Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe

Vlad Dinca¹,², Evgeny V. Zakharov³, Paul D. N. Hebert³ and Roger Vila¹,⁴,*

¹Institute of Evolutionary Biology (UPF-CSIC), Passeig Maritim de la Barceloneta 37-49, 08003 Barcelona, Spain
²Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
³Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada N1G 2W1
⁴Institució Catalana de Recerca i Estudis Avancats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

DNA barcoding aims to accelerate species identification and discovery, but performance tests have shown marked differences in identification success. As a consequence, there remains a great need for comprehensive studies which objectively test the method in groups with a solid taxonomic framework. This study focuses on the 180 species of butterflies in Romania, accounting for about one third of the European butterfly fauna. This country includes five eco-regions, the highest of any in the European Union, and is a good representative for temperate areas. Morphology and DNA barcodes of more than 1300 specimens were carefully studied and compared. Our results indicate that 90 per cent of the species form barcode clusters allowing their reliable identification. The remaining cases involve nine closely related species pairs, some whose taxonomic status is controversial or that hybridize regularly. Interestingly, DNA barcoding was found to be the most effective identification tool, outperforming external morphology, and being slightly better than male genitalia. Romania is now the first country to have a comprehensive DNA barcode reference database for butterflies. Similar barcoding efforts based on comprehensive sampling of specific geographical regions can act as functional modules that will foster the early application of DNA barcoding while a global system is under development.

**Keywords:** biodiversity; DNA barcoding; Lepidoptera; morphology; taxonomy

1. INTRODUCTION

Correct identification and monitoring of global biodiversity is a huge task, one that currently overwhelms the available human resources. DNA-based identifications have the potential to resolve this problem by enabling broader participation in the process. Although mitochondrial DNA (mtDNA) analysis has been employed in molecular studies on animals for more than three decades [1], it is only recently that a short, standardized gene region of mtDNA (5′ segment of mitochondrial cytochrome c oxidase subunit I–COI) was proposed as a ‘DNA barcode’ for discriminating most animal species [2,3]. The application of DNA barcoding using other markers for plants has made significant progress [4,5], as well as for fungi [6], macroalgae [7], protists [8] and bacteria [9]. The main goals of this method are (i) to ensure fast and reliable species identification and (ii) to aid the discovery of undescribed species. These goals complement many potential applications related to biodiversity conservation, pest management, forensics and healthcare.

The proposal to develop an identification system based on a single gene marker attracted early criticism, largely based on theoretical considerations such as mtDNA introgression, incomplete lineage sorting and heteroplasmia [10–14]. The method has now proved its effectiveness for various groups of vertebrates (e.g. [15–17]) and invertebrates [18–25]. However, lower success rates have been reported in certain groups of animals [26–33]. This variable performance may reflect biological differences between taxonomic groups, the sampling coverage (both in terms of geography and taxa) and the quality of the taxonomic framework. Therefore, some studies reporting low success have focused on groups with difficult taxonomy such as ithomine butterflies [29] or groups well known for their explosive speciation such as the butterfly genus *Agrodiaetus* [28]. Such lineages are likely to represent worst case scenarios for DNA barcoding, as they actually often do for morphology-based taxonomy. At the opposite pole, studies reporting the highest success rates for DNA barcoding have included few closely related taxa, or have examined a limited geographical area (e.g. [2,3,34]). By careful comparison of the results from morphology and DNA barcoding, our study provides the data needed to objectively assess the identification success of DNA barcoding.
The European butterfly fauna has several attributes that make it a good test bed: it includes more than 500 species that are a flagship for nature conservation and they have an exceptionally well-established taxonomy compared with other invertebrates. In this study, we barcode nearly all of the butterfly species known from Romania, an area well-suited for analysis because samples can be obtained from varied habitats, altitudes and climatic influences, reflecting the fact that Romania is the only state in the European Union where five eco-regions are present (Pannonian, Continental, Alpine, Stepic and Pontic). Therefore, we had at our disposal a dataset originating from a well-defined region representative for the butterfly fauna of temperate areas and consisting of an unbiased sample composition based on uniform collecting throughout the country’s territory. The careful comparison with morphological traits (using linear and/or geometric morphometry when necessary) often employed in species identification allowed an objective assessment of identification results obtained through different methods and to the conclusion that DNA barcoding is a valuable approach for the identification of temperate Rhopalocera.

2. MATERIAL AND METHODS

Methods are described in a more detailed manner (fully referenced) in the electronic supplementary material, annex 1.

(a) Sampling and collection data

We obtained 1387 COI sequences for 180 butterfly species representing 99 per cent of the species with a confirmed occurrence in Romania within the last 30 years, including two new species for the country [35,36]. These samples were collected from 135 localities across the country from April 2006 to June 2009 (see the electronic supplementary material, annex 1). The bodies were stored in tubes with 100 or 96 per cent ethanol, while the wings were detached from the body and kept in glassine envelopes as vouchers.

(b) Morphology examination

A dedicated effort was made to ensure the correct morphology-based identification of each specimen. Besides careful examination of wing morphology, genitalic preparations were made for more than 400 specimens where external features were considered insufficient for certain identification. For taxa with particularly similar genitalic structures, we employed linear and/or geometric morphometry. Linear morphometric analyses employed digital photographs obtained through a stereomicroscope and measured with the software AXIOVISION. For geometric morphometry, a combination of landmarks and sliding semi-landmarks was applied using the tps (thin plate spline) software package.

(c) COI amplification

A glass fibre protocol was employed to extract DNA from a single or half leg of each specimen (depending on size). A 658 bp fragment of cytochrome c oxidase subunit I (COI) was targeted for amplification following standard procedures for Lepidoptera.

(d) Sequence analysis

Sequences were edited and assembled using either SEQUENCHER 4.5 (GeneCodes Corporation, Ann Arbor, MI) or CODONCODE ALIGNER 3.0. Sequence alignment was done in MEGA 4 software. Genetic distances were calculated in MEGA 4 under the Kimura 2-parameter model of base substitution. MEGA 4 was also used to produce the neighbour-joining tree and to perform bootstrap analysis (100 replicates). Sequences, specimen photographs and associated data are available at the Barcode of Life Data Systems web site (www.barcodinglife.org). Sequences are also available at GenBank (accession numbers HQ003941 to HQ005268).

3. RESULTS

(a) Identification success based on DNA barcodes

Our dataset consisted of 1387 samples representing 180 species, belonging to six families (Hesperiidae, Papilionidae, Pieridae, Lycaenidae, Riodinidae and Nymphalidae).

On average, 7.7 specimens were analysed per species. Only five species (2.8%) (Allancastria cerisyi, Polygonum amandus, Nymphalis l-album, Limenitis reducta and Hipparchia volgensis) were represented by one specimen and most (81.7%) had five or more records. Genetic distance to the nearest-neighbouring taxon varied from 0 to 11.1%, with an average of 4.7 per cent. Fifteen species pairs (16.7%) displayed overlap between their maximum intraspecific variation and the minimum interspecific divergence to another taxon (figure 1a).

The neighbour joining (NJ) tree profile showed that sequence records for 162 of the 180 species formed distinct barcode clusters allowing their unambiguous identification. The other 18 species consisted of four species pairs (4.5%) that formed paraphyletic clusters, two species pairs (2.2%) that were polyphyletic and three species pairs (3.3%) that shared barcodes (figure 1b). These results are unlikely to shift in any dramatic way with further sampling as four of five species represented by just one specimen displayed high sequence divergence to their nearest-neighbour (3.8%, 6.6%, 6.9% and 7%), indicating that COI allows for their reliable identification. The remaining taxon, H. volgensis, represented one of the three cases of barcode sharing (with Hipparchia semela). Therefore, using the criterion of barcode clusters, identification success is 90 per cent (the full NJ tree with bootstrap supports is available in the electronic supplementary material, annex 2).

(b) DNA barcodes separate some sibling taxa

Our results show that DNA barcoding performs well in discriminating between most Romanian butterfly species. This resolution extends to several species that, apart from DNA barcoding, can often be reliably identified only by genitalic examination (e.g. Melitaea athalia, Melitaea aurelia and Melitaea brutonartis or Leptidea sinapis and Leptidea reali (figure 2, see the electronic supplementary material, annex 3).

In fact, DNA barcodes distinguish several very similar taxa that are often impossible to identify based on the morphology of the adult even with genitalic examination. Such cases include Aricia agestis and Aricia artaxes (data on collection locality is needed, but not always sufficient) or Colias hyale and Colias alfacariensis (the larval stage is necessary for reliable identification; see the electronic supplementary material, annex 3).

(c) Cases of DNA barcode sharing

Members of three species pairs showed cases of barcode sharing (see the electronic supplementary material, annex 4). One of these cases involves a species pair...
(H. semele—H. volgensis) with unclear taxonomic status. Although H. volgensis is considered as a good species by some authors [37], detailed studies of male genitalia [38] yielded inconclusive results. The other two cases (Pieris napi—Pieris bryoniae, Colias crocea—Colias erate) involve pairs of species where only typical specimens can be reliably distinguished morphologically. In any reasonably large sample, specimens display considerable intraspecific variability often overlapping with their sister species. Members of both these species pairs are known to frequently hybridize [39–41] and the taxonomic status of P. bryoniae remains controversial.

(d) **Cases of DNA barcode paraphyly**

Four closely related species pairs displayed paraphyly and only two of these pairs can be reliably separated by examination of wing pattern (Apatura ilia—Apatura metis and Coenonympha tullia—Coenonympha rhodopensis). Two other pairs require examination of the genitalia (Hipparchia fagi—H. syriaca, Carcharodus flocciferus—Carcharodus orientalis). In this context, the cases of paraphyly involving taxa with very similar external and internal morphology have been thoroughly analysed in order to test the relationship between DNA barcodes and morphology-based identifications. For example, in the case of C. flocciferus and C. orientalis linear and geometric morphometry of the male genitalia were necessary in order to test the identification success of DNA barcoding (figure 3; for additional information, see the electronic supplementary material, annex 4).

(e) **Cases of DNA barcode polyphyly**

The species pair Polyommatus bellargus—Polyommatus coridon is one of the two cases of polyphyly present in
flies represent 4.4 per cent of the entire dataset of 180 that may represent cryptic species (table 1). However, the bootstrap values more than 50% indicated.

By applying this threshold, we found eight cases of lineages of the taxa involved [42,43]. Such analyses are compulsory because divergent sequences might correspond to actually represent cryptic taxa because the sequences displayed intraspecific divergence of at least 2%.

Cases of deep intraspecific divergence in DNA barcodes

Our studies of barcode divergences in a third of the European butterfly fauna allowed an assessment of potentially cryptic species within one of the best-studied invertebrate groups and regions.

We provide an example considering that a lineage may represent a cryptic species if a sequence or group of sequences displayed intraspecific divergence of at least 2 per cent. This threshold has been repeatedly suggested for various animal groups including Lepidoptera [2,3,22,34]. By applying this threshold, we found eight cases of lineages that may represent cryptic species (table 1). However, the clarification of such cases requires deeper integrative approaches including morphology, biology and nuclear markers of the taxa involved [42,43]. Such analyses are compulsory because divergent sequences might correspond to fully inter fertile lineages of the same species reflecting ancestral polymorphisms or diversity gained by introgression.

The dangers of relying solely on mtDNA data to define species have been proved through the use of amplified fragment length polymorphism markers in the case of Lepidoptera from the genus Mechanitis (Nymphalidae) [33].

The eight cases of deep divergence in Romanian butterflies represent 4.4 per cent of the entire dataset of 180 species (table 1). However, several of these are unlikely to actually represent cryptic taxa because the sequences responsible for the deep divergence are either very similar to those of a closely related taxon, suggesting introgression or incomplete lineage sorting (e.g. P. napi, P. bellargus), or represent extreme haplotypes of a rather gradual intraspecific continuum (e.g. M. aurelia). Preliminary examination of the external (wing pattern) and internal morphology (male genitalia) of the specimens involved failed to reveal obvious differences. It is worth however mentioning that five of the cases display sympathy (figure 4a–d), which would facilitate testing their specific status through the genotypic cluster concept [44].

On the other hand, without applying a fixed threshold, we found that geographically correlated intraspecific

---

**Figure 3. Carcharodus flocciferus** is paraphyletic with respect to C. orientalis. The two species are morphologically very similar and accurate identifications were possible based on the combined results of geometric and linear morphometry of the male genitalia. (a) Scatter plot of relative warp 1 and relative warp 2 explaining most of the variance reveals the presence of two discrete clusters of cuiller tip shape. The upper left image illustrates the landmarks (open circles) and semi-landmarks (black dots) used to describe the shape of the cuiller. (b) NJ tree of COI barcodes of Romanian Carcharodus orientalis and C. flocciferus with bootstrap values more than 50% indicated.

**Table 1.** Taxa containing lineages that display intraspecific divergences of at least 2%.

<table>
<thead>
<tr>
<th>taxon</th>
<th>number of samples barcoded</th>
<th>max. intraspecific distance (%)</th>
<th>sympathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasiomnata maera</td>
<td>12</td>
<td>2.0</td>
<td>no</td>
</tr>
<tr>
<td>Melitaea aurelia</td>
<td>11</td>
<td>2.2</td>
<td>no</td>
</tr>
<tr>
<td>Pieris napi</td>
<td>19</td>
<td>2.3</td>
<td>yes</td>
</tr>
<tr>
<td>Polyommatus bellargus</td>
<td>10</td>
<td>2.3</td>
<td>yes</td>
</tr>
<tr>
<td>Thymelicus sylvestris</td>
<td>17</td>
<td>2.7</td>
<td>yes</td>
</tr>
<tr>
<td>Pyrgus armoricanus</td>
<td>23</td>
<td>2.8</td>
<td>yes</td>
</tr>
<tr>
<td>Hipparchia senele</td>
<td>10</td>
<td>3.5</td>
<td>yes</td>
</tr>
<tr>
<td>Boloria euphyoseyne</td>
<td>8</td>
<td>5.0</td>
<td>no</td>
</tr>
</tbody>
</table>
divergence was present for several Romanian butterfly species. This may not necessarily represent potential cryptic species, but also subspecific taxonomic units of conservation importance. These often involved species that are protected at national or European scale (Pyrgus sidae, Maculinea nausithous, Euphydryas maturna, Erebia sudetica). An interesting example is that of the European endangered M. nausithous, the larvae of which are obligate social parasites of Myrmica ant nests after developing on Sanguisorba officinalis L. [45, 46]. In this case the divergence in DNA barcodes (0.46%) seems to be not only geographically, but also biologically correlated (figure 5). Maculinea nausithous is one of the rarest Romanian butterflies with only two small groups of populations known. These populations (lying in Transylvania, respectively, northern Moldavia) are separated by ca 200 km. It has been recently shown that the M. nausithous populations from Transylvania use Myrmica scabrinodis as a host ant, while the Moldavian populations use only Myrmica rubra [47]. Although the level of divergence is low, such cases may also require deeper studies especially given implications for biodiversity conservation and nature management.

(g) DNA barcoding versus morphology

As DNA barcoding relies on a single gene fragment to identify species, we compared its effectiveness with the two most commonly used approaches for the identification of butterflies: wing morphology and male genitalia (table 2, see the electronic supplementary material, annex 5). For DNA barcoding we used two categories to quantify identification success rates: 'no' means all specimens of a given species can be reliably identified when a particular feature is analysed, 'yes' means that the species is recovered as monophyletic, even if only one specimen is the cause, and 'yes' means that the species is recovered as paraphyletic or polyphyletic, even if only one specimen is the cause, and 'no' means a substantial proportion (i.e. more than 10%) of the specimens for a given species cannot be reliably identified through the targeted feature and 'not for certain
We found that, for the Romanian butterflies, DNA barcoding is more informative than wing morphology (6.1% and 7.8% higher identification success rate for males and females, respectively) and slightly better than the male genitalia (1.1% higher identification success rate). This difference is mainly owing to the fact that morphological characters, especially wing pattern are subject to considerable intraspecific variability that causes overlap in the phenotypes of closely related taxa. Identification based on wing morphology is usually the fastest and most accessible, but it requires taxonomic experience. The lowest success in wing morphology-based identification was found in females, reflecting the fact that females of closely related species are often particularly difficult to distinguish (e.g. *Cupido alcetas*—*C. decolorata*, *Pyrgus alveus*—*P. armoricana*). By contrast, and similarly to birds, secondary sexual characters in males are frequent in Lepidoptera and provide additional elements for species identification. Male genitalia performed well but this approach is less accessible than wing pattern examination and it requires considerable taxonomic expertise to interpret. Moreover, even after combining information on wing patterns and genitalia for male specimens, the identification success only increases to 95.6%. By adding data from DNA barcoding, performance is increased to 97.8 per cent and the only problematic cases that remain are the two species pairs suspected of regular hybridization (*P. napi*—*P. bryoniae* and *Colias crocea*—*C. erate*). However, this example proves that integration of data from multiple sources considerably improves identification success, especially for the most difficult taxa. It is also worth mentioning that many species with difficult morphology-based identification proved to be also problematic for DNA barcoding (2 × 2 χ²-test, p < 0.01). This suggests that, in most closely related species pairs, limited differentiation occurs both in phenotype and genotype.

**4. DISCUSSION**

Our study of the butterflies of Romania showed that DNA barcoding provides certain identification for 90 per cent of the species in this region. The remaining 10 per cent involve cases of paraphyly, polyphyly or shared barcodes between closely related species pairs. This success rate is among the highest reported for butterflies, especially because, by contrast to other studies [24], it does not include cases of paraphyly. Our results can be considered a reliable estimate for temperate Europe owing to comprehensive sampling and detailed morphology-based comparisons. A small number of the species in our study (6.7%), possessed barcode sequences that showed paraphyly or polyphyly. As these patterns of sequence variation can be produced by several types of speciation [48] or other processes such as introgression, they are not rare in nature [49,50]. Moreover, detailed analysis of such cases can provide a better understanding of the
evolutionary history of the species involved. Cases of mtDNA paralogy caused by introgression have been previously reported in Lepidoptera [51], but they can also reflect incomplete lineage sorting in recent speciation events (all cases encountered by us involve recently diverged species pairs). We emphasize that cases of paralogy and polyphyly do not prevent the identification of species unless they share haplotypes. For example, cases of paralogy in Central Asian butterflies were treated as identification successes because the species involved were never found to share haplotypes [24]. The same pattern was observed in our study—all six cases (four paralogy, two polyphyly) displayed very short branches, reflecting low levels of minimum interspecific distances (between 0.15% and 0.58%). Based on our current data, all haplotypes are species-specific, so that specimens could be attributed to the correct taxon. Including these cases, the DNA barcoding identification success via barcoding for Romanian butterflies rises to 96.7% per cent. However, given the small interspecific distances involved, further sampling is needed in order to validate the robustness of this conclusion. Such cases also highlight the importance of comprehensive sampling (across different populations and geographic regions) without which several species pairs in our dataset may have appeared as reciprocally monophyletic, leading to misinterpretations of DNA barcoding performance.

We also emphasize that all three species pairs sharing DNA barcodes are very closely related and that two are known to hybridize regularly (C. crocea—C. erate and P. napi—P. bryoniae; [39–41]). In fact, assigning species identification to hybrids is intrinsically erroneous. Some methods may be able to identify hybrids as such (for example, multilocus markers or morphology) but it is not possible to identify them exclusively through DNA barcoding owing to the lack of recombination of this marker. One taxon involved in barcode sharing (H. volgensis) needs research to clarify its taxonomic status, and exemplifies the effect that an unresolved taxonomy has for the assessment of DNA barcoding performance, a much more severe problem in other groups of organisms. Our results apply exclusively to the Romanian butterflies and it is probable that the overall identification success of DNA barcoding, as well as that of morphology-based methods, would slightly decrease if samples from a broader geographical area were included. This may be particularly the case of species pairs that already display small interspecific divergence in Romania such as Plebejus idas and Plebejus argyrognomon. Europe is generally rather depauperate, and many coexisting European species are relatively distant members of much larger genera often distributed across Eurasia and in some cases North America. Thus, the addition of more species of these genera to the study could affect the performance of all the methods tested.

Our study has shown that DNA barcoding is more effective in identifying the butterflies of Romania than the morphological characters (wing pattern and male genitalia) that are ordinarily employed for identification. Owing to intraspecific variability, identification based on morphological characters often has subjectivity that can generate errors when morphometry is not used and when the discriminating characters are subtle. Such complexities occur in many closely related species owing to overlap between intra and interspecific phenotypic variability. Identification based on wing morphology alone becomes even more difficult if the specimens involved are worn so that diagnostic characters are not clear. Morphology-based identifications are even more difficult for preimaginal stages, especially eggs and pupae which often lack reliable diagnostic characters and are, in many cases, difficult or impossible to identify if not linked to adult the stage. By contrast, DNA barcoding has the same success rate for the identification of all life stages (e.g. [52–54]) thus representing a very promising approach in standardized faunal surveys. The method also proved capable of highlighting cryptic biodiversity (of both specific and infraspecific level) that would have passed unnoticed based on morphological characters alone. In recent years, the number of reported cryptic species has grown considerably in great part owing to an increasing number of studies incorporating DNA-based techniques. A good estimate of cryptic species diversity is of major importance for many aspects of biology and conservation [55]. Recent opinions on cryptic diversity are rather contradictory ranging from the hypothesis of a non-random distribution across taxonomic groups and biomes [55] to homogeneity [56]. In this context, there is great need for a tool that would facilitate the discovery of cryptic biodiversity and allow for a correct evaluation of the problem. DNA barcoding has the potential to highlight lineages that could represent distinct species [18–21,23]. Results provided by DNA barcoding suggested, for example, unexpectedly high levels of cryptic diversity within tropical Lepidoptera, parasitoid flies or parasitoid wasps [15,18,20,21].

Trying to identify potential cryptic species using more or less diverged lineages is limited by several factors that produce variable levels of divergence in mtDNA between taxa of the same rank [28,29,57]. This means that the use of any particular threshold [2,3,15,34] will overlook young taxa. However, thresholds do provide a quick indication of diverged lineages that are candidates for cryptic species. By employing a 2 per cent threshold, we found eight cases that are worth deeper morphological, ecological and molecular studies. Geographically and/or biologically correlated population differentiation was also noted in several cases, some of which involved endangered taxa protected at national and/or European level. Similar cases have also been reported by other studies (e.g. [25]) and indicate that DNA barcoding may serve as a complementary tool for conservation-oriented efforts, given the fact that it could facilitate comparative studies of genetic diversity and help to delineate subspecific taxonomic units of conservation importance [38].

This study has developed a comprehensive DNA barcode reference database for the butterflies of Romania. As a result, any butterfly from this country can be identified through DNA barcoding to a species or, in few cases, to a species-pair, regardless of life stage or specimen quality and without requiring any taxonomic knowledge. We point out the importance of a solid taxonomic framework for the DNA barcode library and the advantages of a region-oriented DNA barcoding strategy which accelerates the applicability of the method by providing non-specialists with a reliable identification method.
We thank C. Corduneanu, S. Cuvelier, M. Goia, A. Heres, J. Hernández-Roldán, Z. Kovács, Z. Kovács, S. Montagud, S. Mihut, L. Székely, C. Stefanescu, J. Viader and S. Viader for their help in collecting samples sequenced for this study. We are grateful to L. Dappporto for advice on geometric morphometry analyses. Support for this research was provided by the Ministerio de Ciencia e Innovación project (CGL2007-60 516/BOS) to R.V. and V.D., and a predoctoral fellowship from Universitat Autonoma de Barcelona to V.D. Support for DNA sequence analysis and bioinformatics was provided through grants to P.D.N.H. from NSERC, from the Ontario Ministry of Research and Innovation and from Genome Canada through the Ontario Genomics Institute.

REFERENCES


54 Dincă, V. & Vila, R. 2008 Improving the knowledge on Romanian Rhopalocera, including the rediscovery of *Polyommatus amandus* (Schneider, 1792) (Lycaenidae) and an application of DNA-based identification. *Nota lepid.* 31, 3–23.


