Mini-barcodes are equally useful for species identification and more suitable for large-scale species discovery in Metazoa than full-length barcodes

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

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

Running Title: mini-barcodes for species discovery

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”).

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 third, species identification or description. Fortunately, centuries of collecting have generated many of the specimens that are needed for large-scale species discovery. Indeed, for many invertebrate groups it is likely that the museum collections contain more specimens of undescribed than described species; i.e., this unsorted collection material represents vast and still underutilized source for species

discovery (Lister & Climate Change Research Group, 2011; Kemp, 2015; Yeates et al., 2016). The second step in species discovery/description is species-level sorting. This step is in dire need for acceleration. Traditionally, it starts with the sorting of unsorted material into major taxa (e.g., order-level in insects). This task can still be accomplished by parataxonomists but may in the future be taken over by machines utilizing neural networks (Valan et al., 2019). In contrast, the subsequent species-level sorting is usually time-limiting because the specimens for many invertebrate taxa have to be prepared and studied by highly-skilled specialists (e.g., dissected; slide-mounted) before the material can be sorted into putative species. This means that the traditional techniques are neither rapid nor cost-effective for many invertebrate groups. This impediment is likely to be responsible for why certain taxa that are known to be abundant and species-rich are particularly poorly studied (Bickel, 1999).

An alternative way to sort specimens to species-level would be with DNA sequences. This approach is particularly promising for metazoan species because most multicellular animal species can be distinguished based on cytochrome c oxidase subunit I (*cox1*) barcode sequences (Hebert et al., 2003). However, such sorting requires that each specimen is barcoded. This creates cost- and scalability problems when the barcodes are obtained with Sanger sequencing (see Taylor and Harris, 2012). Such sequencing is currently still the standard in barcoding studies because the animal barcode was defined as a 658-bp long fragment of *cox1* ("Folmer region": Folmer et al., 1994), although sequences >500-bp with <1% ambiguous bases are also considered BOLD-compliant (BOLDsystems.org). The 658-bp barcode was optimized for ABI capillary sequencers but it has become a burden because it is not suitable for most new sequencing technologies, while Sanger sequencing remains expensive and only scalable when expensive liquid-handling robots are used. This approach is hence unlikely to become widely available in those countries that harbour most of the species diversity. Due to these constraints, very few studies have utilized DNA barcodes to sort entire samples into putative species (but see Fagan-Jeffries et al., 2018). Instead, most studies use a mixed approach where species-level sorting is carried out based on morphology

before a select few specimens per morphospecies are barcoded (e.g., Riedel et al., 2010). This two-step process requires considerable amounts of skilled manpower and time.

Scalability and cost-effectiveness are hallmark features of the new short-read high throughput sequencing technologies. In addition, these technologies are particularly suitable for sequencing the kind of degraded DNA that is typical for museum specimens. Indeed, anchored hybrid enrichment (AHE) has already been optimized for the use with old museum specimens (Bi et al., 2013; Guschanski et al., 2013; Blaimer et al., 2016) and is likely to play a major role for the integration of rare species into taxonomic and systematic projects. It will be difficult, however, to apply AHE to millions of specimens because it requires time-consuming and expensive molecular protocols (e.g., specimen-specific libraries). Fortunately, for most species it is likely that species-level sorting does not require a very large number of markers. We would thus argue that the initial species-level sorting can be achieved using barcodes that can be obtained via “tagged amplicon sequencing” on next-generation sequencing platforms (“NGS barcodes”; Wang et al., 2018; Yeo et al., 2018). Until recently, obtaining full-length NGS barcodes via tagged amplicon sequencing was difficult because the reads of most next-generation-sequencing platforms were too short for sequencing the full-length barcode. This has now changed with the arrival of third-generation platforms (ONT: MinION: Srivathsan et al., 2018; PacBio: Sequel: Hebert et al., 2018). These platforms, however, come with drawbacks; *viz* elevated sequencing error rates and higher cost. These problems are likely to be overcome in the future (e.g., Yang et al., 2018), but such solutions would not solve the main challenge posed by full-length barcodes; i.e., the problems with reliably obtaining amplicons from museum specimens with degraded DNA (e.g., Hajibabaei et al., 2006). We thus submit that one should optimize the barcode length based on empirical evidence; it should be only as long as needed for accurate pre-sorting of specimens into putative species.

Barcodes that are shorter than the full-length barcode are called mini-barcodes. They are obtained with primers that amplify shorter subsets of the original barcode region and have several advantages. Firstly, such amplicons are easier to obtain when the DNA in the sample

is degraded (Hajibabaei & McKenna, 2012). Secondly, mini-barcodes can be sequenced at low cost using tagged amplicon sequencing on short-read sequencing platforms (e.g., Illumina). Thirdly, mini-barcode primers are available for a large number of species-rich metazoan clades (Hajibabaei et al., 2006; Meusnier et al., 2008; Hebert et al., 2013; Little, 2014) as well as for specific taxa such as fruit flies, catfish and sharks (Fan et al., 2009; Bhattacharjee & Ghosh, 2014; Fields et al., 2015). It is thus not surprising that short barcodes are already the barcodes of choice when the template DNA is degraded. This is often the case for museum specimens (Zuccon et al., 2012; Hebert et al., 2013) or for environmental DNA which is usually analysed via metabarcoding (e.g.: processed food: Armani et al., 2015; Shokralla et al., 2015; water, soil, fecal matter: Epp et al., 2012; Srivathsan et al., 2015; Lim et al., 2016). Mini-barcodes were initially obtained via Sanger sequencing, but they can now be sequenced much more efficiently via tagged amplicon sequencing on short-read platforms ("NGS barcoding": Wang et al., 2018: sequencing cost < 4 cents). Wang et al. (2018) could thus implement a "reverse workflow" based on sequencing all specimens without any species-level pre-sorting based on morphology. Four-thousand specimens of ants were barcoded with a 313-bp mini-barcode. The specimens were then pre-sorted into 89 molecular operational taxonomic units (mOTUs) that were largely congruent with morphospecies (86 species). However, it remained unclear whether full-length DNA barcodes would have further improved congruence, whether the results from this study can be generalized, and which mini-barcode is optimal for large-scale pre-sorting of specimens into putative species.

The answers to these questions are elusive, because mini-barcodes remain insufficiently tested despite their ubiquitous use in metabarcoding. Arguably, existing tests suffer from lack of scale (the largest study includes 6695 barcodes for 1587 species: Meusnier et al., 2008) and taxonomic scope (usually only 1-2 family-level taxa: e.g. Hajibabaei et al., 2006; Yu & You, 2010). Furthermore, the tests yielded conflicting results. Hajibabaei et al. (2006) found high congruence with the full-length barcode when species are delimited based on mini-barcodes and Meusnier et al. (2008) find similar BLAST identification rates for mini-barcodes and full-

length barcodes in their *in silico* tests. However, Yu & You (2010) conceded that mini-barcodes may have worse accuracy despite having close structural concordance with the full-length barcode. In addition, Sultana et al. (2018) concluded that the ability to identify species is compromised when the barcodes are too short (<150-bp), but it remained unclear at which length and in which position mini-barcodes stop performing poorly. Furthermore, published tests of mini-barcodes compare their performance to results obtained with full-length barcodes. All conflict is then implicitly considered evidence for the failure of mini-barcodes. However, results obtained with longer barcodes should not automatically be assumed to be accurate given that the Folmer region varies in nucleotide variability (Roe & Sperling, 2007). Lastly, the existing tests of mini-barcodes do not include a sufficiently large number of different mini-barcodes in order to be able to detect positional and lengths effects across the 658-bp barcode region.

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

significantly lower congruence with morphospecies are unlikely to be useful for accurate species-level sorting.

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

Lastly, we examine the utility of mini-barcodes for species identification. This is the original use of DNA barcodes (Hebert et al., 2003). The widespread use of the metazoan *cox1* Folmer region barcode greatly contributed to the usefulness of DNA barcodes for this purpose, with the value of barcodes increasing as the databases grow. While not as relevant to large-scale species discovery, species identification with barcodes is particularly valuable for targeted species detection and community characterization in metabarcoding studies (Ficetola et al., 2008; Lim et al., 2016; Morinière et al., 2019). Most of these studies tend to involve poor quality

genetic material (i.e., gut content, faecal matter, environmental DNA) and hence it would be valuable to know whether short markers can reliably assign species identities to an unknown sequence. We here test the performance of the barcodes in our test datasets using “best close match”, which considers those sequences correctly identified when they match conspecific barcodes within a predefined threshold (Meier et al., 2006).

Materials & Methods

Dataset selection

We surveyed the barcoding literature in order to identify publications that cited the original barcode paper by Hebert et al. (2003) and met the following criteria: 1) have pre-identified specimens where the barcoded specimens were pre-sorted/identified based on morphology and 2) the dataset had at least 500 specimens with *cox1* barcodes >656-bp. We identified 20 most recent datasets dating backwards from 2017 (Table S1); all had >500 barcoded specimens even after removing those that were not sorted to species level (e.g., only identified to genus or higher) or had short sequences <657-bp (the full-length barcode is technically 658-bp long, but a 1-bp concession was made to prevent the loss of too much data). The barcode sequences were downloaded from BOLDSystems or NCBI GenBank and aligned with MAFFT v7 (Katoh & Standley, 2013) with a gap opening penalty of 5.0.

Using a custom python script, we generated three sets of mini-barcodes along a “sliding window”. They were of 100-, 200- and 300-bp lengths. The first iteration begins with the first base pair of the 658-bp barcode and the shifting windows jump 60-bp at each iteration, generating ten 100-bp windows, eight 200-bp windows and six 300-bp windows. Additionally, we identified nine mini-barcodes with published primers within the *cox1* Folmer region (Fig. 1 & Table S2). These mini-barcodes have been repeatedly used in the literature and were used for a broad range of taxa. The primers for the various mini-barcodes were aligned to the homologous regions of each dataset with MAFFT v7 --addfragments (Katoh & Standley, 2013)

in order to identify the precise position of the mini-barcodes within the full-length barcode. The mini-barcode subsets from each barcode were then identified after alignment to full-length barcodes to mark the start and end-point of the amplicon. Note that most of the published primers are in the 5' prime region of the full-length barcode.

Species delimitation

The mini-barcodes and the full-length barcodes were clustered into putative species using three species delimitation algorithms: objective clustering (Meier et al., 2006), ABGD (Puillandre et al., 2012) and PTP (Zhang et al., 2013). For objective clustering, the mOTUs were clustered at 2 – 4% uncorrected p-distance thresholds (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.

In order to test whether barcode length is a significant predictor of congruence, MANOVA tests were carried out in R (R Core Team, 2017) with "match ratio" (species-level congruence) as the response variable and "dataset" and "mini-barcode" as categorical explanatory variables. We found that most of the variance in our study was generated by the "dataset" variable ($P < 0.05$ in MANOVA tests). Given that we were particularly interested in the effect of barcode length and position, "dataset" was subsequently treated as a random effect "mini-barcode" as the explanatory variable (categorical) in a linear mixed effects model (R package *lme4*: Bates, 2010). The *emmeans* R package (Lenth, 2018) was then used to perform pairwise post-hoc Tukey tests between mini- and full-length barcodes so as to assess whether either barcode was performing significantly better/worse. To compare the differences in performance between objective clustering, ABGD and PTP, ANOVA tests were performed in R. After which, pairwise Tukey tests were used to determine which species delimitation method was responsible for significant differences. Lastly, in order to explore the reasons for positional effects, the proportion of conserved sites for each mini-barcode was obtained using MEGA6 (Tamura et al., 2013).

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

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

Species identification

Using the full-length and mini-barcode sets derived from the published primers, we assessed whether the barcodes were yielding conspecific matches according to "best close match" implemented in SpeciesIdentifier program (TaxonDNA: Meier et al., 2006). "Best close match"

examines one barcode at a time. It considers it unidentified and then determines whether the best-matching reference sequence in the dataset is conspecific and within a user-defined distance threshold. The distance threshold used for all assessments was 2%, which corresponds to the best match ratio performance results for objective clustering in this study. Sequences with conspecific matches within 2% are considered correctly identified, while sequences that match allospecific sequences within 2% are considered incorrectly identified. Sequences without matches within 2% are considered unidentified and sequences that have both con- and allo-specific matches within the threshold are considered to have an ambiguous identification. Mixed effects models were again used to assess the performance of barcodes of different length. Similarly, pairwise Tukey tests were used to test for statistical significance with the proportion of correct matches as the response variable in order to determine whether the full-length barcode performed significantly better than the mini-barcodes.

Results

For species delimitation with objective clustering, we found that the 2% p-distance threshold yielded the highest congruence across the datasets. It was hence used as the upper-bound estimator for species- and specimen-level congruence. The corresponding results for the 3 and 4% p-distance clusters are reported in the supplementary materials. For ABGD it was the P=0.001 prior that yielded the highest average match ratio and hence the clusters generated by this prior were used in the main analysis (see supplementary material for results under P=0.01 and P=0.04). PTP does not require parameter choices post the input tree.

The MANOVA tests performed on all treatments (species delimitation method and distance threshold/prior) indicated that the test variable “dataset” was responsible for much of the observed variance in “match ratio”. The choice of mini-barcode or algorithm for generating mOTUs was of secondary importance (Table S3). After accounting for “dataset”, we find that only mini-barcodes <200-bp perform significantly worse than full-length barcodes (Fig. 2); for

all other mini-barcodes (>200-bp) the congruence with morphospecies does not differ significantly and is occasionally superior to what is observed for the full-length barcode. This is evident from the large number of significant differences ($p < 0.05$ & $p < 0.001$) in pairwise post-hoc Tukey tests applied to 100-bp mini- and 657-bp full-length barcodes. Only short barcodes (<100-bp) have a mean performance that is worse (<0 match ratio deviation) than the full-length barcode. Conversely, there is no significant difference between the 200 and 300-bp mini-barcodes and the full-length barcode when objective clustering or PTP are used to estimate mOTUs. Under ABGD, the mini-barcodes outperform the full-length barcodes. For all mOTU delimitation methods, the variance across datasets decline as the mini-barcode increases in length (Fig. 2). The results obtained for *in silico* mini-barcodes are consistent with the performance of mini-barcodes with published primers: the mini-barcodes of 94-bp, 130-bp, and 145-bp lengths tend to perform worse than the longer mini-barcodes (Fig. 3). The results are also similar for specimen-level congruence (Table 1 & S8). However, there are some exceptions including the performance improvements of short mini-barcodes, for example, for European marine fish and Northwest Pacific molluscs when grouped with objective clustering (Fig. 3).

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

Mini-barcodes situated at the 5’ end of the full-length barcode appear to perform somewhat worse than those situated at the middle or at the 3’ end (Fig. 2). For example, the 100-bp mini-barcodes at the 5’ end perform poorly for objective clustering (mini-barcode midpoints at 50, 110 & 170-bp), ABGD (mini-barcode midpoints at 110 & 170-bp) and PTP (mini-barcode midpoint at 110-bp). This effect is, however, only statistically significant when the mini-barcodes are very short (100-bp). This positional effect is present across all species

delimitation techniques. Note that the 5' end of the full-length barcode appears to contain a large proportion of conserved sites, particularly around the 170-bp and 230-bp midpoint of the 100-bp mini-barcode (Fig. 5). This positional effect averages out as the mini-barcodes increase in length.

With regard to specimen-based congruence, we focused on the mini-barcodes with published primers and here report the results for those that a barcode length >200-bp. Approximately three quarters of all specimens are in the "Congruence Class I" (Tables 1 & S8); i.e., their placement is congruent between mOTUs and morphospecies (Average/Median: OC at 2%: 75/75%; ABGD P=0.001: 71/72%; PTP: 75/75%). The remaining specimens are placed in mOTUs that are split, lumped, or split and lumped. The number of specimens that are responsible for the splitting and lumping are classified as Congruence Class II and III specimens. Overall, fewer than 10% of the specimens fall into these categories (Table 1: see Class II specimens across species delimitation methods). These are the specimens that should be studied first when addressing conflict between morphospecies and mOTUs.

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

Discussion

Accelerating species discovery and description is arguably one of the foremost challenges in modern systematics. Specimens for many undescribed species are already in world's natural history museums, but the samples need to be sorted to species-level before they become available for species identification/description and can be used for large-scale analyses of biodiversity patterns. Pre-sorting specimens with DNA barcodes is a potentially promising solution because it is scalable, can be applied to millions of specimens, and much of the specimen handling can be automated. However, in order for this approach to be suitable, a sufficiently large proportion of the pre-sorted units need to accurately reflect species boundaries and the methods for obtaining the sequences need to be suitable for processing specimens that yield degraded DNA.

The main source of variance in congruence: datasets

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

resolution. For our datasets, congruence levels for specimens are slightly lower than for species.

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

Conflict between mOTUs and morphospecies can be caused by technical error or biology. A typical technical factor would be accidental misplacement of specimens due to lab contamination or error during morphospecies sorting. Indeed, the literature is replete with cases where mOTUs that were initially in conflict with morphospecies became congruent once the study of additional morphological characters let to the revision of morphospecies boundaries (e.g., Smith et al., 2008; Tan et al., 2010; Baldwin et al., 2011; Ang et al., 2017). But there are also numerous biological reasons for why one should not expect perfect congruence between mOTUs and species. Lineage sorting, fast speciation, large amounts of intraspecific variability, and introgression are known to negatively affect the accuracy of DNA barcodes (Will & Rubinoff, 2004; Rubinoff et al., 2006; Meier, 2008). It is thus somewhat surprising that regardless of these issues, the final levels of congruence between morphospecies and DNA sequences are often quite high in animals (Ball et a., 2005; Cywinska et al., 2006; Renaud et al., 2012; Landi et al., 2014; Wang et al., 2018). This implies that the pre-sorting specimens to species-level units based on barcodes is worth pursuing for many metazoan clades. High levels of congruence are, however, not a universal observation across all of life. This approach to specimen sorting is unlikely to be useful in groups with widespread

barcode sharing between species. This phenomenon occurs within Metazoa (e.g., Anthozoa: Huang et al., 2008) and is likely to be the default outside of Metazoa (e.g., Chase & Fay, 2009; Hollingsworth et al., 2011).

Barcode length and species delimitation methods

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

We also find that the choice of species delimitation algorithm matters little for mini-barcodes >200-bp (Fig. 4). This is fortunate as objective clustering and ABGD algorithms are less computationally demanding than PTP, which necessitates the reconstruction of a ML trees. However, there are some exceptions. Firstly, when the mini-barcodes are extremely short (~100-bp), objective clustering tends to outperform ABGD and PTP. PTP's poor performance for 100-bp mini-barcodes is not surprising given that it relies on tree topologies which cannot be estimated with confidence based on so little data. ABGD's poor performance is mostly observed for certain priors (e.g., P=0.04: Fig. S5 & S6). Under these priors, ABGD tends to lump most of the 100-bp barcodes into one or few large clusters. Prior-choice also affects ABGD's performance for full-length barcodes. ABGD does not perform well with very low priors (P = 0.001: Fig. 2 & 3 vs. P = 0.01; P = 0.04: Fig. S5). Overall, we conclude that the selection of the best priors and/or clustering thresholds remains a significant challenge for the study of largely unknown faunas that lack morphological information as an *a posteriori* method for selecting priors/thresholds. Overall, we recommend the use of multiple methods and thresholds in order to distinguish robust from labile mOTUs that are heavily dependent on

threshold- or prior-choice. The latter should be targeted in reconciliation studies utilizing another type of data.

Barcode length and species identification

Similar to our tests on species delimitation, we find that mini-barcodes >200-bp have no significant performance difference when compared with full-length barcodes. However, it does appear that full-length barcodes have marginally – albeit statistically not significantly – better performance than >200-bp mini-barcodes (Table 2: 0.00 – 4.15% difference, 0.37% median). This is consistent with findings by Meusnier et al. (2008) who reported at least 95% identification success for 250-bp barcodes while full-length barcodes performed the best (97% success). Overall, the data suggest that mini-barcodes >200-bp are suitable for species identification when matching sequences to barcode databases with identified barcodes. The poorer performance of the shorter mini-barcodes (<200-bp) are largely due to an increase in ambiguous matches rather than incorrect matches. These mini-barcodes are less informative which could be due to short length or the placement of many short barcodes in a more conserved part of the Folmer region as described below.

Positional effects

Overall, mini-barcodes at the 3' end of the Folmer region outperform mini-barcodes at the 5' end. This is consistent across all three species delimitation methods and was also mentioned by Shokralla et al. (2015) who concluded that mini-barcodes at the 5' end have worse species resolution for fish species. This positional effect is apparent when match ratios are compared across a "sliding window" (Fig. 2). The lowest congruence with morphology is observed for 100-bp mini-barcodes with midpoints at the 50, 110 and 170-bp marks. However, this positional effect is only significant when the barcode lengths are very short (<200-bp). Once the mini-barcodes are sufficiently long (>200-bp), there seems to be no appreciable difference in performance, which is not surprising because sampling more nucleotides helps with buffering against regional changes in nucleotide variability. These changes in nucleotide

variability may have functional reasons related to the conformation of the Cox1 protein in the mitochondrion membrane. The Folmer region of Cox1 contains six transmembrane α- helices that are connected by five loops (Tsukihara et al. 1996; Pentinsaari et al. 2016). Pentinsaari et al. (2016) compared 292 Cox1 sequences across 26 animal phyla and found high amino acid variability in helix I and the loop connecting helix I and helix II (corresponding to position 1-102 of *cox1*), as well as end of helix IV and loop connecting helix IV and V (corresponding to positions ~448-498). These regions of high variability are distant from the active sites and thus less likely to affect Cox1 function (Pentinsaari et al. 2016). This may lead to lower selection pressure and thus higher variability in these areas which could impact the performance of mini-barcodes for species delimitation.

Accelerating biodiversity discovery and description

We had earlier argued that species discovery and description can be broken up into three steps (1) obtaining specimens, (2) species-level sorting and (3) species identification or description. We here only address species-level sorting. This means that the impediments caused by slow rates of species identification and description are apparently unresolved. However, this is only partially correct. Firstly, some mOTUs delimited via barcodes can be identified via barcode databases. The proportion of successful identification differs depending on how well a particular fauna has been barcoded. This is illustrated by our recent work on dragon- and damselflies (Odonata), ants (Formicidae), and non-biting midges (Chironomidae) of Singapore (Wang et al., 2018, Yeo et al., 2018; Baloğlu et al., 2018). These studies also document why it is important to distinguish between species- and specimen-level identification because the results differ widely. For odonates, BLAST-searches identified more than half of the 95 mOTUs and >75% of the specimens to species. The corresponding numbers for ants and midges were ca. 20% and 10% at mOTU-level, and 9% and 40% at the specimen-level. Secondly, mOTUs discovered via barcodes can be readily compared across studies and borders (Ratnasingham, et al. 2013). In contrast, species newly discovered based on morphological evidence usually remain unavailable to the scientific community until they are

published. This is a very significant difference because a large amount of downstream biodiversity analysis can be carried out based on mOTUs instead of identified/described species. This includes studying species richness and abundance over time which is a task that is becoming increasingly important in the 21st century. This means that only moderate harm is done if species identification or description are postponed.

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

Conclusions

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

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

Acknowledgements

We would like to acknowledge support from a Ministry of Education grant on biodiversity discovery (R-154-000-A22-112). We would also like to thank Athira Adom and Jonathan Ho for data processing and Emily Hartop and Wendy Wang for commenting on the manuscript.

References

Ahrens, D., Fujisawa, T., Krammer, H. J., Eberle, J., Fabrizi, S., & Vogler, A. P. (2016). Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology*, *65*(3), 478-494.

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

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

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

Ball, S. L., Hebert, P. D., Burian, S. K., & Webb, J. M. (2005). Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society*, *24*(3), 508-524.

Baloğlu, B., Clews, E., & Meier, R. (2018). NGS barcoding reveals high resistance of a hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs. *Frontiers in zoology*, *15*(1), 31.

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, *115*(25), 6506-6511.

Bates, D. M. (2010). *lme4*: Mixed-effects modeling with R.

Bhattacharjee, M. J., & Ghosh, S. K. (2014). Design of Mini-barcode for Catfishes for assessment of archival biodiversity. *Molecular Ecology Resources*, *14*(3), 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*(24), 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*(8), 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*(3), 138-153.

Cafaro, P. (2015). Three ways to think about the sixth mass extinction. *Biological Conservation*, *192*, 387-393.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, *1*(5), e1400253.

Chase, M. W., & Fay, M. F. (2009). Barcoding of plants and fungi. *Science*, *325*(5941), 682-683.

Cooper, A. (1994). DNA from museum specimens. In *Ancient DNA* (pp. 149-165). Springer, New York, NY.

Cywinska, A., Hunter, F. F., & Hebert, P. D. (2006). Identifying Canadian mosquito species through DNA barcodes. *Medical and veterinary entomology*, *20*(4), 413-424.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., ... & Stenøien, H. K. (2012). New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology*, *21*(8), 1821-1833.

Ermakov, O. A., Simonov, E., Surin, V. L., Titov, S. V., Brandler, O. V., Ivanova, N. V., & Borisenko, A. V. (2015). Implications of hybridization, NUMTs, and overlooked diversity for DNA barcoding of Eurasian ground squirrels. *PLoS One*, *10*(1), e0117201.

Fagan-Jeffries, E. P., Cooper, S. J., Bertozzi, T., Bradford, T. M., & Austin, A. D. (2018). DNA barcoding of microgastrine parasitoid wasps (Hymenoptera: Braconidae) using high-throughput methods more than doubles the number of species known for Australia. *Molecular ecology resources*, *18*(5), 1132-1143.

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

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology letters*, *4*(4), 423-425.

Fields, A. T., Abercrombie, D. L., Eng, R., Feldheim, K., & Chapman, D. D. (2015). A novel mini-DNA barcoding assay to identify processed fins from internationally protected shark species. *PloS One*, *10*(2), e0114844.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology, 3*(5), 294-299.

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., ... & Lenglet, G. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Systematic biology*, *62*(4), 539-554.

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, *6*(4), 959-964.

Hajibabaei, M., & McKenna, C. (2012). DNA mini-barcodes. In *DNA barcodes* (pp. 339-353). Humana Press, Totowa, NJ.

Han, T., Lee, W., Lee, S., Park, I. G., & Park, H. (2016). Reassessment of species diversity of the subfamily Denticollinae (Coleoptera: Elateridae) through DNA Barcoding. *PloS one*, *11*(2), e0148602.

Hebert, P. D., Cywinska, A., Ball, S. L., & Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313-321.

Hebert, P. D., Zakharov, E. V., Prosser, S. W., Sones, J. E., McKeown, J. T., Mantle, B., & La Salle, J. (2013). A DNA 'Barcode Blitz': Rapid digitization and sequencing of a natural history collection. *PLoS One*, *8*(7), e68535.

Hebert, P. D., Braukmann, T. W., Prosser, S. W., Ratnasingham, S., Ivanova, N. V., Janzen, D. H., ... & Zakharov, E. V. (2018). A Sequel to Sanger: amplicon sequencing that scales. *BMC genomics*, *19*(1), 219.

Hollatz, C., Leite, B. R., Lobo, J., Froufe, H., Egas, C., & Costa, F. O. (2016). Priming of a DNA metabarcoding approach for species identification and inventory in marine macrobenthic communities. *Genome*, *60*(3), 260-271.

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PloS one*, *6*(5), e19254.

Huang, D., Meier, R., Todd, P. A., & Chou, L. M. (2008). Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *Journal of Molecular Evolution*, *66*(2), 167-174.

Jones, M., Ghoorah, A., & Blaxter, M. (2011). jmOTU and taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. *PLoS one*, *6*(4), e19259.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Kemp, C. (2015). Museums: The endangered dead. *Nature News*, *518*(7539), 292.

Krell, F. T. (2004). Parataxonomy vs. taxonomy in biodiversity studies–pitfalls and applicability of 'morphospecies' sorting. *Biodiversity & Conservation*, *13*(4), 795-812.

Landi, M., Dimech, M., Arculeo, M., Biondo, G., Martins, R., Carneiro, M., ... & Costa, F. O. (2014). DNA barcoding for species assignment: the case of Mediterranean marine fishes. *PLoS One*, *9*(9), e106135.

Lenth, R. (2018). *Emmeans*: Estimated marginal means, aka least-squares means. *R Package Version*, *1*(2).

Lim, N. K., Tay, Y. C., Srivathsan, A., Tan, J. W., Kwik, J. T., Baloğlu, B., ... & Yeo, D. C. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science*, *3*(11), 160635.

Lister, A. M., & Climate Change Research Group. (2011). Natural history collections as sources of long-term datasets. *Trends in ecology & evolution*, *26*(4), 153-154.

Little, D. P. (2014). A DNA mini-barcode for land plants. *Molecular Ecology Resources*, *14*(3), 437-446.

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic biology*, *55*(5), 715-728.

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

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

Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, *9*(1), 214.

Min, X. J., & Hickey, D. A. (2007). Assessing the effect of varying sequence length on DNA barcoding of fungi. *Molecular Ecology Resources*, *7*(3), 365-373.

Morinière, J., Balke, M., Doczkal, D., Geiger, M. F., Hardulak, L. A., Haszprunar, G., ... & Schmidt, S. (2019). A DNA barcode library for 5,200 German flies and midges (Insecta:

Diptera) and its implications for metabarcoding-based biomonitoring. *Molecular Ecology Resources*. doi: https://doi.org/10.1111/1755-0998.13022.

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

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), 1864-1877.

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

Ratnasingham, S., & Hebert, P. D. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PloS One*, *8*(7), e66213.

Renaud, A. K., Savage, J., & Adamowicz, S. J. (2012). DNA barcoding of Northern Nearctic Muscidae (Diptera) reveals high correspondence between morphological and molecular species limits. *BMC ecology*, *12*(1), 24.

Riedel, A., Daawia, D., & Balke, M. (2010). Deep cox1 divergence and hyperdiversity of Trigonopterus weevils in a New Guinea mountain range (Coleoptera, Curculionidae). *Zoologica Scripta*, *39*(1), 63-74.

Roe, A.D. and Sperling, F.A. (2007). Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution, 44*(1), 325-345.

Roskov, Y., Abucay, L., Orrell, T., Nicolson, D., Bailly, N., Kirk, P. M., ... & Nieukerken, E. V. (2018). Species 2000 and ITIS Catalogue of Life 2017. Naturalis, Leiden, the Netherlands. Available at http://www.catalogueoflife.org/annual-checklist/2017.

Rubinoff, D., Cameron, S., & Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for “barcoding” identification. *Journal of heredity*, *97*(6), 581-594.

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

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

Smith, M. A., Rodriguez, J. J., Whitfield, J. B., Deans, A. R., Janzen, D. H., Hallwachs, W., & Hebert, P. D. (2008). Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. *Proceedings of the National Academy of Sciences*, *105*(34), 12359-12364.

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

Srivathsan, A., Sha, J., Vogler, A. P., & Meier, R. (2015). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources*, *15*(2), 250-261.

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W. X., Bertrand, D., Ng, A. H., ... & Meier, R. (2018). A Min ION™-based pipeline for fast and cost-effective DNA barcoding. *Molecular Ecology Resources, 0*, 1-15.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Stork, N. E., McBroom, J., Gely, C., & Hamilton, A. J. (2015). New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proceedings of the National Academy of Sciences*, *112*(24), 7519-7523.

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

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, *30*(12), 2725-2729.

Tan, D. S., Ang, Y., Lim, G. S., Ismail, M. R. B., & Meier, R. (2010). From 'cryptic species' to integrative taxonomy: an iterative process involving DNA sequences, morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae: Diptera). *Zoologica Scripta*, *39*(1), 51-61.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725-2729.

Taylor, H. R., & Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, *12*(3), 377-388.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., ... & Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science*, *272*(5265), 1136-1144.

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

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., & Meier, R. (2018). Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. *Molecular Ecology Resources*, *18*(3), 490-501.

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, *107*(20), 9264-9269.

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

Yeates, D. K., Zwick, A., & Mikheyev, A. S. (2016). Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections. *Current opinion in insect science*, *18*, 83-88.

Yeo, D., Puniamoorthy, J., Ngiam, R. W. J., & Meier, R. (2018). Towards holomorphology in entomology: rapid and cost-effective adult–larva matching using NGS barcodes. *Systematic entomology*, *43*(4), 678-691.

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

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, *29*(22), 2869-2876.

Zuccon, D., Brisset, J., Corbari, L., Puillandre, N., Utge, J., & Samadi, S. (2012). An optimised protocol for barcoding museum collections of decapod crustaceans: a case-study for a 10–40-years-old collection. *Invertebrate Systematics*, *26*(6), 592-600.

Table 1. Proportion of specimens congruent between morphospecies and mOTU clusters under the three stringency classes. Values in brackets represent the 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)

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

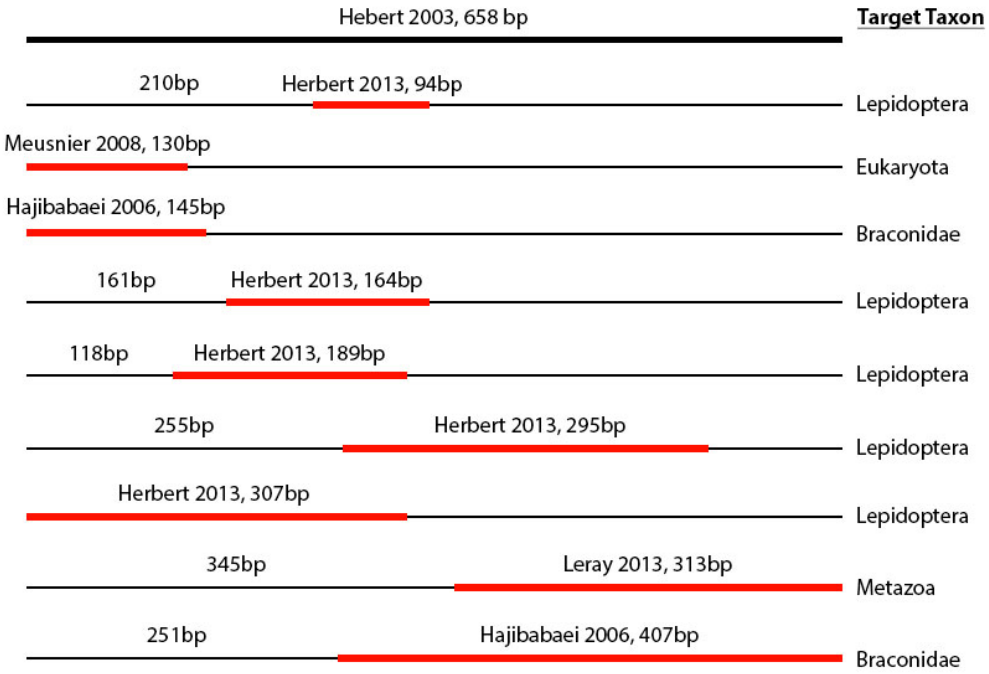

Figure 1. Position of the mini-barcode with established primers in this study.

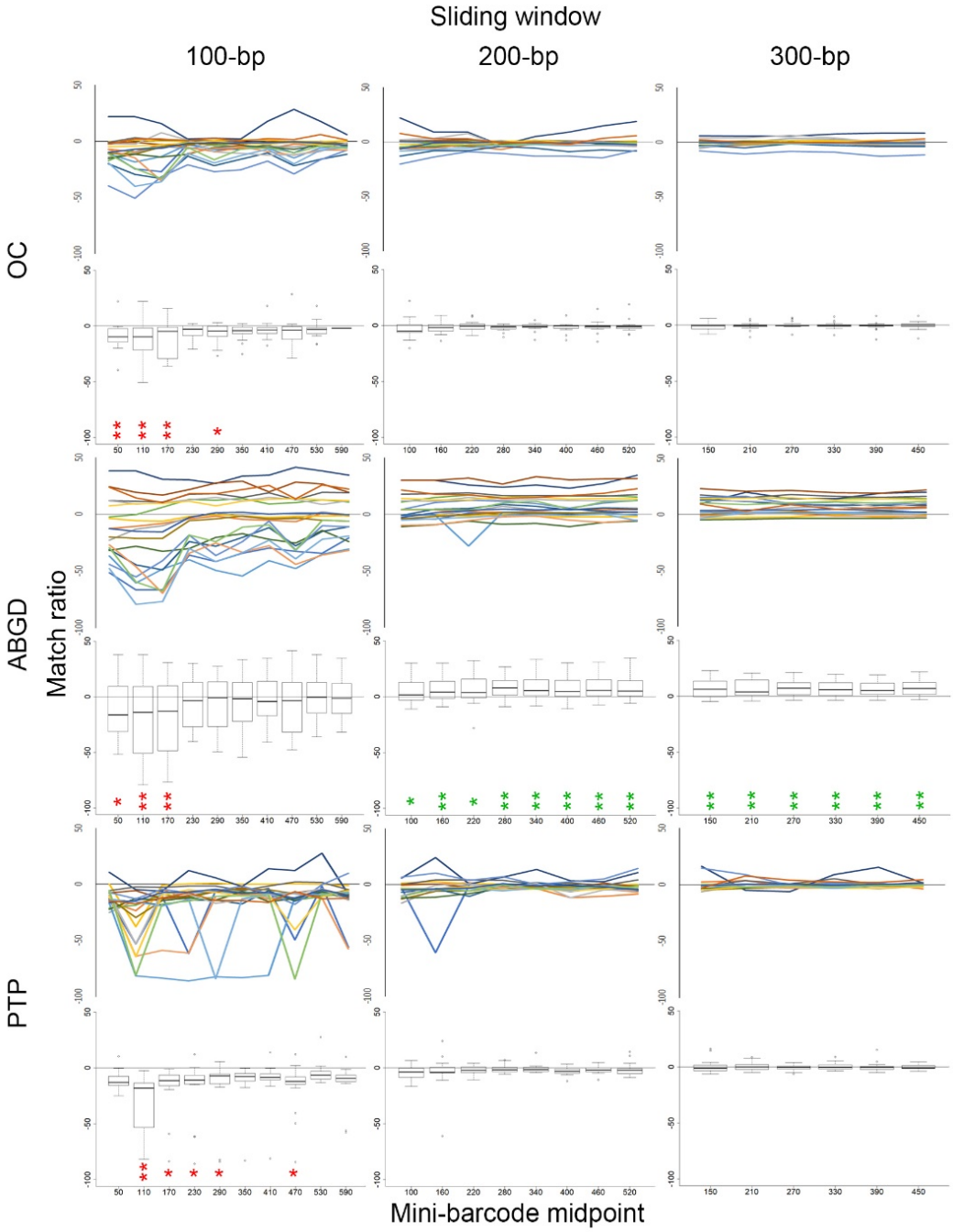

Figure 2. Performance of mini-barcodes along a sliding window (100, 200, 300-bp). Mini-barcode position is indicated on the x-axis and congruence with morphology on the y-axis. mOTUs were obtained with Objective Clustering (2%), ABGD P=0.001 prior), and PTP. Each line represents one data set while the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

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.

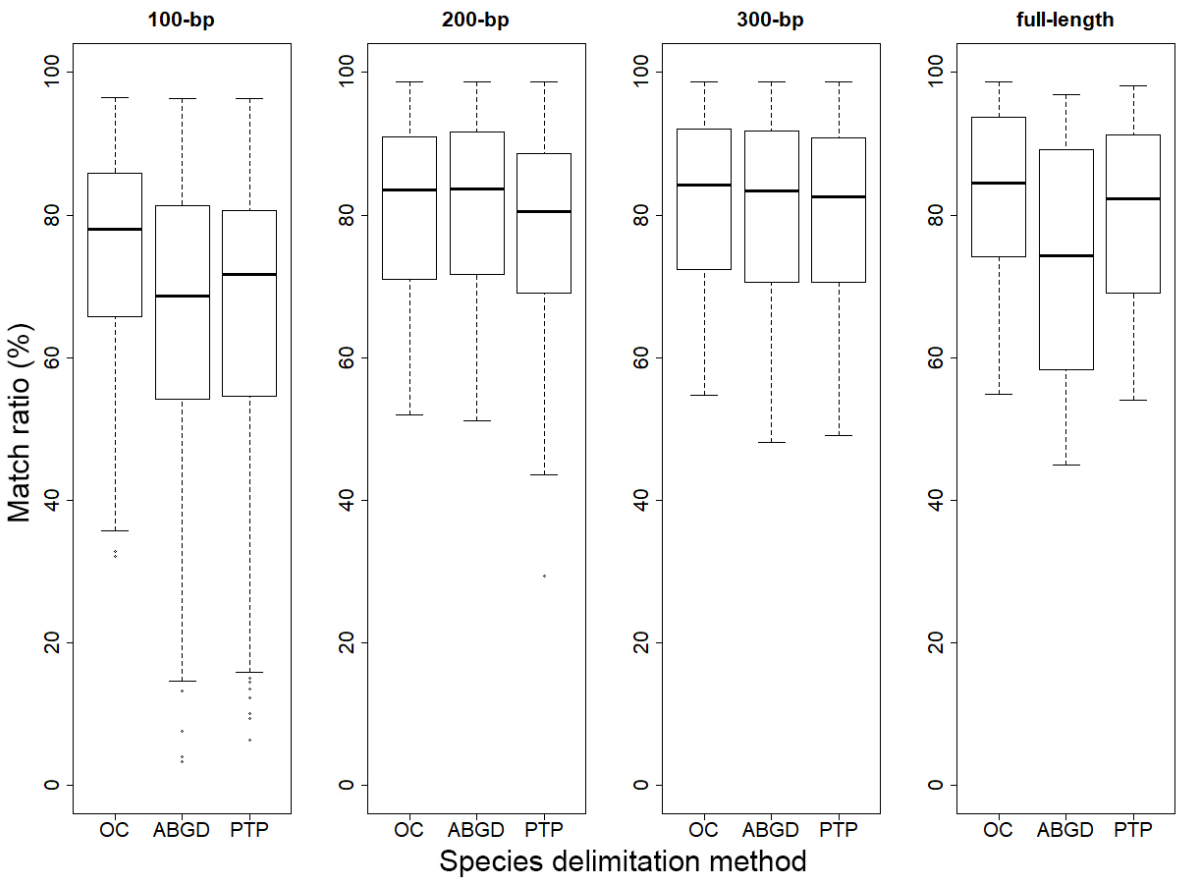

Figure 4. Comparison of species delimitation methods for full-length and mini-barcodes generated by “sliding windows” (100-bp, 200-bp, 300-bp).

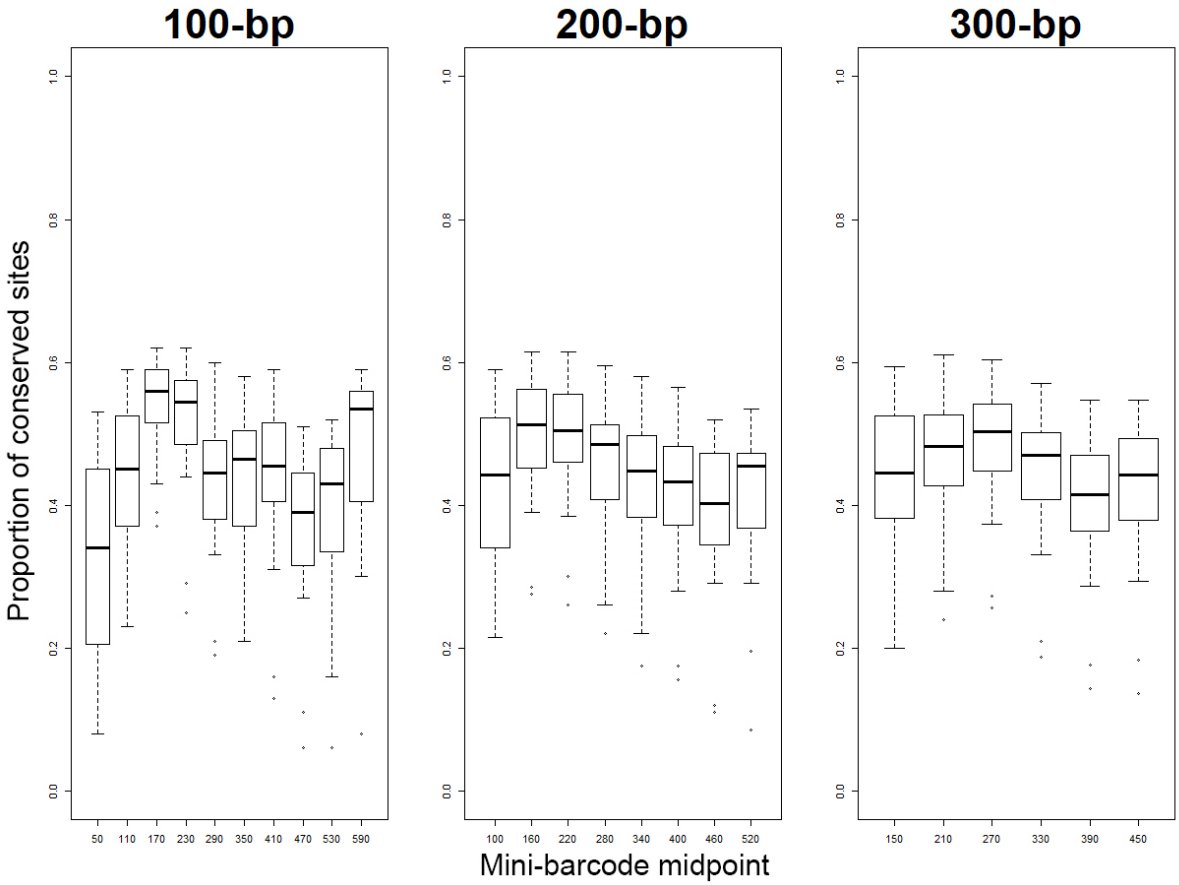

Figure 5. Proportion of conserved sites along the full-length barcode (sliding windows of 100-bp, 200-bp, 300-bp).