A DNA barcode library for the butterflies of North America

Jacopo D'Ercole[1,2], Vlad Dincă[3], Paul A. Opler[4], Norbert Kondla[5], Christian Schmidt[6], Jarrett D. Phillips[2,7], Robert Robbins[8], John M. Burns[8], Scott E. Miller[8], Nick Grishin[9,10], Evgeny V. Zakharov[2], Jeremy R. DeWaard[2], Sujeevan Ratnasingham[2] and Paul D.N. Hebert[1,2]

[1] Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada
[2] Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada
[3] Ecology and Genetics Research Unit, University of Oulu, Oulu, Finland
[4] Colorado State University, Fort Collins, CO, United States of America
[5] Unaffiliated, Calgary, Alberta, Canada
[6] Canadian National Collection of Insects, Arachnids and Nematodes, Agriculture and Agri-Food, Guelph, Ontario, Canada
[7] School of Computer Science, University of Guelph, Guelph, Ontario, Canada
[8] Department of Entomology, Smithsonian Institution, Washington DC, United States of America
[9] Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, United States of America
[10] Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, United States of America

ABSTRACT

Although the butterflies of North America have received considerable taxonomic attention, overlooked species and instances of hybridization continue to be revealed. The present study assembles a DNA barcode reference library for this fauna to identify groups whose patterns of sequence variation suggest the need for further taxonomic study. Based on 14,626 records from 814 species, DNA barcodes were obtained for 96% of the fauna. The maximum intraspecific distance averaged 1/4 the minimum distance to the nearest neighbor, producing a barcode gap in 76% of the species. Most species (80%) were monophyletic, the others were para- or polyphyletic. Although 15% of currently recognized species shared barcodes, the incidence of such taxa was far higher in regions exposed to Pleistocene glaciations than in those that were ice-free. Nearly 10% of species displayed high intraspecific variation (>2.5%), suggesting the need for further investigation to assess potential cryptic diversity. Aside from aiding the identification of all life stages of North American butterflies, the reference library has provided new perspectives on the incidence of both cryptic and potentially over-split species, setting the stage for future studies that can further explore the evolutionary dynamics of this group.

Subjects Biodiversity, Conservation Biology, Entomology, Taxonomy, Zoology
Keywords DNA barcoding, CO1, Barcode library, Butterflies, North America, Quaternary glaciations

Submitted 7 May 2020
Accepted 4 March 2021
Published 19 April 2021

Corresponding author
Jacopo D'Ercole,
jdercole@uoguelph.ca

Academic editor
Ilaria Negri

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.11157**

INTRODUCTION

DNA barcoding is an effective tool for addressing the widely recognized need for an improved understanding of biodiversity. By employing sequence diversity in short,

How to cite this article D'Ercole J, Dincă V, Opler PA, Kondla N, Schmidt C, Phillips JD, Robbins R, Burns JM, Miller SE, Grishin N, Zakharov EV, DeWaard JR, Ratnasingham S, Hebert PDN. 2021. A DNA barcode library for the butterflies of North America. *PeerJ* **9**:e11157
http://doi.org/10.7717/peerj.11157

standardized gene regions, such as the 648 base pair segment of the 5′ region of mitochondrial cytochrome *c* oxidase 1 (CO1) employed for the animal kingdom (*Hebert et al., 2003*), DNA barcoding allows both the identification of specimens and the discovery of new species. Since its introduction, this marker has been adopted in fields ranging from population genetics (*Hajibabaei et al., 2007*), phylogenetics (*Hajibabaei et al., 2007*), and phylogeography (*Dapporto et al., 2019*) to ecology (*Valentini, Pompanon & Taberlet, 2009*) and conservation (*Dincă et al., 2018*). It has also gained application in contexts ranging from the detection of marketplace fraud (*Galimberti et al., 2013*) to the suppression of illegal trade in endangered species (*Rehman et al., 2015*).

DNA barcoding enables the identification of specimens without morphological analysis by querying their CO1 sequences against a reference library of DNA barcodes obtained from carefully identified vouchers. These reference sequences are curated in the Barcode of Life Datasystem (BOLD) (boldsystems.org) (*Ratnasingham & Hebert, 2007*), an informatics platform that also hosts collateral data such as specimen images and collection data. Aside from identifying specimens, DNA barcoding can help to delineate species boundaries (*Cariou et al., 2020*), an important task since species play a central role in biodiversity assessments and conservation actions. Rather than simply presuming that the current taxonomic system is valid, DNA barcoding provides a basis for testing this assertion. Prior studies have shown that closely allied congeneric species of Lepidoptera typically show more than 2% divergence (*Hebert et al., 2003*; *Huemer et al., 2014*; *Dincă et al., 2015*). Although some sister species do show lower divergence (*Burns et al., 2007*; *Cong et al., 2016*), cases where species share barcode sequences can reflect over-splitting (*Vila et al., 2010*) or introgressive hybridization (*Zakharov et al., 2009*; *Cong et al., 2017*). Conversely, when members of a putative species show high sequence divergence, this often signals the presence of overlooked species (*Hebert et al., 2004*; *Burns et al., 2007*; *Dincă et al., 2013*). DNA barcode data has also enabled the development of algorithms that employ sequence information for species delimitation. The latter methods cluster specimens into Molecular Operational Taxonomic Units (MOTUs) that have been shown to correspond closely with recognized species in groups with well-established taxonomy (*Ratnasingham & Hebert, 2013*).

Most studies have tested the capacity of DNA barcodes to discriminate species when viewed from a local or regional context, and only a few have examined resolution at a continental scale (e.g., *Kerr et al., 2007*; *Lukhtanov et al., 2009*; *Bergsten et al., 2012*; *Huemer et al., 2014*; *Zahiri et al., 2017*; *Dincă et al., 2021*). The latter studies are important because they can reveal cases of low interspecific divergences, potentially reducing the effectiveness of DNA barcoding for species delimitation. Prior studies have shown the general effectiveness of DNA barcoding for butterflies (*Lukhtanov et al., 2009*; *Dincă et al., 2011*; *Dincă et al., 2015*; *Lavinia et al., 2017*; *Dincă et al., 2021*), but have also exposed discordances with current taxonomy including probable cases of synonymy (e.g., *Vila et al., 2010*) and frequent instances of overlooked species (e.g., *Hebert et al., 2004*; *Burns et al., 2008*).

The butterfly fauna of North America has seen more intensive morphological study than any other comparably diverse insect lineage on this continent (*Warren et al., 2012*). Despite

this attention, there remains uncertainty in the status of many taxa, often reflecting the subjectivity inherent in decisions on species boundaries based on morphology alone. The present study assembles a comprehensive DNA barcode library for the butterfly fauna of North America, delivering an identification system for most of these species while testing the current taxonomy. This work also provides an overview of patterns of genetic diversity and offers insights on mechanisms responsible for shaping the genetic diversity of the butterflies of this continent.

METHODS

Sampling

This study sought to recover DNA barcodes for the butterfly fauna of Canada and USA. BOLD hosts a checklist for 846 species (CL-NABUT) derived from Pelham's list (*Warren et al., 2012*) with a few changes based on recent publications. Table S1 provides a condensed version of this checklist for the 648 species with persistent populations in North America, excluding those introduced by humans.

The sampling program aimed to capture geographic and phylogenetic diversity for each species in continental North America (i.e., islands beyond the continental shelf were excluded). Specimens from Canada (6,935) and USA (7,037) were sequenced when possible, but this left some gaps which were filled by analyzing specimens from Central America (602), South America (27), Europe (7), Asia (4) and unvouchered (11) records from GenBank (Fig. 1). Overall 14,626 specimens were analyzed, and associated metadata are available on BOLD (v4.boldsystems.org) in the public dataset "DS-USCANLEP" (dx.doi.org/10.5883/DS-USCANLEP). From this total, 10,425 vouchers are held in public natural history collections, 3,864 in private collections, 259 derive from GenBank, and 78 were unvouchered. Permission from all the institutional and private collections was obtained to access and study the records. The Centre for Biodiversity Genomics made the largest contribution (4,474 records), followed by the Canadian National Collection (1,771) and the Smithsonian's National Museum of Natural History (1,010).

iNEXT (INterpolation and EXTrapolation) (*Hsieh, Ma & Chao, 2016*), was employed to estimate sampling completeness using R (*R Studio Team, 2016*). This approach implements the Chao1 diversity estimator (*Chao, 1984*) to generate accumulation curves that can be used to estimate the total haplotype diversity in a species. This asymptotic value was compared with the observed haplotype diversity to quantify sampling completeness, an approach used to estimate coverage for European butterflies (*Dincă et al., 2021*). Because the present study targeted species resident in North America, levels of genetic diversity for introduced species and tropical strays (for which sampling was limited) were likely to be underestimated. As a result, estimates of sampling completeness excluded species whose distribution primarily falls outside North America. Moreover, species represented by fewer than six specimens were also excluded, reducing consideration to 402 of the 648 species (Table S1). For each species in the barcode library, we recorded the number of specimens (N), the number of observed haplotypes (H), the fraction of haplotype diversity retrieved (R), and the number of additional haplotypes which remain to be sampled (L).

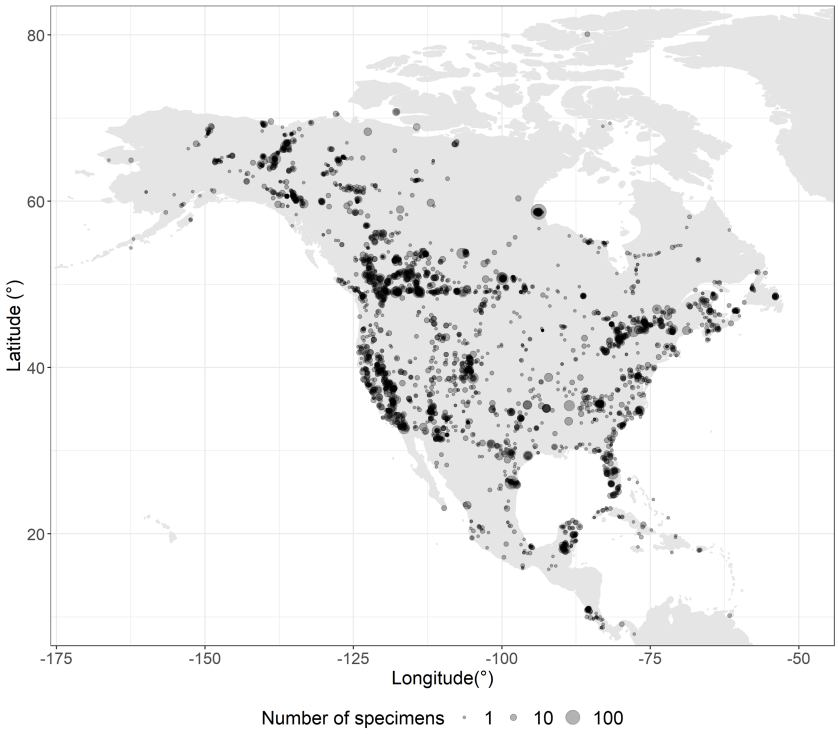

Figure 1 **Sampling coverage.** Overlapping sampling points are shown as darker circles. Twenty-eight records are not shown in the map as they derived from Argentina, Brazil, China, India, Italy, and Peru.
Full-size DOI: 10.7717/peerj.11157/fig-1

Mitochondrial CO1 characterization and quality control

DNA extraction, PCR, and sequencing followed standard protocols at the Canadian Centre for DNA Barcoding (CCDB). DNA was extracted using a silica-based method in 96-well plate format (*Ivanova, DeWaard & Hebert, 2006*). PCR volumes and thermal cycling conditions followed *DeWaard et al. (2008)* or *Hebert et al. (2013)* in the case of older museum specimens. Trace files were assembled into contigs with CodonCode Aligner (CodonCode Corporation, http://www.codoncode.com) to generate a sequence record for each specimen. DNA extracts from museum specimens often contain low concentrations of degraded DNA. Although Sanger sequencing can usually recover DNA barcodes from specimens less than 50 years old (*Hausmann et al., 2009*; *Lees et al., 2011*; *Hebert et al., 2013*), sequence length is often <200 bp (*Allentoft et al., 2012*). Some specimens that failed to generate a Sanger sequence were processed with a next-generation sequencing protocol (*D'Ercole, Prosser & Hebert, 2021*).

All except five of the 14,626 sequence records included at least 500 unambiguous base pairs of CO1. All sequences were examined for stop codons and the few containing them were removed as likely NUMTS.

Although most specimens were identified by taxonomic specialists through analysis of morphological (i.e., external characters and genitalia) and ecological traits, prior studies have revealed that misidentifications are inevitable in any large-scale study (*Mutanen et al., 2016*). As a result, Neighbor-Joining (NJ) trees (one for each family) including all records

were examined to detect cases where two or more species were admixed in a sequence cluster, a pattern that often arises as a result of misidentification or contamination. The external morphology of specimens in such clusters was examined to confirm their identity, an approach that revealed some errors which were corrected prior to further analyses. Because a detailed inspection (i.e., nuclear markers or genitalic dissections) of all specimens was not possible and because of taxonomic uncertainty in some groups, additional cases of misidentification may remain in our dataset. The NJ tree was also employed to aid the identification of unnamed specimens.

Genetic analysis

DNA sequences were aligned on BOLD by employing a Hidden Markov Model (*Eddy, 1998*) based on amino acid sequences. Intraspecific and interspecific genetic distances were calculated using BOLD employing the Kimura two parameter (K2P) distance model (*Kimura, 1980*). The barcode gap was examined by plotting maximum intraspecific distance for each species against the distance to its nearest neighbor (i.e., minimum interspecific distance). Intraspecific distances and barcode gap analysis could only be calculated for the 755 species represented by two or more specimens.

Bayesian phylogenies (one for each butterfly family) were employed to assess the number of species displaying monophyly. BEAST2 (*Bouckaert et al., 2014*) was used to generate the trees. The selection of the best model of molecular evolution for phylogenetic investigation was performed with JModeltest2 (*Guindon & Gascuel, 2003*; *Darriba et al., 2012*). The GTR model of molecular evolution (*Tavaré, 1986*), along with the gamma function discretized in four categories, and a parameter for the proportion of invariable sites, were used to estimate genetic distances. Each branch was assumed to evolve at the same rate, accumulating divergence at 1.5% per million years (*Quek et al., 2004*), following a strict molecular clock. The Markov chain Monte Carlo (MCMC) chain length was 10,000,000 with log frequency every 1,000 samples of the posterior distribution. The pre-burnin was set at 1,000. TreeAnnotator (*Bouckaert et al., 2014*) was employed to combine the Bayesian trees sampled from the posterior distribution while convergence was confirmed with Tracer (*Rambaut et al., 2018*). This analysis was restricted to the 755 species represented by two or more specimens as only they could satisfy the definition of monophyly (*Hennig, 1966*).

The presence of both potentially overlooked and over-split species was tested with three approaches. The first and simplest approach employed a fixed divergence value of 2.5% to discriminate intraspecific from interspecific diversity. Although the application of fixed thresholds is controversial (*Collins & Cruickshank, 2013*), it provides a useful point of reference for comparison with other studies. The other two methods were the General Mixed Yule Coalescent (GMYC) (*Pons et al., 2006*; *Fujisawa & Barraclough, 2013*) and the Barcode Index Number System (BIN) (*Ratnasingham & Hebert, 2013*). These methods for species delimitation have been shown to perform well in recovering species counts congruent with taxon boundaries established through morphological studies (*Ratnasingham & Hebert, 2013*). While GMYC typically generates more MOTUs than morphospecies (*Miralles & Vences, 2013*; *Kekkonen & Hebert, 2014*), the BIN system was designed to provide a conservative estimate of the number of species (*Ratnasingham &*

Hebert, 2013). The likelihood-based GMYC model makes use of the Bayesian ultrametric trees to determine the transition between intra- and interspecific branching patterns. The R package "splits" (*Ezard, Fujisawa & Barraclough, 2009*) was employed for the GMYC analysis. A few specimens that were only identified to a generic level were excluded from analysis, and subspecies designations were stripped from specimens that possessed them. The dataset was then collapsed to retain only unique haplotypes (5,116) as past studies showed this approach optimizes results (*Talavera, Dincă & Vila, 2013*; *Tang et al., 2014*). BOLD was employed to assign each sequence to a BIN (*Ratnasingham & Hebert, 2013*). GMYC and BIN assessments generated results falling into four categories: Match, Merge, Split, and Mixture. A species was assigned to the Match category when all of its specimens were assigned to one MOTU. In cases where two or more species shared the same MOTU, they belonged to the Merge category. A species was placed in the Split category when its component specimens were assigned to two or more MOTUs. Finally, a species characterized by a more complex pattern, including both Match and Split, was assigned to the Mixture category.

Barcode sharing

Barcode sharing describes the situation where individuals of two or more species share identical DNA sequences. As opposed to this, following the character-based definition outlined by *DeSalle, Egan & Siddall (2005)*, species with diagnostic barcodes are those whose sequences show consistent nucleotide differences (or a combination of nucleotide differences) from any other species. As a result, DNA barcodes can be diagnostic even when they derive from species with such low divergence that they are assigned to a single MOTU. To better reflect the differing exposure of species to biogeographic shifts during the Pleistocene, species involved in barcode sharing were partitioned into three categories. The first category (North/alpine hereafter) included species with a geographic distribution north of the last glacial maximum (LGM) and alpine/subalpine species on mountains south of the LGM. The second category (Mid-latitude hereafter) included species with a distribution extending across the LGM. The third category (South hereafter) was composed of species located south of the LGM. The assignment of each species to one of these categories was based on its current distribution (*Scott, 1986*; *Brock & Kaufman, 2006*). Because the probability of detecting barcode sharing is influenced by sampling intensity, iNEXT was used to evaluate sampling completeness by region (North/alpine, Mid-latitude, South).

The Spearman's correlation coefficient was employed to assess the association between the number of species with barcode sharing in a genus and the total number of species in that genus.

RESULTS

Sampling and DNA barcoding performance

The present dataset provides barcode coverage for 96.2% (i.e., 814 of 846) of North American butterfly species with an average of 18 sequences per species (Fig. 1, Figs. S1–S6, dx.doi.org/10.5883/DS-USCANLEP). However, 59 species were represented by singletons, including 34 of the 648 species on the truncated list (Table S1). The coverage rises to

97.2% (630 of 648) when only species with permanent populations in North America are considered (Table S1). Estimates of sampling completeness were performed with iNEXT on 63.8% (402 of 630) of the species in the truncated list (Table S1). This analysis, which considered 12,860 specimens, indicated that their 3,212 unique haplotypes corresponded to 67% of the haplotype diversity in this subset of North American butterflies (Table S1). In order to raise haplotype recovery to 100%, it was estimated that at least another 4,702 haplotypes would need to be recovered, an average of 12 haplotypes per species (Table S1).

Genetic diversity

Maximum intraspecific distances averaged 0.97% (range 0–8.4%) while the nearest neighbor distance averaged 3.7% (range 0–14.3%), almost 4-fold higher than the maximum intraspecific distance (Fig. 2, Table S2). A barcode gap was present in 573 of 755 species (75.9%) represented by at least two individuals (Fig. 2, Table S2).

The Bayesian trees revealed that 604 of the 755 species (80%) represented by two or more individuals formed monophyletic groups, while the other 151 (20%) were either paraphyletic or polyphyletic (Figs. S7–S12, S13 and Table S2). Species with just a single barcode sequence were necessarily excluded from this analysis, but 55 of the 59 possessed barcodes distinct from their nearest neighbor. While discrimination between paraphyly and polyphyly is not essential for specimen identification, it is critical to distinguish those species characterized by overlapping phylogenetic branches from those species sharing barcodes with their nearest neighbor(s). Twenty-six species (3.4%) fell in the first category and 125 (16.6%) species in the latter (Fig. S13).

Use of a fixed distance threshold (2.5%) exposed 79 cases (9.7%) where intraspecific distance was above the set threshold and 324 (39.8%) cases where interspecific divergence was below it. MOTU delineation revealed a variable number of entities depending on the method employed. BIN analysis was performed on all but one species (the sole sequence for *Calephelis rawsoni* did not qualify for analysis) and revealed 772 BINs, comprised of 540 Matches (66.4%), 55 Splits (6.8%), 181 Merges (22.3%), and 37 Mixtures (4.6%). By comparison, GMYC analysis generated 862 taxonomic units, partitioned in 527 Matches (64.7%), 63 Splits (7.8%), 150 Merges (18.4%), and 74 Mixtures (9.1%) (Figs. S7–S12). Overall, the three analyses provided concordant support for 369 species recognized by the current taxonomy, but they also revealed 24 species split into two or more units, 124 grouped with one or more nearest neighbor(s), and 34 in both previous categories. The performance of the three methods is compared in Table S2.

Barcode sharing

In total, 125 of the 814 (15%) species shared their barcode with another species (Table S2). The incidence of barcode sharing on the condensed list (Table S1) was highest at 42.2% (38/90) for the Northern/alpine species, dropped to 25.6% (55/215) for the Mid-latitude species, and was just 9.2% (30/325) for the Southern species.

iNEXT (Table S1) revealed that the analysis of an average of 41 specimens/species in Northern/alpine species detected an average of 8 unique haplotypes/species, corresponding to 66% of the estimated haplotype diversity for these taxa. By comparison, the analysis of

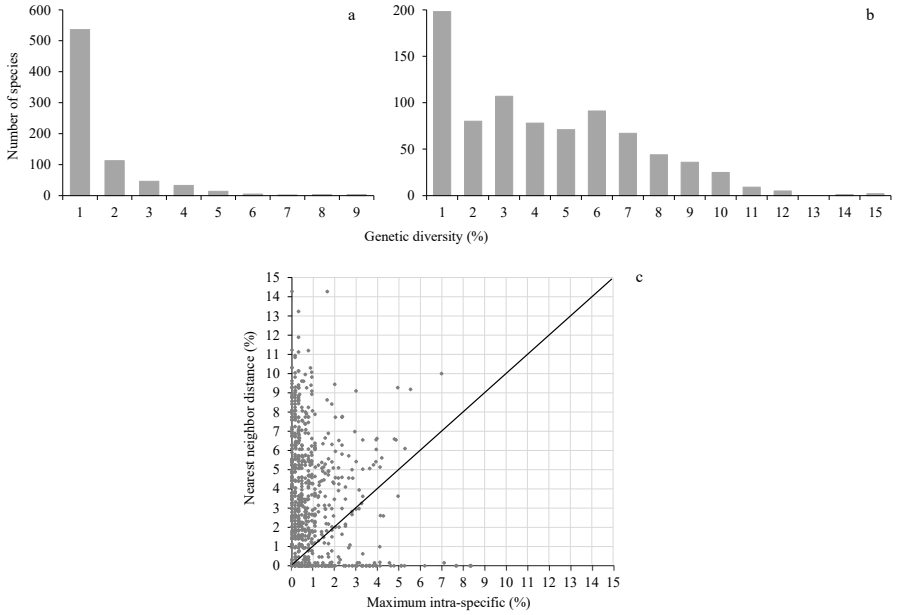

Figure 2 **Genetic diversity.** The upper histograms show the maximum intraspecific distance (A) and the nearest neighbor distance (B); the lower scatterplot (C) shows for each species (dot) the presence (above the diagonal) or absence (below the diagonal) of a barcode gap.

Full-size DOI: 10.7717/peerj.11157/fig-2

an average of 39 specimens/species for the Mid-latitude species detected an average of 10 haplotypes/species and represented 64% of their estimated diversity. Finally, the analysis of 16 specimens for the Southern species revealed an average of 5 unique haplotypes, corresponding to 71% of their estimated haplotype diversity.

The incidence of barcode sharing varied among butterfly families. The Hesperiidae (21/205, 10.2%) and Riodinidae (2/18, 11.1%) were least impacted, followed by Nymphalidae (30/173, 17.3%), Papilionidae (7/31, 22.6%), Lycaenidae (38/137, 27.7%), and Pieridae (25/66, 37.9%) (Table 1, Table S1). Spearman's coefficient revealed a positive but non-significant correlation between the incidence of barcode sharing and the number of species in a genus ($R^2 = 0.37$, $p = 0.056$; Fig. S14).

DISCUSSION

This study generated a DNA barcode reference library for 96% of the North American butterfly fauna (814 of 846 species) with an average of 18 records per species. This level of sampling captured 67% of the estimated haplotype diversity, but at least 4,705 haplotypes await detection. The present library was effective in assigning newly encountered specimens to either a species or to a small number of closely allied species. Because butterflies are key bioindicators (*Syaripuddin, Sing & Wilson, 2015*) and umbrella species (*New, 1997*), this library facilitates their use in tracking the impacts of habitat loss, fragmentation, and climate change. Along with the reference barcodes available for other taxonomic groups (*Ratnasingham & Hebert, 2007*), the present work also provides a basis for exploring species interactions.

Table 1 Distribution of barcode sharing with respect to the LGM by family. Proportions are shown in brackets.

	Above LGM and alpine/subalpine	Across LGM	Below LGM	Total
Papilionidae	2/7 (28.6%)	3/11 (27.3%)	2/13 (15.4%)	7/31 (22.6%)
Pieridae	16/18 (88.9%)	8/15 (53.3%)	1/33 (3%)	25/66 (37.9%)
Hesperiidae	3/11 (27.3%)	10/62 (16.1%)	8/132 (6.1%)	21/205 (10.2%)
Lycaenidae	9/15 (60%)	14/54 (25.9%)	15/68 (22.1%)	38/137 (27.7%)
Riodinidae	0	1/3 (33.3%)	1/15 (6.7%)	2/18 (11.1%)
Nymphalidae	8/39 (20.5%)	19/70 (27.1%)	3/64 (4.7%)	30/173 (17.3%)
Total	38/90 (42.2%)	55/215 (24.9%)	30/325 (9.2%)	123/630 (19.5%)

The sequence records analyzed in this study were mainly obtained during two ways. The first approach involved the collection of fresh material and lasted about two decades (2000–2019), while the latter spanned three years (2015–2017) and consisted of the targeted analysis of specimens held in two major natural history collections—the Smithsonian's National Museum of Natural History and the Canadian National Collection. Previous studies on Lepidoptera, largely regional in scale, indicated that most species are separated from their nearest neighbor by a barcode gap (e.g., *Janzen et al., 2005*; *Hajibabaei et al., 2006*; *Lavinia et al., 2017*). Increased geographical coverage should lead to higher intraspecific genetic variation as a result of the isolation-by-distance effect (*Wright, 1943*), and reduced interspecific divergences as more species are analyzed (*Avise, 2000*). Both factors should reduce the difference between intra- and inter-specific divergence. It needs emphasis that exceptions to this model were observed in butterflies and that narrow suture zones with high diversity alternate with extended regions with little variation (e.g., *Gompert et al., 2010*; *Dapporto et al., 2019*; *Platania et al., 2020*). This continental-scale study revealed that average maximum CO1 divergences within species was nearly four times less than the average minimum distance to the nearest neighbor. Reflecting this fact, 76% of the species displayed a barcode gap ensuring their unambiguous identification. Identification of specimens using the criterion of monophyly raised identification success as 80% of North American butterfly species formed a monophyletic cluster. These results approximate those obtained in a study of European butterflies where monophyly was met for 94% of species at a regional level (Iberia) ($\chi 2 = 23.1$; p-value <0.001) versus 85% for the continent (*Dincă et al., 2015*) ($\chi 2 = 2.71$; p-value $= 0.099$). As the European study only considered approximately 60% of the fauna, the proportion of monophyletic species is likely to decline with increased coverage. In a study examining how increased geographic scale affected the barcode gap in Asian butterflies, *Lukhtanov et al. (2009)* found that the barcode gap decreased with distance, but that species resolution was nearly unaltered. Interestingly, they showed that monophyly, unlike the barcode gap, was less affected by geographic coverage. The evidence for reduced identification success in the present study is likely explained, at least in part, by the 5-fold higher sampling effort (18 specimens/species) than in *Lukhtanov et al. (2009)* (3 specimens/species). Although the presence of a barcode gap or monophyly are sufficient conditions to ensure the correct

assignment of a sequence to its correct species, these are not essential criteria. For example, a species whose component sequences are paraphyletic or polyphyletic can meet neither criteria, but can be perfectly diagnosable (*Ross, Murugan & Li, 2008*; *Bergsten et al., 2012*).

Because the capacity of DNA barcoding (*Ratnasingham & Hebert, 2007*) to deliver a correct identification ultimately depends upon the presence of diagnostic (or a diagnostic combination of) characters, sequence sharing by species indicates that their discrimination will be compromised. This study revealed that 15% of North American butterfly species share their DNA barcodes with another species. Work on European butterflies revealed 3% barcode sharing at a regional level (Iberia), and 7% at a continental scale (*Dincă et al., 2015*). Although the latter value is about half that observed for North America ($\chi 2 = 12.6$; p-value <0.001), it was inferred based on 60% of the European fauna and more comprehensive taxonomic and geographic coverage will almost certainly increase the incidence of barcode sharing for European butterflies. Because the level of barcode sharing does not only depends on geographic coverage, it is important to ascertain the factors underlying this pattern. First, both the incomplete sorting of ancestral polymorphisms and introgression can lead to barcode sharing, particularly between recently diverged species. The former factor should be less important for mitochondrial than nuclear genes because their lower effective population size facilitates the loss of ancestral polymorphisms. The impact of introgression is more controversial. Although Haldane's rule predicts that the heterogametic sex (females in Lepidoptera) of hybrid individuals is not likely to pass on mitochondrial DNA (*Haldane, 1922*), introgression has often been reported in butterflies (e.g., *Sperling, 1993*; *Gompert et al., 2006*; *Wahlberg et al., 2009*). This could be explained by a more general hypothesis, broadly applicable to all organisms, suggesting that low purifying selection on introgressed mitochondrial genes favours the transfer of these elements (over nuclear ones) across species boundaries (*Harrison, 1993*). Another feature that facilitates contact between closely related lineages, increasing the likelihood for introgression, is the high dispersal capability of some butterflies (*Stevens, Turlure & Baguette, 2010*). Second, although butterflies possess a relatively well-established taxonomy, recent studies have exposed taxonomic uncertainty including cases of over-split species (e.g., *Vila et al., 2010*). Such unrecognized cases of synonymy can produce barcode sharing. Third, although the specimens examined in this study were identified by specialists, DNA barcode results revealed a number of misidentifications. In other cases, the discrimination of morphologically similar species is so difficult that diagnostic characters might have been misinterpreted inflating the incidence of barcode sharing. A full investigation of the roles played by these three factors is beyond the scope of this study, however, an exhaustive study of 42,000 specimens representing nearly 5,000 species of European Lepidoptera showed that just 40% of non-monophyletic species were generated by biological factors (i.e., introgression, incomplete lineage sorting) while 60% reflected methodological problems such as misidentifications (*Mutanen et al., 2016*).

The incidence of barcode sharing in North American butterflies varies nearly 5-fold with latitude, being far higher among species found in the North/alpine (42%) than at Mid-latitude (25%) or South (9%) locales. A similar pattern was evident for 1541 North

American noctuoid moth species as the Canadian fauna showed 10% barcode sharing (*Zahiri et al., 2014*), versus 7% for the continent (*Zahiri et al., 2017*).

Although the main scope of this work was to build a comprehensive barcode library for the identification of North American butterflies, effort was made to capture diversity from regions where contact zones, hybrid zones, and phylogeographic breaks congregate (*Swenson & Howard, 2005*). This strategy aimed to maximize the recovery of haplotype diversity, yet ascertainment bias is inevitable and likely impacted both geography (*Yang, Ma & Kreft, 2013*) and taxonomy (*Troudet et al., 2017*). While under-sampling can exaggerate the sequence divergence between species, comprehensive knowledge of intraspecific diversity will decrease interspecific distances and expose barcode sharing (e.g., *Wiemers & Fiedler, 2007*; *Dasmahapatra et al., 2010*). Our sample sizes were similar (40 specimens/species) for two regions (North/alpine, Mid-latitude), but lower in the South (16 specimens/species). However, iNEXT indicated that similar proportions of CO1 diversity were recovered from each bioregion (66%—North/alpine, 64%—Mid-latitude, 71%—South). Although iNEXT brings statistical rigor to such estimates, its accuracy closely depends on sampling quality. For instance, under-sampling biodiversity hotspots would give the illusion of low diversity and inflate estimates of sampling completeness. This result suggests that the observed differences in barcode sharing are not an artefact of varied sampling coverage. This pattern could well reflect the different exposure of species in each bioregion to the impacts of Quaternary glaciation (*Hewitt, 1996*). Species in northern regions could have experienced recurrent cycles of isolation in glacial refugia, leading to subsequent opportunities for secondary contact and mitochondrial exchange (*Hewitt, 2000*). Moreover, the small size of the populations at the leading edge of the species distribution might aid the fixation of introgressed mitochondria (*Kingman, 1982*). Another consequence of rapid expansion into deglaciated habitats would be low density at the leading edge. This situation can create difficulties in finding a conspecific mate and can weaken isolating mechanisms (e.g., *Shelly & Bailey, 1992*; *Alatalo et al., 1998*; *Willis, Ryan & Rosenthal, 2011*), leading to heterospecific mating and introgression (e.g., *Wirtz, 1999*; *Randler, 2002*). Not only scarcity of conspecifics could favor hybrid formation, but it could also enhance their persistence because of decreased competition with parental populations (*Arnold, 1997*).

Aside from this latitudinal pattern, barcode sharing varied nearly 4-fold among butterfly families ($\chi 2 = 25.13$; p-value <0.001), from 10% in Hesperiidae to 38% in Pieridae. Interestingly, a similar pattern was also observed at lower taxonomic rank, among genera, where the incidence of barcode sharing showed a weak and non-significant correlation with the number of species in a genus suggesting taxonomic localization. The genus *Colias* (Pieridae) was particularly impacted as all 22 species shared at least one of their barcode sequences with another species (Table S2), perhaps reflecting their recent radiation (*Chew & Watt, 2006*; *Wheat & Watt, 2008*). Not only this pattern explains the introgression of haplotypes between hybridizing species such as *C. eurytheme*/*C. phildice* in the eastern USA (*Gerould, 1946*; *Jahner, Shapiro & Forister, 2011*), and *C. eurytheme*/*C. eriphyle* in the west (*Taylor Jr, 1972*), but it also lays the foundation for operational issues such as misidentifications reflecting the unsettled taxonomy of the genus (*Wheat & Watt, 2008*).

DNA barcoding combined with species delimitation methods enables rapid, cost-effective surveys of biodiversity. While this is particularly beneficial for poorly-studied taxonomic groups, it can also disclose overlooked diversity in well-studied taxa. BIN analysis indicated that ~11% of North American butterfly species (6.8% Splits, 4.6% Mixture) were split into two or more entities. The incidence of such cases mirrors values (9–12%) reported in prior studies on Lepidoptera (*Huemer et al., 2014*; *Zahiri et al., 2014*). GMYC analysis showed an even higher discordance with current taxonomy, showing that about 17% of species (7.8% Splits, 9.1% Mixture) potentially involve overlooked diversity. Interestingly, when the same approach was applied to 60% of European butterfly species, there was even higher discordance (28%) (*Dincă et al., 2015*). It is probable that the varied habitats in the Mediterranean basin (*Blondel et al., 2010*), coupled with the presence of southern refugia during the Pleistocene (*Schmitt, 2007*), created more genetic structure and/or speciation in Europe (*Vodă et al., 2016*; *Dapporto et al., 2019*; *Scalercio et al., 2020*). Employing a fixed divergence threshold (2.5%), about 10% of North American butterfly species exceeded this criterion. Similar values (8–12% with a 2% threshold) have been reported in other studies on Lepidoptera (*Huemer et al., 2014*; *Zahiri et al., 2014*). Based on these results, it is likely that a considerable number of cryptic species await description or that Evolutionary Significant Units (ESUs) within species deserve protection (*Avise, 1989*). Detailed studies (e.g., nuclear genetic, morphological, ecological) (*DeSalle, Egan & Siddall, 2005*) should be undertaken on these lineages (*Polasky & Solow, 1999*).

CONCLUSION

This study has generated one of the first continental-scale DNA barcode libraries for an entire taxonomic group. Beyond providing an identification system for most (>96%) North American butterflies, it creates the foundation needed to test the current classification. This library also delivers an overview of large-scale patterns of genetic diversity revealing cases of evolutionary interest such as potential hybridization and of importance to biodiversity conservation such as cryptic diversity and evolutionary significant units. As such, this study provides a basis for improving understanding of the mechanisms that have shaped genetic diversity in the North American butterfly fauna.

ACKNOWLEDGEMENTS

We thank Rachel Breese, Ernst Brockmann, Julio Genaro, Angela Gradish, Crispin Guppy, Winnie Hallwachs, Axel Hausmann, Jennifer Heron, Peter Houlihan, Leland Humble, Tea Huotari, Daniel Janzen, Donald Lafontaine, Jean-Francois Landry, Luis Leite, Beverly Mcclenaghan, Jim Moore, Marko Mutanen, Vazrick Nazari, Ezequiel Nunez Bustos, Nicholas Pardikes, Doo-Sang Park, Samuel Pinna, Carmen Pozo de la Tijera, Blanca Prado, Rodolphe Rougerie, Chris Schmidt, Noemy Seraphim, Kimberley Shropshire, Daniel Sigouin, Derek Sikes, Alex Smith, Mark Stoeckle, James Sullivan, Jon Turner, Chithravel Vadivalagan, David Wagner, Karen Wright, Ellen Yerger, Alejandro Zaldivar-Riveron, and Manuel Zumbado for granting use of their specimen records. We also thank curatorial staff at the Smithsonian's National Museum of Natural History for access to specimens, namely

Brian Harris, Don Harvey, Nick Silverson, and Margaret Rosati. We thank the Collections and Informatics Units at the Centre for Biodiversity Genomics for facilitating access to specimens and their genetic analysis, especially Meredith Miller and Allison Brown. We also thank Michelle D'Souza, Leonardo Dapporto, and Robert Young for help with the bioinformatic analysis. Finally, we thank the Canadian Centre for DNA Barcoding (CCDB) for sequence analysis.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by grants to Paul Hebert from NSERC, Genome Canada through Ontario Genomics, the Canada Foundation for Innovation, and the Canada Research Chairs Program. Support for this research was also provided by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (project no. 625997) and by the Academy of Finland to Vlad Dincă (Academy Research Fellow, decision no. 328895). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NSERC.
Ontario Genomics.
Canada Foundation for Innovation.
Canada Research Chairs Program.
7th European Community Framework Programme: 625997.
Academy of Finland to Vlad Dincă: 328895.
Genome Canada.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Jacopo D'Ercole conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Vlad Dincă conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, data collection, and approved the final draft.
- Paul A. Opler, Norbert Kondla, Christian Schmidt and Robert Robbins analyzed the data, authored or reviewed drafts of the paper, data share and collection, and approved the final draft.
- Jarrett D. Phillips analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

- John M. Burns, Scott E. Miller, Evgeny V. Zakharov, Jeremy R. DeWaard and Sujeevan Ratnasingham analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Nick Grishin analyzed the data, authored or reviewed drafts of the paper, data share, and approved the final draft.
- Paul D.N. Hebert conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Final barcode sequences, including specimen metadata for the 14623 samples are available at the BOLD database: dataset "DS-USCANLEP" (dx.doi.org/10.5883/DS-USCANLEP).

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11157#supplemental-information.

REFERENCES

Alatalo RV, Kotiaho J, Mappes J, Parri S. 1998. Mate choice for offspring performance: major benefits or minor costs? *Proceedings of the Royal Society B: Biological Sciences* **265**:2297–2301 DOI 10.1098/rspb.1998.0574.

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert TPM, Willerslev E, Zhang G, Scofield RP, Holdaway RN, Bunce M. 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences* **279**:4724–4733 DOI 10.1098/rspb.2012.1745.

Arnold ML. 1997. *Natural hybridization and evolution*. New York: Oxford University Press.

Avise JC. 1989. A role for molecular genetics in the recognition and conservation of endangered species. *Trends in Ecology & Evolution* **4(9)**:279–281 DOI 10.1016/0169-5347(89)90203-6.

Avise JC. 2000. *Phylogeography: the history and formation of species*. Cambridge: Harvard University Press.

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN, Ribera I, Nilsson AN, Barraclough TG, Vogler AP. 2012. The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology* **61**:851–869 DOI 10.1093/sysbio/sys037.

Blondel J, Aronson J, Bodiou J, Boeuf G. 2010. *The Mediterranean region: biological diversity in space and time*. Oxford: Oxford University Press.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary

analysis. *PLOS Computational Biology* **10(4)**:e1003537 DOI 10.1371/journal.pcbi.1003537.

Brock JP, Kaufman K. 2006. *Kaufman field guide to butterflies of North America*. Boston: Houghton Mifflin Harcourt.

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN. 2007. DNA barcodes of closely related (but morphologically and ecologically distinct) species of skipper butterflies (Hesperiidae) can differ by only one to three nucleotides. *Journal of the Lepidopterists' Society* **61(3)**:138–153.

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN. 2008. DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservacion Guanacaste, Costa Rica. *Proceedings of the National Academy of Sciences of the United States of America* **105(17)**:6350–6355 DOI 10.1073/pnas.0712181105.

Cariou M, Henri H, Martinez S, Duret L, Charlat S. 2020. How consistent is RADseq divergence with DNAbarcode based clustering in insects? *Molecular Ecology Resources* **20**:1294–1298 DOI 10.1111/1755-0998.13178.

Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11(4)**:265–270.

Chew FS, Watt WB. 2006. The green-veined white (*Pieris napi* L.), its Pierine relatives, and the systematics dilemmas of divergent character sets (Lepidoptera, Pieridae). *Biological Journal of the Linnean Society* **88(3)**:413–435 DOI 10.1111/j.1095-8312.2006.00630.x.

Collins RA, Cruickshank RH. 2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources* **13(6)**:969–975 DOI 10.1111/1755-0998.12046.

Cong Q, Shen J, Borek D, Robbins RK, Opler PA, Otwinowski Z, Grishin NV. 2017. When COI barcodes deceive: complete genomes reveal introgression in hairstreaks. *Proceedings of the Royal Society B: Biological Sciences* **284(1848)**:18480161735 DOI 10.1098/rspb.2016.1735.

Cong Q, Shen J, Borek D, Robbins RK, Otwinowski Z, Grishin NV. 2016. Complete genomes of Hairstreak butterflies, their speciation and nucleo-mitochondrial incongruence. *Scientific Reports* **6**:24863 DOI 10.1038/srep24863.

Dapporto L, Cini A, Vodă R, Dincă V, Wiemers M, Menchetti M, Magini G, Talavera G, Shreeve T, Bonelli S, Casacci LP, Balletto E, Scalercio S, Vila R. 2019. Integrating three comprehensive data sets shows that mitochondrial DNA variation is linked to species traits and paleogeographic events in European butterflies. *Molecular Ecology Resources* **19(6)**:1623–1636 DOI 10.1111/1755-0998.13059.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. JModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9(8)**:772 DOI 10.1038/nmeth.2109.

Dasmahapatra KK, Elias M, Hill RI, Hoffman JI, Mallet J. 2010. Mitochondrial DNA barcoding detects some species that are real, and some that are not. *Molecular Ecology Resources* **10(2)**:264–273 DOI 10.1111/j.1755-0998.2009.02763.x.

D'Ercole J, Prosser SW, Hebert PDN. 2021. A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera)—patterns of nucleotide misincorporation. *PeerJ* **9**:e10420 DOI 10.7717/peerj.10420.

DeSalle R, Egan MG, Siddall M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360(1462)**:1905–1916 DOI 10.1098/rstb.2005.1722.

DeWaard JR, Ivanova NV, Hajibabaei M, Hebert PDN. 2008. Assembling DNA Barcodes. Analytical protocols. *Methods in Molecular Biology* **410**:275–293 DOI 10.1007/978-1-59745-548-0_15.

Dincă V, Bálint ZS, Vodă R, Dapporto L, Hebert PDN, Vila R. 2018. Use of genetic, climatic, and microbiological data to inform reintroduction of a regionally extinct butterfly. *Conservation Biology* **32(4)**:828–837 DOI 10.1111/cobi.13111.

Dincă V, Dapporto L, Somervuo P, Vodă R, Cuvelier S, Gascoigne-Pees M, Huemer P, Mutanen M, Hebert PDN, Vila R. 2021. High resolution DNA barcode library for European butterflies reveals continental patterns of mitochondrial genetic diversity. *Communications Biology* **4**:315 DOI 10.1038/s42003-021-01834-7.

Dincă V, Montagud S, Talavera G, Hernández-Roldán J, Munguira ML, García-Barros E, Hebert PDN, Vila R. 2015. DNA barcode reference library for Iberian butterflies enables a continental-scale preview of potential cryptic diversity. *Scientific Reports* **5**:12395 DOI 10.1038/srep12395.

Dincă V, Wiklund C, Lukhtanov VA, Kodandaramaiah U, Noren K, Dapporto L, Wahlberg N, Vila R, Friberg M. 2013. Reproductive isolation and patterns of genetic differentiation in a cryptic butterfly species complex. *Journal of Evolutionary Biology* **26(10)**:2095–2106 DOI 10.1111/jeb.12211.

Dincă V, Zakharov EV, Hebert PDN, Vila R. 2011. Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B: Biological Sciences* **278**:347–355 DOI 10.1098/rspb.2010.1089.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14(9)**:755–763 DOI 10.1093/bioinformatics/14.9.755.

Ezard T, Fujisawa T, Barraclough TG. 2009. Splits: SPecies' LImits by Threshold Statistics. *Available at r-forge.r-project.org/R/?group_id=333* (accessed on 10 October 2019).

Fujisawa T, Barraclough TG. 2013. Delimiting species using single-locus data and the generalized mixed Yule coalescent approach: a revised method and evaluation on simulated data sets. *Systematic Biology* **62(5)**:707–724 DOI 10.1093/sysbio/syt033.

Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S, Labra M. 2013. DNA barcoding as a new tool for food traceability. *Food Research International* **50(1)**:55–63 DOI 10.1016/j.foodres.2012.09.036.

Gerould JH. 1946. Hybridization and female albinism in *Colias philodice* and *C. eurytheme*. A New Hampshire survey in 1943 with subsequent data. *Annals of the Entomological Society of America* **39(3)**:383–396 DOI 10.1093/aesa/39.3.383.

Gompert Z, Lucas LK, Fordyce JA, Forister ML, Nice CC. 2010. Secondary contact between *Lycaeides idas* and *L. melissa* in the Rocky Mountains: extensive admixture and a patchy hybrid zone. *Molecular Ecology* **19(15)**:3171–3192 DOI 10.1111/j.1365-294X.2010.04727.x.

Gompert Z, Nice CC, Fordyce JA, Forister ML, Shapiro AM. 2006. Identifying units for conservation using molecular systematics: the cautionary tale of the Karner blue butterfly. *Molecular Ecology* **15(7)**:1759–1768 DOI 10.1111/j.1365-294X.2006.02905.x.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52(5)**:696–704 DOI 10.1080/10635150390235520.

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 of the United States of America* **103(4)**:968–971 DOI 10.1073/pnas.0510466103.

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 DOI 10.1016/j.tig.2007.02.001.

Haldane JBS. 1922. Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics* **12(2)**:101–109 DOI 10.1007/BF02983075.

Harrison RG. 1993. *Hybrid zones and the evolutionary process*. New York: Oxford University Press.

Hausmann A, Hebert PDN, Mitchell A, Rougerie R, Sommerer M, Edwards T, Young CJ. 2009. Revision of the Australian *Oenochroma vinaria* Guenée, 1858 species-complex (Lepidoptera: Geometridae, Oenochrominae): DNA barcoding reveals cryptic diversity and assesses status of type specimen without dissection. *Zootaxa* **2239**:1–21 DOI 10.5281/zenodo.190505.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* **270(1512)**:313–321 DOI 10.1098/rspb.2002.2218.

Hebert PDN, DeWaard JR, Zakharov EV, Prosser SWJ, Sones JE, McKeown JTA, Mantle B, La Salle J. 2013. A DNA Barcode Blitz: rapid digitization and sequencing of a natural history collection. *PLOS ONE* **8(7)**:e68535 DOI 10.1371/journal.pone.0068535.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America* **101(41)**:14812–14817 DOI 10.1073/pnas.0406166101.

Hennig W. 1966. *Phylogenetic systematics*. Urbana: University of Illinois Press, 72–77.

Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58(3)**:247–276 DOI 10.1111/j.1095-8312.1996.tb01434.x.

Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405**:907–913 DOI 10.1038/35016000.

Hsieh TC, Ma KH, Chao A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* **7**:1451–1456 DOI 10.1111/2041-210X.12613.

Huemer P, Mutanen M, Sefc KM, Hebert PDN. 2014. Testing DNA barcode performance in 1000 species of European Lepidoptera: large geographic distances have small genetic impacts. *PLOS ONE* **9(12)**:e115774 DOI 10.1371/journal.pone.0115774.

Ivanova NV, DeWaard JR, Hebert PDN. 2006. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes* **6**:998–1002 DOI 10.1111/j.1471-8286.2006.01428.x.

Jahner JP, Shapiro AM, Forister ML. 2011. Drivers of hybridization in a 66-generation record of *Colias* butterflies. *Evolution* **66(3)**:818–830 DOI 10.1111/j.1558-5646.2011.01481.x.

Janzen DH, Hajibabaei M, Burns J, Hallwachs W, Remigio E, Hebert PDN. 2005. Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**:835–1845 DOI 10.1098/rstb.2005.1715.

Kekkonen M, Hebert PDN. 2014. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Molecular Ecology Resources* **14(4)**:706–715 DOI 10.1111/1755-0998.12233.

Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, Hebert PDN. 2007. Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes* **7(4)**:535–543 DOI 10.1111/j.1471-8286.2007.01670.x.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111–120 DOI 10.1007/BF01731581.

Kingman JFC. 1982. On the genealogy of large populations. *Journal of Applied Probability* **19(A)**:27–43 DOI 10.2307/3213548.

Lavinia PD, Núñez Bustos EO, Kopuchian C, Lijtmaer DA, García NC, Hebert PDN, Tubaro PL. 2017. Barcoding the butterflies of southern South America: species delimitation efficacy, cryptic diversity and geographic patterns of divergence. *PLOS ONE* **12(10)**:e0186845 DOI 10.1371/journal.pone.0186845.

Lees DC, Lack HW, Rougerie R, Hernandez-Lopez A, Raus T, Avtzis ND, Augustin S, Lopez-Vaamonde C. 2011. Tracking origins of invasive herbivores through herbaria and archival DNA: the case of the horse-chestnut leaf miner. *Frontiers in Ecology and the Environment* **9**:322–328 DOI 10.1890/100098.

Lukhtanov VA, Sourakov A, Zakharov EV, Hebert PDN. 2009. DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources* **9**:1302–1310 DOI 10.1111/j.1755-0998.2009.02577.x.

Miralles A, Vences M. 2013. New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. *PLOS ONE* **8**(**7**):e68242 DOI 10.1371/journal.pone.0068242.

Mutanen M, Kivelä SM, Vos RA, Doorenweerd C, Ratnasingham S, Hausmann A, Huemer P, Dinča V, Van Nieukerken EJ, Lopez-Vaamonde C, Vila R, Aarvik L, Decaëns T, Efetov KA, Hebert PDN, Johnsen A, Karsholt O, Pentinsaari M, Rougerie R, Segerer A, Tarmann G, Zahiri R, Godfray HCJ. 2016. Species-level para- and polyphyly in DNA barcode gene trees: strong operational bias in European Lepidoptera. *Systematic Biology* **65**(**6**):1024–1040 DOI 10.1093/sysbio/syw044.

New TR. 1997. Are Lepidoptera an effective 'umbrella group' for biodiversity conservation? *Journal of Insect Conservation* **1**:5–12 DOI 10.1023/A:101843340.

Platania L, Vodă R, Dincă V, Talavera G, Vila R, Dapporto L. 2020. Integrative analyses on Western Palearctic *Lasiommata* reveal a mosaic of nascent butterfly species. *Journal of Zoological Systematics and Evolutionary Research* **58**:809–822 DOI 10.1111/jzs.12356.

Polasky S, Solow AR. 1999. Conserving biological diversity with scarce resources. In: Klopatek JM, Gardner RH, eds. *Landscape ecological analysis*. New York: Springer Publishing, 154–174.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* **55**(**4**):595–609 DOI 10.1080/10635150600852011.

Quek SP, Davies SJ, Itino T, Pierce NE. 2004. Codiversification in an ant-plant mutualism: stem texture and the evolution of host use in *Crematogaster* (Formicidae: Myrmicinae) inhabitants of *Macaranga* (Euphorbiaceae). *Evolution* **58**(**3**):554–570 DOI 10.1111/j.0014-3820.2004.tb01678.x.

R Studio Team. 2016. *RStudio: integrated development for R*. Boston: RStudio, Inc.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* **67**(**5**):901–904 DOI 10.1093/sysbio/syy032.

Randler C. 2002. Avian hybridization, mixed pairing and female choice. *Animal Behaviour* **63**(**1**):103–119 DOI 10.1006/anbe.2001.1884.

Ratnasingham S, Hebert PDN. 2007. BOLD: the barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes* **7**(**3**):355–364 DOI 10.1111/j.1471-8286.2006.01678.x.

Ratnasingham S, Hebert PDN. 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLOS ONE* **8**(**8**):e66213 DOI 10.1371/journal.pone.0066213.

Rehman A, Jafar S, Raja NA, Mahar J. 2015. Use of DNA barcoding to control the illegal wildlife trade: a CITES case report from Pakistan. *Journal of Bioresource Management* **2**(**2**):19–22 DOI 10.35691/JBM.5102.0017.

Ross HA, Murugan S, Li WLS. 2008. Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology* **57(2)**:216–230 DOI 10.1080/10635150802032990.

Scalercio S, Cini A, Menchetti M, Vodă R, Bonelli S, Bordoni A, Casacci LP, Dincă V, Balletto E, Vila R, Dapporto L. 2020. How long is 3 km for a butterfly? Ecological constraints and functional traits explain high mitochondrial genetic diversity between Sicily and the Italian Peninsula. *Journal of Animal Ecology* **89**:2013–2026 DOI 10.1111/1365-2656.13196.

Schmitt T. 2007. Molecular biogeography of Europe: pleistocene cycles and postglacial trends. *Frontiers in Zoology* **4**:11 DOI 10.1186/1742-9994-4-11.

Scott JA. 1986. *The butterflies of North America: a natural history and field guide*. Stanford: Stanford University Press.

Shelly TE, Bailey WJ. 1992. Experimental manipulation of mate choice by male katydids: the effect of female encounter rate. *Behavioral Ecology and Sociobiology* **30**:277–282 DOI 10.1007/BF00166713.

Sperling FAH. 1993. Mitochondrial DNA variation and Haldane's rule in the *Papilio glaucus* and *P. troilus* species groups. *Heredity* **71**:227–233 DOI 10.1038/hdy.1993.130.

Stevens VM, Turlure C, Baguette M. 2010. A meta-analysis of dispersal in butterflies. *Biological Reviews* **85(3)**:413–683 DOI 10.1111/j.1469-185X.2009.00119.x.

Swenson NG, Howard DJ. 2005. Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American Naturalist* **166(5)**:581–591 DOI 10.1086/491688.

Syaripuddin K, Sing K, Wilson J. 2015. Comparison of butterflies, bats and beetles as bioindicators based on four key criteria and DNA barcodes. *Tropical Conservation Science* **8(1)**:138–149 DOI 10.1177/194008291500800112.

Talavera G, Dincă V, Vila R. 2013. Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. *Methods in Ecology and Evolution* **4**:1101–1110 DOI 10.1111/2041-210X.12107.

Tang CQ, Humphreys AM, Fontaneto D, Barraclough TG. 2014. Effects of phylogenetic reconstruction on the robustness of species delimitation methods using single-locus data. *Methods in Ecology and Evolution* **5**:1086–1094 DOI 10.1111/2041-210X.12246.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**:57–86.

Taylor Jr OR. 1972. Random vs. non-random mating in the sulfur butterflies, Colias eurytheme and *Colias philodice* (Lepidoptera: Pieridae). *Evolution* **26(3)**:344–356 DOI 10.2307/2407010.

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7**:9132 DOI 10.1038/s41598-017-09084-6.

Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. *Trends in Ecology and Evolution* **24(2)**:110–117 DOI 10.1016/j.tree.2008.09.011.

Vila R, Lukhtanov VA, Talavera G, Gil-t F, Pierce NE. 2010. How common are dot-like distributions? Taxonomical oversplitting in western European *Agrodiaetus* (Lepidoptera: Lycaenidae) revealed by chromosomal and molecular markers. *Biological Journal of the Linnean Society* **101(1)**:130–154 DOI 10.1111/j.1095-8312.2010.01481.x.

Vodă R, Dapporto L, Dincă V, Shreeve TG, Khaldi M, Barech G, Rebbas K, Sammut P, Scalercio S, Hebert PDN, Vila R. 2016. Historical and contemporary factors generate unique butterfly communities on islands. *Scientific Reports* **6**:28828 DOI 10.1038/srep28828.

Wahlberg N, Weingartner E, Warren AD, Nylin S. 2009. Timing major conflict between mitochondrial and nuclear genes in species relationships of *Polygonia* butterflies (Nymphalidae: Nymphalini). *BMC Evolutionary Biology* **9**:92 DOI 10.1186/1471-2148-9-92.

Warren AD, Davis KJ, Grishin NV, Pelham JP, Stangeland EM. 2012. Illustrated lists of American butterflies. *Available at www.butterfliesofamerica.com* (accessed on 15 April 2019).

Wheat CW, Watt WB. 2008. A mitochondrial-DNA-based phylogeny for some evolutionary-genetic model species of *Colias* butterflies (Lepidoptera, Pieridae). *Molecular Phylogenetics and Evolution* **47(3)**:893–902 DOI 10.1016/j.ympev.2008.03.013.

Wiemers M, Fiedler K. 2007. Does the DNA barcoding gap exist?—a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* **4**:8 DOI 10.1186/1742-9994-4-8.

Willis PM, Ryan MJ, Rosenthal GG. 2011. Encounter rates with conspecific males influence female mate choice in a naturally hybridizing fish. *Behavioral Ecology* **22(6)**:1234–1240 DOI 10.1093/beheco/arr119.

Wirtz P. 1999. Mother species–father species: unidirectional hybridization in animals with female choice. *Animal Behaviour* **58(1)**:1–12 DOI 10.1006/anbe.1999.1144.

Wright S. 1943. Isolation by distance. *Genetics* **28(2)**:114–138.

Yang W, Ma K, Kreft H. 2013. Geographical sampling bias in a large distributional database and its effects on species richness–environment models. *Journal of Biogeography* **40**:1415–1426 DOI 10.1111/jbi.12108.

Zahiri R, Lafontaine JD, Schmidt BC, DeWaard JR, Zakharov EV, Hebert PDN. 2014. A transcontinental challenge—a test of DNA barcode performance for 1,541 species of Canadian Noctuoidea (Lepidoptera). *PLOS ONE* **9(3)**:e92797 DOI 10.1371/journal.pone.0092797.

Zahiri R, Lafontaine JD, Schmidt BC, DeWaard JR, Zakharov EV, Hebert PDN. 2017. Probing planetary biodiversity with DNA barcodes: the Noctuoidea of North America. *PLOS ONE* **12(6)**:e0178548 DOI 10.1371/journal.pone.0178548.

Zakharov EV, Lobo NF, Nowak C, Hellmann JJ. 2009. Introgression as a likely cause of mtDNA paraphyly in two allopatric skippers (Lepidoptera: Hesperiidae). *Heredity* **102**:590–599 DOI 10.1038/hdy.2009.26.