Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens

CHRISTOPH MAYER1, LARS DIETZ2, ELSA CALL3, SANDRA KUKOWKA4, SEBASTIAN MARTIN5 and MARIANNE ESPELAND2

1Statistical Phylogenetics and Phylogenomics, Zoological Research Museum Alexander Koenig, Bonn, Germany, 2Arthropoda Department, Zoological Research Museum Alexander Koenig, Bonn, Germany, 3Department of Biology, Lund University, Lund, Sweden, 4Molecular Laboratory, Zoological Research Museum Alexander Koenig, Bonn, Germany and 5Computational Genomics, Zoological Research Museum Alexander Koenig, Bonn, Germany

Abstract. Museum collections around the world contain billions of specimens, including rare and extinct species. If their genetic information could be retrieved at a large scale, this would dramatically increase our knowledge of genetic and taxonomic diversity information, and support evolutionary, ecological and systematic studies. We here present a target enrichment kit for 2953 loci in 1753 orthologous nuclear genes + the barcoding region of cytochrome C oxidase 1, for Lepidoptera and demonstrate its utility to obtain a large number of nuclear loci from dry, pinned museum material collected from 1892 to 2017. We sequenced enriched libraries of 37 museum specimens across the order Lepidoptera, many from higher taxa not yet included in high-throughput molecular studies, showing that our kit can be used to generate comparable data across the order, and provides resolution both for shallower and deeper nodes. The filtered datasets (172 taxa, 234 464 amino acid positions and corresponding nucleotides from 1835 CDS regions) were used to infer a phylogeny of Lepidoptera, which is largely congruent in topology to recent phylogenomic studies, but with the addition of some key taxa. We furthermore present our TEnriAn (Target Enrichment Analysis) workflow for processing and combining target enrichment, transcriptomic and genomic data.

Introduction

Billions of animal and plant specimens are housed in museum collections around the world (Wheeler et al., 2012). Collections include rare, endangered and extinct species, some of which can never be collected again. If amenable to DNA sequencing, these legacy specimens already in collections have the potential to dramatically increase the genetic and taxonomic diversity of molecular studies (Novotný & Basset, 2000; Thessen et al., 2012). Such specimens have been an invaluable source for morphological and taxonomic studies for decades, but due to the decay of DNA over time their utility for molecular work has been limited (Yeates et al., 2016). Generating DNA sequence data from old specimens was initially exceedingly laborious involving multiple PCRs of short fragments to assemble a single, short DNA barcode (e.g. Hajibabaei et al., 2006; Strutzenberger et al., 2012; Miller et al., 2013; Hernández-Triana et al., 2014). With the advent of high-throughput sequencing, however, a number of molecular studies have attempted to access the vast amounts of genetic data stored within museum specimens in the burgeoning field coined museomics. Several methods have been tested on old material including shotgun sequencing (Staats et al., 2013; Kanda et al., 2015; Cong et al., 2017a; Li et al., 2019; Jin et al., 2020), target enrichment (Bi et al., 2012; Bailey et al., 2016; Blaimer et al., 2016; Hart et al., 2016; McCormack et al., 2016; StLaurent et al., 2018; Brewer et al., 2019; Derkarebian et al., 2019; Hamilton et al., 2019; Knyshov et al., 2019; Zhang et al., 2020a, b), and combinations of these with restriction site associated sequencing, such as hyRAD and related methods (Tin et al., 2014; Suchan et al., 2016; Linck et al., 2017; Schmid et al., 2017; Gauthier et al., 2020). Several of these studies did not utilize very old specimens (Staats et al., 2013; St Laurent et al., 2018; Wood et al., 2018; Hamilton et al., 2019).
and/or only sequenced mitochondrial genomes (e.g. Timmermans et al., 2016, 2019; Jin et al., 2020; Zhang et al., 2020a). Some methods generate data not comparable with already available data (e.g. available transcriptomes, Sanger sequencing data), such as RADseq-based approaches, which generate data randomly across the genome (Harvey et al., 2016), or ultra-conserved elements (UCE), which utilize mostly non-coding regions (e.g. Blaimer et al., 2016; McCormack et al., 2016; Wood et al., 2018). When using shotgun approaches it might be difficult to assess what coverage would be appropriate, if genome size is unknown. The produced data might therefore be very incomplete and have low cover age if the genome size is large or variable within a group.

For target enrichment methods (Albert et al., 2007; Gnirke et al., 2009; Ng et al., 2009; Meyer &Kircher, 2010; Bi et al., 2012), sequencing efforts are concentrated on the targeted loci, and coverage is therefore generally much higher for the loci of interest compared to random shotgun sequencing approaches. This leads to less missing data and should be ideal for phylogenetic studies (Jones & Good, 2016), especially for museum material. Target enrichment has been shown to work well both for shallow and deep phylogenetics (e.g. Harvey et al., 2016; Espeland et al., 2018; Sann et al., 2018; Dietz et al., 2019; Dömel et al., 2019; Bagley et al., 2020; Choquet et al., 2020). The use of hybrid enrichment methods for sequencing genes of old DNA or museum specimens was proposed by Millar et al. (2008) and Mason et al. (2011), who suggested and used a cross-species capture hybridisation of mitochondrial genomes. Bailey et al. (2016), Blaimer et al. (2016) and McCormack et al. (2016) were among the first to apply the target enrichment method to enrich multiple loci from the nuclear genome of museum specimens, and the recovery rate of genes in Bailey et al. (2016) was on the order of 90%, showing this method to be highly effective. The most widely employed target enrichment methods include exon or gene capture (Gnirke et al., 2009; Ng et al., 2009; Cosart et al., 2011; Bi et al., 2012; Li et al., 2013; Mayer et al., 2016), anchored hybrid enrichment of conserved regions (AHE, Lemmon et al., 2012), and ultraconserved elements (UCE, Faircloth et al., 2012). The methods differ in the targeted regions, which are exons in exon or gene capture, conserved regions in mostly coding genes in AHE studies, and extremely conserved mostly non-coding regions in UCE studies. Consequently, when applying exon capture and to some extent AHE, both the target and the flanking regions (e.g. introns) can be used for phylogenetic analyses (Andermann et al., 2020). If using UCEs, only the flanking regions are useful as there is little to no information in the target region. Since a high proportion of the sequence data obtained with hybrid enrichment methods typically belongs to the target region, the UCE and AHE approaches have the disadvantage that only a small proportion of the generated sequence data is phylogenetically informative. Furthermore, methods that rely on the variable flanking regions have the drawback that the orthology of the flanking regions is only extrapolated from the orthology of the conserved region (Eberle et al., 2020). For old, degraded material, conserved and ultraconserved bait regions are particularly problematic, since their degraded DNA leads to much shorter fragments in the capturing procedure, making it naturally more difficult or impossible to sequence into sufficiently variable flanking regions. Therefore, the usage of AHE and especially UCE is potentially problematic for particularly degraded material, and tend to produce datasets with very little variation when sourcing museum material (e.g. Blaimer et al., 2016; McCormack et al., 2016). Methods that use informative exons as bait regions should not only be more efficient since a much higher proportion of the sequenced data can be used, but also have the important advantage that they can be combined with sequence data mined from transcriptomes. Another advantage of using coding regions is that they can normally be aligned more reliably than non-coding sequences. Among all enrichment approaches mentioned above, only exon capture and to some extent AHE are suited for mixing data from hybrid enrichment and from transcriptomes. Shotgun approaches can be combined with such data if generated with sufficient coverage to recover nuclear genes. Care has to be taken since data from different sources can be systematically biased by the data retrieval step. A potential risk is, for instance, to partially combine the intronic sequences next to the target exon (from enrichment data) and the exon next to the target exon (from transcriptomic data) in the same alignment columns (see Bank et al., 2017).

Since the advent of molecular phylogenetic methods there has been an increasing number of studies attempting to reconstruct the phylogenetic relationships of butterflies and moths (Lepidoptera). Early studies included few genes and few taxa, and as Sanger sequencing techniques developed, the number of taxa increased and eight genes were widely used (e.g. Wahlberg & Wheat, 2008; Mutanen et al., 2010; Sihvonen et al., 2011; Zahiri et al., 2013; Heikkilä et al., 2015; Doorenweerd et al., 2017); although some studies used up to nineteen genes (e.g. Regier et al., 2013, 2015a,b, 2017; Sohn et al., 2013, 2016; Murillo-Ramos et al., 2019). All of these studies had in common that only recently collected and mainly alcohol-preserved material was used for molecular work, since the typical short DNA fragments found in old material are largely not useful for Sanger sequencing. The early attempts using high-throughput sequencing were mainly based on transcriptomes or mitochondrial genomes, which are comparatively easy to sequence in a cost-efficient way. These studies included only few taxa, but often many genes, and especially the transcriptomic studies required fresh material (Bazinet et al., 2013, 2017; Breinholt & Kawahara, 2013; Kawahara & Breinholt, 2014; Timmermans et al., 2014; Kawahara et al., 2019). AHE has also been used successfully for recently collected Lepidoptera material (e.g. Breinholt et al., 2018; Espeland et al., 2018; Homziak et al., 2019; Dowdwy et al., 2020), but see also St Laurent et al. (2018) and Hamilton et al. (2019) for attempts with relatively recent museum material (<35 years old). Shotgun sequencing in Lepidoptera phylogenetic studies has mainly been restricted to butterfly studies (e.g. Cong et al., 2017b; Li et al., 2019; Zhang et al., 2019a,b; Allio et al., 2020), but see Twort et al. (2020). Most recently, a comprehensive higher-level phylogeny of Lepidoptera based on transcriptome data and available genome data was published (Kawahara et al., 2019). Still, higher-level studies have in common that
large parts of the backbone are not well-resolved or supported (Fig. 1). Although taxon sampling has increased, the need for fresh material in transcriptome-based studies or well preserved relatively recent material for AHE, has so far made it impossible to include several highly interesting and rarely collected taxa, for which recently collected material may not be available (Kawahara et al., 2019). Very recently, target enrichment kits specialized for specific Lepidoptera taxa have appeared, such as for butterflies (Espeland et al., 2018), Arctiinae (Erebidae) (Dowdy et al., 2020), Bombycoidea (Hamilton et al., 2019) and Heliconerpa armigera Hüblner (Noctuidae) (McGaughran, 2020), which likely only provide partially overlapping data.

It is now time to add leaves to the Lepidoptera tree using comparable data generated from the millions of specimens deposited in museums around the world. We present a new target enrichment kit (LepZFMK1.0) that can be used to sequence a set of 2953 CDS regions (loci) from 1753 orthologous nuclear genes (and the mitochondrial Cytochrome C Oxidase 1, COI) across Lepidoptera, exhibiting high resolving power at multiple phylogenetic levels. Most loci from Breinholt et al. (2018) and those from Espeland et al. (2018) are included in this kit to assure data comparability. We use the kit to sequence 37 samples of dry, pinned museum specimens with collecting dates ranging from 1892 to 2017, many representing families and superfamilies that have not been included in phylogenomic studies previously. COI is included in the kit and was used to ensure correct identification of specimens. Furthermore, we present an efficient workflow (TEnriAn) for assembling and filtering data, as well as adding data resulting from TE samples to already available transcriptome and genome data. The workflow is available on Github (see data availability statement).

Material and methods

Taxon and gene sampling

Thirty-seven taxa were selected from the Lepidoptera collection of the Zoological Research Museum Alexander Koenig (ZFMK). All specimens were pinned except for the two nymphalid and three lycaenid specimens, which were stored in glassine envelopes. Many were likely rehydrated before pinning, although the exact preparation methods are unknown. Taxa were chosen to cover most major Lepidoptera groups including multiple superfamilies not included in previous phylogenomic studies, both smaller and larger specimens, and a range of sample ages covering more than 100 years, from 1892 to 2017 (See Table S1). Only specimens with information at least on sampling year were chosen, with the exception of one species (Anurapteryx interlineata Walker, Sematuridae), for which no other material was available. In several cases we included two specimens of the same or closely related species, but with different collection years, to confirm the placement of the samples in the phylogeny. For the assignment of species to superfamilies and higher clades we follow Kawahara et al. (2019), and van Nieukerken et al. (2011).

Gene selection for the target enrichment kit

We searched OrthoDB version 9.1 (Zdobnov et al., 2017) for information on orthologous genes for Lepidoptera and Endopterygota and downloaded tab-delimited files with lists of all eukaryotic orthologous groups (EOGs) for both. Note that OrthoDB version 9.1 used the genes annotated within the following seven lepidopteran genomes for inferring and classifying EOGs of Lepidoptera: Melitiaa cinxia (Linnaeus) (Ahola et al., 2014), Danausplexippus (Linnaeus) (Zhan et al., 2011), Heliconiusmelpomene (Linnaeus) (The Heliconius Genome Consortium, 2012), Papilio glaucus Linnaeus (Cong et al., 2015), Plutella xylostella (Linnaeus) (You et al., 2013), Bombyx mori (Linnaeus) (The International Silkworm Genome Consortium, 2008) and Manduca sexta (Linnaeus) (Kanost et al., 2016). Here we refer to this set of seven species as the OrthoDB reference species and their genomes as the set of OrthoDB reference genomes. Using the tab delimited files obtained from OrthoDB for all EOGs we identified 3107 genes/EOGs that are ‘single-copy’ as well as present in all seven Lepidoptera reference genomes. We observed that some of these single-copy genes in Lepidoptera, identified by their gene name, are multi-copy at the level of Endopterygota, and actually have multiple copies in some Lepidoptera species. This is possible if two or more gene variants are very similar but can be distinguished in reciprocal similarity comparisons on the ordinal level in Lepidoptera, but not at a higher level. Generally, clusters of orthologous genes become more and more inclusive on a higher taxonomic level. Since such cases may lead to target enrichment baits binding to multiple variants, we discarded all genes identified as multi-copy at the level of Endopterygota with respect to Lepidoptera species in the final step of the bait design. We refer to the initial set of 3107 genes as the SCPA (single copy, present in all) candidate set of EOGs. This is the first set of candidate genes in the bait design workflow.

Reference alignments for bait design

We downloaded the official gene sets (OGS) of the seven lepidopteran reference genomes listed above. This included FASTA files of all annotated genes as amino acid sequences as well as of the corresponding coding nucleotide sequences. If multiple versions of the OGS were available, we downloaded the OGS files in which the gene names correspond to those used in OrthoDB 9.1. For each gene sequence in the amino acid and nucleotide FASTA files we identified the gene name used in OrthoDB and renamed the sequence accordingly. If multiple variants of the same gene existed in the OGS, e.g. different splicing variants, we included only the longest of all variants in our modified OGS FASTA files. Mitochondrial genes were removed. Renaming was necessary in order to be able to use these OGS gene files in Orthograph.

We designed baits for the selected SCPA genes with the BaitFisher software v. 1.2.8 (Mayer et al., 2016). For each locus the baits should reflect the full variability of the gene of interest in the whole taxonomic group of interest in order to successfully
Fig. 1. Comparison of inferred ditrysian relationships on the superfamily level between major molecular phylogenetic studies and the current study. Green circles indicate branches with at least 95% support in that study (bootstrap or posterior probabilities, depending on what was used). Branch lengths are arbitrary. [Colour figure can be viewed at wileyonlinelibrary.com].
enrich this gene. Although the seven OrthoDB reference species capture some variety across Lepidoptera, several are butterflies and do not fully reflect the expected variation. In order to include more species as templates for bait design, we followed Mayer et al. (2016) and Sann et al. (2018) and used Orthograph, version 0.6.2, (Petersen et al., 2017) to mine the SCPA genes from 19 samples with genome or transcriptome assemblies published on NCBI, or by Breinholt et al. (2018) or Kawahara & Breinholt (2014). These taxa were selected to cover the entire Lepidoptera from Micropterigidae to Noctuidae (see highlighted taxa in Table S2). We added the SCPA genes and the modified OGS files of the seven reference species as amino acids and nucleotides into an Orthograph reference database. Using this database, Orthograph was used with default parameters to mine orthologous sequences within the specified transcriptomes and genomes. As output we obtained the corresponding amino acid and nucleotide sequence files for each SCPA gene for the OrthoDB reference species and the sequences mined from the transcriptomes. Sequence names in the Orthograph output were modified by replacing spaces with underscores, since programs in the downstream analysis workflow clipped sequence names at the first space.

For each gene, amino acid sequences obtained from Orthograph were aligned with MAFFT 7.305b (Katoh & Standley, 2013), using the E-INS-i algorithm. Nucleotide alignments based on these amino acid alignments were generated using a modified version of PAL2NAL 14.1 (Suyama et al., 2006; Misof et al., 2014). Outliers in the alignments were identified with the inhouse software OliInSeq version 0.9.1 (Martin et al. unpublished) with default parameters except for the parameters -f 2.0 -I 0.563 -w 20. OliInSeq masks or removes outliers simultaneously in the amino acid and the nucleotide datasets. The program is included as part of our TEnriAn workflow.

Including genes for backward compatibility with the BUTTERFLY1.0 kit

In order for the data of the present study to be backward compatible with the data obtained in Espeland et al. (2018, 2019), Toussaint et al. (2018) and other studies using the BUTTERFLY1.0 kit by Espeland et al. (2018) or variations thereof, we decided to also include all genes used in this kit in our study. As in the BUTTERFLY1.0 kit, most loci traditionally used for Sanger sequencing are included in our kit. The BUTTERFLY1.0 kit was produced by modifying and adding to the Lepidoptera-wide Lepidoptera Lep1.0 kit by Breinholt et al. (2018). Most of the Lep1.0 loci are also included in the BUTTERFLY1.0 kit, except those that were deemed too short (<150 bp), too little variable, or were found to be multi-copy, so including the BUTTERFLY1.0 loci also assures partial compatibility with the Lep1.0 kit. For identifying genes included in the earlier kit we used the BLAST+ software suite version 2.2.29 (Camacho et al., 2009) to conduct BLAST searches with default parameters except -outfmt 6 -task blastn -evalue 0.001’ of the gene sequences used in Espeland et al. (2018) against the official gene sets of the reference taxa. In this way, we obtained the names of the genes used in Espeland et al. (2018) and in turn their EOG identifiers on the Lepidoptera level. Finally we determined the genes used in Espeland et al. (2018) that are not covered by our SCPA dataset. Using the same procedure described above for the SCPA genes, including sequence extraction, Orthograph, alignment and outlier detection, we obtained 167 additional gene template alignments for the bait design.

Bait design

The BaitFisher software was used to design hybrid enrichment baits for three different gene sets: (i) the SCPA genes, (ii) the BUTTERFLY1.0 genes and (iii) the barcoding region of the COI gene. For the datasets (i) and (ii), we excised individual CDS regions from the nucleotide multiple sequence alignments of the EOGs using the annotated CDS sequences of the Danaus plexippus genome as a reference (Mayer et al., 2016) in order to avoid spanning baits across exon-exon boundaries in the concatenated CDS regions in the gene alignments. We specified a bait length of 120 bp, a clustering threshold of 0.15, and a tiling design of three baits per bait region with an overlap of 60 bp for two consecutive baits. This resulted in bait regions with a length of 240 bp. The reference sequences of Danaus were not removed from the alignments in the bait design process. Baits were inferred using the heuristic implementation of the unweighted Hamming 1-center DNA sequence search algorithm. BaitFisher designed baits for all CDS template alignments, if they have the required minimum length of 240 bp. From the set of 3107 SCAP genes BaitFisher was able to design baits for 3600 CDS regions, belonging to 1991 genes. For the additional 167 genes of the BUTTERFLY1.0 set, BaitFisher was able to design baits in 226 CDS regions belonging to 144 genes. BaitFisher designs baits at every potential start position of a bait region. The resulting redundancy allowed us to filter potential bait regions and to choose an optimal bait region for every locus. We first used the BaitFilter version 1.0.6 program to search for and exclude baits that possibly enrich multiple loci. We ran BaitFilter with the following options: ‘-m blast -I blast-first-hit-evalue 0.00000001 -blast-second-hit-evalue 0.000001’. With these options, BaitFilter searched all potential baits against a reference genome, in this study the genome of D. plexippus, with the aid of the BLAST+ software suite. The BLAST result was used to remove all bait regions for which at least one bait exhibited significant sequence similarity with more than one position in the reference genome, i.e. the best and second best hit had e-values smaller than 0.00000001 and 0.00001, respectively (see the BaitFilter manual for more details). The BLAST filter reduced the number of target loci in the SCPA set to 3268 CDS regions in 1898 genes and the number of additional genes from the BUTTERFLY1.0 set to 209 CDS regions in 137 genes. Finally, we used BaitFilter in a separate run to choose the optimal bait region within each CDS region among all remaining bait regions with the option ‘-m fs’. Specifically, we chose the start position within each CDS region at which the highest sequence coverage in the template alignments was found.
BaitFisher and BaitFilter were also used to design hybrid enrichment baits for the barcode region of the COI gene for lepidopteran species. We created a single alignment file containing 341 sequences of the barcode COI region across Lepidoptera. This file was used as a template in a BaitFisher analysis. We specified a clustering threshold of 0.15 and a tiling design of only one bait per bait region. Since it is known that mitochondrial baits are far more efficient than baits for nuclear genes, we manually reduced the number of COI baits to target only three bait windows in the beginning, middle and end of the barcode region, namely at positions 179, 340, 536.

As mentioned above, we removed after the bait design step all bait regions belonging to genes that are multi-copy for Endopterygota for at least one Lepidoptera species according to OrthoDB 9.1. This filtering step resulted in a reduction of the number of target loci from 3268 to 2744 CDS regions, now in 1616 genes. Combined with the herein designed baits for the BUTTERFLY1.0 gene set, our final bait set targets 2953 CDS regions in 1753 different nuclear genes with the aid of 94996 baits plus the COI gene with the aid of 221 baits. The final LepZFMK1.0 kit targets 248 exons that are also targeted in the BUTTERFLY1.0 kit to achieve backward compatibility. The target enrichment bait kit is available on Zenodo (doi: 10.5281/zenodo.4273959).

Molecular laboratory methods

DNA was extracted from dry abdomens of our 37 specimens (Table S1) without damaging the genitalia, with the exception of the Eirmocides Braby, Espeland & Müller (Lycaenidae) and Parnassius Latreille (Papilionidae) specimens, where a single leg was used. The dry tissue was incubated a few minutes in nuclease-free water to avoid soaking up too much lysis buffer during extraction. Samples were lysed at 56°C overnight shaking with 350 rpm using an Eppendorf Thermomixer comfort for approximately 12–18 h. We used the DNeasy Blood & Tissue kit (Qiagen, Hilden, Germany) by following the steps from the standard extraction protocol, except that we included a RNase-digestion step and eluted the DNA in 100 μL of prepared Dynabeads MyOne Streptavidin T1, DNA Repair Mix (NEB, USA). This repair mix is designed for formalin-fixed, paraffin-embedded (FFPE) tissues and repairs deamination of cytosine to uracil, nicks and gaps, oxidized bases and blocked 3’ ends, which are common problems in samples with low DNA quality (e.g. Zimmermann et al., 2008; Staats et al., 2011). Preliminary tests showed that it performs equally well on samples derived from old, dry tissue. The reaction mix was prepared on ice, according to the manufacturer’s protocol. The reactions were incubated for 15 min at 20°C. Purification of the reactions was carried out with Agencourt AMPure XP beads, and the products were quantified again using the Quantus Fluorometer. End-repair and A-Tailing reactions were conducted using the Agilent Sure Select XT2 Library Prep Kit on the basis of their 100 ng Agilent SureSelect XT2 protocol, but with the reaction volume reduced by 50%, as in Bank et al. (2017). After the A-Tailing reaction we purified the samples with Agencourt AMPure XP Beads in a ratio 1:1 and again quantified them. For the following steps we used the New England Biolabs reaction kits to enable a dual indexed library. Adaptor ligation was performed with the NEBNext Quick Ligation Module and the adaptors from the NEBNext Multiplex Oligos for Illumina (Dual Index Set1) Kit. The reaction mix was prepared on ice, according to the manufacturer’s protocol with a 1:25 dilution of the adaptors. The reactions were incubated for 15 min at 20°C, and after adding the USER Enzyme contained in the NEBNext Multiplex Oligos for Illumina (Dual Index Set1), the samples were incubated for 15 min at 37°C. Purification of the reactions was carried out with Agencourt AMPure XP beads in a ratio 1:1. For the pre-capture library amplification we used NEBNext Multiplex Oligos for Illumina (Dual Index Set1) and NEBNext Q5 Hot-Start HiFi PCR Master Mix, to dual-index the libraries. Reaction mix was prepared on ice with 15 μL, Adaptor-tagged DNA, 5 μL NEBNext i7 Primer, 5 μL NEBNext i5 Primer and 25 μL NEB-Next Q5 HotStart HiFi PCR Master Mix. The amplification was carried out under the following conditions: 98°C for 30 s, followed by cycles of 98°C for 10 s and 65°C for 75 s, and finally 65°C for 5 min. The number of cycles was chosen based on the concentration measured after the A-Tailing reaction, in our case 12 or 16 cycles. After PCR reaction we purified the samples with Agencourt AMPure XP beads in a ratio of 1:0.9.

The resulting libraries were quantified with a Quantus Fluorometer (Promega) and quality checked with a Fragment Analyzer (Advanced Analytical, now Agilent Technologies Inc.). For the following enrichment and capture steps we continued with the Agilent SureSelect XT2 protocol, as in Bank et al. (2017), with additional modifications: Specimens were pooled in absolute amounts of 93 ng in a pool of eight samples and the volume was reduced to 3.5 μL with a SpeedVac R SPD 111 V (ThermoFisher Scientific, Waltham, MA; USA). Hybridisation was performed on each pool with the SureSelect XT2 Pre-Capture ILM Module Box 2 (Agilent Technologies) and the SureSelect Custom Baits 6–11.9 Mb kit (Agilent Technologies), following the manufacturer’s protocol for captured library size >3.0 Mb with a volume reduced by 50%. Library and baits were hybridized at 65°C in a GeneAmp Thermal Cycler 2720 for nearly 48 h. We captured the enriched libraries with 50 μL of prepared Dynabeads MyOne Streptavidin T1,
We then processed data using two different approaches, the default parameters to remove adapters and low-quality regions. Filtering, preprocessing and assembly of the sequencedata obtained from hybrid enrichment was done at StarSEQ GmbH (Mainz, Germany) on an Illumina Nextseq 500 System with a read length of 150 bp and an estimated output of 21.67 million reads per pool or 0.4 Gbp per sample.

Filtering, preprocessing and assembly of the sequence data obtained from hybrid enrichment

Reads were trimmed with fastq-mcf (Aronesty, 2011) using default parameters to remove adapters and low-quality regions. We then processed data using two different approaches, the Iterated Bait Assembly (IBA) pipeline (Breinholt et al., 2018) and our own TEEnriAn workflow described below, and compared the results.

1. IBA: Sequences of target loci were extracted from the reads using IBA with default parameters, except that the pair gap length was set to 100 (−g 100). As a reference or seed for the IBA assembly we used the genomic sequences of the target regions of D. plexippus. In IBA, reads similar to the reference/seed sequence are identified with USEARCH (Edgar, 2010) and assembled with Bridger (Chang et al., 2015). The result of the assembly is then used as a reference sequence for another run of USEARCH and the process is repeated three times.

2. The TEEnriAn (target enrichment analysis) workflow presented in this paper: Reads were assembled with the de-novo assembler Trinity version 2.9.0 (Grabherr et al., 2011). Since this assembler is made for assembling transcriptomes it works well for data with unequal coverage, as expected for target enrichment data (Grabherr et al., 2011). Trinity crashed on one sample (Sample ID 49_S26). Analyses of the raw data for this sample revealed a particularly high amount of reads with long poly-G repeats in the read sequences. The cause for this problem is DNA of poor quality which is expected for some of the old museum samples. Unfortunately, the Illumina platform can assign these reads high quality scores. Since existing tools remove poly-G repeats only if they occur at the ends of the read, with no or a limited number of mismatches we had to solve this problem with an in house script (available upon request) that removes all read pairs for which one of the two reads had a poly-G repeat with a length longer than 20 bp, irrespective of the position of the repeat in the read. Furthermore, reads with a length shorter than 50 bp were removed.

Cross-contamination, both in the laboratory and during sequencing, has been shown to be a common problem in high-throughput sequencing studies, and especially so in studies targeting old samples with very low amounts of DNA (Kircher et al., 2012; Laurin-Lemay et al., 2012; Misof et al., 2014; Ballenghien et al., 2017). In order to be able to conduct the cross-contamination check introduced in Misof et al. (2014), we estimated the read coverage of all contigs of the Trinity assemblies of the hybrid enrichment data with bowtie2 and the RSEM algorithm (Li & Dewey, 2011). Abundance estimation was performed with the script align_and_estimate_abundance.pl, which is part of the Trinity utils. These estimated contig coverage values were used to detect and remove potential cross-contamination from the sequencing step. This was done using the script ContamCheck5, which performs all-versus-all BLAST (−evalue = 1e-50, −perc_identity = 99) comparisons of assemblies of all samples that were sequenced together. It removes sequences with 99% or higher identity found in both samples (see Misof et al. (2014) for more details). We, however, made sure that highly identical sequences in comparisons between samples of the same species were not filtered. The number of filtered contigs per sample can be found in Table S1.

Orthology prediction

The identification of sequences orthologous to the CDS regions of interest was achieved with the tool Orthograph. Orthograph applies a best reciprocal hit search strategy using profile hidden Markov models and maps nucleotide sequences to the globally best-matching cluster of orthologous genes, which allows a delineation of orthologs and paralogs from transcriptomic and genomic sequence data. This was applied to the assembled target enrichment data as well as a set of published transcriptomic and genomic lepidopteran sequences.

Building an extended reference dataset

To increase the mining sensitivity of Orthograph across all Lepidoptera, we extended the set of seven reference genomes available in OrthoDB at that time with 14 Lepidoptera species, for which official gene sets have been published on NCBI (Amyleolis transitella Walker, Spodoptera litura (Fabricius), Pieris rapae (Linnaeus), Papilio machaon Linnaeus, Bicyclus anynana (Butler), Oopheroptera bruma (Linnaeus), Chilo suppressalis (Walker), Ostrinia furnacalis (Gueneé), Trichoplusia ni (Hübner), Papilio xuthus Linnaeus, Eumeta japonica Heylaerts, Helicoverpa armigera (Hübner), Heliolithis virescens (Fabricius) and Galleria mellonella (Linnaeus), see Table S2 for further details). This was done by using Orthograph with the seven OrthoDB reference species to mine the CDS regions of interest in the 14 additional reference species. Since we only wanted to mine the CDS regions of interest and not the whole genes, we selectively replaced the whole genes with the CDS regions of interest in the gene set files given as input to Orthograph. As multiple CDS regions per gene are possible, we replaced gene names with CDS names which consisted...
of the gene name and the CDS feature number. These names were used in the gene set files and the tab-delimited file with CDS features of interest. Furthermore we encoded the feature number in modified EOG names in order to tell Orthograph to distinguish CDS regions. For the additional reference sequences obtained with Orthograph, we created extended tab delimited files. The OGS files of the now 21 references were used to create an extended Orthograph database that we applied in all further Orthograph analyses.

Mining sequences of interest

With the aid of this extended reference set of 21 Lepidoptera species we used Orthograph as described above to mine 138 published genomes and transcriptomes and the 37 target enrichment assemblies (this study) for sequences orthologous to the CDS regions of interest, i.e. CDS regions of the genes of the SCPA and BUTTERFLY 1.0 gene sets. A full list of the species mined in this way is given in Table S2. Orthograph reported the CDS sequences of interest as amino acid and nucleotide sequences from these species. These sequences were combined to individual and initially unaligned CDS region files using in-house scripts. Data from the 21 reference taxa were included by mining them together with the other genomes and transcriptomes with Orthograph. We could have included them by simply using the sequences excised from the genomes to build the extended Orthograph reference database, but we decided to retrieve them by mining them in the same way as all other sequences in order to avoid systematic biases resulting from data being obtained in different ways. Our complete dataset contained 175 taxa. See Tables S1 and S2.

Alignment and filtering of orthologous CDS regions

In the mining step, Orthograph generated HMM profiles for all CDS regions of interest. These HMM profiles were extracted from the Orthograph output and used to create reference alignments of the retrieved CDS regions on the amino acid level with the hmmlalign software (HMMER package, version 3.3, http://www.hmmer.org) with the ‘–amino –informat FASTA –outformat Stockholm’ options. The resulting Stockholm-format files were converted to FASTA-format. Nucleotide alignments corresponding to the amino acid alignments were built with a modified version of pal2nal (based on version 14.1, see Misof et al., 2014) and trimmed to the HMM region with an inhouse perl script (available as part of the workflow). By doing this we avoid including intron sequences, which would introduce systematic biases (see Bank et al., 2017). In these alignments we detected and masked outlier sequences with the OliInSeq software v0.9.5 software (available as part of the workflow) on the amino acid level using the following parameters: – remove-gap-ambig-lowerCase-sites, --mask-lower-case -w = 15; --IQR-factor = 2.0. This software searches for outlier sequences with a sliding window approach, here with windows of size 15. Sequence segments identified as outliers were soft masked with lower case symbols. Sequences for which more than 50% of the valid windows are identified as outliers were removed entirely from the corresponding CDS region alignment. OliInSeq automatically removes and masks the outlier sequence segments in the corresponding nucleotide alignments. This was followed by filtering each pair of amino acid and nucleotide alignments to remove all positions with less than 5 target enrichment taxa. Additionally, all individual sequences shorter than 10 amino acids were removed from all alignments. Next, we determined the pairwise sequence identity for all CDS region alignments on the amino acid level. If the mean pairwise BLOSUM62 scores were below 0.5 for more than 10 pairs of sequences, the alignment was excluded from further analyses. In order to identify random similarity within our multiple sequence alignments we used Aliscore (Misof & Misof, 2009) with default parameters except for using the -r 1 000 000 option, which ensures that all sequence pairs are compared. This program uses Monte Carlo resampling within a sliding window to identify alignment segments which cannot be distinguished from segments containing randomised data. The identified problematic positions were removed using an in-house script available as part of the workflow. Alignments shorter than 50 amino acids were excluded from further analyses to reduce the amount of missing data. After this step we again ensured that a minimum of five target enrichment taxa were present in each position in each alignment, and that the minimum length of sequences was at least 10 amino acids. A presence/absence matrix of all included taxa in the final alignments can be found in Table S3.

A graphical outline of the workflow used in this study is shown in Fig. 2. The post-sequencing steps including trimming, assembly, mining of orthologous loci, alignment and filtering (including in-house scripts) are combined in a Snakemake workflow (TenriAn, Target Enrichment Analysis) to ensure reproducibility and scalability. Snakemake is a popular workflow engine, which can be seamlessly scaled to server, cluster, grid and cloud environments, without the need to modify the workflow definition (Köster & Rahmann, 2012). The workflow is described in a human readable, Python-based language, and is provided in a Singularity container (Kurtzer et al., 2017), which can be installed anywhere independent of the operating system. The workflow is available on Github (see Data availability statement).

Extraction of COI reads

We used the COI barcoding gene to verify the identification of the different species, and extracted COI sequences of the barcode region from the target enrichment data. As a reference, we used an alignment of COI amino acid sequences of the barcode region from 341 species of Lepidoptera downloaded from NCBI (Table S4, alignment available on Zenodo, see data availability statement). Reads were aligned (nucleotide to amino acid) to a consensus sequence of this alignment with Exonerate 2.4.0 (Slater & Birney, 2005), and consensus sequences of the aligned reads for all specimens were created with Geneious 9.1.6 (Biomatters Ltd, Auckland, New Zealand). For these sequences, BLASTn searches were conducted against the NCBI
Capturing genes from pinned insect museum material

nucleotide and the BOLD (Ratnasingham & Hebert, 2007) databases. Resulting COI sequences were aligned together with the sequences used for bait design using MAFFT with the L-INS-i algorithm, and a phylogenetic analysis was conducted with IQ-TREE 1.6.3 (Nguyen et al., 2015) using the GTR + I + G model.

Determining enrichment efficiency

In order to determine the efficiency with which our kit enriched the DNA for our samples, we mapped the sequenced and quality trimmed reads back onto the CDS regions. CDS sequences have undergone multiple filtering steps. Thus, in order to be able to achieve an unbiased mapping, we mapped the reads onto the CDS regions for each specimen directly after masking and removing outlier sequences with OliInSeq, but before any further filtering steps. We mapped the read onto the CDS regions with bwa-mem2 version 2.0 (Vasimuddin et al., 2019) with default parameters. Resulting SAM files were sorted, converted to the bam format and the read coverage was estimated with samtools version 1.10 (Li et al., 2009). Mean species coverage values were obtained by computing a weighted mean of all CDS region coverage values, using the lengths of the CDS regions as weights.

Phylogenetic analyses

For the phylogenetic analyses, we concatenated the amino acid sequences and the nucleotide sequences of all 1835 excised target regions that passed the above filtering steps using AMAS.
v. 1.0 (Borowiec, 2016). The following datasets were generated and used for phylogenetic analyses: (i) NT2 (172 taxa, 234,464 bp) – nucleotide alignment including only second codon positions, (ii) NT12 (172 taxa, 468,928 bp) – nucleotide alignment including only first and second codon positions, (iii) AA (172 taxa, 234,464 AA) – amino acid alignment, and (iv) AA_red (172 taxa, 106,585 AA) – amino acid alignment including only those alignments with 120 or more sequences. Alignment statistics for the individual alignments for each dataset can be found in Tables S5–S7. Partition merging and substitution models selection was done using ModelFinder (Chernomor et al., 2016; Kalyaanamoorthy et al., 2017) in IQ-TREE 2.1.2 (Minh et al., 2020) with the -rclusterf algorithm (Lanfear et al., 2014) and the option -rclusterf 10. For the amino acid datasets, model selection and partition merging was conducted by testing all nuclear amino acid substitution models available in IQ-TREE together with the standard FreeRate models plus the default set of model extensions by specifying the ‘-m MF+MERGE, -msub nuclear’ command line parameters. Nucleotide partition merging and model finding was conducted with the ‘-m MF+MERGE’ option. Alternative models were compared using the default Bayesian information criterion. After determining the optimal partition scheme and model for each partition, phylogenetic trees were computed in separate IQ-TREE analyses using the determined models and partitioning schemes as input. For each dataset, 50 likelihood searches were performed in IQ-TREE 2.1.2, and 1000 ultrafast bootstrap replicates (Hoang et al., 2018) were computed after specifying the -bnni option to reduce the risk of overestimating support values. Furthermore, a SH-like approximate likelihood ratio test (SH-LRT) was conducted with 1000 replicates (Guindon et al., 2010) to obtain another set of support values. The -safe option was specified to avoid numerical underflow in large datasets. From the 50 likelihood searches, the tree with the highest likelihood was selected as the best tree for each dataset. Additionally, we inferred gene trees from the amino acid alignment of each locus using IQ-TREE as described above, with 1000 ultrafast bootstrap replicates. In the gene trees, branches with support values lower than 10 were collapsed using the Newick Utilities v.1.6 (Junier & Zdobnov, 2010), which has been shown to improve accuracy in the following species tree analysis (Zhang et al., 2017). Additionally, we used treeshrink v.1.3.6 (Mai & Mirarab, 2017) to remove abnormally long branches in the gene trees. Outgroup taxa were whitelisted using the -x option to avoid these being removed, and the -b option was set to 20. These filtered gene trees were used as input for ASTRAL-III v.5.7.4 (Zhang et al., 2017) to infer a species tree. Local posterior probabilities (Sayyari & Mirarab, 2016) were calculated as a measure of branch support.

**Graphical comparison of phylogenetic trees**

Pairwise comparisons of tree topologies were visualized using the cophyl function from the R-package *phytools* (Revell, 2012), with additional modifications in ggtree (Yu et al., 2017). Here the nodes of both trees were rotated to optimize vertical matching of the tips. Branches that are unique to each tree were identified using the phylo.diff command from the R-package *distory* 1.4.1 (Chakerian & Holmes, 2013). Additional manual modifications were performed in Inkscape v. 0.91 (Inkscape Project, 2020).

**Evaluating factors affecting sequencing success**

We investigated the relationships between sample age and several factors potentially affecting sequencing success by producing scatter plots with linear regression trend lines and calculating the Spearman rank correlation. Plots were produced using the R-package ggplot2 (Wickham, 2011). Additionally, we evaluated whether initial DNA concentration, body length (as proxy for specimen size, and amount of tissue available), the proportion of detected cross-contamination, and K2P distance to the nearest relative used for bait design, affected the number of aligned nucleotides recovered, using the same methods as above. The pairwise K2P distances (Kimura, 1980) were computed between the CDS sequences derived from our samples and those derived from the closest relative (according to the NT12 tree) among the 26 taxa used for bait design. In cases, where a sequenced sample had multiple closest relatives in the bait kit, we used the mean distance. K2P distances were calculated using the dist.DNA function in the R package ape v.5.4 (Paradis & Schliep, 2019).

**Results and discussion**

**COI results**

COI sequences were enriched, sequenced and extracted to confirm the identification of samples by their DNA barcode. A sufficient number of reads for creating a reliable consensus COI sequence was found for all specimens, ranging from 949 to 283,056 reads, see Table S1. For some specimens the amount of COI reads was exceedingly high even though the number of baits used to enrich COI was already chosen to be low. The number of COI baits should be reduced in future versions of the kit to avoid obtaining more reads than necessary for COI. We confirmed the correct determination of all samples belonging to species for which a COI sequence was already available in the GenBank or BOLD database, except for one specimen which was shown to be misidentified. Sample 46_S20 was initially identified as *Heliozela sericella* (Haworth) (*Heliozelidae*) but our COI sequence was 100% identical to *Perittia herrichiiella* (Herrich-Schäffer) (*Elachistidae*) when comparing it with barcode sequences in the BOLD database. This assignment was confirmed by reassessing the voucher specimen. In other cases, the identification was confirmed at least to genus or family level both by BLAST and phylogenetic analyses. As expected, the phylogenetic tree based merely on COI (Fig. S1) did not recover most generally recognized higher taxa, but taxa known to be closely related (e.g. within the same genus) almost always grouped together.
Capturing genes from pinned insect museum material

(A) Number of loci recovered using the IBA pipeline and our TEnriAn workflow; (B–I) factors potentially affecting sequencing success plotted against sample collection year: (B) number of reads after initial trimming; (C) proportion of detected cross-contaminated contigs; (D) mean length of DNA fragments (a trend line was not added here since the relationships are not linear); (E) mean coverage depth; (F) proportion of mapped reads; (G) median CDS length; (H–J), factors potentially affecting the amount of recovered aligned data: (H) K2P distance between sequenced samples and the samples used for bait design; (I) initial DNA concentration; (J) body length as a proxy of specimen size and the amount of tissue used (abdomen). The five samples for which legs only were used, are not included in this comparison. The black lines are linear regression trend lines and the grey shading is the 95% confidence interval.

IBA pipeline versus TEnriAn workflow

A list of all specimens with the number of recovered loci for both approaches, sample age and other statistics can be found in Table S1. Initial pre-library DNA concentration ranged from 0.01 to 83 ng/μL and median DNA fragment length, determined with the Fragment Analyzer, ranged from 69 to 313 bps. The TEnriAn workflow almost always recovered a higher number of loci than IBA, and especially so for older samples (Fig. 3A). Only 15 loci were recovered from *Semioptila flavidiscata* Hampson (Himantopteridae, collected in 1946) using IBA, whereas 499 loci were recovered using our TEnriAn
work. Our results also highlight many other larger discrepancies in performance, such as the *Brachodes funebris* specimen (Brachodidae, collected in 1970), where IBA recovered 78 loci and TEnriAn workflow 768 loci, a tenfold increase. Our workflow recovered more loci than IBA in most cases, but there are some exceptions, such as *Epicea polydora* (Westwood), for which IBA recovered 1451 loci, whereas our TEnriAn workflow only recovered 808 loci. The lower number of loci recovered in some cases by our workflow, might be due to the thorough cross-contamination check we perform as part of the workflow. For the *E. polydora* sample, for example, 19% of the contigs were discarded as cross-contamination, and in our oldest successful sample 32% of the contigs were discarded (Table S1). We have not checked whether the identified cross-contamination contigs were ‘on target’. Reads would, for example, also be removed in this step if two old specimens have been contaminated with the same fungus. Due to the altogether higher number of recovered loci, we used the results from our own workflow in further analyses. A presence-absence matrix for all samples and included loci can be found in Table S3. Alignment statistics for all individual alignments for all datasets can be found in Tables S5 to S7.

**Performance**

Between 1 and 1793 loci were recovered per sample (Table S1). For three samples very few loci were recovered: *Heterogynis penella* (Hübner) (1 locus), *Heliodines roesella* (Linnaeus) (41 loci) and *Hylabaeza puera* (Cramer) (43 loci). A high number of reads were sequenced for the latter two samples (Table S1), but most of these turned out to be of fungal origin, indicating that these samples might have been infested by mould at some time. For the former sample almost no reads were recovered, for unknown reasons, although initial DNA concentration was reasonable. For all other samples we obtained more than 450 targeted loci. These included a *Parabraxas davidi* specimen (Epiceopidae) collected in Tibet in 1892, for which 468 loci (115320 aligned bps) were recovered, *Chrysoptoma variegata* (Chrysoptolomidae, Zimbabwe, 1935), with 828 recovered loci (212892 aligned bps), *Heterogymna ochrogramma* (Carposinidae, China, 1935) with 866 recovered loci (228258 aligned bps), and *Zygaea afghana* Moore (Zygaelidae, Afghanistan, 1971) with 1256 loci (358209 aligned bps). After filtering alignment positions, and at the end excluding locus alignments which contain less than 50 taxa in order to lower the amount of missing data, the included samples contained from 446 to 1676 loci, and from 115000 to over 600000 bps (Table S1) out of a total alignment length of 703392 bps in our dataset. Two samples (*P. davidi* and *S. flavidiscata*) contained more than 80% missing data in our final dataset, and several samples contained more than 50%. The exclusion or inclusion of taxa or genes with missing data from phylogenetic analyses has frequently been discussed (e.g. Lemmon et al., 2009; Cho et al., 2011; Wiens & Morrill, 2011; Wiens & Tiu, 2012; Roure et al., 2013; Hosner et al., 2016; Streicher et al., 2016). Too large an amount of missing data might cause weakening of support and unstable phylogenetic relationships. It has, however, also been shown that missing data should not be problematic if distributed randomly and therefore should not be the primary guiding principle for the inclusion or exclusion of taxa and genes. Furthermore, larger, sparser matrices often have higher support values, than smaller, more complete matrices (Wiens & Tiu, 2012; Roure et al., 2013; Wagner et al., 2013; Huang & Knowles, 2014; Jiang et al., 2014; Hosner et al., 2016; Streicher et al., 2016). Data generated from museum specimens will always be incomplete. Our museum samples, including those with large amounts of missing data, were all placed consistently in the phylogeny with high support. Museum specimens are an invaluable resource, including many species not available anywhere else. Only a small part of the specimens is available for sequencing, and once tissue has been taken for molecular work, there might not be enough tissue left for a second try without destroying the specimen. Whether or not to exclude such samples from analysis should be carefully evaluated on a case by case basis.

The hybrid enrichment kit presented here has good recovery of loci from a wide taxonomic range of families, from Psychidae (Tineoidea) to Noctuidae (Noctuoidea). Furthermore, we generated data for several families not included in any recent phylogenomic studies due to a lack of suitable material (e.g. Kawahara et al., 2019). This includes the family Pseudobistonidae, which was mentioned as a key taxon by Kawahara et al. (2019), members of superfamilies of uncertain placement such as Carposinoidea and Epermenioidea, and other smaller families including Brachodidae, Metabeliidae, Chrysopolomidae, Anomoetidae and Himantopteridae. For several families we included closely related species, with different collection years in order to confirm that a potential contamination of old material (e.g. fungal or bacterial sequences) is not obscuring the placement. If contamination was a major issue, it could be expected that the samples do not group together. Examples are two closely related chrysopolomids (collected in 1935 and 1976), or two specimens of *Chalcocelis albistriatus* (Snellen) (Limacodidae, collected 1937 and 1972). In all cases, the taxa group together in the phylogenetic tree (Figs 4, S1–S5). All specimens remained in the same place in the AA_red dataset which was characterized by a considerably lower amount of missing data for most taxa. Call et al. (2021) furthermore show that our target enrichment approach also works at a lower level, such as resolving phylogenetic relationships within a single family (Epiceopidae), by only using museum specimens.

Our full enrichment kit included baits to enrich 2953 CDS regions (loci). Thus, in the most successful taxa we were able to recover only 61% of the loci which is considerably lower than for fresh, alcohol preserved samples (Mayer et al., 2016). Interestingly, the enrichment success did strongly depend on the specific loci (Table S3). There was a substantial set of loci, which were found in none, or only a small number of taxa, and there were a considerable number of loci which were found in a large number of taxa. Possible reasons include: (i) baits of certain loci might be less effective, (ii) baits for one locus might impede baits for other loci, (iii) some loci might belong to genomic regions that are less conserved. Additionally, in many cases more reads
than necessary were captured for COI (Table S1). By removing loci with low capture rate, and lowering the amount of baits for COI, the target enrichment kit could be considerably improved for future studies, and a new version of the kit alleviating several of these issues is in production.

Factors affecting the amount of recovered data

As in other studies (e.g. McCormack et al., 2016; Brewer et al., 2019; McGaughran, 2020), we unsurprisingly found a strong negative correlation between sample age and the amount of data recovered (Fig. 3A, B, Table 1). The amount of detected cross-contamination during sequencing increased slightly with sample age (Fig. 3C, Table 1), which is also expected since samples containing little DNA are generally more prone to be affected by contamination. The correlation of mean DNA fragment length (as measured on the Fragment Analyzer) with sample age was not linear (Fig. 3D, Table 1). The DNA fragment length does not seem to degrade very much after approximately the first 20 years. A group of more recent samples (collected between 2003 and 2017) showed much longer fragment lengths than the rest of the samples, which is not unexpected, but interestingly, some samples of approximately
the same age appeared also in the group with shorter fragment lengths. This difference likely reflects discrepancies in sample preparation and preservation. The two recent nymphalid and the three recent lycaenid specimens that had longer fragment lengths, were not pinned but stored in envelopes. The two papilionid specimens, which also fell into the longer fragment group were pinned, but likely have not gone through any other specific treatments. Many treatments of museum specimens are expected to negatively affect sequencing success (e.g. Dean & Ballard, 2001; Espeland et al., 2010; Burrell et al., 2015; Vaudo et al., 2018; Forrest et al., 2019; Gibb & Oseto, 2020). Individual specimen treatment steps are, however, most often unknown for old museum material. Some frequently used treatments, such as relaxing specimens before pinning are likely influencing the quality and quantity of DNA extracted from the specimens. All specimens used here, with exception of those mentioned above, were pinned, and several of them have therefore likely gone through at least one of these deteriorating treatments. The older pinned specimens have also, for many years, been exposed to dichlorvos, formerly a frequently used pesticide in collections (remains are often still present), which has been shown to have a negative effect on DNA quality and quantity (Espeland et al., 2010). Since we use de novo assembly and do not map to a reference genome, we subsequently mapped the reads on the recovered loci to infer information. Mean coverage depth when mapping reads to the CDS region, the proportion of mapped reads, and the median CDS length are all negatively correlated with sample age (Fig. 3E–G, Table 1). The proportion of mapped reads and mean coverage depth are comparable to other studies using target capture approaches on museum specimens (Blaimer et al., 2016; Brewer et al., 2019; McLaughran, 2020). Although the proportion of mapped reads is not very high (0.01–0.25 for successful samples), it is sufficient to generate a useful average coverage (18–257×) for 34 of the 37 samples (Table S1). A low enrichment efficiency is expected for old material. Short fragments should more likely bind randomly to hybrid enrichment probes than long fragments, since forces to overcome a random pairing are stronger for longer than for shorter fragments.

There was no significant correlation between the number of loci recovered and the pairwise K2P genetic distance between the sequenced samples and the nearest relative used for bait design (rho = −0.17, p = 0.34, Fig. 3H, Table 2), indicating that the limited set of taxa chosen for bait design does not negatively impact sequencing success. Our bait kit works well across Lepidoptera, and does not only capture taxa close to the samples used to generate the kit. The correlation between initial DNA concentration and recovered data is significant (Fig. 3I), but not very strong (rho = 0.39, p = 0.02), and there was no correlation (Fig. 3J, Table 2) between body length, a proxy for the amount of tissue used, and recovered data (rho = 0.14, p = 0.46). In many cases we recovered more data from very small specimens than from large specimens of similar age (Table S1). Sample preservation is probably once again an important factor here. Small specimens are more typically pinned directly, otherwise mounting is often next to impossible. Larger specimens are more often relaxed prior to pinning by placing them in relaxation chambers with water and acetic acid, for hours or days. Often mould repellents or other chemicals are added. Larger specimens are also often killed by injecting with ammonium, with boiling water and/or treated with a variety of solvents for degreasing (e.g. Oman, 1952, Peterson, 1964, Gibb & Oseto, 2020). The effect of these treatments of insect specimens on DNA quality is not well understood, but likely detrimental (Kanda et al., 2015; Gibb & Oseto, 2020). We recovered a large amount of DNA sequence data from most specimens, small and large, regardless of preservation conditions.

### Alignment statistics

Alignment statistics for the individual alignments for each dataset can be found in Tables S5–S7. The complete dataset used for phylogenetic analyses contained 1835 loci and 234 464 amino acids (AA), 468 928 nucleotides (NT12), 234 464 nucleotides (NT2) and 106 585 amino acids (AA_red). Hetrogynis penella, Hyblaea puera and Heliodines roesella were excluded from all further analyses to reduce the amount of missing data, and the total phylogenetic dataset included 172 taxa. Initial analyses showed that these three taxa could not be reliably placed in the phylogeny. The total amount of missing data in our dataset (AA: 41%, AA_red: 26%) is comparable to Kawahara et al. (2019) (AA: 36%) based on high-quality transcriptomic data. Two samples, Semioptila flavidiscata and Parabraxus davidi (Epicopeidae), contained more than 80% missing data in the AA dataset and only slightly less in the AA_red dataset. The mean identity of nucleotide sequence pairs in the final supermatrices was

---

**Table 1.** Spearman rank correlation of sample age in years and metrics potentially affecting sequencing success.

| Metric 1 | S   | p       | rho |
|----------|-----|---------|-----|
| Number of bases recovered | 12 046 | 5.07E−04** | −0.55 |
| Number of loci | 11 645 | 1.96E−03** | −0.5 |
| Number of reads after trimming | 12 157 | 3.34E−04** | −0.57 |
| Mean DNA fragment length | 12 238 | 1.46E−06** | −0.71 |
| Proportion cross-contamination | 4699.9 | 0.02* | 0.4 |
| Coverage depth | 11 856 | 9.82E−04** | −0.53 |
| Proportion mapped reads | 11 650 | 1.93E−03** | −0.5 |
| Median CDS length | 13 301 | 1.13E−06** | −0.71 |

Highly significant correlations are marked with **P values in the range 0.01−0.05 are marked with an *.

---

**Table 2.** Spearman rank correlation of number of aligned bases recovered and metrics potentially affecting sequencing success.

| Metric 1 | S   | p       | rho |
|----------|-----|---------|-----|
| Body length | 4722 | 0.46 | 0.14 |
| K2P distance between sequenced sample and closest relative used for bait design | 7653.5 | 0.34 | −0.17 |
| Initial DNA concentration | 5128.6 | 0.02* | 0.39 |

P values in the range 0.01−0.05 are marked with an *.
91.1% for matrix NT2, 87.0% for matrix NT12 and 72.9% for matrix NT123 (not analysed). Blaimer et al. (2016) or McCormack et al. (2016) generated large matrices using UCEs on museum material with a mean sequence identity of 97 and 99%, respectively. It is, however, difficult to directly compare these numbers, since in both of these studies more closely related taxa were included.

Phylogenetic relationships

The inferred phylogenetic relationships for the AA and NT12 datasets were similar (Fig. 4) and differences, such as the placement of Gelechioidae, were generally not well supported. The tree inferred based on the NT2 dataset (Fig. S2, S3) was similar to the tree of the AA dataset, which is expected since the second codon position is most important in determining the amino acids. Tests of compositional heterogeneity often show less conflict in phylogenetic analyses of second codon positions only (Vasilikopoulos et al., 2019). There were, however some differences between the AA and NT2 tree that are discussed below. The tree recovered from the smaller AA_red dataset with a reduced amount of missing data was similar to the tree from the AA dataset (Fig. S4), but some rearrangements within certain clades were apparent. Bootstrap and SH-alrt support were similar between trees, but slightly lower for some clades in the AA_red dataset, which might be due to the smaller size of the dataset. All our included museum specimens appeared in or near the families or superfamilies where they were expected based on other studies (see Fig. 1), indicating that there is useful signal in the data also for the samples for which relatively little data was obtained. Overall our results are consistent with those inferred by Kawahara et al. (2019), but we generally obtain better support among several parts of the backbone. This might be due to our inclusion of key taxa not included in Kawahara et al. (2019), as well as the more thorough filtering of outliers in the alignments, performed as part of our workflow. A similar cross-contamination check as performed here was also applied in Kawahara et al. (2019), so cross-contamination should not be a problem in either study.

Relationships among non-ditrysiain moths are supported by a range of synapomorphies (Kristensen, 2003; Regier et al., 2015b) and were stable and highly supported in all trees, except for the sister group relationship between the clades consisting of Neosseustidae+Eriocranidae and Lophocoronidae+Hepialidae, which was only well supported in the Astral tree (Fig. S5). Our inferred relationships are similar to those found by Kawahara et al. (2019), Bazinet et al. (2017) and Regier et al. (2015b), indicating that this part of the Lepidoptera phylogeny seems to have largely stabilized.

For Ditrysia, by contrast, a number of different superfamily relationships have been suggested in molecular studies over the years, with varying support (Fig. 1). Although we used part of the same data used by Kawahara et al. (2019), and similar analyses, there are still some differences between the results of each study. The base of Ditrysia is stable in both studies starting with a non-monophyletic Tineoidea, which has similarly been found in all studies where more than one member was included (Mutanen et al., 2010; Regier et al., 2013, 2015b; Heikkilä et al., 2015; Breinholt et al., 2018; Kawahara et al., 2019). There are also no synapomorphies supporting Tineoidea as currently circumscribed (Davis & Robinson, 1998). We included one member of Psychidae (Eochorica balcanica [Rebel]), collected in 1986), which consistently grouped with other members of Psychidae (based on published transcriptome/genome data).

In our NT12 and NT2 trees as well as in Kawahara et al. (2019) the non-monophyletic Tineoidea were followed by a clade consisting of Gracillarioidea and Yponomeutoidea, which are not reciprocally monophyletic due to the placement of Roeslerstammiidae (Gracillarioidea) as sister to Yponomeutoidea. In our AA tree this family was the sister group to Gracillarioidea+Yponomeutoidea. Together these superfamilies formed a monophyletic group as found in most other studies (Mutanen et al., 2010; Bazinet et al., 2013; Regier et al., 2013; Kawahara & Breinholt, 2014; Breinholt et al., 2018), but see Heikkilä et al. (2015). Interestingly, our inclusion of Lyoneta Hübner (Lyonetiidae) and Bedellia Stainton (Bedelliidae), both families not included by Kawahara et al. (2019), rendered Lyonetiidae paraphyletic in all ML trees.

Almost all possible combinations of phylogenetic relationships in the so-called non-Obtectomeran Apoditrysiain superfamilies have been suggested in previous studies, often also including traditional Obtectomerans such as Immonoeida, but there has rarely been support for these suggested relationships (Fig. 1). Kawahara et al. (2019) proposed a clade consisting of Choreutoidea+Uroidea followed by a clade including Immoidea+Tortricoidea. This relationship is corroborated and relatively well-supported here, but we also find Dougasioidae in this group as the sister to Uroidea. Kawahara et al. (2011) found Douglasiidae as sister group to Choreutidae, although lacking support. Other studies including Dougasioida recovered them either as sister to Pterophoroidea (Mutanen et al., 2010; Heikkilä et al., 2015), or Schreckensteinioidea (Regier et al., 2013). Based on morphology, Douglasiidae was previously placed in Gracillarioidea (Davis & Robinson, 1998), but was left as an unplaced Apoditrysin in van Nieukerken et al. (2011).

Within Sesioidea, Brachodidae have traditionally been seen as the sister group to Sesioidea+Castniidae, based on certain features of the head and thorax morphology (Minet, 1986, 1991; Edwards et al., 1998). Cossoidea have been treated as the sister group to Sesioidea based on synapomorphies in the larval ground plan (Minet, 1991) and it was suggested that these two were the sister to Zygaenoidea (Kristensen & Skalski, 1998). These three superfamilies frequently form a monophyletic group in molecular studies (Regier et al., 2013; Heikkilä et al., 2015; Kawahara et al., 2019). Regier et al. (2013) and Kawahara et al. (2019) found Castniidae well-supported within Cossoidea, which is corroborated here. Heikkilä et al. (2015) found the traditional Sesioidea to be monophyletic, with a sister group relationship between Castniidae and Brachodidae. Regier et al. (2013) found Brachodidae to be placed within Cossoidea as sister to Cossidae. The relationships in both studies are however largely unsupported. In all our ML trees we found...
Brachodidae as the sister group to Sesioidea. Interestingly this was supported by a very high alt score (90–100), but ultrafast bootstrap support was low (see Newick tree files available on Zenodo). Our AA tree agrees with Kawahara et al. (2019) in that Cossoidea was the sister to Sesioidea+Zygaenoidea, but in the NT12 tree Sesioidea was the sister to Cossoidea+Zygaenoidea, and in the AA_red tree Zygaenoidea was the sister group to Sesioidea+Cossoidea, so all possible relationships were represented in our trees. The relationships were, however, not well-supported in either tree. The Metabeliidae are a small family sometimes included in Cossoidea (Minet, 1991), although it apparently does not share any synapomorphies with this family (Lehmann, 2019). Regier et al. (2013) found it as sister to the cossid subfamily Hypotinae and these were in turn most closely related to Castniidae. Here we found a monophyletic Cossoidea, with some support in all trees, but Castniidae were rendered paraphyletic by Dudgeonidae, Castniidae and Metabeliidae. Metabeliidae was here found to be the sister group to Cossulinae, and not Hypotinae, and this group was relatively far away from Castniidae, but support was low within this superfamily. The butterfly-like Ratardidae are also likely to be found in this clade, possibly even within Cossoidea (Edwards et al., 1998); more cossoids need to be included to better understand these relationships.

Zygaenoidea are not supported by any synapomorphies and many families (e.g. the Limacodid group) have moved between Zygaenoidea and Cossoidea (Brock, 1971; Heppner, 1984, 1995). Zygaenoidea as currently circumscribed were reunited by Epstein et al. (1998). In addition to the taxa found in Kawahara et al. (2019) we included the small and infrequently collected families Chrysopolomidae, Himantopteridae and Anomoeotidae in our study based on material collected in 1935 and 1976, 1946 and 1984, respectively. Anomoeotidae was included in Zygaenoidea until raised to family rank by Fletcher & Nye (1982), but the systematic position is unclear and Epstein et al. (1998) suggested it might need to be moved to another superfamily once the biology of larvae and adults is better known. In other studies, this family is treated as a subfamily of Himantopteridae (van Nieukerken et al., 2011; Heikkilä et al., 2015). We always recovered Anomoeotidae + Himantopteridae as a well-supported sister group to the Megalopygidae in both trees. This corresponds to relationships found by Heikkilä et al. (2015) where Megalopygidae was found to be the sister group to a clade including Aidinae (not included here), Himantopteridae and Anomoeotidae (as Anomoeotinae) with relatively high support. Chrysopolomidae was included as a subfamily of the Limacodidae (Epstein et al., 1998). Zolotuhin et al. (2014) found differences in the genitalia musculature between Limacodidae and Chrysopolomidae and suggested this clade should be treated as a separate family close to Limacodidae. We found Chrysopolomidae highly supported in a clade with Limacodidae+Dalceridae in all trees, so either Chrysopolomidae should be treated as a family, or Dalceridae need to be treated as a subfamily of Limacodidae. Characters of the larvae might be helpful to assess this situation (Epstein et al., 1998).

The Gelechioidea were thought to be the sister group to the Yponomeutoidea based on morphology of the pupa and the palpi (Minet, 1991), but were then placed in the non-Obtectomeran Apoditry sia by Kaila (2004) based on an extensive morphological study. In Kawahara et al. (2019), on the other hand, the Gelechioidea are placed within Obtectomera in a clade also containing Alucitoidea, Pterophoroidea, Calliduloidea and Thyridoidea. A similar placement is also indicated in Bazinet et al. (2013) and Kawahara & Breinholt (2014) with lower taxon sampling, but support is low in all studies. Heikkilä et al. (2015) inferred a similar clade, but Pterophoroidea was not part of this clade, but placed with Urodoidea and Douglassioidea. We found a well-supported clade consisting of Gelechioidea, Alucitoidea, Pterophoroidea, Calliduloidea, Thyridoidea, Epermenioidea and Carposinoidea (here called the GAPTECC clade), and internal relationships within the group were identical in the AA and NT12 trees, but differed slightly in the AA_red tree. Twort et al. (2020), based on maximum likelihood analyses of a 332 gene dataset derived from shotgun sequencing also found a non-monophyletic GAPTECC clade, but except for a clade consisting of Calliduloidea, Thyridoidea and Whalleyaneoidea, relationships are not supported. Kawahara et al. (2019) found Alucitoidea (many-plumed moths) + Pterophoroidea (plume moths) to form a relatively well-supported monophyletic group as suggested by Minet (1991) based on their resting posture and the configuration of the meso-precoxal suture, but Carposinoidea and Epermenioidea were not included in that study. Interestingly, Pterophoroidea and Alucitoidea were never sister groups in any of our four trees, indicating that the peculiar plumed wing morphology and resting posture have either evolved twice or have been independently lost in the Carposinoidea and Epermenioidea, which seem to be key taxa to include in this clade. In our AA tree, the GAPTECC clade was the sister group to Pyraloidea+Macroheterocera, which was also found by Kawahara et al. (2019), while in the NT12 tree the clade was the sister group to Papilionoidea+Pyralidae+Macroheterocera. Furthermore, in our NT2 and Astral trees the GAPTECC clade was not monophyletic, with Gelechioidea being the sister to Pyraloidea+Macroheterocera, and the remainder of the clade being sister to Papilionoidea (as a grade in the Astral tree), but all of these relationships were weakly supported. Earlier studies have never found a monophyletic GAPTECC clade, and at least part of it was placed outside Obtectomera (e.g. Mutanen et al., 2010; Regier et al., 2013; Heikkilä et al., 2015).

The remaining Obtectomera consisting of Pyraloidea+Mimallonioidea+Macroheterocera are stable and well-supported in most studies. Pseudobistonidae were described in 2015 based on a single species and it was suggested that it is the sister group to Epicopeiidae within Geometroidea (Rajaee et al., 2015). The family was mentioned as a key taxon to be included in future phylogenomic studies to resolve the relationships within this group (Kawahara et al., 2019). We consistently found Pseudobistonidae as the sister to Sematuridae, and these in turn sister to Epicopeiidae, with all relationships being well-supported. Interestingly, Call et al. (2021) using museum specimens, a subset of the loci used here and increased taxon sampling, found Pseudobistonidae to be the sister group to Epicopeiidae, always with high support. This relationship was also indicated by Rajaee et al. (2015). The impact of gene and taxon sampling
on support needs to be further investigated by including more epicopeiids and sematurids. Additionally, *Heracula discivitta* Moore, recently moved from Lymantriidae to Pseudobistonidae (Wang et al., 2019), should be included in a large scale molecular dataset to further investigate these relationships. Zhang et al. (2020b) using a 90 marker PCR-based target capture found *Heracula* to be sister to Epicopeiidae, but did not include *Pseudobiston* Inoue, or Sematuridae, making comparison of relationships difficult.

A small amount of conflict between the amino acid and nucleotide trees for the same underlying dataset can be found in our study. This is interesting since both types of data are inherently connected. A similar discordance between amino acid and nucleotide trees was found in Gillung et al. (2018), and could not be explained by compositional heterogeneity or saturation of the third codon positions. A failure to distinguish serine residues encoded by different codons (Zwick et al., 2012), or substitution models not being equally realistic for both data types could possibly cause such discrepancies. In addition to a tree reconstruction of the N12 dataset we also reconstructed a tree for the N2 data (Fig. S3), which in some studies has been favoured, since tests of compositional heterogeneity often show less conflict in phylogenetic analyses of second codon positions only (Vasilikopoulos et al., 2019).

**Conclusion**

Our target capture kit provides a large amount of informative data for clarifying Lepidoptera relationships, and was successfully used to add data from old, pinned museum samples to existing data from genomes and transcriptomes. We recovered from 468 to over 1700 loci from all but three specimens with collection years ranging from 1892 to 2017, and the samples could be placed in the phylogeny with reasonable confidence. This indicates that at least 130 year old material can be sequenced successfully using our method. Such a museomics approach allows us to add more leaves to better resolve the Lepidoptera tree, by using the large amount of material that has been collected during the last few hundred years. The proposed pipeline allows to process and combine hybrid enrichment data as well as data from transcriptomes and genomes available from previous studies.

**Author contributions**

ME and CM conceived the study. CM, LD and ME produced the target enrichment kit. SK and EC did the molecular work. SM, LD and ME processed data and did the phylogenetic analyses. ME, CM and SM wrote the paper with input from the other authors. All authors read, commented on and approved the final manuscript.

**Supporting Information**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

---

**Fig. S1.** Tree inferred using the COI sequences obtained in this study and sequences from Genbank.

**Fig. S2.** Comparison of the ML trees inferred from the AA (left) and NT2 (right) datasets

**Fig. S3.** Comparison of the ML trees inferred from the NT12 (left) and NT2 (right) datasets.

**Fig. S4.** Comparison of the ML (left) and Astral (right) inferred from the AA dataset. LPP = local posterior probabilities as shown on the nodes on the Astral tree.

**Fig. S5.** Comparison of the trees inferred from the AA and AA_red datasets.

**Table S1.** Sample information and statistics for the hybrid enrichment samples included in this study.

**Table S2.** Information on published transcriptome and genome data included in this study. Samples highlighted in yellow were used for bait design. Some of these species were not included in the data mining for phylogenetic analyses, and the total number of samples in the table is therefore higher than 138.

**Table S3.** Presence/Absence matrix of taxa in the final alignments. Target Enrichment taxa are highlighted in yellow. The first column contains the name (EOG number) of each individual alignment. The first row contains the name of the samples used. The samples sequenced as part of this study are highlighted in yellow. The second row shows the total number of alignments per sample.

**Table S4.** Accession numbers for COI sequences taken from Genbank.

**Table S5.** Alignment statistics for the individual amino acid alignments. This includes all alignments, also those not included in the phylogenetic analyses since they contained less than 50 taxa (AA) or 120 taxa (AA_red).

**Table S6.** Alignment statistics for the individual nucleotide alignments including first and second codon positions (NT12). This includes all alignments, also those not included in the phylogenetic analyses since they contained less than 50 taxa.

**Table S7.** Alignment statistics for the individual nucleotide alignments including second codon positions only (NT2). This includes all alignments, also those not included in the phylogenetic analyses since they contained less than 50 taxa.

**Acknowledgements**

We are grateful to Claudia Etzbauer for help with ordering lab supplies. We want to thank Niklas Wahlberg, Thomas Simonsen, Shaun Winterton and two anonymous reviewers, who provided...
useful comments, which greatly improved the manuscript. The Colombian specimens were collected as part of a Colombia Bio expedition. Gonzalo Andrade and Mariela Osorno Muñoz kindly helped with organization of fieldwork and the arrangement of permits (export permit 01191, Autoridad Nacional de Licencias Ambientales). We are highly indebted to everyone who has deposited samples in the Lepidoptera collection at the ZFMK. The study was funded by the Zoological Research Museum Alexander Koenig, and EC received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 6422141. The authors declare no conflict of interest.

Open Access funding enabled and organized by Projekt DEAL.

Data availability statement

Raw sequence read data have been uploaded to the NCBI Short Read Archive (BioProject PRJNA638273). Assemblies, individual alignments for all datasets, concatenated datasets with substitution model files, as well as all tree files, and the target enrichment kit are freely available on Zenodo (doi: 10.5281/zenodo.4273959). Further data are available in the supplementary information. The TEAniAn workflow is available on Github: https://github.com/ZFMK/TEAniAn.

References

Ahola, V., Lehtonen, R., Somervuo, P. et al. (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications*, 5, 4737.

Albert, T.J., Molla, M.N., Muzny, D.M. et al. (2007) Direct selection of human genomic loci by microarray hybridization. *Nature Methods*, 4, 903–905.

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A.-L., Sperling, F.A. & Condamine, F.L. (2020) Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, 69, 38–60.

Andermann, T., Torres Jiménez, M.F., Matos-Maraví, P. et al. (2020) A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics*, 10, 1407. http://dx.doi.org/10.3389/fgene.2019.01407.

Aronesty, E. (2011) ea-utils: Command-line Tools for Processing Biological Sequencing Data URL. https://github.com/ExpressionAnalysis/ea-utils [accessed on 15 July 2019].

Bagley, J.C., Uribe-Convers, S., Carlsen, M.M. & Muchhala, N. (2020) Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical *Burseimastera* belflowers as a case study. *Molecular Phylogenetics and Evolution*, 152, 106769.

Bailey, S.E., Mao, X., Strubeig, M. et al. (2016) The use of museum samples for large-scale sequence capture: a study of congeneric horseshoe bats (family Rhinolophidae). *Biological Journal of the Linnean Society*, 117, 58–70.

Ballenghien, M., Faivre, N. & Galtier, N. (2017) Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology*, 15, 25.

Bank, S., Sann, M., Mayer, C. et al. (2017) Transcriptome and target DNA enrichment sequence data provide new insights into the phylogeny of vespid wasps (Hymenoptera: Aculeata: Vespoidea). *Molecular Phylogenetics and Evolution*, 116, 213–226.

Bazinet, A.L., Cummings, M.P., Mitter, K.T. & Mitter, C.W. (2013) Can RNA-seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysa)? An exploratory study. *PLoS One*, 8, e26155.

Bazinet, A.L., Mitter, K.T., Davis, D.R., Nieuwerkerk, E.J.V., Cummings, M.P. & Mitter, C. (2017) Phylogenotranscriptomics resolves ancient divergences in the Lepidoptera. *Systematic Entomology*, 42, 305–316.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, 13, 403.

Blaimer, B.B., Lloyd, M.W., Guilflory, W.X. & Brady, S.G. (2016) Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS One*, 11, e0161531.

Borowiec, M.L. (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, 4, e1660.

Breinholt, J.W. & Kawahara, A.Y. (2013) Phylogenotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. *Genome Biology and Evolution*, 5, 2082–2092.

Breinholt, J., Cason, C., Lemmon, A.R., Lemmon, E.M., Xiao, L. & Kawahara, A.Y. (2018) Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Systematic Biology*, 67, 78–93.

Brewer, G.E., Clarkson, J.J., Maurin, O. et al. (2019) Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science*, 10, 1102.

Brock, J.P. (1971) A contribution towards an understanding of the morphology and phylogeny of the Ditrysian Lepidoptera. *Journal of Natural History*, 5, 29–102.

Burrell, A.S., Disotell, T.R. & Bergey, C.M. (2015) The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, 79, 35–44.

Call, E., Mayer, C., Twort, V.G., Wahlberg, N. & Espeland, M. (2021) Museomics: phylogenogenomics of the moth family Epicopeiidae (Lepidoptera) using target enrichment. *Insect Systematics and Diversity*. Camacho, C., Coulouris, G., Avagyan, V. et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.

Chakerian, J. & Holmes, S. (2013) Distory: Distance between Phylogenetic Histories. *R* package. https://cran.r-project.org/web/packages/distory/index.html [accessed on 15 July 2020].

Chang, Z., Li, G., Liu, J. et al. (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, 16, 30.

Chernomor, O., von Haeseler, A. & Minh, B.Q. (2016) Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, 65, 997–1008.

Cho, S., Zwick, A., Regier, J.C. et al. (2011) Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? *Systematic Biology*, 60, 782–796.

Choquet, M., Smolina, I., Dhanasiri, A.K.S. et al. (2020) Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton *Calanus finmarchicus*. *Royal Society Open Science*, 6, 180608.

Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N.V. (2015) Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Reports*, 10, 910–919.
Cong, Q., Shen, J., Borek, D. et al. (2017a) When COI barcodes deceive: complete genomes reveal introgression in hairstreaks. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20161735.

Cong, Q., Shen, J., Li, W., Borek, D., Orwinowski, Z. & Grishin, N.V. (2017b) The first complete genomes of metalmarks and the classification of butterfly families. *Genomics*, 109, 485–493.

Cosart, T., Reja-Pereira, A., Chen, S., Ng, S.B., Shendure, J. & Luikart, G. (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, 12, 347.

Davis, D.R. & Robinson, G.S. (1998) The Tineoidea and Gracillarioidae. *Handbook of Zoology, Vol. IV Arthropoda: Insecta Part 35, Lepidoptera, Moths and Butterflies. Volume 1: Evolution, Systematics, and Biogeography* (ed. by N.P. Kristensen), pp. 91–117. Walter de Gruyter, Berlin.

Dean, M.D. & Ballard, J.W.O. (2001) Factors affecting mitochondrial DNA quality from museum preserved *Drosophila simuli* simulans. *Entomologia Experimentalis et Applicata*, 98, 279–283.

Derkarabetian, S., Benavides, L.R. & Giribet, G. (2019) Sequence capture phylogeny of historical ethanol-preserved museum specimens: unlocking the rest of the vault. *Molecular Ecology Resources*, 19, 1531–1544.

Dietz, L., Dömel, J.S., Leese, F., Mahon, A.R. & Mayer, C. (2019) Phylogenetics of the longitarsal Colossenedae: the evolutionary history of an Antarctic Sea spider radiation. *Molecular Phylogenetics and Evolution*, 136, 206–214.

Dömel, J.S., Macher, T.H., Dietz, L. et al. (2019) Combining morphological and genomic evidence to resolve species diversity and study speciation processes of the *Pulalenopsis patagonica* (Pycnogonida) species complex. *Frontiers in Zoology*, 16, 36.

Doorenweerd, C., Nieukerken, E.J.V. & Hoare, R.J.B. (2017) Phylogeny, classification and divergence times of pygmy leaf-mining moths (Lepidoptera: Nepticulidae): the earliest lepidopteran radiation on earth (Lepidoptera: Erebidae: Arctiinae). *Evolutionary Biology*, 45, 874–893.

Eberle, J., Ahrens, D., Mayer, C., Niehuis, O. & Misof, B. (2020) A plea for standardized nuclear markers in Metazoan DNA taxonomy. *Trends in Ecology & Evolution*, 35, 336–345.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.

Edwards, E.D., Gentili, P., Horak, M., Kristensen, N.P. & Nielsen, E.S. (1998) The Cossoid/Sesoid assemblage. *Handbook of Zoology, Vol. IV Arthropoda: Insecta Part 35, Lepidoptera, Moths and Butterflies. Volume 1: Evolution, Systematics, and Biogeography* (ed. by N.P. Kristensen), pp. 181–197. Walter de Gruyter, Berlin.

Epstein, M.E., Geersema, H., Naumann, C.M. & Tarnam, G.M. (1998) The Zygaenoidea. *Handbook of Zoology, Vol. IV Arthropoda: Insecta Part 35*, pp. 159–175. Walter de Gruyter, Berlin.

Espeland, M., Irestedt, M., Johanson, K.A., Åkerlund, M., Bergh, J.-E. & Källersjö, M. (2010) Dichlorvos exposure impedes extraction and amplification of DNA from insects in museum collections. *Frontiers in Zoology*, 7, 2.

Espeland, M., Breinholt, J., Willmott, K.R. et al. (2018) A comprehensive and dated phylogenetic analysis of butterflies. *Current Biology*, 28, 770–778.e5.

Espeland, M., Breinholt, J., Barbosa, E.P. et al. (2019) Four hundred shades of brown: higher level phylogeny of the problematic Euphychiini (Lepidoptera, nymphalidae, Satyrinae) based on hybrid enrichment data. *Molecular Phylogenetics and Evolution*, 131, 116–124.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T. & Glenn, T.C. (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61, 717–726.

Fletcher, D.S. & Nye, J.W.B. (1982) *The Generic Names of Moths of the World 4. Bombychoidea, Castnioidae, Cossioidea, Mimallonidae, Sesioidea, Sphingoidea, Zygaenoidea*. Trustees of the British Museum (Natural History), London.

Forrest, L.L., Hart, M.L., Hughes, M. et al. (2019) The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. *Frontiers in Ecology and Evolution*, 7, 439.

Gauthier, J., Pajkovci, M., Neuenschwander, S., Kaila, L., Schmid, S., Orlando, L. & Alvarez, N. (2020) Museomics identifies genetic erosion in two butterfly species across the 20th century in Finland. *Molecular Ecology Resources*, 20, 1191–1205.

Gibb, T.J. & Osato, C. (2020) Insect Collection and Identification: *Techniques for the Field and Laboratory*. London: Academic Press.

Gilling, J.P., Winterton, S.L., Bayless, K.M. et al. (2018) Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids. *Molecular Phylogenetics and Evolution*, 128, 233–245.

Gnirke, A., Melnikov, A., Maguire, J. et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27, 182–189.

Grabb, M.G., Haas, B.J., Yassour, M. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644–652.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59, 307–321.

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B. & Hebert, P.D.N. (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, 6, 959–964.

Hamilton, C.A., St Laurent, R.A., Dexter, K. et al. (2019) Phylogenomics resolves major relationships and reveals significant diversification rate shifts in the evolution of silk moths and relatives. *BMC Evolutionary Biology*, 19, 182.

Hart, M.L., Forrest, L.L., Nicholls, J.A. & Kidner, C.A. (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon*, 65, 1081–1092.

Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C. & Brumfield, R.T. (2016) Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, 65, 910–924.

Heikilä, M., Mutanen, M., Wahlberg, N., Silvonen, P. & Kaila, L. (2015) Elusive ditrysian phylogeny: an account of combining systematized morphology with molecular data (Lepidoptera). *BMC Evolutionary Biology*, 15, 260.

Heppner, J.B. (1984) Outline of classification of the Lepidoptera. *Atlas of Neotropical Lepidoptera. Checklist: Part I Micropterigoidea – Innoidea* (ed. by J.B. Heppner), pp. xix–xxiii. W. Junk, The Hague.

Heppner, J.B. (1995) Lacturidae, new family (Lepidoptera: Micropterigoidea–Innoidea). *Entomological News*, 106, 518–522.
Meyer, M. & Kircher, M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, 2010, pdb.prot5448.

Millar, C.D., Huynen, L., Subramanian, S., Mohandesan, E. & Lambert, D.M. (2008) New developments in ancient genomics. Trends in Ecology & Evolution, 23, 386–393.

Miller, J.A., Beentjes, K.K., van Helsdingen, P. & IJland, S. (2013) Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. ZooKeys, 365, 245–261.

Minet, J. (1986) Ébauche d’une classification modérée de l’ordre des Lepidoptères. Alexanor, 14, 291–313.

Minet, J. (1991) Tentative reconstruction of the ditrysian phylogeny (Lepidoptera: Glossata). Insect Systematics & Evolution, 22, 69–95.

Minh, B.Q., Schmidt, H.A., Chernomor, O., et al. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37, 1530–1534.

Misof, B. & Misof, K. (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Systematic Biology, 58, 21–34.

Misérová, H., Greve, C., Letsch, H. et al. (2015) Advances in Geometroidea phylogeny, with characterization of a new family based on Pseudobiston pirinatarai (Lepidoptera, Glossata). Zootaxa Scripta, 44, 418–436.

Regier, J.C., Mitter, C., Zwick, A. et al. (2013) A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). PLoS One, 8, e55856.

Reinhard, C., Mitter, C., Davis, D.R. et al. (2015a) A molecular phylogeny and revised classification for the oldest ditrysian moth lineages (Lepidoptera: Tineoidea), with implications for ancestral feeding habits of the mega-diverse Ditrysia. Systematic Entomology, 40, 409–432.

Regier, J.C., Mitter, C., Kristensen, N.P. et al. (2015b) A molecular phylogeny for the oldest (nonditrysian) lineages of extant Lepidoptera, with implications for classification, comparative morphology and life-history evolution. Systematic Entomology, 40, 671–704.

Regier, J.C., Mitter, C., Mitter, K. et al. (2017) Further progress on the phylogeny of Noctuoidea (Insecta: Lepidoptera) using an expanded gene sample. Systematic Entomology, 42, 82–93.

Revell, L.J. (2015) Phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3, 217–223.

Roure, B., Baurain, D. & Philippe, H. (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Molecular Biology and Evolution, 30, 197–214.

Sann, M., Niehuis, O., Peters, R.S. et al. (2018) Phylogenetic analysis of Apoidea sheds new light on the sister group of bees. BMC Evolutionary Biology, 18, 71.

Sayyari, E. & Mirarab, S. (2016) Fast coalescent-based computation of local branch support from quartet frequencies. Molecular Biology and Evolution, 33, 1654–1668.

Schmid, S., Genevest, R., Gobet, E. et al. (2017) HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genotypic information from ancient DNA. Methods in Ecology and Evolution, 8, 1374–1388.

Silhowneu, P., Mutanen, M., Kaila, L., Brehm, G., Haussmann, A. & Staudte, H.S. (2011) Comprehensive molecular sampling yields a robust phylogeny for geometrid moths (Lepidoptera: Geometridae). Systematic Entomology, 36, 69–93.

Sohn, J.-C., Regier, J.C., Mitter, C. et al. (2013) A molecular phylogeny for Yponomeutoidea (Insecta, Lepidoptera, Ditrysia) and its implications for classification, biogeography and the evolution of host plant use. PLoS One, 8, e55066.

Sohn, J.-C., Regier, J.C., Mitter, C. et al. (2016) Phylogeny and feeding trait evolution of the mega-diverse Gelechioidea (Lepidoptera: Obiectomeram): new insight from 19 nuclear genes. Systematic Entomology, 41, 112–132.

St Laurent, R.A., Hamilton, C.A. & Kawahara, A.Y. (2018) Museum specimens provide phylogenomic data to resolve relationships of sack-bearer moths (Lepidoptera, Mimallonidae, Mimallonidae). Systematic Entomology, 43, 729–761.

Staude, H.S. (2011) Comprehensive molecular sampling yields a robust phylogeny for the oldest ditrysian lineages of extant Lepidoptera, with implications for classification, comparative morphology and life-history evolution. Systematic Entomology, 40, 671–704.

Revell, L.J. (2012) Phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3, 217–223.

Roure, B., Baurain, D. & Philippe, H. (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Molecular Biology and Evolution, 30, 197–214.

Sann, M., Niehuis, O., Peters, R.S. et al. (2018) Phylogenetic analysis of Apoidea sheds new light on the sister group of bees. BMC Evolutionary Biology, 18, 71.

Sayyari, E. & Mirarab, S. (2016) Fast coalescent-based computation of local branch support from quartet frequencies. Molecular Biology and Evolution, 33, 1654–1668.

Schmid, S., Genevest, R., Gobet, E. et al. (2017) HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genotypic information from ancient DNA. Methods in Ecology and Evolution, 8, 1374–1388.

Silhowneu, P., Mutanen, M., Kaila, L., Brehm, G., Haussmann, A. & Staudte, H.S. (2011) Comprehensive molecular sampling yields a robust phylogeny for geometrid moths (Lepidoptera: Geometridae). Systematic Entomology, 36, 69–93.
Suchan, T., Pitteloud, C., Gerasimova, N.S. et al. (2016) Hybridization
capture using RAD probes (hyRAD), a new tool for performing
genomic analyses on collection specimens. *PLoS One*, 11, e0151651.

Suyama, M., Torrents, D. & Bork, P. (2006) PAL2NAL: robust con-
version of protein sequence alignments into the corresponding codon
alignments. *Nucleic Acids Research*, 34, W609–W612.
The Heliconius Genome Consortium (2012) Butterfly genome reveals
promiscuous exchange of mimicry adaptations among species.
*Nature*, 487, 94–98.
The International Silkworm Genome Consortium (2008) The genome of a
lepidopteran model insect, the silkworm *Bombyx mori*. *Insect
Biochemistry and Molecular Biology*, 38, 1036–1045.
Thessen, A.E., Patterson, D.J. & Murray, S.A. (2012) The taxonomic
significance of species that have only been observed once: the genus
*Gynnoedinium* (Dinoflagellata) as an example. *PLoS One*, 7, e44015.
Timmermans, M.J.T.N., Lees, D.C. & Simonsen, T.J. (2014) Towards
a mitogenomic phylogeny of Lepidoptera. *Molecular Phylogenetics
and Evolution, 79*, 169–178.
Timmermans, M.J.T.N., Viberg, C., Martin, G., Hopkins, K. & Vogler,
A.P. (2016) Rapid assembly of taxonomically validated mitochondrial
genomes from historical insect collections. *Biological Journal of the
Linean Society*, 117, 83–95.
Timmermans, M.J.T.N., Daghmoumi, S.M., Glass, D., Hamilton, C.A.,
Kawahara, A.Y. & Kitching, I.J. (2019) Phylogeny of the hawkmoth
tribe Ambulycini (Lepidoptera: Sphingidae): Mitogenomes from
museum specimens resolve major relationships. *Insect Systematics and
Diversity, 3*, 1–8.
Tin, M.M.-Y., Economo, E.P. & Mikheyev, A.S. (2014) Sequencing
degraded DNA from non-destructively sampled museum specimens
for RAD-zagging and low-coverage shotgun phylogenetics. *PLoS
One*, 9, e96793.
Toussaint, E.F.A., Breinholt, J.W., Earl, C.J. et al. (2018) Anchored
phlogenomics illuminates the skipper butterfly tree of life. *BMC
Evolutionary Biology*, 18, 101.
Twort, V.G., Minet, J., Wheat, C.W. & Wahlberg, N. (2020) Museomics
of a rare taxon: placing Whalleyanidae in the Lepidoptera tree of life.
*bioRxiv*, 2020.08.18.255182.
Vasilikopoulos, A., Balke, M., Beutel, R.G. et al. (2019) Phylogenomics
of the superfamly Dytiscoidea (Coleoptera: Adephaga) with an
evaluation of phylogenetic conflict and systematic error. *Molecular
Phylogenetics and Evolution, 135*, 270–285.
Vasimuddin, M., Misra, S., Li, H. & Aluru, S. (2019) Efficient
architecture-aware acceleration of BWA-MEM for multicore systems.
In 2019 *IEEE International Parallel and Distributed Processing Sym-
opsmum (IPDPS)*. Presented at the 2019 *IEEE International Parallel
and Distributed Processing Symposium (IPDPS)*, Rio de Janeiro,
Brazil, pp. 314–324.
Vuado, A.D., Fritz, M.L. & López-Uribe, M.M. (2018) Opening the
door to the past: accessing phylogenetic, pathogen, and population
data from museum curated bees. *Insect Systematics and Diversity, 2*,
1–14.
Wagner, C.E., Keller, I., Wittwer, S. et al. (2013) Genome-wide RAD
sequence data provide unprecedented resolution of species boundaries
and relationships in the Lake Victoria cichlid adaptive radiation.
*Molecular Ecology*, 22, 787–798.
Wahlberg, N. & Wheat, C.W. (2008) Genomic outposts serve the
phylogenomic pioneers: designing novel nuclear markers for
genomic DNA extractions of Lepidoptera. *Systematic Biology*, 57,
231–242.
Wang, H., Holloway, J.D., Wahlberg, N., Wang, M. & Nylin, S. (2019)
Molecular phylogenetic and morphological studies on the systematic
position of *Heraca discivitita* reveal a new subfamily of Pseudo-
bistonidae (Lepidoptera: Geometroidea). *Systematic Entomology, 44*,
211–225.
Wheeler, Q.D., Knapp, S., Stevenson, D.W. et al. (2012) Mapping the
biosphere: exploring species to understand the origin, organization
and sustainability of biodiversity. *Systematics and Biodiversity, 10*,
1–20.
Wickham, H. (2011) ggplot2. *Wiley Interdisciplinary Reviews: Compu-
tational Statistics*, 3, 180–185.
Wiens, J.J. & Morrill, M.C. (2011) Missing data in phylogenetic
analysis: reconciling results from simulations and empirical data.
*Systematic Biology*, 60, 719–731.
Wiens, J.J. & Liu, J. (2012) Highly incomplete taxa can rescue phyloge-
etic analyses from the negative impacts of limited taxon sampling.
*PLoS One*, 7, e49225.
Wood, H.M., González, V.L., Lloyd, M., Coddington, J. & Scharff,
N. (2018) Next-generation museum genomics: phylogenetic relation-
ships among palpimanoid spiders using sequence capture techniques
(Araneae: Palpimanoida). *Molecular Phylogenetics and Evolution,
127*, 907–918.
Yeates, D.K., Zwick, A. & Mikheyev, A.S. (2016) Museums are biobanks:
unlocking the genetic potential of the three billion speci-
mens in the world’s biological collections. *Current Opinion in Insect
Science*, 18, 83–88.
You, M., Yue, Z., He, W. et al. (2013) A heterozygous moth genome
provides insights into herbivory and detoxification. *Nature Genetics,
45*, 220–225.
Yu, G., Smith, D.K., Zhu, H., Guan, Y. & Lam, T.T.-Y. (2017) Ggtree:
an r package for visualization and annotation of phylogenetic trees
with their covariates and other associated data. *Methods in Ecology
and Evolution, 8*, 26–36.
Zahiri, R., Lafontaine, D., Schmidt, C. et al. (2013) Relationships among
the basal lineages of Noctuidae (Lepidoptera, Noctuoidea) based on
eight gene regions. *Zeotologica Scripta*, 42, 488–507.
Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D. et al. (2017) OrthoDB
v9.1: cataloging evolutionary and functional annotations for animal,
fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids
Research, 45*, D744–D749.
Zhan, S., Merlin, C., Boone, J.L. & Reppert, S.M. (2011) The monarch
butterfly genome yields insights into long-distance migration.
*Cell*, 147, 1171–1185.
Zhang, C., Sayyari, E. & Mirarab, S. (2017) ASTRAL-III: increased
scalability and impacts of contracting low support branches. *Compara-
tive Genomics, Lecture Notes in Computer Science. Presented at
the RECOMB International Workshop on Comparative Genomics,
pp. 53–75. Springer, Cham.
Zhang, J., Cong, Q., Shen, J., Brockmann, E. & Grishin, N.V. (2019a)
Genomes reveal drastic and recurrent phenotypic divergence in firetip
skipper butterflies (Hesperidae: Pyrrhopyginae). *Proceedings of the
Royal Society B: Biological Sciences, 286*, 20190609.
Zhang, J., Shen, J., Cong, Q. & Grishin, N.V. (2019b) Genomic analysis
of the tribe Emesini (Lepidoptera: Riodinidae). *Zootaxa, 4668*,
475–488.
Zhang, J., Lees, D.C., Shen, J. et al. (2020a) The mitogenome of a
Malagasy butterfly *Malaca fasciatus* (Mabille, 1884) recovered from
the holotype collected over 140 years ago adds support for a new
subfamily of Hesperidae (Lepidoptera). *Genome, 63*, 195–202.
Zhang, Y., Huang, S., Liang, D., Wang, H. & Zhang, P. (2020b) A
multilocus analysis of Epicopeiidae (Lepidoptera, Geometroidea)
provides new insights into their relationships and the evolutionary
history of mimicry. *Molecular Phylogenetics and Evolution, 149*,
106847.
Zimmermann, J., Hajibabaei, M., Blackburn, D.C. et al. (2008) DNA damage in preserved specimens and tissue samples: a molecular assessment. *Frontiers in Zoology*, 5, 18.

Zolotuhin, V.V., Kurshakov, P.A. & Solovyev, A.V. (2014) The use of muscle morphology of Chrysopolomidae (Lepidoptera) genitalic appendages in developing a family system. *Entomological Review*, 94, 181–189.

Zwick, A., Regier, J.C. & Zwickl, D.J. (2012) Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. *PLoS One*, 7, e47450.

Accepted 8 March 2021
First published online 29 March 2021