Make every species count: FASTACHAR software for rapid determination of molecular diagnostic characters to describe species

Lucas M. Merckelbach¹* | Luísa M. S. Borges²*

¹Centre for Materials and Coastal Research, Helmholtz-Zentrum Geesthacht, Geesthacht, Germany
²L3 Scientific Solutions, Geesthacht, Germany

Correspondence
Lucas M. Merckelbach, Helmholtz-Zentrum Geesthacht, Centre for Materials and Coastal Research, Geesthacht, 21502, Germany.
Email: lucas.merckelbach@hzg.de

Abstract
Only a fraction of species found so far has been described, particularly cryptic species uncovered by molecular data. The latter might require the use of molecular data for its diagnosis, but it is important to make use of the diagnostic content of the molecular data itself. The molecular character-based model provides discrete molecular diagnostic characters within DNA sequences that can be used in species descriptions fulfilling the requirement of most codes of nomenclature for a character-based description of species. Here, we introduce FASTACHAR, a software developed to extract molecular diagnostic characters from one or several taxonomically informative DNA markers of a selected taxon compared with those of other taxa in a single step. The input data consist of a single file with aligned sequences in the fasta format, which can be created using alignment software such as MEGA or GENEIOUS. FASTACHAR is an easy-to-use software with a graphical interface. Thus, the software does not require the user to have any knowledge of the underlying programming environment (Python). We hope this software, based on the method proposed by Jörger and Schrödl (Frontiers in Zoology, 10, 59, 2013) to describe cryptic species, will encourage researchers to take the final step in taxonomy: the formal description of species. We propose the use of this method and FASTACHAR also for the inclusion of molecular data in the description of any species. FASTACHAR is released as open-source software under GNU General Public License V3 and is freely available for all major operating systems from https://github.com/smerckel/FastaChar.

KEYWORDS
cryptic species, DNA barcode, DNA markers, FASTACHAR, molecular diagnostic characters, molecular taxonomy, species descriptions

*Lucas Merckelbach and Luisa Borges should be considered joint first author.
1 | Introduction

Taxonomy encompasses identifying, discovering and formally describing species and other taxa (Valdecasas et al., 2008), with the major objective of inventory all life on earth. Although taxonomic research has been carried out for over 250 years, only a fraction of species found has been described (Coleman, 2015). The introduction of the DNA barcode concept by Hebert et al. (2003) has given a new breath to taxonomy, the so-called molecular taxonomy. In a strict sense, DNA barcoding aims at identification by comparing DNA barcodes of unknown specimens with DNA barcodes of a priori defined taxonomic entities in databases such as GenBank (https://www.ncbi.nlm.nih.gov/genbank/) or BOLD systems (http://www.boldsystems.org/). Several methods and software programs, referred to as Sequence Identification Engines (SIDEs), were developed to accomplish identification in a robust way (e.g., BLAST; Altschul et al., 1990; DNA-BAR; DasGupta, Konwar, Mándoiu, Shvartsman, 2005; BRONX; Little, 2011; BLOG 2.0; Weitschek et al., 2013). As an identification tool, DNA barcoding was considered extremely exciting and potent, but it was considered a poor approach to species discovery and description (Wheeler, 2004; Rach et al., 2008). However, there is no denying of the utility of DNA barcodes in species discovery by first flagging putative new species such as cryptic species, morphologically identical but genetically distinct, in many nominal taxa. Examples can be found in disparate groups of organisms such as Rotifera (Fontaneto et al., 2011), Mollusca (Borges et al., 2012; Jörg er et al., 2012; Neusser et al., 2011), Arthropoda (Hebert et al., 2004), fish (Azpelicueta et al., 2019) and turtles (Reid et al., 2011).

The discovery of putative new species is usually followed by molecular species delimitation. The goal is to build a taxonomic scheme for DNA sequence data, which can be single locus or multilocus, to infer a de novo delimitation of the putative new species, sometimes referred to as molecular operational taxonomic units (MOTUs; Blaxter et al., 2005). The distance-based method advanced by Hebert et al. (2003) has become the workhorse used in DNA barcoding (Reid et al., 2011), aiming for the detection of a ‘barcoding gap’ between intra- and interspecific variation (Hebert et al., 2004; Puillandre et al., 2012). However, this approach has been criticized because of the high variation and occurrence of overlap between intra- and interspecific variation in many groups of organisms (Meyer & Paulay, 2005). More recently, several analytical methods were developed for molecular species delimitations. Some support species delimitation with single-locus data, partitioning sequences into MOTUs without adopting a rigid sequence threshold (Kekkonen & Hebert, 2014). One popular approach is the general mixed Yule-coalescent (GMYC; Fujisawa & Barraclough, 2013; Pons et al., 2006), which is used to discriminate between population and speciation processes to delineate species. Other popular methods are the Automatic Barcode Gap Discovery (Puillandre et al., 2012), the Barcode Index Numbers (BIN; Ratnasingham & Hebert, 2013) and the Poisson tree processes (PTP; Kapli et al., 2017; Zhang, Kapli, Pavlidis, Stamatakis, 2013). The ABGD and BIN methods apply clustering algorithms to distinguish partitions in the genetic distance to delineate species that are clustered into MOTUs. The PTP method is based on a gene tree inferred from molecular sequences.

In the case of new species in cryptic complexes, the multilocus approach has been favoured to provide more information and stability in the species descriptions (Neusser et al., 2011; Jörg er et al., 2012; Borges & Merckelbach, 2018; Trevisan et al., 2020). The multilocus approach will also address the well-known fact that gene trees are not necessarily identical to species trees, for instance because of incomplete lineage sorting or horizontal gene transfer (Jörg er & Schrödl, 2013). Several methods and software programs have been developed to analyse multilocus data, for example the Bayes factor delimitation (Grummer, Bryson, Reeder, 2014; Leaché, Fujita, Minin, & Bouckaert, RR, 2014), and phylogeographic inference using approximate likelihoods (Jackson, Carstens, Morales, & O’Meara, 2017). But by far the most popular software is the Bayesian Phylogenetics and Phylogeography (BPP; Yang & Rannala, 2010; Yang, 2015), a Bayesian Markov chain Monte Carlo program for analysing multilocus sequence data under the multispecies coalescent (MSC) model with and without introgression (Flouri et al., 2020). For reviews and comparisons of methods for molecular species delimitation, see, for instance, Fujita et al. (2012), Sauer & Hausdorf (2012), Carstens et al. (2013), and Luo et al. (2018).

Many newly discovered species remain, however, undescribed even in cases where the species hypothesis is highly supported by numerous lines of evidence (Pante et al., 2015), hindering estimates of biodiversity, conservation efforts and attempts to control diseases and invasive species (Schlick-Steiner et al., 2007). Taxonomy remains incomplete if species are flagged as merely putative rather than formally described and thus full established. In many cases, the transition from species delimitation to species descriptions is the major task to be achieved (Jörg er & Schrödl, 2013). Recently, new methods have been proposed to speed up species description using a combination of molecular data, concise morphological descriptions and other sources of information such as geographic distribution (Butcher et al., 2012; Summers et al., 2014). However, molecular data have been included in species descriptions in various forms, such as lists of GenBank accession numbers, inclusion of entire DNA barcode sequences or presentation of raw distance measures, but seldom making use of the diagnostic content of the molecular data itself (Goldstein & DeSalle, 2011; Jörg er & Schrödl, 2013).

In this context, the molecular character-based model proposed by DeSalle et al. (2005) provides discrete molecular diagnostic characters (MDCs) within DNA sequences that are especially useful for the diagnosis of species, in particular of cryptic species that cannot make use of morphological characters for diagnosis (Jörg er & Schrö dl, 2013). The molecular diagnosis involves the listing of the DNA (usually from a combination of mitochondrial and nuclear markers) or protein characters of a species which are different (apomorphic) from that of its closest relatives (e.g., congeneric species). This allows for a direct comparison between species and standardizes the use of molecular data (Goldstein & DeSalle, 2011), which can be used as diagnostic features in a similar fashion as traditional morphological diagnostic characters are used for
species diagnosis (Bauer et al., 2011; Bergmann et al., 2009). In addition, using MDCs allows the detection of differences that reflect lineage independence, decreasing taxonomic ambiguity and instability (Bauer et al., 2011).

In this work, we present fastachar, a software package designed to extract MDCs from one or several taxonomically informative DNA markers of a selected taxon compared with those of other taxa in a single step. The software was developed to determine MDCs for the description of *Lyrodus mersinensis* (Borges & Merckelbach, 2018) because none of the existing software packages met our needs.

2 | A Brief Review of Existing Software

The Character Attribute Organization System (CAOS; Sarkar et al., 2008) was developed following previous research (Sarkar et al., 2002; Rach et al., 2008) for species identification. However, it was pointed out by the authors that the first function (P-Gnome) could also be useful to establish diagnostics from DNA sequences for species descriptions (Sarkar et al., 2008). However, being not designed for this purpose, the workflow to determine MDCs requires a time-consuming iterative process (Jörger & Schrödl, 2014). Furthermore, Kühn and Haase (2019) note that the software considers masked entries in the DNA as valid characters, yielding potentially incorrect results. SPIDER (Brown et al., 2012) is an open-source software package implemented in r and provides a number of functions for use in analyses of DNA barcoding studies, including determining MDCs (Azpelicueta et al., 2019; Escobar et al., 2019). SPIDER, however, does not adhere to the definition of MDCs as proposed by Jörger and Schrödl (2013) and does not treat masked entries in the DNA correctly, as pointed out by Kühn and Haase (2019). In addition, the use of SPIDER requires a workable knowledge of the programming environment R, and lacking this, the time and effort required to obtain literacy in R seem disproportional.

A third software package, quiddich (Kühn & Haase, 2019), is presented as a tool specifically for determining MDCs, but this too, being an r package, presents a conundrum for non-R users. Compared to CAOS and fastachar, for example, quiddich defines two additional types of molecular diagnostic characters. These types (3 and 4) are plesiomorphic (not unique to the species being analysed) and therefore cannot be used as diagnostic or defining characters of that species. Thus, they are not meaningful for the purpose of a formal description and do not adhere to the definition proposed by Jörger and Schrödl (2013) either.

3 | Overview of fastachar

3.1 | Implementation

fastachar is implemented as a package for the scripting language Python. We chose Python mainly because (a) Python code is easy to read by design, so that nonexperts in programming still can work out how the algorithms work, (b) the Python environment can be installed easily, and iii) the Python environment includes by default a robust cross-platform toolkit to create graphical user interfaces.

The fastachar package consists of a number of Python modules, which implement the necessary data types and functionality to read and compare DNA sequences, and report the results. Using these modules, simple Python scripts can be written that process pre-aligned DNA sequences, generating output in text or spreadsheet files for further analysis. A cross-platform (Linux/Mac/Windows) graphical user interface (GUI) is implemented in an additional module, offering a workflow that does not require knowledge of Python. The software package is published as open-source software under the GNU Public License version 3 (GPL v3). The source code and manual can be obtained from https://github.com/smerckel/fastachar or https://www.cubic-lscience/. The software version number of fastachar as used in this manuscript is 0.2.4 (Merckelbach & Borges, 2020).

3.2 | Algorithms

Within the context of fastachar, MDCs, as defined by Jörger and Schrödl (2013), are determined by comparing two lists consisting of sequences of a single alignment. List A contains sequences belonging to one taxon A, whereas list B can contain those of one or more taxa. Per position k in the alignment, (mathematical) sets A_k and B_k are created from the characters of the sequences in the lists A and B, respectively. The characters specified in the IUPAC ambiguity code for nucleotides (http://www.bioinformatics.org/sms/iupac.html) are expanded into all the base nucleotides they represent. For example, a set containing M expands into a set containing both A and C, and so does N expand to A, C, G and T.

3.2.1 | Evaluation criteria for molecular diagnostic characters

FastaChar’s algorithm yields an MDC for taxon A at position k, if and only if the following conditions are met.

1. A_k contains one character only and
2. None of the elements in set B_k appear in A_k

---

1 A Python module can be considered as a library of functions and classes that are meant to be called from other programs.

2 A mathematical set is a collection of distinct elements. This means that if a list of a number of identical elements is converted into a set, then the set will contain only one such an element.
This is equivalent to the ‘single pure characters’ used in CAOS (Sarkar et al., 2008) and ‘type 1 diagnostic characters’ used in QUIDDICH (Kühn & Haase, 2019).

3.2.2 | Evaluation criterion for potential molecular diagnostic characters

If a taxon A has one or more ambiguous character codes in a given position $k$, then the position disqualifies to yield a molecular diagnostic character. In some of these cases, however, the user might still consider taxon A to have an MDC at this position. To single out these cases for further inspection, a second algorithm yields the potential MDCs by replacing Condition 1 by the condition

1. $A_k$ contains at least two characters

This is equivalent to the ‘private characters’ used in CAOS (Sarkar et al., 2008) and ‘type 2 diagnostic characters’ used in QUIDDICH (Kühn & Haase, 2019).

3.2.3 | Noninformation characters

Not all sequences obtained for a certain marker are of the same length because of low-quality readings, usually at the beginning and/or at the end of the sequence. After aligning, the shorter sequences are padded, yielding continuous blocks of noninformative dash characters at the beginning and/or the end of the sequence. Masking programs are also sometimes used to mask parts of sequences that are highly variable or of bad quality, using (mostly) N as masking character. To avoid noninformative characters to be treated as informative, which could lead to erroneous results, a dash (or N) is excluded from the sets $A_k$ and $B_k$ if it appears in a continuous block of dashes (or N’s) at the start and/or at the end of a sequence.

4 | FASTACHAR workflow

4.1 | Preliminary considerations on how to use molecular data in species descriptions

The use of molecular data as the partial or even as the major source of a species description needs consideration on the best practice to follow. It is important to keep in mind that each species description is a hypothesis (not a fact) about the discontinuous distribution of unique combination of characters (Nixon & Wheeler, 1990; Valdecasas et al., 2008) may they be morphological and/or molecular. The final decision of the taxonomist may always be falsified and is testable forever (Haszprunar, 2011). Given that many species have wide and complex ranges, to sample thousands of specimens from many populations of several species can be done but is prohibitively time-consuming and cost-inefficient (Wheeler, 2004). Thus, when using molecular sequence data (or morphological data) for species descriptions the hypotheses of the apomorphic characters (derived traits referred to in the FASTACHAR context as MDCs) are theorized (Haszprunar, 2011) and may contain errors in the form of incorrectly assumed apomorphic character states (Jörger & Schrödl, 2013). This may be the case when working with sparsely sampled species or when there is a low number of sequences per species. When new sequences, for instance from new cryptic species or other closely related congeneric species, are added to the analysis at a later stage, the putative molecular apomorphies of described species may have to be reconsidered as plesiomorphies (ancestral traits not unique to the species being analysed). It may also happen that when sequence data of new specimens of the same species are included in the analysis, some putative molecular apomorphies may vanish in intraspecific variation. Thus, it is important to include in the analysis two or more independently evolving markers (at least two independent informative markers per species) and several sequences of each species to account for intraspecific variation. This is to ensure that at least some of the potentially apomorphic nucleotides found in different markers are truly unique mutations accumulated due to the absence of gene exchange (Jörger & Schrödl, 2013).

4.2 | How to select and prepare data for the analysis

The most important decision in the analyses using FASTACHAR is to select the most adequate evolutionary group which will serve for comparison. One possibility is to compare the species of interest with some or all its congeners. If that is not possible, it can be compared with other congeneric species. In the latter case, however, care must be taken not to include distant congeneric species in the comparison. When distant species are compared, there is a higher probability of including homoplastic character states, which increases with evolutionary distance (Rach et al., 2008) given that there are only four potential possibilities in a homologous position of a nucleotide alignment. This is particularly important when dealing with fast-evolving markers. On the other hand, if a species is only compared to its known direct sister species this increases the risk of including plesiomorphies as MDCs. It also depends entirely on the validity of this phylogenetic sister group relationship (Jörger & Schrödl, 2014).

After selecting the group to be analysed, the sequences must be aligned with a dedicated alignment program, such as MUSCLE (Edgar, 2004) or GENEIOUS (Kearse et al., 2012). Since the determination of the diagnostic characters relies on the positional homology assumption, it is important to compare the results of different alignment programs in FASTACHAR and choose the one that takes the most conservative approach. This results in fewer diagnostic characters, but their reliability will be higher (Jörger & Schrödl, 2013). The aligned
sequences are then exported into a standard fasta file format, where each sequence must be accompanied by a species name, and any associated laboratory code or GenBank information.

4.3 | How to process the data using fastachar's GUI

When fastachar is started, a GUI is brought up. The user first opens the fasta file prepared in the previous step, and all species contained are listed in text field ‘Unselected Species’ (see Figure 1). From this field, the required species are dragged into the text fields of ‘Selected species list A and B’, respectively. Then, the desired operation is selected choosing between determining MDCs or potential MDCs. After clicking ‘Process’ button, the results are shown in the bottom text field, which allows for a direct inspection of the results. If desired, the results can be saved as an ASCII file or as a spreadsheet file for further analysis.

4.4 | Demonstration of fastachar using six data sets with several taxonomically informative markers

The software can be used to determine MDCs from one or several taxonomically informative molecular markers for any taxon. Thus, below we provide examples of determination of MDCs from several molecular markers in disparate taxa including algae, fungi, invertebrates and vertebrates. Each data set is associated with a published peer-reviewed article (Table 1). The results are shown in Annex 1.

For demonstrational purposes only, we determine MDCs for only one species in each data set (Annex 1). The 18S rRNA and 28S rRNA sets, from the work by Jörger and Schrödl (2013), were used to compare the performance of fastachar with that of caos (Sarkar et al., 2008). Similarly, we used the COI data set from Azpelicueta et al. (2019) to compare the performance of fastachar with that of spider (Brown et al., 2012). The MDCs determined by fastachar were identical to those determined by caos and spider.

![FIGURE 1](image_url)

In this example of a fastachar main window, an alignment of sequences of 10 species is loaded. Then, the number of sequences per species is denoted in parentheses. The MDCs of Lyrodus mersinensis (list A) are determined by a comparison with the species in list B. The result is (partly) shown in the text area at the bottom of the main window, showing for example that Lyrodus mersinensis has a molecular diagnostic character C at position 61, while sequences from other species have either a G a T or A at this position.
An application to the COI sets in the Supplement of Kühn & Haase (2019) yielded identical results for fastachar and quiddich, except for one position where fastachar finds a MDC not reported by quiddich. The discrepancy is triggered by a single sequence out of 50 that is padded with trailing masking Ns. Since trailing masking Ns do not contain information, fastachar’s algorithm ignores this sequence for the positions where it has masking Ns (see also Section 3.2, non-informative characters), allowing it to find one more MDC from the comparison of the 49 remaining sequences. Quiddich on the other hand does not make the distinction between informative and non-informative characters and therefore discards a priori all MDCs on positions where one or more sequences have leading or trailing masking Ns.

### 4.5 Presentation of the results

The presentation of the results obtained from the analysis in fastachar is of great importance. This is particularly the case when MDCs are the partial or even as the major data source used in a species description. Thus, the researcher has to ensure that the data are traceable, reproducible and testable. In their method, Jörger & Schrödl (2013) advised that to make a diagnostic position in a sequence traceable and the results reproducible by other researchers the following information should be included: (a) primers and alignment program used; (b) how position 1 was determined (for instance the first base after the primer sequence); (c) deposit the sequences in public databases (e.g., GenBank or the BOLD system or as additional material in the article); and (d) in new species descriptions the provided reference sequences should be generated from type material. For further details, see Jörger & Schrödl (2013). We further advise that MDCs and respective positions should be presented not only for the species of interest but also for all species to which it was compared (see Tables S1–S6 in Annex). This will ensure that future researchers can reproduce and test the results.

### 5 Discussion and Conclusions

Molecular taxonomy is a burgeoning field, and several methods and software programs have been developed to facilitate the three major processes: identification, molecular species delimitation (discovery) and formal descriptions of species. Most programs are used only for one of these functions, for example BLOG 2.0 for identification (Weitschek et al., 2013) BPP for delimitation (Yang & Rannala, 2010). But others such as spider (Brown et al., 2012) and caos (Sarkar et al., 2008) may be used in more than one process. For example, spider can be used for identification, delimitation and description, while caos was designed for species identification but it can be also used for species description. In this context, fastachar could be viewed as a software not only for species description but also for species delimitation. This is because the presence of MDCs may reflect differences that indicate separately evolving lineages (species) according to the concept of species proposed by De Queiroz (2007). However, to use fastachar for species delimitation would be inferior to the use of, for example, a coalescent-based delimitation method such as BPP (Yang & Rannala, 2010), as there is no objective criterion on the number of MDCs to delimit a species (Jörger & Schrödl, 2013). In addition, there is no simple formula that can predict the length of the sequence that must be analysed to ensure species diagnosis because rates of molecular evolution vary between different segments of the genome and across taxa (Hebert et al., 2003). Therefore, fastachar should be used only to extract MDCs to formally describe species, when hypotheses for those species were congruently supported across delimitation methods. This is consistent with the principles of integrative taxonomy, in that taxonomic inference should be based on congruence across analyses that utilize multiple sources of data (Carstens et al., 2013).

Software packages to extract MDCs exist, but either have a cumbersome and time-consuming workflow, or require a working knowledge of the programming environment R, which not everyone may have. This may be one of the reasons why in many species’ descriptions molecular data are used but not in the form of MDCs.

Fastachar is easy and quick to install, and runs on all major operating systems. The intuitive graphical user interface provides a convenient way to determine the MDCs from any number of sequences of a selected taxon compared with those of other taxa (as many as required by the user) in a single step, whereas it removes the requirement to have any knowledge of any programming environment. Therefore, we hope it helps to standardize the use of molecular data and stimulate researchers to proceed to the final step of molecular taxonomy, that is describe the new species, particularly cryptic
species, after the exploratory step of delimitation, making them available for biodiversity research.

ACKNOWLEDGEMENTS
We thank the three anonymous reviewers for their constructive feedback that helped improving the manuscript.

AUTHORS’ CONTRIBUTIONS
L.M.M. wrote the software. L.M.S.B. tested the software for different taxa and molecular markers. Both authors wrote, reviewed and accepted the manuscript.

ORCID
Lucas M. Merckelbach https://orcid.org/0000-0003-4677-8670
Luisa M. S. Borges https://orcid.org/0000-0001-8582-0013

REFERENCES
Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410.

Azpelcueta, M. M., Delpiani, S. M., Cione, A. L., Oliveira, C., Marceníuk, A. P. et al (2019). Morphology and molecular evidence support the validity of Pogonias courbina (Lacepède, 1803) (Teleostei: Sciaenidae), with a redescription and neotype designation. PLoS ONE, 14, e0216280.

Bauer, A. M., Parham, J. F., Brown, R. M., Stuart, B. L., Grismer, L. et al (2011). Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. Proceedings, Biological Sciences/The Royal Society, 278, 490–492.

Bergmann, T., Hadrys, H., & Breves, G. (2009). Character-based DNA barcoding: a superior tool for species classification. Berliner Münchener Tierärztliche Wochenchrift, 122, 446–450.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C. et al (2005). Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society B: Biological Sciences, 360, 1935–1943.

Borges, L. M. S., & Merckelbach, L. M. (2018). Lyrodus mersinius sp. nov. (Bivalvia: Terediniidae) another cryptic species in the Lyrodus pedicelatus (Quatrefages, 1849) complex. Zootaxa, 4421, 441–457.

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M. C., Malumbres-Olarte, J. et al (2012). SPIDER: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources, 12, 562–565.

Bucher, B. A., Smith, M. A., Sharkey, M. J., & Quicke, D. L. J. (2012). A turbo-taxonomic study of Thai Aleiodes (Aleiodes) and Aleiodes (Aralaeides) (Hymenoptera: Braconidae: Roganiidae) based largely on COI barcoded specimens, with rapid description of 179 new species. Zootaxa, 3457, 1–232.

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. Molecular Ecology, 22, 4369–4383.

Coleman, C. O. (2015). Taxonomy in times of the taxonomic impediment - Examples from the community of experts on amphipod crustaceans. Journal of Crustacean Biology, 35, 729–740.

DasGupta, B., Konwar, K. M., Mândouli, I., & Shvartsman, A. A. (2005). DNA-BAR: Distinguisher selection for DNA barcoding. Bioinformatics, 21, 3424–3426.

De Queiroz, K. (2007). Species concepts and species delimitation. Systematic Biology, 56, 879–886.

DeSalle, R., Egan, M. G., & Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences, 360, 1905–1916.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5, 113.

Escobar, L. M. D., Ota, R. P., Machado-Allison, A., Andrade-López, J., Farias, I. P., & Hrbek, T. (2019). A new species of Piaractus (Characiformes: Serrasalmidae) from the Orinoco Basin with a re-description of Piaractus brachypomus. Journal of Fish Biology, 95, 411–427.

Feau, N., Vialle, A., Allaire, M., Tanguay, P., Joly, D. L. et al (2009). Fungal pathogen (mis)identifications: A case study with DNA barcodes on Melampsora rusts of aspen and white poplar. Mycological Research, 113, 713–724.

Flouri, T., Rannala, B., & Yang, Z. (2020). A tutorial on the Use of BPP for species tree estimation and species delimitation. In F. Delsuc, N. Galtier, C. Scornavacca (Ed.), Phylogenetics in the genomic era (pp. 1–16).

Fontaneto, D., Iakovenko, N., Eyres, I., Kaya, M., Wyman, M. et al (2011). Cryptic diversity in the genus Adinia Hudson & Gosses, 1886 (Rotifera: Bdelloidea: Adinetidae): A DNA taxonomy approach. Hydrobiologia, 662, 27–33.

Fujiwara, T., & Barraclough, T. G. (2013). Delimiting species using single-locus data and the generalized mixed yule coalescent approach: A revised method and evaluation on simulated data sets. Systematic Biology, 62, 707–724.

Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., & Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology and Evolution, 27, 480–488.

Goldstein, P. Z., & DeSalle, R. (2011). Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. BioEssays, 33, 135–147.

Gummer, J. A., Bryson, R. W., & Reeder, T. W. (2014). Species delimitation using Bayes factors: Simulations and application to the sceloporus scalaris species group (Squamata: Phrynosomatidae). Systematic Biology, 63, 119–133.

Haszprunar, G. (2011). Species delimitations—not “only descriptive”. Organisms Diversity and Evolution, 11, 249–252.

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society B: Biological Sciences, 270, 596–599.

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America, 101, 14812–14817.

Jackson, N. D., Carstens, B. C., Morales, A. E., & O’Meara, B. C. (2017). Species delimitation with gene flow. Systematic Biology, 66, 799–812.

Jörgér, K. M., Norenburg, J. L., Wilson, N. G., & Schrödl, M. (2012). Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. BMC Evolutionary Biology, 12, 245.

Jörgèr, K. M., & Schrödl, M. (2013). How to describe a cryptic species? Practical challenges of molecular taxonomy. Frontiers in Zoology, 10, 59.

Jörgèr, K. M., & Schrödl, M. (2014). How to use COAS software for taxonomy? A quick guide to extract diagnostic nucleotides or amino acids for species descriptions. Spxiana, 37, 21–26.

Kapli, P., Lutteropp, S., Zhang, J., Kober, K., Pavlidis, P., Stamatakis, A., & Flouri, T. (2017). Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. Bioinformatics, 33, 1630–1638.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). **generous basic**: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647–1649.

Kekkonen, M., & Hebert, P. D. N. (2014). DNA barcode-based delimitation of putative species: Efficient start for taxonomic workflows. *Molecular Ecology Resources*, 14, 706–715.

Kühn, A. L., & Haase, M. (2019). **Quick IDentification of Diagnostic ChAracters. Journal of Zoological Systematics and Evolutionary Research**, 00, 1–5.

Lakra, W. S., Verma, M. S., Goswami, M., Lal, K. K., Mohindra, V., Punia, P., Gopalakrishnan, A., Singh, K. V., Ward, R. D., & Hebert, P. (2011). DNA barcoding Indian marine fishes. *Molecular Ecology Resources*, 11, 60–71.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP Data. *Systematic Biology*, 63, 534–542.

Little, D. P. (2011). DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS ONE*, 6.

Luo, A., Ling, C., Ho, S. Y. W., & Zhu, C. D. (2018). Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Systematic Biology*, 67, 830–846.

Merckelbach, L. M., & Borges, L. M. S. (2020). **fastachar**: software for rapid determination of molecular diagnostic characters to describe species, v0.2.4. https://doi.org/10.5281/zenodo.3906478

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, 3, e422.

Nakada, T., Soga, T., & Nozaki, H. (2007). A taxonomic study of Hafnimonas (Chlorophyceae) based on a comparative examination of cultured material. *Journal of Phycology*, 43, 397–411.

Nakada, T., Soga, T., Tomita, M., & Nozaki, H. (2010). *Chlorogonium complex* sp. nov. (Volvocales, Chlorophyceae), and morphological evolution of *Chlorogonium*. *European Journal of Phycology*, 45, 97–106.

Neisser, T. P., Jörger, K. M., & Schrödl, M. (2011). Cryptic species in tropic sands - interactive 3D anatomy, molecular phylogeny and evolution of meiofaunal Pseudundulidae (Gastropoda, Acochlidiida). *PLoS ONE*, 6, e23313.

Nixon, K. C., & Wheeler, Q. D. (1990). An amplification of the phylogenetic species concept. *Cladistics*, 6, 211–223.

Nozaki, H., Ohta, N., Morita, E., & Watanabe, M. M. (1998). Toward a natural system of species in *Chlorogonium* (Volvocales, Chlorophyta): a combined analysis of morphological and rbcL gene sequence data. *Journal of Phycology*, 34, 1024–1037.

Pante, E., Schoelinx, C., & Puillandre, N. (2015). From integrative taxonomy to species description: One step beyond. *Systematic Biology*, 64, 152–160.

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., ... Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55, 595–609.

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

Rach, J., DeSalle, R., Sarkar, I. N., Schierwater, B., & Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences*, 275, 237–247.

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE*, 8, e23313.

Reid, B. N., Le, M., McCord, W. P., Iverson, J. B., Georges, A. et al. (2011). Comparing and combining distance-based and character-based approaches for barcoding turtles. *Molecular Ecology Resources*, 11, 956–967.

Sarkar, I. N., Planet, P. J., & Desalle, R. (2008). **cado**: software for use in character-based DNA barcoding. *Molecular Ecology Resources*, 8, 1256–1259.

Sarkar, I. N., Thornton, J. W., Planet, P. J., Figurski, D. H., Schierwater, B. & DeSalle, R. (2002). An automated phylogenetic key for classifying homeoboxes. *Molecular Phylogenetics and Evolution*, 24, 388–399.

Sauer, J., & Hausdorf, B. (2012). A comparison of DNA-based methods for delimiting species in a Cretan land snail radiation reveals short-comings of exclusively molecular taxonomy. *Cladistics*, 28, 300–316.

Schlick-Steiner, B. C., Seifert, B., Stauffer, C., Christian, E., Crozier, R. H., & Steiner, F. M. (2007). Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in Ecology and Evolution*, 22, 391–392.

Summers, M. M., Al-Hakim, I. I., & Rouse, G. W. (2014). Turbo-taxonomy: 21 new species of *Myzostomida* (Annelida). *Zootaxa*, 3873, 301–344.

Trevisan, C. C., Batalha-Filho, H., Garda, A. A., Menezes, L., Dias, I. R., Solé, M., ... Napoli, M. F. (2020). Cryptic diversity and ancient diversification in the northern Atlantic Forest Pristimantis (Amphibia, Anura, Craugastoridae). *Molecular Phylogenetics and Evolution*, 148, 106811.

Valdecasas, A. G., Williams, D., & Wheeler, Q. D. (2008). “Integrative taxonomy” then and now: A response to Dayrat (2005). *Biological Journal of the Linnean Society*, 93, 211–216.

Weitschek, E., Van Velzen, R., Felici, G., & Bertolazzi, P. (2013). **blog**: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. *Molecular Ecology Resources*, 13, 1043–1046.

Wheeler, Q. D. (2004). Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359, 571–583.

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

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, 61, 854–865.

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

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.