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

Running Title: mini-barcodes for species discovery

Darren Yeo¹, Amrita Srivathsan¹, Rudolf Meier^{1*}

¹Department 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 are needed in order to accelerate species discovery, determine how many species live on earth, and develop efficient biomonitoring techniques. These sorting methods should be reliable, scalable and cost-effective, as well as being largely insensitive to low-quality genomic DNA, given that this is usually all that can be obtained from museum specimens. Mini-barcodes seem to satisfy these criteria, but it is unclear how well they perform for species-level sorting when compared to full-length barcodes. This is here tested based on 20 empirical datasets covering ca. 30,000 specimens and 5,500 species, as well as six clade-specific datasets from GenBank covering ca. 98,000 specimens for over 20,000 species. All specimens in these datasets had full-length barcodes and had been sorted to species-level based on morphology. Mini-barcodes of different lengths and positions were obtained *in silico* from full-length barcodes using a sliding window approach (3 windows: 100-bp, 200-bp, 300-bp) and by excising nine mini-barcodes with established primers (length: 94 – 407-bp). We then tested whether barcode length and/or position reduces species-level congruence between morphospecies and molecular Operational Taxonomic Units (mOTUs) that were obtained using three different species delimitation techniques (PTP, ABGD, objective clustering). Surprisingly, we find no significant differences in performance for both species- or specimen-level 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. The mean congruence between morphospecies and mOTUs is ca. 75% for barcodes >200-bp and the congruent mOTUs contain ca. 75% of all specimens. Most conflict is caused by ca. 10% of the specimens that can be

45 identified and should be targeted for re-examination in order to efficiently resolve
 46 conflict. Our study suggests that large-scale species discovery, identification, and
 47 metabarcoding can utilize mini-barcodes without any demonstrable loss of
 48 information compared to full-length barcodes.

49

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

Introduction

The question of how many species live on earth has intrigued biologists for centuries, but we are nowhere close to having a robust answer. We do know that fewer than 2 million have been described and that there are an estimated 10-100 million multicellular species on the planet (Roskov et al. 2018). We also know that many are currently being extirpated by the “sixth mass extinction” (Ceballos et al. 2015; Sánchez-Bayo and Wyckhuys 2019), with potentially catastrophic consequences for the environment (Cafaro 2015). Monitoring, halting, and perhaps even reversing this process is frustrated by the “taxonomic impediment”. This impediment is particularly severe for “invertebrates” that collectively contribute much of the animal biomass (e.g., 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 very likely requires the development of new molecular sorting methods because the traditional approach involving parataxonomists or highly-trained taxonomic experts is either too imprecise (Krell 2004) or too slow and expensive. However, any replacement method based on DNA sequences should 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 will likely require the processing of >500 million specimens, even under the very conservative assumption that there are only 10 million species (Stork 2018) 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 limited funding and only basic research facilities. On the positive side, many species are already represented as unsorted material 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.

High-throughput species discovery with barcodes

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 gathered 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 et al. 2011; Kemp 2015; Yeates et al. 2016). The second step in species discovery/description is species-level sorting, which is in dire need of acceleration. Traditionally, it starts with sorting unsorted material into major taxa (e.g., order-level in insects). This task can be accomplished by parataxonomists but may in the future be guided by robots 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 dissected and slide-mounted before they can be sorted to species-level by highly-skilled specialists; i.e., the traditional techniques are neither rapid nor cost-effective. 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 every 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 many 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 (Barcode Of Life Data System: BOLDsystems.org). The 658-bp barcode was optimized for ABI capillary sequencers but it has become a burden because it is not suitable for cost-effective sequencing with Illumina platforms.

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). Unfortunately, this two-step process requires considerable amounts of skilled manpower and time and does not allow for an unbiased assessment of congruence-levels between morphology and DNA barcodes.

Obtaining Barcodes with High Throughput Sequencing

Fortunately, scalability and cost-effectiveness are hallmark features of the new high throughput sequencing (HTS) technologies. These technologies are particularly suitable for sequencing the kind of degraded DNA that is typical of museum specimens. Indeed, hybrid capture 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 and expensive, however, to apply enrichment methods to millions of specimens because such methods require time-consuming and expensive molecular protocols (e.g., specimen-specific libraries). Fortunately, for most species the initial species-level pre-sorting can be achieved using barcodes that can be obtained via “tagged amplicon sequencing” on a variety of 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, 2019; 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 this would still not solve the main challenge posed by full-length barcodes; i.e., problems with reliably obtaining amplicons from museum specimens with degraded DNA (e.g. Hajibabaei et al. 2006). We thus submit that it is time to use empirical evidence to determine at which length and in which position mini-barcodes are as effective for pre-sorting specimens into putative species as full-length barcodes.

Mini-barcodes

Barcodes that are shorter than the full-length barcode are often referred to as “mini-barcodes”. They are obtained with primers that amplify shorter subsets of the original barcode region and have three key advantages. Firstly, mini-barcodes are predicted to amplify more readily when the DNA in the sample is degraded. This was

confirmed by several studies comparing amplification success rates for mini- and full-length barcodes for the same DNA template (Hajibabaei et al. 2006; Meusnier et al. 2008; Hajibabaei and McKenna 2012). This property makes mini-barcodes the preferred choice for barcoding museum specimens that had been stored under suboptimal conditions for decades. The ease of amplification is also one of the reasons why mini-barcodes are the default for metabarcoding projects that rely on environmental DNA with widely varying DNA quality [e.g., gut content, faeces (Deagle et al. 2006)].

A second benefit of mini-barcodes is that they can be sequenced at low cost on short-read sequencing platforms (e.g., Illumina). The maximum amplicon lengths that such platforms can accommodate is ~450-bp including primer length (using 250-bp paired-end sequencing libraries), which necessitates the amplification and sequencing of two overlapping regions to recover the entire 658-bp barcode (Shokralla et al. 2015b). However, given the additional time and resources required, it begs the question if there is much to be gained by investing in full-length barcodes. Full-length barcodes are often automatically assumed to improve performance but more data does not automatically yield better results. There is often a saturation point beyond which additional data have little impact on results while increasing consumable and manpower cost. Occasionally, more data can even be detrimental. For example, fast-evolving genes or third positions in protein-encoding genes can worsen the results of phylogenetic analyses addressing the relationships between old clades. Note that this could also be a concern for barcodes, because sequence variability changes across the Folmer region of COI (Roe and Sperling 2007; Pentinsaari et al. 2016). Overall, it is important that one should use empirical data to determine how long barcodes have to be in order to be informative.

Unfortunately, the performance of mini-barcodes remains insufficiently tested despite their ubiquitous use in metabarcoding. The existing tests suffer from lack of scale (the largest study includes 6695 barcodes for 1587 species: Meusnier et al. 2008) or taxonomic scope (usually only 1-2 family-level taxa: e.g. Hajibabaei et al. 2006; Yu and 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 and 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/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 considered evidence for the failure of mini-barcodes, even though the Folmer region varies in nucleotide variability (Roe and Sperling 2007) and conflict alone does not settle which result is correct. 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 length effects across the 658-bp barcode region.

Here, we address the lack of scale in previous tests by including 20 empirical studies covering 5500 species (ca. 30,000 barcodes) and Genbank data for 20,673 species (ca. 98,000 barcodes). We 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 established 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 more accurate than those obtained with mini-barcodes. Instead, we assess whether mOTUs obtained with different-length barcodes have different levels of congruence with morphospecies; i.e. morphology is treated as a constant and we only test whether barcode length and/or position influence the number of morphospecies that are recovered. All specimens that are mis-sorted based on morphology (e.g. due to the presence of cryptic species) will be equally mis-sorted for the full-length barcodes and the mini-barcodes that are derived from the former via shortening; i.e., misidentified specimens become noise. What we assess is whether there is an additional loss of congruence between barcodes and morphology as the barcode length shrinks or mini-barcodes are in a different position within the Folmer region of COI.

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 and Rannala 2010; jmOTU: Jones et al. 2011; ABGD: Puillandre et al. 2012; BINs: Ratnasingham and Hebert 2013; PTP: Zhang et al. 2013; etc.). For the purposes of this study, we selected three algorithms that represent 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. Note that PTP is frequently applied to trees derived from barcode data (e.g. Ermakov et al. 2015; Han et al. 2016; Hollatz et al. 2016). There are numerous additional techniques for species delimitation, but most require multiple markers and thus do not scale easily to millions of specimens.

In addition to species delimitation, we also examine the utility of mini-barcodes for species identification, which is the original use of DNA barcodes (Hebert et al. 2003). Species identification with barcodes is particularly valuable for the detection of specific species and community characterization via 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 be reliably used to assign species names to unidentified mini-barcodes. We here test the performance of the barcodes 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, alignment, excising in silico mini-barcodes

We surveyed the barcoding literature in order to identify 20 recent publications from 2007 – 2017 that cited the original barcode paper by Hebert et al. (2003) and met the following criteria: 1) have 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 (Table S1a). In addition, we compiled six

additional datasets based on sequences from GenBank for species-rich metazoan taxa: Actinopterygii, Arachnida, Coleoptera, Diptera, Hymenoptera and Lepidoptera. The clade names were used for taxonomy searches in NCBI and all mitochondrial nucleotide sequences of length 657-1000 bp were downloaded. To these, we added annotated *cox1* sequences from mitochondrial genomes (mitochondrial sequences of length 10,000-20,000 bp). Subsequently, we filtered for the *cox1* gene using an alignment to a reference full-length *cox1* barcode with MAFFT's --add function (Kato and Standley 2013), where any non-target sequences are easily identified by large indels and discarded. We then removed all sequences that were shorter than 657-bp, introduced indels in the multiple sequence alignment, were not identified to species, or included ambiguous species names. Four of the datasets (Actinopterygii: 21,390 barcodes, Coleoptera: 44,403 barcodes, Diptera: 56,344 barcodes, Lepidoptera: 198,441 barcodes), were too large for a full evaluation using all species delimitation methods. We therefore randomly selected 20,000 sequences (=RAND() in Microsoft Excel) for further analysis. These six datasets comprised 97,758 barcodes for 20,673 morphospecies after subsampling (Table S1b).

The barcode sequences were downloaded from BOLDSystems or NCBI GenBank and aligned with MAFFT v7 (Kato and 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" for the 20 empirical datasets. 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

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 (Kato and 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' region of the full-length barcode. The nine mini-barcodes with published primers were assessed for the 20 empirical and for the six GenBank datasets.

Species delimitation

Mini-barcodes and 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 and Meier 2012) using a Python script which re-implements 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, 2016; Ratnasingham and Hebert 2013). 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. In the ABGD algorithm, they are refined by recursive application. Lastly, in order to use PTP, we generated maximum likelihood (ML) trees in RAxML v.8 (Stamatakis 2014) via rapid

bootstrapping (-f a) under the GTRCAT model. For the 20 empirical datasets, the best tree generated for each dataset was then used for species delimitation with PTP (Zhang et al. 2013) under default parameters. For the six much larger clade-based datasets, we used mPTP (Kapli et al. 2017) with the --single parameter as the original algorithm took too long to yield results. In order to reduce computational time for RAxML and PTP, haplotype representatives were used for the six Genbank dataset instead of the entire dataset. The remaining barcodes with identical haplotypes were then mapped back into the same mOTU clusters using their respective haplotypes.

Performance assessment

The performance of mini-barcodes was assessed 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 $\frac{2 \times N_{\text{match}}}{N_1 + N_2} \times 100$, where N_{match} is the number of clusters identical across both mOTU delimitation methods/thresholds (N_1 & N_2). Consolidated match ratios are derived by first summing relevant all numerator and denominator values before calculating the ratio. 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 variable “dataset” ($P < 0.05$ in MANOVA tests). Given that we were interested in the effect of barcode length and position, “dataset” was subsequently treated as a random effect and “mini-barcode” as the explanatory variable (categorical) in a linear mixed effects model (R package *lme4*: Bates et al. 2014). 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 into congruent units. Note that 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 $A = B$ then number of congruent specimens is $N_{c1} = |A| \text{ OR } |B|$.

Lack of congruence is caused by morphospecies that are split, lumped, or both split and lumped in the mOTUs. This means that a single mis-sorted specimen placed into a single, 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-assigning the mis-sorted specimen after re-examination. 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 $A_1 \cup A_2 \cup \dots \cup A_x = B : Nc_2 = \max(|A_1|, |A_2| \dots |A_x|)$

(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 $A_1 \cup A_2 \cup \dots \cup A_x = B_1 \cup B_2 \cup \dots \cup B_y : Nc_3 = \max(|A_1 \cap B_1|, |A_2 \cap B_1| \dots |A_x \cap B_1| \dots |A_x \cap B_y|)$ only if $\max(|A_1 \cap B_1|, |A_2 \cap B_1| \dots |A_x \cap B_y|) > \frac{1}{2}(|A_1 \cup A_2 \dots A_x|)$.

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

Species identification

Using the full-length and mini-barcode sets corresponding to 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.

Amplification success rates for 313-bp and 658-bp amplicons

We tested amplification success rates for 47-48 specimens (~1mm) from six Malaise trap samples of different age (2012 and 2018). The same genomic DNA was used for each specimen for the amplification of the full-length barcode and a mini-barcode of 313-bp length in two separate PCRs using different tagged primers: HCO2198 (5'-TAAACTTCAGGGTGACCAAAAAATCA-3') and LCO1490 (5'-GGTCAACAAATCATAAAGATATTGG-3') (658-bp; Folmer et al. 1994), mICOLintF (5'-GGWACWGGWTGAACWGTWTAYCCYCC-3') and jgHCO2198 (5'-TAIACYTCIGGRTGICCRARAAYCA-3') (313-bp; Leray et al. 2013). The PCR reagents used were 5.0µl Taq MasterMix (CW0682; CWBio, Beijing, China), 1.0µl RNase-free water, 0.5µl bovine serum albumin (1.0 mM), 0.5µl MgCl₂, 1.0µl of each primer and 1.0µl of genomic DNA extract. The PCRs were run at 95°C for 5min, with

35 cycles of 94°C for 1min, 45°C for 2min and 72°C for 1min, and a final extension at 72°C for 5min. PCR amplification success was assessed using a 1% agarose gel, with the electrophoresis run at 90V for 30min. Band presence was tabulated and chi-squared performance testing was performed in R (R Core Team 2017).

Results

Evaluation of mini-barcode congruence with morphology

For species delimitation with objective clustering, we found that the 2% p-distance threshold yielded the highest congruence-levels across the datasets. It was hence used as the upper-bound estimator for assessing species- and specimen-level congruence (see supplementary materials for corresponding results for the 3 and 4% p-distances). For ABGD, average congruence was maximized for the prior $P=0.001$ and this prior was used in the main analysis (see supplementary material for results under $P=0.01$ and $P=0.04$). PTP does not require *a priori* parameter choices. Overall consolidated match ratio values range from 68 – 74% for these selected parameters (Fig. 3), with the 20 empirical datasets generally performing better than the six clade-based datasets (empirical/clade-based: OC at 2%: 81/72%; ABGD $P=0.001$: 76/69%; PTP: 76/66%). Overall match ratios for the other objective clustering and ABGD parameters tested range from 61 – 74% (Fig. S6).

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 our measure of congruence (“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). 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, for mini-barcodes >200-bp, congruence with morphospecies does not differ significantly and is occasionally superior to what is observed for the full-length barcode. Across all datasets (20 empirical and 6 clade-based), species delimitation methods, and clustering parameters (Fig. 3 & S6), the full-length barcode have the highest congruence values in 49 of the 182 tests, but in 8 of these cases it is a tie and in an additional 19 some mini-barcodes are within 0.5%. Furthermore, there is no significant difference between the 200-bp and 300-bp mini-barcodes and the full-length barcode when objective clustering or PTP are used to estimate mOTUs. However, for all mOTU delimitation methods, the variance across datasets declines as the mini-barcode increases in length (Fig. 2). The results obtained for *in silico* mini-barcodes are broadly consistent with the performance of mini-barcodes corresponding to 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). Congruence at the species level is similar to congruence at the specimen level (Tables 1 & S8) with some exceptions such as the performance improvements of short mini-barcodes, for example, for the “French Guiana earthworms” when grouped with objective clustering.

The results obtained for the 20 empirical datasets are similar to what is found for the taxon-specific datasets consisting of sequences from GenBank. There are small differences in match ratios for barcodes >200-bp and only very short barcodes

perform poorly. In many cases, the full-length barcode does not yield the best performance compared to the mini-barcodes. An exception is the Coleoptera where ABGD behaves erratically (Fig. 3). Short barcodes outperform long barcodes under the lowest prior ($P=0.001$), but this trend reverses for higher priors.

With regard to specimen-based congruence, we focused on the mini-barcodes with published primers that are >200-bp. For the 20 empirical datasets, 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%: 77/77%; ABGD $P=0.001$: 74/74%; PTP: 77/77%). The remaining specimens are placed in mOTUs that are split, lumped, or split and lumped. The number of specimens that are responsible for splitting and lumping are classified as Congruence Class II and III specimens (see Materials and Methods). Overall, fewer than 10% of the specimens fall into these categories (Table 1: see Class II specimens across species delimitation methods); i.e., a fairly small number of specimens have to be studied when addressing conflict between morphospecies and mOTUs. For the six clade-based datasets, there is generally lower specimen congruence (Class I Average/Median: OC at 2%: 57/57%; ABGD $P=0.001$: 53/54%; PTP: 57/57%), although this improves to ca. 75% at Class II and ca. 80% at Class III. Note that a single mis-identified *D. melanogaster* could result in all correctly identified specimens of this species to be considered “incongruent” because they are no longer member of a congruent mOTU.

Species delimitation method

When the performance of the three different clustering methods was compared, significant differences ($p < 0.05$ in ANOVA test) were found only for 100-

bp mini-barcodes (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).

Positional effects

Mini-barcodes situated at the 5' end of the full-length barcode 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 observed 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).

Species identification performance

When comparing the suitability of full-length and mini-barcodes for identification purposes, we find that full-length barcodes frequently have the highest number of correct matches (Table 2), but the differences are only statistically significant for short mini-barcodes <200-bp: 94-bp: $p = <0.0001$, 130-bp: $p = <0.0001$, 145-bp: $p = <0.0001$, 164-bp: $p = 0.0243$. All the remaining mini-barcodes: (189 – 407-bp) have non-significant differences. Note that this is not due to scale because we are using fairly large datasets here ($p = 0.0691 - 0.9996$). However, for many datasets (e.g., Northwest Pacific Molluscs, North American Birds, etc.), the

proportion of ambiguous matches is higher for very short mini-barcodes (94-bp – 164-bp).

Amplification success rates

The 313-bp mini-barcode amplified significantly better than the full-length barcode (472/570: 82.8% vs. 361/570: 63.3%; chi-squared test: $p = <0.001$) (Fig. S9). Overall, the performance differences were greater for samples with overall low amplification success rates (but see sample 4822).

Discussion

Accelerating species discovery and description are arguably foremost challenges for the systematics of the 21st century because biodiversity is in steep decline while systematists are still grappling with establishing the size and distribution of the biodiversity that needs protection. Specimens for many undescribed species are already in the 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 viable, a sufficiently large proportion of the pre-sorted units need to accurately reflect species boundaries. Furthermore, the methods should be suitable for processing specimens that only yield degraded DNA. Mini-barcodes appear attractive on technical grounds because sequencing short amplicons is cheap (Wang et al. 2018: sequencing cost < 4 cents) and such amplicons have higher PCR amplification rates

especially if the DNA template is degraded. The latter has been repeatedly discussed in the literature (Hajibabaei et al. 2006; Meusnier et al. 2008; Hajibabaei and McKenna 2012) and was also confirmed here by comparing identification success rates for a mini-barcode (313bp) and the full-full length barcode using the same template (Fig. S9).

However, the use of mini-barcodes for species-level pre-sorting would nevertheless have to be discouraged if full-length barcodes were to significantly outperform mini-barcodes with regard to the accuracy of sorting specimens into species. This is what we tested here and we find that mini-barcodes of moderate length (>200 bp) perform as well as full-length barcodes.

The main source of variance of congruence: datasets

Overall, we here find that the average congruence between mOTUs and morphospecies is 80% for all barcodes >200-bp (median: 83%) in the 20 empirical datasets, with the median being higher (83%) because of a few outlier datasets with low congruence <65% (OC at 2%; ABGD $P=0.001$, PTP). Similar variance is also observed for the six clade-specific GenBank datasets. Both types of datasets were tested because GenBank data are less likely to be affected by the idiosyncrasies of individual study designs, but is known to contain misidentified sequences (Mioduchowska et al. 2018). We overall find that the results are nevertheless similar (Fig. 3 & S6) although the average congruence between mOTUs and morphospecies is somewhat lower for GenBank datasets (mean=72%, median=71% for all barcodes >200-bp across all three species delimitation methods and parameters). Full-length barcodes only have the highest congruence level in 13 of the 42

treatments, but in 10 of those, the second highest congruence level for a mini-barcode was only <0.5% lower. Mini-barcodes <200-bp perform poorly with the possible exception of those for Coleoptera where mini-barcodes <200-bp perform well under ABGD under two priors (Figs. 3 & S6: $P=0.001$, $P=0.01$).

Most studies assessing congruence levels between barcodes and morphology focus on species-level congruence. However, specimen-level congruence is equally important because the basic units in a museum collection or ecological surveys are specimens. 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 with species-level resolution. We find that for the 20 empirical datasets, 74-77% (median) of the ca. 30,000 specimens are assigned to species that are supported by molecular and morphological data even under the strictest conditions (Table 1). Overall, this is a very high proportion compared to species-level congruence obtained with species-level sorting by parataxonomists (Krell 2004).

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 misidentification or contamination of a PCR product) 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 using additional data. The proportion across the

20 datasets in our study is fairly low and ranges from 9-11% (median) depending on which mOTU delimitation technique is used.

The six GenBank datasets however, have notably lower congruence-levels for specimens (Class I median: 54-57%), with the Actinopterygii (47-48% median) and Hymenoptera (45-47% median) having the poorest performance (Table S8). The overall lower congruence could be caused by the misidentification of a few specimens belonging to large mOTUs. This is supported by the greater difference between Class I and Class III for the clade-based datasets as compared to the empirical ones. Identifications from the empirical datasets are also more likely to be standardized within the study and consequently have fewer potential conflicts caused by differing taxon concepts associated with the same name (Franz 2005; Meier 2017).

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 led 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 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 and Rubinoff 2004; Rubinoff et al. 2006; Meier 2008). It is thus somewhat heartening that regardless of these issues, the final levels of congruence between morphospecies and DNA sequences are often quite high in animals (Ball et

al. 2005; Cywinska et al. 2006; Renaud et al. 2012; Landi et al. 2014; Wang et al. 2018). This implies that pre-sorting specimens to species-level units based on barcodes is worth pursuing for many metazoan clades. This realization led to the proposal of the “reverse workflow” in Wang et al. (2018) who tested pre-sorting based on sequences for 4,000 ant specimens. They found that 86 of the 89 molecular operational taxonomic units (mOTUs) were congruent with morphospecies after reconciliation.

High levels of congruence levels are, however, not a universal observation across all of life. There are groups with widespread barcode sharing between species. Within Metazoa this is known for taxa such as Anthozoa (Huang et al. 2008) and is likely to be the default outside of Metazoa (e.g. Chase and Fay 2009; Hollingsworth et al. 2011). Indeed, congruence levels between mOTUs and morphospecies also vary widely in the 20 empirical and 6 Genbank datasets studied here (Fig. 3). Eighteen of the 26 datasets have congruence levels >75% and are arguably doing well. With regard to the remaining eight, there are four where the low levels may be due to the fact that they cover taxa or faunas that are comparatively poorly known and sorting with morphology may be prone to error (French Guiana Earthworms, Ecuadorian Chrysomelidae, chironomid midges in the genus *Tanytarsus*). For a 5th data set, “Congolese fishes”, the authors also mention identification problems due to the lack of identification tools (Decru et al. 2016). However, these problems should be lacking for Iberian butterflies where the authors consider it likely that introgression and cryptic species are responsible for the lack of congruence (Dincă et al. (2015). Note, however, that these problems do not appear to affect the remaining Lepidoptera data set for which we observe high levels of congruence (5 data sets). The picture is similarly confusing for fish (Actinopterygii)

where low congruence is observed for the Genbank dataset while marine fishes are doing well (3 datasets). Similarly, inconsistent patterns are observed for Hymenoptera and Arachnida.

We suspect that these inconsistent results are largely due to the lack of rigorous studies that establish congruence with confidence. Such studies should be based on datasets with dense taxon and geographic sampling where both morphological and DNA sequence information is obtained for all specimens. Subsequently, all specimens that cause conflict should be re-examined in order to determine the proportion of specimens that were initially misplaced based on morphology and the proportion that were initially misplaced based on DNA barcodes (=pre-sorting error). Lastly, one should establish the number of specimens that cannot be resolved because there is genuine conflict between the data sources. Such studies are lacking. This also affects the quality of the morphospecies sorting in the 26 datasets that are analyzed in our study. However, it should not affect our overall conclusions with regard to the effect of barcode length and position on congruence levels because morphospecies sorting errors affect both full-length barcodes and the mini-barcodes that were excised from the former.

But what is gained if all specimens are sorted based on morphology and barcodes? Using both sources of data would allow for higher quality work based on the principles of integrative taxonomy (Dayrat 2005; Schlick-Steiner et al. 2010), but it would not accelerate species discovery. Acceleration can only be achieved if either a subset of all specimens have to be studied with morphology or the morphological work requires less time. Both is the case when pre-sorting with barcodes is adopted. This is shown in Wang et al. (2018) and Srivathsan et al. (2019), who pre-sorted 4000 ant and 7000 phorid specimens, respectively, with barcodes. Checking for

congruence between mOTUs and morpho-species was fast because the specimens had already been pre-grouped into putative species based on barcodes. In addition, determining the species limits between putatively closely related species was simplified because the taxon-specialist could identify closely related species/populations based on barcode distances. Lastly, for abundant species, only a subset of the specimens had to be studied with morphology. This subset could be selected based on genetic, temporal, and geographic gradients (Wang et al. 2018; Srivathsan et al. 2019); i.e., the highly-skilled specialists could target a small proportion of specimens for dissection and slide-mounting.

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 and Hickey 2007). If this assumption was confirmed, then the use of mini-barcodes for pre-sorting 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 substantially from the performance of full-length barcodes. In 19/140 of the empirical and 3/42 of the clade-based dataset treatments, the full-length barcode outperform all mini-barcodes, after discounting ties (8/140 and 0/42) and cases where the difference between the next highest congruence is <0.5% (9/140 and 10/42). Indeed, compared to the dataset effect, the choice of barcode length is largely secondary. Note that this conclusion is robust across 26 datasets and holds for different clustering algorithms.

We also find that the choice of species delimitation algorithm matters little for mini-barcodes longer than 200-bp (Fig. 4). This is fortunate as objective clustering

and ABGD algorithms are less computationally demanding than PTP, which necessitates the reconstruction of ML trees. However, there are some exceptions. Firstly, when mini-barcodes are extremely short (~100-bp), ABGD and PTP tend to underperform relative to objective clustering. 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 a few large clusters. Prior-choice also affects ABGD's performance for full-length barcodes. Overall, ABGD does not perform well with very high priors ($P = 0.04$: Fig. S6 vs. $P = 0.001$: Fig. 2 & 3) and 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 or other gene information that can be used as an *a posteriori* criterion for selecting priors/thresholds. In the absence of these data, we would recommend the use of multiple methods and thresholds in order to distinguish robust from labile mOTUs that are heavily dependent on threshold- or prior-choices. The latter mOTUs should be targeted in reconciliation studies utilizing another type of data. As illustrated earlier, these reconciliation studies can focus on only 10-15% of all specimens that cause species-level conflict.

Barcode length and species identification

Similar to our tests of species delimitation, we find that mini-barcodes >200-bp have no significant performance difference when compared with full-length barcodes. Full-length barcodes have only marginally – albeit statistically non-significantly – higher identification success rates than mini-barcodes >200-bp (Table 2: 0.00 – 4.15% difference, 0.37% median). This is consistent with the findings of 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, our data suggest that mini-barcodes >200-bp are suitable for species identification. The poorer performance of shorter mini-barcodes (<200-bp) is largely due to an increase of ambiguous matches rather than incorrect matches. These mini-barcodes are less informative which is probably due to their short length and the placement of many short barcodes in a more conserved part of the Folmer region (see 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 noticed by Shokralla et al. (2015a) 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-bp, 110-bp 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 is 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

In our study we test whether species discovery with mini-barcodes is as accurate as discovery with full-length barcodes. However, the taxonomic impediment is also caused by the lack of taxonomic capacity for identifying specimens to species and describing new species. Fortunately, the species discovery step can also help with species identification because some specimens placed in mOTUs can be identified via barcode databases. Common species are here more likely to benefit because they are covered by the databases. This is illustrated by our recent work on dragon- and damselflies (Odonata), ants (Formicidae), and non-biting midges (Chironomidae) of Singapore (Baloğlu et al. 2018; Wang et al. 2018; Yeo et al. 2018). These studies 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; i.e., despite low mOTU identification success for midges, a fairly large number of specimens could be identified because very common species were already represented in a barcode database.

In addition, mOTUs are also useful for understanding biodiversity patterns because they can be readily compared across studies and borders (Ratnasingham and Hebert 2013). In contrast, a new species identified based on morphology usually remains unavailable to the scientific community until the descriptions are published unless associated molecular information is also present. This is a very significant difference because a large amount of downstream biodiversity analysis can be carried out based on newly discovered species regardless of whether they were discovered with morphological or molecular means. However, cross-study comparisons are only straightforward for mOTUs because sequences are readily shared and compared. This allows for the study of species richness and abundances before taxa are described. Even studies covering different time periods can be carried out, which is increasingly important in the 21st century.

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 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.

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, Lee Wan Ting for performing the molecular work 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. *Syst. Biol.* 65:478–494.
- Ang Y., Meier R., Su K.F.-Y., Rajaratnam G. 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*. 698:95–111.
- Baldwin C., Castillo C., Weigt L., Victor B. 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*. 79:21–72.
- Ball S.L., Hebert P.D.N., Burian S.K., Webb J.M. 2005. Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *J. North Am. Benthol. Soc.* 24:508–524.

788 Baloğlu B., Clews E., Meier R. 2018. NGS barcoding reveals high resistance of a
789 hyperdiverse chironomid (Diptera) swamp fauna against invasion from
790 adjacent freshwater reservoirs. *Front. Zool.* 15:31.

791 Bar-On Y.M., Phillips R., Milo R. 2018. The biomass distribution on Earth. *Proc. Natl.*
792 *Acad. Sci.* 115:6506–6511.

793 Bates D., Mächler M., Bolker B., Walker S. 2014. Fitting linear mixed-effects models
794 using *lme4*. ArXiv14065823 Stat.

795 Bi K., Linderoth T., Vanderpool D., Good J.M., Nielsen R., Moritz C. 2013. Unlocking
796 the vault: next-generation museum population genomics. *Mol. Ecol.* 22:6018–
797 6032.

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

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

803 Burns J.M., Janzen D.H., Merhadd Hajibabai, Winnie Hallwachs, Paul D. N. Hebert.
804 2007. DNA barcodes of closely related (but morphologically and ecologically
805 distinct) species of skipper butterflies (Hesperiidae) can differ by only one to
806 three nucleotides. *J. Lepidopterists' Soc.* 61:138–153.

807 Cafaro P. 2015. Three ways to think about the sixth mass extinction. *Biol. Conserv.*
808 192:387–393.

809 Ceballos G., Ehrlich P.R., Barnosky A.D., García A., Pringle R.M., Palmer T.M.
810 2015. Accelerated modern human-induced species losses: Entering the sixth
811 mass extinction. *Sci. Adv.* 1:e1400253.

812 Chase M.W., Fay M.F. 2009. Barcoding of plants and fungi. *Science.* 325:682–683.

813 Cooper A. 1994. DNA from museum specimens. In: Herrmann B., Hummel S.,
814 editors. *Ancient DNA: Recovery and Analysis of Genetic Material from*
815 *Paleontological, Archaeological, Museum, Medical, and Forensic Specimens.*
816 New York, NY: Springer New York. p. 149–165.

817 Cywinska A., Hunter F.F., Hebert P.D.N. 2006. Identifying Canadian mosquito
818 species through DNA barcodes. *Med. Vet. Entomol.* 20:413–424.

819 Dayrat B. 2005. Towards integrative taxonomy. *Biol. J. Linn. Soc.* 85:407–415.

820 Deagle B.E., Eveson J.P., Jarman S.N. 2006. Quantification of damage in DNA
821 recovered from highly degraded samples – a case study on DNA in faeces.
822 *Front. Zool.* 3:11.

823 Decru E., Moelants T., De Gelas K., Vreven E., Verheyen E., Snoeks J. 2016.
824 Taxonomic challenges in freshwater fishes: a mismatch between morphology

825 and DNA barcoding in fish of the north-eastern part of the Congo basin. *Mol.*
826 *Ecol. Resour.* 16:342–352.

827 Dincă V., Montagud S., Talavera G., Hernández-Roldán J., Munguira M.L., García-
828 Barros E., Hebert P.D.N., Vila R. 2015. DNA barcode reference library for
829 Iberian butterflies enables a continental-scale preview of potential cryptic
830 diversity. *Sci. Rep.* 5:12395.

831 Ermakov O.A., Simonov E., Surin V.L., Titov S.V., Brandler O.V., Ivanova N.V.,
832 Borisenko A.V. 2015. Implications of hybridization, NUMTs, and overlooked
833 diversity for DNA barcoding of Eurasian ground squirrels. *PLoS ONE*.
834 10:e0117201.

835 Fagan-Jeffries E.P., Cooper S.J.B., Bertozzi T., Bradford T.M., Austin A.D. 2018.
836 DNA barcoding of microgastrine parasitoid wasps (Hymenoptera: Braconidae)
837 using high-throughput methods more than doubles the number of species
838 known for Australia. *Mol. Ecol. Resour.* 18:1132–1143.

839 Ficetola G.F., Miaud C., Pompanon F., Taberlet P. 2008. Species detection using
840 environmental DNA from water samples. *Biol. Lett.* 4:423–425.

841 Folmer O., Black M., Hoeh W., Lutz R., Vrijenhoek R. 1994. DNA primers for
842 amplification of mitochondrial cytochrome c oxidase subunit I from diverse
843 metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3:294–299.

844 Franz N.M. 2005. On the lack of good scientific reasons for the growing
845 phylogeny/classification gap. *Cladistics.* 21:495–500.

846 Guschanski K., Krause J., Sawyer S., Valente L.M., Bailey S., Finstermeier K., Sabin
847 R., Gilissen E., Sonet G., Nagy Z.T., Lenglet G., Mayer F., Savolainen V.
848 2013. Next-generation museomics disentangles one of the largest primate
849 radiations. *Syst. Biol.* 62:539–554.

850 Hajibabaei M., McKenna C. 2012. DNA Mini-barcodes. In: Kress W.J., Erickson D.L.,
851 editors. *DNA Barcodes*. Totowa, NJ: Humana Press. p. 339–353.

852 Hajibabaei M., Smith M.A., Janzen D.H., Rodriguez J.J., Whitfield J.B., Hebert
853 P.D.N. 2006. A minimalist barcode can identify a specimen whose DNA is
854 degraded. *Mol. Ecol. Notes.* 6:959–964.

855 Han T., Lee W., Lee S., Park I.G., Park H. 2016. Reassessment of species diversity
856 of the subfamily Denticollinae (Coleoptera: Elateridae) through DNA
857 barcoding. *PLoS ONE.* 11:e0148602.

858 Hebert P.D.N., Braukmann T.W.A., Prosser S.W.J., Ratnasingham S., deWaard
859 J.R., Ivanova N.V., Janzen D.H., Hallwachs W., Naik S., Sones J.E.,
860 Zakharov E.V. 2018. A Sequel to Sanger: amplicon sequencing that scales.
861 *BMC Genomics.* 19:219.

862 Hebert P.D.N., Cywinska A., Ball S.L., deWaard J.R. 2003. Biological identifications
863 through DNA barcodes. *Proc. R. Soc. Lond. B Biol. Sci.* 270:313–321.

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

867 Hollingsworth P.M., Graham S.W., Little D.P. 2011. Choosing and using a plant DNA
868 barcode. *PLoS ONE*. 6:e19254.

869 Huang D., Meier R., Todd P.A., Chou L.M. 2008. Slow mitochondrial COI sequence
870 evolution at the base of the Metazoan tree and its implications for DNA
871 barcoding. *J. Mol. Evol.* 66:167–174.

872 Jones M., Ghoorah A., Blaxter M. 2011. jMOTU and Taxonerator: Turning DNA
873 barcode sequences into annotated operational taxonomic units. *PLoS ONE*.
874 6:e19259.

875 Kapli P., Lutteropp S., Zhang J., Kobert K., Pavlidis P., Stamatakis A., Flouri T. 2017.
876 Multi-rate Poisson Tree Processes for single-locus species delimitation under
877 Maximum Likelihood and Markov Chain Monte Carlo. *Bioinformatics*.:btx025.

878 Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version
879 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

880 Kemp C. 2015. The billions of specimens in natural-history museums are becoming
881 more useful for tracking Earth’s shrinking biodiversity. But the collections also
882 face grave threats. *Nat. News*. 518:292–294.

883 Krell F.-T. 2004. Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and
884 applicability of ‘morphospecies’ sorting. *Biodivers. Conserv.* 13:795–812.

885 Landi M., Dimech M., Arculeo M., Biondo G., Martins R., Carneiro M., Carvalho G.R.,
886 Brutto S.L., Costa F.O. 2014. DNA Barcoding for species assignment: The
887 case of mediterranean marine fishes. *PLoS ONE*. 9:e106135.

888 Lenth R. 2018. Emmeans: Estimated marginal means. Aka Least-Sq. Means R.

889 Leray M., Yang J.Y., Meyer C.P., Mills S.C., Agudelo N., Ranwez V., Boehm J.T.,
890 Machida R.J. 2013. A new versatile primer set targeting a short fragment of
891 the mitochondrial COI region for metabarcoding metazoan diversity:
892 application for characterizing coral reef fish gut contents. *Front. Zool.* 10:34.

893 Lim N.K.M., Tay Y.C., Srivathsan A., Tan J.W.T., Kwik J.T.B., Baloğlu B., Meier R.,
894 Yeo D.C.J. 2016. Next-generation freshwater bioassessment: eDNA
895 metabarcoding with a conserved metazoan primer reveals species-rich and
896 reservoir-specific communities. *R. Soc. Open Sci.* 3:160635.

897 Lister A.M., Stephen J. Brooks, Phillip B. Fenberg, Adrian G. Glover, Karen E.
898 James, Kenneth G. Johnson, Ellinor Michel, Beth Okamura, Mark Spencer,
899 John R. Stewart, Jonathan A. Todd, Eugenia Valsami-Jones, Jeremy Young.
900 2011. Natural history collections as sources of long-term datasets. *Trends*
901 *Ecol. Evol.* 26:153–154.

902 Meier R. 2008. DNA sequences in taxonomy, opportunities and challenges. *New*
903 *Taxon.*:65–127.

904 Meier R. 2017. Citation of taxonomic publications: the why, when, what and what
905 not: Species citations. *Syst. Entomol.* 42:301–304.

906 Meier R., Shiyang K., Vaidya G., Ng P.K.L. 2006. DNA barcoding and taxonomy in
907 *Diptera*: A tale of high intraspecific variability and low identification success.
908 *Syst. Biol.* 55:715–728.

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

912 Meusnier I., Singer G.A., Landry J.-F., Hickey D.A., Hebert P.D., Hajibabaei M. 2008.
913 A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics.*
914 9:214.

915 Min X.J., Hickey D.A. 2007. Assessing the effect of varying sequence length on DNA
916 barcoding of fungi. *Mol. Ecol. Notes.* 7:365–373.

917 Mioduchowska M., Czyż M.J., Gołdyn B., Kur J., Sell J. 2018. Instances of erroneous
918 DNA barcoding of metazoan invertebrates: Are universal *cox1* gene primers
919 too “universal”? *PLOS ONE.* 13:e0199609.

920 Morinière J., Balke M., Doczkal D., Geiger M.F., Hardulak L.A., Haszprunar G.,
921 Hausmann A., Hendrich L., Regalado L., Rulik B., Schmidt S., Wägele J.,
922 Hebert P.D.N. 2019. A DNA barcode library for 5,200 German flies and
923 midges (Insecta: *Diptera*) and its implications for metabarcoding-based
924 biomonitoring. *Mol. Ecol. Resour.* 19:900–928.

925 Pentinsaari M., Salmela H., Mutanen M., Roslin T. 2016. Molecular evolution of a
926 widely-adopted taxonomic marker (COI) across the animal tree of life. *Sci.*
927 *Rep.* 6:35275.

928 Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode
929 Gap Discovery for primary species delimitation. *Mol. Ecol.* 21:1864–1877.

930 Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species:
931 the Barcode Index Number (BIN) system. *PLoS ONE.* 8:e66213.

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

935 Riedel A., Daawia D., Balke M. 2010. Deep *cox1* divergence and hyperdiversity of
936 *Trigonopterus* weevils in a New Guinea mountain range (*Coleoptera*,
937 *Curculionidae*). *Zool. Scr.* 39:63–74.

938 Roe A.D., Sperling F.A.H. 2007. Patterns of evolution of mitochondrial cytochrome c
939 oxidase I and II DNA and implications for DNA barcoding. *Mol. Phylogenet.*
940 *Evol.* 44:325–345.

941 Roskov Y., Abucay L., Orrell T., Nicolson D., Bailly N., Kirk P.M., Bourgoin T.,
942 DeWalt R.E., Decock W., De Wever A. 2018. Species 2000 & ITIS Catalogue
943 of Life. 2017, Annual Checklist. Digital resource at www.catalogueoflife.org/annual-checklist/2017. Species 2000: Naturalis, Leiden, the Netherlands.
944 ISSN 2405-884X. .
945

946 Rubinoff D., Cameron S., Will K. 2006. A genomic perspective on the shortcomings
947 of mitochondrial DNA for “Barcoding” identification. *J. Hered.* 97:581–594.

948 Sánchez-Bayo F., Wyckhuys K.A.G. 2019. Worldwide decline of the entomofauna: A
949 review of its drivers. *Biol. Conserv.* 232:8–27.

950 Schlick-Steiner B.C., Steiner F.M., Seifert B., Stauffer C., Christian E., Crozier R.H.
951 2010. Integrative Taxonomy: A Multisource Approach to Exploring
952 Biodiversity. *Annu. Rev. Entomol.* 55:421–438.

953 Shokralla S., Hellberg R.S., Handy S.M., King I., Hajibabaei M. 2015a. A DNA mini-
954 barcoding system for authentication of processed fish products. *Sci. Rep.*
955 5:15894.

956 Shokralla S., Porter T.M., Gibson J.F., Dobosz R., Janzen D.H., Hallwachs W.,
957 Golding G.B., Hajibabaei M. 2015b. Massively parallel multiplex DNA
958 sequencing for specimen identification using an Illumina MiSeq platform. *Sci.*
959 *Rep.* 5:9687.

960 Smith M.A., Rodriguez J.J., Whitfield J.B., Deans A.R., Janzen D.H., Hallwachs W.,
961 Hebert P.D.N. 2008. Extreme diversity of tropical parasitoid wasps exposed
962 by iterative integration of natural history, DNA barcoding, morphology, and
963 collections. *Proc. Natl. Acad. Sci.* 105:12359–12364.

964 Srivathsan A., Baloğlu B., Wang W., Tan W.X., Bertrand D., Ng A.H.Q., Boey E.J.H.,
965 Koh J.J.Y., Nagarajan N., Meier R. 2018. A MinION™-based pipeline for fast
966 and cost-effective DNA barcoding. *Mol. Ecol. Resour.* 18:1035–1049.

967 Srivathsan A., Hartop E., Puniamoorthy J., Lee W.T., Kuttu S.N., Kurina O., Meier R.
968 2019. 1D MinION sequencing for large-scale species discovery: 7000 scuttle
969 flies (Diptera: Phoridae) from one site in Kibale National Park (Uganda)
970 revealed to belong to >650 species. *bioRxiv*:622365.

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

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

975 Stork N.E. 2018. How many species of insects and other terrestrial arthropods are
976 there on Earth? *Annu. Rev. Entomol.* 63:31–45.

977 Stork N.E., McBroom J., Gely C., Hamilton A.J. 2015. New approaches narrow
978 global species estimates for beetles, insects, and terrestrial arthropods. *Proc.*
979 *Natl. Acad. Sci.* 112:7519–7523.

- 980 Sultana S., Ali Md.E., Hossain M.A.M., Asing, Naquiah N., Zaidul I.S.M. 2018.
981 Universal mini COI barcode for the identification of fish species in processed
982 products. Food Res. Int. 105:19–28.
- 983 Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. 2013. MEGA6: molecular
984 evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30:2725–2729.
- 985 Tan D.S.H., Ang Y., Lim G.S., Ismail M.R.B., Meier R. 2010. From ‘cryptic species’ to
986 integrative taxonomy: an iterative process involving DNA sequences,
987 morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma*
988 (Sepsidae: Diptera). Zool. Scr. 39:51–61.
- 989 Taylor H.R., Harris W.E. 2012. An emergent science on the brink of irrelevance: a
990 review of the past 8 years of DNA barcoding. Mol. Ecol. Resour. 12:377–388.
- 991 Tsukihara T., Aoyama H., Yamashita E., Tomizaki T., Yamaguchi H., Shinzawa-Itoh
992 K., Nakashima R., Yaono R., Yoshikawa S. 1996. The whole structure of the
993 13-subunit oxidized cytochrome c oxidase at 2.8 Å. Science. 272:1136–1144.
- 994 Valan M., Makonyi K., Maki A., Vondráček D., Ronquist F. 2019. Automated
995 taxonomic identification of insects with expert-level accuracy using effective
996 feature transfer from convolutional networks. Syst. Biol. 0:1–20.
- 997 Wang W.Y., Srivathsan A., Foo M., Yamane S.K., Meier R. 2018. Sorting specimen-
998 rich invertebrate samples with cost-effective NGS barcodes: Validating a
999 reverse workflow for specimen processing. Mol. Ecol. Resour. 18:490–501.
- 1000 Will K.W., Rubinoff D. 2004. Myth of the molecule: DNA barcodes for species cannot
1001 replace morphology for identification and classification. Cladistics. 20:47–55.
- 1002 Yang C., Tan S., Meng G., Bourne D.G., O’Brien P.A., Xu J., Liao S., Chen A., Chen
1003 X., Liu S. 2018. Access COI barcode efficiently using high throughput Single-
1004 End 400 bp sequencing. .
- 1005 Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence
1006 data. Proc. Natl. Acad. Sci. 107:9264–9269.
- 1007 Yeates D.K., Zwick A., Mikheyev A.S. 2016. Museums are biobanks: unlocking the
1008 genetic potential of the three billion specimens in the world’s biological
1009 collections. Curr. Opin. Insect Sci. 18:83–88.
- 1010 Yeo D., Puniamoorthy J., Ngiam R.W.J., Meier R. 2018. Towards holomorphology in
1011 entomology: rapid and cost-effective adult-larva matching using NGS
1012 barcodes: Life-history stage matching with NGS barcodes. Syst. Entomol.
1013 43:678–691.
- 1014 Yu H.J., You Z.H. 2010. Comparison of DNA truncated barcodes and full-barcodes
1015 for species identification. Int. Conf. Intell. Comput.:108–114.
- 1016 Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation
1017 method with applications to phylogenetic placements. Bioinformatics.
1018 29:2869–2876.

Figure Captions

Table 1. Proportion of specimens congruent between morphospecies and mOTU clusters (20 empirical datasets/6 clade-based datasets) under the three stringency classes. Values in brackets represent an upper bound estimate of number of specimens causing conflict.

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

Figure 1. Positions of the mini-barcodes with established primers assessed in this study.

Figure 2. Performance of mini-barcodes along a sliding window (100-bp, 200-bp, 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 20 empirical datasets (rows) are sorted by average match ratio. The six clade-specific datasets from GenBank are grouped below. Match ratios above average are coloured green, while those below average are in pink. The highest match ratio values are indicated with white text.

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

1044 **Figure 5.** Proportion of conserved sites along the full-length barcode (sliding
1045 windows of 100-bp, 200-bp, 300-bp).

1046 **Table 1.** Proportion of specimens congruent between morphospecies and mOTU clusters (20 empirical datasets/6 clade-based
1047 datasets) under the three stringency classes. Values in brackets represent an upper bound estimate of number of specimens causing
1048 conflict.

<u>Objective clustering, 2% p-distance</u>							
Length	295	307	313	407	657	Average	Median
Midpoint	405	154	502	455	329		
Class I	77/57% (7475/41367)	76/56% (7511/42052)	77/57% (7275/40857)	77/57% (7246/40849)	77/58% (7118/40327)	77/57% (7325/41090)	77/57% (7372/40941)
Class II	90/75% (3283/23846)	90/75% (3333/24794)	90/75% (3131/23554)	90/76% (3078/23304)	90/76% (3006/23349)	90/75% (3166/23769)	90/76% (3112/23469)
Class III	91/82% (3021/17875)	91/81% (2954/18981)	91/82% (2822/17666)	91/82% (2813/17511)	91/82% (2759/17509)	91/82% (2874/17908)	91/82% (2862/17745)
<u>ABGD, P=0.001</u>							
Length	295	307	313	407	657	Average	Median
Midpoint	405	154	502	455	329		
Class I	75/55% (7916/42576)	76/55% (7685/42403)	75/55% (8169/42787)	73/53% (8583/45131)	69/47% (9833/49848)	74/53% (8437/44549)	74/54% (8126/43386)
Class II	89/75% (3305/24555)	89/75% (3442/24541)	89/75% (3336/24267)	89/74% (3526/25153)	88/72% (3901/26691)	89/74% (3502/25041)	89/74% (3382/24625)
Class III	91/81% (2981/18303)	91/81% (2906/18509)	90/81% (2984/18352)	90/80% (3041/18978)	89/78% (3482/20860)	90/80% (3079/19000)	91/81% (2927/18494)
<u>PTP</u>							
Length	295	307	313	407	657	Average	Median
Midpoint	405	154	502	455	329		
Class I	76/57% (7643/40692)	76/56% (7744/42608)	77/57% (7369/41588)	78/57% (7130/41056)	78/58% (7207/40234)	77/57% (7419/41236)	77/57% (7318/40839)
Class II	89/75% (3322/23946)	89/75% (3549/24598)	90/75% (3230/23857)	90/76% (3109/23583)	90/76% (3127/22912)	90/75% (3267/23779)	90/76% (3198/23659)
Class III	90/81% (3091/18208)	90/81% (3334/19189)	91/82% (2926/18179)	91/81% (2813/18968)	91/82% (2859/17317)	91/81% (3005/18372)	91/81% (2937/18193)

1049

1050 **Table 2.** Proportion of correct/incorrect/ambiguous/unidentified specimens in the “best close match” analysis.

Dataset	Published Primers									
	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
Canadian 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
Ecuadorian 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
German 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 American 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
German 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
Amazonian 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 American 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 Guianan 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
Pakistani 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
<i>Tanytarsus</i>	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 European 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
Congolese 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 American 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
Ecuadorian 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
Iberian 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

1051

Figure 1. Positions of the mini-barcode regions with established primers assessed in this study.

Figure 2. Performance of mini-barcodes along a sliding window (100-bp, 200-bp, 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 2%												ABGD P=0.001												PTP											
Length	94	130	145	164	189	295	307	313	407	657	Average	Length	94	130	145	164	189	295	307	313	407	657	Average	Length	94	130	145	164	189	295	307	313	407	657	Average
Midpoint	257	65	73	243	213	405	154	502	455	329		Midpoint	257	65	73	243	213	405	154	502	455	329		Midpoint	257	65	73	243	213	405	154	502	455	329	
<u>20 Empirical Datasets</u>												<u>20 Empirical Datasets</u>												<u>20 Empirical Datasets</u>											
Ecuadorian Geometridae	94	95	96	98	98	99	99	99	99	99	98	North Sea Molluscs	94	90	92	94	94	89	90	89	86	80	90	Great Barrier Reef Fish	89	85	87	91	91	95	90	95	96	96	91
Great Barrier Reef Fish	94	91	92	95	95	97	94	97	97	97	95	South China Sea Fish	92	78	83	89	93	94	92	93	91	90	89	Pakistani Lepidoptera	95	83	82	87	90	95	95	95	95	94	91
Pakistani Lepidoptera	88	87	90	91	91	95	93	95	95	97	92	Great Barrier Reef Fish	79	75	79	90	89	97	92	95	95	94	88	Ecuadorian Geometridae	14	89	91	94	97	98	97	96	99	98	87
South China Sea Fish	92	86	88	94	94	94	94	94	94	95	92	Canadian Echinoderms	87	78	76	91	91	87	91	95	92	89	88	North Sea Molluscs	78	71	82	94	96	86	92	86	93	93	87
South American Butterflies	83	89	88	90	90	92	93	92	93	93	90	Pakistani Lepidoptera	61	67	75	84	89	95	91	95	96	94	85	South China Sea Fish	86	75	77	87	88	92	89	89	92	92	87
Amazonian Moths	87	88	89	90	91	91	90	91	91	91	90	German EPT	85	84	85	86	87	83	85	84	78	67	82	Canadian Echinoderms	75	71	78	83	90	89	90	87	91	89	84
European Marine Fish	81	74	84	92	90	97	93	94	97	97	90	Ecuadorian Geometridae	48	58	57	96	70	99	99	99	99	97	82	German EPT	82	80	81	81	83	79	83	82	79	80	81
Canadian Echinoderms	81	89	85	91	88	91	91	91	94	95	90	South American Butterflies	60	70	77	84	84	89	91	91	91	81	82	North European Tachinidae	77	70	73	82	80	84	83	83	82	85	80
North Sea Molluscs	77	85	84	90	90	89	92	84	85	87	86	European Marine Fish	80	72	72	85	86	83	84	80	78	69	79	Amazonian Moths	32	88	38	89	88	91	89	91	90	90	79
German EPT	78	85	84	85	84	83	85	83	83	84	83	North American Birds	61	67	73	74	76	81	83	83	84	87	77	German Aranea & Opiliones	72	73	77	77	79	82	80	77	81	79	78
North European Tachinidae	81	81	82	81	83	83	83	83	83	84	82	Amazonian Moths	72	43	42	88	75	87	90	90	91	89	77	North American Birds	72	67	66	77	78	81	82	81	85	85	77
North American Birds	84	77	77	82	81	85	82	84	85	85	82	North European Tachinidae	71	60	66	77	80	77	84	82	82	84	76	French Guianan Earthworms	71	83	69	85	80	81	62	78	75	69	75
German Aranea & Opiliones	69	77	76	79	78	80	82	82	80	81	78	French Guianan Earthworms	78	89	93	81	81	73	80	70	59	55	76	European Marine Fish	71	60	61	71	78	80	74	79	79	80	73
Northwest Pacific Molluscs	72	74	73	77	77	78	77	77	78	78	76	Northwest Pacific Molluscs	76	75	75	77	77	76	77	76	76	63	75	Northwest Pacific Molluscs	72	67	68	75	74	72	74	76	77	78	73
North American Pyraustinae	61	56	64	69	77	74	75	78	77	84	71	German Aranea & Opiliones	76	18	79	85	82	71	74	73	65	52	67	Congolese Fish	57	44	50	67	69	68	64	68	71	69	63
Congolese Fish	63	61	60	68	70	72	66	70	71	70	67	Congolese Fish	68	44	55	67	71	70	63	67	71	65	64	Tanytarsus	51	58	64	61	58	68	63	64	66	63	62
French Guianan Earthworms	52	74	74	54	63	65	65	65	65	60	64	Tanytarsus	67	71	71	61	62	50	56	48	47	45	58	North American Pyraustinae	45	58	54	56	49	58	71	80	76	58	61
Iberian Butterflies	51	55	56	59	63	61	62	63	61	65	60	North American Pyraustinae	25	29	41	67	76	70	74	67	58	59	57	Ecuadorian Chrysomelidae	55	57	55	56	55	55	54	55	55	55	55
Ecuadorian Chrysomelidae	54	54	55	55	56	55	55	55	55	55	55	Ecuadorian Chrysomelidae	55	55	54	56	56	57	57	55	57	58	56	Iberian Butterflies	32	47	48	47	48	59	58	61	58	54	51
Tanytarsus	47	51	49	57	57	58	59	55	56	57	55	Iberian Butterflies	26	36	42	53	57	60	65	59	55	52	50	South American Butterflies	0	0	0	0	0	89	90	93	91	91	45
Consolidated:	77	78	79	82	82	83	83	83	83	84	81	Consolidated:	69	63	71	80	79	80	82	81	79	74	76	Consolidated:	63	68	66	74	75	81	81	82	82	82	76
<u>6 Clade-based Datasets</u>												<u>6 Clade-based Datasets</u>												<u>6 Clade-based Datasets</u>											
Coleoptera	76	78	78	80	80	80	81	80	80	81	79	Coleoptera	77	80	80	81	82	77	79	76	73	67	77	Coleoptera	74	73	73	78	78	81	81	80	81	80	78
Lepidoptera	72	73	74	77	78	80	79	80	81	81	78	Diptera	69	73	74	76	77	73	75	72	69	63	72	Arachnida	69	70	70	70	69	70	68	68	68	66	69
Diptera	72	74	74	76	76	77	78	76	77	77	76	Lepidoptera	42	54	62	74	77	79	78	79	77	71	69	Diptera	37	71	69	55	71	76	76	75	75	75	68
Arachnida	63	66	65	65	66	68	68	67	67	67	66	Arachnida	71	71	71	69	68	66	64	64	64	57	67	Lepidoptera	39	65	41	64	33	79	78	77	79	78	63
Actinopterygii	64	60	61	65	65	66	64	66	66	66	64	Actinopterygii	59	54	55	64	64	66	66	65	65	64	62	Hymenoptera	44	63	63	62	64	66	66	65	65	65	62
Hymenoptera	60	63	63	63	63	65	65	65	64	64	64	Hymenoptera	62	64	65	64	65	62	63	62	59	55	62	Actinopterygii	22	49	51	61	61	64	61	64	65	65	56
Consolidated:	69	70	70	72	73	74	73	73	73	74	72	Consolidated:	63	66	68	72	73	71	72	71	69	64	69	Consolidated:	45	64	59	65	61	73	72	72	73	72	66
Overall Consolidated:	70	72	72	74	75	76	75	75	75	76	74	Overall Consolidated:	65	65	68	74	74	73	74	73	71	66	70	Overall Consolidated:	50	65	61	67	64	75	74	74	75	74	68

1063

1064

1065

1066

1067

Figure 3. Match ratios across three different species delimitation methods. Mini-barcodes (columns) are sorted by primer length while the 20 empirical datasets (rows) are sorted by average match ratio. The six clade-specific datasets from GenBank are grouped below. Match ratios above average are coloured green, while those below average are in pink. The highest match ratio values are indicated with white text.

Figure 4. Comparison of species delimitation methods for full-length and mini-barcode 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).