DNA barcodes combined with multi-locus data of representative taxa can generate reliable higher-level phylogenies

Gerard Talavera[1,2,*], Vladimir Lukhtanov[3], Naomi E. Pierce[2], Roger Vila[4]

[1] Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Passeig del Migdia s/n, 08038 Barcelona, Catalonia, Spain.

[2] Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, United States

[3] Department of Karyosystematics, Zoological Institute of Russian Academy of Sciences, Universitetskaya nab. 1, 199034 St. Petersburg, Russia

[4] Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 08003 Barcelona, Catalonia, Spain.

KEYWORDS: DNA barcoding, *incertae sedis*, missing data, phylogeny, Polyommatina, supermatrix, systematics, taxonomy, Lycaenidae, phylogenomic, *Kipepeo, Birabiro.*

RUNNING TITLE: DNA barcodes on higher-level phylogenies

*Corresponding author. E-mail: gerard.talavera@csic.es

ABSTRACT

Taxa are frequently labeled *incertae sedis* when their placement is debated at ranks above the species level, such as their subgeneric, generic or subtribal placement. This is a pervasive problem in groups with complex systematics due to difficulties in identifying suitable synapomorphies. In this study, we propose combining DNA barcodes with a multi-locus backbone phylogeny in order to assign taxa to genus or other higher-level categories. This sampling strategy generates molecular matrices containing large amounts of missing data that are not distributed randomly: barcodes are sampled for all representatives, and additional markers are sampled only for a small percentage. We investigate the effects of the degree and randomness of missing data on phylogenetic accuracy using simulations for up to 100 markers in 1000-tips trees, as well as a real case: the subtribe Polyommatina (Lepidoptera: Lycaenidae), a large group including numerous species with unresolved taxonomy. Our simulation tests show that when a strategic and representative selection of species for higher-level categories has been made for multi-gene sequencing (approximately one per simulated genus), the addition of this multi-gene backbone DNA data for as few as 5-10% of the specimens in the total dataset can produce high-quality phylogenies, comparable to those resulting from 100% multi-gene sampling. In contrast, trees based exclusively on barcodes performed poorly. This approach was applied to a 1365-specimen dataset of Polyommatina (including ca. 80% of described species), with nearly 8% of representative species included in the multi-gene backbone and the remaining 92% included only by mitochondrial COI barcodes, a phylogeny was generated that highlighted potential misplacements, unrecognized major clades, and placement for *insertae sedis* taxa. We use this information to make systematic rearrangements within

Polyommatina, and to describe two new genera. Finally, we propose a systematic workflow to assess higher-level taxonomy in hyperdiverse groups. This research identifies an additional, enhanced value of DNA barcodes for improvements in higher-level systematics using large datasets.

INTRODUCTION

The impact of missing data in modern phylogenetics is highly debated. It is well accepted that phylogenetic accuracy improves with greater sampling of taxa and more informative characters. In practice, possible detrimental effects of imbalanced sampling for phylogenetic inference are often circumvented by excluding taxa and/or genes when the former have problematic placements or the latter have been poorly sampled. However, increasing evidence suggests that inclusion of incomplete taxa (that have not been sequenced for all markers) or incomplete markers (that have not been sequenced for all taxa) may increase phylogenetic accuracy, or at worst be inconsequential provided that a sufficient number of informative characters are sampled overall (Wiens 2003; Philippe et al. 2004; Wiens 2006; de Queiroz and Gatesy 2007; Wiens and Morril 2011; Wiens and Tiu 2012; Roure et al. 2013; Grievink et al. 2013; Jiang et al. 2014). According to this view, complete matrices are not essential for optimal phylogenetic performance, and incomplete taxa can still be placed correctly. Nevertheless, this approach is not without controversy (Lemmon et al. 2009; but see Wiens and Morill 2011; Simmons 2012a; Simmons 2012b).

Given that comprehensive datasets in terms of both taxa and characters are hard to obtain, especially for hyperdiverse taxon groups, two strategies are commonly used to

explore biodiversity: (1) sampling multiple loci (in the hundreds in the case of phylogenomics) for representatives of higher-level taxonomic categories (the "phylogenetic/-omic approach"), which explores deeper relationships but potentially misses recent diversification; or (2) sampling only one or two loci, such as the mitochondrial COI DNA-barcode, for as many taxa as possible, and trying to cover the entire group's biodiversity at the possible expense of accurate inference of deep relationships (the "barcoding approach"). These approaches are analogous to the "bottom up" (many characters, few taxa) and "top down" (many taxa, few characters) analyses described by Wiens (2005). Option (1) should facilitate resolving higher-level relationships. However, while phylogenetic accuracy is improved by the addition of informative characters, this approach can sometimes create model violations if it fails to detect multiple substitutions due to long branches (Poe 2003; Wiens 2005). In addition, other potential issues such as gene tree discordance and 'the anomaly zone' may also challenge phylogenetic accuracy (Jeffroy et al. 2006; Degnan and Rosenberg 2006; Galtier and Daubin 2008; Mendes and Hahn 2018). Option (2) has the benefit of breaking long branches and thus improves the detection of multiple substitutions, but the smaller number of informative characters, as well as the linkage of those characters in a single mitochondrial gene region, may result in weakly supported phylogenetic hypotheses (DeSalle et al. 2005). Moreover, the resulting single-gene phylogenetic histories are likely to not reflect the species trees (Pamilo and Nei 1988; Maddison 1997; Nichols 2001; Degnan and Rosenberg 2006). The debate regarding the costs and benefits of sampling more taxa on the one hand or more characters on the other has a long history and is still unresolved (GrayBeal 1998; Rannala et al. 1998; Zwickl and Hillis 2002; Poe 2003; Rosenberg and Kumar 2001; Wiens and Morrill 2011; Philippe et al. 2011; Wiens and Tiu 2012; Zheng and Wiens 2016).

In theory, sampling many loci for many taxa would be the best solution, but this remains a costly option for many, often due to the difficulty of obtaining samples with well-preserved DNA. In practice, most phylogeneticists have to cope with the problem of missing data or imbalanced sets of sequences in the attempt to build large phylogenies by merging multiple datasets. The availability of public molecular data increases exponentially, but these data are remarkably heterogeneous. Heterogeneity comes from the varied sampling strategies followed in different studies, from uneven sequencing of various genetic markers and/or from sampling biases involving particular clades or genes.

A number of studies have attempted to evaluate the performance of patchy supermatrices in phylogenetics (Wiens et al. 2005; de Queiroz and Gatesy 2007; Cho et al. 2011; Kawahara et al. 2011; Roure et al. 2013; Hovmöller et al. 2013; Streicher et al. 2015; Philippe et al. 2017). Patchy supermatrices may differ not only in their completeness, but also in their randomness, a rarely assessed but possibly important parameter. Randomness in matrix patchiness can vary in both dimensions: taxon or marker. Previous studies tested scenarios with maximum randomness: probabilities of representation are equal for all taxa and characters (i.e., all taxa may equally have 1,2,3... to the maximum number of markers). It is still unclear whether partial matrices with guided missing data (those where only a set of selected samples are fully represented by all genes; see Fig. 1) can benefit phylogenetic accuracy. We here test the effects of missing data on partial matrices where some taxa have a maximum number of markers and some have only one and always the same one, the standard COI mitochondrial DNA barcode. We also test the relevance of the selected taxa having the maximum number of markers by using a random strategy versus a taxonomically-

guided strategy that prioritizes having at least one representative per higher-level clade (genus in this case). The motivating question is to evaluate whether partial matrices can be strategically designed by combining multiple genes for relatively few representative taxa (the phylogenetic/-omic approach) to resolve higher-level relationships, with barcode data from many taxa mostly informing species-level or shallow relationships (the barcoding approach).

We analyze different scenarios through simulation experiments and use this approach to assess a challenging empirical dataset, the Polyommatina butterflies. The subtribe Polyommatina (Lycaenidae, Polyommatinae) is a species-rich group (ca. 480 species) that is the product of one or more radiations (Kandul et al. 2004; Wiemers et al. 2010; Vila et al. 2011; Talavera et al. 2013a; Talavera et al. 2015, Stradomsky 2016). Butterflies in this group are morphologically highly similar and their taxonomy has been unstable. Species diversity in the Polyommatina has been classified with 82 formally described generic names in a wide array of taxonomic combinations. Prior to this work, we addressed a higher-level taxonomic revision after reconstructing the first comprehensive molecular phylogeny of the group, based on three mitochondrial genes plus six nuclear markers (Talavera et al. 2013a). This dataset included 109 specimens representing nearly all genera and subgenera described within the subtribe. The resulting phylogeny uncovered several polyphyletic genera. We develop objective criteria for a systematic arrangement that could best accommodate pre-existing generic nomenclature to the new phylogenetic framework, and, after applying a flexible temporal scheme, we delimited 32 genera.

The controversial taxonomy of this group mirrors the high evolutionary lability of most morphological characters. It's possible that the group contains cryptic diversity, and that taxa not characterized genetically so far might be assigned to the wrong genus. In fact, a remarkable number of species in this group have been assigned to multiple genera by different authors. For example, the North American taxon *acmon* Westwood, [1851] (originally described as *Lycaena acmon*) has been placed in the genera *Plebejus* (Pelham 2008), *Aricia* (Bálint and Johnson 1997), and *Icaricia* (Layberry et al. 1998, Talavera et al. 2013a). The enigmatic and morphologically distinct taxon *avinovi* Stshetkin 1980 has been placed in the genera *Polyommatus* (Bálint and Johnson 1997), *Rimisia* (Zhdanki 2004; Eckweiler and Bozano 2016) and *Afarsia* (Shapoval and Lukhtanov 2016). This situation is not unique to the Polyommatina, but extends to many other insect groups where rare or morphologically similar taxa provide challenging taxonomic assignments due to difficulties in finding diagnostic synapomorphies.

In this study we increase taxon sampling for the Polyommatina to 1360 specimens, comprising about 80% of putative species. We combine DNA barcodes with the genus-level phylogenetic backbone in a supermatrix where specimens having only COI barcodes (658 bp) represent ca. 92% of the total matrix, and specimens with multiple markers (6666 bp) represent ca. 8%. With this approach, we aim to screen the phylogenetic diversity of the group, assign species or subspecific taxa to genera, identify unrecognized major clades and re-evaluate the phylogenetic history of the group with a nearly complete taxon sampling.

We also design a battery of simulations to evaluate phylogenetic accuracy for partial matrices, with particular emphasis on testing whether strategic selections of fully-

sequenced representatives improve accuracy over random selections. Our simulations test two scenarios: 1) a phylogenetic dataset resembling our empirical data, and 2) a phylogenomic dataset (sampling 100 genes per taxon) to test the possible effect of backbone-barcode imbalances in large-scale studies. We propose a systematic workflow to assess higher-level taxonomy in hyperdiverse groups. In so doing, we also reinforce the value of the COI DNA-barcode in higher systematics when combined with a minimal, but well-designed, multi-locus framework.

MATERIALS AND METHODS

Empirical molecular datasets

We gathered molecular data for as many taxa as possible within the Polyommatina butterflies (genera, subgenera, species and subspecies), sampling as many populations as possible within the distribution range of each taxon (Supplementary Table S1). Our phylogenetic approach involved building two different molecular datasets. First, we took advantage of a multi-locus matrix assembled for an earlier study (Talavera et al 2013a), that included a mitochondrial DNA fragment containing three gene regions, plus six nuclear markers (6,666 bp, hereafter referred to as the backbone dataset, Fig. 1a). This dataset included 109 specimens with at least one representative of each of the 82 formally described genera in Polyommatina (with the exception of *Xinjiangia* Huang & Murayama, 1988 and *Grumiana* Zhdanko, 2004). The markers included in the backbone dataset were mitochondrial *cytochrome oxidase* I (COI), *leucine transfer* RNA (leu-tRNA) and *cytochrome oxidase* II (COII), and nuclear *elongation factor*- 1 *alpha* (EF-1a), 28S ribosome unit (28S), *histone* H3 (H3), *wingless* (wg) – *carbamoyl-*

phosphate synthetase 2 / aspartate transcarbamylase / dihydroorotase (CAD) and *internal transcribed spacer* 2 (ITS2).

A second dataset (hereafter referred to as the barcode dataset, Fig. 1b) was generated by assembling a single gene matrix (658 bp) for the universal barcode fragment of mitochondrial COI. This dataset exemplifies a molecular matrix with a one-gene phylogenetic history, often involving a limited number of informative characters. A total of 1365 barcodes were retrieved from multiple sources: 109 from the backbone dataset, 1100 from the public repositories GenBank and BOLD and 156 from specimens collected in the field or obtained from collections and sequenced specifically for this research. New collection efforts specifically targeted taxa and populations that are difficult to obtain and/or are not sampled in previous studies (collection data in Supplementary Table S1). The barcode dataset included representatives of approximately 80% of the roughly 480 species of Polyommatina currently recognized (Bálint and Johnson 1997; Talavera et al. 2013a). Both backbone and barcode datasets included as outgroup taxa four representatives for the sister subtribe Everina and one for Leptotina based on Talavera et al (2013a). All specimens used in this study are listed in the Supplementary Table S1.

Based on unexpected taxonomic placements or divergences observed from preliminary phylogenetic inspections of the barcode dataset, we increased sequencing coverage by sequencing multiple markers for four additional taxa (*Chilades kedonga*, *Chilades elicola*, *Kretania psyorita* and *Neolysandra corona*) (Supplementary Table S1), thus increasing the backbone dataset to 113 specimens.

A matrix merging barcode and backbone datasets (hereafter referred to as the "combined" dataset) was also built for downstream analyses. This consisted of a matrix of 1365 specimens, where approximately 8% (113 specimens) were completely sequenced for all markers, and 92% (1252 specimens) were represented by *COI* barcodes uniquely. In this asymmetric matrix of characters only one leading marker is complete, and the presence/missing data of other markers is intentionally guided towards particular taxa (Fig. 1c). This model contrasts with that of a partial matrix where presence/missing data of other markers is randomly distributed across taxa (Fig. 1d).

DNA extraction, amplification and sequencing for both barcode and backbone datasets followed standard protocols used for Lycaenidae (Vila et al. 2011; Talavera et al. 2013a). Newly sequenced specimens are stored in the DNA and Tissues Collection of the Institut de Biologia Evolutiva (CSIC-UPF) in Barcelona and the sequences obtained were submitted to GenBank (Supplementary Table S1).

Phylogenetics and divergence times (empirical dataset)

Both barcode and backbone datasets were re-aligned based on available matrices from Talavera et al. (2013a), using Geneious 10.0.3. The barcode matrix consisted of 1365 sequences of 658 bp. The final backbone matrix consisted of 113 tips and 6672 bp: 2172 bp of COI + leu-tRNA + COII, 1171 bp of EF-1a, 745 bp of CAD, 811 bp of 28S, 370 bp of Wg, 1075 bp of ITS2, and 328 bp of H3. Three datasets, backbone alone, barcode alone and backbone and barcode combined, were used for phylogenetics.

Bayesian inference (BI) was used to simultaneously infer evolutionary relationships and divergence times with the software BEAST 1.8.0 (Drummond et al. 2012). Data in the backbone and combined datasets were partitioned by six markers, considering COI + leu-tRNA + COII a single evolutionary unit in the mitochondrial genome. Models for DNA substitution for each marker were chosen according to the Akaike information criterion (AIC) in JModeltest (Guindon and Gascuel 2003; Darriba et al. 2012). As a result, the HKY model was used for H3, the TN model for CAD, and a GTR model for the rest of the markers, in all cases with a gamma distribution (+G) and a proportion of invariants (+I) to account for heterogeneity in evolutionary rates among sites. The gamma distribution was estimated automatically from the data using six rate categories. Normally distributed tmrca priors including maximum and minimum ages within the 95% HPD distribution were established on four well-supported nodes according to Talavera et al. (2013a). The uncorrelated relaxed clock (Drummond et al. 2006) and a constant population size under a coalescent model were established as priors. The rest of the settings and priors were set by default. Two independent chains were run for 50 million generations each, sampling values every 1000 steps. All parameters were analysed using the program Tracer ver. 1.7 (Rambaut and Drummond 2007) to check for stationarity and convergence between runs. Burn-in values were applied accordingly. Independent runs were combined in LogCombiner ver. 1.6.0 and tree topologies were assessed in TreeAnnotator ver. 1.6.0 to generate a maximum clade credibility tree of all sampled trees with median node heights.

Maximum Likelihood (ML) tree inference was performed using two methods, RAxML v.8.2.12 (Stamatakis, 2014) and IQtree v.2 (Minh et al., 2020). For RAxML, a general GTRCAT substitution model for all genes was chosen and 100 rapid bootstrap

277 inferences were executed. For IQtree inference, a general best-fit model for all genes
278 was automatically selected by ModelFinder (Kalyaanamoorthy et al. 2017) and clade
279 support was assessed using ultrafast likelihood bootstrap with 1000 replicates (Hoang
280 et al. 2018). To test for possible effects of different modelling approaches and
281 partitioning schemes, we also inferred ML trees for the combined dataset partitioning
282 characters by codon position, where best substitution models were selected by
283 ModelFinder in IQtree and by PartitionFinder 2.1.1 (Lanfear et a. 2016) for RAxML.
284
285 For the resulting BEAST trees, nodes for genera (as reviewed in Talavera et al. 2013a)
286 were collapsed into a single branch, producing a genus-level tree for subsequent
287 topological comparisons of inter-generic cladogenetic events between the three
288 different datasets. Genus-level trees were produced to discriminate between topological
289 differences belonging to inter- or intra-generic relationships, which are not possible to
290 evaluate from the whole trees. The resulting backbone phylogeny, improved in four
291 relevant taxa, was taken as a reference to re-evaluate generic classifications in
292 Polyommatina by applying the flexible temporal scheme (4-5 Myr) proposed in
293 Talavera et al. (2013a).
294
295 **Simulations**
296
297 We designed simulations to test the performance of combined datasets in both resolving
298 higher-level relationships and placing barcodes within the correct genera. A schematic
299 experimental design is shown in Figure 2. Ten reference trees were first simulated using
300 the function "sim.bd.taxa.age" in the R package TreeSim (Stadler 2011). Parameters
301 were set using information from the Polyommatina phylogeny, including number of

tips, evolutionary time and the flexible temporal scheme delimiting the number of genera. With these parameters, trees were simulated to generate 1000 tips evolving in 15 Myr, λ was set to 0.9 and μ to 0.05. An approximate stem age interval between 4 and 5 Myr was then used to delimit 34 monophyletic clades or hypothetical genera, each of which randomly included a number of tips, ranging from 1 to 166.

Next, DNA sequence evolution was simulated along the 10 generated trees. We simulated two scenarios: 1) a phylogenetic dataset resembling our empirical data, and 2) a phylogenomic dataset to test the backbone-barcode imbalances in large-scale studies. We used the software Seq-Gen (Rambaut and Grassly 1997) to simulate evolution across molecular markers. For the phylogenetic dataset, Seq-Gen was run independently eight times to simulate evolution in the molecular markers commonly used in Polyommatina, *COI*, *COII*, *EF*, *CAD*, *Wg*, *H3*, *ITS2* and *28S* (with the exception of the short mitochondrial leu-tRNA fragment). Parameters for each marker were extracted from likelihood estimations in JModeltest in the empirical dataset, and are shown in the Supplementary Table S2. The eight generated alignments per tree were then concatenated in matrices of 6660 bp, as a complete (backbone) matrix model (Fig. 1a). Barcode datasets were also generated with COI alignments, as a single-gene (barcode) matrix model (Fig. 1b). For the phylogenomic dataset, we simulated evolution in 100 genes, where values for Seq-Gen parameters for each marker were randomly assigned to values within the range of those used in the empirical dataset. The concatenation of the 100 generated alignments resulted in backbone matrices of 81,354 bp.

In order to test for effects of non-barcode presence/missing data on phylogenetic (-omic) performance, we also built 5 datasets where we progressively increased the percentage of representation by additional (non-barcode) markers by 5%, 10%, 25%, 50% and 75% (Fig. 2). The selection of tips represented by these markers followed two strategies: 1) a random selection per each percentage and 2) a guided selection per each percentage prioritizing one addition per genus, thus discarding already represented genera (until all genera were represented). Thus, for each of the 10 simulated trees, we produced 12 matrices ranging from 0% of to 100% of non-barcode data. Specifically, we generated one barcode matrix, one complete matrix, 5 matrices with random selection of tips with multi-gene data and 5 matrices with guided selection of tips with multi-gene data. Overall, this procedure generated a total of 120 simulated molecular matrices for the phylogenetic dataset, and 120 for the phylogenomic dataset.

Phylogenetic inference for all matrices in the simulated phylogenetic dataset were conducted using Maximum Likelihood in RAxML v.8 (Stamatakis 2014). We used the GTRCAT model of nucleotide evolution and conducted a rapid bootstrap analysis with 100 iterations and a search for the best scoring tree in a single run (-f a). For the phylogenomic dataset, Maximum Likelihood phylogenetic inference was conducted using IQtree v.2 (Minh et al. 2020), as described for the empirical dataset. All resulting trees were also posteriorly collapsed into genus-level trees (where intrageneric tips were collapsed into a single branch) according to each of the reference simulated trees.

Tree evaluation (empirical and simulated)

The resulting phylogenetic trees, both from empirical and simulated datasets, were evaluated along four different axes: 1) percentage of nodes correctly resolved, 2) relative branch lengths differences using the K tree score (K) (Soria-Carrasco et al. 2007), 3) bootstrap values as an average of all nodes (for simulations only) and 4) degree of success in recovering monophyletic genera.

For the empirical datasets, we scored the percentage of matching nodes and K score between the combined and barcode trees. We also scored these metrics for the genus-level barcode tree and for the genus-level combined tree, always taking the genus-level backbone tree as a reference. Values for both genus-level and species-level trees allowed us to discern higher-level (between genera or deeper) and lower-level (intra-generic) topological differences. The degree of success in recovering monophyletic genera was also compared between the combined and barcode trees, using the function "AssessMonophyly" in the R package MonoPhy (Schwery and O'Meara 2016). For the battery of simulations, the percentage of nodes correctly resolved and K were also retrieved for both genus-level and species-level trees, taking each corresponding simulated tree as a reference (Fig. 2).

RESULTS

Empirical phylogenetics

At the genus-level, the percentage of nodes matching the backbone tree was higher for the combined trees (81.25% in BEAST, 71.87% in IQtree and 43.75% in RAxML) than for the barcode trees (12.5% in BEAST, 25% in IQtree and 21.87% in RAxML)

(Supplementary Table S3). Assuming that the backbone tree provides the best phylogenetic hypothesis, these results indicate a substantial improvement in phylogenetic resolution for each of the three methods when comparing the combined tree with the barcode tree, even though only 8% of the specimens, representing all genera, incorporated additional, non-barcode data. A similar trend of improvement was observed for relative branch length comparisons, where lower K scores and scale-factors closer to one indicate branch lengths that are more similar to each other between two trees (Table S3). According to this metric, the combined tree was also more similar to the backbone tree (K=1.57/0.006/0.10; scale-factor=1.05/0.92/0.70) than was the barcode tree (K=15.85/0.08/0.14; scale-factor=0.91/0.54/0.30).

At the species-level, the percentage of nodes recovered in both the combined tree and the barcode tree was 42.33% in BEAST, 65.81% in IQtree and 57.52% in RAxML, indicating that there were meaningful differences between the two datasets in the phylogenetic relationships recovered within each genus. Differences between the three tree inference methods used to resolve topologies and relative branch lengths were appreciable, which may be related to the number of unresolved nodes. No supported changes in topology at the genus-level could be detected in ML trees of the combined dataset when we compared non-partitioned analyses with analyses partitioned by codon position (Supplementary Figs. S2, S3). The only observed differences were associated with nodes that repeatedly showed low support across all methods used.

When testing for inconsistencies in resolving monophyletic genera using data from only the barcode tree, we determined that the barcode tree failed to cluster 5 of the 34 genera in the BEAST tree, 4 genera in the IQtree tree, and 3 genera in the RAxML tree, while

400 the combined tree failed to cluster only one genus in the RAxML tree (Supplementary
401 Table S4).
402
403 In an initial exploratory step, the combined tree recovered 4 taxa that each had an
404 unexpected placement or divergence that violated the criteria applied to delimit genera
405 in Polyommatina suggested by Talavera *et al*. 2013 (i.e. divergencies of <4-5 Myr).
406 These four taxa were represented only by barcodes in the analysis. Since taxonomic
407 changes might be required for these four taxa, we increased their molecular
408 representation by sequencing the same additional genes for them that were included in
409 the backbone database. Taxonomic decisions were then applied based on a tree
410 incorporating these additional sequences (Fig. 3, Supplementary Fig. S1).
411
412 *Neolysandra corona* was confirmed to be nested within *Polyommatus*, and thus we
413 transferred the taxon *corona* to *Polyommatus*.
414
415 *Kretania psylorita*'s divergence (4.02 Myr) fell within the flexible temporal scheme of
416 4-5 Myr, and thus we retained *psylorita* together with the rest of the taxa within
417 *Kretania*, as defined here. However, the genus was not well supported and relationships
418 shifted depending upon the method of phylogenetic reconstruction. Since *psylorita* is
419 the type-species of the genus *Kretania*, this could have taxonomic consequences, but
420 for now we have opted for the topology most frequently recovered, which is also in
421 keeping with the morphology-based classification.
422
423 Divergences for *C. elicola* (6.68 [4.58-9.01] Myr) and *C. kedonga* (8.21 [5.65-10.77]
424 Myr) were considerably older than 5 Myr, ages that in both cases indicated the need for

a description of new, monotypic, genus. We describe these two new genera as *Birabiro* **gen. nov.** (type species *elicola*) and *Kipepeo* **gen. nov.** (type species *kedonga*) (See Appendix).

Finally, we use these results to propose a full division into subgenera of the large genus *Polyommatus*, including the description of three new subgenera: *Escherilycaena* **subgen. nov**., *Amandolycaena* **subgen. nov**., and *Iranolysandra* **subgen. nov.** This new phylogenetic classification helps to resolve other debated cases such as that of *Chilades parrhasius*, which is transferred to *Luthrodes* (see Supplementary Information for the full taxonomic description and discussion).

Simulations

The phylogenetic consequences of combining different sequencing strategies to infer higher-level systematics were further evaluated using simulated experiments. The proportion of nodes that were correctly resolved in genus-level trees increased with the percentage of fully-sequenced tips in the matrices (Fig. 4a). For the phylogenetic dataset, the proportion of nodes that were correctly resolved was 68.75% on average for the barcode datasets and reached a peak value of 94.06% for the combined datasets, whereas for the phylogenomic dataset, these values ranged from 73.44% for the barcode datasets to 99.38% for the combined datasets.

The improvement curve was optimized when the selection of tips was guided to include one tip per genus (Fig. 4a). In these cases, combined trees having only 5% of fully-sequenced tips (90.94% and 99.38% of correct nodes in the phylogenetic and

450 phylogenomic datasets, respectively), or 10% of fully-sequenced tips (93.12% and
451 98.75% of correct nodes) already produced comparable topologies to the ones resulting
452 from complete matrices with 100% of fully-sequenced tips (93.44% and 99.38% of
453 correct nodes) (Fig. 4a). This was not the case using a random selection strategy, where
454 equivalent topologies were achieved only when 50% of fully-sequenced tips were
455 included (93.12% and 98.44% of correct nodes), percentages that were likely to have
456 included at least one representative per genus by chance (Fig. 4a). In species-level trees,
457 a progressive improvement of phylogenetic accuracy was also observed, but only
458 reached the optimal when trees were reconstructed using 100% of the data (Fig. 4a).
459 The percentage of correctly resolved nodes ranged from 86.1% in the barcode trees to
460 96.8% in the complete trees for the phylogenetic dataset, and from 86% to 99.45% in
461 the phylogenomic dataset.
462
463 Tree shape as indicated by relative branch length assessments performed similarly to
464 topological assessments (Fig. 4b). In these comparisons, higher K scores indicate more
465 disparate branch lengths than lower K scores. For genus-level trees, an improvement
466 (decrease) of K was generally observed when guided fully-sequenced tips were
467 progressively added: when none of these were present, K=7.07 and K=6.58 for
468 phylogenetic and phylogenomic datasets respectively, whereas when only 5% of fully-
469 sequenced tips representing each genus were added, these values dropped to K= 2.97
470 and K= 1.25 respectively. The latter values were already quite close to those obtained
471 when 100% of taxa were fully sequenced, with K=2.44 and K=0.85 respectively (Fig.
472 4b).
473

474 This rapid convergence in branch length differences did not occur for randomly selected,
475 fully-sequenced tips, where K only approached optimal values when 100% of fully-
476 sequenced tips were included (K=2.43 and K=0.85) (Fig. 4b).
477
478 When assessing lower-level phylogenetic relationships with species-level trees, K
479 scores showed a similar pattern with a progressive improvement from 0% of fully-
480 sequenced tips where K=13.74 and K=13.35 for phylogenetic and phylogenomic
481 datasets respectively, to 100% of fully-sequenced tips where K=5.77 and K=2.12 (Fig.
482 4b). Random and guided selections did not show substantial differences in this case,
483 suggesting that a guided selection of fully-sequenced tips is mainly of benefit in
484 resolving deeper level phylogenetic relationships.
485
486 Bootstrap values rapidly increased on average from the barcode datasets (57.88%) to
487 the combined datasets, with 5% of fully-sequenced tips (87.16% for guided selection
488 and 77.16% for random selection) in the genus-level trees of the phylogenetic dataset
489 (Fig. 4c). Bootstrap values of the phylogenomic datasets increased from 92.51% to
490 98.12% (guided selection) and 94.83% (random selection) (Fig. 4c). Bootstrap values
491 of the species-level trees showed a progressive improvement as fully-sequenced tips
492 were incorporated, independent of the sampling strategy (Fig. 4c). Although the
493 average bootstrap does not provide information about which set of nodes contribute the
494 most to the topological changes observed, the patterns are consistent between all of
495 these indices, and give no indication that a few nodes might be strongly biasing the
496 results.
497

The number of monophyletic genera in the phylogenetic datasets increased with the number of fully-sequenced tips, starting with an average of 9.7% of non-monophyletic genera out of 34 (involving on average 13.4% of affected tips) in barcode trees to 0.6% non-monophyletic genera (involving 0.9% of affected tips) in complete (100% gene sampling) trees (Fig. 4d). Values in the phylogenomic datasets ranged from an average of 9.1% of non-monophyletic genera (involving 11.1% of affected tips) to none (Fig. 4d). No substantial differences were detected between randomly and guided selection strategies. Fewer genera were recovered as non-monophyletic in the simulations than in the empirical dataset, highlighting the simplicity of simulated evolution against the complexity of real evolutionary processes in nature. Nevertheless, the simulations show cases of tips that are hard to place into the right genera, possibly due to effects of short internode branching patterns or of 'singletons', genera represented by a single terminal species, either because of poor sampling or because monotypic lineages can be grouped together erroneously due to long branch attraction.

DISCUSSION

Robustness of the combined approach

All tree evaluation methods assessed, both for empirical and simulated data, show important improvements in phylogenetic accuracy when progressively increasing fully-sequenced tips (Fig. 4, Supplementary Table S3). Topology, bootstrap support, and concordance in relative branch lengths are particularly strengthened when fully-sequenced tips are not added randomly, but are selected with the goal of representing at least one tip per genus (Fig. 4). Taxonomically balanced, multi-gene phylogenetic

information seems efficient at counteracting the leading signal of the single-gene COI history in the combined phylogenies. Trees with 5%-10% fully-sequenced tips are comparable to those with 100% fully-sequenced tips, but not to trees inferred from only barcodes. Interestingly, this effect mostly applies to deeper level phylogenetic relationships (i.e., genus-level trees) (Fig. 4).

The K score can be interpreted as a proxy for divergence time estimates. Missing data have previously been estimated to have little influence in the accuracy of divergence dating in BEAST (Zheng and Wiens 2015). Our empirical results also show little difference in divergence times when comparing the backbone and the taxonomy-guided combined datasets. This is also reflected in the simulations, which achieve near optimal values at 10% sampling provided fully-sequenced tips are selected to be representative of each genus. However, datasets where fully-sequenced tips are added randomly do not achieve optimal values until sampling is 100% complete (Fig. 4).

The placement of species into genera with which they are traditionally associated is reflected by the number of monophyletic genera recovered by an analysis. Our empirical data show that taxa are likely to be misplaced into genera with which they are not normally associated in phylogenies based exclusively on data from COI-based barcodes, with up to five genera recovered as non-monophyletic (affecting 768 of the tips of the tree) (Supplementary Table S4). However, inaccurate placements are reduced in phylogenies based on the combined datasets. The same result is obtained with simulations, where the number of monophyletic genera improves progressively with the addition of fully-sequenced tips (Fig. 4d).

548 Studies carried out by Cho et al. (2011) and Kawahara et al. (2011) show at the order
549 and family level respectively that increased gene sampling improves estimates of deep
550 relationships as indicated by higher support values. Our simulated findings are
551 generally compatible with these results (Fig. 4), which have also been observed in
552 multiple other phylogenies when increasing the number of characters (Rokas et al.
553 2003; Bapteste et al. 2002; Dunn et al. 2008; Zwick et al. 2011; Wilson 2011; Wilson
554 et al. 2011; Kuntner et al. 2019).
555
556 The simulated phylogenomic analyses (Fig. 4) show that datasets with large barcode
557 representation can be successfully combined with modern genomic datasets where taxa
558 have been sampled for a large number of genes. The overall performance of the
559 simulated combined phylogenomic dataset (100 genes + barcode) is better than that of
560 the phylogenetic combined dataset (7 genes + barcode), as expected by the much greater
561 number of characters. Again, in order to produce the best possible trees, it is key that
562 taxa with genomic data represent a diversity of higher-level taxonomic categories. Thus,
563 phylogeneticists are encouraged —and many do so instinctively— to strategically
564 design their sampling to include 1) taxonomically distributed and representative species
565 characterized with genomic data as well as 2) well sampled barcode data from
566 individuals representing as many species as possible in order to recover large-scale
567 phylogenetic relationships.
568
569 **DNA barcodes as a tool for higher-level systematics: a new value**
570
571 A great many DNA barcodes representing a wide array of organisms have been
572 generated and deposited in public repositories in recent years. Several markers function

as DNA barcodes, with mitochondrial COI typically representing animals, and others such as ITS2 representing fungi, *rcbl* or *matK* representing plants, and 16S rRNA representing bacteria. To date, nearly 9.2 million barcoded specimens are available on the BOLD database, and nearly 4 million can be extracted from GenBank for COI.

Potential applications of DNA barcodes are varied. First, they have been used as references for species-level identification since their conception (Hebert et al. 2003), and their impact on taxonomy is undeniable (Miller 2007; Huber and Hanner 2015; Dincă et al. 2015; Miller et al. 2016). Conceptual variations of the initial DNA barcode idea such as metabarcoding have expanded into many other fields of molecular ecology and community ecology (Creer et al. 2016). DNA barcodes are also widely applied in phylogeography and surveys of intraspecific variability. After much initial debate, it is now well established that DNA barcoding (and any other single-marker approach) can be a useful tool to identify potential cryptic species, although an integrative approach is necessary for confirmation (e. g. nuclear markers, morphology, ecology) (Will et al. 2005; DeSalle et al. 2005; Talavera et al. 2013b; Dincă et al. 2015; Hernández-Roldán et al. 2016; Lukhtanov et al. 2016; Gaunet et al. 2019).

Few studies have assessed whether DNA barcodes can be helpful at placing unidentified species into higher-level taxonomic categories (Wilson et al. 2011; Coddington et al. 2016). Here we show that DNA barcoding can potentially be applied to assign taxa to genera (or higher categories) provided a solid and representative higher-level backbone phylogeny exists. Our results indicate that large datasets of barcodes can be used to identify cases where taxa have been wrongly assigned to higher-level taxonomic categories, a frequent problem in diverse groups with complex

598 taxonomy, where synapomorphies helping to delineate genera have been difficult to
599 find.
600
601 In the case where potential higher-level cryptic taxa are indicated by the results, these
602 can be the focus of further taxonomic assessments following standard principles of
603 phylogenetic systematics, such as the addition of molecular characters that aid in
604 phylogenetic placement. Our proposed workflow for phylogenetic systematic
605 assessments (Fig. 5, Supplementary Information) takes advantage of the huge number
606 of sequenced specimens available in public databases with the aim of accelerating
607 taxonomic resolution at higher-taxonomic levels. It may facilitate molecular-based
608 taxonomy in research labs where phylogenomic techniques are not yet easily available
609 and, ultimately, benefit the common goal of taxonomic stability.
610
611 **Figure 5**. Diagram of the proposed workflow for higher-level systematic assessments.
612
613 **Conclusions**
614 Phylogenetic inference based exclusively on DNA barcodes has been shown, both here
615 and elsewhere, to perform poorly. However, we show how in combination with a
616 backbone of carefully sampled, representative taxa for which a large number of
617 additional markers have been sequenced, these short barcode sequences can
618 nevertheless be used effectively to produce reliable phylogenies and improve higher-
619 level systematics in large datasets. Our simulation tests show that a multi-gene
620 sampling for as few as 5-10% of the specimens in the total dataset can produce high-
621 quality phylogenies, comparable to those resulting from 100% multi-gene sampling,
622 provided a strategic selection has been made of higher-level representatives for multi-

gene sequencing (approximately one per genus). These results are found at both a phylogenetic and phylogenomic scale, thus accounting for a wide range of imbalance in the number of characters between the combined barcode and backbone matrices. Thus, as long as backbone matrices are taxonomically representative, data coming from probe capture, transcriptomic or genomic techniques can be effectively combined with barcodes to generate phylogenetically accurate, large-scale molecular characterizations of biodiversity.

ACKNOWLEDGEMENTS

We thank many colleagues who collected material used in this study, including: D. Benyamini, F. Bolland, C. Castelain, S. Cuvelier, L. Dapporto, V. Dincă, V. Doroshkin, V. Doroshkin, K. Dovgailo, Ph. George, J. Hernández-Roldán, E. Ivanova, M. Khaldi, R. Khellaf, T. Larsen, M. Markhasiov, D.J. Martins, I.N. Osipov, O. Pak, V. Patrikeev, N. Rubin, P. Stamer, M.R. Tarrier, V. Tikhonov, S. Toropov, A. Ugarte, J. Verhulst and R. Vodă. Our special thanks are to Blanca Huertas for taking pictures of type specimens in the Natural History Museum in London. This work was funded by projects PID2019-107078GB-I00 / AEI / 10.13039/501100011033 and 2017-SGR-991 (Generalitat de Catalunya) to RV and GT, and by the Committee for Research and Exploration of the National Geographic Society (grant WW1-300R-18) to GT, by the Putnam Expeditionary Fund of the Museum of Comparative Zoology (to all authors) and the U.S. National Science Foundation under DEB-0447244, and DEB-1541560 (to NEP). GT was supported by the Ramón y Cajal programme of the Spanish Ministry of Science and Innovation (RYC2018-025335-I). Taxonomic studies and descriptions of new genera and subgenera were supported by the Russian Science Foundation grant N 19-14-00202 to the Zoological Institute of the Russian Academy of Sciences (to VL).

REFERENCES

Bálint Z., Jonson K. 1997. Reformation of the Polyommatus section with a taxonomic and biogeographic overview (Lepidoptera, Lycaenidae, Polyommatini). Neue Ent. Nachr. 40:1–68.

Bapteste E., Brinkmann H., Lee J.A., Moore D.V., Sensen C.W., Gordon P., Duruflé L., Gaasterland T., Lopez P., Müller M., Philippe H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. U.S.A. 99:1414–1419.

Cho S., Zwick A., Regier J.C., Mitter C., Cummings M.P., Yao J., Du Z., Zhao H., Kawahara A.Y., Weller S., Davis D.R., Baixeras J., Brown J.W., Parr C. 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60:782–796.

Creer S., Deiner K., Frey S., Porazinska D., Taberlet P., Thomas K., Potter C., Bik H. 2016. The ecologist's field guide to sequence-based identification of biodiversity. Meth. Ecol. Evol. 7:1008–1018.

Coddington J.A., Agnarsson I., Cheng R-C., Candek K., Driskell A., Frick H., Gregoric M., Kostanjsek R., Kropf C., Kweskin M., Lokovsek R., Pipan M., Vidergar N., Kuntner M. 2016. DNA barcode data accurately assign higher spider taxa. PeerJ. 4:e2201.

673 Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new
674 heuristics and parallel computing. Nature Methods. 9:772.
675
676 Degnan J., Rosenberg N.A. 2006. Discordance of species trees with their most likely
677 gene trees. PLoS Genet. 2:e68.
678
679 de Queiroz A., Gatesy J. 2006. The supermatrix approach to systematics. Trends Ecol.
680 Evol. 22:34–41.
681
682 DeSalle R., Egan M.G., Siddall M. 2005. The unholy trinity: taxonomy, species
683 delimitation and DNA barcoding. Philos. Trans. Royal Soc B. 360:1905–1916.
684
685 Dincă V., Montagud S., Talavera G., Hernández-Roldán J., Munguira M.L., García-
686 Barros E., Hebert P.D.N., Vila R. 2015. DNA barcode reference library for Iberian
687 butterflies enables a continental-scale preview of potential cryptic diversity. Scientific
688 Reports. 5:12395
689
690 Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics
691 and dating with confidence. PLOS Biol. 4, e88.
692
693 Drummond A.J., Suchard M.A., Xie D., Rambaut, A. 2012 Bayesian phylogenetics
694 with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.
695

696 Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse
697 GW, Obst M, Edgecombe GD, et al: Broad phylogenomic sampling improves
698 resolution of the animal tree of life. Nature 2008. 452:745–749.
699
700 Eckweiler W., Bozano G.C. 2016. Guide to the Butterflies of the Palearctic Region:
701 Lycaenidae. Part IV. Milano, pp.132.
702
703 Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Phil.
704 Trans. Roy. Soc. B 363:4023–4029.
705
706 Gaunet A., Dincă V., Dapporto L., Montagud S., Vodă R., Schär S., Badiane A., Font,
707 E., Vila R. 2019. Two consecutive *Wolbachia*-mediated mitochondrial introgressions
708 obscure taxonomy in Palearctic swallowtail butterflies (Lepidoptera, Papilionidae).
709 Zool. Scr. 48:507–519
710
711 Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic
712 problem? Syst. Biol. 47:9–17.
713
714 Grievink L., Penny D., Holland B.R. 2013. Missing data and influential sites: Choice
715 of sites for phylogenetic analysis can be as important as taxon sampling and model
716 choice. Genome Biol. Evol. 5:681–687.
717
718 Guindon S., Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large
719 phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
720

721 Hebert P.D.N., Cywinska A., Ball S.L., deWaard J.F. 2003. Biological identifications
722 through DNA barcodes. Proc. R. Soc. Lond. B 270:313–321.
723
724 Hernández-Roldán J.L., Dapporto L., Dinca V., Vicente J.C., Hornett E.A., Šichová J.,
725 Lukhtanov V., Talavera G., Vila R. 2016. Integrative analyses unveil speciation linked
726 to host plant shift in *Spialia* butterflies. Mol. Ecol. 25:4267–4284.
727
728 Hey J., Waples R.S., Arnold M.L., Butlin R.K., Harrison R.G. 2003. Understanding
729 and confronting species uncertainty in biology and conservation. Trends Ecol. Evol.
730 18:597–603.
731
732 Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh, L.S. 2018. UFBoot2:
733 Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522
734
735 Hovmöller R., Knowles L.L., Kubatko L.S. 2013. Effects of missing data on species
736 tree estimation under the coalescent. Mol. Phylogenet. Evol. 69:1057–1062.
737
738 Hubert N., Hanner R. (2015). DNA Barcoding, species delineation and taxonomy: a
739 historical perspective. DNA Barcodes. 3:44–58.
740
741 Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning
742 of incongruence? Trends Genet. 22:225–231.
743
744 Jiang W., Chen S.Y., Wang H., Li D.Z., Wiens J.J. 2014. Should genes with missing
745 data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol. 80:308–318.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., L.S. Jermiin L.S. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods, 14:587–589.

Kandul N.P., Lukhtanov V.A., Dantchenko A.V., Coleman J.W.S., Sekercioglu C.H., Haig D., Pierce N.E. 2004. Phylogeny of *Agrodiaetus* Hübner 1822 (Lepidoptera: Lycaenidae) Inferred from mtDNA Sequences of COI and COII and Nuclear Sequences of EF1-α: Karyotype Diversification and Species Radiation. Systematic Biology. 53:278–298.

Kawahara A.Y., Ohshima I., Kawakita A., Regier J.C., Mitter C., Cummings M.P., Davis D.R., Wagner D.L., De Prins J., Lopez-Vaamonde C. 2011. Increased gene sampling strengthens support for higher-level groups within leaf-mining moths and relatives (Lepidoptera: Gracillariidae). BMC Evol. Biol. 11:182.

Kuntner M., Hamilton C.A., Ren-Chung C., Gregoric M., Lupse N., Lokovsek T., Lemmon E.M., Lemmon A.R., Agnarsson I., Coddington J.A., Bond, J. 2019. Golden Orbweavers ignore biological rules: phylogenomic and comparative analyses unravel a complex evolution of sexual size dimorphism. Syst. Biol. 68:555–572.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott,B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution formolecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

771 Layberry R.A., Hall P.W., Lafontaine J.D. 1998. The butterflies of Canada. University
772 of Toronto Press, Toronto, Buffalo, London. 280 pp.
773
774 Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of
775 ambiguous data on phylogenetic estimates obtained by maximum likelihood and
776 Bayesian inference. Syst. Biol. 58:130–145.
777
778 Lukhtanov V.A., Sourakov A., Zakharov E.V. 2016. DNA barcodes as a tool in
779 biodiversity research: testing pre-existing taxonomic hypotheses in Delphic Apollo
780 butterflies (Lepidoptera, Papilionidae). Syst. Biodivers. 14:599–613.
781
782 Maddison W.P. 1997. Gene trees in species trees. Syst Biol 46:523–536.
783
784 Mendes F.K., Hahn M.W. 2018. Why concatenation fails near the anomaly zone. Syst.
785 Biol. 67:158–169.
786
787 Minh B.Q., Schmidt H.A, Chernomor O., Schrempf D., Woodhams M.D., von Haeseler
788 A., Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic
789 inference in the genomic era. Mol. Biol. Evol., msaa015
790
791 Miller S.E. 2007. DNA barcoding and the renaissance of taxonomy. Proc. Natl. Acad.
792 Sci. U.S.A. 104:4775–4776.
793
794 Miller S.E., Hausmann A, Hallwachs W, Janzen D.H. 2016. Advancing taxonomy and
795 bioinventories with DNA barcodes. Philos. Trans. Royal Soc B. 371:20150339.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Pelham J.P. 2008, A catalogue of the butterflies of the United States and Canada. J. res. lepid. 40: I-XIII, 1–658.

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of Eukaryotes: Impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Philippe H., Vienne D.M. de, Ranwez V., Roure B., Baurain D., Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. Eur. J. Taxon. 283:1–25.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biology. 9:e1000602.

Poe S. 2003. Evaluation of the strategy of long branch subdivision to improve accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Rambaut A., Grassly N.C. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard MA. (2018) Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.

Rannala, B., Huelsenbeck, J.P., Yang, Z., Nielsen, R., 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rokas A., Williams BL., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Rosenberg M.S., Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. U.S.A. 98:10751–10756.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Rubinoff D., Holland B.S. 2005. Between two extremes: Mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. Syst. Biol. 54:952–961.

Schwery O., O'Meara B.C. 2016. MonoPhy: a simple R package to find and visualize monophyly issues. PeerJ Comput. Sci. 2:e56.

Shapoval N., Lukhtanov V. 2016. On the generic position of *Polyommatus avinovi* (Lepidoptera: Lycaenidae). Folia Biol. (Krakow) 64: 267–273.

Simmons M.P. 2012. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Molecular Phylogenetics and Evolution. 62:472–484.

Simmons M.P. 2012. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics. 28:208–222.

Stadler T. 2011. Simulating trees on a fixed number of extant species. Syst. Biol. 60:676–84.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Streicher J.W., Schulte J.A.II., Wiens J.J. 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65:128–145.

869 Soria-Carrasco V., Talavera G., Igea J., Castresana J. 2007. The K tree score:
870 quantification of differences in the relative branch length and topology of phylogenetic
871 trees. Bioinformatics. 23:2954–2956.
872
873 Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-
874 analysis of large phylogenies. Bioinformatics, 30:1312–1313.
875
876 Stradomsky B.V. 2016. A molecular phylogeny of the subfamily Polyommatinae
877 (Lepidoptera: Lycaenidae). Caucas. Entomol. Bull. 12:145–156
878
879 Talavera G., Lukhtanov V.A., Pierce N.E., Vila R. 2013a. Establishing criteria for
880 higher-level classification using molecular data: the systematics of *Polyommatus* blue
881 butterflies (Lepidoptera, Lycaenidae). Cladistics. 29:166–192.
882
883 Talavera G., Dincă V., Vila R. 2013b. Factors affecting species delimitations with the
884 GMYC model: insights from a butterfly survey. Methods Ecol. Evol. 4:1101–1110.
885
886 Talavera G., Kaminski L.A., Freitas A.V.L., Vila R. 2015. One-note samba: the
887 biogeographical history of the relict Brazilian butterfly *Elkalyce cogina*. J. Biogeogr.
888 43:727–737.
889
890 Vila R., Bell C.D., Macniven R., Goldman-Huertas B., Ree R.H., Marshall C.R., Balint
891 Z., Johnson K., Benyamini D., Pierce N.E. 2011. Phylogeny and palaeoecology of
892 *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the
893 New World. Proc. R. Soc. Lond., B, Biol. Sci. 278:2737–2744.

Wiemers M., Stradomsky B.V., Vodolazhsky D.I. 2010. A molecular phylogeny of *Polyommatus* s. str. and *Plebicula* based on mitochondrial COI and nuclear ITS2 sequences (Lepidoptera: Lycaenidae). Eur. J. Entomol. 107:325–336.

Wiens J.J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiens J.J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54:731–742.

Wiens J.J., Fetzner J.W., Parkinson C.L., Reeder T.W. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54, 719–748.

Wiens J.J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39:34–42.

Wiens J.J., Morrill M.C. 2011. Missing data in phylogenetic analysis: Reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.

Wiens J.J., Tiu J. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS ONE. 7:e42925.

Will K.W., Mishler B.D., Wheeler Q.D. 2005. The perils of DNA Barcoding and the need for integrative taxonomy. Syst. Biol. 54:844–851.

Wilson J., Rougerie R., Schonfeld J., Janzen D.H., Hallwachs W., Hajibabaei M., Kitching I.J., Haxaire J., Hebert P.D. 2011. When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. BMC Ecol. 11:18.

Wilson J.J. 2011. Assessing the Value of DNA Barcodes for molecular phylogenetics: Effect of increased taxon sampling in Lepidoptera. PLoS ONE. 6:e24769.

Zhdanko A.B. 2004. A revision of the supraspecific taxa of the lycaenid tribe Polyommatini (Lepidoptera, Lycaenidae). Entomol. Rev. 84:782–796.

Zheng Y., Wiens J.J. 2016. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. 94:537–547.

Zheng Y., Wiens J.J. 2015. Do missing data influence the accuracy of divergence-time estimation with BEAST? Mol. Phylogenet. Evol. 85:41–49.

Zwick A., Regier J.C., Mitter C., Cummings M.P. 2011. Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera). Syst. Entomol. 36:31–43.

Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

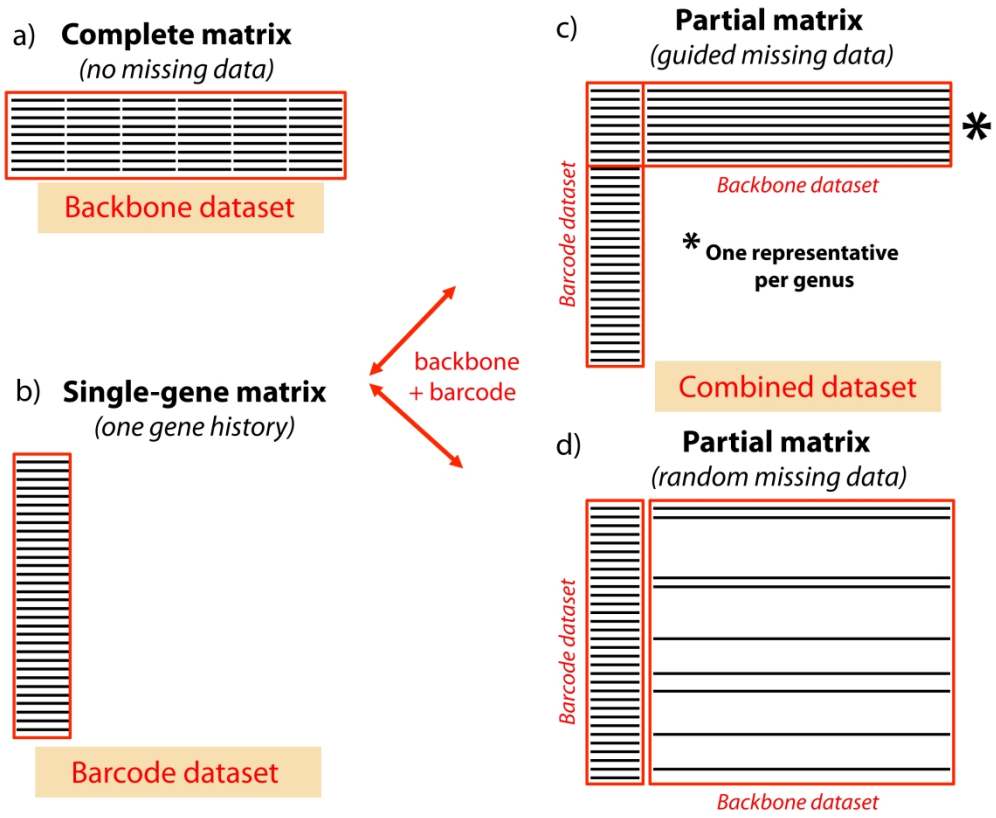

Figure 1. Distribution of missing data in molecular matrices. a) A complete matrix, where no missing data are involved (referred as the backbone dataset). b) A single-gene matrix, including only one molecular marker and therefore providing information about only one gene history (referred as the barcode dataset). c) The combined matrix, the product of merging a backbone and a barcode dataset, where one marker (DNA-barcode) is present for all specimens, but other markers are entirely sampled only for a reduced percentage of specimens that have been selected by prioritizing representatives of higher-level taxonomic categories. d) A combined matrix, where the selection of fully-sequenced specimens, and therefore the distribution of missing data, is randomly sampled (usually as a result of merging datasets from various sources).

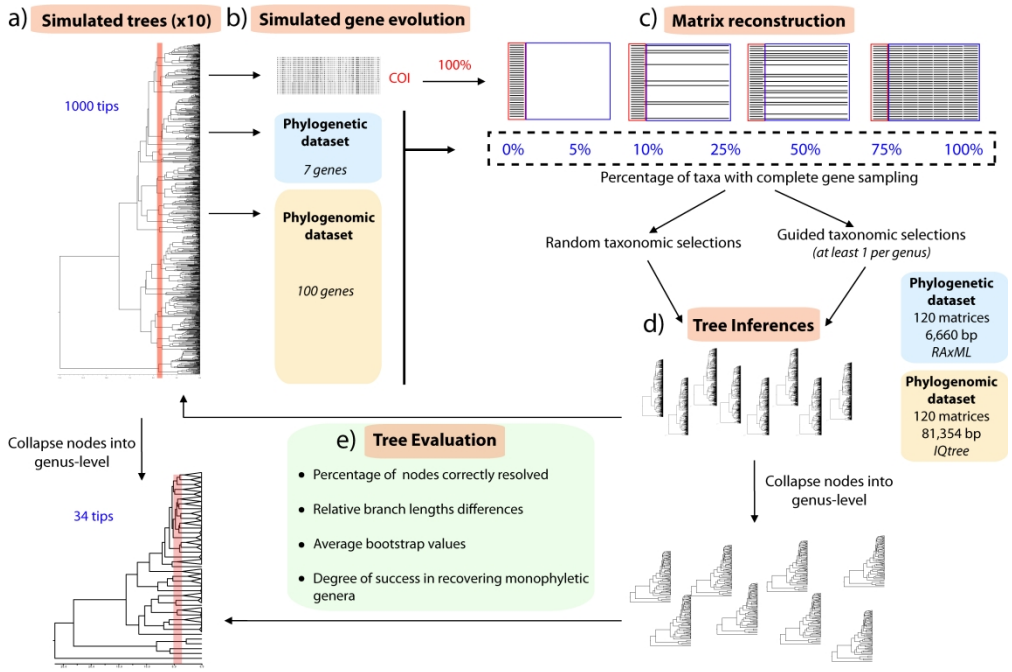

Figure 2. Diagram representing the designed simulation experiments. a) Ten reference trees of 1000 tips were simulated using TreeSim. b) Sequence evolution for independent markers were simulated along trees with SeqGen to generate two datasets: phylogenetic datasets including 7 markers + 1 barcode, and phylogenomic datasets including 100 markers + 1 barcode. c) A matrix reconstruction procedure produced matrices including different fractions of non-barcode fully-sequenced tips (0%, 5%, 10%, 25%, 50%, 75%, 100%). Two strategies in selecting these tips were tested: random selections vs guided selections, where new additions of fully-sequenced tips prioritized representatives for each genus. All matrices included barcode data for all tips. d) Tree inference for all generated matrices were performed, using RAxML for phylogenetic datasets (120 matrices, 6,660 bp each), and IQtree for phylogenomic datasets (120 matrices, 81,354 bp each). Trees were collapsed into single branches at nodes defining genera, thus generating genus-level trees. e) Phylogenetic performance for both species-level and genus-level trees were evaluated against the originally simulated reference trees. Tree evaluation metrics included the proportion of correctly resolved nodes, relative branch-length differences, averaged bootstrap values and degree of success in recovering monophyletic genera (for species-level trees only).

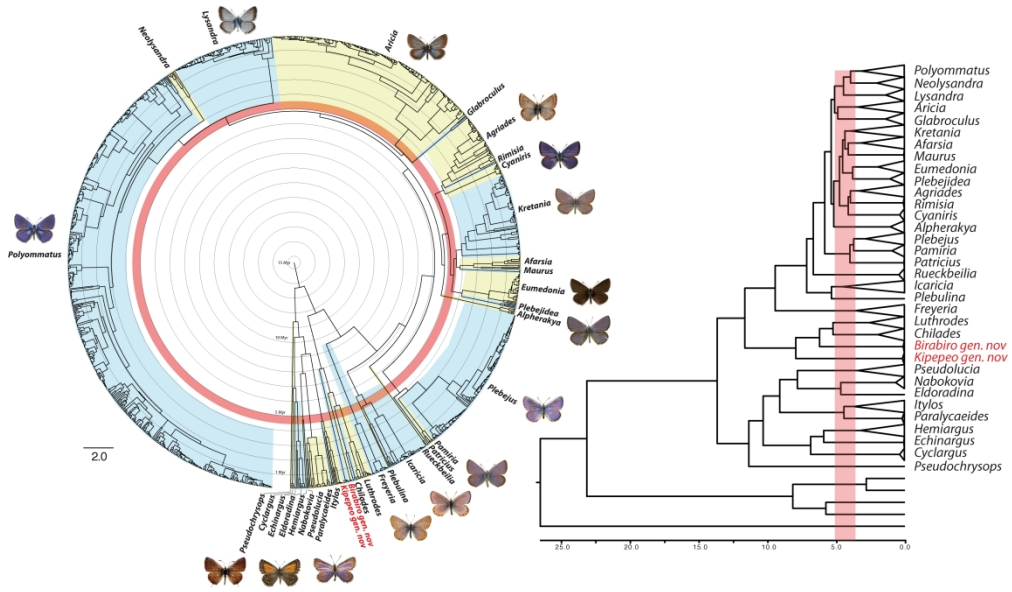

Figure 3. BEAST tree for the species-level dataset of Polyommatina butterflies (1365 specimens – ca. 80% of all taxa (left), and genus-level tree where nodes are collapsed into a single branch per genus (right), both showing the temporal banding used as a threshold for genus delimitation.

295x172mm (300 x 300 DPI)

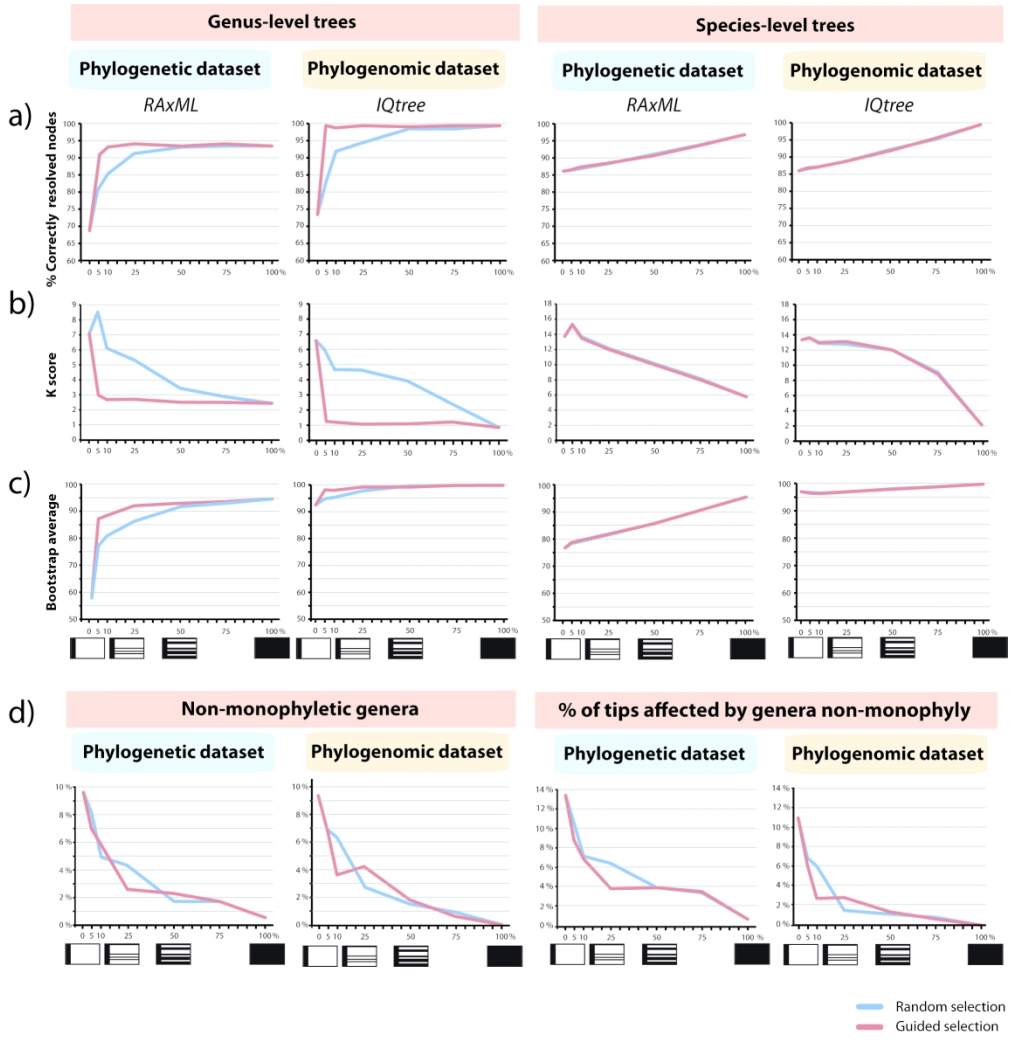

Figure 4. Phylogenetic performance evaluation from simulation experiments along a progressive percentage of fully-sequenced representatives in combination with barcode datasets: a) Percentage of correctly resolved nodes, b) relative branch-length/divergence times (K score), c) average bootstrap values and d) monophyly assessment for genera (number of non-monophyletic genera and number of affected tips). Evaluation metrics are shown for both the species-level trees (1000 tips) and the genus-level trees (34 tips, after collapsing nodes defining genera into a single branch). Values from randomly selected representatives with all markers (in blue) and guided selection strategy using one representative per genus (in red) are compared. Results for a phylogenetic dataset (7 markers + 1 barcode), and a phylogenomic dataset (100 markers + 1 barcode) are shown. Trees were inferred using RAXML for the phylogenetic dataset and IQtree for the phylogenomic dataset.

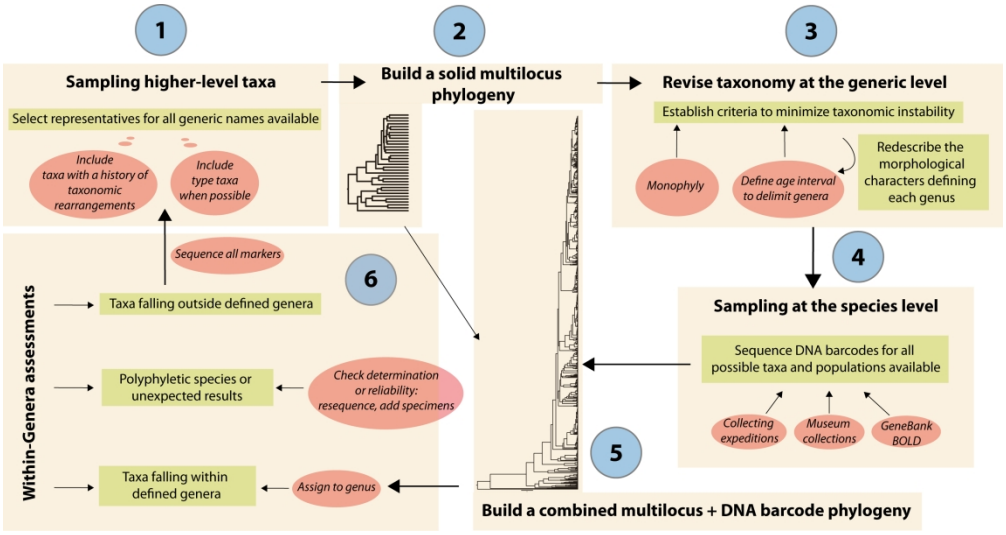

Figure 5. Diagram of the proposed workflow for higher-level systematic assessments.