Limited performance of DNA barcoding in a diverse community of tropical butterflies

Marianne Elias¹,6,*, Ryan I. Hill², Keith R. Willmott³, Kanchon K. Dasmahapatra⁴, Andrew V. Z. Brower⁵, James Mallet⁴ and Chris D. Jiggins⁶

¹Institute of Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK
²Department of Integrative Biology, University of California, Berkeley, CA 94720, USA
³McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA
⁴Department of Biology, University College London, Weldon House, Stephenson Way, London NW1 2HE, UK
⁵Department of Biology, Middle Tennessee State University, Murfreesboro, TN 37132, USA
⁶Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

DNA ‘barcoding’ relies on a short fragment of mitochondrial DNA to infer identification of specimens. The method depends on genetic diversity being markedly lower within than between species. Closely related species are most likely to share genetic variation in communities where speciation rates are rapid and effective population sizes are large, such that coalescence times are long. We assessed the applicability of DNA barcoding (here the 5’ half of the cytochrome c oxidase I) to a diverse community of butterflies from the upper Amazon, using a group with a well-established morphological taxonomy to serve as a reference. Only 77% of species could be accurately identified using the barcode data, a figure that dropped to 68% in species represented in the analyses by more than one geographical race and at least one congener. The use of additional mitochondrial sequence data hardly improved species identification, while a fragment of a nuclear gene resolved issues in some of the problematic species. We acknowledge the utility of barcodes when morphological characters are ambiguous or unknown, but we also recommend the addition of nuclear sequence data, and caution that species-level identification rates might be lower in the most diverse habitats of our planet.

Keywords: DNA barcoding; Amazon; biodiversity; Lepidoptera; mimicry

1. INTRODUCTION

There has been considerable recent interest in the use of short sequence tags known as ‘barcodes’ for the documentation and identification of species (Blaxter 2003; Hebert et al. 2003a,b; Hebert & Gregory 2005; Savolainen et al. 2005; Dasmahapatra & Mallet 2006; Hajibabaei et al. 2007). In arthropods and vertebrates the DNA sequence used as a barcode is the 5’ half of the mitochondrial gene cytochrome c oxidase I (Col). The utility of barcodes to accurately pigeonhole species has been successfully demonstrated in a number of studies (e.g. birds, Hebert et al. 2004a; Kerr et al. 2007), and has been offered as a tool for the discovery of cryptic butterfly and dipteran species (Hebert et al. 2004a; Smith et al. 2006, 2007; van Velzen et al. 2007). However, barcoding has been criticized from both theoretical and practical perspectives (Will & Rubinoff 2004; DeSalle et al. 2005; Ebach & Holdrege 2005; Meyer & Paulay 2005; Will et al. 2005; Brower 2006; Meier et al. 2006). Notably, successful barcode identification depends upon genetic diversity being markedly lower within than between species (Hebert et al. 2004b). This assumption is likely to be broken sometimes, especially when rates of speciation are greater than coalescence times of the gene in question (Pamilo & Nei 1988; Brower et al. 1996; Monaghan et al. 2006; Nielsen & Matz 2006; Wiemers & Fiedler 2007). While the success rate of barcoding undoubtedly varies among groups, some taxa and ecosystems are particularly likely to be subject to difficulties—in particular, groups in which recent speciation rates are high and effective population sizes large and reasonably stable, as is probable in many tropical insects. Here we carry out one of the first community-level barcoding studies in the most diverse terrestrial ecosystem in the world—the upper Amazon basin.

Our study group are the Ithomiinae, an entirely Neotropical subfamily of butterflies containing approximately 360 species confined to moist forest habitats. Adults are distasteful to predators and have warning wing colour patterns, and virtually all species are involved in Müllerian mimicry (Müller 1879; Brown 1984). Mimicry is likely to cause speciation (Jiggins et al. 2001, 2006) and thus to be involved in the diversification of ithomine genera. The alpha taxonomy of ithomines has a relatively long history of study and seems to be fairly well resolved, being largely based on abundant morphological characters including wing pattern, venation, androconia and genitalia.

* Author for correspondence (marianne.elias@ed.ac.uk).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2007.1035 or via http://www.journals.royalsoc.ac.uk.

Received 29 July 2007
Accepted 16 August 2007

2881 This journal is © 2007 The Royal Society
structures (Lamas 2004; Willmott & Freitas 2006). The Ithomiinae are thus a good test case to assess the effectiveness of barcoding.

Our primary aim is to assess the utility of DNA barcoding in the identification of specimens from an exceptionally diverse ithomiine community in eastern lowland Ecuador. Previous studies based on a larger fragment of mitochondrial DNA (mtDNA) revealed that some ithomiine species in Amazonian Peru were not monophyletic (Whinnett et al. 2005), thus questioning the utility of barcoding in this group. Here we measure the frequency of misidentification at the species and genus levels, based on Col barcode sequences only. To take into account geographical diversity, we include also conspecifics from more distant sampling sites, and congeners and other ithomiine genera not represented in the community. We also investigate whether the accuracy of identification is improved by increasing the length of mtDNA sequence and by including a nuclear gene.

2. MATERIAL AND METHODS

(a) Study sites and sampling

Our main sampling effort is focused on two study sites located in eastern Ecuador, across the approximately 1 km wide Rio Napo—Garza Cocha/La Selva Jungle Lodge near the Quechua communities of Piliche and Sani Isla on the north bank of the Añangu Quechua community/Napo Wildlife Center on the south bank. The Ithomiine populations were studied between 2000 and 2007 at Garza Cocha and in 2005 and 2007 at Añangu. We collected all the 58 species present locally (table 1, electronic supplementary material 1). Specimens were identified by the authors using morphology of wing pattern, venation and genitalia, and were collected during each field season for genetic analyses. Bodies or legs were preserved in salt-saturated DMSO (20% DMSO, 0.25 M EDTA, saturated with NaCl), and wings were kept in envelopes. Rarely, individuals were dried as complete specimens and a single leg was used for genetic analysis. Two to nine local individuals per species were used in the analyses (except for the rare species Brevoiraria seba and Diricenna dero, table 1).

(b) DNA extraction and sequencing, and GenBank sequences

DNA was extracted using the QIAGEN DNeasy Kit, according to the manufacturer’s protocol. Two hundred and seventy three local specimens and 80 non-local specimens belonging to the same species or genera (electronic supplementary material 1) were sequenced at the arthropod ‘barcode’ region, namely the 5′ half of the mitochondrial gene 16S (653 bp fragment). Primers, PCR and sequencing reaction conditions are given in the electronic supplementary material 2. In addition, 70 sequences of conspecifics, congeners or other ithomiine genera were downloaded from GenBank and aligned with our sequences (electronic supplementary material 1). Tellervio zeilis (Tellerviinae) was used as an out-group in the analyses detailed below.

To test whether increasing information improves the accuracy of identification, a larger mitochondrial fragment (the entire Col, 1464 bp, tRNA-Leucine, 62 bp and ColII, 716 bp) and a fragment of a nuclear gene, Elongation factor 1a (EFla, 1215 bp), were sequenced or downloaded from GenBank for a subset of the previous specimens (electronic supplementary material 1). Sequences were aligned using CodonCode Aligner v. 1.6.3 (CodonCode Corporation) and checked for reading frame errors using MacClade v. 4.07 (Maddison & Maddison 1997). All sequences have been deposited in GenBank under accession numbers EU068763–EU069266.

(c) Data analyses

Relationships among barcode sequences were inferred with three clustering methods. A neighbour-joining (NJ) tree was based on Kimura 2-parameter (K2P) distances was computed using the software MEGA v. 3.1 (Kumar et al. 2004). Branch support was assessed with 1000 bootstrap replicates. A maximum likelihood (ML) tree was generated with the program PhyML (Guindon & Gascuel 2003) using a GTR+I substitution model. Branch support was assessed with 100 bootstrap replicates. We also performed a Bayesian analysis using the program MrBayes v. 3.1.2 (Huelsenbeck & Ronquist 2001). We performed two runs of four simultaneous Markov chains each for 1 000 000 generations starting from random initial trees and under a GTR+I model, and sampled a tree every 100 generations. Data from the first 900 000 generations (9000 trees) were discarded, after confirming that likelihood values had stabilized well before that. The consensus tree and posterior probability of nodes were calculated from the remaining 1000 trees.

Accuracy of identification was also tested by using the BLAST algorithm (Altschul et al. 1997) as follows. A FASTA file of the barcode data was first formatted as a BLAST database and a local BLASTn search (v. 2.2.15) was carried out on the entire dataset against this database.

Specimens that did not cluster with conspecifics or congeners in our barcode analysis were re-examined carefully to confirm identification (real specimen or photo) and were sometimes sequenced again to confirm initial results.

Two phylogenetic trees based on Col–ColII and EF1α were obtained with MrBayes v. 3.1.2 (Huelsenbeck & Ronquist 2001) as described previously. Mitochondrial sequence data were divided into eight partitions: ColI codon positions 1, 2 and 3, ColII codon positions 1, 2 and 3, tRNA-Leucine, and non-coding DNA upstream of ColI, EF1α sequence data were partitioned by codon position. The best model of substitution for each region was selected using MrModelTest v. 2.2 (Nylander 2004; electronic supplementary material 3). ML and parsimony analyses were also performed, and they showed identical clustering patterns at the genus and species levels (results not shown).

3. RESULTS

(a) Analyses of barcode sequences

The three clustering methods (NJ, ML and Bayesian) gave comparable results for all but three species: Brevoiraria arzalia, Napeogenes sylphis and Pseudoscala timna (table 1). Up to 42 out of 56 species present locally and represented by more than one individual (75%) formed well-supported monophyletic groups (bootstrap or Bayesian probability greater than 50%, table 1, figure 1, electronic supplementary material 4a–c). Species monophyly increased to 77% (44 out of 57) when Napeogenes larina aethra and Mechanitis mazaeus deceptus were considered species separate from Napeogenes larina otases and Mechanitis mazaeus mazaeus, respectively, as these and
## Table 1. Checklist and number of local and geographically distant conspecifics, support values (bootstrap or Bayesian posterior probabilities, in percentage) of the three clustering methods of the barcode sequences for genera and species represented by more than one species or individual.

<table>
<thead>
<tr>
<th>genus (ML, Bayesian and NJ support values)</th>
<th>species</th>
<th>local subspecies</th>
<th>no. local individuals</th>
<th>no. distant conspecifics</th>
<th>support values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aeria</strong> (only one species)</td>
<td>eurimeta</td>
<td>nevricola</td>
<td>6</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td><strong>Brevioleria</strong> (no, no, no)</td>
<td>arzalia</td>
<td>ssp.</td>
<td>5</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td><strong>Callithonia</strong> (100, 100, 100)</td>
<td>aletna</td>
<td>oculata</td>
<td>1</td>
<td>no</td>
<td>78</td>
</tr>
<tr>
<td><strong>Ceratinia</strong> (no, no, no)</td>
<td>tutia</td>
<td>poecila</td>
<td>6</td>
<td>4</td>
<td>99</td>
</tr>
<tr>
<td><strong>Dirusina</strong> (only one species)</td>
<td>dero</td>
<td></td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Episelia</strong> (no, no, 54)</td>
<td>sulphurea</td>
<td>ssp. 1</td>
<td>4</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td><strong>Forbestra</strong> (98, 100, 84)</td>
<td>equicola</td>
<td>equicoloides</td>
<td>3</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Godas</strong> (95, 100, 99)</td>
<td>zavaleta</td>
<td>matronalis</td>
<td>5</td>
<td>3</td>
<td>95</td>
</tr>
<tr>
<td><strong>Heterosais</strong> (only one species)</td>
<td>nephele</td>
<td>nephele</td>
<td>5</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hyallyris</strong> (no, no, no)</td>
<td>sp.</td>
<td></td>
<td>4</td>
<td>no</td>
<td>95</td>
</tr>
<tr>
<td><strong>Hypholyris</strong> (only one species)</td>
<td>lavinia</td>
<td>chrysonoma</td>
<td>5</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hyposcada</strong> (no, no, 21)</td>
<td>anchiasa</td>
<td>ssp.</td>
<td>5</td>
<td>2</td>
<td>84</td>
</tr>
<tr>
<td><strong>Hypothyris</strong> (no, no, no)</td>
<td>anastasia</td>
<td>honesta</td>
<td>5</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hypothonis</strong> (no, no, no)</td>
<td>euclea</td>
<td>intermedia</td>
<td>5</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hypothrys</strong> (no, no, no)</td>
<td>fluonia</td>
<td>berna</td>
<td>5</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hypotheis</strong> (no, no, no)</td>
<td>mamercus</td>
<td>mamerics</td>
<td>7</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hyposcolia</strong> (no, no, no)</td>
<td>moebiusi</td>
<td>moebiusi</td>
<td>5</td>
<td>no</td>
<td>100</td>
</tr>
<tr>
<td><strong>Hyposcuta</strong> (no, no, no)</td>
<td>semiflucta</td>
<td>satura</td>
<td>5</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td><strong>Ithorhinia</strong> (27, 50, 29)</td>
<td>agniosia</td>
<td>agniosia</td>
<td>6</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Melinaeia</strong> (100, 100, 100)</td>
<td>maraesa</td>
<td>macaria/mothone</td>
<td>9</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td><strong>Melinaeis</strong> (100, 100, 100)</td>
<td>menophilus</td>
<td>menophilus</td>
<td>5</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td><strong>Melinaeis</strong> (100, 100, 100)</td>
<td>satevis</td>
<td>maonim</td>
<td>6</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td><strong>Methona</strong> (100, 100, 98)</td>
<td>confusa</td>
<td>confusa</td>
<td>3</td>
<td>5</td>
<td>98</td>
</tr>
<tr>
<td><strong>Napeogenes</strong> (66, 100, 72)</td>
<td>acusa</td>
<td>acusa</td>
<td>3</td>
<td>no</td>
<td>100</td>
</tr>
<tr>
<td><strong>Oleria</strong> (18, 100, no)</td>
<td>agarista</td>
<td>agarista</td>
<td>2</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td><strong>Pteronymia</strong> (65, 100, 56)</td>
<td>primula</td>
<td>primula</td>
<td>5</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td><strong>Pseudoscada</strong> (no, no, no)</td>
<td>florula</td>
<td>aureola</td>
<td>4</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Procris</strong> (100, 100, 99)</td>
<td>zizia</td>
<td>batesi</td>
<td>5</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

(Continued.)
other data suggest (Elias et al. 2007; R. I. Hill et al. 2006–2007, unpublished data). Consequently, Nauphoeta \textit{larina} \textit{aethra} and \textit{M. mazaetus} \textit{dezractus} will be considered ‘good’ species in the analyses below. Considering only those species from the community dataset represented by more than one congener and including remote conspecifics in the analysis (i.e. situations where barcoding would be most useful), only 28 out of 41 species (68\%) were monophyletic. Similar results were achieved using the BLAST algorithm. In 13 of the species (23\% of the species, 32\% of the species with congeners and remote conspecifics), at least one non-conspecific individual showed a higher BLAST bit score as compared with other conspecifics in the dataset.

In the genus \textit{Melinaea}, the three local species were entirely undifferentiated, confirming previous observations from Peru (Whinnett et al. 2005). Similarly, sister species pairs such as \textit{Forbesia olivencia} and \textit{Forbesia procersis}, \textit{Pteronymia vestilla} and \textit{Pteronymia primula}, \textit{Callithomia lenea} and \textit{Callithomia alexirhooe} were partly or completely undifferentiated with respect to one another. In contrast, \textit{M. mazaetus} \textit{dezractus} and \textit{Oleria agarista} were split into several lineages, thus becoming para- or polyphyletic with respect to closely related taxa. Failure to form monophyletic groups occurred between local conspecifics in 10 species, while only divergence between geographical populations was involved in the remaining three cases (figure 1). There was no differentiation between the two banks of the Rio Napo, the only exception being \textit{O. agarista} (electronic supplementary material 4a-c).

Genetic distances within species present locally (0.0085±0.0137) were on average much lower than distances between congenic species (0.0602±0.0292, figure 2a). However, the distributions overlapped, with intraspecific distances over 0.08 in \textit{Oleria} and \textit{Hyposcada} contrasting with no interspecific divergence in \textit{Melinaea} and \textit{Forbesia} (figure 2a,b).

At the generic level, 9 (ML) to 11 (Bayesian analysis) of the 18 genera present locally and represented by more than one species (50–61\%) were monophyletic (figure 1, table 1, electronic supplementary material 4a-c). Paraphyletic genera were \textit{Brevoirleria}, \textit{Ceratinia}, \textit{Episca}, \textit{Hyalyris}, \textit{Hyposcada}, \textit{Hypothyris}, \textit{Ithomia}, \textit{Oleria} and \textit{Pseudocada}.

(b) Improved accuracy by using more genes

Inclusion of the entire Col-ColI region resulted in the same relationships within species as with the barcode alone (electronic supplementary material 5a). The only exception to this was \textit{O. agarista}, which was monophyletic in the full mtDNA dataset. The nuclear gene \textit{EF1\alpha} did contribute more to identification accuracy (electronic supplementary material 5b). \textit{Pteronymia primula}, \textit{F. procersis}, \textit{F. olivencia} and \textit{O. agarista} were monophyletic, sequences of the last species being virtually identical. All \textit{M. mazaetus} clustered together, but there was no distinction between \textit{M. mazaetus} \textit{mazaetus} and \textit{M. mazaetus} \textit{dezractus}.

At the generic level, \textit{Ceratinia} remained paraphyletic with respect to \textit{Episca} with both gene regions. Species from the genera \textit{Greta} and \textit{Pseudocada} were intermixed in the mtDNA analysis, but with \textit{EF1\alpha} \textit{Pseudocada} species formed a monophyletic group, close to a paraphyletic \textit{Greta}. \textit{Oleria} became monophyletic with the longer mitochondrial region but not with \textit{EF1\alpha}. \textit{Hyalyris}, \textit{Hyposcada} and \textit{Ithomia} formed monophyletic groups using each of the additional regions.

4. DISCUSSION

Barcoding has triggered a passionate debate between proponents who aim to ‘tag’ the diversity of life (Hebert et al. 2003a,b; Hebert & Gregory 2005; Savolainen et al. 2005; Hajibabaei et al. 2007), and detractors who highlight the pitfalls of the single-gene approach and are concerned about competition for funds with traditional taxonomy (Will & Rubinoff 2004; Ebach & Holdrege 2005; Wheeler 2005; Will et al. 2005; Brower 2006; Cameron et al. 2006; Rubinoff 2006). It has been claimed that mtDNA barcodes achieve an accuracy close to 100\% in delimiting species in some groups (Hebert et al. 2004b; Hajibabaei et al. 2006a; Clare et al. 2007; Ekrem et al. 2007; Kerr et al. 2007), although in some cases it has proved less successful (Meyer & Pauly 2005; Meier et al. 2006; Wiemers & Fiedler 2007; Whitworth et al. 2007). Our own study shows a rather poor success of DNA barcoding, with only 77\% of species unambiguously identified.

Previous barcoding work in the tropics has mainly focused on the Guanacaste National Park in Costa Rica (Hebert et al. 2004a; Janzen et al. 2005; Hajibabaei et al. 2006a; Smith et al. 2006, 2007), a mosaic of mainly dry forest and cloud forest habitats which is less diverse than the habitat studied here. For example, in the whole of Costa Rica (51 100 km$^2$) there are 61 species of ithomiine butterflies (DeVries 1997), similar to the diversity found in the 21 km$^2$ sampled in the Napo community. The relatively aseasonal and continuous habitat of the upper Amazon probably maintains large and stable population sizes of these butterflies, with correspondingly longer coalescence times when compared with more seasonal habitats such as dry forest, or temperate regions where populations have been reduced to small refugia during glacial periods. Moreover, butterfly genera of the upper Amazon tend to
contain more species (Lamas 2004). Both factors point to a greater challenge for gene-based identification in the Amazon. Not surprisingly, the six species without congener in our analysis were accurately diagnosed. However, adding congener and geographical populations decreased the success of barcoding: 32% of such...
species could not be diagnosed. In other words, the barcode method of identification becomes significantly less reliable when groups of closely related (congeneric) species are examined, or geographical populations of the same species are included, exactly the circumstances where barcoding would be most useful as a tool. Thus, biogeographic and evolutionary history also plays a role in the success of the identification of specimens using barcoding.

Hebert et al. (2004b) suggested the use of a threshold in sequence divergence in the discovery of new species, claiming that barcoding could be an outstanding tool for this purpose (Gomez et al. 2007). Yet the existence of the so-called ‘barcode gap’ has been challenged both on theoretical (Hickerson et al. 2006) and empirical grounds (Cognato 2006; Wiemers & Fiedler 2007). The pattern of barcode evolution in ithomiine butterflies, with highly variable levels of divergence within different species complexes and genera (Whinnett et al. 2005; figure 2), adds fuel to these criticisms. We agree that barcoding should not be relied on as the sole basis for species discovery, except as a preliminary effort in the case where ecological, morphological or additional genetic data are absent.

(a) Generic level identification
At best 61% of currently recognized genera represented by more than one species formed stable clusters. Some generic circumscriptions are likely to be revised in the near future on the basis of ongoing morphological and molecular work (Brower et al. 2006; Willmott & Freitas 2006; K. R. Willmott 2001–2007, unpublished data), and some taxa could be moved from one genus to another. This might affect the pairs of genera Pseudosenda and Greta, Hypothyris and Hyaluris, and Ceratina and Epicasta. In the most optimistic hypothesis where these pairs of genera in our study are pooled, only 11–13 of the rearranged genera represented locally (69–81%) are monophyletic. Thus, even at the generic level, the barcode performs rather poorly in assigning names to species.

(b) Limitations of existing databases
DNA barcodes cannot be a useful identification tool without a comprehensive and reliable reference database (Meyer & Paulay 2005; Brower 2006; Hajibabaei et al. 2006; Scheffer et al. 2006). To investigate the extent to which existing databases can be used to identify specimens in our study, we tested the Barcode of Life Data Systems identification engine (BoLD-ID) in August 2007 (Ratnasingham & Hebert 2007; http://www.barcodinglife.com/views/idrequest.php), with mixed results. We used a subset of 61 sequences of local specimens representing all species and major lineages within species. While 74% of the specimens were correctly identified at the genus level using the full database (but only 41% using the reference database), a number of specimens (18 and 44%, using the full and reference databases, respectively) were assigned to an incorrect (and unrelated) genus with a probability of placement of 1. The remaining specimens could not be assigned to any genus. In the latter cases, the paucity of ithomiine sequences in the database (15 and 9 ithomiine genera found in the full and reference databases, respectively) explains the failure to hit the correct genera. Moreover, the decision system that generates the probabilities of placement tends to overestimate these values for taxa poorly represented in the database. Thus the BoLD-ID engine is currently not suitable for a reliable identification of ithomiines (and that of other relatively poorly represented groups), although the identification accuracy should improve when the sequences of this study are included in the BoLD database. As a comparison when BLASTing the same sequences against the GenBank nucleotide database, where all the tested genera are represented, the correct genus was identified in 88% of the cases.

(c) Analytical methods for barcode data
As no consensus has been reached yet concerning the analysis of barcode data, there is a growing literature testing and describing analytical methods (e.g. Little & Stevenson 2007). A number of authors have proposed or

---

**Figure 2.** Distribution of within-species (white) and between congeneric species (black) K2P distances of barcode sequences. Only species and genera represented locally are considered. (a) All pairwise distances and (b) only maximum intraspecific and minimum interspecific distances.
tested novel methods based on diagnostic molecular characters relying either on the presence/absence of short fragments of sequence (DasGupta et al. 2005; Little & Stevenson 2007) or on parsimony informative sites or combinations of sites referred to as ‘characteristic attributes’ (Sarkar et al. 2002; Kelly et al. 2007). We have not used any of these methods as we believe that the use of diagnostic molecular characters to identify species is likely to result in lower accuracy due to incomplete conspecific reference material (see previous paragraph), especially from across the geographical ranges of the species (Little & Stevenson 2007). To establish a sequence database for the Ithomiinae over the entire range of the subfamily would require collecting and sequencing more than 1600 morphologically divergent subspecific taxa, which is a daunting task.

Here, we compared three alignment-based clustering methods and an alignment-free similarity method, which gave similar results. Since the NJ clustering and the similarity method (BLAST) perform considerably faster than the others, we believe they are a good choice for barcode analyses (Little & Stevenson 2007). NJ clustering has indeed been used in the great majority of published barcoding studies. However, the most promising area for future research is methods that take into account population size and therefore coalescence times in the estimation of barcoding probabilities (Nielsen & Matz 2006).

(d) The value of nuclear data

Despite extensive use of the mitochondrial gene COI in vertebrates, arthropods and, more recently, fungi (Seifert et al. 2007), there are potential drawbacks to using the mitochondrial genome (Rubinoff et al. 2006). Notably, maternal inheritance, the absence of recombination or the spread of cytoplasmically inherited symbionts can lead to misleading identification using the mtDNA barcode (Hurst & Jiggins 2005; Rubinoff et al. 2006; Whitworth et al. 2007).

Our data have demonstrated the value of adding nuclear sequence data to the mtDNA barcode. Despite the relatively slow rate of evolution at EF1α, it has nonetheless resolved several cases where the mtDNA barcode failed. Notably, species with very divergent mtDNA sequences between geographical populations showed nearly identical EF1α sequences. As already suggested (Dasmahapatra & Mallet 2006; Smith et al. 2006, 2007; Gomez et al. 2007) we would therefore recommend that the mtDNA barcode be supplemented where possible by one or more nuclear genes. EF1α has been used here as it is widely studied in insect phylogenetics and easily amplified across many insect groups (Wahlberg et al. 2005; Danforth et al. 2006), although its slow rate of evolution means that other genes, such as rapidly evolving introns (Beltrán et al. 2002), might be more informative for closely related species.

5. CONCLUSION

Our work on ithomiine butterflies has demonstrated that the barcode method of identification, as currently applied, may be significantly less successful in certain circumstances than proponents have suggested. In particular, diverse tropical invertebrate faunas with multiple congeneric species and geographical races, in other words the bulk of biodiversity, pose a particular challenge. Despite the limitations of barcoding for species identification and discovery, the use of mitochondrial sequence data is clearly a valuable tool for taxonomy. Insect molecular systematics employing mtDNA has enjoyed great success for two decades (DeSalle et al. 1987; Brower 1994; Caterino et al. 2000). Unexpected mtDNA COI patterns have led to the discovery of cryptic species (Brower 1996; Hebert et al. 2004a; Malletaro et al. 2005; Smith et al. 2006, 2007). Similarly, barcode patterns have encouraged us to search for covarying morphological, genetic and ecological characters in the genus Mechanitis (R. I. Hill et al. 2006–2007, unpublished data) and in the species Napoegenes larina (Elias et al. 2007), where divergent lineages probably represent unrecognized cryptic species. We are also using barcode sequences to recover the identity of immature stages and thus host plant records, where larvae were collected in the field but died before reaching the adult stage (K. R. Willmott et al. 2005–2007, unpublished data), and to identify the sources of blood meals from engorged mosquitoes (Townzen et al. 2005, unpublished data). Finally, mitochondrial barcode sequences are sometimes the best sequence data that can be obtained from old, degraded museum specimens. The growing public databases containing a homologous sequence region from a diversity of taxa will undoubtedly continue to be a valuable tool for the study of biodiversity.

We thank Raul Aldaz, Julia Robinson Willmott, Aniko Zölei, Gabor Papp, Alexander Toporov and Carlos Sanchez for their help in the field. We thank Gerardo Lamas, Mathieu Joron, Mark Blaxter and two anonymous referees for their useful comments and taxonomic advice and Fraser Simpson and Lisa Leadbeater for providing specimens. We thank the Ministerio del Ambiente and the Museo Ecuatoriano de Ciencias Naturales in Ecuador, and the Instituto Nacional de Recursos Naturales (INRENA), the Museo de Historia Natural and the Universidad Nacional Mayor de San Marcos in Peru for collecting permits and support. The Napo Wildlife Centre and La Selva Jungle Lodge provided logistic support for our fieldwork. This research was funded by the Leverhulme trust and the Royal Society (M.E., K.R.W. and C.D.J.), by the Margaret C. Walker fund for teaching and research in systematic entomology (R.I.H.), NERC (K.K.D. and J.M.), and by the US National Science Foundation grants DEB 0089886 and DEB 0640301 (A.V.Z.B.).

REFERENCES

Brower, A. V. Z. 1996 A new mimic species of Heliconius (Lepidoptera: Nymphalidae), from southeastern


Elias, M., et al. 2007 Phylogenetic hypothesis, pattern of speciation and evolution of wing pattern in neotropical Napeages butterflies (Lepidoptera: Nymphalidae). In Seventh Int. Workshop on the Molecular Biology and Genetics


Nylander, J. A. A. 2004 MrMODELTEST v2. Program distributed by the author. Uppsala University, Sweden: Evolutionary Biology Center.


