Syst. Biol. 65(6):1024–1040, 2016

DOI:10.1093/sysbio/syw044
Advance Access publication June 10, 2016

Species-Level Para- and Polyphyly in DNA Barcode Gene Trees: Strong Operational Bias in European Lepidoptera

MARKO MUTANEN[1,*], SAMI M. KIVELÄ[2], RUTGER A. VOS[3], CAMIEL DOORENWEERD[3], SUJEEVAN RATNASINGHAM[4], AXEL HAUSMANN[5], PETER HUEMER[6], VLAD DINCĂ[4,7], ERIK J. VAN NIEUKERKEN[3], CARLOS LOPEZ-VAAMONDE[8,9], ROGER VILA[7], LEIF AARVIK[10], THIBAUD DECAËNS[11], KONSTANTIN A. EFETOV[12], PAUL D. N. HEBERT[4], ARILD JOHNSEN[10], OLE KARSHOLT[13], MIKKO PENTINSAARI[1], RODOLPHE ROUGERIE[14], ANDREAS SEGERER[5], GERHARD TARMANN[6], REZA ZAHIRI[4,15], AND H. CHARLES J. GODFRAY[16]

[1]*Department of Genetics and Physiology, University of Oulu, Finland;* [2]*Department of Ecology, University of Oulu, Finland;* [3]*Naturalis Biodiversity Center, Leiden, The Netherlands;* [4]*Centre for Biodiversity Genomics, Biodiversity Institute of Ontario, University of Guelph, Canada;* [5]*SNSB – Bavarian State Collection of Zoology, Munich, Germany;* [6]*Tiroler Landesmuseen-Betriebsgesellschaft m.b.H., Innsbruck, Austria;* [7]*Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain;* [8]*INRA, UR633 Zoologie Forestière, 45075 Orléans, France;* [9]*Institut de Recherche sur la Biologie de l'Insecte, CNRS UMR 7261, Université François-Rabelais de Tours, UFR Sciences et Techniques, 37200 Tours, France* [10]*Natural History Museum University of Oslo, Norway;* [11]*Centre d'Écologie Fonctionnelle et Évolutive, UMR 5175 CNRS / University of Montpellier / University of Montpellier 3 / EPHE / SupAgro Montpellier / INRA / IRD, 1919 Route de Mende, 34293 Montpellier Cedex 5, France;* [12]*Crimean Federal University, Simferopol, Crimea;* [13]*Zoologisk Museum, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark;* [14]*Département Systématique et Evolution, Muséum National d'Histoire Naturelle, Institut de Systématique, Evolution, Biodiversité, ISYEB–UMR 7205 MNHN, CNRS, UPMC, EPHE, Sorbonne Universités, Paris, France;* [15]*Ottawa Plant Laboratory, Canadian Food Inspection Agency, Canada;* [16]*Department of Zoology, University of Oxford, UK;*
**Correspondence to be sent to: Department of Genetics and Physiology, University of Oulu, PO Box 3000, Oulu FI-90014, Finland; E-mail: marko.mutanen@oulu.fi*

Received 14 December 2015; reviews returned 18 May 2016; accepted 18 May 2016
Associate Editor: Karl Kjer

Abstract.—The proliferation of DNA data is revolutionizing all fields of systematic research. DNA barcode sequences, now available for millions of specimens and several hundred thousand species, are increasingly used in algorithmic species delimitations. This is complicated by occasional incongruences between species and gene genealogies, as indicated by situations where conspecific individuals do not form a monophyletic cluster in a gene tree. In two previous reviews, non-monophyly has been reported as being common in mitochondrial DNA gene trees. We developed a novel web service "Monophylizer" to detect non-monophyly in phylogenetic trees and used it to ascertain the incidence of species non-monophyly in COI (a.k.a. cox1) barcode sequence data from 4977 species and 41,583 specimens of European Lepidoptera, the largest data set of DNA barcodes analyzed from this regard. Particular attention was paid to accurate species identification to ensure data integrity. We investigated the effects of tree-building method, sampling effort, and other methodological issues, all of which can influence estimates of non-monophyly. We found a 12% incidence of non-monophyly, a value significantly lower than that observed in previous studies. Neighbor joining (NJ) and maximum likelihood (ML) methods yielded almost equal numbers of non-monophyletic species, but 24.1% of these cases of non-monophyly were only found by one of these methods. Non-monophyletic species tend to show either low genetic distances to their nearest neighbors or exceptionally high levels of intraspecific variability. Cases of polyphyly in COI trees arising as a result of deep intraspecific divergence are negligible, as the detected cases reflected misidentifications or methodological errors. Taking into consideration variation in sampling effort, we estimate that the true incidence of non-monophyly is ~23%, but with operational factors still being included. Within the operational factors, we separately assessed the frequency of taxonomic limitations (presence of overlooked cryptic and oversplit species) and identification uncertainties. We observed that operational factors are potentially present in more than half (58.6%) of the detected cases of non-monophyly. Furthermore, we observed that in about 20% of non-monophyletic species and entangled species, the lineages involved are either allopatric or parapatric—conditions where species delimitation is inherently subjective and particularly dependent on the species concept that has been adopted. These observations suggest that species-level non-monophyly in COI gene trees is less common than previously supposed, with many cases reflecting misidentifications, the subjectivity of species delimitation or other operational factors. [DNA barcoding; gene tree; Lepidoptera; mitochondrial COI; mitochondrial cox1; paraphyly; polyphyly; species delimitation; species monophyly.]

There has been endless debate over the definition of a species and whether the concept has any biological reality (Wheeler and Meier 2000; De Queiroz 2007; Mallet 2007; Hausdorf 2011). While there is now a broad consensus that more inclusive taxonomic categories are defined solely following cladistic principles (Hennig 1966) (i.e., by monophyly criterion and hierarchical order) and are largely arbitrary, species are generally viewed as natural entities with observable distances between them, resulting from the differentiation of lineages through speciation (Wright 1940; Mayr 1942; Coyne and Orr 2004). However, species boundaries are often much harder to discern when individuals are sampled across geographical scales or through time (Baselga et al. 2013), and the complexity in gathering direct evidence on the potential for interbreeding creates challenges for rigorous testing of species boundaries. Nonetheless, the species rank has maintained its status as a central concept in virtually all fields of biology, one with particular societal relevance because of its centrality in conservation, legislation, or food trade (e.g., Avise 1989; Isaac et al. 2004). Although there are species

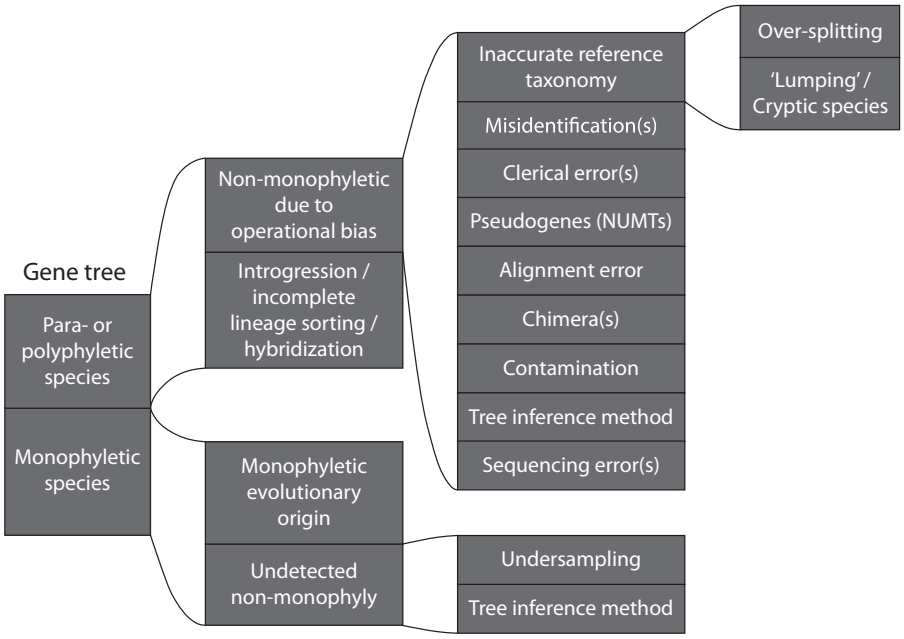

FIGURE 1. Schematic overview of the different potential reasons for species to be classified as para- or polyphyletic, or false-positively monophyletic in a gene tree.

concepts that do not perceive species necessarily as monophyletic entities (for reviews see De Queiroz 2007; Mallet 2007), monophyly is a central criterion in most of them (e.g., the phylogenetic species concept) (Cracraft 1989).

A phylogenetic tree depicting relationships among species is known as a species tree. Because evolution is a unique and nonrecurrent process, there is only a single true topology that reflects evolutionary relationships among species. Systematists have traditionally inferred tree topologies based on morphological, ecological, or other life history characters, but now this is most often done based on DNA sequences. A much-discussed issue concerning the use of DNA sequences is that the evolution of a gene is not necessarily congruent with that of a species (Pamilo and Nei 1988; Maddison 1997). Because nuclear genes of sexually reproducing organisms are subject to recombination, their coalescence histories differ. With rare exceptions, mitochondrial genes are inherited uniparentally (usually maternally), show very limited recombination, have population genetics governed by an effective population size (N_e) that is one quarter of that for the nuclear genome, and are particularly susceptible to selective sweeps (Hurst and Jiggins 2005; Rubinoff et al. 2006).

A species tree, or a phylogenetic tree of any taxonomic rank, is ideally presented as the consensus of a "cloud" of several gene trees (Maddison 1997; Steel and Velasco 2014). This is now common practice for higher-level phylogenies. Assessing whether a species is monophyletic becomes possible when a tree, built from one or more DNA markers, includes multiple individuals per species (when using phrases like "monophyletic species," we do so for convenience only, and actually refer to the pattern observed in a gene tree that includes multiple individuals per species). If all species form reciprocally monophyletic groups then species delimitation is usually straightforward. However, the frequency of non-monophyly may be underestimated if insufficient specimens are surveyed or sampling is geographically restricted. All individuals in a monophyletic species have a common ancestor (otherwise, the species is polyphyletic) that is shared by individuals of no other species (otherwise it is paraphyletic). In practice, paraphyly and polyphyly are closely related phenomena that are often difficult to distinguish (Platnick 1977). There are biological and nonbiological reasons for species appearing as non-monophyletic in a gene tree (Fig. 1). The former include incomplete lineage sorting, introgression, and hybridization; all of which are more likely to occur among recently diverged species than older lineages, and they are not mutually exclusive. Nonbiological reasons include errors in trait categorization and analysis (Ross 2014) including inaccurate reference taxonomy, misidentifications, clerical errors, amplification of paralogous genes, alignment errors, inclusion of chimeras (resulting from the erroneous combination of sequences derived from multiple species), contamination, and methodological issues in phylogenetic inference (e.g., long-branch attraction). Inaccurate taxonomy may result from two contrasting phenomena: oversplitting of species, or lumping of (often cryptic) species. We refer to all these nonbiological processes as sources of operational bias.

Rapidly accumulating DNA barcode libraries, such as the Barcode of Life Data System (BOLD; Ratnasingham and Hebert 2007), are boosting the number of species

being described as new to science (Olave et al. 2014). Typically, DNA barcode gene trees are used as an important source of information in species delimitation. Several algorithmic species delimitation tools have been developed, some of them specifically for use with DNA barcodes (Pons et al. 2006; Puillandre et al. 2012; Fujisawa and Barraclough 2013; Ratnasingham and Hebert 2013; Zhang et al. 2013; Jones et al. 2015). However, these methods can misdiagnose species boundaries when gene trees are non-monophyletic and hence it is important to know the frequency of paraphyly and polyphyly. A benchmark review concluded that on average 23% of species express non-monophyly in mitochondrial DNA markers (Funk and Omland 2003). In 146 studies of arthropods, the average percentage of non-monophyletic species was 26.5% and in other invertebrates as high as 38.6% (Funk and Omland 2003). Recently, Ross (2014) estimated the incidence of paraphyly (including polyphyly) with an extensive, but not completely validated, data set of the animal COI barcodes accessed from BOLD, and concluded that 19% of species were non-monophyletic. For Lepidoptera, he reported the level of 17% of non-monophyly.

Here, we studied the incidence of para- and polyphyly in trees built using the standard COI barcode gene. The data set we analyzed included 41,583 specimens of 4977 species of European Lepidoptera, 50.6% of all known species in this order from this continent. We developed a novel tool (named Monophylizer) that permits the automated identification of paraphyletic and polyphyletic lineages based on tree topology. Prior to carrying out the non-monophyly assessment, we paid particular attention to potential sources of operational bias, in particular by cross-checking identifications (based on current morphology-based taxonomy of all studied taxa) as carefully as possible. We then continued by assessing various other operational factors, particularly taxonomic uncertainties and identification difficulties. Our study is the first to attempt to estimate the significance of these effects. We also applied statistical modeling approaches not previously used in this context to test the effect of sampling effort.

Material and Methods

Target Group

The insect order Lepidoptera constitutes one of the most diverse animal groups; it contains approximately 157,500 described extant species (van Nieukerken et al. 2011). Lepidoptera is likely the best-studied insect order, although with strong geographic and taxonomic biases. Families with large species are generally better known than those dominated by small species, and species-rich tropical faunas are generally poorly investigated compared to those in temperate regions (Lees et al. 2013). At a global scale, Lepidoptera is the best represented order in the International Barcode of Life Project (iBOL) with approximately 1 million sequences on BOLD associated with 100,000 species. By way of context, 9846 lepidopteran species have been reported from Europe up to 2011 (Karsholt and van Nieukerken 2013).

Material Collection and Delimitation of Sampling Area

The data used in this study were largely collected within the framework of the iBOL as part of national or regional initiatives such as the Fauna Bavarica project (http://www.faunabavarica.de, last accessed June 2, 2016), the Finnish Barcode of Life project (http://www.finbol.org/, last accessed June 2, 2016), the "Lepidoptera of the Alps" campaign, the Norwegian Barcode of Life (http://www.norbol.org/, last accessed June 2, 2016) and the Nature of The Netherlands project, or as individual research projects.

Our analysis focused on specimens from Europe as defined by current political boundaries but excluding European Turkey, Cyprus, and most of the Macaronesian islands. Most tissue samples analyzed were from identified pinned specimens collected in the past 15 years, because older material was less likely to generate sequence data. Altogether 41,583 specimens, representing 4977 species, yielded a DNA sequence of over 500 base pairs (bp) in length. Specimens with shorter sequences were excluded from the analyses. From one to 146 specimens were sampled per species with an average of 8.4. Among this total, 697 species (14%) were represented by a single DNA barcode: the so-called singletons. Although singletons cannot show non-monophyly themselves, they were included in the analyses because they may render other species para- or polyphyletic by becoming "entangled" with them. Almost all sequenced species were included, several of them from species or species groups already known to be subject to species delimitation and identification difficulties. We excluded some specimens that could not be associated with any described species, but included many species likely or possibly encompassing cryptic species. In one case (the *Stigmella salicis* group, Nepticulidae), we included undescribed species and applied interim names, because in this case the presence of several species has been convincingly demonstrated (van Nieukerken et al. 2012). The only species group not wholly included is the *Dahlica*/*Siederia* complex of species (Psychidae), because the current taxonomy of this group is known to be largely inaccurate, parthenogenesis is frequent, and morphological characters for many species are misleading (Elzinga et al. 2014), rendering identification of species currently impossible by anything other than molecular data. Sampling was geographically somewhat biased, with 60% of the specimens collected in only 10 of the 51 European countries: Finland (9619), Germany (7922), Italy (4829), United Kingdom (4005), Austria (3184), France (2774), Spain (1440), Romania (1384), the Netherlands (1167), and Norway (868). Full taxonomic and collection information of specimens is available in BOLD through

individual specimen pages within the public data set DS-MARKALL (dx.doi.org/10.5883/DS-MARKALL) in the BOLD (www.boldsystems.org) barcode data repository. Collection localities are also available in .klm format (viewable with Google Earth) on Dryad at http://dx.doi.org/10.5061/dryad.k3mr1.

Sequencing

A 500–658 bp long amplicon of the 5′ terminus of the mitochondrial COI gene (the standard DNA barcode region for animals) was sequenced for all specimens. A single codon deletion occurs in three species of Scardiinae (Tineidae), but otherwise the target gene region does not show length variation in European Lepidoptera. Sequencing was predominantly conducted at the Canadian Centre for DNA Barcoding (CCDB), but also at Naturalis Biodiversity Center (the Netherlands) and laboratories of the authors' research organizations. The CCDB's sequencing protocol is described in detail in deWaard et al. (2008). The primer pair LepF1 and LepF2 (Brower and Egan 1997) was primarily used to amplify the barcode region, but, in cases of failure, other primer sets were also attempted. Full primer details, laboratory reports, trace files, sequences, and GenBank accession numbers can be retrieved from the sequence page of each record in BOLD and can be downloaded at dx.doi.org/10.5883/DS-MARKALL.

Verifying Identifications and Taxonomic Names

Although specimens were generally identified to species level by taxonomic experts based on morphology prior to sequencing, the resulting DNA barcodes provided an efficient way to cross-check the identifications. Since misidentifications easily produce false cases of non-monophyly, we carefully examined all anomalous cases. This necessitated at least the superficial re-examination of voucher specimens, but in many occasions also the dissection of their genitalia, whose morphology often carries important diagnostic features. This process revealed many misidentifications, which were corrected in BOLD prior to final analyses of species-level monophyly. Clerical errors and the application of different nomenclatures can similarly lead to false observations of non-monophyly, especially when performed using automated detection of non-monophyly as done here. We, therefore, harmonized names throughout the complete data set following the nomenclature of Fauna Europaea (http://www.fauna-eu.org/, last accessed June 2, 2016). This revealed hundreds of cases where two or more names had been applied to a single species, but also cases of homonymy, (the application of a single name to several species). Despite careful cross-checking of identifications, it is likely that some misidentifications remain in the data because of identification problems or taxonomic uncertainties in several species groups. We attempt to estimate the effect of this below.

Detection of Contamination, NUMTs and Chimeric Sequences

Prior to analysis, several validation steps were performed to increase the reliability of the results. First, Sanger sequencing trace electropherograms were reviewed for quality, excising sequences associated with a mean trace quality "phred" score below 30 and where more than 10% of the bases showed a quality score below 20 after trimming of the primer sequences. Sequences that met these quality criteria were reviewed to excise those that are likely pseudogenes (NUMTs) or chimeric in origin. Pseudogenes were detected by comparing each sequence to a Hidden Markov Model (Eddy 1998) of the COI protein (Finn et al. 2010). Low-scoring sequences contained either unusual amino acid substitutions, stop codons or reading frame shifts, all indicators of pseudogenization. Tests for chimeras involved dividing sequences into 100 bp fragments with each fragment independently searched against the barcode reference library. Resulting hits were compared to ensure that all fragments match similar reference records in the library. Sequences failing this test were manually evaluated and discarded if a chimeric origin was confirmed. Finally, sequences were compared against a reference library of common laboratory contaminants, discarding those that matched.

Phylogenetic Analyses

Distance-based NJ and optimality criterion-based ML phylogenetic methods were used to reconstruct DNA barcode gene trees. These methods are capable of analyzing large (>5000 sequences) data sets and were used to estimate the effect of the inference method on the incidence of paraphyly and polyphyly. In general, ML is expected to yield a more correct tree topology because of limitations inherent in the NJ method. These include especially the sensitivity of the method to the input order of specimens and the correctness of the distance matrix (Huelsenbeck and Hillis 1993; Farris et al. 1996). Despite these problems, NJ has repeatedly been shown to perform well for species delimitation and to approximate phylogenetic relationships (Huelsenbeck and Hillis 1993; Kumar and Gadagkar 2000; Mihaescu et al. 2009).

Since NJ is computationally less demanding than ML and permits rapid construction of trees with thousands of specimens, NJ trees were constructed without the exclusion of redundant (identical) haplotypes, which is expected to have minimal effect on the tree topology estimated by this method. In contrast, haplotype collapsing was done for the more demanding ML analyses to increase computational efficiency. However, redundant haplotypes were not removed when they occurred between different species (barcode-sharing). Haplotype collapsing was conducted using ALTER (Glez-Peña et al. 2010).

Distance matrices for NJ trees were calculated under both the Kimura 2-parameter (K2P) (Kimura 1980) and P-distance model, using the BOLD alignment of

sequences (amino acid based HMM). To estimate non-monophyly in NJ trees, we applied K2P because it is usually used for DNA barcode data, although it is not necessarily the best-fit model of nucleotide evolution of the COI gene (Srivathsan and Meier 2012). Trees generated with P-distance showed very similar topologies. Trials were conducted to estimate the effect of different nucleotide substitution models available in BOLD, but no effect on incidence of non-monophyly was detected (results not shown). Trees were rooted on a specimen representing a sister group (where known) or a closely related group as based on recent comprehensive Lepidoptera phylogenies (Mutanen et al. 2010; R Core Team 2013). Analyses were performed mostly per family or by grouping several related families, but due to large numbers of specimens (exceeding the maximum number permitted by BOLD), by subfamilies in Geometridae and by separating Noctuinae from the rest of Noctuidae. This partitioning is unlikely to lead to any case of non-monophyly remaining undetected, because non-monophyletic species are only in exceptional cases tangled outside a single genus (we have never observed this in our data). A few families or subfamilies (Bedelliidae, Urodidae, Schreckensteiniidae, Heterogyniidae, Riodinidae, Thyrididae, Orthostixinae in Geometridae) include only a single species in this study and our analysis thus makes them by definition monophyletic. But all seven species have highly divergent barcodes and would remain monophyletic, however, treated.

ML trees were constructed using RAxML v. 8 (Stamatakis 2014) via the Black Box web server (http://embnet.vital-it.ch/raxml-bb/index.php, last accessed June 2, 2016). The analyses were conducted under the GTR+G model of nucleotide evolution (Tavaré 1986). Node support values were estimated with 100 bootstrap replicates. Analyses were mostly performed using the division applied in NJ analyses, except that families with very few specimens were combined in three groups, the first including two non-ditrysian families; the second the ditrysian families excluding the non-macroheteroceran families except for Riodinidae (a single monophyletic species); and the third Macroheterocera plus Riodinidae. This division is phylogeny based except for the placement of Riodinidae, which is not currently included in Macroheterocera. All trees were saved in Newick format for the detection of monophyly and are deposited in Dryad.

Detection of Non-Monophyly

Non-monophyly can be detected by eye in a graphical representation but is prone to human error. In trees with hundreds or thousands of terminals, internal branches are often short and detection by eye can be very difficult. Also, polyphyletic species dispersed among many other species might remain undetected. For these reasons, we developed a web service called "Monophylizer" that detects cases of non-monophyly. The service accepts Newick, Nexus, NeXML, and PhyloXML trees. The Monophylizer was designed to be rather permissive in the Newick syntax it allows because BOLD can omit syntactically invalid Newick tree descriptions. However, some of the database fields that BOLD includes allow text fields that may contain parentheses or commas, which file readers cannot distinguish from the commas and parentheses used by the Newick syntax. These must be avoided. Trees are parsed by the service using the Bio::Phylo toolkit (Vos et al. 2011), which can accept many tree format "dialects," including most of the idiosyncrasies produced by BOLD.

Before proceeding, the web service applies an auto-incrementing integer index to each node both in a pre- and a postorder traversal. In a preorder tree traversal parent nodes are processed before their children, whereas in postorder children are processed before their parents. Thus, in this indexing scheme, each node is assigned the value of the incrementing index both before and after visiting its children, such that the tree ((A,B),C); is indexed as ((A{3.4},B{5.6}){2.7},C{8.9}){1.10}; if we signify the pre- and postorder node index as, respectively, the first and second integer of each statement between braces. The web service assesses monophyly using the following algorithm, which is applied to all distinct species in the tree:

1. Based on the species name, all leaf nodes that belong to the focal species are collected.
2. The most recent common ancestor (MRCA) of the collected leaf nodes is identified.
3. All leaf nodes subtended by the identified MRCA are collected.
4. If this set is the same as the set of leaf nodes in step 1, the species is monophyletic. If not, continue to step 5.
5. All internal nodes in the tree that subtend leaf nodes from the focal species as well as at least one other species are collected and sorted by their postorder index.
6. The collected, sorted internal nodes from step 5 are grouped into distinct root-to-tip paths. Internal nodes that are nested in each other are identified (and collected in the same group) by checking that the preorder index of the focal node is larger, and the postorder index of the focal node is smaller than that of the next node.
7. If there is more than one distinct root-to-tip path (i.e., group), the taxon is considered polyphyletic, otherwise paraphyletic.
8. For each first (i.e., most recent) node in each group, all subtended species are collected. The union of these sets across groups forms the set of entangled species.

The web service can be accessed at http://monophylizer.naturalis.nl/, last accessed June 2, 2016 and the

source code is freely available at https://github.com/naturalis/monophylizer, last accessed June 2, 2016. The output of the web service can be configured to be either a table in a web page, or a tab-separated spreadsheet for high-throughput applications, for example, when combined with automated web clients. We used this web service to analyze the topologies of our gene trees.

Estimating the Effects of Sampling Effort and Intra- and Interspecific Divergence

The frequency of observed species monophyly is strongly influenced by intraspecific genetic variation represented in the data and the genetic distance to related species. Both measures are affected by sampling intensity. Higher sampling effort will reveal more intraspecific variation and will tend to identify more closely related species. We explored the effect of both these factors and their interaction on the frequency of monophyly as determined by Monophylizer. The measures we used of maximum intraspecific divergence and minimum genetic distance to the nearest neighbor were based on the K2P model of nucleotide substitution and were calculated using the "barcode gap analysis" tool of BOLD using pairwise deletion setting for missing nucleotides. Sequences were aligned using the BOLD sequence aligner (amino acid based HMM). The analysis was carried out with a slightly reduced data set of 4921 species (56 species excluded) since Barcode Gap Analysis currently treats records with infraspecific names as different species.

We analyzed the occurrence of non-monophyly by fitting a generalized linear model using the function "glm" in R 3.0.0 (R Core Team 2013). Species monophyly versus non-monophyly was treated as a binary response variable while the explanatory variables were distance to the nearest neighbor, maximum intraspecific genetic variation and the number of specimens analyzed. All interactions among the explanatory variables were included in the model. We assumed a binomial error distribution and used a logistic link function. For visualization and inferences, the fitted values of the model were transformed to probabilities by using the inverse of the link function.

Some fitted values of the above model were either zero or one, which results in problems in applying the Wald approximation used in deriving *P*-values for parameters of the model (Venables and Ripley 2002). Therefore, we performed a permutation test to derive empirical *P*-values for the model parameters. We randomly reordered the observations 10,000 times and fitted the statistical model described above to each permutated data set. An empirical *P*-value can be calculated by comparing the estimate derived from the true data to the distribution of estimates produced by permutation.

To assess the predictive power of the statistical model, we performed a cross-validation analysis at the family level: each of the 71 families was, in turn, used to test the model fitted to the data on the remaining 70 families. We chose this approach because the incidence of non-monophyly may vary among families, potentially biasing the predictive power of the model toward a subset of families. The overall performance of the model was then assessed by comparing the predictions to observations with the area under a Receiver Operating Characteristic (ROC) curve method [function "AUC" (LeDell et al. 2014)].

Statistical Relationship between Sampling Effort and Non-Monophyly

Because a limited and variable number of specimens of each species was available for analysis, and because sampling is nearly always incomplete geographically, the observed numbers of para- and polyphyletic species underestimate the actual frequency of species-level non-monophyly. We explored sampling bias by dividing the data set into classes, each class containing all the species with equal numbers of specimens. We included in the statistical analysis only those 35 classes that included at least 10 species, and calculated the proportion of non-monophyletic species for each class. Standard errors were placed on these estimates using bootstrap resampling (with 5000 replicates) with the R function "boot" (Canty and Ripley 2013). We fitted a nonlinear function,

$$y = \gamma(1 - e^{-(x-1)e^{\delta}}) \quad (1)$$

to the data with the R function "nls." This function is constrained to go through the point (1,0) hence implicitly assuming the absence of non-monophyly when only a single specimen is available. The starting values for γ and δ were the estimated values of α and β from the model $y = \alpha(1 - e^{-x \cdot \exp(\beta)})$, fitted with the self-starting function "SSasympOrig." When estimating the parameters of model 1, we weighted the observed proportions of non-monophyletic species by the inverses of their bootstrap standard errors (the weight for the class including single specimens was arbitrarily set to 200 as the standard error is zero though this choice did not affect the parameter estimates). Finally, we derived 95% confidence intervals for γ and δ by using the R function "confint" (Venables and Ripley 2002)

Barcode-Sharing

If a species shares its barcode with another species, and both show no intraspecific variation, they are treated as monophyletic by Monophylizer. However, these species could equally be considered para- or polyphyletic. Furthermore, such species pairs would no longer be reciprocally monophyletic if even a single nucleotide substitution were to occur. We investigated the frequency of this phenomenon by searching for species that showed sequence identity to their nearest neighbor but had been ranked monophyletic. As the same issues can arise when

neighboring species are very similar but not identical, we also searched for monophyletic species differing by less than 1% from their nearest neighbor. Searches were performed using the "barcode gap analysis" tool in BOLD.

Estimation of Taxonomic Uncertainty and Misidentifications

Taxonomic inaccuracy and misidentifications are likely to yield many "false positives" cases of non-monophyly due to the incorrect assignment of a specimen to a species. Although we carefully cross-checked identifications of doubtful records prior to the analyses, there are many species groups where unclear morphological limits among species can produce misidentifications. A more significant effect is the likely inaccuracy of the taxonomy itself in many groups. To avoid circular reasoning we did not remove such groups from our data (with the exception of the psychid *Dahlica/Siederia* group where the inadequacy of taxonomy is widely acknowledged), but instead attempted to estimate the magnitude of this effect. As the authors include many of Europe's leading experts on Lepidoptera, this was done by asking the relevant specialist to categorize each non-monophyletic species as "species identification straightforward" or "species identification problematic," and separately as "species limits well-defined" or "species limits poorly defined." We also used expert judgment to assess the occurrence of potential cryptic species in the data. While we acknowledge that assessing these effects involves some subjectivity, we find their impact potentially significant. In making their assessments the taxon specialists were asked to be conservative and include only the most obvious cases of synonymy and misidentifications, while the presence of potential cryptic species was only accepted when additional independent evidence, such as morphological or ecological differences, supported the genetic differences (thus we excluded cases of deep intraspecific barcode splits lacking further evidence that cryptic species may be involved).

Estimating Effects of Allopatry and Parapatry

Estimating the effect of geography is especially challenging because geographic information is often used to delimit different species. This is particularly problematic when a species is composed of spatially isolated populations, which will show additional structure in their degree of genetic differentiation; their allocation to species is both subjective and depends on the species concept employed (Mutanen et al. 2012). As pointed out by McKay and Zink (2010), where such species clusters involve paraphyly, it could simply be eliminated by elevating allopatric populations to valid species. We estimated the effect of allopatry on the incidence of non-monophyly, an exercise that was greatly facilitated by the distributional data for European Lepidoptera, which is superb in comparison with any other diverse invertebrate group or faunal region.

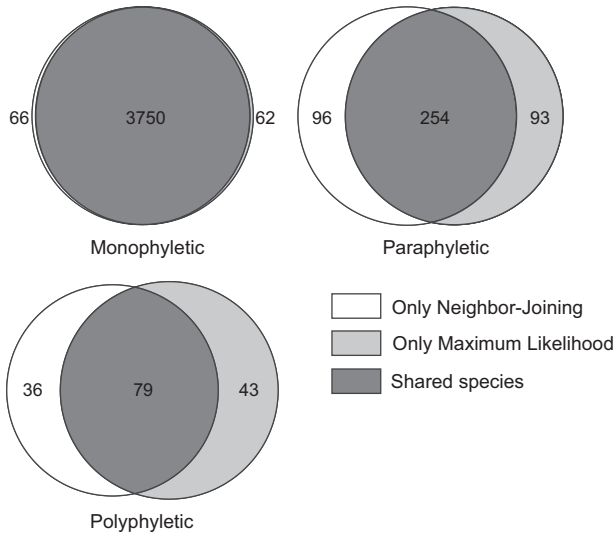

FIGURE 2. Overlap in species classified as mono-, para-, and polyphyletic using either NJ or ML methods. The number of species is indicated in each partition (the counts for monophyly exclude species represented by singletons).

RESULTS

Incidence of Non-Monophyly in NJ and ML Trees

NJ and ML methods resulted in almost equal estimates of the number of non-monophyletic species, though the two sets did not completely overlap. NJ found 465 (350 para- and 115 polyphyletic) and ML 469 (347 para- and 122 polyphyletic) non-monophyletic species. This equates to 12.2% and 12.3% incidence of non-monophyly in NJ and ML trees, respectively, with singleton species excluded from both calculations. Altogether 531 species were classed as non-monophyletic by at least one method, and of these 24.1% were identified by just one. 57.3% of paraphyletic species were identified by both methods, while the equivalent percentage for polyphyletic species was 50.0% (Fig. 2). Thus the tree-building method made a substantial difference to whether a species was classed as non-monophyletic.

Non-monophyletic species unique to ML show both a higher average intraspecific K2P variability (ML: mean = 1.93, 95% adjusted bootstrap percentile [BCa] confidence interval [CI] = 1.50–2.52; NJ: mean 1.29, 95% BCa CI = 0.882–1.905) and greater average minimum genetic distance to their nearest neighbor (ML: mean = 1.79, 95% BCa CI = 1.45–2.22; NJ: mean = 0.95, 95% BCa CI = 0.64–1.40). Only three non-monophyletic species unique to ML fully shared their barcode sequence (0.0% K2P distance) with their nearest neighbor, whereas with NJ this occurred in 29 species. A closer investigation of these cases showed that the difference is largely due to the tendency of the NJ method to place sequences that are identical except for length at slightly different nodes, a known pitfall of this method. As NJ, however, yielded fewer cases of non-monophyly (465 in NJ vs. 469 in ML) and several of these cases were due to the presence of haplotypes identical except for sequence

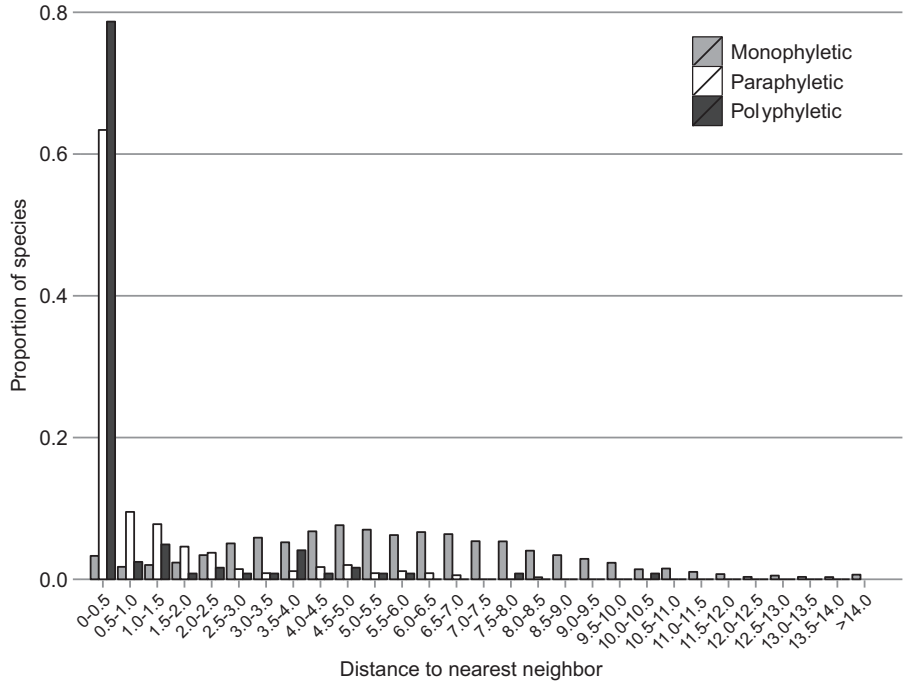

FIGURE 3. Proportions of species with different minimum K2P distances to their nearest neighbor in mono-, para-, and polyphyletic species. For monophyletic species, singletons were excluded.

length, ML seems to identify more species as non-monophyletic. Non-monophyletic species discovered only by ML have higher mean intraspecific variation because many of these species represent taxa with deep intraspecific splits, suggesting that ML recovers such species more frequently as non-monophyletic than NJ.

In 97.9% of species classed as non-monophyletic by ML, the polyphyly or paraphyly was due to one or more species from the same genus, but seven (1.5%) involved moths in closely related genera. Three of the latter cases involved a pair of genera that are currently being proposed for synonymy (*Sciadia* and *Elophos*, Geometridae, Ennominae—see also Huemer and Hausmann 2009), whereas a fourth involves two genera (*Crombrugghia* and *Oxyptilus*, Pterophoridae) separated by such minor characters that they are not accepted by all authorities (Kullberg et al. 2001). Three species (0.6%) showed phylogenetic "tangles" involving 5 (*Glacies coracina*, Geometridae), 36 (*Dryobotodes monochroma*, Noctuidae), and 71 (*Deltote incognita*, Noctuidae) genera, but these were very likely misidentification artifacts (see Supplementary Table 1 available on Dryad). Non-monophyly was not observed among species in different subfamilies or higher-level ranks.

Para- and polyphyly are presumed to be more prevalent in young, recently diverged species than in older species. Although mitochondrial introgression may obscure the gene history, genetic distance to the closest relative is likely a good proxy for species' coalescence time. We found that non-monophyletic species often showed a low genetic distance to their nearest neighbor, and that this pattern was more pronounced in polyphyletic than paraphyletic species, although the difference was not statistically significant. Monophyletic species show an average K2P minimum distance of 5.66 (95% BCa CI 5.57–5.76) to their nearest neighbor, whereas the corresponding values for paraphyletic and polyphyletic species are 0.89 (95% BCa CI 0.751.07) and 0.74 (95% BCa CI 0.50–1.11), respectively. Singletons were excluded as they cannot exhibit non-monophyly and a further 56 monophyletic species were omitted because their nearest neighbor is a subspecies of the same species in the BOLD database. Of monophyletic species (singletons excluded), 7.0% show less than 0.5% minimum K2P distance to their nearest neighbor, whereas the equivalent numbers for paraphyletic and polyphyletic species are 63.4% and 78.7% (Fig. 3). However, many monophyletic species have low genetic distances to their nearest neighbor, including those represented by singletons which could become non-monophyletic with increased sampling (Fig. 4). In both para- and polyphyletic species, low K2P distance to the nearest neighbor is frequently caused by one or more operational factors, such as overlooked synonymies or potential cryptic species (Supplementary Table 1 available on Dryad).

High intraspecific variability is associated with the presence of non-monophyly. Of 3807 monophyletic species represented by more than one individual, the mean K2P intraspecific maximum variability is 0.99% (95% BCa CI 0.95–1.04; mean $n = 8.99$). In paraphyletic species, the mean maximum intraspecific variation is

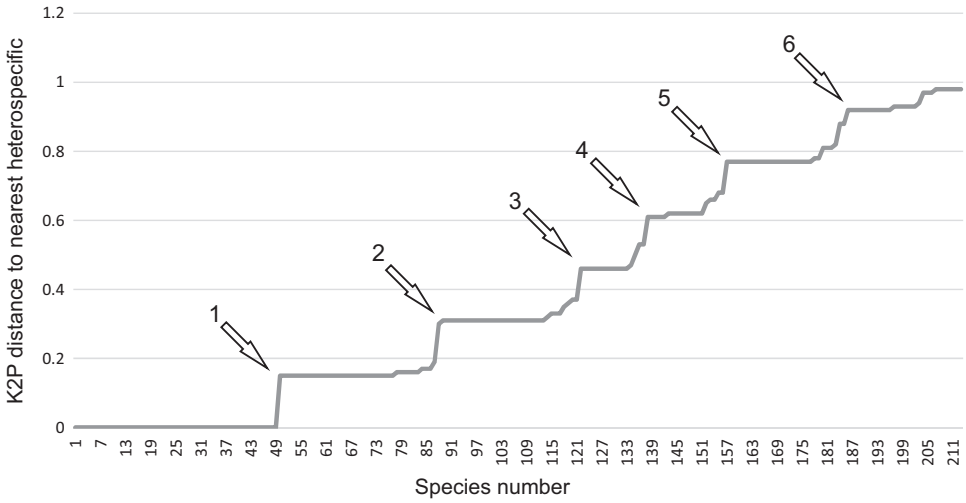

FIGURE 4. Monophyletic species (215 in total) showing less than 0.01 minimum K2P distance (or less than 7 nucleotide substitutions difference) to their closest species. The number of nucleotide substitutions to the nearest neighbors are indicated with arrows. The curve is not cleanly stepped because of slight variation in sequence lengths and because the substitution model employed does not assume equal likelihoods of all substitutions. Forty-eight species having K2P divergence of zero to the closest heterospecific would be rendered non-monophyletic by a single nucleotide substitution.

2.37% (95% BCa CI 2.14–2.62; mean $n = 14.58$) and in polyphyletic species it rises to 3.26% (95% BCa CI 2.77–3.87; mean $n = 12.78$). These differences remain after controlling for sampling effort across categories by dividing the mean maximum intraspecific variation for each species by the number of specimens (monophyletic: mean = 0.14, 95% BCa CI 0.13–0.15; paraphyletic: mean = 0.26, 95% BCa CI 0.23–0.30; polyphyletic: mean = 0.53, 95% BCa CI 0.40–0.74).

The remaining analyses are based only on results obtained through ML. Identical haplotypes were not considered except in assessing the effects of sampling bias.

Effects of Intra- and Interspecific Divergence and Sampling Effort

The results of the statistical analysis of the factors affecting the probability of monophyly were rather complex as there was a significant three-way interaction between the distance to the nearest neighbor, the maximum intraspecific genetic variation and the number of specimens analyzed (Table 1). This interaction is illustrated by plotting the fitted regression surfaces involving the distance to the nearest neighbor and the maximum intraspecific genetic variation for different numbers of specimens analyzed (Fig. 5). The probability of non-monophyly increases very steeply from zero to one at a threshold whose shape and location depends on the number of specimens analyzed. The statistical model predicts non-monophyly to be most likely when the distance to the nearest neighbor is small and there is considerable intraspecific genetic variation. As the number of specimens increases, non-monophyly is predicted to occur in an increasing region of parameter space (Fig. 5). The statistical model performed very well according to the cross-validation analysis, which resulted in an area under the ROC curve of 0.988 (Supplementary Fig. 1 available on Dryad). This result was not sensitive to the "leave-one-family-out" method as a 5-fold cross-validation analysis resulted in the same mean area under the ROC curve [0.988; 95% CI estimated with the function "ci.cvAUC" (LeDell et al. 2014): 0.984–0.992) when the random division of species into the five groups was repeated 100 times. The proportion of non-monophyletic species tends to increase as more specimens are included in the analysis (Fig. 6). The asymptote of the regression function fitted to these data indicates that the proportion of non-monophyletic species approaches 0.23 (95% confidence interval: 0.16–0.48) as the number of specimens analyzed becomes large. Hence, increasing sampling is expected to increase the observed level of non-monophyly in European Lepidoptera by between 16% and 48%.

Effect of Taxonomic and Identification Uncertainty

The taxonomic specialists among the authors estimate that 31.8% of non-monophyly may not be valid, as they are likely to reflect "over-splitting" of species (Supplementary Table 1 available on Dryad). These include (i) cases where highly similar but allopatric populations have been considered distinct species, (ii) parapatric species pairs with little information on the extent of gene flow between populations, (iii) ecological (e.g., altitudinal, latitudinal, habitat, or food-plant associated) forms or potentially polymorphic species, and (iv) sympatric pairs or groups of variable species separated by uncertain boundaries.

In 15.1% of non-monophyletic species, there was further independent evidence of cryptic diversity

TABLE 1. Parameter estimates from a binomial generalized linear model (with a logistic link function) explaining the probability of non-monophyly

Parameter	Estimate	Std. Error	z	P-value	Empirical P-value
Intercept	−0.354	0.190	−1.86	0.063	< 0.0001
Dist. to NN	−2.22	0.189	−11.8	< 0.0001	< 0.0001
Intraspec. var.	2.10	0.177	11.9	< 0.0001	< 0.0001
Specimens	0.0417	0.0148	2.82	0.0048	0.00020
Dist. to NN × intraspec. var.	−0.0413	0.0258	−1.60	0.11	0.0010
Dist. to NN × specimens	−0.0117	0.0109	−1.079	0.28	0.00010
Intraspec. var. × specimens	−0.0257	0.00672	−3.83	0.00013	0.00030
Dist. to NN × intraspec. var. × specimens	0.00464	0.00150	3.10	0.0019	< 0.0001

Notes: The explanatory variables were the genetic distance to the nearest neighbor species (dist. to NN), maximum intraspecific K2P variation (intraspec. var.), and the number of specimens analyzed (specimens). Empirical P-values were derived from a permutation test (see text for details) because some fitted probabilities were numerically either zero or one, which may result in overestimated P-values when using the Wald approximation.

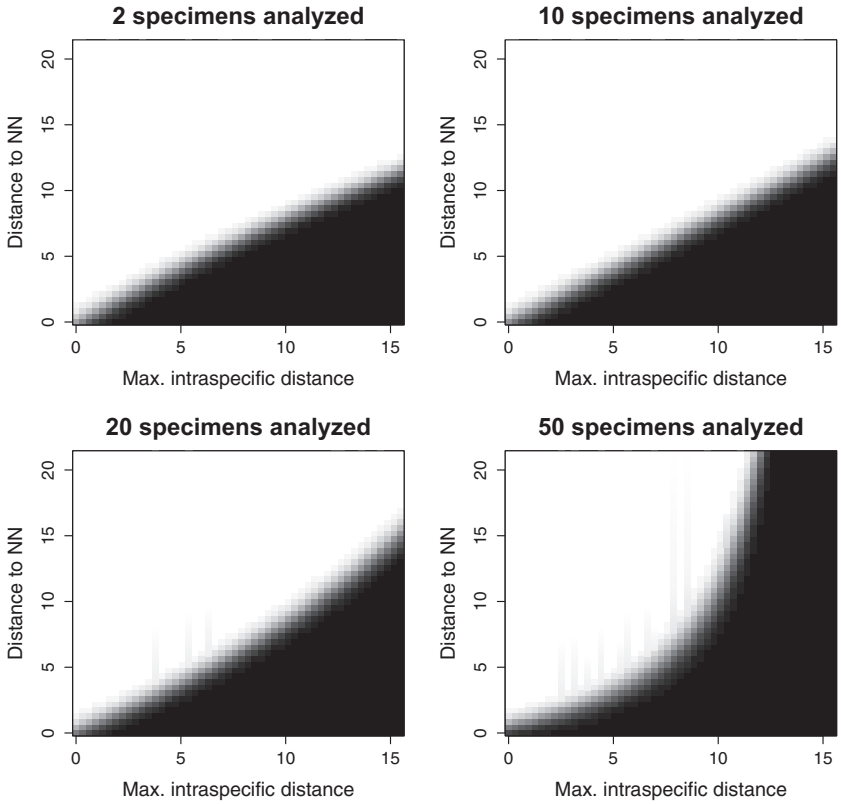

FIGURE 5. The estimated probability of species non-monophyly from a generalized linear model including as explanatory factors: genetic distance (in percent with 1% distance equaling to 0.01 K2P divergence) to nearest neighbor (vertical axes), maximum within-species K2P genetic distance (horizontal axis) and the number of specimens included in the analysis (the figures show four values). Probability values are indicated by a grayscale gradient with black = 1 and white = 0.

(undersplitting), in most cases from morphological differences associated with barcode divergence. Of these, 78.9% involve sympatric splits, indicating that many cases are likely to represent reproductively isolated, but morphologically similar species pairs or groups. Several of these cases are undergoing taxonomic revision using an integrative approach, but with a single exception (*Stigmella salici*s group, see above) we followed currently recognized taxonomic boundaries.

In 31.6% of non-monophyletic species, we identified problems with species identification, despite our initial careful validation. In many but not all cases, these difficulties were associated with possible cases of oversplitting. In some cases, such as in *Yponomeuta* (Yponomeutidae), the problems arose because reliable identifications require larval characters and are much harder for adults. Often, but not always, the difficulty in identifying species is linked to the likely presence of cryptic species or oversplit species.

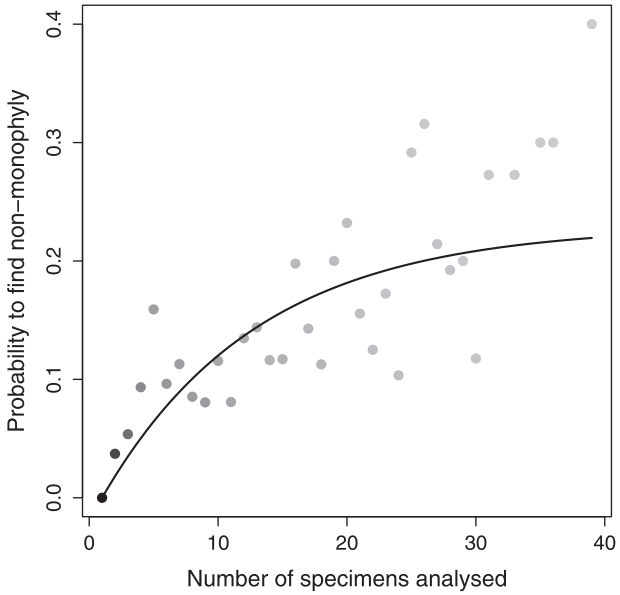

FIGURE 6. Probability of finding non-monophyly as a function of the number of specimens per species included in the analysis. Points indicate proportions of non-monophyletic species in groups of species with equal number of analyzed specimens. The darkness of points indicates weights (inverses of bootstrap standard errors) used in fitting the regression curve, darker colors indicating higher weights. The curve $[y = 0.23 \times (1 - e^{-(x-1)e^{-2.5}})]$ is fitted with nonlinear asymptotic regression (see text for details).

Altogether 58.6% of non-monophyletic species were estimated as potentially being affected at least by one taxonomic uncertainty (undersplitting, oversplitting, or identification difficulties). In fact, 19.9% of species were estimated as being affected by more than one taxonomic issue.

Effect of Geography-Related Patterns among Species

Altogether, 78.7% of the non-monophyletic species are sympatric with at least one of the species responsible for polyphyly or paraphyly. An entirely allopatric relationship was detected in 14.5% of species and parapatry in 5.8% of species. Of the species suspected of being non-monophyletic due to oversplitting, at least one of the species responsible was sympatric in 61.5% of the cases, allopatric in 21.5% of the cases and parapatric in 14.8% of the cases. In four cases, the geographic relationships between species are uncertain due to poor distributional data.

DISCUSSION

Our survey examined 41,583 specimens belonging to 4977 lepidopteran species, the largest data set of DNA barcodes so far published for any taxonomic group. Moreover, our data are of higher quality than previously published arthropod data sets due to the efforts made to resolve issues such as uncertainties in identification. For example, the benchmark review of species non-monophyly (Funk and Omland 2003) included 2319 species, whereas Ross (2014) examined 21,337 species. Both of these studies examined a broad range of taxa, but relied largely upon unpublished and poorly validated data, without examining the reliability of the identifications.

Funk and Omland (2003) reported the occurrence of non-monophyly to be 26.5% in arthropods (figure for Lepidoptera not estimated separately), whereas Ross (2014) reported an 18% non-monophyly in Arthropoda and 17% in Lepidoptera (16% with interim operational names included). We observed a considerably lower level of 12.3% of non-monophyly in our "raw" data. This is not explained by differences in sampling effort, since Ross sampled an average of 8.1 specimens per species among Arthropoda, whereas our average was 8.4. Ross' Lepidoptera data are based partially on the same cleaned data used in our study, indicating that the level of non-monophyly in data not used in our study was likely higher than 17%. Since Funk and Omland's data were retrieved from GenBank while Ross's data came from BOLD, the higher incidence may be due to the higher proportion of misidentifications in GenBank (Harris 2003; Bidartondo 2008; Groenenberg et al. 2011). A second possible cause is that before the advent of systematic species sampling through the iBOL, data sets were biased because of more concentrated efforts on difficult species groups, which in turn are likely to express unusually high levels of non-monophyly. We also believe that the lower level of non-monophyly in our study compared with that by Ross reflects our considerable effort in data validation. Our data, however, are certainly not free of identification errors. The true incidence of non-monophyly is further exaggerated by several other operational factors, but attenuated by limited sampling in many species. We attempted to assess the significance of all these factors in our study.

Our results are remarkably congruent with a detailed study of paraphyly on birds (McKay and Zink 2010). Funk and Omland (2003) found that 16.7% of birds exhibited non-monophyly, a value close to the 14.3% reported by McKay and Zink (2010). However, a detailed examination of 856 species by the latter authors revealed that 55.7% of these cases were due to incorrect taxonomy, a value close to our estimate of 58.6% of cases generated by taxonomic issues in our Lepidoptera data set.

Two genetic factors play a crucial role in influencing the likelihood of species monophyly: the degree of intraspecific variation and the extent of divergence between species. Our data clearly demonstrate that relatively high intraspecific variation and relatively low genetic distance from the closest species characterize virtually all non-monophyletic species, and that where both occur the species concerned is nearly always non-monophyletic. Monophyletic species have, on average, nearly six times higher genetic distances to their nearest neighbors than paraphyletic species and

almost seven times higher than polyphyletic species. Similarly, their intraspecific variation is less than half that of paraphyletic species and less than one-third of that of polyphyletic species. The statistical analysis indicates that the probability of non-monophyly increases practically as a step function from zero to one with increasing intraspecific genetic variation and decreasing distance to the nearest neighbor, and that non-monophyly becomes more likely with increasing sampling effort.

Since sampling of insect populations is never complete, only a fraction of the total genetic diversity is represented in any data set. Both the study by Ross (2014) and our work shows a clear positive correlation between sampling effort and the incidence of non-monophyly. Previous studies, with the exception of that of McKay and Zink (2010) in birds, did not attempt to estimate the actual level of non-monophyly as we did. However, many species in our study have distributions that extend beyond Europe and sampling across their entire range will likely reveal additional genetic variation that might affect estimates of non-monophyly. Based on our data, we estimated that the actual level of non-monophyly in European Lepidoptera would be about 23% (95% confidence interval 16–48%) without considering the impact of operational factors (see below). The point estimate is slightly less than the 26.5% reported by Funk and Omland (2003), although their value falls within the confidence limits of our estimate.

We detected identification issues in 31.6% of non-monophyletic species. Cases of oversplitting are strikingly frequent in our data, since we estimated that up to 31.8% of all non-monophyletic species may represent "false species." The taxonomic issues affecting the distribution of non-monophyly in our study occur for two main interconnected reasons. First, lepidopteran taxonomy has a long tradition in Europe, where the fauna is well investigated in many areas, leading to the situation in which "taxonomic resolution" eventually gets very fine in those groups that have been studied by many workers. While this effort undoubtedly helps to reveal many cryptic species, the side effect is that species that fail to match standard criteria are more likely to be considered as valid. This is exemplified by the many cases of allopatric populations of European Lepidoptera that over time and with increased taxonomic scrutiny have been accorded species status (Mutanen et al. 2012). Second, taxonomic tradition favors species splitting at the expense of species lumping. This is exemplified by *Euxoa tritici* (Noctuidae), which was split into three species in a non-peer-reviewed revision of the group (Fibiger 1997). Despite subsequent morphometric studies indicating broad overlap and the poor performance of the diagnostic characters separating the proposed species (Mutanen 2005), many checklists still list them as distinct (DNA barcodes also do not support the presence of more than one species). The International Code of Zoological Nomenclature (ICZN 1999) does not require peer review for a new name to be valid, while synonymization of species typically requires thorough studies, which then have to become generally accepted by the taxonomic community. This situation leads to the gradual accumulation of poorly justified species over time.

We estimated that 14.9% of all non-monophyletic species in our study are actually a species pair or group. This estimate is likely conservative since we included only cases with independent evidence aside from DNA barcodes supporting the presence of cryptic species. In several cases, the description of one or several new species is underway (e.g., Huemer and Mutanen 2015). It is likely that many more cases with deep intraspecific splits will eventually be revealed as species complexes since DNA barcodes are very effective in revealing potential cryptic species (Hausmann et al. 2009a, 2009b, 2013; Dincă et al. 2011, 2015; Huemer and Hebert 2011; Huemer et al. 2013, 2014; Mutanen et al. 2013; Huemer and Timossi 2014; Kirichenko et al. 2015). On the other hand, some other studies have not found evidence of cryptic species among taxa showing unusually high intraspecific barcode variation (Webb et al. 2011; Hogner et al. 2012; Kvie et al. 2013;.

The assignment of allopatric populations to species in a standardized way is one of the most difficult challenges of alpha taxonomy (Mutanen et al. 2012). Geographic distance often leads to a breakdown of gene flow between populations, resulting in the gradual differentiation of populations over time and eventually speciation. Under these circumstances, the taxonomic delimitation of populations is inherently subjective and greatly affected by the underlying species concept. Of the non-monophyletic species in this study, 14.5% involved two or more allopatric species and 5.8% one or more parapatric species. Many other non-monophyletic species are allopatric or parapatric in relation to some, but not all, associated species. This suggests that issues related to the geographic relationships of the species and resulting taxonomic difficulties play a significant role in species poly- and paraphyly.

Altogether 58.6% of non-monophyletic species detected in this study are likely to be due to methodological rather than biological causes. Thus, the observed level of 12.3% non-monophyly in our data would drop to 5.1% if these methodological issues were taken into account. Similarly, our extrapolation to estimate the actual level of non-monophyly (23%) would drop to 9.5, albeit with relatively broad confidence intervals. It is not possible to precisely classify all cases of non-monophyly as due to methodological or biological causes because there are different interpretations of species concepts, but our data indicate the two are of roughly equal importance. Furthermore, we are not able to provide any estimate of the frequency of hybrid specimens in our data. Hybrid specimens could easily be misidentified or even described as valid species. Recent in-depth studies have revealed several such cases (Anderson et al. 2007; Rougerie et al. 2012). Because of maternal inheritance, hybrid specimens cannot be identified by mitochondrial DNA markers such as the barcode locus (COI).

NJ and ML methods yielded similar estimates of the incidence of paraphyly or polyphyly, suggesting that any differences with prior studies were not caused by the tree-construction method. Most studies included in Funk and Omland (2003) were based on NJ analysis. We deliberately used the K2P nucleotide substitution model in NJ analyses because it is employed by most DNA barcoding studies, although it is not always the best-fit model (Srivathsan and Meier 2012). Based on our trials (results not shown), the substitution model very seldom had any effect on the tree topology using NJ. NJ is known to be prone to a variety of artifacts, such as the order of specimens in the input file (Farris et al. 1996) and long-branch attraction (Felsenstein 1978). For this reason, its use in molecular phylogenetic studies is often disfavored. We also observed that NJ tends to place sequences that are identical except for length at different nodes. We, therefore, adopted ML as the basis for most of our analyses. Although the numbers of non-monophyletic species recovered by the two methods were similar there was only a 76% overlap in species composition. This highlights the importance of method selection in DNA barcode-related studies, especially when topology-related questions are addressed.

Based on our observations, species-level polyphyly in COI gene trees involving deep genetic divergence is very rare in Lepidoptera, but the situation may be different in groups with different biology (e.g., with haplodiploid genetic system, see Patten et al. 2015). Non-monophyly above the genus level is exceptional, and is likely to involve either misidentifications or oversplitting of genera while non-monophyly involving higher taxonomic groups was never observed. Cases of COI barcode-sharing between closely related species have been reported in many taxa, although their frequency is usually low and is often due to oversplitting (e.g., Hausmann et al. 2013; Pentinsaari et al. 2014). Under these circumstances, paraphyly, polyphyly, and monophyly are not distinct phenomena as a single nucleotide substitution can change the type of relationship. For the same reason, some of the species observed as monophyletic in this study may be revealed as non-monophyletic with increased sampling (cf. Fig. 4). Actually, under perfect barcode-sharing between two or more species with no intraspecific variation, species could equally be considered mono-, para-, or polyphyletic. We considered such cases (usually species pairs) reciprocally monophyletic, but even a single mutation would negate this. A significant fraction of non-monophyletic species represents cases of high genetic similarity between two or more species. Many species having deep intraspecific variation (usually with a deep split) appear paraphyletic because other species are nested within them.

Is the prevalence of non-monophyly in Lepidoptera typical of other groups? Both Funk and Omland (2003) and Ross (2014) compared the incidence of non-monophyly among major animal groups. They detected significant differences amongst taxa but with Arthropoda typically being near the mean. We found that the probability of non-monophyly declined as the average genetic distance to the nearest neighbor increased. Several recent DNA barcode data release papers enable us to explore whether this pattern extends to comparisons across higher taxa. Pentinsaari et al. (2014) studied 1972 Coleoptera species and found that the mean K2P difference to the nearest neighbor was over twice that of Lepidoptera (11.99% vs. 5.80%; 5.22% in our data for Lepidoptera). Their estimate of the frequency of non-monophyly was only 2.2% without adjustment for methodological issues such as cryptic species. The data are geographically more limited and the average sampling effort lower, but there is little doubt that the level of non-monophyly in beetles is lower than in Lepidoptera. Ward et al. (2005) reported a mean interspecific divergence from the nearest neighbor of 22.03% in fishes, Hebert et al. (2004) and Kerr et al. (2009) 11.82% and 12.64% in birds, respectively, Chang et al. (2009) 18.66% in earthworms, Zhou et al. (2009) 15.54% in Trichoptera, 23.89% in Ephemeroptera, and 19.24% in Plecoptera, Ball et al. (2005) 25.02% in Ephemeroptera, Shaffield et al. (2009) 13.81% in Hymenoptera (Apoidea), Blagoev et al. (2009) 6.77% in Araneae, and Hogg and Hebert (2004) 21.03% in Collembola.

DNA barcoding has great potential to accelerate taxonomic workflows by enabling rapid sorting of specimens into tentative species or operational taxonomic units (Zhou et al. 2007; Kekkonen and Hebert 2014). Molecular data have the advantage of potentially permitting species delimitation in a quantitative and standardized way (Tautz et al. 2003; Leaché et al. 2014). While a final taxonomic framework has to be based on more comprehensive genomic (Leaché et al. 2014) or broadly integrative data (Padial et al. 2010; Schlick-Steiner et al. 2010), DNA barcodes are very valuable because they are easy to obtain at low cost (including from older museum specimens) and existing reference libraries with broad taxonomic coverage already exist. Therefore, an increasing number of taxonomic revisions is based in whole or in part on DNA barcodes. Several quantitative species delimitation algorithms for molecular data have been developed over the past decade, including approaches dedicated to DNA barcodes such as ABGD and BIN (Puillandre et al. 2012; Ratnasingham and Hebert 2013). Other approaches such as GMYC (Pons et al. 2006), bGMYC (Fujisawa and Barraclough 2013), DISSECT (Jones et al. 2015), and PTP (Zhang et al. 2013) permit the analysis of varied genetic markers, whereas other methods enable species delimitation based on multi-marker or even genome-wide SNP data (Yang and Rannala 2010; Leaché et al. 2014; Pante et al. 2014). Regardless of the method, species-level non-monophyly forms a major challenge as no method can correctly delimit species showing polyphyly in a gene tree and only exceptionally can they deal with paraphyly. While our study suggests that species-level non-monophyly is less frequent than previously thought, problems in

algorithmic DNA-based species delimitation remain. More attention should be paid to separating the true cases of non-monophyly from those resulting from technical and methodological causes. We hope a schematic stepwise chart (Fig. 1) will serve as a general blueprint for taxonomic studies. Our open-access tool "Monophylizer" should help taxonomists to rapidly and confidently separate monophyletic, paraphyletic, and polyphyletic species from each other based on phylogenetic trees in a variety of common file formats. Cases of non-monophyly should be flagged for careful reappraisal and deep-level species polyphyly studied for the presence of overlooked species.

Conclusions

Species delimitation is increasingly based on molecular data. However, non-monophyly represents a major challenge for algorithmic species delimitations. Processes such as incomplete lineage sorting and introgression give rise to biological non-monophyly that cannot be resolved by increased geographic or genetic sampling. However, our results suggest that current estimates overestimate the extent of non-monophyly in trees based on mitochondrial DNA. We found that a very high fraction of cases of non-monophyly reflects methodological issues, such as misidentifications, oversplitting of species, overlooked species, and the inherent subjectivity of species delimitations, especially when allopatric populations are concerned. Species polyphyly in mtDNA is rare and mostly attributable to cases of very shallow divergence between species, but in rare cases it may also result from mitochondrial introgression. Whether or not a species appears monophyletic in a tree is also affected by the method used to build the tree. Overall, our study supports the argument that, when used with care and in conjunction with other techniques, DNA barcodes are a valuable addition to the tools available for taxonomic work on animals.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.k3mr1.

Funding

Most of the sequences used in this study were generated at the Biodiversity Institute of Ontario under the International Barcode of Life Project, funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute. The generation of German data was funded by grants from the Bavarian State Ministry of Education, Culture, Science and the Arts (Barcoding Fauna Bavarica, BFB) and the German Federal Ministry of Education and Research (German Barcode of Life GBOL2: BMBF #01LI1101B). Molecular laboratory infrastructure and sequencing within the Nature of The Netherlands project was funded by a FES grant from the Dutch Ministry of Finance. The Finnish Barcode of Life project was funded by the Kone Foundation, the Finnish Cultural Foundation, and the University of Oulu. Support for this research was provided by the Spanish Ministerio de Economía y Competitividad [projects CGL2010-21226/BOS and CGL2013-48277-P to R.V.], by a Région Haute-Normandie post-doctoral fellowship [to R.R.], and by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme [project no. 625997 to V.D.]ă Sequencing of Norwegian material was supported by the Natural History Museum, University of Oslo, and the Norwegian Barcode of Life Network (NorBOL). Sequencing within the framework of the Lepidoptera of the Alps Campaign was supported by the Promotion of Educational Policies, University and Research Department of the Autonomous Province of Bolzano—South Tyrol with funds to the project "Genetic biodiversity archive - DNA barcoding of Lepidoptera of the central Alpine region (South, East and North Tyrol)," the Austrian Federal Ministry of Science, Research and Economics with funds to ABOL (Austrian Barcode of Life), and by the regional institutions Tiroler Landesmuseen, inatura and Landesmuseum Kärnten. S.M.K. was funded by the international fellowship program at Stockholm University and Finnish Cultural Foundation. Sampling of Lepidoptera from Upper-Normandy (France) was supported by a grant by Conseil Régional de Haute-Normandie to Thibaud Decaëns, then member of the ECODIV laboratory at the University of Rouen.

Acknowledgments

We are indebted to a large number of taxonomic experts and collectors who have contributed to the Lepidoptera Barcode of Life campaign in multiple ways, especially by providing material for DNA barcoding and by identifying specimens. We are grateful to the ICT staff at Naturalis, and especially David Heijkamp, for hosting the Monophylizer web service through their infrastructure. We are very grateful to staff at the Biodiversity Institute of Ontario for their key role in generating sequences, photographing specimens, entering data elements into BOLD, and aiding the curation of this information. We thank Jess Johansson and Megan Milton for help with GenBank submissions and BOLD data set. Finally, we thank Frank Andersson, Karl Kjer, and an anonymous referee for their comments on an earlier version of the study.

References

Anderson S.J., Gould, P., Freeland, J.R. 2007. Repetitive flanking sequences (ReFS): novel molecular markers from microsatellite families. Mol. Ecol. Notes 7:374–376.

Avise J.C. 1989. A role for molecular genetics in the recognition and coservation of endangered species. Trends Ecol. Evol. 4:279–281.
Ball S.L., Hebert P.D.N., Burian S.K., Webb J.M. 2005. Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. J. N. Am. Benthol. Soc. 24:508–524.
Baselga A., Fujisawa T., Crampton-Platt A., Bergsten J., Foster P.G., Monaghan M.T., Vogler A.P. 2013. Whole-community DNA barcoding reveals a spatiotemporal continuum of biodiversity at species and genetic levels. Nat. Commun. 4:1892.
Bidartondo M.I. 2008. Preserving accuracy in GenBank. Science 319:1616.
Blagoev G., Hebert P., Adamowicz S., Robinson E. 2009. Prospects for using DNA barcoding to identify spiders in species-rich genera. ZooKeys 16:27–46.
Brower A.V.Z., Egan M.G. 1997. Cladistic analysis of *Heliconius* butterflies and relatives (Nymphalidae: Heliconiiti): a revised phylogenetic position for *Eueides* based on sequences from mtDNA and a nuclear gene. Proc. R. Soc. B 264:969.
Canty A., Ripley B. 2013. boot: Bootstrap R (S-Plus) functions. R package version 1.3–9. https://cran.r-project.org/web/packages/boot/citation.html, last accessed June 2, 2016.
Chang C., Rougerie R., Chen J. 2009. Identifying earthworms through DNA barcodes: pitfalls and promise. Pedobiologia 52: 171–180.
Coyne J.A., Orr H.A. 2004. Speciation. Sunderland (MA): Sinauer Associates.
Cracraft J. 1989. Speciation and its ontology: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In: Otte D., Endler J.A., editors. Speciation and its consequences. Sunderland (MA): Sinauer Associates. p. 28–59.
De Queiroz K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.
deWaard J.R., Ivanova N.V., Hajibabaei M., Hebert P.D.N. 2008. Assembling DNA barcodes: analytical protocols. In: Cristofre Martin C., editor. Methods in molecular biology: environmental genetics. Totowa (NJ): Humana Press Inc. p. 275–293.
Dincă V., Lukhtanov A.V., Talavera G., Vila R. 2011. Unexpected layers of cryptic diversity in Wood White *Leptidea* butterflies. Nat. Commun. 2:324.
Dincă V., Montagud S., Talavera G., Hernández-Roldán J., Munguira M.L., García-Barros E., Hebert P.D.N., Vila R. 2015. DNA barcode reference library for Iberian butterflies enables a continental-scale preview of potential cryptic diversity. Sci. Rep. 5:12395.
Eddy S.R. 1998. Profile hidden Markov models. Bioinformatics 14: 755–763.
Elzinga J.A., Mappes J., Kaila L. 2014. Pre- and post-mating reproductive barriers drive divergence of five sympatric species of Naryciinae moths (Lepidoptera: Psychidae). Biol. J. Linn. Soc. 112:584–605.
Farris J.S., Albert V.A., Kallersjo M., Lipscomb D., Kluge A.G. 1996. Parsimony jackknifing outperforms Neighbor-Joining. Cladistics 12:99–124.
Felsenstein J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Fibiger M. 1997. Noctuinae III. In: Tremewan W.G., Honey M., Lyneborg L. Noctuidae europaeae, Vol. 3. Sorø: Entomological Press. p. 1–418.
Finn R.D., Mistry J., Tate J., Coggill P., Heger A., Pollington J.E., Gavin O.L., Gunasekaran P., Ceric G., Forslund K., Holm L., Sonnhammer E.L.L., Eddy S.R., Bateman A. 2010. The Pfam protein families database. Nucleic Acids Res. 38:D211–D222.
Fujisawa T., Barraclough T.G. 2013. Delimiting species using single-locus data and the generalized mixed Yule coalescent approach: a revised method and evaluation on simulated data sets. Syst. Biol. 62:707–724.
Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.
Glez-Peña D., Gómez-Blanco D., Reboiro-Jato M., Fdez-Riverola F., Posada D. 2010. ALTER: program-oriented conversion of DNA and protein alignments. Nucleic Acids Res. 38:W14–W18.
Groenenberg D.S.J., Neubert E., Gittenberger E. 2011. Reappraisal of the "Molecular phylogeny of Western Palaearctic Helicidae s.l. (Gastropoda: Stylommatophora)": when poor science meets GenBank. Mol. Phyl. Evol. 61:914–923.
Harris D.J. 2003. Can you bank on GenBank? Trends Ecol. Evol. 18: 317–319.
Hausdorf B. 2011. Progress toward a general species concept. Evolution 65:923–931.
Hausmann A., Godfray H.C.J., Huemer P., Mutanen M., Rougerie R., van Nieukerken E.J., Ratnasingham S., Hebert P.D.N. 2013. Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. PLoS One 8:e84518.
Hausmann A., Hebert P.D.N., Mitchell A., Rougerie R., Sommerer M., Young, C. 2009a. 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.
Hausmann A., Sommerer M., Rougerie R., Hebert, P.D.N. 2009b. *Hypobapta tachyhalotaria* n. sp. from Tasmania – an example of a new species revealed by DNA barcoding (Lepidoptera, Geometridae). Spixiana 32:237–242.
Hebert P.D.N., Stoeckle M.Y., Zemlak T.S., Francis C.M. 2004. Identification of birds through DNA barcodes. PLoS Biol. 2:e312.
Hennig W. 1966. Phylogenetic systematics. Urabana (IL): University of Illinois Press.
Hogg I.D., Hebert P.D.N. 2004. Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. Can. J. Zool. 82:749–754.
Hogner S., Laskemoen T., Lifjeld J.T., Porkert J., Kleven O., Albayrak T., Kabasakal B., Johnsen A. 2012. Deep sympatric mitochondrial divergence without reproductive isolation in the common redstart *Phoenicurus phoenicurus*. Ecol. Evol. 2:2974–2988.
Huelsenbeck J.P., Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.
Huemer P., Elsner G., Karsholt O. 2013. Review of the Eulamprotes wilkella species-group based on morphology and DNA barcodes, with descriptions of new taxa (Lepidoptera, Gelechiidae). Zootaxa 3746:069–100.
Huemer P., Hausmann A. 2009. A new expanded revision of the European high mountain *Sciadia tenebraria* species group (Lepidoptera: Geometridae). Zootaxa 2117:1–30.
Huemer P., Hebert P.D.N. 2011. Cryptic diversity and morphology of high alpine *Sattleria* – a case study combining DNA barcodes and morphology (Lepidoptera: Gelechiidae). Zootaxa 2981: 1–22.
Huemer P., Karsholt O., Mutanen M. 2014. DNA barcoding as a screening tool for cryptic diversity: an example from *Caryocolum*, with description of a new species (Lepidoptera, Gelechiidae). ZooKeys 404:91–111.
Huemer P., Mutanen M. 2015. Alpha taxonomy of the genus *Kessleria* Nowicki, 1864, revisited in light of DNA-barcoding (Lepidoptera, Yponomeutidae). ZooKeys 503:89–133.
Huemer P., Timossi G. 2014. *Sattleria* revisited: unexpected cryptic diversity on the Balkan Peninsula and in the south-eastern Alps (Lepidoptera: Gelechiidae). Zootaxa 3780:282–296.
Hurst G.D.D., Jiggins F.M. 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc. R. Soc. B 272: 1525–1534.
ICZN. 1999. International code of zoological nomenclature. 4th ed. London: The International Trust for Zoological Nomenclature. 306 pp.
Isaac N.J.B., Mallet J., Mace G.M. 2004. Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol. Evol. 19:464–469.
Jones G., Aydin Z., Oxelman B. 2015. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31:991–998.
Karsholt O., van Nieukerken E.J., editors. 2013. Fauna Europaea: Lepidoptera. Fauna Europaea version 2.6. Available from http://www.faunaeur.org/, last accessed June 2, 2016.
Kekkonen M., Hebert P.D.N. 2014. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Mol. Ecol. Res. 14:706–715.

Kerr K.C.R., Birks S.M., Kalyakin M.V., Red'kin Y.A., Koblik E.A., Hebert P.D.N. 2009. Filling the gap - COI barcode resolution in eastern Palearctic birds. Front. Zool. 6:29–42.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.
Kirichenko N., Huemer P., Deutsch H., Triberti P., Rougerie R., Lopez-Vaamonde C. 2015. Integrative taxonomy reveals a new species of *Callisto* (Lepidoptera, Gracillariidae) in the Alps. ZooKeys 473: 157–176.
Kullberg J., Albrecht A., Kaila L., Varis V. 2001. Checklist of Finnish Lepidoptera. Sahlbergia 6:45–190.
Kumar S., Gadagkar S.R. 2000. Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. J. Mol. Evol. 51:544–53.
Kvie K.S., Hogner S., Aarvik L., Lifjeld J.T., Johnsen A. 2013. Deep sympatric mtDNA divergence in the autumnal moth (*Epirrita autumnata*). Ecol. Evol. 3:126–144.
Leaché A.D., Fujita M.K., Minin V.N., Bouckaet R.R. 2014. Species delimitation using genome-wide SNP data. Syst. Biol. 63: 534–542.
LeDell E., Petersen M., van der Laan M. 2014. cvAUC: cross-validated area under the ROC curve confidence intervals. R Package Version 1.1.0. http://CRAN.R-project.org/package=cvAUC, last accessed June 2, 2016.
Lees D.C., Kawahara A.Y., Bouteleux O., Ohshima I., Kawakita A., Rougerie R., De Prins J., Lopez-Vaamonde C. (2013). DNA barcoding reveals a largely unknown fauna of Gracillariidae leaf-mining moths in the Neotropics. Mol. Ecol. Res. 14:286–296.
Maddison W.P. 1997. Gene trees in species trees. Syst Biol. 46: 523–536.
Mallet J. 2007. Species, concepts of. In: Levin S.A., editor. Encyclopedia of biodiversity, Vol. 5. Elsevier, Oxford: Academic Press. p. 427–440.
Mayr, E. 1942. Systematics and the origin of species. New York: Columbia University Press.
McKay B.D., Zink R.M. 2010. The causes of mitochondrial DNA gene tree papraphyly in birds. Mol. Phyl. Evol. 54:647–650.
Mihaescu R., Levy D., Pachter L. 2009. Why neighbor-joining works. Algorithmica (New York). 54:1–24.
Mutanen M. 2005. Delimitation difficulties in species splits: a morphometric case study on the *Euxoa tritici* complex (Lepidoptera, Noctuidae). Syst. Entomol. 30:632–643.
Mutanen M., Hausmann A., Hebert P.D.N., Landry J.F., de Waard J.R., Huemer P. 2012. Allopatry as a Gordian knot for taxonomists: patterns of DNA barcode divergence in Arctic-Alpine Lepidoptera. PLoS One 7:e47214.
Mutanen M., Kaila L., Tabell J. 2013. Wide-ranging barcoding aids discovery of one-third increase of species richness in presumably well-investigated moths. Sci. Rep. 3: 2901.
Mutanen M.,Wahlberg N., Kaila L. 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc. R. Soc. B 277:2839–2848.
Olave M., Solà E., Knowles L.L. 2014. Upstream analyses create problems with DNA-based species delimitation. Syst. Biol. 63: 262–271.
Padial J.M., Miralles A., De la Riva I., Vences M. 2010. Review: the integrative future of taxonomy. Front. Zool. 7:16.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Pante E., Abdelkrim J., Viricel A., Gey D., France S.C., Boisselier M.C. Samadi S. 2014. Use of RAD sequencing for delimiting species. Heredity 114:450–459.
Patten M.M., Carioscia S.A., Linnen C.R. 2015. Biased introgression of mitochondrial and nuclear genes: a comparison of diploid and haplodiploid systems. Mol. Ecol. 24:5200–5210.
Pentinsaari M., Hebert P.D.N., Mutanen M. 2014. Barcoding beetles: a regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. PLoS One 9:e108651.
Platnick N.I. 1977. Paraphyletic and polyphyletic groups. Syst. Biol. 26:195–200.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol. Ecol. 21:1864–1877.
Ratnasingham S., Hebert P.D.N. 2007. BOLD: the Barcode of Life Data System (www.barcodinglife.org). Mol. Ecol. Notes 7:355–364.
Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS One 8:e66213.
R Core Team 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. http://www.R-project.org/, last accessed June 2, 2016.
Regier J.C., Mitter C., Zwick A., Bazinet A.L., Cummings M.P., Kawahara A.Y., Sohn J.C., Zwickl D.J., Cho S., Davis D.R., Baixeras J., Brown J., Parr C., Weller S., Lees D.C., Mitter K.T. 2013. A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (Moths and Butterflies). PLoS One 8: e58568.
Ross H.A. 2014. The incidence of species-level paraphyly in animals: a re-assessment. Mol. Phylogenet. Evol. 76:10–17.
Rougerie R., Haxaire J., Kitching I.J., Hebert P.D.N. 2012. DNA barcodes and morphology reveal a hybrid hawkmoth in Tahiti (Lepidoptera: Sphingidae). Inv. Syst. 26:445–450.
Rubinoff D., Cameron S., Will K. 2006. A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J. Hered. 97:581–594.
Schlick-Steiner B.C., Steiner F.M., Seifert B., Stauffer C., Christian E., Crozier R.H. 2010. Integrative taxonomy: a multisource approach to exploring biodiversity. Annu. Rev. Entomol. 55:421–438.
Shaffield C.S., Hebert P.D.N., Kevan P.G., Packer L. 2009. DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. Mol. Ecol. Res. 9:196–207.
Srivathsan A., Meier R. 2012. On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. Cladistics 28:190–194.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.
Steel M., Velasco J.D. 2014. Axiomatic opportunities and obstacles for inferring a species tree from gene trees. Syst. Biol. 63:772–778.
Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18:70–74.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Am. Math. Soc. Lect. Math. Life Sci. 17:57–86.
van Nieukerken E.J., Kaila L., Kitching I, J., Kristensen N.P., Lees D.C., Minet J., Mitter C., Mutanen M., Regier J.C., Simonsen T.J., Wahlberg N., Yen S-H., Zahiri R., Adamski D., Baixeras J., Bartsch D., Bengtsson B.Å., Brown J.W., Bucheli S.R., Davis D.R., De Prins J., De Prins W., Epstein M.E., Gentili-Poole P., Gielis C., Hättenschwiler P., Hausmann A., Holloway J. D., Kallies A., Karsholt O., Kawahara A., Koster S.J.C., Kozlov M., Lafontaine J.D., Lamas G., Landry J-F., Lee S., Nuss M., Park K.T., Penz C., Rota J., Schmidt B.C., Schintlmeister A., Sohn J.C., Solis M.A., Tarmann G.M., Warren A.D., Weller S., Yakovlev R.V., Zolotuhin V.V., Zwick A. 2011. Order Lepidoptera Linnaeus, 1758. In: Zhang Z.-Q., editor. Animal biodiversity: an outline of higher-level classification and survey of taxonomic richness. Zootaxa 3148:212–221.
van Nieukerken E.J., Mutanen M., Doorenweerd C. 2012. DNA barcoding resolves species complexes in *Stigmella salicis* and *S. aurella* species groups and shows additional cryptic speciation in *S. salicis* (Lepidoptera: Nepticulidae). Entomol. Tidsk. 132:235–255.
Venables W.N., Ripley B.D. 2002. Modern applied statistics with S. New York: Springer.
Vos R.A., Caravas J., Hartmann K., Jensen M.A., Miller C. 2011. BIO::phylo-phyloinformatic analysis using perl. BMC Bioinform. 12:63.
Ward R.D., Zemlak T.S., Innes B.H., Last P.R., Hebert P.D.N. 2005. DNA barcoding Australia's fish species. Philos. Trans. R. Soc. B 360: 1847–1857.

Webb W.C., Marzluff J.M., Omland K.E. 2011. Random interbreeding between cryptic lineages of the common Raven: evidence for speciation in reverse. Mol. Ecol. 20:2390–2402.
Wheeler Q.D., Meier R., editors. 2000. Species concept and phylogenetic theory: a debate. New York: Columbia University Press.
Wright, S. 1940. The statistical consequences of Mendelian heredity in relation to speciation. In: Huxley J., editor. The new systematics. London: Oxford University Press. p. 161–183.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl Acad. Sci. USA 107:9264–9269.
Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876.
Zhou X., Adamowicz S.J., Jacobus L.M., Dewalt R.E., Hebert P.D.N. 2009. Towards a comprehensive barcode library for arctic life - Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. Front. Zool. 6:30.
Zhou X., Kjer K.M., Morse J.C. 2007. Associating larvae and adults of Chinese hydropsychidae caddisflies (Insecta: Trichoptera) using DNA sequences. J. N. Am. Benthol. Soc. 26:719–742.