the largest subset of congruently placed specimens can be determined as follows. If $\boldsymbol{A_1} \cup \boldsymbol{A_2} \cup \ldots \cup \boldsymbol{A_x} = \boldsymbol{B}$: $Nc_2 = \max(|A_1|, |A_2| \ldots |A_x|)$

(3) Congruence Class III specimens: This covers specimens in sets of clusters that are both split and lumped relative to morphospecies. Here, only those specimens are considered potentially congruent that (1) are in one mOTU and one morphospecies and (2) combined exceed the number of the other specimens in the set of clusters. In detail, if $\boldsymbol{A_1} \cup \boldsymbol{A_2} \cup \ldots \cup \boldsymbol{A_x} = \boldsymbol{B_1} \cup \boldsymbol{B_2} \cup \ldots \cup \boldsymbol{B_y}$: $Nc_3 = \max(|A_1 \cap B_1|, |A_2 \cap B_1| \ldots |A_x \cap B_y|)$ only if $\max(|A_1 \cap B_1|, |A_2 \cap B_1| \ldots |A_x \cap B_y|) > \frac{1}{2}(|A_1 \cup A_2 \ldots A_x|)$.

Note that these are approximate estimates because we did not re-cluster specimens after the removal of the specimens that were causing incongruence during in the original cluster sets.

359 *Species identification*

360 Using the full-length and mini-barcode sets derived from the published
361 primers, we assessed whether the barcodes were yielding conspecific matches
362 according to “best close match” implemented in SpeciesIdentifier program
363 (TaxonDNA: Meier et al. 2006). “Best close match” examines one barcode at a time.
364 It considers it unidentified and then determines whether the best-matching reference
365 sequence in the dataset is conspecific and within a user-defined distance threshold.
366 The distance threshold used for all assessments was 2%, which corresponds to the
367 best match ratio performance results for objective clustering in this study. Sequences
368 with conspecific matches within 2% are considered correctly identified, while
369 sequences that match allospecific sequences within 2% are considered incorrectly
370 identified. Sequences without matches within 2% are considered unidentified and
371 sequences that have both con- and allo-specific matches within the threshold are
372 considered to have an ambiguous identification. Mixed effects models were again
373 used to assess the performance of barcodes of different length. Similarly, pairwise
374 Tukey tests were used to test for statistical significance with the proportion of correct
375 matches as the response variable in order to determine whether the full-length
376 barcode performed significantly better than the mini-barcodes.

377

378 **Results**

379 For species delimitation with objective clustering, we found that the 2% p-
380 distance threshold yielded the highest congruence across the datasets. It was hence
381 used as the upper-bound estimator for species- and specimen-level congruence.
382 The corresponding results for the 3 and 4% p-distance clusters are reported in the

383 supplementary materials. For ABGD it was the P=0.001 prior that yielded the highest
384 average match ratio and hence the clusters generated by this prior were used in the
385 main analysis (see supplementary material for results under P=0.01 and P=0.04).
386 PTP does not require parameter choices post the input tree.

387 The MANOVA tests performed on all treatments (species delimitation method
388 and distance threshold/prior) indicated that the test variable “dataset” was
389 responsible for much of the observed variance in “match ratio”. The choice of mini-
390 barcode or algorithm for generating mOTUs was of secondary importance (Table
391 S3). After accounting for “dataset”, we find that only mini-barcodes <200-bp perform
392 significantly worse than full-length barcodes (Fig. 2); for all other mini-barcodes
393 (>200-bp) the congruence with morphospecies does not differ significantly and is
394 occasionally superior to what is observed for the full-length barcode. This is evident
395 from the large number of significant differences ($p < 0.05$ & $p < 0.001$) in pairwise
396 post-hoc Tukey tests applied to 100-bp mini- and 657-bp full-length barcodes. Only
397 short barcodes (<100-bp) have a mean performance that is worse (<0 match ratio
398 deviation) than the full-length barcode. Conversely, there is no significant difference
399 between the 200 and 300-bp mini-barcodes and the full-length barcode when
400 objective clustering or PTP are used to estimate mOTUs. Under ABGD, the mini-
401 barcodes outperform the full-length barcodes. For all mOTU delimitation methods,
402 the variance across datasets decline as the mini-barcode increases in length (Fig. 2).
403 The results obtained for *in silico* mini-barcodes are consistent with the performance
404 of mini-barcodes with published primers: the mini-barcodes of 94-bp, 130-bp, and
405 145-bp lengths tend to perform worse than the longer mini-barcodes (Fig. 3). The
406 results are also similar for specimen-level congruence (Table 1 & S8). However,
407 there are some exceptions including the performance improvements of short mini-

408 barcodes, for example, for European marine fish and Northwest Pacific molluscs
409 when grouped with objective clustering (Fig. 3).

410 When the performance of the three different clustering methods was
411 compared, significant differences ($p < 0.05$ in ANOVA test) were found only for the
412 100-bp mini-barcode set (Fig. 4). Here, pairwise post-hoc Tukey tests find that
413 objective clustering performs significantly better than the other delimitation methods
414 ($p < 0.001$) while ABGD and PTP do not differ significantly ($p = 0.88$) but behave
415 erratically for short mini-barcodes (Fig. 2).

416 Mini-barcodes situated at the 5' end of the full-length barcode appear to
417 perform somewhat worse than those situated at the middle or at the 3' end (Fig. 2).
418 For example, the 100-bp mini-barcodes at the 5' end perform poorly for objective
419 clustering (mini-barcode midpoints at 50, 110 & 170-bp), ABGD (mini-barcode
420 midpoints at 110 & 170-bp) and PTP (mini-barcode midpoint at 110-bp). This effect
421 is, however, only statistically significant when the mini-barcodes are very short (100-
422 bp). This positional effect is present across all species delimitation techniques. Note
423 that the 5' end of the full-length barcode appears to contain a large proportion of
424 conserved sites, particularly around the 170-bp and 230-bp midpoint of the 100-bp
425 mini-barcode (Fig. 5). This positional effect averages out as the mini-barcodes
426 increase in length.

427 With regard to specimen-based congruence, we focused on the mini-
428 barcodes with published primers and here report the results for those that a barcode
429 length >200-bp. Approximately three quarters of all specimens are in the
430 "Congruence Class I" (Tables 1 & S8); i.e., their placement is congruent between
431 mOTUs and morphospecies (Average/Median: OC at 2%: 75/75%; ABGD P=0.001:

432 71/72%; PTP: 75/75%). The remaining specimens are placed in mOTUs that are
433 split, lumped, or split and lumped. The number of specimens that are responsible for
434 the splitting and lumping are classified as Congruence Class II and III specimens.
435 Overall, fewer than 10% of the specimens fall into these categories (Table 1: see
436 Class II specimens across species delimitation methods). These are the specimens
437 that should be studied first when addressing conflict between morphospecies and
438 mOTUs.

439 When comparing the suitability of full-length and mini-barcodes for
440 identification purposes, we find that full-length barcodes often have the highest
441 number of correct matches (Table 2), but the difference between mini-
442 barcodes >200-bp is not statistically significant and only very short mini-barcodes
443 had significantly lower proportions of correct matches when compared with the full-
444 length barcode: 94-bp: $p = <0.0001$, 130-bp: $p = <0.0001$, 145-bp: $p = <0.0001$, 164-
445 bp: $p = 0.0243$. The other mini-barcodes (189 – 407-bp) did not show statistically
446 significant differences ($p = 0.0691 – 0.9996$) to the full-length barcode in terms of
447 correct matches. However, for many datasets (e.g., Northwest Pacific Molluscs,
448 North American Birds, etc.), the proportion of ambiguous matches is higher for
449 shorter mini-barcodes (94 – 164-bp) compared to longer barcodes (189 – 657-bp).

450

451 **Discussion**

452 Accelerating species discovery and description is arguably one of the
453 foremost challenges in modern systematics. Specimens for many undescribed
454 species are already in world's natural history museums, but the samples need to be
455 sorted to species-level before they become available for species

456 identification/description and can be used for large-scale analyses of biodiversity
457 patterns. Pre-sorting specimens with DNA barcodes is a potentially promising
458 solution because it is scalable, can be applied to millions of specimens, and much of
459 the specimen handling can be automated. However, in order for this approach to be
460 suitable, a sufficiently large proportion of the pre-sorted units need to accurately
461 reflect species boundaries and the methods for obtaining the sequences need to be
462 suitable for processing specimens that yield degraded DNA.

463 *The main source of variance in congruence: datasets*

464 We here find that the average congruence between mOTUs and
465 morphospecies is 80% for all barcodes >200-bp (median: 83%), with the median
466 being higher (83%) because of outlier datasets with congruence <65% (OC at 2%;
467 ABGD P=0.001, PTP). These outlier datasets are also likely to be responsible for the
468 observation that much of the variance in congruence throughout our study is
469 explained by the variable "dataset". Despite the outliers, 72-75% (median) of the ca.
470 30,000 specimens are assigned to species that are supported by molecular and
471 morphological data. Overall, this is a very high proportion compared to species-level
472 congruence obtained with species-level sorting by parataxonomists (Krell 2004). The
473 specimen-based perspective on congruence between morphology and molecular
474 data is often underappreciated. However, specimen-level congruence is an important
475 criterion for evaluating the suitability of species-level sorting with barcodes. After all,
476 the basic units in a museum collection or an ecological survey are specimens and
477 not species. The correct placement of specimens into species is thus important for
478 systematists and biodiversity researchers alike given that the former would like to
479 see most of the specimens in a collection correctly placed and the latter often need

480 abundance and biomass information at species-level resolution. For our datasets,
481 congruence levels for specimens are slightly lower than for species.

482 The remaining ca. 25% of specimens are placed in mOTUs whose boundaries
483 do not agree with morphospecies. One may initially consider this an unacceptably
484 high proportion, but it is important to keep in mind that the misplacement of one
485 specimen (e.g., due to a contamination of a PCR) will render two mOTUs
486 incongruent; i.e., all specimens in these mOTUs will be considered incongruent and
487 contribute to the 25% of "incongruently" placed specimens. Arguably, one may
488 instead want to investigate how many specimens are causing the conflict. These are
489 the specimens that should be targeted for re-investigation in reconciliation studies.
490 The proportion across the 20 datasets in our study is fairly low and ranges from 10-
491 12% (median) depending on which mOTU delimitation technique is used.

492 Conflict between mOTUs and morphospecies can be caused by technical
493 error or biology. A typical technical factor would be accidental misplacement of
494 specimens due to lab contamination or error during morphospecies sorting. Indeed,
495 the literature is replete with cases where mOTUs that were initially in conflict with
496 morphospecies became congruent once the study of additional morphological
497 characters let to the revision of morphospecies boundaries (e.g., Smith et al. 2008;
498 Tan et al. 2010; Baldwin et al. 2011; Ang et al. 2017). But there are also numerous
499 biological reasons for why one should not expect perfect congruence between
500 mOTUs and species. Lineage sorting, fast speciation, large amounts of intraspecific
501 variability, and introgression are known to negatively affect the accuracy of DNA
502 barcodes (Will & Rubinoff 2004; Rubinoff et al. 2006; Meier 2008). It is thus
503 somewhat surprising that regardless of these issues, the final levels of congruence
504 between morphospecies and DNA sequences are often quite high in animals (Ball et

505 al. 2005; Cywinska et al. 2006; Renaud et al. 2012; Landi et al. 2014; Wang et al.
506 2018). This implies that the pre-sorting specimens to species-level units based on
507 barcodes is worth pursuing for many metazoan clades. High levels of congruence
508 are, however, not a universal observation across all of life. This approach to
509 specimen sorting is unlikely to be useful in groups with widespread barcode sharing
510 between species. This phenomenon occurs within Metazoa (e.g., Anthozoa: Huang
511 et al. 2008) and is likely to be the default outside of Metazoa (e.g., Chase & Fay
512 2009; Hollingsworth et al. 2011).

513 *Barcode length and species delimitation methods*

514 We here tested the widespread assumption that mOTUs based on full-length
515 barcodes are more reliable than those based on mini-barcodes (Burns et al. 2007;
516 Min & Hickey 2007). If this assumption was confirmed, then the use of mini-barcodes
517 would have to be discouraged despite higher amplification success rates and lower
518 cost. However, we find that the performance of *cox1* mini-barcodes with a
519 length >200-bp do not differ significantly from the performance of full-length
520 barcodes. Indeed, compared to the dataset effect, the choice of barcode length is
521 largely secondary. This conclusion is robust across 20 diverse datasets and holds for
522 different clustering algorithms.

523 We also find that the choice of species delimitation algorithm matters little for
524 mini-barcodes >200-bp (Fig. 4). This is fortunate as objective clustering and ABGD
525 algorithms are less computationally demanding than PTP, which necessitates the
526 reconstruction of a ML trees. However, there are some exceptions. Firstly, when the
527 mini-barcodes are extremely short (~100-bp), objective clustering tends to
528 outperform ABGD and PTP. PTP's poor performance for 100-bp mini-barcodes is not

529 surprising given that it relies on tree topologies which cannot be estimated with
530 confidence based on so little data. ABGD's poor performance is mostly observed for
531 certain priors (e.g., P=0.04: Fig. S5 & S6). Under these priors, ABGD tends to lump
532 most of the 100-bp barcodes into one or few large clusters. Prior-choice also affects
533 ABGD's performance for full-length barcodes. ABGD does not perform well with very
534 low priors (P = 0.001: Fig. 2 & 3 vs. P = 0.01; P = 0.04: Fig. S5). Overall, we
535 conclude that the selection of the best priors and/or clustering thresholds remains a
536 significant challenge for the study of largely unknown faunas that lack morphological
537 information as an *a posteriori* method for selecting priors/thresholds. Overall, we
538 recommend the use of multiple methods and thresholds in order to distinguish robust
539 from labile mOTUs that are heavily dependent on threshold- or prior-choice. The
540 latter should be targeted in reconciliation studies utilizing another type of data.

541 *Barcode length and species identification*

542 Similar to our tests on species delimitation, we find that mini-barcodes >200-
543 bp have no significant performance difference when compared with full-length
544 barcodes. However, it does appear that full-length barcodes have marginally – albeit
545 statistically not significantly – better performance than >200-bp mini-barcodes (Table
546 2: 0.00 – 4.15% difference, 0.37% median). This is consistent with findings by
547 Meusnier et al. (2008) who reported at least 95% identification success for 250-bp
548 barcodes while full-length barcodes performed the best (97% success). Overall, the
549 data suggest that mini-barcodes >200-bp are suitable for species identification when
550 matching sequences to barcode databases with identified barcodes. The poorer
551 performance of the shorter mini-barcodes (<200-bp) are largely due to an increase in
552 ambiguous matches rather than incorrect matches. These mini-barcodes are less

553 informative which could be due to short length or the placement of many short
554 barcodes in a more conserved part of the Folmer region as described below.

555 *Positional effects*

556 Overall, mini-barcodes at the 3' end of the Folmer region outperform mini-
557 barcodes at the 5' end. This is consistent across all three species delimitation
558 methods and was also mentioned by Shokralla et al. (2015) who concluded that
559 mini-barcodes at the 5' end have worse species resolution for fish species. This
560 positional effect is apparent when match ratios are compared across a "sliding
561 window" (Fig. 2). The lowest congruence with morphology is observed for 100-bp
562 mini-barcodes with midpoints at the 50, 110 and 170-bp marks. However, this
563 positional effect is only significant when the barcode lengths are very short (<200-
564 bp). Once the mini-barcodes are sufficiently long (>200-bp), there seems to be no
565 appreciable difference in performance, which is not surprising because sampling
566 more nucleotides helps with buffering against regional changes in nucleotide
567 variability. These changes in nucleotide variability may have functional reasons
568 related to the conformation of the Cox1 protein in the mitochondrion membrane. The
569 Folmer region of Cox1 contains six transmembrane α- helices that are connected by
570 five loops (Tsukihara et al. 1996; Pentinsaari et al. 2016). Pentinsaari et al. (2016)
571 compared 292 Cox1 sequences across 26 animal phyla and found high amino acid
572 variability in helix I and the loop connecting helix I and helix II (corresponding to
573 position 1-102 of *cox1*), as well as end of helix IV and loop connecting helix IV and V
574 (corresponding to positions ~448-498). These regions of high variability are distant
575 from the active sites and thus less likely to affect Cox1 function (Pentinsaari et al.
576 2016). This may lead to lower selection pressure and thus higher variability in these
577 areas which could impact the performance of mini-barcodes for species delimitation.

578 *Accelerating biodiversity discovery and description*

579 We had earlier argued that species discovery and description can be broken
580 up into three steps (1) obtaining specimens, (2) species-level sorting and (3) species
581 identification or description. We here only address species-level sorting. This means
582 that the impediments caused by slow rates of species identification and description
583 are apparently unresolved. However, this is only partially correct. Firstly, some
584 mOTUs delimited via barcodes can be identified via barcode databases. The
585 proportion of successful identification differs depending on how well a particular
586 fauna has been barcoded. This is illustrated by our recent work on dragon- and
587 damselflies (Odonata), ants (Formicidae), and non-biting midges (Chironomidae) of
588 Singapore (Wang et al. 2018, Yeo et al. 2018; Baloğlu et al. 2018). These studies
589 also document why it is important to distinguish between species- and specimen-
590 level identification because the results differ widely. For odonates, BLAST-searches
591 identified more than half of the 95 mOTUs and >75% of the specimens to species.
592 The corresponding numbers for ants and midges were ca. 20% and 10% at mOTU-
593 level, and 9% and 40% at the specimen-level. Secondly, mOTUs discovered via
594 barcodes can be readily compared across studies and borders (Ratnasingham, et al.
595 2013). In contrast, species newly discovered based on morphological evidence
596 usually remain unavailable to the scientific community until they are published. This
597 is a very significant difference because a large amount of downstream biodiversity
598 analysis can be carried out based on mOTUs instead of identified/described species.
599 This includes studying species richness and abundance over time which is a task
600 that is becoming increasingly important in the 21st century. This means that only
601 moderate harm is done if species identification or description are postponed.

602 Unfortunately, the analyses of biodiversity patterns will be impacted by
603 incorrectly delimited mOTUs and it is thus important to estimate this error. In our
604 study, we find that ca. 80% of the mOTUs are congruent with morphospecies. This is
605 prior to a reconciliation stage where the morphology of specimens with a conflicting
606 assignment is revisited in order to rule out that the morphological evidence was
607 misinterpreted and/or an insufficient number of characters was studied; i.e., we
608 would predict that the overall congruence levels after reconciliation will be higher.
609 Ideally, we would like to know which proportion of mOTUs that are in conflict with
610 morphospecies will eventually be rejected, but unfortunately we still know fairly little
611 about the congruence levels between morphology and barcodes after reconciliation.
612 This is because rigorous studies would have to be based on datasets with dense
613 taxon and geographic sampling where first morphological and DNA sequence
614 information is obtained for all specimens and then all cases of conflict are re-studied.
615 Unfortunately, there are very few datasets that satisfy these criteria. This is
616 presumably because the high cost of full-length barcodes has prevented biologists
617 from sequencing all specimens.

618 *Conclusions*

619 We here illustrate that mini-barcodes are as reliable for pre-sorting specimens
620 into putative species as full-length barcodes. Mini-barcodes are also suitable for
621 identifying unidentified barcodes to species based on barcode reference databases.
622 We would thus argue that mini-barcodes should be preferred over full-length
623 barcodes because (1) they can be obtained more readily for specimens that only
624 contain degraded DNA (Hajibabaei et al. 2006) and (2) are much more cost-
625 effective. In particular, we recommend the use of mini-barcodes >200-bp at the 3'
626 end of the Folmer region. It is encouraging that such mini-barcodes perform well

627 across a large range of metazoan taxa. These conclusions are based on three
628 species delimitation algorithms (objective clustering, ABGD and PTP) which, overall,
629 have no appreciable effect on the performance of mini-barcodes. If the DNA of the
630 specimens is so degraded that very short mini-barcodes have to be obtained, we
631 advise against the use of PTP and ABGD (especially with high priors) in order to
632 reduce the likelihood that morphospecies are lumped.

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

Funding

This work was supported by a Ministry of Education grant on biodiversity discovery (R-154-000-A22-112).

Acknowledgements

We would like to acknowledge Athira Adom and Jonathan Ho for data processing and Emily Hartop and Wendy Wang for commenting on the manuscript.

668

669

670 **References**

671 Ahrens D., Fujisawa T., Krammer H.J., Eberle J., Fabrizi S., & Vogler A.P. 2016.
672 Rarity and incomplete sampling in DNA-based species delimitation. *Systematic*
673 *Biology*. 65:478–494.

674 Ang Y., Rajaratnam G., Su K.F., & Meier R. 2017. Hidden in the urban parks of New
675 York City: *Themira lohmanus*, a new species of Sepsidae described based on
676 morphology, DNA sequences, mating behavior, and reproductive isolation
677 (Sepsidae, Diptera). *ZooKeys*. 698:95.

678 Armani A., Guardone L., Castigliego L., D'Amico P., Messina A., Malandra R.,
679 Gianfaldonia D. & Guidi A. 2015. DNA and Mini-DNA barcoding for the identification
680 of Porgies species (family Sparidae) of commercial interest on the international
681 market. *Food Control*. 50:589–596.

682 Baldwin C.C., Castillo C.I., & Weigt L.A. 2011. Seven new species within western
683 Atlantic *Starksia atlantica*, *S. lepicoelia*, and *S. sluiteri* (Teleostei, Labrisomidae),
684 with comments on congruence of DNA barcodes and species. *ZooKeys*. 79:21.

685 Ball S.L., Hebert P.D., Burian S.K., & Webb J.M. 2005. Biological identifications of
686 mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American*
687 *Benthological Society*. 24:508–524.

688 Baloğlu B., Clews E., & Meier R. 2018. NGS barcoding reveals high resistance of a
689 hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent
690 freshwater reservoirs. *Frontiers in zoology*. 15:31.

Bar-On Y.M., Phillips R., & Milo R. 2018. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*. 115:6506–6511.

Bates D., Mächler M., Bolker B., & Walker S. 2015. Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*. 67:1–48

Bhattacharjee M.J., & Ghosh S.K. 2014. Design of Mini-barcode for Catfishes for assessment of archival biodiversity. *Molecular Ecology Resources*. 14:469–477.

Bi K., Linderoth T., Vanderpool D., Good J.M., Nielsen R., & Moritz C. 2013. Unlocking the vault: next-generation museum population genomics. *Molecular ecology*. 22:6018–6032.

Bickel D.J. 1999. What museum collections reveal about species accumulation, richness, and rarity: an example from the Diptera. *The other 99%: the conservation and biodiversity of invertebrates*. 174–181.

Blaimer B.B., Lloyd M.W., Guillory W.X. & Brady S.G. 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS One*. 11:e0161531.

Burns J.M., Janzen D.H., Hajibabaei M., Hallwachs W. & Hebert P.D. 2007. DNA barcodes of closely related (but morphologically and ecologically distinct) species of skipper butterflies (Hesperiidae) can differ by only one to three nucleotides. *Journal of the Lepidopterists Society*. 61:138–153.

Cafaro P. 2015. Three ways to think about the sixth mass extinction. *Biological Conservation*. 192:387–393.

712 Ceballos G., Ehrlich P.R., Barnosky A.D., García A., Pringle R.M. & Palmer T.M.
713 2015. Accelerated modern human–induced species losses: Entering the sixth mass
714 extinction. *Science Advances*. 1:e1400253.

715 Chase M.W. & Fay M.F. 2009. Barcoding of plants and fungi. *Science*. 325:682−683.

716 Cooper A. 1994. DNA from museum specimens. *Ancient DNA*. 149−165. Springer,
717 New York, NY.

718 Cywinska A., Hunter F.F. & Hebert P.D. 2006. Identifying Canadian mosquito
719 species through DNA barcodes. *Medical and Veterinary Entomology*. 20:413−424.

720 Epp L.S., Boessenkool S., Bellemain E.P., Haile J., Esposito A., Riaz T., Erséus C.,
721 Gusarov V.I., Edwards M.E., Johnsen A., Stenøien H.K., Hassel K., Kauserud H.,
722 Yoccoz N.G., Bråthen K.A., Willerslev E., Taberlet P., Coissac E. & Stenøien H.K.
723 2012. New environmental metabarcodes for analysing soil DNA: potential for
724 studying past and present ecosystems. *Molecular Ecology*. 21:1821−1833.

725 Ermakov O.A., Simonov E., Surin V.L., Titov S.V., Brandler O.V., Ivanova N.V. &
726 Borisenko A.V. 2015. Implications of hybridization, NUMTs, and overlooked diversity
727 for DNA barcoding of Eurasian ground squirrels. *PLoS One*. 10:e0117201.

728 Fagan-Jeffries E.P., Cooper S.J., Bertozzi T., Bradford T.M. & Austin A.D. 2018.
729 DNA barcoding of microgastrine parasitoid wasps (Hymenoptera: Braconidae) using
730 high-throughput methods more than doubles the number of species known for
731 Australia. *Molecular Ecology Resources*. 18:1132−1143.

732 Fan J.A., Gu H., Chen S., Mo B., Wen Y., He W., Liu W & Zeng X. 2009. Species
733 identification of 36 kinds of fruit flies based on minimalist-barcode. *Chinese Journal*
734 *of Applied & Environmental Biology.* 2:215−219.

735 Ficetola G.F., Miaud C., Pompanon F. & Taberlet P. 2008. Species detection using
736 environmental DNA from water samples. *Biology Letters*. 4:423–425.

737 Fields A.T., Abercrombie D.L., Eng R., Feldheim K. & Chapman D.D. 2015. A novel
738 mini-DNA barcoding assay to identify processed fins from internationally protected
739 shark species. *PloS One*. 10:e0114844.

740 Folmer O., Black M., Hoeh W., Lutz R. & Vrijenhoek R. 1994. DNA primers for
741 amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan
742 invertebrates. *Molecular Marine Biology and Biotechnology.* 3:294–299.

743 Guschanski K., Krause J., Sawyer S., Valente L.M., Bailey S., Finstermeier K., Sabin
744 R., Gilissen E., Sonet G., Nagy Z.T., Lenglet G., Mayer F. & Lenglet G. 2013. Next-
745 generation museomics disentangles one of the largest primate radiations.
746 *Systematic Biology*. 62:539–554.

747 Hajibabaei M., Smith M.A., Janzen D.H., Rodriguez J.J., Whitfield J.B. & Hebert P.D.
748 2006. A minimalist barcode can identify a specimen whose DNA is degraded.
749 *Molecular Ecology Notes*. 6:959–964.

750 Hajibabaei M. & McKenna C. 2012. DNA mini-barcodes. *DNA barcodes.* 339–353.
751 Humana Press, Totowa, NJ.

752 Han T., Lee W., Lee S., Park I.G. & Park H. 2016. Reassessment of species
753 diversity of the subfamily Denticollinae (Coleoptera: Elateridae) through DNA
754 Barcoding. *PloS One*. 11:e0148602.

755 Hebert P.D., Cywinska A., Ball S.L. & Dewaard J.R. 2003. Biological identifications
756 through DNA barcodes. *Proceedings of the Royal Society of London. Series B:*
757 *Biological Sciences*. 270:313–321.

758 Hebert P.D., Zakharov E.V., Prosser S.W., Sones J.E., McKeown J.T., Mantle B. &
759 La Salle J. 2013. A DNA 'Barcode Blitz': Rapid digitization and sequencing of a
760 natural history collection. *PLoS One*. 8:e68535.

761 Hebert P.D., Braukmann T.W., Prosser S.W., Ratnasingham S., Ivanova N.V.,
762 Janzen D.H., Hallwachs W., Naik S., Sones J.E. & Zakharov E.V. 2018. A Sequel to
763 Sanger: amplicon sequencing that scales. *BMC Genomics*. 19:219.

764 Hollatz C., Leite B.R., Lobo J., Froufe H., Egas C. & Costa F.O. 2016. Priming of a
765 DNA metabarcoding approach for species identification and inventory in marine
766 macrobenthic communities. *Genome*. 60:260–271.

767 Hollingsworth P.M., Graham S.W. & Little D.P. 2011. Choosing and using a plant
768 DNA barcode. *PloS One*. 6:e19254.

769 Huang D., Meier R., Todd P.A. & Chou L.M. 2008. Slow mitochondrial COI sequence
770 evolution at the base of the metazoan tree and its implications for DNA barcoding.
771 *Journal of Molecular Evolution*. 66:167–174.

772 Jones M., Ghoorah A. & Blaxter M. 2011. jmOTU and taxonerator: turning DNA
773 barcode sequences into annotated operational taxonomic units. *PLoS One.*
774 6:e19259.

775 Katoh K. & Standley D.M. 2013. MAFFT multiple sequence alignment software
776 version 7: improvements in performance and usability. *Molecular Biology and*
777 *Evolution*. 30:772–780.

778 Kemp C. 2015. Museums: The endangered dead. *Nature News*. 518:292.

779 Krell F.T. 2004. Parataxonomy vs. taxonomy in biodiversity studies–pitfalls and
780 applicability of 'morphospecies' sorting. *Biodiversity & Conservation*. 13:795–812.

781 Landi M., Dimech M., Arculeo M., Biondo G., Martins R., Carneiro M., Carvalho G.R.,
782 Brutto S.L. & Costa F.O. 2014. DNA barcoding for species assignment: the case of
783 Mediterranean marine fishes. *PLoS One*. 9:e106135.

784 Lenth R. 2018. *Emmeans*: Estimated marginal means, aka least-squares means. *R*
785 *Package Version*. 1.

786 Lim N.K.M., Tay Y.C., Srivathsan A., Tan J.W.T., Kwik J.T.B., Baloğlu B., Meier R &
787 Yeo D.C.J. 2016. Next-generation freshwater bioassessment: eDNA metabarcoding
788 with a conserved metazoan primer reveals species-rich and reservoir-specific
789 communities. *Royal Society Open Science*. 3:160635.

790 Lister A.M. & Climate Change Research Group. 2011. Natural history collections as
791 sources of long-term datasets. *Trends in Ecology & Evolution*. 26:153–154.

792 Little D.P. 2014. A DNA mini-barcode for land plants. *Molecular Ecology Resources*.
793 14:437–446.

794 Meier R., Shiyang K., Vaidya G. & Ng P.K. 2006. DNA barcoding and taxonomy in
795 Diptera: a tale of high intraspecific variability and low identification success.
796 *Systematic Biology.* 55:715–728.

797 Meier R. 2008. DNA sequences in taxonomy: opportunities and challenges. *The New*
798 *Taxonomy (ed. Wheeler QD)*. 7:95–127. CRC Press, New York.

799 Meier R., Wong W., Srivathsan A. & Foo M. 2016. $1 DNA barcodes for
800 reconstructing complex phenomes and finding rare species in specimen-rich
801 samples. *Cladistics*. 32:100–110.

802 Meusnier I., Singer G.A., Landry J.F., Hickey D.A., Hebert P.D. & Hajibabaei M.
803 2008. A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*.
804 9:214.

805 Min X.J. & Hickey D.A. 2007. Assessing the effect of varying sequence length on
806 DNA barcoding of fungi. *Molecular Ecology Resources*. 7:365–373.

807 Morinière J., Balke M., Doczkal D., Geiger M.F., Hardulak L.A., Haszprunar G.,
808 Hausmann A., Hendrich L., Regalado L., Rulik B., Schmidt S., Wägele J.W. & Hebert
809 P.D.N. 2019. A DNA barcode library for 5,200 German flies and midges (Insecta:
810 Diptera) and its implications for metabarcoding-based biomonitoring. *Molecular*
811 *Ecology Resources*. doi: https://doi.org/10.1111/1755-0998.13022.

812 Pentinsaari M., Salmela H., Mutanen M. & Roslin T. 2016. Molecular evolution of a
813 widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific*
814 *Reports*. 6:35275.

815 Puillandre N., Lambert A., Brouillet S. & Achaz G. 2012. ABGD, Automatic Barcode
816 Gap Discovery for primary species delimitation. *Molecular Ecology*. 21:1864–1877.

817 R Core Team. 2017. R: A language and environment for statistical computing. *R*
818 *Foundation for Statistical Computing*. Vienna, Austria. URL http://www.R-project.org.

819 Ratnasingham S. & Hebert P.D. 2013. A DNA-based registry for all animal species:
820 the Barcode Index Number (BIN) system. *PloS One*. 8:e66213.

821 Renaud A.K., Savage J. & Adamowicz S.J. 2012. DNA barcoding of Northern
822 Nearctic Muscidae (Diptera) reveals high correspondence between morphological
823 and molecular species limits. *BMC Ecology*. 12:24.

824 Riedel A., Daawia D. & Balke M. 2010. Deep cox1 divergence and hyperdiversity of
825 Trigonopterus weevils in a New Guinea mountain range (Coleoptera,
826 Curculionidae). *Zoologica Scripta*. 39:63–74.

827 Roe A.D. & Sperling F.A. 2007. Patterns of evolution of mitochondrial cytochrome c
828 oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics*
829 *and Evolution.* 44:325–345.

830 Roskov Y., Abucay L., Orrell T., Nicolson D., Bailly N., Kirk P.M., Bourgoin T.,
831 DeWalt R.E., Decock W., De Wever A., Nieukerken E. van, Zarucchi J., Penev L.
832 2017. Species 2000 & ITIS Catalogue of Life, 2017 Annual Checklist. Digital
833 resource at www.catalogueoflife.org/annual-checklist/2017. *Species 2000: Naturalis*.
834 Leiden, the Netherlands. ISSN 2405-884X.

835 Rubinoff D., Cameron S. & Will K. 2006. A genomic perspective on the shortcomings
836 of mitochondrial DNA for “barcoding” identification. *Journal of Heredity*. 97:581–594.

837 Sánchez-Bayo F. & Wyckhuys K.A. 2019. Worldwide decline of the entomofauna: A
838 review of its drivers. *Biological Conservation.* 232:8–27.

839 Shokralla S., Hellberg R.S., Handy S.M., King I. & Hajibabaei M. 2015. A DNA mini-
840 barcoding system for authentication of processed fish products. *Scientific Reports*.
841 5:15894.

842 Smith M.A., Rodriguez J.J., Whitfield J.B., Deans A.R., Janzen D.H., Hallwachs W. &
843 Hebert P.D. 2008. Extreme diversity of tropical parasitoid wasps exposed by iterative
844 integration of natural history, DNA barcoding, morphology, and
845 collections. *Proceedings of the National Academy of Sciences*. 105:12359–12364.

846 Srivathsan A. & Meier R. 2012. On the inappropriate use of Kimura-2-parameter
847 (K2P) divergences in the DNA-barcoding literature. *Cladistics*. 28:190–194.

848 Srivathsan A., Sha J., Vogler A.P. & Meier R. 2015. Comparing the effectiveness of
849 metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey
850 (*Pygathrix nemaeus*). *Molecular Ecology Resources*. 15:250–261.

851 Srivathsan A., Baloğlu B., Wang W., Tan W.X., Bertrand D., Ng A.H.Q., Boey E.J.H.,
852 Koh J.J.Y., Nagarajan N. & Meier R. 2018. A Min ION™-based pipeline for fast and
853 cost-effective DNA barcoding. *Molecular Ecology Resources*. 0:1–15.

854 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-
855 analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.

856 Stork N.E., McBroom J., Gely C. & Hamilton A.J. 2015. New approaches narrow
857 global species estimates for beetles, insects, and terrestrial arthropods. *Proceedings*
858 *of the National Academy of Sciences*. 112:7519–7523.

859 Sultana S., Ali M.E., Hossain M.M., Naquiah N. & Zaidul I.S.M. 2018. Universal mini
860 COI barcode for the identification of fish species in processed products. *Food*
861 *Research International*. 105:19–28.

862 Tamura K., Stecher G., Peterson D., Filipski A. & Kumar S. 2013. MEGA6: molecular
863 evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*.
864 30:2725–2729.

865 Tan D.S., Ang Y., Lim G.S., Ismail M.R.B. & Meier R. 2010. From 'cryptic species' to
866 integrative taxonomy: an iterative process involving DNA sequences, morphology,

867 and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae:
868 Diptera). *Zoologica Scripta*. 39:51–61.

869 Tamura K., Stecher G., Peterson D., Filipski A. & Kumar S. 2013. MEGA6: molecular
870 evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*.
871 30:2725–2729.

872 Taylor H.R. & Harris W.E. 2012. An emergent science on the brink of irrelevance: a
873 review of the past 8 years of DNA barcoding. *Molecular Ecology Resources.* 12:377–
874 388.

875 Tsukihara T., Aoyama H., Yamashita E., Tomizaki T., Yamaguchi H., Shinzawa-Itoh
876 K., Nakashima R., Yaono R. & Yoshikawa S. 1996. The whole structure of the 13-
877 subunit oxidized cytochrome c oxidase at 2.8 Å. *Science*. 272:1136–1144.

878 Valan M., Makonyi K., Maki A., Vondráček D. & Ronquist F. 2019. Automated
879 Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective
880 Feature Transfer from Convolutional Networks. *Systematic Biology*, *syz014*,
881 https://doi.org/10.1093/sysbio/syz014.

882 Wang W.Y., Srivathsan A., Foo M., Yamane S.K. & Meier R. 2018. Sorting
883 specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a
884 reverse workflow for specimen processing. *Molecular Ecology Resources*. 18:490–
885 501.

886 Yang Z. & Rannala B. 2010. Bayesian species delimitation using multilocus
887 sequence data. *Proceedings of the National Academy of Sciences*. 107:9264–9269.

888 Yang C., Tan S., Meng G., Bourne D.G., O'brien P.A., Xu J., Liao S., Chen A., Chen
889 X. & Liu S. 2018. Access COI barcode efficiently using high throughput Single End
890 400 bp sequencing. *BioRxiv*. 498618. doi: http://dx.doi.org/10.1101/498618.

891 Yeates D.K., Zwick A. & Mikheyev A.S. 2016. Museums are biobanks: unlocking the
892 genetic potential of the three billion specimens in the world's biological
893 collections. *Current Opinion in Insect Science*. 18:83–88.

894 Yeo D., Puniamoorthy J., Ngiam R.W.J. & Meier R. 2018. Towards holomorphology
895 in entomology: rapid and cost-effective adult–larva matching using NGS
896 barcodes. *Systematic Entomology*. 43:678–691.

897 Yu H.J. & You Z.H. 2010. Comparison of DNA truncated barcodes and full-barcodes
898 for species identification. *Advanced Intelligent Computing Theories and Applications.*
899 *With Aspects of Artificial Intelligence.* 108-114. Springer, Berlin, Heidelberg.

900 Zhang J., Kapli P., Pavlidis P. & Stamatakis A. 2013. A general species delimitation
901 method with applications to phylogenetic placements. *Bioinformatics*. 29:2869–2876.

902 Zuccon D., Brisset J., Corbari L., Puillandre N., Utge J. & Samadi S. 2012. An
903 optimised protocol for barcoding museum collections of decapod crustaceans: a
904 case-study for a 10–40-years-old collection. *Invertebrate Systematics*. 26:592–600.

905 **Figure Captions**

906 **Table 1.** Proportion of specimens congruent between morphospecies and mOTU
907 clusters under the three stringency classes. Values in brackets represent the
908 estimated number of specimens causing conflict.

909 **Table 2.** Proportion of correct/incorrect/ambiguous/unidentified specimens in the
910 “best close match” analysis.

Figure 1. Position of the mini-barcode with established primers in this study.

Figure 2. Performance of mini-barcodes along a sliding window (100, 200, 300-bp). Mini-barcode position is indicated on the x-axis and congruence with morphology on the y-axis. mOTUs were obtained with Objective Clustering (2%), ABGD P=0.001 prior), and PTP. Each line represents one data set while the boxplots summarise the values across datasets. Significant deviations from the results obtained with full-length barcodes are indicated with color-coded asterisks (* = $p < 0.05$; ** = $p < 0.001$; red = poorer and green = higher congruence with morphology).

Figure 3. Match ratios across three different species delimitation methods. Mini-barcodes (columns) are sorted by primer length while the datasets (rows) are grouped into 4 classes according to average match ratio. Colours are applied separately to each class.

Figure 4. Comparison of species delimitation methods for full-length and mini-barcodes generated by “sliding windows” (100-bp, 200-bp, 300-bp).

Figure 5. Proportion of conserved sites along the full-length barcode (sliding windows of 100-bp, 200-bp, 300-bp).

928 **Table 1.** Proportion of specimens congruent between morphospecies and mOTU
929 clusters under the three stringency classes. Values in brackets represent the
930 estimated number of specimens causing conflict.

			Objective clustering, 2% p-distance				
Length	**295**	**307**	**313**	**407**	**657**	**Average**	**Median**
Midpoint	**405**	**154**	**502**	**455**	**329**		
Class I	74% (7475)	74% (7511)	75% (7275)	75% (7246)	76% (7118)	75% (7325)	75% (7372)
Class II	89% (2049)	89% (2214)	89% (2079)	89% (2030)	90% (1967)	89% (2068)	89% (2079)
Class III	90% (975)	90% (743)	90% (746)	90% (786)	91% (795)	90% (809)	90% (802)
			ABGD, P=0.001				
Length	**295**	**307**	**313**	**407**	**657**	**Average**	**Median**
Midpoint	**405**	**154**	**502**	**455**	**329**		
Class I	73% (7916)	74% (7685)	72% (8169)	71% (8583)	66% (9833)	71% (8437)	72% (8126)
Class II	89% (2178)	88% (2014)	89% (2138)	88% (2129)	87% (2604)	88% (2213)	88% (2057)
Class III	90% (806)	90% (895)	90% (849)	90% (915)	88% (881)	89% (869)	90% (862)
			PTP				
Length	**295**	**307**	**313**	**407**	**657**	**Average**	**Median**
Midpoint	**405**	**154**	**502**	**455**	**329**		
Class I	74% (7643)	74% (7744)	75% (7369)	76% (7130)	75% (7207)	75% (7419)	75% (7318)
Class II	89% (2362)	88% (2433)	89% (2227)	89% (2095)	89% (1965)	89% (2216)	89% (2179)
Class III	89% (732)	89% (904)	90% (702)	90% (721)	90% (897)	90% (791)	90% (733)

931
932
933
934
935
936

937 **Table 2.** Proportion of correct/incorrect/ambiguous/unidentified specimens in the "best close match" analysis.

	Published Primers									
Dataset	**94**	**130**	**145**	**164**	**189**	**295**	**307**	**313**	**407**	**657**
Great Barrier Reef Fish	85/1/0/15	83/1/3/13	83/1/2/14	85/0/0/14	86/1/0/13	86/0/0/14	85/1/0/14	86/0/0/14	86/0/0/14	86/0/0/14
South China Sea Fish	93/0/1/6	91/1/3/6	91/1/3/6	93/0/1/6	93/0/1/6	94/0/0/6	93/0/1/6	93/1/1/6	94/1/0/6	94/0/0/6
North Sea Molluscs	90/0/0/10	91/0/0/9	92/0/0/8	92/0/0/8	92/0/0/8	92/0/0/8	92/0/0/8	91/0/0/9	91/0/0/9	92/0/0/8
Canada Echinoderms	95/0/0/5	95/0/0/4	95/0/0/5	96/0/0/4	96/0/0/4	95/0/0/5	96/0/0/4	95/0/0/5	96/0/0/4	96/0/0/4
Ecuador Geometridae	84/0/0/15	85/1/0/13	85/0/0/14	85/0/0/14	85/0/0/14	86/0/0/14	86/0/0/14	86/0/0/14	86/0/0/14	86/0/0/14
Germany Aranea & Opiliones	88/0/5/7	89/0/5/5	89/0/5/6	90/1/4/5	90/0/4/5	92/0/2/6	91/1/3/5	92/0/2/5	92/1/2/5	93/1/2/5
European Marine Fish	92/0/7/1	91/0/7/1	92/0/6/2	96/0/3/1	96/0/3/1	97/0/1/1	97/0/1/1	96/0/2/1	97/0/1/1	98/0/1/1
South America Butterflies	89/0/4/7	91/1/2/6	92/0/1/6	92/0/1/6	93/0/0/6	93/0/0/6	94/0/0/6	94/0/0/6	94/0/0/6	94/0/0/6
Germany EPT	92/0/1/7	93/0/1/5	94/0/1/6	94/0/0/6	94/0/0/6	93/0/0/6	94/0/0/5	94/0/0/6	94/0/0/6	94/0/0/6
Northwest Pacific Molluscs	69/2/15/14	75/2/10/13	74/2/10/13	73/2/12/12	73/2/12/12	77/2/8/12	77/2/9/12	76/3/9/12	77/3/8/12	78/4/6/12
Amazon Moths	62/2/1/36	63/3/1/34	63/3/1/34	62/2/1/35	63/2/1/35	63/1/1/35	63/2/1/34	63/1/1/35	63/1/1/35	63/2/1/34
North America Birds	77/1/11/11	72/2/15/10	73/2/15/11	81/2/7/10	80/1/9/10	83/2/5/11	81/2/6/11	84/2/3/10	85/2/3/10	85/2/2/11
French Guiana Earthworms	96/0/0/4	97/0/0/3	97/0/0/3	96/0/0/4	96/0/0/4	97/0/0/3	97/0/0/3	97/0/0/3	97/0/0/3	96/0/0/4
Pakistan Lepidoptera	88/1/3/8	90/1/1/8	90/1/1/8	91/1/1/8	91/1/0/8	91/0/0/8	91/0/0/8	91/0/0/8	91/0/0/8	91/0/0/8
Tanytarsus	81/1/5/12	83/2/5/11	82/2/5/12	86/2/2/11	85/2/3/10	83/2/3/12	86/2/3/9	85/2/2/10	86/2/1/11	87/2/1/11
North Europe Tachinidae	64/2/14/19	65/3/16/17	65/3/15/17	66/3/13/18	65/3/15/17	67/3/11/18	68/3/12/17	68/3/10/18	69/3/10/18	71/4/8/17
Congo Fish	80/1/10/9	78/1/11/9	79/1/11/10	84/2/6/7	86/2/5/7	85/2/5/8	86/2/4/8	86/3/4/7	86/3/4/7	88/3/2/7
North America Pyraustinae	93/1/4/3	94/1/4/1	96/1/2/1	96/1/1/2	97/0/0/2	97/0/0/3	98/1/0/2	97/0/0/3	97/0/0/2	98/2/0/2
Ecuador Chrysomelidae	74/0/1/25	76/0/1/23	76/0/1/23	76/0/1/23	76/0/1/24	75/0/1/24	76/0/1/23	75/0/1/24	75/0/1/24	75/0/1/24
Iberia Butterflies	77/0/22/1	85/0/15/0	86/0/13/0	87/0/12/0	89/0/10/0	93/0/6/0	94/1/5/0	94/0/5/0	96/0/4/0	97/1/3/0

938

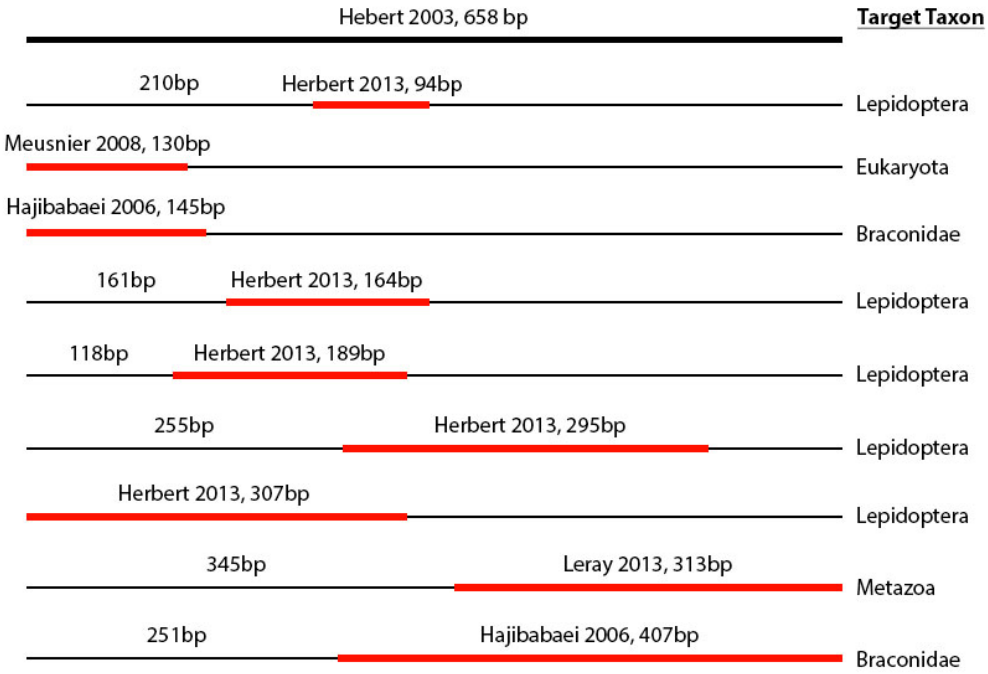

Figure 1. Position of the mini-barcode with established primers in this study.

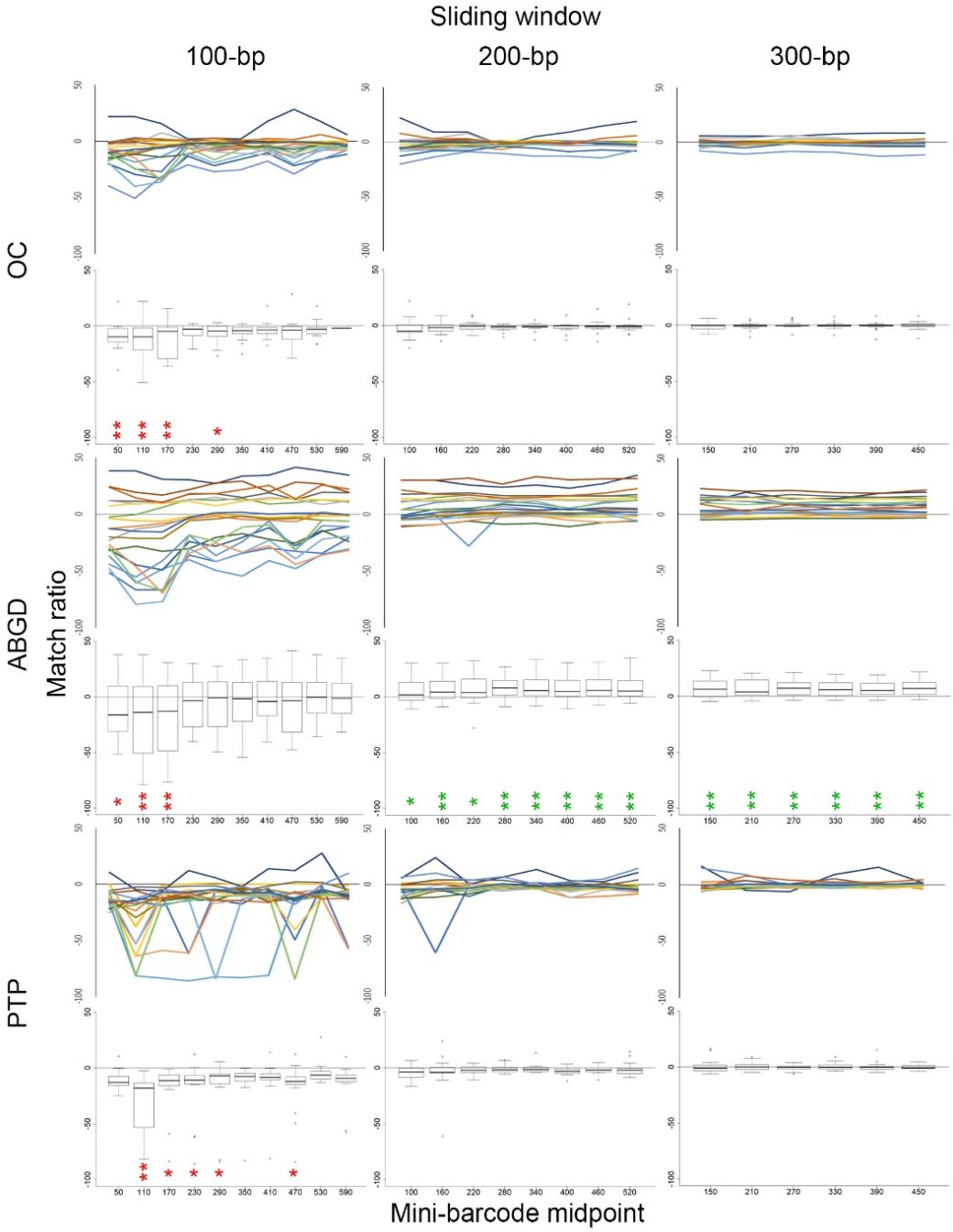

Figure 2. Performance of mini-barcodes along a sliding window (100, 200, 300-bp). Mini-barcode position is indicated on the x-axis and congruence with morphology on the y-axis. mOTUs were obtained with Objective Clustering (2%), ABGD P=0.001 prior), and PTP. Each line represents one data set while the boxplots summarise the values across datasets. Significant deviations from the results obtained with full-

947 length barcodes are indicated with color-coded asterisks (* = $p < 0.05$; ** = $p <$

948 0.001; red = poorer and green = higher congruence with morphology).

Objective clustering

Length	94	130	145	164	189	295	307	313	407	657	
Midpoint	**257**	**65**	**73**	**243**	**213**	**405**	**154**	**502**	**455**	**329**	**Average**
Great Barrier Reef Fish	94	91	92	95	95	97	94	97	97	97	95
South China Sea Fish	94	95	96	98	88	95	93	95	95	97	95
North Sea Molluscs	91	86	87	94	91	94	94	93	93	95	92
Canada Echinoderms	81	89	85	91	93	91	91	91	94	95	90
Ecuador Geometridae	78	70	73	85	98	99	99	99	99	99	90
	Set Mean:			**92**	**Set Range:**			**70**	**-**	**99**	
Germany Aranea & Opiliones	83	89	88	90	84	84	85	84	85	85	86
European Marine Fish	87	88	89	90	81	83	85	83	83	84	85
South America Butterflies	72	74	73	77	90	92	93	92	93	93	85
Germany EPT	78	85	84	85	85	85	82	84	85	85	84
Northwest Pacific Molluscs	88	87	90	91	83	78	77	77	78	78	83
	Set Mean:			**84**	**Set Range:**			**72**	**-**	**93**	
Amazon Moths	69	77	76	79	91	89	92	84	85	87	83
North America Birds	84	77	77	82	77	83	83	83	83	84	81
French Guiana Earthworms	77	85	84	90	78	74	75	78	77	84	80
Pakistan Lepidoptera	52	74	74	54	90	91	90	91	91	91	80
Tanytarsus	81	81	82	81	63	61	62	63	61	65	70
	Set Mean:			**79**	**Set Range:**			**52**	**-**	**92**	
North Europe Tachinidae	47	51	49	57	77	80	82	82	80	81	68
Congo Fish	63	61	60	68	70	65	65	65	65	60	64
North America Pyraustinae	54	54	55	55	57	72	66	70	71	70	63
Ecuador Chrysomelidae	61	56	64	69	63	58	59	55	56	57	60
Iberia Butterflies	51	55	56	59	56	55	55	55	55	55	55
	Set Mean:			**62**	**Set Range:**			**47**	**-**	**82**	

ABGD

Length	94	130	145	164	189	295	307	313	407	657	
Midpoint	**257**	**65**	**73**	**243**	**213**	**405**	**154**	**502**	**455**	**329**	**Average**
North Sea Molluscs	94	90	92	94	94	89	90	89	86	80	90
South China Sea Fish	92	78	83	89	93	94	92	93	91	90	89
Great Barrier Reef Fish	79	75	79	90	89	97	92	95	95	94	88
Canada Echinoderms	87	78	76	91	91	87	91	95	92	89	88
Pakistan Lepidoptera	61	67	75	84	89	95	91	95	96	94	85
	Set Mean:			**88**	**Set Range:**			**61**	**-**	**97**	
Germany EPT	85	84	85	86	87	83	85	84	78	67	82
Ecuador Geometridae	48	58	57	96	70	99	99	99	99	97	82
South America Butterflies	60	70	77	84	84	89	91	91	91	81	82
European Marine Fish	80	72	72	85	86	83	84	80	78	69	79
North America Birds	61	67	73	74	76	81	83	83	84	87	77
	Set Mean:			**80**	**Set Range:**			**48**	**-**	**99**	
Amazon Moths	72	43	42	88	75	87	90	90	91	89	77
North Europe Tachinidae	71	60	66	77	80	77	84	82	82	84	76
French Guiana Earthworms	78	89	93	81	81	73	80	70	59	55	76
Northwest Pacific Molluscs	76	75	75	77	77	76	77	76	76	63	75
Germany Aranea & Opiliones	76	18	79	85	82	71	74	73	65	52	67
	Set Mean:			**74**	**Set Range:**			**18**	**-**	**93**	
Congo Fish	68	44	55	67	71	70	63	67	71	65	64
Tanytarsus	67	71	71	61	62	50	56	48	47	45	58
North America Pyraustinae	25	29	41	67	76	70	74	67	58	59	57
Ecuador Chrysomelidae	55	55	54	56	56	57	57	55	57	58	56
Iberia Butterflies	26	36	42	53	57	60	65	59	55	52	50
	Set Mean:			**57**	**Set Range:**			**25**	**-**	**76**	

PTP

Length	94	130	145	164	189	295	307	313	407	657	
Midpoint	**257**	**65**	**73**	**243**	**213**	**405**	**154**	**502**	**455**	**329**	**Average**
Great Barrier Reef Fish	89	85	87	91	91	95	90	95	96	96	91
Pakistan Lepidoptera	95	83	82	87	90	95	95	95	95	94	91
Ecuador Geometridae	14	89	91	94	97	98	97	96	99	98	87
North Sea Molluscs	78	71	82	94	96	86	92	86	93	93	87
South China Sea Fish	86	75	77	87	88	92	89	89	92	92	87
	Set Mean:			**89**	**Set Range:**			**14**	**-**	**99**	
Canada Echinoderms	75	71	78	83	90	89	90	87	91	89	84
Germany EPT	82	80	81	81	83	79	83	82	79	80	81
North Europe Tachinidae	77	70	73	82	80	84	83	83	82	85	80
Amazon Moths	32	88	38	89	88	91	89	91	90	90	79
Germany Aranea & Opiliones	72	73	77	77	79	82	80	77	81	79	78
	Set Mean:			**80**	**Set Range:**			**32**	**-**	**91**	
North America Birds	72	67	66	77	78	81	82	81	85	85	77
French Guiana Earthworms	71	83	69	85	80	81	62	78	75	69	75
European Marine Fish	71	60	61	71	78	80	74	79	79	80	73
Northwest Pacific Molluscs	72	67	68	75	74	72	74	76	77	78	73
Congo Fish	57	44	50	67	69	68	64	68	71	69	63
	Set Mean:			**72**	**Set Range:**			**44**	**-**	**85**	
Tanytarsus	51	58	64	61	58	68	63	64	66	63	62
North America Pyraustinae	45	58	54	56	49	58	71	80	76	58	61
Ecuador Chrysomelidae	55	57	55	56	55	55	54	55	55	55	55
Iberia Butterflies	32	47	48	47	48	59	58	61	58	54	51
South America Butterflies	0	0	0	0.5	0	89	90	93	91	91	45
	Set Mean:			**55**	**Set Range:**			**0**	**-**	**93**	

Figure 3. Match ratios across three different species delimitation methods. Mini-barcodes (columns) are sorted by primer length while the datasets (rows) are grouped into 4 classes according to average match ratio. Colours are applied separately to each class.

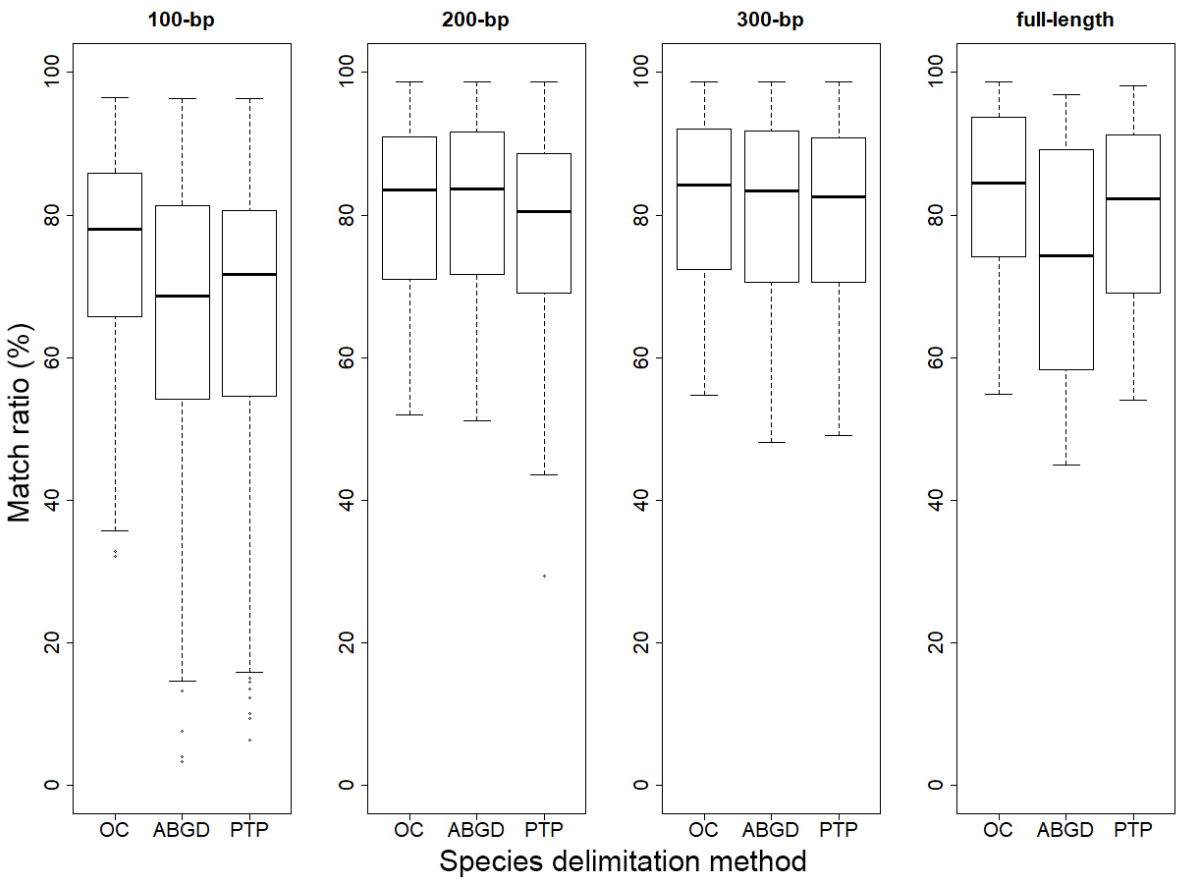

Figure 4. Comparison of species delimitation methods for full-length and mini-barcodes generated by “sliding windows” (100-bp, 200-bp, 300-bp).

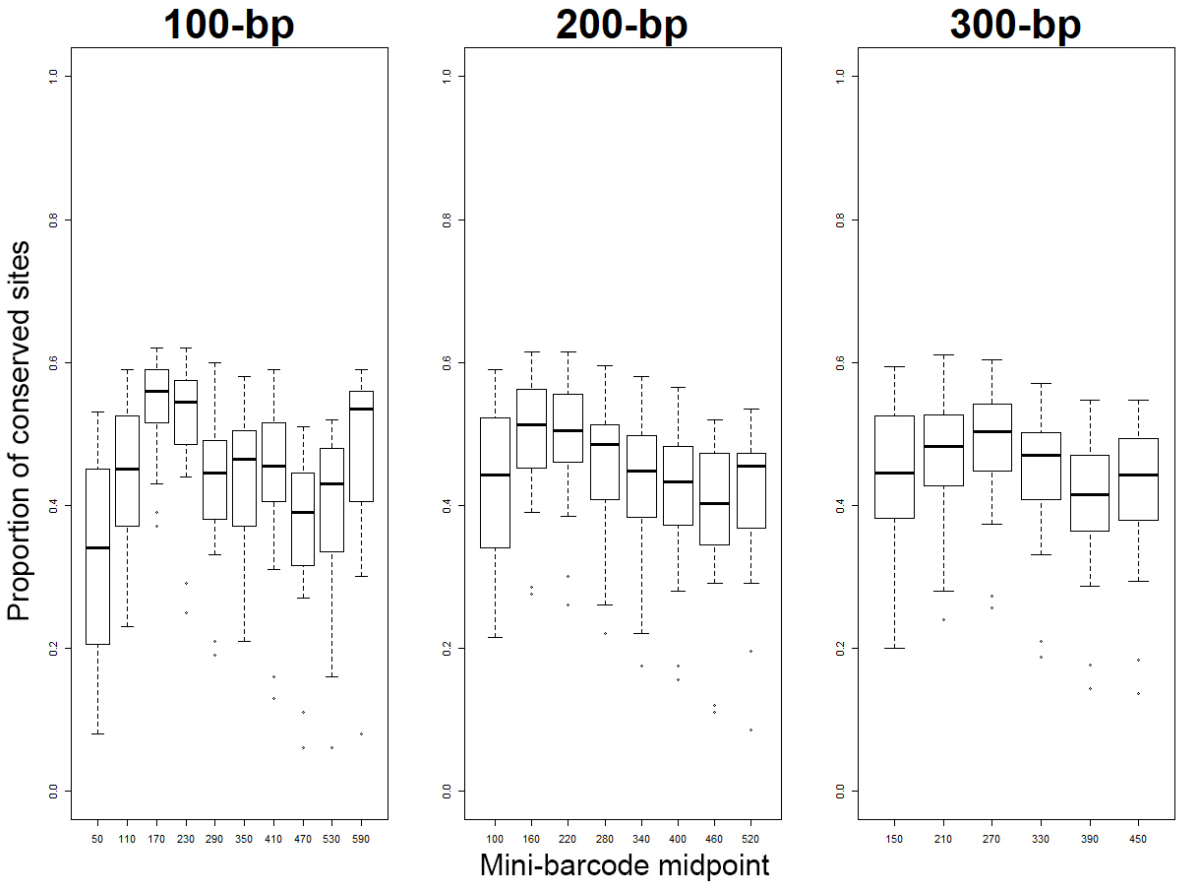

Figure 5. Proportion of conserved sites along the full-length barcode (sliding windows of 100-bp, 200-bp, 300-bp).