OPINION

The seven deadly sins of DNA barcoding

R. A. COLLINS* and R. H. CRUICKSHANK†

*Bio-Protection Research Centre, Lincoln University, PO Box 84, Lincoln, 7647, Canterbury, New Zealand, †Department of Ecology, Faculty of Agriculture and Life Sciences, Lincoln University, Lincoln, 7647, Canterbury, New Zealand

Abstract

Despite the broad benefits that DNA barcoding can bring to a diverse range of biological disciplines, a number of shortcomings still exist in terms of the experimental design of studies incorporating this approach. One underlying reason for this lies in the confusion that often exists between species discovery and specimen identification, and this is reflected in the way that hypotheses are generated and tested. Although these aims can be associated, they are quite distinct and require different methodological approaches, but their conflation has led to the frequently inappropriate use of commonly used analytical methods such as neighbour-joining trees, bootstrap resampling and fixed distance thresholds. Furthermore, the misidentification of voucher specimens can also have serious implications for end users of reference libraries such as the Barcode of Life Data Systems, and in this regard we advocate increased diligence in the a priori identification of specimens to be used for this purpose. This commentary provides an assessment of seven deficiencies that we identify as common in the DNA barcoding literature, and outline some potential improvements for its adaptation and adoption towards more reliable and accurate outcomes.

Keywords: bootstrap, DNA barcoding, hypothesis testing, neighbour joining, reference library, species identification

Received 5 June 2012; revision received 27 August 2012; accepted 9 November 2012

Introduction

DNA barcoding is now a well established method, and has been shown to bring substantial benefits to applications such as food product regulation (Becker et al. 2011), conservation (Francis et al. 2010), and investigating species interactions (Valentini et al. 2009). Nevertheless, we feel that a number of misconceptions have pervaded the DNA barcoding literature, and these are manifested in shortcomings of experimental design and analytical procedure. We have loosely identified seven such problems, and here we highlight each of them and suggest how improvements could be made. Most of these concerns have been raised previously in the literature, but as DNA barcoding approaches its tenth year, a timely synthesis is perhaps required, especially when DNA barcoding may in future provide admissible evidence in wildlife crime cases (Alacs et al. 2010; Linacre & Tobe 2011).

Our main concern is over the goal of DNA barcoding (Moritz & Cicero 2004; De Salle 2006; Rubinoff et al. 2006; Goldstein & De Salle 2011; Taylor & Harris 2012). We acknowledge that DNA barcoding (sensu Hebert et al. 2003) can comprise two distinct aims: specimen identification and species discovery (Schindel & Miller 2005). Specimen identification involves assigning taxonomic names to unknown specimens using a DNA reference library of morphologically preidentified vouchers. Here, DNA barcoding is particularly useful for applications such as biosecurity (Collins et al. 2012a) or marketplace substitution (Lowenstein et al. 2010). Species discovery with DNA barcodes, on the other hand, is best thought of as a quick and dirty 'molecular parataxonomy' process, analogous to physically sorting specimens into morphospecies (Brower 2006). Due to the relative speed at which COI sequence data can be generated and analysed, DNA barcoding therefore represents a powerful triage tool for biodiversity assessment, quickly sorting collections into species-like units (Schindel & Miller 2005). We use the term 'species-like units' to acknowledge that although short-length single-locus markers such as COI are often effective proxies for species, they are frequently not representative of full phylogenetic history (i.e. the species tree), due to their idiosyncratic behaviour (Brower et al. 1996; Rubinoff et al. 2006; Dasmahapatra et al. 2010;Dupuis et al. 2012; Fujita et al. 2012). Of course, subsequent to an initial DNA barcode triage, COI data can then be incorporated into more sophisticated species delimitation systems using multiple loci (Dupuis et al. 2012; Fujita...
et al. 2012), and ultimately into an integrative (Will et al. 2005; Padial et al. 2010; Goldstein & De Salle 2011) or DNA-based taxonomy framework (Tautz et al. 2003; Vogler & Monaghan 2007). We specifically make the point of distinguishing relatively crude single-locus methods such as DNA barcoding as ‘species discovery’, and multiplex/integrative methods as ‘species delimitation’ (Sites & Marshall 2004).

We feel that both of the outlined aims of DNA barcoding (specimen identification and species discovery) are uncontroversial, provided that they are clearly defined. However, several authors have raised repeated concerns regarding the blurring of these boundaries (De Salle et al. 2005; De Salle 2006; Vogler & Monaghan 2007; Meier 2008; Goldstein & De Salle 2011), and it seems impossible to separate these objectives in many examples from the DNA barcoding literature. This provides the basis for many of the criticisms herein.

**Sin number one: failure to test clear hypotheses**

Perhaps one of the gravest sins in many barcoding studies is the lack of clearly stated, objective hypotheses. A ‘typical’ barcoding study—‘DNA barcoding the [insert taxon] of [insert geographic region]’—aims to: (i) assemble a DNA reference library from specimens identified to species using morphological characters; (ii) test how effective this DNA library is for future identification purposes and then (iii) explore previously unrecognized diversity apparent in the DNA barcodes. However, it is in regard to these three steps that we feel there is often confusion in how hypotheses are generated and tested. Too frequently, objectives (ii) and (iii) are conflated, and methodological approaches do not appear to reflect these different goals (Meier 2008; Goldstein & De Salle 2011). Analytical techniques presented in many studies do not explicitly set out to test identification success (objective ii) by simulating a quantified identification scenario. Rather, they tend to employ the same method (usually a neighbour-joining tree) to test both objectives (ii) and (iii), and usually present a descriptive summary rather than a hypothesis-based test of the data. If the data collected are intended to be used as an identification tool, then they should be tested as such. Conversely, if a study aims to test the suitability of DNA barcoding as a biodiversity assessment tool (species discovery), then hypotheses of species richness should be estimated independently of the taxonomic names, and then compared a posteriori.

In light of these distinct goals, we recommend defining each objective more clearly in the methods section of the work, and explicitly separating the experimental procedures used to achieve each aim.

**Sin number two: inadequate a priori identification of specimens**

A serious limitation to the utility of DNA barcoding as a practical resource for regulation and molecular diagnostics is human error and uncertainty in creating and curating reference libraries. Becker et al. (2011) identify this as the primary source of error in FISH-BOL barcode data. Conflicting identifications can be made when multiple labs are working on the same taxa, and in the process of their morphological identifications they ascribe different taxonomic names to the same species.

As an example of the potential severity of the problem, we investigated each BIN cluster (Barcode Index Number) in the ‘RCYY’ project of ornamental cyprinid fishes (Collins et al. 2012a) on BOLD, the Barcode of Life Data System (Ratnasingham & Hebert 2007, URL: http://v3.boldsystems.org/). A total of 54 BINs in the project contained data from external projects (as of February 13, 2012). Of the 54, 19 BINs (35%) contained more than one species name, meaning that BOLD was unable at that time to offer an unambiguous species-level identification for these taxa. Even more worryingly, the same analysis was repeated some months later (June 26, 2012); the number of BINs found to contain data from external projects increased to 70, and the proportion of these BINs comprising taxonomic conflict (more than one species name) increased to 53% (=37).

A crucial aspect of DNA barcoding is the maintenance of records, supporting information and voucher specimens; this is what sets BOLD apart from GenBank (Ratnasingham & Hebert 2007). However, there are currently few safeguards against misidentifying a specimen, and once a name has been added to a database, it may be difficult for a third party to convince data managers that it should be changed. A new feature of BOLD v3.0 is a wiki-like framework for community-based annotation of barcode data (http://v3.boldsystems.org), but preemp- tive solutions are perhaps preferable. To this effect, a system of identification confidence has been proposed, which rates identifications according to the degree of expertise used and effort made (Steinke & Hanner 2011). This should encourage increasing diligence over how identifications are generated and justified.

The importance of accurate identification is obvious (Bortolus 2008), and providing a bibliography of reference material and morphological characters used for identification should be mandatory for publication (Vink et al. 2012). These additional metadata may be extremely valuable for correcting mistakes without recourse to the effort of loaning and reexamining voucher material. To quantify the extent to which identifications are unjusti- fied, all articles published in Molecular Ecology Resources
from 2010 to 2012 (up to August 16), with either ‘DNA barcoding’ or ‘DNA barcodes’ in the title were downloaded (total 61 articles). Of these, 46 presented empirical DNA barcode data, but of those only 16 (35%) cited literature to support their identifications. Only three presented morphological characters to directly support each taxon identification.

**Sin number three: the use of the term ‘species identification’**

The term ‘species identification’ is ubiquitous in the DNA barcoding literature, but in our view this terminology is misleading, and reflects a long-standing confusion between the two subdisciplines of DNA barcoding (specimen identification vs. species discovery; see above). We interpret ‘species identification’ here as shorthand for: identification of biological material—a specimen—to the level of species. However, we are concerned that it can be seen as species discovery or species delimitation (as used in Ferguson 2002). One way to minimize this confusion and to clarify the distinct role of each of the two separate objectives is to use the terms ‘specimen identification’ or ‘species discovery’ in place of ‘species identification’, as appropriate. This more objectively states what hypotheses are being tested, and better ensures that identification of individuals is not confused with discovery of groups. Both of these aims fall within the purview of DNA barcoding, but they should be clearly distinguished as they require different methodological and analytical approaches.

**Sin number four: inappropriate use of neighbour-joining trees**

Almost every DNA barcoding study presents a neighbour-joining (NJ) tree as part of the standard analytical procedure (Casiraghi et al. 2010). NJ trees were initially presented in the DNA barcoding literature by Hebert et al. (2003), although the method had been used prior to that for similar purposes (e.g. Hsieh et al. 2001). NJ trees use a distance matrix of sequence similarity to produce a hierarchical clustering phenogram (Little & Stevenson 2007). They are fast and easy to compute for large data sets, and are always fully resolved (i.e. do not produce ambiguous polytomies). However, several authors have expressed concerns as to whether this kind of phylogenetic approach to DNA barcoding is appropriate (Will & Rubinoff 2004; Meier et al. 2008; Goldstein & De Salle 2011).

For specimen identification purposes, it has been well documented both empirically and theoretically that NJ trees perform poorly (Meier et al. 2006; Virgilio et al. 2010; Little 2011; Zhang et al. 2012). They can be misleading to interpret, especially in conjunction with an incompletely sampled reference library. Unless nested directly within a cluster, the tree alone presents no way to discern if an unknown belongs to the closest topological species or not; i.e. it is unable to provide a ‘no identification’ result when an exact match is not present (Will et al. 2005; Collins et al. 2012a). It is important to note at this point that problems with NJ trees are not resolved by using any other tree inference method such as maximum likelihood or parsimony. The problem is with relying on tree topologies and monophyly (in a topological rather than cladistic sense) as an identification criterion. In situations of incomplete lineage sorting and species-level paraphyly, tree-based identification methods will result in ambiguous or incorrect identifications (Lowenstein et al. 2010). Few species concepts require reciprocal monophyly (Meier 2008), and in any case, monophyly of mtDNA lineages can be an unrealistic scenario in many closely related groups (Funk & Omland 2003; Zhang et al. 2012).

Despite the popularity and intuitiveness of NJ trees, identification success generally improves when using more accurate tree-free techniques, which are usually based directly on the genetic distance matrix. The ‘best close match’ method (Meier et al. 2006) has been shown to be reliable, predictable, and computationally tractable (Virgilio et al. 2010; Collins et al. 2012a). Intra-/interspecific variation is discriminated by a user-set threshold (but see sin number six). The method is easy to interpret, being one of the most simple available, but is able to make identifications even in the presence of paraphyly. Alternatively, many other criteria are also available for measuring identification success (see Casiraghi et al. 2010), and comparisons of performance between some of these have already been made (Meier et al. 2006; Little & Stevenson 2007; Ross et al. 2008; Austerlitz et al. 2009; Virgilio et al. 2010; Zhang et al. 2012).

It is important to note, however, that a quantification of monophyly still remains a useful description of the data, and should still be used in conjunction with other methods. Problems mostly occur when NJ trees are the only analytical method presented, and identification success rates are not quantified (Little & Stevenson 2007). Likewise, for the purpose of graphically summarizing the data, NJ trees can be considered appropriate. It is also worth mentioning that the ubiquitous use of the Kimura two-parameter (K2P) model for constructing initial distance matrices is also questionable; uncorrected $p$ distances should be used instead (Collins et al. 2012b; Srivathsan & Meier 2012).

In some scenarios, character-based methods using diagnostic nucleotide combinations may be preferable (Davis & Nixon 1992; Brower 1999; De Salle 2007), and this is particularly the case for small groups of closely
related taxa where similarity methods perform poorly (Lowenstein et al. 2009). However, character-based approaches such as those implemented in the CAOS software (Sarkar et al. 2008), have yet to be fully characterized in terms of their sensitivity to taxon sampling and homoplasy, and are therefore at present perhaps limited to restricted applications (Kerr et al. 2009). The use of discrete characters could be seen in terms of ‘DNA barcoding 2.0’, potentially offering additional benefits after sampling is extended beyond simply collecting baseline data.

Ultimately, phenetic (similarity) methods using genetic distances may be regarded as something of a stop-gap solution. In the near future, the problem of accurately assigning identifications is likely to be addressed by either likelihood-based information-theoretic approaches, or machine learning and statistical tools such as supervised classification and pattern recognition (e.g. Zhang et al. 2008; Austerlitz et al. 2009). Bayesian coalescent methods also promise statistical advantages, but may be too computationally inefficient in their current incarnations (Zhang et al. 2012).

When evaluating DNA barcoding as a species discovery tool, a method is required that can approximate the number of species in a sample directly from the DNA sequence data, and independently from the preassigned taxonomic names (i.e. the data set used to subsequently measure congruence between the two approaches). NJ trees are a poor choice in this respect, as on their own, they are unable to make a quantitative assessment. Techniques such as the Automatic Barcode Gap Discovery method (Puillandre et al. 2012), the general mixed Yule coalescent model (Monaghan et al. 2009), or the BOLD v3.0 BIN system are able to use genetic information to generate an estimate of the number of species-like units, and should therefore be used in preference to NJ trees for this kind of study.

**Sin number five: inappropriate use of bootstrap resampling**

The use of bootstrap resampling in DNA barcoding studies typifies the confusion between species discovery and specimen identification. The bootstrap, along with reciprocal monophyly, is one method among many that can be used to test whether a species-like cluster is well supported, and can be used as part of a species discovery/delimitation process (i.e. a test of genetic distinctiveness for a potentially new taxon). Bootstrapping in this situation also helps address problems with NJ trees such as taxon-order bias and tied trees (Meier 2008; Lowenstein et al. 2009).

However, the use of bootstrapping for specimen identification is somewhat perplexing. The aim here is to maximize congruence with a priori defined species, *viz.* the taxonomic names from a morphological identification process. A species with low bootstrap support does not falsify a species hypothesis when this assessment was based on independent data (i.e. morphology from the original description). In many cases, recently diverged sister species on short branches will have low support and therefore fail to be identified, even if they are morphologically distinct and diagnosable by unique mutations (Lowenstein et al. 2009). Thus, using a bootstrap value as a cut-off for correct identification severely compromises the efficacy of a reference library (Collins et al. 2012a; Zhang et al. 2012), and exacerbates the previously outlined weaknesses of using tree-based methods in general. On top of this, bootstrap resampling does not make an assessment of the uncertainty in identification; an unknown can group with a reference specimen at 100% bootstrap support, and yet be an entirely different species. Perhaps an improved way to measure uncertainty in identification is to better understand interspecific threshold distances (see below), calculate group membership probabilities (Zhang et al. 2012) and make explicit ‘caveats in relation to the breadth [sic] of sampling’ (Moritz & Cicero 2004).

**Sin number six: inappropriate use of fixed distance thresholds**

The use of distance thresholds has been extensively debated (see references in Puillandre et al. 2012; Virgilio et al. 2012; Zhang et al. 2012), so our aim is not to cover old ground, but to reemphasize points already made. A threshold is essential when identifying specimens using genetic distance data; in the absence of complete sampling, distance thresholds aim to minimize misidentifications of unknowns that do not have conspecifics represented in the reference library (Virgilio et al. 2012). However, there is no a priori reason to assume that a universal threshold is applicable, as coalescent depths among species will vary considerably due to differences in population size, rate of mutation and time since speciation (Monaghan et al. 2009; Fujita et al. 2012).

A generic threshold such as 1% is perhaps not an unreasonable heuristic in some cases, but it can be considered arbitrary, and is likely to suffer from varying rates of false-positive and false-negative error, depending on the data. Rather than relying on prescribed cut-offs, optimized thresholds should be generated directly from the data (Meyer & Paulay 2005; Virgilio et al. 2012). Software and protocols are now available to calculate these optimized thresholds, and for species discovery, can even be generated in the absence of taxonomic names (Brown et al. 2012; Puillandre et al. 2012; Virgilio et al. 2012). However, a newly developed
fuzzy-set-theory technique (Zhang et al. 2012) perhaps obviates the requirement for distance thresholds, offering a more sophisticated group membership parameter that provides additional information lacking in threshold-based implementations.

**Sin number seven: incorrectly interpreting the barcoding gap**

The barcoding gap as proposed by Meyer & Paulay (2005) can represent two distinct scenarios: one for specimen identification, with an individual being closer to a member of its own species than a different species (i.e. a ‘local’ barcoding gap); and one for species discovery, a distance that equates to a threshold applicable to all species (i.e. a ‘global’ barcoding gap). Although this has been previously discussed (Wiemers & Fiedler 2007), the two scenarios are frequently confused, and this again demonstrates conflation of the two objectives of DNA barcoding.

Many DNA barcoding studies present histograms showing frequency distributions of both intra- and interspecific divergences for all pooled species analysed in a study. Overlap between the two distributions can be interpreted as a failure of DNA barcoding, but the only failure demonstrated in this case is that of defining a universal cut-off value. In this regard, and as stated previously, it is widely acknowledged that coalescent depths vary among species, and substantial overlap between intra- and interspecific distances may be the rule, rather than the exception (Wiemers & Fiedler 2007; Virgilio et al. 2010). Therefore, for specimen identification purposes this type of presentation is wholly uninformative, as intraspecific distances for one species can exceed interspecific distances for other species in the analysis, but without necessarily compromising identification success (the local gap).

A better display of distance data for specimen identification is a dotplot in which, for each individual in the data set, the distance to the furthest conspecific is plotted against the distance to the nearest nonconspecific, with a 1:1 slope representing the point at which the difference between the two is zero (i.e. no local barcoding gap). An example of this method is illustrated in Fig. 4 of Robinson et al. (2009). It is also important to note that these statistics should be generated using the smallest

<table>
<thead>
<tr>
<th>Problem</th>
<th>Consequence</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to test clear hypotheses</td>
<td>Choice of inappropriate or suboptimal analytical method due to confusion as to the objectives of the study</td>
<td>Explicitly state each hypothesis, and for each distinct aspect of the study present separate headings in methods and results sections</td>
</tr>
<tr>
<td>Inadequate a priori identification of specimens</td>
<td>Conflicting identifications made by different labs using the method of Virgilio et al. (2010), or GMYC (Monaghan et al. 2009) or BOLD’s BIN system (<a href="http://v3.boldsystems.org">http://v3.boldsystems.org</a>)</td>
<td>Present a bibliography of references, as well as the distinguishing morphological characters used in the identification process. Follow recommendations outlined by Steinke &amp; Hanner (2011)</td>
</tr>
<tr>
<td>The use of the term ‘species identification’</td>
<td>Confusion between identification of individuals, and delimitation/discovery of species</td>
<td>To clarify objectives, use the term ‘specimen identification’ or ‘species discovery’ where appropriate</td>
</tr>
<tr>
<td>Inappropriate use of neighbour-joining trees</td>
<td>(a) Relying on strict monophyly for identification can reduce the apparent effectiveness of DNA barcoding as an identification tool. This can be due to either mtDNA paraphyly or misidentification of specimens. (b) For biodiversity assessment and species discovery, NJ trees cannot estimate the number of species independently with respect to the taxonomic names</td>
<td>(a) Alternative criteria such as ‘best close match’ are readily available, and have higher rates of identification success. This method can be implemented using the free software packages TaxonDNA (Meier et al. 2006) or Spider (Brown et al. 2012). (b) Estimate species richness using ABGD (Puillandre et al. 2012), GMYC (Monaghan et al. 2009) or BOLD’s BIN system (<a href="http://v3.boldsystems.org">http://v3.boldsystems.org</a>)</td>
</tr>
<tr>
<td>Inappropriate use of bootstrap resampling</td>
<td>For specimen identification purposes, bootstrap resampling can further reduce the already low identification success rates associated with NJ trees</td>
<td>Only use bootstrapping where appropriate: e.g. as part of a species delimitation process on preestimated groups</td>
</tr>
<tr>
<td>Inappropriate use of fixed distance thresholds</td>
<td>For specimen identification purposes, a generic threshold which is set too low or high can reduce or bias identification error rates</td>
<td>Thresholds can now be optimized for specific data sets using the method of Virgilio et al. (2012), or with software such as ABGD (Puillandre et al. 2012) and Spider (Brown et al. 2012)</td>
</tr>
<tr>
<td>Incorrectly interpreting the barcoding gap</td>
<td>Overlapping distributions of intra-/interspecific distances do not necessarily mean that barcodes perform poorly for identification</td>
<td>For specimen identification studies, dotplots of intra-/interspecific distances are a better way to illustrate the barcoding gap (e.g. Robinson et al. 2009)</td>
</tr>
</tbody>
</table>
rather than the mean interspecific distance (Meier et al. 2008).

Summary
We hope that future barcoding studies will push forward improvements in data analysis, and make more use of alternative methods. One possible cause of the limited uptake of some new approaches may be the lack of a common platform for carrying out analyses (Sarkar & Trizna 2011). Exploring different analytical methods is important, but many software applications have quite different input and output formats, making comparison between them difficult. Fortunately, it is now increasingly possible to conduct a variety of analyses in a universal, open-source environment such as R (R Development Core Team 2011), which could potentially supersede the inflexible and piecemeal software applications that are currently available (Freckleton 2009). This is one solution that could ultimately encourage improved dissemination, sharing and benchmarking of new techniques among labs.

In conclusion, we feel that more care should be taken in setting clear hypotheses for DNA barcoding studies, and choosing appropriate methods for testing each distinct hypothesis. Given the opportunities DNA barcoding can offer to both biodiversity research and regulatory science, it would be unfortunate to compromise its potential due to misconceptions or inadequacies in study design. Table 1 provides an outline of possible solutions to some of the problems outlined herein.

Acknowledgements
This work was supported by funding from the Ministry for Primary Industries Biosecurity New Zealand and the Bio-Protection Research Centre for RAC. The authors thank Karen Armstrong, Rudolf Meier, Laura Boykin and two reviewers for suggesting improvements to earlier versions of the manuscript.

References