Syst. Biol. 63(6):1005–1009, 2014

DOI:10.1093/sysbio/syu060
Advance Access publication August 12, 2014

Known Knowns, Known Unknowns, Unknown Unknowns and Unknown Knowns in DNA Barcoding: A Comment on Dowton et al.

RUPERT A. COLLINS[1]* AND ROBERT H. CRUICKSHANK[2]

[1]*Laboratório de Evolução e Genética Animal, Departamento de Biologia, Universidade Federal do Amazonas, Av. Rodrigo Otávio, Manaus, Amazonas, Brazil and* [2]*Department of Ecology, Faculty of Agriculture and Life Sciences, Lincoln University, Lincoln 7647, Canterbury, New Zealand*
**Correspondence to be sent to: Departamento de Biologia, Universidade Federal do Amazonas, Av. Rodrigo Otávio, Manaus, Amazonas, Brazil; E-mail: rupertcollins@gmail.com*

Received 5 June 2014; reviews returned 1 August 2014; accepted 4 August 2014
Associate Editor: Tanya Stadler

In a recent commentary, Dowton et al. (2014) propose a framework for "next-generation" DNA barcoding, whereby multilocus data sets are coupled with coalescent-based species delimitation methods to make specimen identifications. They claim single-locus DNA barcoding is outdated, and a multilocus approach superior, with their assertions supported by an analysis of 33 species of *Sarcophaga* flesh flies. Here, we reanalyze their data and show that a standard DNA barcode analysis is in fact capable of identifying 99.8% (all but one) of their *Sarcophaga* specimens, and that their conclusions misrepresent their data. We also discuss the benefits and drawbacks to their vision of "next-gen" barcoding.

Identifying Biological Specimens using DNA Barcodes

DNA barcoding for identifying biological specimens relies on a reference library of sequences with *a priori* identifications. Adult life stages are identified using diagnostic morphological characters, and then the DNA barcodes from immature or incomplete specimens can be matched to these, a technique highly applicable to accurate forensic identification of Diptera larvae (Meiklejohn et al. 2011). However, the breadth and depth of the reference library has implications for the degree of identification success that can be obtained from it (Bergsten et al. 2012; Virgilio et al. 2012; Zhang et al. 2012b). Once the DNA barcode library has been assembled, the sequences contained and not contained in the library can be formulated as follows: "known knowns" (described and identified species with DNA barcodes in the library); "known unknowns" (described species without DNA barcodes in the library); "unknown knowns" (divergent lineages among the described species present in the library = possibly cryptic or overlooked species); and "unknown unknowns" (undescribed or cryptic species without DNA barcodes in the library). This formulation helps us characterize an incompletely sampled reference library.

Sequence data generated from query specimens can then be identified using the library by sequence similarity matching (Meier et al. 2006), usually in a simulated scenario (database members acting as both reference library and query sequences). When there is 100% genetic similarity between the query and a matching, unambiguous sequence(s), the identification is trivial. However, in the absence of complete haplotype sampling, nonidentical sequences pose problems in differentiating intra- from inter-specific variation (Virgilio et al. 2012), and obtaining a correct identification or correct nonidentification—if no match is present for singletons—relies upon heuristic solutions. Early studies on DNA barcoding (e.g., Hebert et al. 2004) proposed a "ten times" rule to determine a threshold value of genetic distance—interspecific variation being 10 times or more the intraspecific variation—but this was later discredited (Hickerson et al. 2006). More often, thresholds are generated by taking a database of specimens with a known taxonomy, and calculating the threshold that minimizes the cumulative identification failure incorporating false-positive error (no matches within the threshold value but conspecific samples available) and false-negative error (more than one species recorded within threshold) (Meyer and Paulay 2005). Based on this approach, the Barcode of Life Data System (BOLD) Barcode Index Number (BIN) algorithm (Ratnasingham and Hebert 2013) uses 2.2% *p*-distance threshold (with subsequent refinement using Markov clustering). However, in many studies, unjustified and arbitrary threshold values are frequently copied from the literature without the consideration of alternatives—a practice that was criticized by Collins and Cruickshank (2013)—despite the availability of easy-to-use tools to optimize thresholds empirically (Brown et al. 2012; Puillandre et al. 2012b; Virgilio et al. 2012; Sonet et al. 2013).

A Reanalysis of Dowton et al.

Dowton et al. (2014) (hereafter Dowton et al.) reported that single-locus COI barcodes were able to identify just 27 of 31 *Sarcophaga* species (87%). When compared to the 100% correct identification using the multispecies coalescent (hereafter 'MSC') software BPP (Yang and Rannala 2010), this appears to validate their claim that multilocus data sets and highly parameterized analyses are indeed required. However, for their standard DNA barcode analysis they used a K2P divergence threshold of 4%, but no references were provided to support the application of this value to their taxa. We reanalyzed the data of Dowton et al. (http://dx.doi.org/10.5061/dryad.st467, last accessed May 27, 2014) to establish if 4% was an accurate threshold for distinguishing inter- and intraspecific variation. First, we used the taxonomic-name-based threshold optimization technique *threshOpt* in the Spider package for R (Brown et al. 2012), followed by a name-free estimation of discontinuities in amalgamated genetic distances using the *localMinima* function, also in Spider (full details to repeat our analysis is presented as an R script in Supplementary Material available on Dryad; http://dx.doi.org/10.5061/dryad.h3g63). The two methods agreed on an optimum threshold of 2.13% (*p*-distance), and therefore lower than the 4% used by Dowton et al. We then applied this optimized threshold to simulate the identification of the 405 specimens—from the 33 *Sarcophaga* species used in their "testing" data set—with the "best close match" (BCM) technique of Meier et al. (2006).

We report that 401 of 405 specimens were either correctly identified to species (if conspecifics were present in the library), or correctly reported as having no conspecific in the library (i.e., singletons). Of the four misidentified specimens, one (KM592)—considered as *S. bancroftorum* "clade 1" according to the taxon labeling in the data set provided and Table 1 of Dowton et al.—was identified here as *S. taenionota*. However, this appears to be simply a mislabeled sequence, as the same individual is labeled as *S. taenionota* in Fig. 2. of Meiklejohn et al. (2013). Three further specimens could not be identified despite having conspecifics in the library; it was proposed by Meiklejohn et al. (2012) that sequences from two of these specimens (*S. omikron* KM311 and *S. spinigera* KM260) could represent paralogous nuclear copies, which might explain the large divergences from their closest conspecifics (5.73 and 2.51%, respectively). However, upon examining the alignment and the trace files on BOLD, a more simple explanation is apparent, whereby sequencing errors due to "dye blobs" have introduced an additional nucleotide, thus frame-shifting part of the alignment and causing the divergences. When these errors are corrected, the sequences are identified correctly (Supplementary Material available on Dryad; http://dx.doi.org/10.5061/dryad.h3g63). The remaining individual (*S. australis* KM673) appears to be the only genuine case of an unidentifiable individual, being 3.35% divergent from its closest putative conspecific. Thus, excluding the probable errors, this represents a 99.8% success rate, and calls into question the 87% success rate reported by Dowton et al. as justification for the superiority of the MSC. Importantly, the single individual that could not be placed (*S. australis* KM673) was not included in the BPP analysis of Dowton et al., so it cannot currently be established if this specimen would have been included or not as the correct species by the MSC approach.

This discrepancy between success rates reported here and in Dowton et al. is due to two factors: 1) their identification method was applied to species rather than individuals, a general problem in much of the DNA barcoding literature whereby species discovery is conflated with specimen identification (Collins and Cruickshank 2013); and 2) the use of an arbitrary 4% threshold which compromised identification ability due to higher potential error: 16.09% cumulative rate for the 4% threshold in contrast to 10.57% for the 2.13% threshold (error rates for full data set, not just the 33 query species). Therefore, the aim here is to illustrate that there is no "one-size-fits-all" threshold for all taxa, and optimizing thresholds based on the library is a straightforward way to dramatically increase the ability of DNA barcodes to identify specimens. This was reiterated by Ratnasingham and Hebert (2013), who stated: "performance, as measured by the number of correctly recognized species, dropped steeply when the threshold deviated on either side of optimality."

Pros and Cons of Multilocus Methods

Regardless, an optimized threshold while superior to an arbitrary one such as 4%, remains nevertheless a heuristic approach that averages—albeit a more accurate average—over all species in that data set, and will depend upon substitution rate variation among lineages, the depth and breadth of taxon sampling, and the maturity of the taxonomy upon which the thresholds are derived. Where there are many species with multiple divergent lineages within them—due to either incomplete taxonomy, cryptic speciation, or geographically isolated populations—intraspecific divergences may be overestimated by an approach to threshold optimization that relies on taxonomic names. In this regard, we see the potential that MSC methods can offer if applied *a priori* to the data set to derive posterior probabilities of speciation events within a putatively single species (i.e., turning "unknown knowns" into "known knowns"). The advantage of an MSC approach is that it helps provide a stronger biological basis for recognizing this branch length variation as a potential speciation event rather than an artefact due to geographic sampling bias (Bergsten et al. 2012). However, it is questionable whether such statistics would be reliable where they would be most useful—that is, for singletons such as *S. australis* KM673—due to the sampling and parameter estimation problems associated

with taxon rarity in species delimitation methods (Lim et al. 2012).

An MSC approach does have the potential to improve on current DNA barcoding practices in the case of mixed clusters of haplotypes due to nonmonophyly. However, before investing resources, it is worth estimating how prevalent these processes/patterns really are. A recent study (Ross 2014) reassessed the findings of an influential paper (Funk and Omland 2003) stating that species-level mitochondrial paraphyly occurs in roughly 23% of animal species; he reported lower, but similar levels (19%) based on a larger data set. As to the causes of this nonmonophyly, a review of bird data (McKay and Zink 2010) reported 14.3% of bird species as nonmonophyletic, with 6% nonmonophyletic due to either introgression or incomplete lineage sorting, and 8% due to uncertain taxonomy. Misidentified specimens have also been cited as a potential problem particularly in DNA barcode data, and especially for taxa where multiple workers are submitting data for the same taxa independently (Collins and Cruickshank 2013).

So, in which of these situations would the MSC be superior? As stated previously, the MSC can be applied to robustly delimit species boundaries, but in the absence of integrative character data it is unlikely to be successful in completely resolving cases of uncertain taxonomy, due to the unavailability of node-based names derived from these kinds of analyses (Bauer et al. 2010). Likewise, misidentified voucher specimens are in most cases just as likely to be detected with single-locus markers as with multiple (Becker et al. 2011; Ko et al. 2013). On the other hand, in the cases of introgression and incomplete lineage sorting, an MSC approach would in theory be superior to single-locus barcoding. However, including a single nuclear gene as carried out by Dowton et al., is unlikely to provide much additional information as the power of nuclear loci generally lies in their multiplicity (Edwards and Bensch 2009). This is confirmed by our analysis of Dowton et al.'s CAD data (Supplementary Material available on Dryad; http://dx.doi.org/10.5061/dryad.h3g63); only 67% of the total species were monophyletic, and the BCM identification success of their 33 test species was just 51% for this gene.

Practical Issues

In a practical sense, MSC methods are incredibly computationally intensive with genomic-scale data. Even with the minimum of two genes and a very modest subset of just 136 individuals, Dowton et al. reported that the analysis remained "prohibitively slow", and they were unable to analyze all their individuals as queries, only testing a random single individual. Due to the lengthy exploratory analyses and inevitable model and prior selection issues when running highly parameterized species-tree analyses with hundreds of loci, we are highly doubtful that this vision of DNA barcoding can be used for the types of routine identification where speed is important (Armstrong and Ball 2005). Of course, we recognize that the field should not disregard philosophical and theoretic advances due to purely practical issues such as computational burden, but it seems that despite the increased availability of cloud computing facilities, there appears to be an even greater increase in the volumes of data now generated by genomic sequencing. Clearly, until considerably faster species delimitation methods can be developed, a pragmatic attitude cannot be entirely avoided. Therefore, to benchmark the efficiency and accuracy of species delimitation methodologies, it should now be a priority to highlight exemplar data sets—empirical and/or simulated—for which MSC methods clearly outperform simpler mtDNA analyses. Although this was one of the intentions of Dowton et al., our results show that the data set they used does not represent such a demonstration.

Improvements on Current Methods

Despite some clear theoretical advantages of MSC techniques in certain situations, the question remains as to whether the problems with mtDNA data have been quantified sufficiently to demand a new paradigm of specimen identification, and whether current methods/data could be improved or adapted without recourse to "overkill." As an example, there is still a widespread assumption that reciprocal monophyly is required for DNA barcode identifications (Goldstein and DeSalle 2011), but this is not the case (Meier 2008), and correct identifications are possible as long as haplotypes are not shared (Meier et al. 2006). There are also several identification methods now available for single loci that are substantially more sophisticated than sequence matching or tree building, and these may also offer improvements: Nielsen and Matz (2006) and Abdo and Golding (2007) developed coalescent theory to obtain statistical evidence for species membership; Zhang et al. (2008) and Zhang et al. (2012a) used machine learning and neural network methods; Zhang et al. (2012b) used fuzzy-set theory; and Weitschek et al. (2013) employed a character-based logic method. All of these techniques make different assumptions, resulting in specific advantages and disadvantages, but they have unfortunately been largely ignored in favor of the simpler tree-building methods. However, rather than concentrating on analytical methods, perhaps it is better to recognize that comprehensive sampling and complete reference libraries (Boykin et al. 2012; Virgilio et al. 2012), bring arguably the single biggest improvement to DNA barcode identification success (Ekrem et al. 2007). It is also important not to forget that the voucher specimens can also be examined for morphological characters in cases of ambiguity (although this may not be appropriate for incomplete or juvenile specimens).

Conclusions and Future Directions

Ultimately, researchers will develop the best methods for their questions and study system, so if MSC methods

are required for some difficult taxa then they should be used (Dupuis et al. 2012). For forensic and legal applications in particular, posterior probabilities from multilocus data offer a more robust and certainly more attractive approach. However, it is important to note that generating extra data and performing additional analyses on a case-by-case basis is not DNA barcoding. The power of DNA barcoding relies in its standardization and the subsequent scalability associated with that. To achieve standardization for genomic data, there are considerable costs associated with re-sequencing the millions of specimens already barcoded, as well as in curating this data. Taylor and Harris (2012) also suggested that post-Sanger genomic-sequencing technologies have rendered DNA barcoding obsolete. Parallel sequencing has certainly increased the cost efficiency of generating nucleotide data—and it can readily be used to generate barcode libraries (Shokralla et al. 2014)—but in our opinion the more significant costs of DNA barcoding come with collection, pre-laboratory processing, identification, databasing, and curation of the vouchers in a collection (Gregory 2005; Borisenko et al. 2009; Puillandre et al. 2012a). Perhaps a more tractable option to better capitalize on new sequencing technologies is through the generation of whole mitochondrial genomes (see Gillett et al. 2014). The additional gene sampling may reduce stochastic error and improve resolution of some closely related taxa, but remain computationally efficient due to mtDNA being effectively a single locus. Furthermore, these data would remain compatible with the existing DNA barcode system, and can also be readily reused for phylogenetic purposes (Gillett et al. 2014).

In conclusion, our opinion is that DNA barcoding need not assume "gene trees and species trees are synonymous," but rather that DNA barcodes are an effective proxy for species that can achieve high rates of identification success when appropriate methods are used to analyze them. So, although it is important to know where the limitations of the method lies, it is also important to realize its strengths. There should be a good justification for abandoning a proven technique, and the case study presented by Dowton et al. does not sufficiently demonstrate the failure of standard DNA barcoding due to the choice of a poorly performing identification technique. We reiterate Karl et al. (2012) in questioning the prevailing orthodoxy, namely that "more data are always better," and "one needs to do a Bayesian analysis". Therefore, we agree that researchers should be looking into "smarter" methods for taxon identification, but only in cases where the "dumb" methods have first been comprehensively shown not to work.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.h3g63

Funding

This work was supported by a Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Ciência sem Fronteiras fellowship (400813/2012-2) for R.A.C.

References

Abdo Z., Golding G.B. 2007. A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. Syst. Biol. 56:44–56.

Armstrong K. F., Ball S.L. 2005. DNA barcodes for biosecurity: invasive species identification. Philos. T. Roy. Soc. B. 360:1813–1823.

Bauer A.M., Parham J.F., Brown R.M., Stuart B.L., Grismer L., Theodore J., Böhme W., Savage J.M., Carranza S., Grismer J.L., Wagner P., Schmitz A., Ananjeva N.B., Inger R.F. 2010. Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. P. Roy. Soc. B. 278(1705):490–492.

Becker S., Hanner R., Steinke D. 2011. Five years of FISH-BOL: brief status report. Mitochondrial DNA 22 Suppl 1:3–9.

Bergsten J., Bilton D.T., Fujisawa T., Elliott M., Monaghan M.T., Balke M., Hendrich L., Geijer J., Herrmann J., Foster G.N., Ribera I., Nilsson A.N., Barraclough T.G., Vogler A.P. 2012. The effect of geographical scale of sampling on DNA barcoding. Syst. Biol. 61:851–869.

Borisenko A.V., Sones J.E., Hebert P.D.N. 2009. The front-end logistics of DNA barcoding: challenges and prospects. Mol. Ecol. Resour. 9:27–34.

Boykin L.M., Armstrong K., Kubatko L., De Barro P. 2012. DNA barcoding invasive insects: database roadblocks. Invertebr. Syst. 26:506–514.

Brown S.D.J., Collins R.A., Boyer S., Lefort M.-C., Malumbres-Olarte J., Vink C.J., Cruickshank R.H. 2012. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol. Ecol. Resour. 12:562–565.

Collins R.A., Cruickshank R.H. 2013. The seven deadly sins of DNA barcoding. Mol. Ecol. Resour. 13:969–975.

Dowton M., Meiklejohn K., Cameron S.L., Wallman J. 2014. A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. Syst. Biol. 63:639–644.

Dupuis J.R., Roe A.D., Sperling F.A.H. 2012. Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. Mol. Ecol. 21:4422–4436.

Edwards S., Bensch S. 2009. Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. Mol. Ecol. 18:2930–2933.

Ekrem T., Willassen E., Stur E. 2007. A comprehensive DNA sequence library is essential for identification with DNA barcodes. Mol. Phylogenet. Evol. 43:530–542.

Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.

Gillett C.P.D.T., Crampton-Platt A., Timmermans M.J.T.N., Jordal B., Emerson B.C., Vogler A.P. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). Mol. Biol. Evol. 31:2223–2237.

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

Gregory T.R. 2005. DNA barcoding does not compete with taxonomy. Nature 434:1067.

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S., Francis C.M. 2004. Identification of Birds through DNA Barcodes. PLoS Biology 2:e312.

Hickerson M.J., Meyer C.P., Moritz C. 2006. DNA barcoding will often fail to discover new animal species over broad parameter space. Syst. Biol. 55:729–739.

Karl S.A., Toonen R.J., Grant W.S., Bowen B.W. 2012. Common misconceptions in molecular ecology: echoes of the modern synthesis. Mol. Ecol. 21:4171–4189.

Ko H.-L., Wang Y.-T., Chiu T.-S., Lee M.-A., Leu M.-Y., Chang K.-Z., Chen W.-Y., Shao K.-T. 2013. Evaluating the accuracy of morphological identification of larval fishes by applying DNA barcoding. PLoS ONE 8:e53451.

Lim G.S., Balke M., Meier R. 2012. Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Syst. Biol. 61:165–169.

McKay B.D., Zink R.M. 2010. The causes of mitochondrial DNA gene tree paraphyly in birds. Mol. Phylogenet. Evol. 54:647–650.

Meier, R. 2008. DNA Sequences in Taxonomy: Opportunities and Challenges. In Wheeler Q.D., editor, The New Taxonomy chap. 7. New York: CRC Press. pp. 95–127.

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

Meiklejohn K.A., Wallman J.F., Cameron S.L., Dowton M. 2012. Comprehensive evaluation of DNA barcoding for the molecular species identification of forensically important Australian Sarcophagidae (Diptera). Invertebr. Syst. 26:515–525.

Meiklejohn K.A., Wallman J.F., Dowton M. 2011. DNA-based identification of forensically important Australian Sarcophagidae (Diptera). Int. J. Legal Med. 125:27–32.

Meiklejohn K.A., Wallman J.F., Pape T., Cameron S.L., Dowton M. 2013. Utility of COI, CAD and morphological data for resolving relationships within the genus Sarcophaga (sensu lato) (Diptera: Sarcophagidae): a preliminary study. Mol. Phylogenet. Evol. 69:133–141.

Meyer C.P., Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biology 3:e422.

Nielsen R., Matz M. 2006. Statistical approaches for DNA barcoding. Syst. Biol. 55:162–169.

Puillandre N., Bouchet P., Boisselier-Dubayle M.-C., Brisset J., Buge B., Castelin M., Chagnoux S., Christophe T., Corbari L., Lambourdière J., Lozouet P., Marani G., Rivasseau A., Silva N., Terryn Y., Tillier S., Utge J., Samadi S. 2012a. New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Mol. Ecol. Resour. 12:396–402.

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

Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS ONE 8:e66213.

Ross H.A. 2014. The incidence of species-level paraphyly in animals: a re-assessment. Mol. Phylogenet. Evol. 76:10–17.

Shokralla S., Gibson J.F., Nikbakht H., Janzen D.H., Hallwachs W., Hajibabaei M. 2014. Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Mol. Ecol. Resour. 14(5): 892–901.

Sonet G., Jordaens K., Nagy Z.T., Breman F.C., De Meyer M., Backeljau T., Virgilio M. 2013. Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification. ZooKeys 336:329–335.

Taylor H.R., Harris W.E. 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Mol. Ecol. Resour. 12:377–388.

Virgilio M., Jordaens K., Breman F.C., Backeljau T., De Meyer M. 2012. Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7:e31581.

Weitschek E., Van Velzen R., Felici G., Bertolazzi P. 2013. BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. Mol. Ecol. Resour. 13:1043–1046.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. P. Nat. Acad. Sci. USA 107: 9264–9269.

Zhang A.-B., Feng J., Ward R.D., Wan P., Gao Q., Wu J., Zhao W.-Z. 2012a. A new method for species identification via protein-coding and non-coding DNA barcodes by combining machine learning with bioinformatic methods. PLoS ONE 7:e30986.

Zhang A.-B., Muster C., Liang H.-B., Zhu C.-D., Crozier R., Wan P., Feng J., Ward R.D. 2012b. A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. Mol. Ecol. 21:1848–1863.

Zhang A.-B., Sikes D.S., Muster C., Li S.Q. 2008. Inferring species membership using DNA sequences with back-propagation neural networks. Syst. Biol. 57:202–215.