Molecular Ecology Notes (2006) **6**, 959–964 doi: 10.1111/j.1471-8286.2006.01470.x

BARCODING

A minimalist barcode can identify a specimen whose DNA is degraded

MEHRDAD HAJIBABAEI,* M. ALEX SMITH,* DANIEL H. JANZEN,† JOSEPHINE J. RODRIGUEZ,‡ JAMES B. WHITFIELD‡ and PAUL D. N. HEBERT*

**Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada, N1G 2W1, †Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA, ‡Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

Abstract

A DNA barcode based on 650 bp of mitochondrial gene cytochrome *c* oxidase I is proving to be highly functional in species identification for various animal groups. However, DNA degradation complicates the recovery of a full-length barcode from many museum specimens. Here we explore the use of shorter barcode sequences for identification of such specimens. We recovered short sequences — i.e. ~100 bp — with a single PCR pass from more than 90% of the specimens in assemblages of moth and wasp museum specimens from which full barcode recovery was only 50%, and the latter were usually less than 8 years old. Short barcodes were effective in identifying specimens, confirming their utility in circumstances where full barcodes are too expensive to obtain and the identification comparisons are within a confined taxonomic group.

Keywords: COI, DNA barcoding, DNA degradation, fish, Lepidoptera, museum specimens, parasitic wasps, taxonomy

Received 3 March 2006; revision received 21 April 2006; accepted 24 May 2006

Introduction

DNA barcoding employs standardized genomic fragments to facilitate species identification and discovery (Hebert *et al*. 2003; Kress *et al*. 2005; Savolainen *et al*. 2005). Studies on various groups of animals have shown that a 650-bp fragment of the mitochondrial gene, cytochrome *c* oxidase I (COI, *cox1*) is generally effective as a barcode sequence, delivering more than 95% species-level resolution (Hebert *et al*. 2003, 2004a, 2004b; Barrett & Hebert 2005; Meyer & Paulay 2005; Hajibabaei *et al*. 2006; Smith *et al*. 2006). Based on these results, studies are already underway to build DNA barcode libraries for all birds and fishes (Marshall 2005) and substantial arrays of Lepidoptera (Hajibabaei *et al*. 2006).

It is generally difficult to quickly and cheaply recover barcode sequences from museum specimens that are more than a decade old, since their DNA is degraded (Whitfield 1999; Hajibabaei *et al*. 2005). As a result, major barcode library construction currently focuses on the analysis of recently collected specimens or on samples that have been protected from degradation by freezing, ethanol, or other DNA-friendly preservation methods. However, it will ultimately be necessary for sequences from fresh specimens to be compared with sequences from millions of older museum specimens. For example, such comparison is critical when barcoding reveals several cryptic species within what had been viewed as one species, and it is not morphologically evident which of them matches the holotype (Hebert *et al*. 2004a; Janzen *et al*. 2005). Equally, the ultimate validation for a modern barcode record should involve its comparison with the barcode record from the holotype specimen for that species. Aside from the need for such comparisons, it is evident that the barcoding of old museum specimens will provide a cost-effective way of building barcode libraries with broad geographical coverage of individual taxa.

While the recovery of full barcode sequences from aged specimens currently requires costly and time-consuming forensic/ancient DNA protocols, short sequences can regularly be obtained from century-old museum specimens

Correspondence: Mehrdad Hajibabaei, Fax: (519) 767-1656; E-mail: mhajibab@uoguelph.ca

Table 1 A comparison of full-length barcodes and mini-barcodes in two exemplar data sets

DNA barcode	Length (bp)	Variability (in %)†	Parsimony (in %)‡	% intraspecific (SE)§	% intrageneric (SE)¶
Fishes of Australia (697 individuals, 204 species, 112 genera)					
Full barcode	654	52.0	49.1	0.5 (0.1)	6.0 (0.6)
Mini-barcode-218-1	218	52.3	48.6	0.4 (0.1)*	4.8 (0.5)*
Mini-barcode-218-2	218	47.7	47.7	0.4 (0.1)	6.7 (0.6)*
Mini-barcode-218-3	218	56.0	51.0	0.6 (0.1)	6.6 (0.7)*
Mini-barcode-109-1	109	58.7	53.2	0.4 (0.1)	4.8 (0.6)*
Mini-barcode-109-2	109	45.9	44.0	0.3 (0.1)*	4.8 (0.5)*
Mini-barcode-109-3	109	47.7	47.7	0.4 (0.1)	6.5 (0.6)
Mini-barcode-109-4	109	47.7	47.7	0.4 (0.1)	7.0 (0.7)*
Mini-barcode-109-5	109	57.8	57.8	0.6 (0.1)	6.8 (0.7)
Mini-barcode-109-6	109	54.1	44.0	0.6 (0.1)	6.4 (0.6)
Lepidoptera of ACG (522 individuals, 61 species, 4 genera)					
Full barcode	654	43.4	39.8	0.2 (0.0)	7.2 (0.4)
Mini-barcode-218-1	218	40.8	34.9	0.1 (0.0)	5.8 (0.6)
Mini-barcode-218-2	218	41.3	39.9	0.2 (0.1)	7.7 (0.5)
Mini-barcode-218-3	218	48.2	44.5	0.3 (0.1)	8.2 (0.6)
Mini-barcode-109-1	109	48.6	36.7	0.1 (0.0)	5.4 (0.7)
Mini-barcode-109-2	109	33.0	33.0	0.1 (0.0)	6.2 (0.6)
Mini-barcode-109-3	109	39.5	36.7	0.2 (0.1)	7.4 (0.4)
Mini-barcode-109-4	109	43.1	43.1	0.2 (0.1)	8.1 (0.7)
Mini-barcode-109-5	109	52.3	48.6	0.2 (0.1)	7.6 (0.5)
Mini-barcode-109-6	109	44.0	40.4	0.4 (0.1)	8.9 (0.8)

Mini-barcodes are numbered according to their position relative to the full-length (5′–3′) barcode region. For example, mini-barcode-218-1 indicates the first 218 nt of the 5′–3′ full-length barcode, mini-barcode 218-2 indicates nt 219–436, etc.

†Per cent of sites that varied in the sequences.

‡Per cent of parsimony informative sites in the alignment.

§Average pairwise intraspecific Kimura-2 parameter distances (Kimura 1980).

¶Average pairwise intrageneric Kimura-2 parameter distances. P values were calculated in a paired t-test with Bonferroni correction. The test was conducted to evaluate if the distance measures are different in full-length vs. mini-barcodes by calculating P for each mini-barcode in comparison to the full-length barcode.

*$P < 0.05$. Genetic distances were calculated using MEGA 3.1 software (Kumar *et al*. 2004).

ACG = Area de Conservación Guanacaste.

using routine protocols (i.e. Goldstein & Desalle 2003). Here we examine, using both *in silico* and empirical tests, the accuracy of such short DNA fragments in species identification.

In silico analysis

We first tested the *in silico* performance of 'mini-barcodes' (~200 bp and ~100 bp) in species-level identifications of two barcode data sets. The first data set consisted of Australian fishes (Ward *et al*. 2005) including 204 species represented by an average of 3.4 individuals and 19% of the species were represented by a single individual. The second data set consisted of four species-rich genera of Lepidoptera (Janzen *et al*. 2005; Hajibabaei *et al*. 2006) including 61 species represented by an average of 8.6 individuals and 13% of the species were represented by a single individual (see Supplementary material for GenBank accession numbers and molecular methodology). We divided the full-length barcode region into 3 or 6 equal sizes (218 bp or 109 bp each) from the 5′–3′ end (e.g. nt 1–218, 219–436, etc.). We then compared the percentage of variable and parsimony informative sites, and both intraspecific and intrageneric divergences, in mini- vs. full-length barcodes (Table 1). The mini-barcodes generally provided measures of sequence variability and divergence similar to those of full barcodes at both intraspecific and intrageneric levels. However, significant shifts in divergence values were noted in comparisons of some mini- vs. full-length barcodes in fishes (Table 1). The results suggest that the identification of fish or Lepidoptera would generally have been as accurate with mini-barcodes as with full-length barcodes. In addition, we carried out a neighbour-joining (NJ) analysis (Saitou & Nei 1987) with either 218–2 or 109–3 mini-barcode data sets (see Table 1) and counted the number of species with non-overlapping barcodes (i.e. barcodes that uniquely identify a species) as a measure of resolution. We compared these results

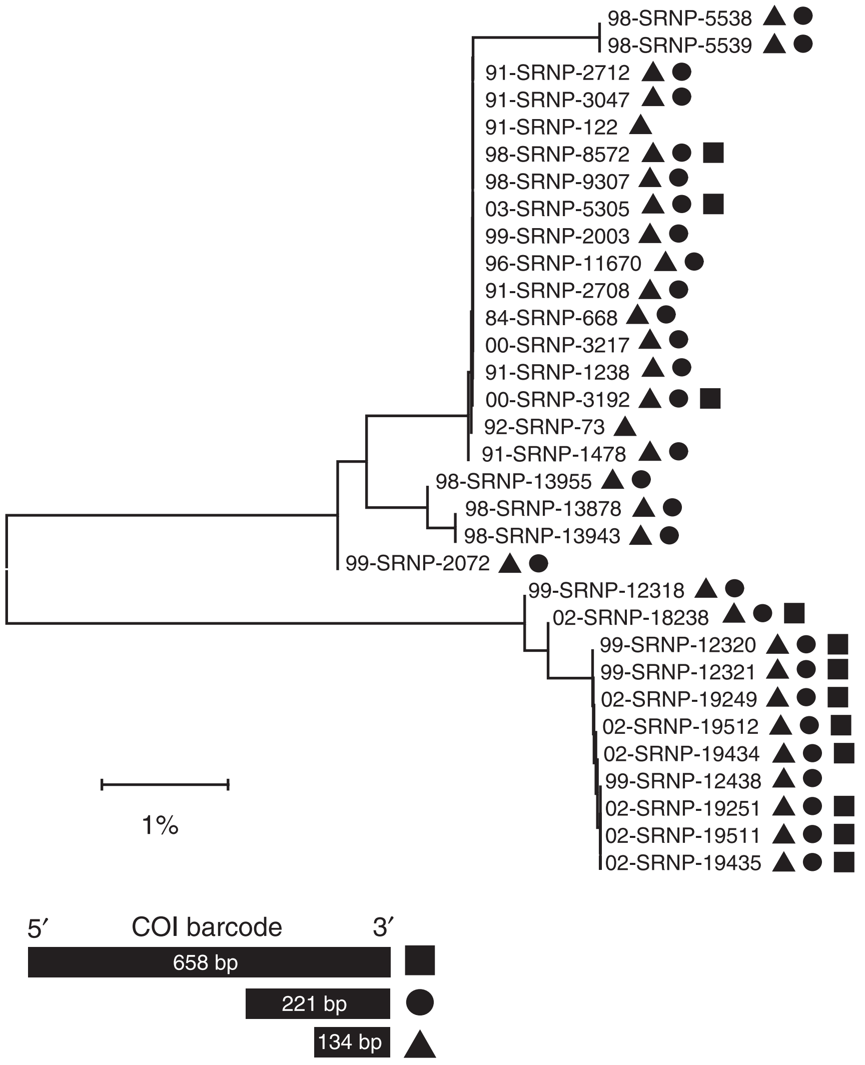

Fig. 1 The ability of mini-barcodes compared to full-length barcodes to discriminate among 1–2-decade-old specimens of two cryptic species within *Xylophanes libya* in neighbour-joining (NJ) analysis (Saitou *et al.* 1987). Tree is based on the 32 specimens (of a total of 33) that produced 134 bp mini barcodes — shown by triangular markers. Individuals with circular markers yielded 221 bp mini-barcodes and individuals with square shape markers yielded full-length barcodes (658 bp). The first two digits of each voucher code show the date of collection of the specimen; so 00 indicates year 2000. The size and position of each amplicon with reference to the full COI barcode region is indicated below the tree. The analysis was performed using MEGA 3.1 software (Kumar *et al.* 2004) with Kimura-2 parameter distances (Kimura 1980).

with the results obtained from full-length barcodes. For instance, 93% and 92% of the species were correctly identified with the 218–2 and 109–3 mini-barcodes, respectively, compared to 95% with the full-length barcode. A similar analysis with mini-barcodes with significant differences in their distance measure show a somewhat lower resolution in the NJ trees (results not shown).

Empirical tests

We generated mini-barcodes for two sets of museum specimens of varied age. We selected these two exemplar cases to address different sample preservations (either oven dried or ethanol preserved) and two different taxonomic scopes. In the first case, *Xylophanes libya*, a neotropical sphingid moth, has been found to consist of two species in northwestern Costa Rica by barcoding (Janzen *et al*. 2005; Hajibabaei *et al*. 2006; Fig. 1a) (these two cryptic species were subsequently found to be morphologically and microgeographically distinguishable as well). We used this available sequence information to design PCR primers to amplify 221 and 134 bp amplicons from the 3′ end of the full-length barcode region (see Supplementary material). We then attempted amplification of these mini-barcodes from 33 oven-dried specimens ranging in age from 2 to 21 years (average age of 7.5 years). Full-length barcodes were recovered from only 39% of the specimens, all younger than 8 years. However, there was 94% and 97% success in obtaining the 221 and 134 bp mini-barcodes, respectively. These mini-barcodes contained 16 and 10 diagnostic characters, respectively, as compared to 38 in full-length barcodes. An NJ analysis was carried out with full-length, 221 bp, and 134 bp data sets. The mini-barcodes produced species-level resolution congruent with full-length barcodes, and allowed all of

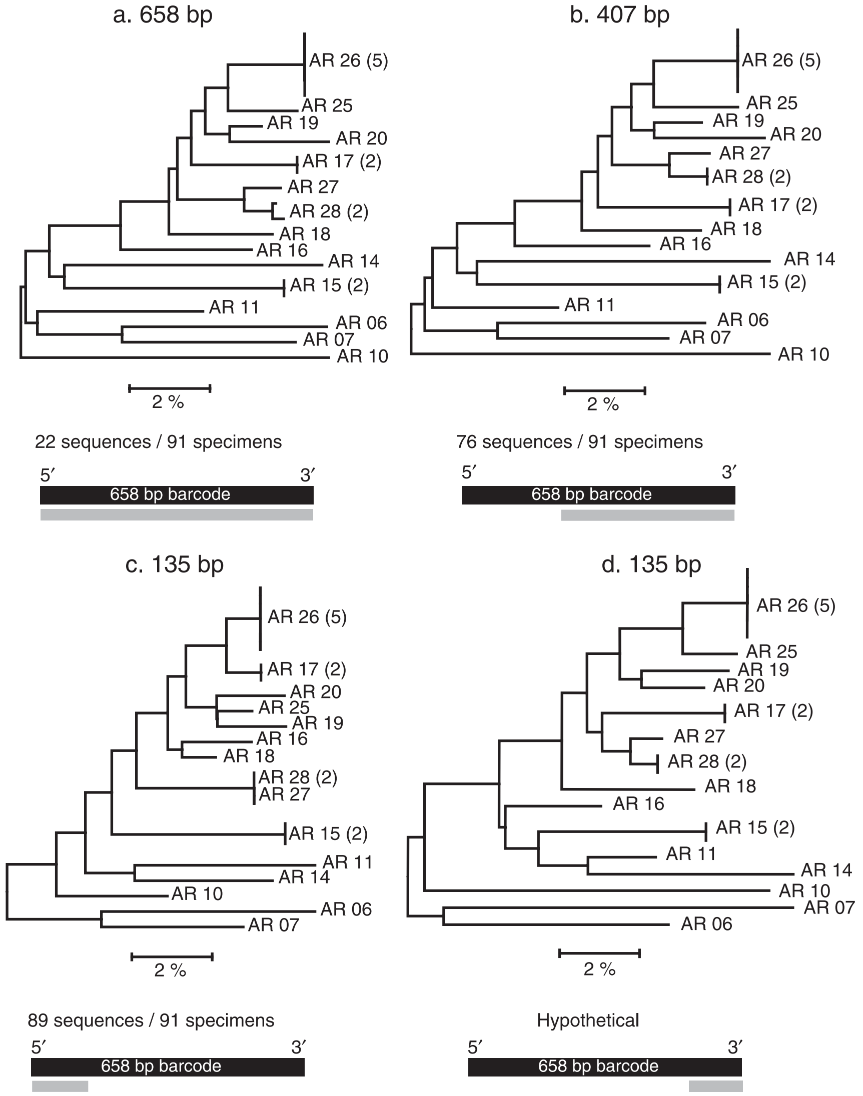

Fig. 2 Species-level resolution of COI barcodes of varying lengths for microgastrine braconids in NJ analysis. AR = *Apanteles* Rodriguezxx species. The specimens shown are restricted to those that produced a barcode for each amplicon length. The amplicon's approximate size and position with reference to the full COI barcode region is indicated in grey. Number of sequences recovered for each amplicon size is shown beneath each tree. The analysis was performed using MEGA 3.1 software (Kumar *et al*. 2004) with Kimura-2 parameter distances (Kimura 1980). Numbers in the parentheses indicate number of individuals analysed. Panel d represents a hypothetical 135 bp mini-barcode — for the same specimens shown in other panels — extracted from the 3′ region of the full-length barcode of panel a.

the older museum specimens to be assigned to one or the other species (Fig. 1).

Our second test involved reared microgastrine parasitic wasps (Hymenoptera, Braconidae, *Apanteles*) from Costa Rica. Full-length barcodes and two mini-barcodes (407 bp from the 3′ end, 135 bp from the 5′ end) were amplified from 91 ethanol-preserved specimens varying in age from 1 to 14 years (average age of 4.4 years), revealing the presence of 15 species (Fig. 2). Only 24% of these specimens yielded full-length barcodes (all from specimens less than 6 years old), but there was a 84% and 98% success rate in recovering 407 and 135 bp mini-barcodes, respectively. The NJ analysis indicated that the short sequences produced species-level resolution as effectively as that of full-length barcodes (Fig. 2) with one exception. Two species from different host caterpillars (*Apanteles* Rodriguez27 and *A.* Rodriguez28) showed divergences of 1.79% (11 diagnostic characters) and 1.93% (8 diagnostic characters) using full and 407 bp barcodes, respectively, but were not separable with the 135 bp barcode.

Discussion and conclusions

Our *in silico* analysis of barcode sequences for fishes and Lepidoptera revealed the potential ability of mini-barcodes to discriminate among species (Table 1). While mini-barcodes produced divergence values comparable to full-length barcodes in both data sets, they were some what less effective in discriminating among the species in large assemblages (e.g. 204 species of Australian fishes). However, most applications of mini-barcodes will not involve cases that seek to place a specimen among all known species, but rather within a small assemblage, as exemplified by our two empirical sets. In both cases, we obtained as many barcode sequences as possible from young specimens and subsequently gathered mini-barcodes from old specimens to link the old specimens to the younger ones. Mini-barcodes were effective in species identification in both cases. However, the choice of length and position of mini-barcodes is important in their ability to discriminate among species. In the two cryptic species of *Xylophanes libya*, mini-barcodes were positioned specifically to discriminate this species pair, and their relatively short lengths took into account the old age of the specimens. In addition, the fact that the forward primers in the two mini-barcodes were designed specifically for this species complex might have positively influenced the chance of amplifying potentially degraded DNA. By contrast, in the case of the microgastrine wasps, the two mini-barcodes were used as routine alternate amplicons to increase the chance of gaining barcode information from an old specimen. In this case, we found that a 135-bp mini-barcode situated at the 5′ end did not allow discrimination between two of the 15 species. Interestingly, a similarly sized fragment from the other end of the barcode region would have achieved identification (Fig. 2d).

Full-length barcode sequences can be easily and cheaply obtained from recently collected tissue or from those preserved for DNA extraction (Hajibabaei *et al*. 2005). The resultant records provide a 'gold standard' with high confidence for species discrimination within large species pools (Hebert *et al*. 2003, 2004b; Smith *et al*. 2005, 2006; Ward *et al*. 2005; Hajibabaei *et al*. 2006). Here we confirm that very short barcode sequences are also valuable for the identification of old specimens from selected narrow taxonomic arrays, such as comparing a newly collected specimen with an old holotype. Similarly, it may well be possible to employ mini-barcodes for the identification of formalin-fixed samples, which often contain highly fragmented DNA (Schander & Kenneth 2003). Mini-barcodes may also provide the option to employ alternative sequencing methods, such as pyrosequencing (Fakhrai-Rad *et al*. 2002), that yield only short sequences, but offer lower costs and higher throughput than standard approaches.

Acknowledgements

Funding for this study was provided by grants from the Natural Sciences and Engineering Research Council (Canada), the Canada Research Chairs program, the Gordon and Betty Moore Foundation, and Genome Canada through Ontario Genomics Institute to PDNH, and by grants from the US National Science Foundation (DEB 0072730 and 0515699), Guanacaste Dry Forest Conservation Fund, and Area de Conservación Guanacaste to DHJ. MAS was supported by a Fonds québécois de la recherche sur la nature et les technologies B3 post-doctoral fellowship. We thank Stephanie Kirk and Taika von Königslöw for assistance with molecular work, Angela Holliss for DNA sequencing, Tanya Dapkey for preparing specimens, ACG parataxonomists for rearing the caterpillars and wasps, and Carolyne Bardeleben for constructive review of the manuscript.

Supplementary material

The supplementary material is available from http://www.blackwellpublishing.com/products/journals/suppmat/MEN/MEN1470/MEN1470sm/htm

References

Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology*, **83**, 481–491.

Fakhrai-Rad H, Pourmand N, Ronaghi M (2002) Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. *Human Mutation*, **19**, 479–485.

Goldstein PZ, Desalle R (2003) Calibrating phylogenetic species formation in a threatened insect using DNA from historical specimens. *Molecular Ecology*, **12**, 1993–1998.

Hajibabaei M, deWaard JR, Ivanova NV *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1959–1967.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences, USA*, **103**, 968–971.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, 313–321.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences, USA*, **101**, 14812–14817.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b) Identification of birds through DNA barcodes. *Public Library of Science Biology*, **2**, 1657–1663.

Janzen DH, Hajibabaei M, Burns JM *et al.* (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1835–1845.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA*, **102**, 8369–8374.

Kumar S, Tamura K, Nei M (2004) MEGA 3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.

Marshall E (2005) Taxonomy. Will DNA bar codes breathe life into classification? *Science*, **307**, 1037.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *Public Library of Science Biology*, **3**, E422.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1805–1811.

Schander C, Kenneth HM (2003) DNA, PCR and formalinized animal tissue — a short review and protocols. *Organisms Diversity and Evolution*, **3**, 195–205.

Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1825–1834.

Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PDN (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proceedings of the National Academy of Sciences, USA*, **103**, 3657–3662.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1847–1857.

Whitfield JB (1999) Destructive sampling and information management in molecular systematic research: an entomological perspective. In: *Managing the Modern Herbarium: An Interdisciplinary Approach* (Ed. Byers S, Metsger D), p. 384. Society for Preservation of Natural History Collections and Royal Ontario Museum.