MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2009) **9**, 1302–1310 doi: 10.1111/j.1755-0998.2009.02577.x

DNA BARCODING

DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification

VLADIMIR A. LUKHTANOV,*† ANDREI SOURAKOV,‡ EVGENY V. ZAKHAROV§ and PAUL D. N. HEBERT§

*Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St. Petersburg, Russia, †Department of Entomology, St. Petersburg State University, Universitetskaya nab. 7/9, 199034 St. Petersburg, Russia, ‡McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA, §Biodiversity Institute of Ontario, University of Guelph, Guelph, ON, Canada N1G 2W1

Abstract

DNA barcoding employs short, standardized gene regions (5′ segment of mitochondrial cytochrome oxidase subunit I for animals) as an internal tag to enable species identification. Prior studies have indicated that it performs this task well, because interspecific variation at cytochrome oxidase subunit I is typically much greater than intraspecific variation. However, most previous studies have focused on local faunas only, and critics have suggested two reasons why barcoding should be less effective in species identification when the geographical coverage is expanded. They suggested that many recently diverged taxa will be excluded from local analyses because they are allopatric. Second, intraspecific variation may be seriously underestimated by local studies, because geographical variation in the barcode region is not considered. In this paper, we analyse how adding a geographical dimension affects barcode resolution, examining 353 butterfly species from Central Asia. Despite predictions, we found that geographically separated and recently diverged allopatric species did not show, on average, less sequence differentiation than recently diverged sympatric taxa. Although expanded geographical coverage did substantially increase intraspecific variation reducing the barcoding gap between species, this did not decrease species identification using neighbour-joining clustering. The inclusion of additional populations increased the number of paraphyletic entities, but did not impede species-level identification, because paraphyletic species were separated from their monophyletic relatives by substantial sequence divergence. Thus, this study demonstrates that DNA barcoding remains an effective identification tool even when taxa are sampled from a large geographical area.

Keywords: Central Asia, DNA barcoding, geographical diversity, Lepidoptera, species identification, species paraphyly

Received 25 October 2008; revision accepted 23 December 2008

Introduction

DNA barcoding aims to provide an efficient method for species-level identification of biological specimens using short, standardized gene regions (Hebert *et al*. 2003; Hajibabaei *et al*. 2007). The effectiveness of a cytochrome oxidase subunit I (COI)-based identification system has been demonstrated in several groups of animals, such as birds (Hebert *et al*. 2004b), fishes (Ward *et al*. 2005), butterflies (Hebert *et al*. 2004a; Janzen *et al*. 2005; Hajibabaei *et al*. 2006), flies (Smith *et al*. 2007), and spiders (Barrett & Hebert 2005). DNA barcoding systems are now also being established for plants (Kress *et al*. 2005), macroalgae (Saunders 2005), fungi (Summerbell *et al*. 2005), protists (Scicluna *et al*. 2006), and bacteria (Sogin *et al*. 2006). The utility of barcoding relies on the fact that genetic variation within species is smaller than that between species. This gap was found in the studies above

Correspondence: Vladimir A. Lukhtanov, Fax: +7 (812) 4507310; E-mail: lukhtanov@mail.ru

and produced a 95–98% success rate in species identification. However, other investigations have shown lower success in recently diverged groups of butterflies (Kandul *et al*. 2004; Lukhtanov *et al*. 2005, 2008; Elias *et al*. 2007; Wiemers & Fiedler 2007) and flies (Meier *et al*. 2006). Some of these cases of compromised resolution are due, at least in part, to taxonomic errors. For example, Meier *et al*. (2006) based their taxonomic assignments on records extracted from GenBank, which is widely acknowledged to contain many specimens with wrong species-level assignments.

Most past barcoding studies have concentrated on the analysis of local faunas. Critics have suggested that the method will be much less effective when allopatric species and additional populations are included in the study (Moritz & Cicero 2004; Meyer & Paulay 2005; Dasmahapatra & Mallet 2006; Meier *et al*. 2006; Elias *et al*. 2007; Wiemers & Fiedler 2007). This prediction of reduced efficiency has two theoretical bases. First, many recently diverged taxa may be excluded from local analyses 'as they do not necessarily occur in the areas over which the sampling was carried out' (Dasmahapatra & Mallet 2006), since allopatric speciation is the most common mode of speciation (Coyne & Orr 2004) and many recently diverged species are expected to be allopatric. Second, it has been argued that intraspecific variation will be seriously underestimated by local studies because intraspecific geographical variation of barcoding DNA region is not considered. A wider geographical analysis presented here will examine the effectiveness of DNA barcoding.

The present study tests if the utility of DNA barcoding diminishes when the geographical scope is increased by analysing Central Asian butterflies collected from an area of about 6 million km^2 including Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan and the southwestern part of Russian Asia (Fig. 1). The number of butterfly species in this region is relatively small (*c*. 500 species), but the fauna does include numerous diverse endemic genera, subgenera and species groups, enabling comparison of many recently evolved sister species (Lukhtanov & Lukhtanov 1994; Tuzov 1997–2000).

More specifically, we addressed two issues:

1 Do sympatric species really display on average more sequence differentiation than allopatric species?

2 How does the inclusion of an additional, geographically separated population influence the effectiveness of DNA barcoding?

Materials and methods

Study sites and sampling

We obtained 880 COI sequences from specimens representing 370 populations and 318 species of Central Asian butterflies collected by V.A.L. and A.S. between 1984 and 2001. The expeditions collected most species known from Central Asia, often from several widely separated localities. None of the specimens were subjected to any chemical treatment before desiccation. The climate of the regions ensured quick drying of specimens which were stored at a room temperature (18–25 °C) for 5–20 years. A single leg was removed from each specimen before mounting on a pin and wing spreading. Photographs of all specimens used in the analysis as well as collecting data are available in the project 'Butterflies of Central Asia' on the Barcode of Life Data System (BOLD) at http://www.barcodinglife.org/.

All voucher specimens are deposited at the McGuire Center for Lepidoptera and Biodiversity (Florida Museum of Natural History, University of Florida) and are identified with field numbers and with the corresponding unique BOLD Process ID, which is automatically generated by BOLD at the time of first data submission.

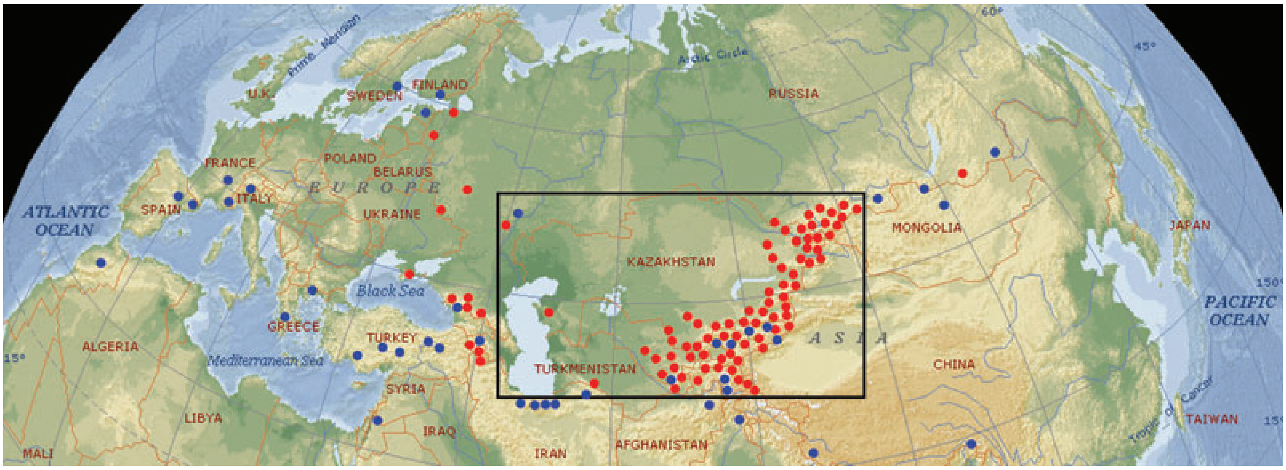

Fig. 1 Map showing both our collection sites (red circles) and collection localities for specimen records from GenBank (blue circles). The target territory (Central Asia) is boxed.

Specimens were identified by comparing vouchers with type specimens at the Natural History Museum (BMNH, London, UK), Zoological Institute of Russian Academy of Science (St. Petersburg, Russia), Zoologisches Museum an der Humboldt-Universität (Berlin, Germany), Zoologische Staatssamlung (Munich, Germany) and Muséum National d'Histoire Naturelle (Paris, France), and by comparison with original descriptions and by researching relevant secondary literature. Species names follow the most recent taxonomic checklist (Tuzov 1997–2000) for butterfly nomenclature and species classification.

Whenever possible, several individuals of each species were analysed to assess intraspecific variation and in such cases, we tried to analyse specimens from two or more populations located as far apart as possible. We increased sample sizes when individuals of different species had similar barcode sequences, or when two or more populations of a species were considered. To enhance geographical coverage, we included 153 COI sequences from the Genetic Sequence Data Bank (GenBank) in our analysis, representing additional species from the studied region and additional populations of Central Asian species from adjacent territories of Eurasia. We only extracted those GenBank records for which we were confident that the species identification was correct. Such confidence could be justified if the images of the voucher specimens were available online (see Table S1, Supporting information), or if species identification was unambiguous due to absence of morphologically similar taxa. Additional selection criteria were (i) nomenclature/taxonomy assigned to the GenBank records was in accordance with Tuzov (1997–2000), and (ii) sequences were at least 600 bp and overlapped with the barcode region in at least 500 bp. We included only those sequences that satisfied all of the above criteria. The list of the GenBank specimens used in the analysis and the corresponding references can be found in Table S1.

COI amplification

DNA was extracted from a single leg removed from each voucher specimen employing a glass fibre protocol (Ivanova *et al*. 2006). All polymerase chain reactions (PCR) and DNA sequencing were carried out following standard DNA barcoding procedures for Lepidoptera as described previously (Hajibabaei *et al*. 2005; deWaard *et al*. 2008). For 78.9% of the samples (684 specimens), the primers LepF (5′-ATTCAACCAATCATAAAGATATTGG-3′) and LepR (5′-TAAACTTCTGGATGTCCAAAAAATCA-3′) amplified the target 658-bp fragment of COI. In 20.5% of the cases (180 individuals) where these primers did not produce a PCR product, we used primer Enh_LepR (5′-CTCCWCCAGCAGGATCAAAA-3′) as a reverse primer. The combination of this primer and LepF amplifies a 609-bp fragment of COI. Finally, for 0.6% of the samples (six individuals) that were recalcitrant, most of which were 23 years old, we amplified shorter overlapping fragments by using the primer combinations LepF + MH-MR1 (5′-CCTGTTCCAGCTCCATTTTC-3′) (307-bp amplicon) and MH-MF1 (5′-GCTTTCCCACGAATAAATAATA-3′) + LepR (407-bp amplicon). Sequences were obtained by using either an ABI PRISM 377 (25% of total sequences, unidirectional read) or an ABI 3730 (75% of total sequences, bidirectional read) sequencer following manufacturer's recommendations.

Sequence analysis

Sequences were edited to remove ambiguous base calls and were assembled using Sequencher (Gene Codes). Sequences were then aligned using ClustalW (Thompson *et al*. 1994) software and manually edited. Sequence information was entered in the Barcode of Life Data System (http://www.barcodinglife.org) along with an image and collateral information for each voucher specimen. The detailed specimen records and sequence information, including trace files, are available in the LOWA project file on BOLD. All sequences are also available through GenBank (Accession numbers FJ663211–FJ664096).

Additional sequences of west Palearctic butterflies from GenBank were included in our analysis when they were at least 600 bp and overlapped with barcode region in at least 500 bp Table S1. The Kimura 2-parameter model of base substitution was used to calculate genetic distances in MEGA 4 software (Tamura *et al*. 2007). MEGA 4 was also used to produce the neighbour-joining tree and to perform bootstrap analysis (2000 replicates).

Results and discussion

Our study examined 1033 individuals representing 353 species and 97 genera, about 70% of the traditionally recognized species of Papilionidae, Pieridae, Nymphalidae and Lycaenidae from Central Asia (Lukhtanov & Lukhtanov 1994; Tuzov 1997–2000). For most taxa, between two and 10 barcode sequences were obtained; just 19% of the taxa (68 species) were represented by a single individual. For the genera *Alpherakya*, *Glabroculus*, *Farsia*, *Triphysa* and *Zegris*, and for certain monophyletic species groups of *Aricia*, *Athamanthia*, *Colias*, *Cupido*, *Erebia*, *Euchloe*, *Hyponephele*, *Karanasa*, *Melitaea*, *Neolycaena*, *Oeneis*, *Parnassius*, *Pieris*, *Pontia* and *Superflua*, we obtained sequences from all known species.

We found that 318 (90.1%) of the 353 species were unambiguously distinguishable from all other species because their barcode sequences formed distinct, non-overlapping monophyletic (86.9%) or paraphyletic (3.2%) clusters in a neighbour-joining (NJ) analysis (for the tree and the bootstrap support, see Fig. S1, Supporting information). Only 34 species (9.6%) could not be distinguished due to undifferentiated barcodes (see Table S2, Supporting information for these species). One taxon (*Melitaea didyma*) (0.3%) appeared

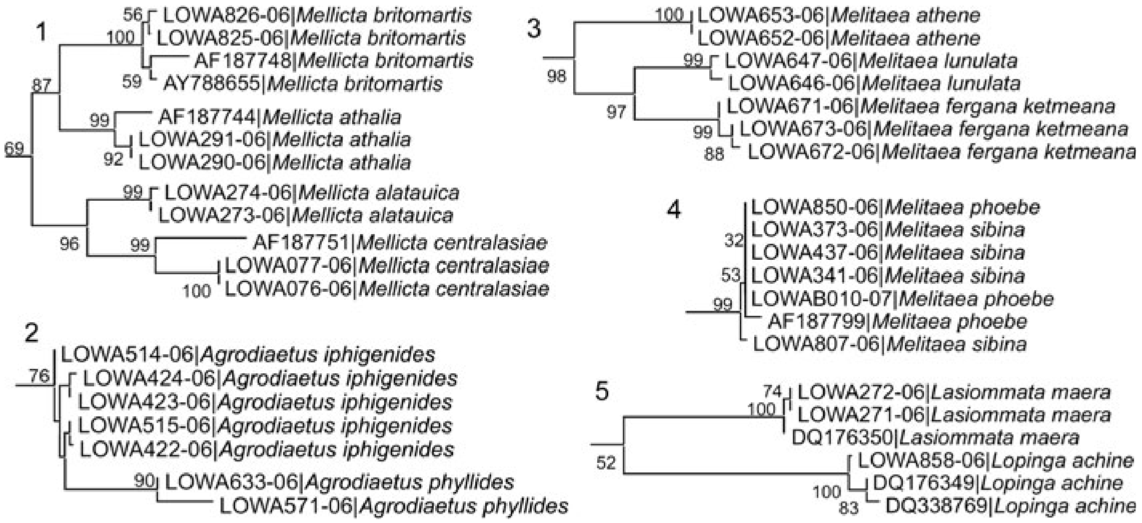
1
56
100
59
87
99
92
69
99
96
99
100
LOWA826-06|Mellicta britomartis
LOWA825-06|Mellicta britomartis
AF187748|Mellicta britomartis
AY788655|Mellicta britomartis
AF187744|Mellicta athalia
LOWA291-06|Mellicta athalia
LOWA290-06|Mellicta athalia
LOWA274-06|Mellicta alatauica
LOWA273-06|Mellicta alatauica
AF187751|Mellicta centralasiae
LOWA077-06|Mellicta centralasiae
LOWA076-06|Mellicta centralasiae
2
76
90
LOWA514-06|Agrodiaetus iphigenides
LOWA424-06|Agrodiaetus iphigenides
LOWA423-06|Agrodiaetus iphigenides
LOWA515-06|Agrodiaetus iphigenides
LOWA422-06|Agrodiaetus iphigenides
LOWA633-06|Agrodiaetus phyllides
LOWA571-06|Agrodiaetus phyllides
3
100
98
99
97
99
88
LOWA653-06|Melitaea athene
LOWA652-06|Melitaea athene
LOWA647-06|Melitaea lunulata
LOWA646-06|Melitaea lunulata
LOWA671-06|Melitaea fergana ketmeana
LOWA673-06|Melitaea fergana ketmeana
LOWA672-06|Melitaea fergana ketmeana
4
32
53
99
LOWA850-06|Melitaea phoebe
LOWA373-06|Melitaea sibina
LOWA437-06|Melitaea sibina
LOWA341-06|Melitaea sibina
LOWAB010-07|Melitaea phoebe
AF187799|Melitaea phoebe
LOWA807-06|Melitaea sibina
5
74
100
52
100
83
LOWA272-06|Lasiommata maera
LOWA271-06|Lasiommata maera
DQ176350|Lasiommata maera
LOWA858-06|Lopinga achine
DQ176349|Lopinga achine
DQ338769|Lopinga achine

Fig. 2 Types of phylogenetic relationships between species of Central Asian butterflies (see Fig. S1 for the complete tree).
2.1 Closely related species. Two pairs of closely related species (*Mellicta britomartis*/*M. athalia*, and *M. alatauica*/*M. centralasiae*) that appear as sister species are shown in this case.
2.2 Paraphyly. *Agrodiaetus iphigenides* is paraphyletic with respect to *A. phyllides* in this case.
2.3 Complex sister taxa. A two-species clade (*Melitaea lunulata* and *M. fergana*) forms the sister taxon to *Melitaea athene* in this case.
2.4 Unresolved phylogenetic relationships. *Melitaea phoebe* and *M. sibina* are unresolved in this case because they share barcodes.
2.5 Divergent sister taxa. *Lasiommata maera* and *Lopinga achine* appear as sister species in this case although they belong to different genera.

in the NJ tree as a clearly polyphyletic entity, consisting of two strongly differentiated lineages (Fig. S1).

Interspecific barcode divergence: comparative analysis of sympatric and allopatric pairs of closely related species

As closely related species, we considered the 68 pairs of terminal taxa that formed dichotomies on the NJ tree (Fig. 2.1), 13 pairs of species that formed paraphyletic–monophyletic assemblages (Fig. 2.2), and 17 pairs of species in which phylogenetic relationships were unresolved because of shared barcodes (Fig. 2.4). For each of the remaining 157 species, the sister group consisted of a multispecies cluster (Fig. 2.3) or another genus (Fig. 2.5). Thus, the first group (196 species) included only pairs of recently diverged species, while the second group (157 species) included more distantly related taxa.

Each of the latter 157 species was separated from its nearest relative by a well-defined barcoding gap, an observation congruent with previous barcode studies on Lepidoptera (Hebert *et al*. 2003, 2004a; Janzen *et al*. 2005; Hajibabaei *et al*. 2006). Consequently, we concentrated our study on the 196 species that formed the 98 most recently diverged species pairs. We first examined whether the geographical distributions of the closely related taxa were overlapping (i.e. sympatric) or not (i.e. allopatric). This analysis revealed that 110 species occurred as sympatric species pairs, while 86 species only occurred in allopatry. Among the 55 sympatric pairs, nine (18 species, 16.4%) were indistinguishable by barcoding (see Table S3, Supporting information). Among the 43 allopatric pairs of closely related species, eight (16 species, 18.6%) were indistinguishable, a value that was not significantly different from that in sympatric pairs ($t = 0.4$, $P = 0.69$).

To gain a better understanding of these unexpected results, we examined these 98 pairs of most closely related species in more detail. This analysis revealed a fundamental difference between the allopatric and sympatric species pairs with shared barcodes (see Table S2 for more details). Based on prior literature and our knowledge of the alpha taxonomy of this fauna, we conclude that every pair of allopatric taxa with undifferentiated barcodes actually represents a case of 'over-splitting.' In short, these cases of 'compromised' barcode resolution actually represent a lack of divergence among conspecific populations.

This re-consideration of species status following barcoding analysis might be viewed as subjective. However, this conclusion is also supported by 40 years of our study of this fauna (e.g. Lukhtanov & Lukhtanov 1994). Furthermore, as noted in the supporting information (Table S3), the questionable status of several of these species pairs has already been noted by other authors. By contrast, all sympatric taxa with shared barcodes appear to be well-defined biological species (e.g. *Colias alpherakyi*–*C. wiskotti*, *Parnassius actius*–*P. tianschanicus*) that occupy distinct ecological niches and have clear morphological differences without intermediates. Unlike the allopatric species pairs, these taxa have never been considered conspecific in the taxonomic literature. Hence, in contrast to our expectations, the percentage of species pairs with undifferentiated barcodes is significantly higher ($P < 0.001$) in sympatric (17.3%) than allopatric (0%) taxa. We conclude that sympatric species pairs are more likely to share barcode sequences than allopatric pairs.

There are at least three explanations for this unexpected observation. First, the existence of sympatric taxa with undifferentiated barcodes could reflect recent sympatric speciation. Second, speciation could begin in allopatry, but be completed in sympatry by means of ecological character displacement or reinforcement mechanisms. The latter two processes are driven by natural selection and could lead to a rapid formation of pre-zygotic isolation that may depend on changes at just a few gene loci (Coyne & Orr 2004; Rundle & Nosil 2005; Hendry *et al*. 2007), while the general genetic background (including the barcode region of mtDNA) might remain unchanged. Conversely, allopatric populations may need greater genetic divergence to achieve reproductive isolation simply by chance. Third, this observation might reflect mitochondrial introgression between sympatric taxa. In our study, introgression seems a likely explanation for barcode sharing in some species pairs, such as *Colias crocea–C. erate* or *Parnassius actius–P. tianschanicus*, because natural interspecific hybrids have been recorded (Eisner 1976; Descimon & Mallet 2008).

Although we lack sufficient evidence to select among these three explanations, evidence for sympatric speciation is scant (Coyne & Orr 2004; Bolnick & Fitzpatrick 2007) (but see Berlocher & Feder 2002). By contrast, it is accepted that rapid speciation linked to secondary sympatry is common in both plants and animals (Rundle & Nosil 2005; Hendry *et al*. 2007). If so, the pattern of greater genetic differences between allopatric vs. sympatric species should be frequent in nature. Mitochondrial introgression is also frequent (Ballard & Whitlock 2004) although Haldane's rule (Haldane 1922) suggests that it should be rare in butterflies because of their female heterogameity (Kandul *et al*. 2007). Despite this barrier, cases of mitochondrial introgression are known in Lepidoptera (Sperling 1993; Jiggins 2006).

To conclude, despite theoretical predictions, the inclusion of allopatric species did not reduce the ability of mitochondrial barcodes to distinguish Central Asian butterflies, and there are reasons to think that this will be the case in other organisms.

Intraspecific variation: how does the inclusion of geographically separated populations influence DNA barcoding?

To provide a critical test of this question, we focused analysis on 37 pairs of closely related species in which at least one species was represented by two geographically distant populations. First, we calculated the maximum intraspecific pairwise genetic distance for each species pair for two instances: (i) when only one local population was considered, and (ii) when an additional population was included in the analysis. We then analysed how adding a distant population changed the barcoding gap by comparing the minimum interspecific distances with maximum intraspecific distances within (i) and (ii) (see Table S3).

As expected, expansion of geographical coverage significantly increased intraspecific variation. The mean value of maximum intraspecific genetic distance increased fivefold: from $x \pm$ S.E. $= 0.26 \pm 0.04\%$ (when one population of each species was considered) to $x \pm$ S.E. $= 1.35 \pm 0.19\%$ (when individuals from different populations were included)

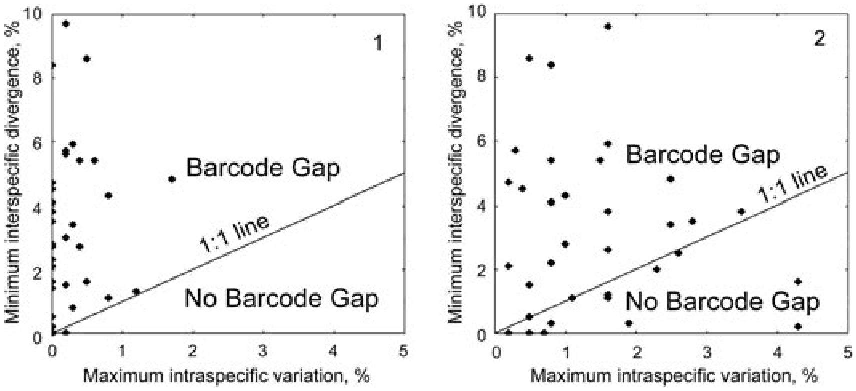

Fig. 3 Influence of geographical coverage on the barcode gap for 37 sister species pairs of Central Asian butterflies.
Each point represents a pair of sister species (Table S3b lists the species analysed). Genetic distances among all individuals of each pair were calculated and the maximum intraspecific distance was plotted against minimum interspecific distances for two situations:
(1) a single local population of each species was considered.
(2) a second population for one sister species was added.
Points above the 1:1 line represent pairs of sister species separated by a barcoding gap, i.e. genetic variation within species is smaller than genetic divergence between species.

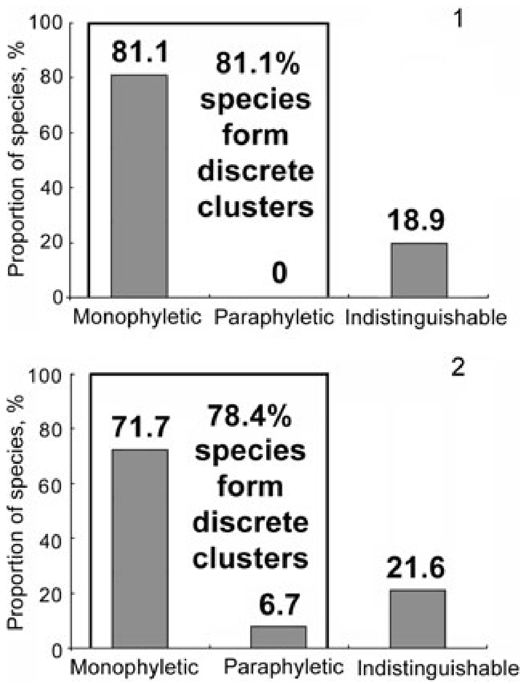

Fig. 4 Influence of increased geographical representation on species clustering (see Table S3b for the list of species analysed and Fig. S1 for their position on the NJ tree).
(1) When one local population of each species is considered, 60 of 74 species form monophyletic non-overlapping entities on a NJ tree (boxed), while 14 species are indistinguishable.
(2) When an additional population is included for each sister species pair, 58 of 74 species remain distinct (boxed), and only two additional species become indistinguishable.

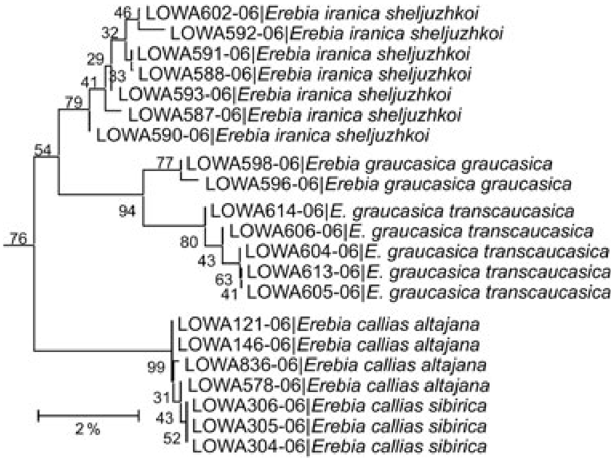
46
LOWA602-06|*Erebia iranica sheljuzhkoi*
32
LOWA592-06|*Erebia iranica sheljuzhkoi*
29
LOWA591-06|*Erebia iranica sheljuzhkoi*
33
LOWA588-06|*Erebia iranica sheljuzhkoi*
41
LOWA593-06|*Erebia iranica sheljuzhkoi*
79
LOWA587-06|*Erebia iranica sheljuzhkoi*
LOWA590-06|*Erebia iranica sheljuzhkoi*
54
77
LOWA598-06|*Erebia graucasica graucasica*
LOWA596-06|*Erebia graucasica graucasica*
94
LOWA614-06|*E. graucasica transcaucasica*
76
80
LOWA606-06|*E. graucasica transcaucasica*
43
LOWA604-06|*E. graucasica transcaucasica*
63
LOWA613-06|*E. graucasica transcaucasica*
41
LOWA605-06|*E. graucasica transcaucasica*
LOWA121-06|*Erebia callias altajana*
LOWA146-06|*Erebia callias altajana*
99
LOWA836-06|*Erebia callias altajana*
LOWA578-06|*Erebia callias altajana*
31
LOWA306-06|*Erebia callias sibirica*
43
LOWA305-06|*Erebia callias sibirica*
2%
52
LOWA304-06|*Erebia callias sibirica*

Fig. 5 Fragment of NJ tree with barcode clusters of the *Erebia iranica-graucasica-callias* group.

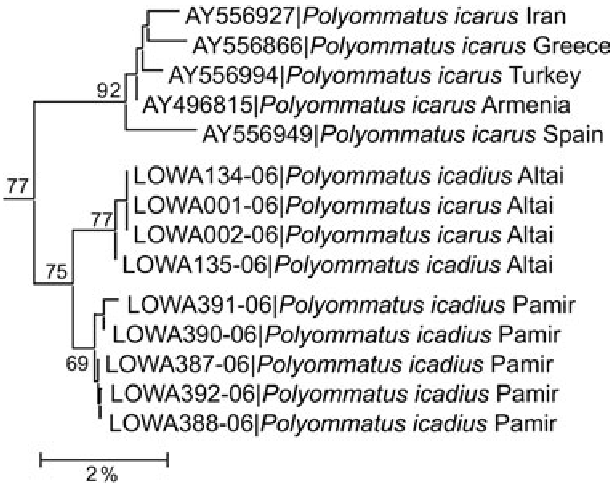
AY556927|*Polyommatus icarus* Iran
AY556866|*Polyommatus icarus* Greece
AY556994|*Polyommatus icarus* Turkey
92
AY496815|*Polyommatus icarus* Armenia
AY556949|*Polyommatus icarus* Spain
77
LOWA134-06|*Polyommatus icadius* Altai
LOWA001-06|*Polyommatus icarus* Altai
77
LOWA002-06|*Polyommatus icarus* Altai
75
LOWA135-06|*Polyommatus icadius* Altai
LOWA391-06|*Polyommatus icadius* Pamir
LOWA390-06|*Polyommatus icadius* Pamir
69
LOWA387-06|*Polyommatus icadius* Pamir
LOWA392-06|*Polyommatus icadius* Pamir
LOWA388-06|*Polyommatus icadius* Pamir
2%

Fig. 6 Fragment of NJ tree with barcode clusters of *Polyommatus icarus* and *P. icadius* from different populations.

(t-value = 5.74, $P < 0.001$), significantly reducing the barcode gap. When one population from each species was considered, most species pairs (30/37, 81.1%) had a clear barcode gap, but it was absent in nearly half of the species pairs (16/37, 43.2%) when two widely separated populations were considered (Fig. 3).

This effect makes species identification using a distance metric problematic as genetic differences between conspecific individuals can often be greater than those between individuals of different species. However, inclusion of additional populations did not seriously impede the ability of the barcodes to identify specimens using NJ clustering because the number of species that formed distinct, non-overlapping clusters on the NJ tree was practically unaffected by geographical coverage (Fig. 4). This result, although seemingly counterintuitive, can be easily explained: clustering does not depend on a barcode gap between species, and genetic differentiation among populations of a species, although it increases intraspecific variability, does not 'fill' the inter-species hiatus. This pattern can be clearly observed within the *Erebia callias-iranica-graucasica* group: maximum genetic distance between individuals of *E. graucasica* (2.6%) exceeds the minimum genetic distance between *E. graucasica* and *E. iranica* (2.5%), but it does not prevent the separation of these species into specific clusters on a NJ tree (Fig. 5).

Our results do not imply that NJ clustering is completely insensitive to geographical variability. Among our 37 species pairs, *Polyommatus icarus* and *P. icadius* were distinct at all localities excepting the Altai region where these two species share barcodes (Fig. 6). Thus, broad geographical sampling can discover populations with shared barcodes due to incomplete lineage sorting or local gene introgression. However, our data indicate that such cases are uncommon. Thus, when NJ clustering is used, the move from local to geographically dispersed sampling does not seriously reduce the ability of DNA barcoding to delineate butterfly species of Central Asia.

Increased geographical coverage does, however, often change the clustering pattern of conspecific individuals.

These clusters of individuals may be mono- or paraphyletic with respect to their relatives (Fig. 2). In our data set, 5 of 37 species shifted from monophyletic to paraphyletic after inclusion of additional populations (Fig. 4), a pattern previously reported in ithomiine butterflies (Elias *et al*. 2007). However, our study suggests that paraphyly is not an obstacle for identification because the separation of paraphyletic species from their monophyletic relatives was supported by high bootstrap values and deep sequence divergences (Fig. 2.2, see also Fig. S1 and Table S4, Supporting information). Barcodes of such species still form non-overlapping clusters (Fig. 2.2), and the species still possess a species-specific combination of diagnostic molecular characters. Identification of such paraphyletic species is possible and can be conducted as follows:

Specimens that constitute the monophyletic complex (sp. 1) are characterized by at least one unique (characteristic for that species) molecular synapomorphy. This synapomorphy (or synapomorphies) not only lies in the basis of monophyletic clustering, but also is used for species identification. The specimens that form the paraphyletic remainder of the paraphyletic–monophyletic assemblage constitute the paraphyletic species (sp. 2). Ascribing barcode sequences to this second paraphyletic species is possible based on the fact that, first, they share molecular synapomorphies with sp. 1 (which clusters all specimens in the assemblage together on a tree) and, second, that these specimens are missing the molecular characteristics that are unique for sp. 1. Hence, a species that is represented by a paraphyletic remainder of the cluster is characterized by a species-specific combination of diagnostic molecular characters (presence of synapomorphies that it shares with sp. 1 and absence of synapomorphies that are characteristic for sp. 1). For instance, this combination of characters allows us to identify species in the paraphyletic–monophyletic assemblages found in our study (*Callophrys rubi*/*Ahlbergia frivaldszkyi*; *Melitaea latonigena*/*sutschana*; *Parnassius staudingeri*/*cardinal*; *Chazara briseis*/*heydenreichi*; *Hyponephele laeta*/*pamira*; *Hyponephele maureri*/*pseudokirgisa*).

Many types of speciation can produce paraphyletic species (Coyne & Orr 2004), so it is no surprise that they are quite common in nature (Avise 2000). DNA barcoding helps to discover such cases, enabling more detailed investigations of their genesis. In any case, paraphyletic–monophyletic complexes contain much more information about the relationships between taxa than complexes of genetically undifferentiated species. Paraphyly will introduce interpretational complexities because the delimitation of species boundaries in paraphyletic taxa cannot be based solely on barcodes. Hence, in such cases, one should always look beyond a purely molecular approach to a decisionary system that employs as wide an array of morphological and ecological characters as possible.

Conclusions

Our study surveyed nearly 70% (353 of 500 species) of the Central Asian butterfly fauna. Our investigations revealed that more than 90% of these species were unambiguously distinguished by their barcode sequences. Sixteen allopatric species could not be differentiated using DNA barcoding. However, critical taxonomic evaluation suggests that those species represent likely cases of over-splitting raising the success in species identification to 95.5%. Despite theoretical predictions, allopatric sister species in our study did not have less COI sequence divergence than sympatric taxa. Expanded geographical coverage did substantially increase intraspecific variation of the barcoding region, reducing the barcoding gap between species, but this increase did not substantially reduce the success of species identification using neighbour-joining clustering. The inclusion of multiple populations did increase the number of paraphyletic entities, but this did not impede species-level identification, because paraphyletic species were separated from their monophyletic relatives by substantial sequence divergence. Thus, we demonstrate that DNA barcoding remains a highly effective identification tool even in nonlocal taxonomic studies.

Acknowledgements

Funding for this study was provided by Genome Canada through the Ontario Genomics Institute, and by a NSERC Discovery Grant to P.D.N.H. This research was also supported by the Russian Foundation for Basic Research (grants RFFI 06-04-49362, 08-04-00295 and 09-04-01234), by grant NSH-963.2008.4 (Leading Scientific Schools) and by the programs of the Presidium of Russian Academy of Science 'Dynamics of Gene Pools in Animals, Plants and Humans' and 'Origin and evolution of biosphere' to V.A.L. We thank the McGuire Family Foundation and Thomas Emmel for facilitating this international collaboration and our many colleagues at the Canadian Centre for DNA Barcoding. Marianne Elias, Daniel Janzen, and Keith Willmott provided valuable comments on earlier versions of this manuscript.

References

Avise JC (2000) *Phylogeography: the History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

Ballard JW, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729–744.

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

Berlocher SH, Feder JL (2002) Sympatric speciation in phytophagous insects: moving beyond controversy? *Annual Reiew of Entomology*, **47**, 773–815.

Bolnick DI, Fitzpatrick BM (2007) Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution and Systematics*, **38**, 459–487.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer & Associates, Sunderland, Massachusetts.

Dasmahapatra KK, Mallet J (2006) DNA barcodes: recent successes and future prospects. *Heredity*, **97**, 254–255.

Descimon H, Mallet J (2008) Bad species. In: *European Butterflies* (eds Settele J, Konvicka M, Shreeve T, Dennis R and Van Dyck H). Cambridge University Press, Cambridge, UK.

Eisner C (1976) Parnassiana nova XLIX. Die Arten und Unterarten der Parnassiinae (Lepidoptera) (Zweiter Teil). *Zoologische Verhandelingen*, **146**, 99–266.

Elias M, Hill RI, Willmott KR *et al*. (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 2881–2889.

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.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.

Haldane JBC (1922) Sex ratio and unisexual sterility in animal hybrids. *Journal of Genetics*, **12**, 101–109.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society 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. *PLoS Biology*, **2**, 1657–1663.

Hendry AP, Nosil P, Rieseberg LH (2007) The speed of ecological speciation. *Functional Ecology*, **21**, 455–464.

Ivanova NV, deWaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high quality DNA. *Molecular Ecology Resources*, **6**, 998–1002.

Janzen DH, Hajibabaei M, Burns JM (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*, **1462**, 1835–1846.

Jiggins CD (2006) Reinforced butterfly speciation. *Heredity*, **96**, 107–108.

Kandul NP, Lukhtanov VA, Dantchenko AV *et al*. (2004) Phylogeny of *Agrodiaetus* Hubner 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences of *COI* and *COII* and nuclear sequences of *EF1-alpha*: karyotype diversification and species radiation. *Systematic Biology*, **53**, 278–298.

Kandul NP, Lukhtanov VA, Pierce NE (2007) Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution*, **61**, 546–559.

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.

Lukhtanov VA, Lukhtanov AG (1994) *Die Tagfalter Nordwestasiens (Lepidoptera, Diurna)*. Herbipoliana, Marktleuthen, Germany.

Lukhtanov VA, Kandul NP, Plotkin JB *et al*. (2005) Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies. *Nature*, **436**, 385–389.

Lukhtanov VA, Shapoval NA, Dantchenko AV (2008) *Agrodiaetus shahkuhensis* sp. n. (Lepidoptera, Lycaenidae), a cryptic species from Iran discovered by using molecular and chromosomal markers. *Comparative Cytogenetics*, **2**, 99–114.

Meier R, Shiyang K, Gaurav V, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.

Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*, **2**, e354.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1879–1888.

Scicluna SM, Tawari B, Clark CG (2006) DNA barcoding of *Blastocystis*. *Protist*, **157**, 77–85.

Smith MA, Wood DM, Janzen DH, Hallwachs W, Hebert PDN (2007) DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proceedings of the National Academy of Sciences, USA*, **104**, 4967–4972.

Sogin ML, Morrison HG, Huber JA *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proceedings of the National Academy of Sciences, USA*, **103**, 12115–12120.

Sperling FAH (1993) Mitochondrial DNA variation and Haldane's rule in the *Papilio glaucus* and *P. troilus* species groups. *Heredity*, **71**, 227–233.

Summerbell RC, Levesque CA, Seifert KA *et al*. (2005) Microcoding: the second step in DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1897–1903.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA 4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Tuzov VK (ed.) (1997–2000). *Guide to the Butterflies of Russia and Adjacent Territories*. Pensoft Publishers, Sofia, Bulgaria and Moscow, Russia.

deWaard JR, Ivanova NV, Hajibabaei M, Hebert PDN (2008) Assembling DNA barcodes: analytical protocols. In: *Environmental Genomics, Methods in Molecular Biology, vol. 410* (ed. Martin CC), pp. 275–283. Humana Press, Totowa, New Jersey.

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.

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist? — a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, **4**, 8.

Supporting information

Additional supporting information may be found in the online version of this article:

Table S1 List of 153 COI sequences from GenBank, representing additional species from Central Asia and additional populations from adjacent territories of Eurasia

Table S2 Taxonomic analysis of sister species pairs that were indistinguishable by their barcodes

Table S3 Lists of sister species pairs analysed

Table S4 List of paraphyletic–monophyletic species assemblages found in Central Asian butterflies

Fig. S1 Bootstrap neighbour-joining tree of all individuals used in the analysis

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.