Systematics and Biodiversity **4** (2): 127–132
doi:10.1017/S147720000500191X Printed in the United Kingdom

Issued 7 June 2006

Andrew V.Z. Brower
Department of Zoology,
Oregon State University,
Corvallis, OR 97331, USA
Email: browera@science.oregonstate.edu

submitted October 2005
accepted December 2005

PERSPECTIVE

Problems with DNA barcodes for species delimitation: 'ten species' of *Astraptes fulgerator* reassessed (Lepidoptera: Hesperiidae)

Abstract Hebert and colleagues (2004) used a short region of the mitochondrial Cytochrome oxidase subunit I gene as a delimiter for ten putative species from among 466 individuals of the skipper butterfly currently known as *Astraptes fulgerator* from Guanacaste, Costa Rica. Their data are reanalysed to assess cluster stability and clade support using Neighbor-Joining bootstrap, population aggregation analysis and cladistic haplotype analysis. At least three, but not more than seven mtDNA clades that may correspond to cryptic species are supported by the evidence. Additional difficulties with Hebert *et al.*'s interpretation of the data are discussed.

Key words mtDNA, COI, molecular systematics, skipper butterfly

Introduction

For some years now, the limiting stage in study of DNA data has not been the generation of sequences themselves, which is now routinely performed at an industrial scale, or even outsourced to private companies, much like sending a roll of film to be developed. Instead, as is evidenced by the birth of bioinformatics as a discipline, it is the careful analysis and interpretation of sequences that is the most time-consuming and labour-intensive step in the process, and the step at which the greatest value is added, raw data being transformed into useful knowledge. A parallel may be drawn with the building of an entomological research collection. Mass sampling of insects in the field is only the first of many stages in the conversion of a bunch of dead bugs into a well-curated, authoritatively identified resource for science. The 'taxonomic impediment' (Wheeler *et al*., 2004) exists just as much for molecular data as it does for traditional collections.

The idea that a 'DNA barcode', a short stretch of mitochondrial DNA sequence, can be used as a universal identifier for animal taxa has lately been promoted by Hebert and colleagues (2003a, b; 2005), and has met with substantial criticism (Lipscomb *et al.,* 2003; Sperling, 2003; Will & Rubinoff, 2004; Meyer & Paulay 2005; DeSalle *et al*., 2005; Will *et al*., 2005). Nevertheless, major museums and funding agencies have invested in the concept, and proceedings of a meeting devoted to the subject have recently been published in a thematic issue of the *Philosophical Transactions of the Royal Society of London* Series B, Biological Sciences (Savolainen *et al*., 2005, *et seq*.).

Using molecular data for species identification has been used in forensics for almost 20 years (Higuchi *et al*., 1988; Li *et al*., 1988; Sperling *et al*., 1994; Baker & Palumbi, 1994; DeSalle & Birstein, 1996), and has been employed for identifying various closely related insects and associating holometabolous life-stages by a number of authors (Sperling *et al*. 1995; Armstrong *et al*., 1997; Stern *et al*., 1997). Most of the fulmination over the recent popularization of DNA barcodes has been provoked by Hebert *et al*.'s hopeful prediction that the method will replace the expertise of traditional systematists as the primary mode of species identification. This paper addresses some methodological and philosophical weaknesses of the DNA barcoding approach as a proxy for the arduous, painstaking work of genuine systematics.

Association of an unknown specimen with a known species by a DNA barcode is accomplished by finding an identical or similar COI sequence in a pre-established data base of sequences from authoritatively identified specimens through pairwise comparison or a clustering algorithm. The question of how similar a sequence from an unidentified organism and a known sequence must be to be considered to belong to 'the same species' is a metaphysical one. One can only say that the sequence from unknown specimen x is more similar to the sequence of species y than it is to the sequence of any other species currently in the database. As has been demonstrated by many authors (e.g., Crochet *et al*., 2003; Penton *et al*., 2004; Meyer & Paulay, 2005), mtDNA (or any other single feature) does not necessarily provide a precise reflection of species boundaries as they might be implied by a broader sampling

of nuclear genes, morphology, mating preferences, and other biological attributes, so even if the diversity of mtDNA COI sequences were exhaustively sampled (which, of course, it has not been), the closest mtDNA match may not identify its bearer's species correctly.

In a recent paper (Hebert *et al*., 2004, hereafter HPBJH), the scope of DNA barcoding is expanded beyond mere identification of unknowns to the delimitation of multiple new species out of an entity formerly considered to be a single species or species-complex based on traditional morphological characters. This is a bold step because it implies that the information content of a 648 bp fragment of mtDNA reveals more than simply the affinity of its bearer to the most similar reference sequence. Now HPBJH claim that DNA barcodes can themselves be used to delimit formal taxa. Below, I will evaluate how the particular example of the skipper butterfly *Astraptes fulgerator* was used to hypothesize 'ten species in one', and show that both as executed by HPBJH in this case, and as a general principle, the delimitation of species by analysis of a short segment of a single gene is ill-conceived and non-operational.

Methods and materials

As the data proved difficult to extract from Hebert's BoLD web site (www.barcodinglife.com, accessed 12/04), the partial mtDNA COI sequences for 466 members of the *Astraptes fulgerator* complex were individually downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=310673&lvl=3&lin=f&keep=1&srchmode=1&unlock, accessed 12/04). An executable NEXUS matrix is available at www.science.oregonstate.edu/systematics/browera/datasets/Astraptes.html. Individuals are identified by their Genbank accession code and 'species' designation as given by HPBJH (these taxa were not formally described in the publication, so they will be referred to here in quotes). Several individuals are identified only by code and '*fulgerator*'.

The sequences were screened for identity by examination of pairwise distances using PAUP* (Swofford, 2000). One representative of each identical sequence set was retained in a reduced matrix of 71 unique haplotypes (not the '137 different sequences' reported by HPBJH, who apparently counted sequences with missing data as distinct from otherwise identical sequences). A *Pyrgus communis* sequence (AF170857) was used as an outgroup (HPBJH did not include an outgroup in their NJ analysis).

Population aggregation analysis (PAA; Davis & Nixon, 1992) was performed manually by inspection of the aligned NEXUS file. Neighbor-Joining (NJ) bootstrap using the HPBJH weighting parameters and cladistic haplotype analysis (CHA; Brower, 1999) using parsimony with equal weights were performed using PAUP* on a Mac G4 laptop. It should be noted that the low number of phylogenetically informative characters in the data set, even after the elimination of identical duplicate sequences, resulted in computer memory saturation by a multiplicity of equally parsimonious trees prior to completion of heuristic parsimony searches. Cladograms and branch support values were inferred based on these aborted searches, but may be overestimates.

Results

The 'species' identified by HPBJH are terminal clusters in a neighbor-joining tree. It is well-known that NJ analyses are sensitive to the order of the terminals in the matrix (Farris *et al*., 1996), but there is no indication in their published methods that HPBJH tested the stability of groups with multiple runs. Nor were any measures of group support or stability, such as a bootstrap, performed. While I am not an advocate of either NJ or bootstrapping (Neighbor-Joining does not hypothesize monophyletic groups in the Hennigian sense), it is instructive to consider the result of a bootstrap analysis of the data (K-2 weighted, as per HPBJH). Figure 1 shows a reduced cartoon of the bootstrap tree. Only TRIGO and LONCHO are supported at $> 95\%$, and YESSEN, SENNOV, MYST and INGCUP form an undifferentiated bush.

Filtering identical sequences from the 467 terminal matrix resulted mostly in elimination of multiple individuals from the same 'species' (as is not surprising, given that the 'species' were determined on the basis of their sequences). One exception to this is AY666878, which is identified as YESSEN, but is identical to a group of SENNOV sequences. Although this sequence was excluded from PAA and CHA analyses conducted here, its existence requires the combination of YESSEN and SENNOV as a single entity under the criteria of both PAA and CHA.

Population aggregation analysis (Davis & Nixon, 1992) of the reduced 71-unique-terminal matrix, using the HPBJH 'species' designations (CELT, LOHAMP, etc.) reveals 41 polymorphic nucleotide sites that differentiate putative groups within the *A. fulgerator* complex. Of these, 17 sites differentiate TRIGO from the remaining groups. The next most differentiated group is NUMT, which differs from the rest by six sites, then, in decreasing order, CELT+NUMT (4), LONCHO (3), LOHAM (3), CELT (2) and CELT+LOHAM, CELT+YESSEN, CELT+NUMT+YESSEN, CELT+NUMT+SENNOV, INGCUP+LOHAM+MYST, INGCUP+LOHAM+SENNOV, HIHAM (1 each). Note that some of these implied groupings contradict others, suggesting that some of the PAA characters must represent homoplasy rather than homology (Brower, 1999). FABOV, SENNOV, MYST, INGCUP, BYTTNER and YESSEN are not differentiated as distinct taxa by any mtDNA character sensu PAA. Given that the hypothetical 'species' were identified based on the sequence data, it is quite remarkable that there is so little unambiguous, non-homoplastic support present.

Cladistic haplotype analysis (Brower, 1999) provides somewhat more resolution (Fig. 2). Most of the 'species' represented by more than a single terminal appear as distinct terminal groups (at or below the species level, it is inappropriate to refer to clades of mtDNA haplotypes as monophyletic entities; Davis & Nixon 1992). Branch support values range from 17 (TRIGO) to 2 (YESSEN). INGCUP and SENNOV form nonterminal grades, and the distinctness of BYTTNER and HIHAMP is not tested because each is represented by a

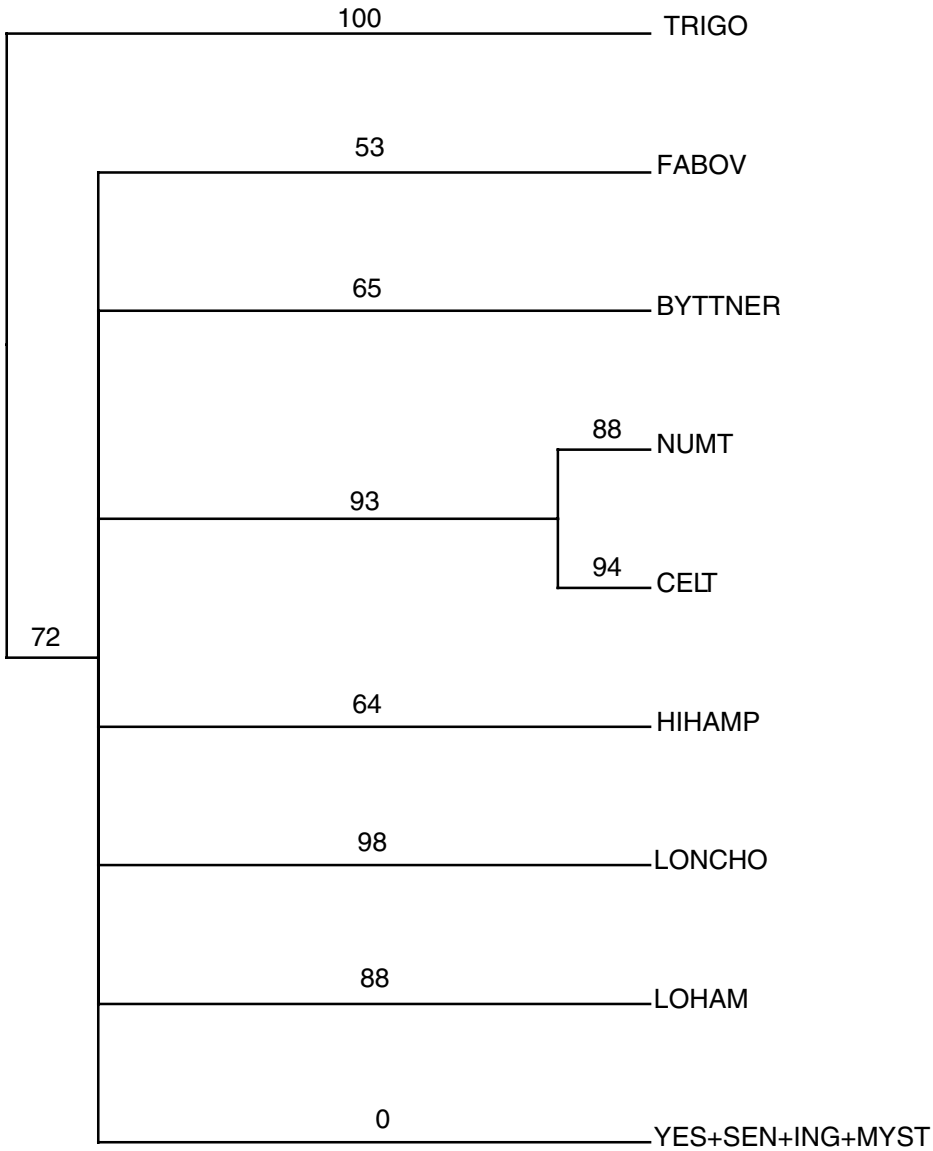

Figure 1 Reduced illustration of NJ bootstrap consensus tree (100 replications) for 466 *A. fulgerator* complex sequences. Only nodes differentiating hypothetical species are shown. YESSEN, INGCUP and SENNOV and MYST are not resolved as distinct entities.

single sequence in the reduced data set. Again, it should be noted that these PAA and CHA analyses do not represent independent tests of the hypothesized 'species', since the 'species' were based on the patterns of similarity among the sequences in the first place.

HPBJH report that 13 butterflies exhibited polymorphic electropherogram bands at certain sites in the COI region. They hypothesized the presence of a second sequence differing from the 'typical' sequence at each polymorphic site, and interpreted these as co-amplifying nuclear pseudogenes (Numts). Four other individual specimens amplified only for the putative Numt sequence. Inspection of these 'Numt' sequences shows that the pseudogene explanation is very unlikely to be correct. Nuclear copies of mitochondrial genes are not constrained by selection like their functional templates, and therefore are expected to accumulate mutations irrespective of nucleotide position (Lopez *et al*., 1997; Bensasson *et al*., 2001). Thus, if these inferred sequences are indeed Numts, an equal number of nucleotide polymorphisms is expected to be seen in first, second and third codon positions. In comparison to members of the most similar non-pseudogene sequence cluster (CELT), 15 of 17 differences occur in normally silent third positions, and the other two are first position T/C transitions that are also silent. The probability of 17/17 mutations in a pseudogene being 'silent' is about one in a million. Alternative hypotheses to explain these apparent mitochondrial heterozygotes are that the individual butterflies are actually heteroplasmic, or that some error or contamination took place in the laboratory. Neither of these alternatives bodes well for the practical success of DNA barcoding as a means to unequivocally identify taxa.

Discussion

A basic flaw of the HPBJH methodology is their failure to explicitly hypothesize the distinctness of putative groups a priori, the existence of which is subsequently tested by analysis of

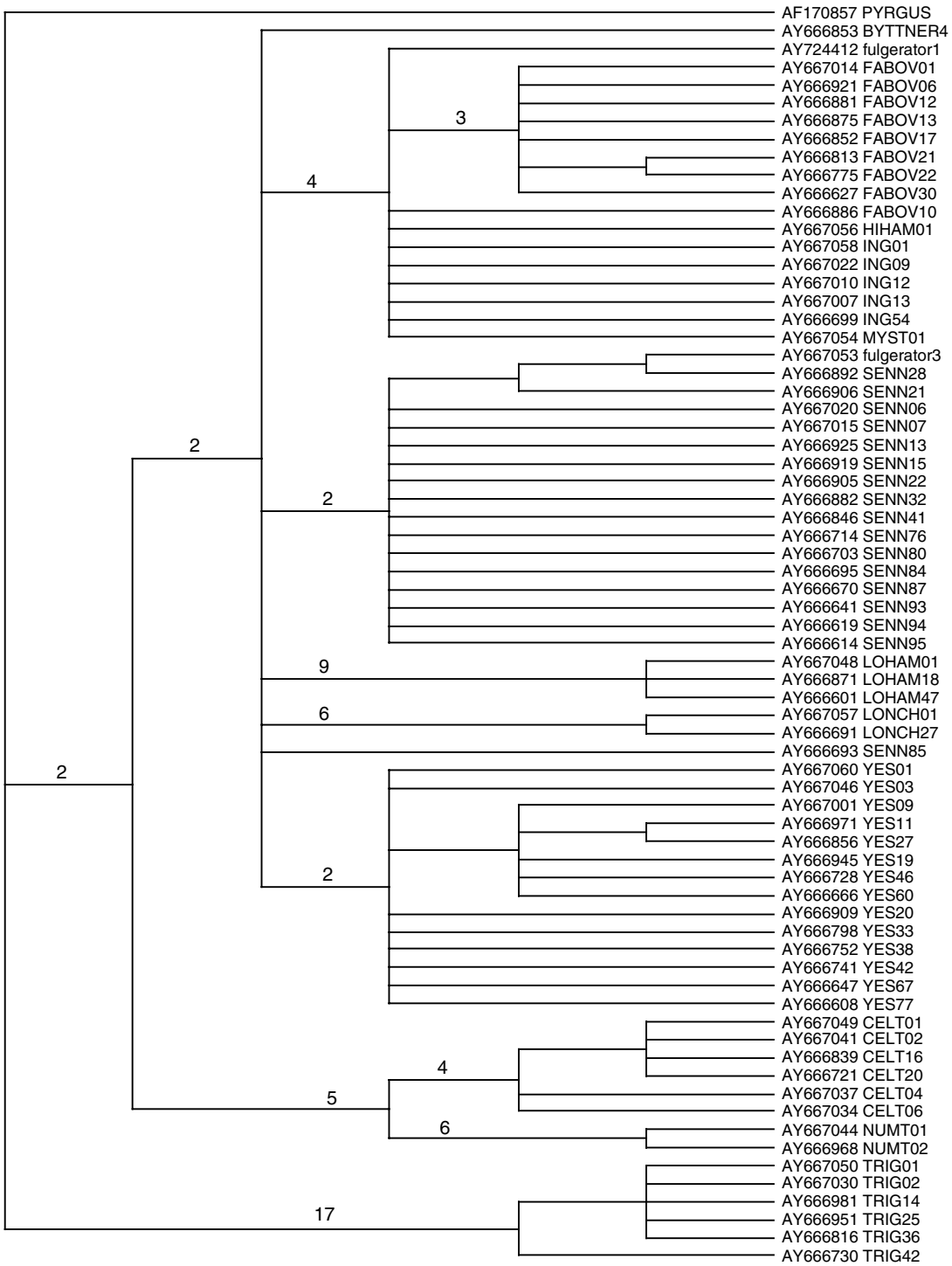

Figure 2 Strict consensus tree of reduced data set of 71 unique *A. fulgerator* sequences, with *Pyrgus* as outgroup. Length = 234 steps, CIx = .5849, RI = .9041. Branch support values are indicated above branches leading to supported 'species' or groups of 'species.' INGCUP and SENNOV are nonterminal grades.

the new DNA evidence (Davis & Nixon, 1992; Brower *et al*., 1996; Brower, 1999; DeSalle *et al*., 2005). Without employing the mtDNA COI evidence as a test of an a priori hypothesis of grouping, conclusions based upon their phylogenetic analysis are tautologous, since any NJ analysis of any variable data will yield a bifurcating topology, the terminal clusters of which can be circumscribed as distinct taxa. As is alluded to in HPBJH's introduction and suggested by the interim names of their hypothesized species, there are morphological and ecological characters that could have been used to formulate testable hypotheses, but instead, these features were discussed in the context of the mtDNA dendrogram, post hoc.

Furthermore, the names associated with these 'species' imply particular ecological patterns that are not cleanly

reflected in the distribution of those associations on the published dendrogram. For example, SENNOV is reported to feed 'chiefly on *Senna hayesiana*', and YESENN 'chiefly on *Senna papillosa*'. While 55 of 76 YESSEN caterpillars were found on *S. papillosa*, only 45 of 100 SENNOV caterpillars were found on *S. hayesiana*. LONCHO and LOHAMP larvae were found on both *Lonchocarpus* and *Hampea*, and caterpillars of two or more 'species' were found on *Canavalia*, *Cassia*, *Cupania*, *Dioclea* and *Inga*. Discovery of a wild larva on a plant means not only that the larva is able to feed successfully on that plant, but also that the adult female selected that plant as a host. The lack of host specificity within and among most of the haplotype clusters suggests that if there are indeed multiple species here, they are not obviously differentiated by larval food plant choice (note that the two 'species' most divergent in their mtDNAs, TRIGO and CELT, were recorded from distinctive plant taxa and do not occur on the hosts of the remaining 'species'). Other sequences that do not fit the general host plant pattern were simply dismissed with an ad hoc and manifestly incorrect explanation.

What conclusions may be drawn from this reexamination of the HPBJH data? First, there are probably at least three species in Guanacaste Preserve within the current circumscription of *Astraptes fulgerator*, but probably not more than the six or seven HPBJH suspected based upon their morphological and ecological characters. Without more extensive sampling from a broader geographical range (*A. fulgerator* s. l. occurs throughout tropical Latin America), it is difficult to interpret the patterns of mtDNA diversity discovered at a single site. Funk & Omland (2003) found that some 23% of animal species (535 out of 2319 records) are polyphyletic as implied by their mtDNA. If this is a general pattern, it means that even under the best of circumstances, a circumscription of terminal clusters as 'species' based on DNA barcoding will be ambiguous or wrong about a quarter of the time (Meyer & Paulay, 2005). This is not to suggest that mtDNA or other DNA sequences are not useful in the discovery of new taxa. The showy mimetic butterfly *Heliconius tristero* was detected initially by its position in a mtDNA cladogram, but the pattern suggested by the molecular evidence was corroborated with morphological characters before the species was described (Brower, 1996). There could well be ten species of *Astraptes* among HPBJH's 466 Guanacaste samples, but the limited information borne by a short fragment of COI sequenced does not support that hypothesis, and further evidence should be presented to corroborate the claim. The era when every sport with a novel colour pattern was described as a new species has happily drawn to a close. It would be unfortunate indeed to diminish the scope of our ongoing taxonomic endeavor to dependence upon a few silent nucleotide substitutions.

Finally, it is worth noting that the BLAST search function of NCBI GenBank (Altschul *et al*., 1990; but see Koski & Golding, 2001) accomplishes the task of associating an 'unknown' sequence with its closest 'known' cognate in the GenBank database efficiently, quickly, and without proprietary complications. Deposition of sequences in this public database will facilitate the professed goal of achieving comprehensive availability of comparative data across a broad taxonomic range more readily than reinventing initiatives such as BoLD with similar missions and functions.

Perhaps endeavouring to mend a perceived schism in the systematics community, some authors have revised and expanded the scope of DNA barcoding to the point where it becomes a synonym for species-level molecular systematics (DeSalle *et al*., 2005; Monaghan *et al*., 2005). Employment of large taxon samples, multiple gene regions, rigorous analyses, and integration of molecular results with evidence from morphology and biogeography, have long been central aims of modern, empirical systematics and are unobjectionable (Will *et al*., 2005). The trouble arises when DNA barcoding is marketed as a substitute for or short cut around these efforts, particularly when its analyses are performed in a perfunctory and cavalier manner. The reason to identify an organism is to connect it to existing knowledge about the group of which it is a member, and to integrate new observations into that general context. In that context, DNA barcoding is a tool, not a research programme. Stated plainly, if resources are cannibalized from systematics to support molecular parataxonomy, systematic training and research programmes will languish, the loss of systematic expertise will be accelerated, and the framework of natural history to which DNA barcodes are intended to link will be impoverished.

Acknowledgements

This work was inspired by a symposium at the 2004 Entomological Society of America meeting. It was supported by NSF DEB-0089870 and the Harold and Leona Rice Endowment for Systematic Entomology.

References

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

Armstrong, K.F., Cameron, C.M. & Frampton, E.R. 1997. Fruit fly (Tephritidae) species identification: a rapid molecular diagnostic technique for quarantine application. *Bulletin of Entomological Research* **87**, 111–118.

Baker, C.S. & Palumbi, S.R. 1994. Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science* **265**, 1538–1539.

Bensasson, D., Zhang, D.-X., Hartl, D.L. & Hewitt, G.M. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution* **16**, 314–321.

Brower, A.V.Z. 1996. A new mimetic species of *Heliconius* (Lepidoptera: Nymphalidae), from southeastern Colombia, as revealed by cladistic analysis of mitochondrial DNA sequences. *Zoological Journal of the Linnean Society* **116**, 317–332.

Brower, A.V.Z. 1999. Delimitation of phylogenetic species with DNA sequences: a critique of Davis and Nixon's Population Aggregation Analysis. *Systematic Biology* **48**, 199–213.

Brower, A.V.Z., DeSalle, R. & Vogler, A. 1996. Gene trees, species trees and systematics: a cladistic perspective. *Annual Review of Ecology and Systematics* **27**, 423–450.

Crochet, P.-A., Chen, J.Z., Pons, J.-M., Lebreton, J.-D., Hebert, P.D.N. & Bonhomme, F. 2003. Genetic differentiation at nuclear and mitochondrial loci among large white-headed gulls: sex-biased interspecific gene flow. *Evolution* **57**, 2865–2878.

DAVIS, J.I. & NIXON, K.C. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology* **41**, 421–435.

DESALLE, R. & BIRSTEIN, V.J. 1996. PCR identification of black caviar. *Nature* **381**, 197–198.

DESALLE, R., EGAN, M.G. & SIDDALL, M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society of London*, Series B, Biological Sciences **360**, 1905–1916.

FARRIS, J.S., ALBERT, V.A., KÄLLERSJÖ, M., LIPSCOMB, D. & KLUGE, A.G. 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12**, 99–124.

FUNK, D.J. & OMLAND, K.C. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology and Systematics* **34**, 397–423.

HEBERT, P.D.N., CYWINSKA, A., BALL, S.L. & DEWAARD, J.R. 2003a. Biological identifications through DNA barcodes. *Philosophical Transactions of the Royal Society of London*, Series B, Biological Sciences **270**, 313–321.

HEBERT, P.D.N. & GREGORY, T.R. 2005. The promise of DNA barcoding for taxonomy. *Systematic Biology* **54**, 852–859.

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 USA* **101**, 14812–14817.

HEBERT, P.D.N., RATNASINGHAM, S. & DEWAARD, J.R. 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Philosophical Transactions of the Royal Society of London*, Series B, Biological Sciences (suppl.) DOI 10.1098/rsbl.2003.0025.

HIGUCHI, R., VON BEROLDINGEN, C.H., SENSABAUGH, G.F. & ERLICH, H.A. 1988. DNA typing from single hairs. *Nature* **332**, 543–546.

KOSKI, L.B. & GOLDING, G.B. 2001. The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* **52**, 540–542.

LI, H., GYLLENSTEN, U.B., CUI, X., SAIKI, R.K., ERLICH, H.A. & ARNHEIM, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* **335**, 414–417.

LIPSCOMB, D., PLATNICK, N. & WHEELER, Q. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends in Ecology and Evolution* **18**, 65–66.

LOPEZ, J.V., CULVER, M., STEPHENS, J.C., JOHNSON, W.E. & O'BRIEN, S.J. 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. *Molecular Biology and Evolution* **14**, 277–286.

MEYER, C.P. & PAULAY, G. 2005. DNA barcoding: error rates based on comprehensive sampling. *Public Library of Science Biology* **3**, e422 (10 pp.).

MONAGHAN, M.T., BALKE, M., GREGORY, T.R. & VOGLER, A.P. 2005. DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* **360**, 1925–1933.

PENTON, E.H., HEBERT, P.D.N. & CREASE, T.J. 2004. Mitochondrial DNA variation in North American populations of *Daphnia obtusa*: continentalism or cryptic endemism? *Molecular Ecology* **13**, 97–107.

SAVOLAINEN, V., COWAN, R.S., VOGLER, A.P., RODERICK, G.K. & 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 **1462**, 1805–1812.

SPERLING, F.A.H. 2003. DNA barcoding: Deus ex machina. *Newsletter of the Biological Survey of Canada (Terrestrial Arthropods)* **22**, 50–53.

SPERLING, F.A.H., ANDERSON, G.S. & HICKEY, D.A. 1994. A DNA-based approach to the identification of insect species used for postmortem interval estimation. *Journal of Forensic Sciences* **39**, 418–427.

SPERLING, F.A.H., LANDRY, J.-F. & HICKEY, D.A. 1995. DNA-based identification of introduced ermine moth species in North America (Lepidoptera: Yponomeutidae). *Annals of the Entomological Society of America* **88**, 155–162.

STERN, D.L., AOKI, S. & KUROSU, U. 1997. Determining aphid taxonomic affinities and life cycles with molecular data: a case study of the tribe Cerataphidini (Hormaphididae: Hemiptera). *Systematic Entomology* **22**, 81–96.

SWOFFORD, D.L. 2000. *PAUP* Phylogenetic Analysis Using Parsimony (*and other methods)*. Sinauer Associates, Sunderland, MA.

WHEELER, Q.D., RAVEN, P.H. & WILSON, E.O. 2004. Taxonomy: impediment or expedient? *Science* **303**, 285.

WILL, K.W., MISHLER, B.D. & WHEELER, Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology* **54**, 844–851.

WILL, K.W. & RUBINOFF, D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* **20**, 47–55.