Systematic Entomology (2021), **46**, 926–937 DOI: 10.1111/syen.12503

Museomics of a rare taxon: placing Whalleyanidae in the Lepidoptera Tree of Life

VICTORIA G. TWORT[1,2], JOËL MINET[3], CHRISTOPHER W. WHEAT[4] and NIKLAS WAHLBERG[1]

[1]Department of Biology, Lund University, Lund, Sweden, [2]The Finnish Museum of Natural History Luomus, Zoology Unit, University of Helsinki, Helsinki, Finland, [3]Muséum National d'Histoire Naturelle, ISYEB, Paris, France and [4]Department of Zoology, Stockholm University, Stockholm, Sweden

Abstract. Museomics is a valuable approach that utilizes the diverse biobanks that are natural history museums. The ability to sequence genomes from old specimens has expanded not only the variety of interesting taxa available to study but also the scope of questions that can be investigated in order to further knowledge about biodiversity. Here, we present whole genome sequencing results from the enigmatic genus *Whalleyana* (comprising two species – occurring in drier biomes of Madagascar – previously placed in a monotypic superfamily, Whalleyanoidea), as well as from certain species of the families Callidulidae and Hyblaeidae (Calliduloidea and Hyblaeoidea, respectively). Library preparation was carried out on four museum specimens and one existing DNA extract and sequenced with Illumina short reads. *De novo* assembly resulted in highly fragmented genomes with the N50 ranging from 317 to 2078 bp. Mining of a manually curated gene set of 331 genes from these draft genomes had an overall gene recovery rate of 64–90%. Phylogenetic analysis places *Whalleyana* as sister to Callidulidae and *Hyblaea* as sister to Pyraloidea. Since the former sister-group relationship turns out to be also supported by ten morphological synapomorphies, we propose to formally assign the Whalleyanidae to the superfamily Calliduloidea. These results highlight the usefulness of not only museum specimens but also existing DNA extracts, for whole genome sequencing and gene mining for phylogenomic studies.

Introduction

Natural history museums represent a diverse biobank of many interesting extant, rare and extinct taxa, making them an important scientific resource (Shaffer *et al.*, 1998; Graham *et al.*, 2004; Suarez & Tsutsui, 2004; Yeates *et al.*, 2016). Additionally, many species present in these collections are more accessible than in their original habitats due to a variety of factors (Wandeler *et al.*, 2007; Thessen *et al.*, 2012), such as remote geographical distributions, taxa being rare or endangered, as well as taxa that have since gone extinct or have not been seen again following their initial collection. Although natural history collections have primarily been used for traditional morphological and taxonomic studies, ongoing advances in DNA sequencing technologies have expanded their role into the realm of genetics, and many other fields (Shaffer *et al.*, 1998; Paijmans *et al.*, 2013; Wiley *et al.*, 2013). Despite the fact that these collections are now being recognized as an important genetic resource, most of the specimens in museums were collected prior to the use of DNA sequencing technology and were thus not preserved with the conservation of DNA in mind. Hence, the DNA from these samples is often damaged and degraded and they were considered unsuitable for traditional molecular methods (Wandeler *et al.*, 2007). In spite of this, museum specimens have been and continue to be used for molecular studies (Bi *et al.*, 2013; Besnard *et al.*, 2014, 2016; Chang *et al.*, 2017; Call *et al.*, 2021). Initial studies focused on using PCR and Sanger sequencing of short fragments of genes (Houde & Braun, 1988; Thomas *et al.*, 1990; Cooper *et al.*, 2006); however, this approach not only requires the development of very specific primers for each gene, hence relying on prior genetic knowledge, but can also be cost prohibitive, and laborious (Soltis & Soltis, 1993; Wandeler *et al.*, 2007).

Correspondence: Victoria G. Twort, Department of Biology, Lund Universitet, Sölvegatan 37, Lund SE 223 62, Sweden. E-mail: niklas.wahlberg@biol.lu.se

The development of high-throughput sequencing (HTS) technologies offers a promise of more efficient ways of sequencing DNA from museum specimens (Rizzi *et al.*, 2012; Hofreiter *et al.*, 2015). This is because HTS involves the sequencing of short fragments of DNA, which is a typical characteristic of DNA extracted from museum specimens, and results in large volumes of sequence data from relatively small amounts of starting material that provides good genome-wide genetic data. The publication of the Mammoth genome (Poinar *et al.*, 2006) followed by the Neanderthal genome (Green *et al.*, 2010; Meyer *et al.*, 2012) showed the promise of HTS technologies using fragmented ancient DNA. Since then, HTS has slowly been applied to a more diverse range of taxa. One particular application being widely used is targeted sequence capture, in which focal regions of the genome are isolated and sequenced. The regions targeted are typically those that are well conserved across the taxa of interest, such as exons and ultraconserved elements, so that only a single probe set is required to be designed for the group(s) of interest. These target-based approaches have been applied to a variety of taxa and specimen ages (Bi *et al.*, 2013; Staats *et al.*, 2013; Bailey *et al.*, 2016; Blaimer *et al.*, 2016; Prosser *et al.*, 2016; Call *et al.*, 2021; Mayer *et al.*, 2021) and have proven relatively successful in recovering the regions of interest for phylogenomic studies. Despite these advances, application of these methods requires prior genetic knowledge of the taxa of interest in order to facilitate probe design before sequencing even begins.

An alternative approach to targeted sequence capture is whole-genome sequencing of the extracted DNA. Although initial studies focused on taxa for which a reference genome, or one closely related, already existed (Rowe *et al.*, 2011; Staats *et al.*, 2013), *de novo* based approaches are starting to be used. Nevertheless, these studies tend to be carried out with taxa for which large volumes of starting material are available for DNA extraction, with very few studies using whole genome-based approaches on insect specimens older than a few decades (Staats *et al.*, 2013; Heintzman *et al.*, 2014; Maddison & Cooper, 2014; Tin *et al.*, 2014; Kanda *et al.*, 2015). Initial studies focused on recovering specific regions, such as mitochondrial or ribosomal DNA, which are present in multiple copies per cell (Staats *et al.*, 2013; Heintzman *et al.*, 2014). Kanda *et al.* (2015) highlighted that the recovery of low-copy regions of the genome is possible from a diversity of museum specimens spanning a variety of ages, preservation methods and DNA quality. The use of both *de novo* and reference-based assemblies for gene recovery showed that although more loci can be recovered if one has an existing reference, many loci are still obtained with a *de novo* approach (Kanda *et al.*, 2015). Additionally, Sproul & Maddison (2017) presented a nondestructive DNA extraction method for historical beetle samples and successfully prepared and sequenced libraries from the low amounts of degraded DNA recovered with good results. These studies show that for historical samples of interest, it is possible to get enough DNA for sequencing and gene recovery without the need for any prior genetic knowledge.

Here, we explore the application of museomics, which is defined as the use of HTS to generate useful genomic data from museum specimens, to resolve the phylogenetic position of the enigmatic *Whalleyana* moths. The genus *Whalleyana* is endemic to Madagascar and was first described in 1977 as an odd member of the Thyrididae (Viette, 1977). Later Minet (1991) placed the genus in its own family Whalleyanidae, subsequently assigned to the monotypic superfamily Whalleyanoidea (Carter & Kristensen, 1998), as it was not clear from the morphology what other families of Lepidoptera were closely related to *Whalleyana*. Very little is known about the two species (*Whalleyana vroni* Viette and *Whalleyana toni* Viette) that make up this genus, including their phylogenetic placement within Lepidoptera. In addition, neither species seem to have been collected after the 1990s. Thus, in order to get a better understanding of this taxon, existing museum samples are the only resource available. The aim of this study is to utilize low-coverage whole genome sequencing of interesting museum specimens and existing DNA extracts to highlight the usefulness of such approaches for answering questions, such as where in the Lepidoptera phylogeny *Whalleyana* belongs. We also sequence museum specimens or existing genomic DNA extracts of potentially related taxa belonging to the families Callidulidae and Hyblaeidae (possible sister groups according to Regier *et al.* (2013)). Such approaches not only allow us access to taxa we may not have had the opportunity to investigate previously, but by utilizing a whole genome sequencing approach, we are able to generate a rich dataset for future researchers using diverse approaches. We assessed the robustness of our results with morphological data from only the adult stage, as the early stages of the Whalleyanidae remain completely unknown.

Material and methods

Taxon sampling for molecular analyses

Whole genome sequencing data were generated from museum specimens of the following species: *W. vroni* Viette, 1977 (collected in 1969), *Helicomitra pulchra* Butler (collected in 1974), *Griveaudia vieui* Viette (collected in 1969), *Hyblaea madagascariensis* Viette, 1961 (collected in 1975), as well as a DNA extract of *Hyblaea puera* (Cramer) from Mutanen *et al.* (2010). *Helicomitra* and *Griveaudia* belong to different subfamilies, Pterothysaninae and Griveaudiinae, respectively, of the family Callidulidae, and *Hyblaea* to Hyblaeidae. Both families are potentially related to *Whalleyana*, and the latter has no genomic resources available, with the exception of a partial DNA barcode, prior to our study. All museum specimens are from the Muséum National d'Histoire Naturelle, Paris (MNHN). The DNA extract of *H. puera* was kindly loaned by Marko Mutanen (University of Oulu, Finland).

DNA extraction

DNA was extracted from the abdomens of four specimens, using the QIAamp DNA Micro kit (Qiagen) following the manufacturer's protocol, with the following modifications: no crushing of the samples was carried out prior to incubation with

lysis buffer, following overnight incubation samples were centrifuged and the supernatant carried forward for extraction with the remaining tissue being placed in ethanol, and elution buffer was incubated in the columns for 20 min at room temperature prior to elution. We took reasonable measures to avoid contamination, including the use of filter tips and sterilized work areas that are physically separated from areas where fresh specimens are prepared. The resulting DNA extracts were visualized on 0.8% w/v agarose gels stained with SYBR safe (Fisher Scientific) to determine DNA fragmentation levels. In the case of the *H. puera* extract, due to high molecular weight DNA, DNA was sonicated to approximately 200–300 bp fragments using a Bioruptor® with the following settings: (M) medium power output, 30 s ON/90 s OFF pulses for 30 min in a 4°C water bath, followed by vacuum centrifugation and resuspension in 50 μL of elution buffer.

Library preparation and sequencing

Library preparation followed a modified protocol of Meyer & Kircher (2010). All reagent distributors and catalogue numbers are given in Table S1. Firstly, DNA was blunt-end repaired. The reaction mix consisted of 1x Tango buffer, dNTP (100 μM each), 1 mM ATP, 0.5 U/μL PNK and 0.1 U/μL T4 DNA Polymerase. The reaction was incubated for 15 min at 25°C, followed by 5 min at 12°C. Purification of the reaction was carried out with the MinElute purification kit (Qiagen), and elution in 22 μL EB buffer. Adapter ligation followed purification with a reaction mix containing: 1× T4 ligation buffer, 5% PEG-4000, 0.125 U/μL T4 Ligase and an adapter mix of P7 and P5 adapters (Meyer & Kircher, 2010) 2.5 μM each. Reactions were incubated for 30 min at 22°C. After purification with the MinElute purification kit (Qiagen), adapter fill-in was performed using the following reaction mix: 1× Isothermopol amplification buffer, dNTP (250 μM) and 0.3 U/μL Bst polymerase. Incubation at 37°C for 20 min was followed by the final heatkill performed by incubation for 20 min at 80°C.

Indexing and amplification of each library was carried out with 3 μL of library template and a unique dual indexing strategy. The amplification mix consisted of 0.05 U/μL AccuPrime Pfx DNA Polymerase, 2.5 μL AccuPrime reaction mix, 200 nM IS4 primer (Meyer & Kircher, 2010) and 200 nM of indexing primer. Amplification was carried out under the following conditions: 95°C for 2 min, 18 cycles of 95°C for 15 s, 60°C for 30 s and 68°C for 60 s, which were carried out in six independent reactions, to avoid amplification bias, and pooled prior to purification. Purification along with size selection was carried out using a two-step process with Agencourt AMPure XP beads. An initial bead concentration of 0.5× was used to remove long fragments that are likely to represent contamination from fresh DNA, libraries were selected with a bead concentration of 1.8× to size select the expected library range of 100–300 bp. The resulting libraries were quantified and quality checked with Quanti-iT™ PicoGreen™ dsDNA assay and with a DNA chip on a Bioanalyzer 2100, respectively. Multiplexed libraries were pooled as follows: *W. vroni* was pooled at 50% molar concentration in a pool of nine samples and sequenced over two runs, while the remaining specimens were pooled in equimolar concentrations in a pool of six samples and sequenced using the Illumina HiSeq2500 technology with 150 bp paired-end reads. A detailed laboratory protocol is available at Figshare DOI: https://doi.org/10.6084/m9.figshare.12927500.

Genome assembly

Raw reads were quality checked with FASTQC v0.11.8 (Andrews, 2010). Sequencing reads resulting from samples with highly degraded DNA were treated from this point as single-end reads. This approach was chosen, as degraded DNA is likely to randomly ligate together during the adapter ligation stage of library preparation, resulting in chimeras of different genomic regions (Willerslev & Cooper, 2005). Nevertheless, the sequencing information contained in the reads is still reliable, as chimera formation typically results in DNA inserts larger than read length; therefore, more reliable results are obtained by treating data as single-end (Rowe *et al.*, 2011). For the sample that underwent sonication (*H. puera*), reads were carried forward as paired-end. Reads with ambiguous bases (N's) were removed from the dataset using Prinseq 0.20.4 (Schmieder & Edwards, 2011). Trimmomatic 0.38 (Bolger *et al.*, 2014) was used to remove low-quality bases from their beginning (LEADING:3) and end (TRAILING:3), by removing reads below 30 bp and by evaluating read quality with a sliding window approach. Quality was measured for sliding windows of four base pairs and had to be greater than PRHED 25 on average. The resulting cleaned reads were used for *de novo* genome assembly with spAdes v3.13.0 (Bankevich *et al.*, 2012) with kmer values of 21, 33 and 55. The completeness of each assembly was assessed using BUSCO 3.0.2b (Simão *et al.*, 2015) using the Insecta lineage set. However, due to the fragmented nature of the genomes, BUSCO has trouble identifying orthologs; therefore, the genomes were searched for the Insecta lineage set using a tblastn approach (e-value threshold 1e-5, minimum identity of 60%) with standalone BLAST 2.9.0 (Camacho *et al.*, 2009).

Orthologue identification and alignment

Orthologues were identified from the fragmented genome assemblies using MESPA v1.3 (Neethiraj *et al.*, 2017), which is an exon aware amino acid to DNA genome aligner, which attempts to scaffold exons together from highly fragmented genome assemblies. For the amino acids used for mining the genomes, we used a custom set of 331 representative gene markers (11 of which are mitochondrial), which have been manually vetted for alignment and orthology based on their amino acid sequences from a set of 195 taxa of Lepidoptera. The details of the vetting process are described in Rota *et al.* (2021). The resulting DNA sequences obtained from MESPA were aligned to pre-existing reference alignments for these genes (taken from [Rota *et al.*, 2021]) using their translated amino

acid sequences using MAFFT v7.310 (Katoh *et al.*, 2002) using the 'add fragments' and 'auto' options, which keeps existing gaps in the alignments and chooses the most appropriate alignment strategy. The resulting amino acid alignments were manually checked to ensure accuracy, screened for the presence of pseudogenes, reading frame errors and alignment errors using Geneious 11.0.3 (https://www.geneious.com). The amino acid alignments were then converted back to nucleotide alignments, and the aligned DNA sequences were curated and maintained using the Voseq database (Peña & Malm, 2012), which allows users to custom-make datasets for downstream phylogenetic analyses in chosen formats (e.g. FASTA, Nexus or Phylip formats). Raw sequencing data can be found under Bioproject PRJNA631866, while genome assemblies can be accessed from Zenodo, DOI: https://doi.org/10.5281/zenodo.3629334.

Phylogenetic analysis

In order to investigate the phylogenetic placement of *Whalleyana*, the new sequences were added to a manually curated dataset derived from published transcriptomes and genomes of ditrysian Lepidoptera (Rota *et al.*, 2021). The final dataset consisted of a total of 337 gene fragments, spanning 331 genes, across 169 taxa (164 taxa were taken from Rota *et al.* (2021), see Table S2 for the full list of included taxa and Table S3 for the genes recovered from the specimens in this study), with the final alignment file being created in Voseq.

The resulting nucleotide and amino acid (aa) sequence alignments were analysed in a maximum likelihood framework using the program IQ-TREE (Nguyen *et al.*, 2014). For the nucleotide dataset, third codon positions were removed (nt12). Nucleotide data were also analysed using degen1 coding (Regier *et al.*, 2010; Zwick *et al.*, 2012). Each dataset was analysed partitioned by gene, with ModelFinder (Kalyaanamoorthy *et al.*, 2017) run first, and then the maximum likelihood search run after based on the optimal models found for each gene. The robustness of our phylogenetic hypotheses was assessed with 1000 ultrafast bootstrap (UFBoot2) approximations (Hoang *et al.*, 2017) in IQ-TREE. Analyses were run on the CIPRES portal (Miller *et al.*, 2010).

Morphological analyses

Adult morphology was investigated using a collection of specimens from MNHN dissected by one of us (JM). These specimens had been chosen to represent most ditrysian families within the framework of several previous publications (Minet, 1991; Rajaei *et al.*, 2015). Among the families that are more precisely the focus of the present study, specimens that have been entirely dissected belong to the following genera: *Striglina* Guenée (Thyrididae: Striglininae), *Marmax* Rafinesque (Thyrididae: Charideinae), *Thyris* Laspeyres (Thyrididae: Thyridinae), *Chrysotypus* Butler (Thyrididae: Siculodinae), *Rhodoneura* Guenée (same subfamily), *Whalleyana* Viette (Whalleyanidae), *Helicomitra* Butler (Callidulidae: Pterothysaninae), *Griveaudia* Viette (Callidulidae: Griveaudiinae), *Callidula* Hübner (Callidulidae: Callidulinae) and *Hyblaea* Fabricius (Hyblaeidae). After removal of their wings, these imagos were macerated in a hot 10% potassium hydroxide solution (KOH), rinsed in demineralized water, then cleaned, descaled, stained (with Chlorazol Black E) and dissected in 70% ethanol (following methods expounded by Brock (1971) and Robinson (1976)). Afterwards, the various parts of the body were severed from adjacent regions and either stored intact in 70% ethanol or preserved as permanent slide mounts in Euparal (following standard techniques: [Robinson, 1976]). Structures of possible phylogenetic interest were photographed and/or examined using an Olympus SZH stereo microscope with a linear magnification range of ×7.5 to ×128. In the search of apomorphic traits suited to support, or not, our molecular phylogeny, special attention was paid to the less homoplastic characters, nevertheless without neglecting any character easy to polarize through outgroup comparisons. External characters whose observation does not require dissections were surveyed on a large scale and full account was taken of published morphological data, especially in the case of the hyblaeoid family Prodidactidae (Epstein & Brown, 2003; Kaila *et al.*, 2013).

Results

The molecular approach

DNA extraction of the four museum specimens was successful with DNA fragments ranging from 70 to 300 bp in size, while sonication of the *H. puera* extract resulted in DNA fragment lengths of 300 bp (results not shown). Sequencing resulted in a total of 754 million reads across the runs, ca. 462 million reads belonged to *W. vroni*, and an average of 72 million reads for the other four samples (Table 1). Of these reads >86% passed adapter and quality trimming. As there is no reference genome available for any related taxa, the genome of each sample was *de novo* assembled. The resulting genome assemblies were highly fragmented with average contig lengths of 321 bp and N50's ranging from 317 to 2078 bp (Table 1). Assessment of the completeness of the resulting assemblies with BUSCO highlighted the difficulty the program has in finding orthologues in fragmented genomes (results not shown). However, the orthologue set can still be used with a BLAST approach to assess the presence of the conserved genes. The blast search for the Insecta orthologues showed that the majority of orthologues are present in at least fragmented form with between 74% and 87% being present in the genomes (Table 1).

Identification of the curated Lepidoptera gene set with MESPA had a recovery rate of between 64% and 90% (Tables 1 and S3). The resulting sequences were uploaded to an in-house database (Voseq: Peña & Malm, 2012), and a final concatenated dataset comprising a total of 162 taxa and 291 516 nucleotides in length was used for analyses. Analysis of both the nucleotide and amino acid datasets shows stable placement of *Whalleyana* as sister to Callidulidae and *Hyblaea* as sister to Pyraloidea (Fig. 1). Both of

Table 1. Genome assembly and gene recovery statistics.

Sample	*Griveaudia vieui*	*Hyblaea madagascariensis*	*Helicomitra pulchra*	*Whalleyana vroni*	*Hyblaea purea*
Code	VT58	VT57	VT56	VT11	MM07227
Data treated as	SE	SE	SE	SE	PE
Raw reads (paired)	73 451 727	81 670 390	68 286 636	462 344 781	68 569 202
Cleaned read pairs	–	–	–	–	23 863 949
Cleaned read unpaired R1	65 227 946	71 533 273	61 719 435	422 057 592	37 539 693
Cleaned reads unpaired R2	63 764 833	69 860 500	58 962 343	407 003 611	391 589
Contigs	700 194	746 054	1 155 426	1 639 567	985 209
Max contig length	5413	4792	5441	7920	65 231
Minimum contig length	56	56	56	56	56
Average contig length	284.0 ± 147.9	330.2 ± 183.0	278.5 ± 197.7	295.5 ± 330.1	417.4 ± 1180.6
Median contig length	264	291	250	209	108
Total contig length	198 848 300	246 324 962	321 785 346	484 482 679	411 261 707
% Non-ATCG characters	0.001	0.001	0.005	0.01	0.287
Contigs ≥ 100 bp	610 227	683 788	941 766	1 104 568	545 487
Contigs ≥ 200 bp	577 430	653 684	735 144	835 486	301 242
Contigs ≥ 500 bp	51 490	101 611	135 739	266 889	139 271
Contigs ≥ 1 kbp	1429	6611	8871	65 308	84 638
Contigs ≥ 10 kbp	–	–	–	–	2639
N50 value	317	367	361	478	2078
Number of genes recovered (/338)	215 (63%)	244 (72%)	244 (72%)	255 (75%)	304 (90%)
Number of BUSCO Insecta genes identified (/1658)	1201 (72%)	1415 (85%)	1396 (84%)	1435 (87%)	1400 (84%)

SE: single end; PE: paired end.

these relationships have 100% UFbootstrap support regardless of data form, except for nt12, where the sister relationship of *Hyblaea* and Pyraloidea receives only 86%.

Morphological synapomorphies and autapomorphies

The interpretation of the following characters is based on Fig. 1, but takes also into account the uncertainties mentioned hereafter (see Discussion) about the interrelationships between Gelechioidea, Papilionoidea and the Thyridoidea + Callidu-loidea lineage (Rota *et al.*, 2021). It should be noted that several losses have been regarded as possible synapomorphies of certain groups although they may represent homoplastic traits relatively difficult to interpret (e.g. loss of the male retinaculum in Callidulidae may be regarded either as a callidulid autapomorphy, with independent loss in one of the two whalleyanid species, or as a synapomorphy of Callidulidae and Whalleyanidae, with reappearance of this retinaculum (reversal) in the other species of Whalleyanidae; nevertheless, such a reversal is not very likely for functional reasons insofar as both whalleyanid species have similar wing shapes).

Thyrididae, Whalleyanidae and Callidulidae share six imaginal synapomorphies: (i) on the head, the ocellus is either absent or devoid of a distinct lens (as also in all Papilionoidea and Hyblaeoidea); (ii) in the forewing, vein M2 arises closer to M3 than to M1 (a frequently encountered apomorphy, which also occurs in Hyblaeoidea + Pyraloidea but does not pertain to the ground plan of Papilionoidea nor to that of Gelechioidea: for instance, in the latter superfamily, M2 arises closer to M1 than to M3 in such genera as *Hypertropha* Meyrick and *Donacostola* Meyrick); (iii) at the base of the forewing, the spinarea is absent or extremely small (an apomorphy also present in all Papilionoidea but absent in the genus *Hyblaea* Fabricius, which retains a large spinarea (Common, 1990: fig. 108), and in many Gelechioidea and Pyraloidea); (iv) both fore- and hindwings lack a distinct, tubular CuP (this vein being replaced by a fold, which may resemble a vein in certain large Thyrididae [e.g. in the genus *Draconia* Hübner]); by contrast a true vein CuP is preserved in both pairs of wings of Prodidactidae (Hyblaeoidea) (Epstein & Brown, 2003: fig. 9) and in the hindwing of *Hyblaea*; (v) in the hindwing, vein Sc+R is approximated to, or fused with, vein Rs (beyond wing base and either before or beyond the upper angle of the discal cell) (in the genus *Prodidactis* Meyrick, only the base of Sc is approximated to the upper edge of the hindwing discal cell); (vi) in the male genitalia, the juxta is provided with a pair of erect 'arms' that are directed caudad or dorsad (Fig. 2A, arrow; since Whalleyanidae and Callidulidae appear as sister groups, the absence of these erect arms in Callidulidae should represent a loss rather than a primary condition). A larval trait may represent a seventh synapomorphy of these families (when the larva of *Whalleyana* is discovered), namely, the presence of just one seta in the L group of segment A9 (see fig. 7 in Chistyakov *et al.* (1994)). Although this apomorphy also occurs in *Hyblaea* and many Pyraloidea, two L setae are preserved (on A9, laterally) in Prodidactidae (Epstein & Brown, 2003: fig. 14) and three in the ground plan of the pyraloid larva (Neunzing, 1987: 463). Thyrididae and Callidulinae also share the following pupal apomorphy: the mandibles (pilifers sensu Mosher, 1916) are distinctly adjacent on the meson (Nakamura, 2011: figs. 1, 2 and 5). Nevertheless, they are not adjacent in the subfamily Pterothysaninae of Callidulidae

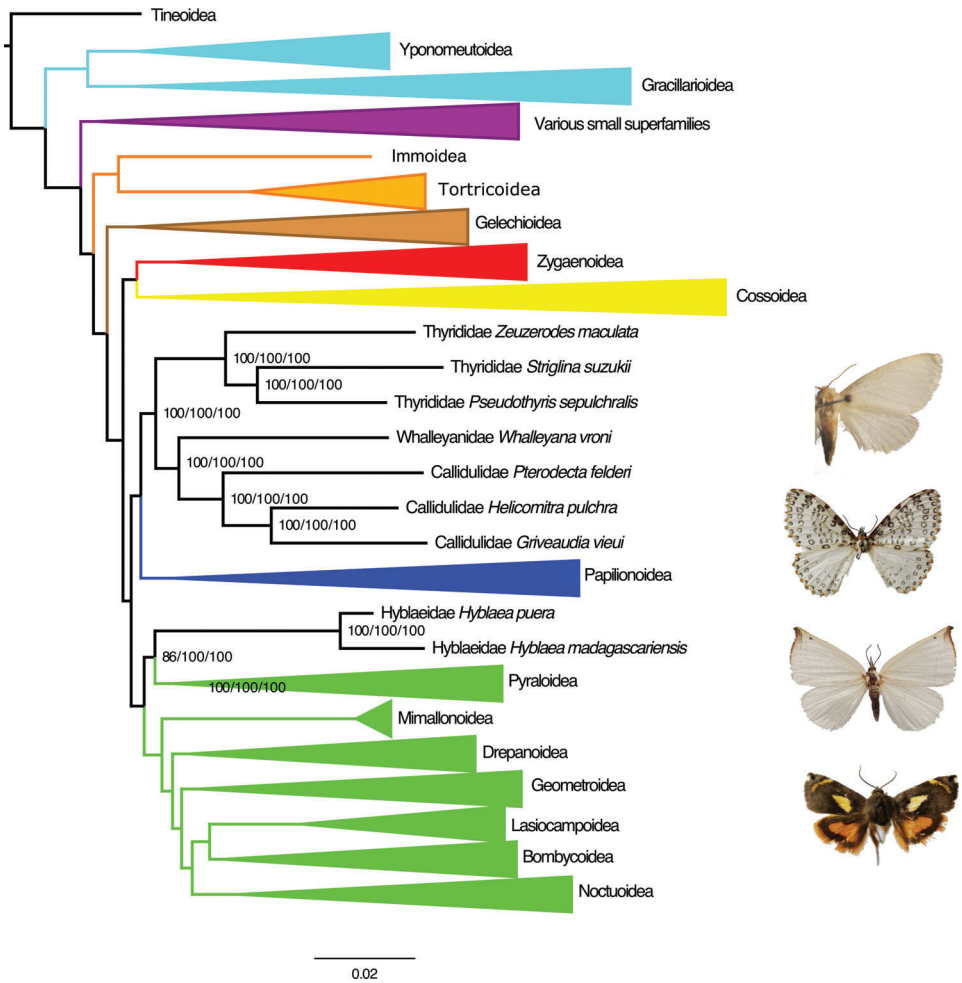

Fig. 1. Phylogenetic relationships of *Whalleyana*, *Helicomitra*, *Griveaudia* and *Hyblaea* based on 331 genes. Superfamilies whose internal relationships are not relevant to this study are shown as collapsed. Numbers to the right of each node give the ultrafast bootstraps for each dataset analysed: nt12/degen1/aa. Pictures to the left are of the sequenced specimens, from top to bottom: *Whalleyana vroni*, *Helicomitra pulchra*, *Griveaudia vieui* and *Hyblaea madagascariensis*.

(Nakamura, 2011: fig. 3) while the pupa of the Griveaudiinae remains unknown to date.

Ten synapomorphies from adult morphology clearly support a sister-group relationship of Whalleyanidae and Callidulidae: (vii) in the antennae of dried specimens, the flagellum is simple (i.e. neither dentate nor pectinate) but has its distal section somewhat sinuous and turned up apically (the original description of the Whalleyanidae (Minet, 1991: 89) states 'flagellum … on distal section curved as in the Callidulidae'; this antennal trait can also be seen in live adults (Wang, 1993, photo of a *Tetragonus catamitus* Geyer), although often less distinctly; (viii) on vertex, the chaetosemata are large and include minute scales between their setae; (ix) veins Rs2 and Rs3 are stalked in the forewing (all Rs veins are 'free' in many Thyrididae, *Prodidactis* (Janse, 1964: pl. 5), *Hyblaea*, and in the ground plan of the Papilionoidea (cf. Hesperiidae)); (x) in the forewing, veins Rs3 and Rs4 run to the termen, reaching it below the apex (Viette, 1977: fig. 1; Minet, 1998: fig. 15.1, B and C) (by contrast, only Rs4 runs to the termen in *Prodidactis*, *Hyblaea*, and in the thyridid ground plan (cf. Common, 1990: fig. 109.4); nevertheless, through parallel evolution, several Thyrididae also possess the apomorphy (10): Common, 1990: fig. 109); (xi) in the basal region of the hindwing, there is a recurrent humeral spur, or fold (*Whalleyana*), between Sc and the frenulum (Minet, 1998: fig. 15.1, B and C); (xii) in the hindwing, vein M2 arises much closer to M3 than to M1 (*Hyblaea* and most Thyrididae also have the hindwing vein M2 arising closer to M3 than to M1 but this vein arises midway between M1 and M3 in the thyridid ground plan (illustrated, for this character, by the genus *Addaea* Walker: Common, 1990: fig. 109.4) and arises slightly closer to M1 than to M3 in the hyblaeid genus *Erythrochrus* Herrich-Schäffer); (xiii) at the base of the abdomen, the marginotergites (term used by Brock (1971)) are anteriorly connected to the anterior angles of sternum A2 through complete tergosternal sclerites (Fig. 2B, long arrow) (in most Thyrididae, sternum A2 has

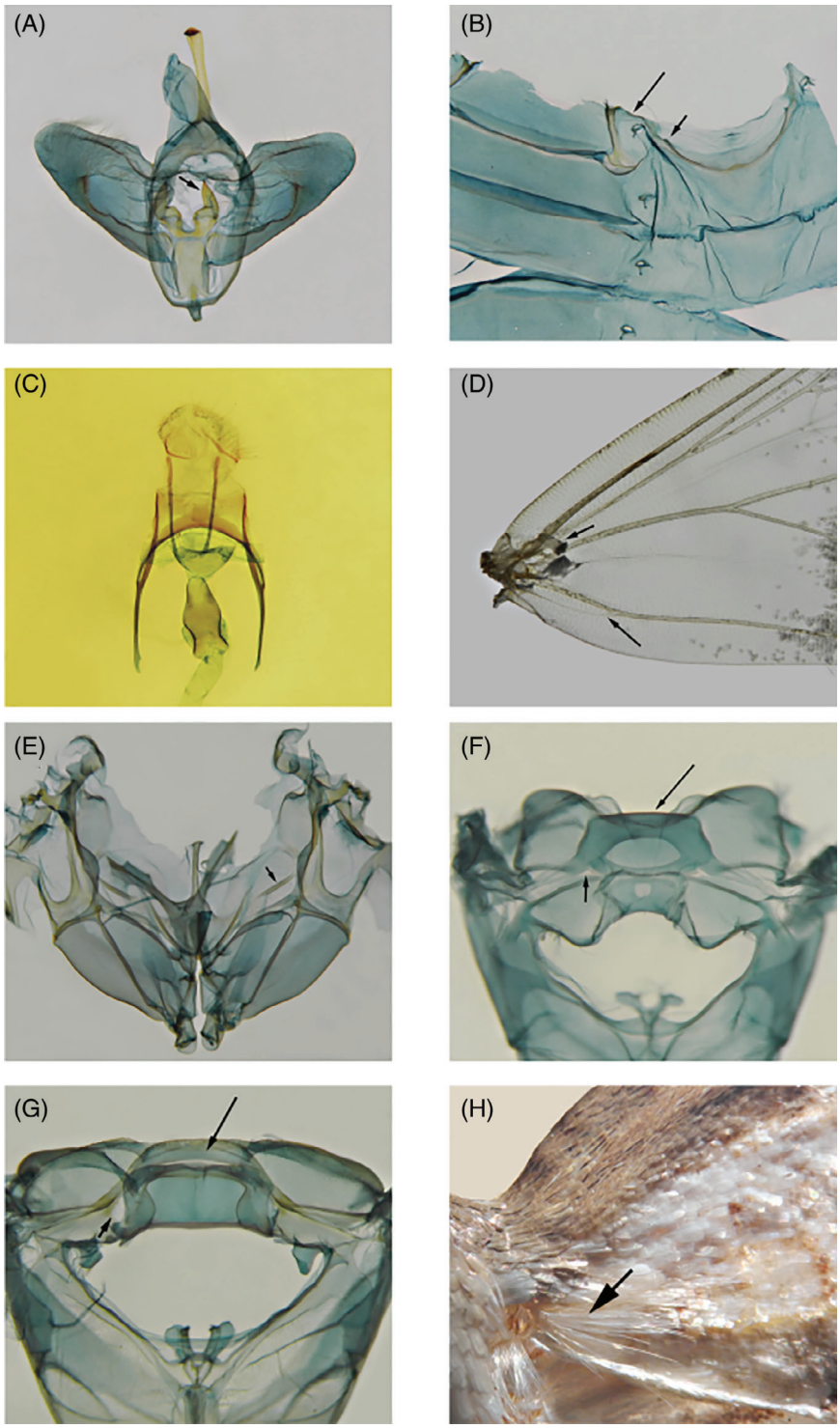

Fig. 2. (A) *Whalleyana vroni*, male genitalia (arrow: one of the two free, dorsally directed, arms of the juxta); (B) *W. vroni*, first three segments of the male abdomen, with the sterna on the right (short arrow: apodeme; long arrow: tergosternal sclerite); (C) *Whalleyana toni*, posterior region of the female genitalia (preparation P. Viette 5451); (D) *W. vroni*, base of the male forewing (ventral surface) after removal of most scales (short arrow: retinaculum; long arrow: lower branch of the 'anal fork'); (E) *W. vroni*, mesothoracic pleurosternum in anterior view (arrow: precoxal sulcus); (F) *Griveaudia vieui*, metathorax in posterior view (short arrow: left-hand fenestra lateralis; long arrow: scutellum); (G) *W. toni*, metathorax in posterior view (short arrow: ventral edge of the left-hand fenestra lateralis; long arrow: scutellum); H, *Rhodoneura opalinula*, forewing base in dorsal view (arrow: bunch of piliform scales arising from the base of vein A1).

just variously developed anterolateral processes that do not reach the marginotergites); (xiv) the apodemes of sternum A2 are short or reduced (Fig. 2B, short arrow) (although reduction of the apodemes also occurs in several Thyrididae, these structures are sometimes large or elongate in this family: for example, fig. 12 in Minet (1983)); (xv) the male genitalia lack a complete gnathos (retained in many Thyrididae); (xvi) in the female genitalia, the eighth sternum is transversely elongate and distinctly arched (concave cephalad) (Fig. 2C; see also, for Callidulidae, several figures in Holloway (1998)). Among these ten derived traits, we regard (vii), (viii), (xi), and (xvi) as really significant synapomorphies, which tend to support the results obtained with our molecular phylogenetic analysis. Therefore, we formally propose here to assign Whalleyanidae to the superfamily Calliduloidea (*revised concept*, with a definition based on the above-mentioned apomorphies (vii)–(xvi)).

Within the thus redefined Calliduloidea, nine autapomorphies support the monophyly of the family Callidulidae (= Pterothysaninae + Griveaudiinae + Callidulinae), namely, (xvii) foreleg with an apical pair of stronger spines on tarsomere 4, but with at most a few minute spines on the ventral surface of tarsomere 5 (Minet, 1990: figs. 4–6); (xviii) male forewing without a subcostal retinaculum (while *W. vroni* retains this retinaculum (Fig. 2D, short arrow); through parallel evolution, the male forewing of *W. toni* has lost this structure); (xix) in the forewing, anal vein simple, devoid of 'basal fork' (A2 being at most a very short veinlet parallel to the base of vein A1 (Minet, 1998: fig. 15.1 B); by contrast, *Whalleyana* retains this 'basal fork', although with a weak lower branch: Fig. 2D, long arrow); (xx) mesopleurosternum with the precoxal sulcus faintly indicated to wholly absent (unlike that observed in the two species of *Whalleyana*: Fig. 2E, arrow); (xxi) metascutellum less elongate, in posterior view (Fig. 2F, long arrow), than in *Whalleyana* (Fig. 2G, long arrow) and most moths; (xxii) fenestrae laterales very small (Fig. 2F, short arrow) (while they are well developed in both species of *Whalleyana* (Fig. 2G, short arrow) and rather large in most Thyrididae); (xxiii) in the male genitalia, juxta without 'erect arms' (a loss, as mentioned earlier: see (vi)); (xxiv) male genitalia with a short, sclerotized bridge, which unites the sacculi ventrad of the juxta (Minet, 1990: fig. 23); (xxv) female genitalia with a characteristic – flat and quadrilobate – ovipositor (Minet, 1990: figs. 27–29; see also figs. 21–25 in Holloway, 1998). Since the callidulid subfamilies Pterothysaninae and Griveaudiinae appear as sister groups based on molecular evidence, it should be noted that the male genitalia also provide a synapomorphy for these two subfamilies, namely, the presence of a few conspicuous setae in the membranous area situated just below the base of the uncus (Minet, 1990: figs. 20 and 21).

Discussion

Molecular data

Here, we present the results for low-coverage whole genome sequencing of Lepidoptera museum specimens and existing DNA extracts. With the exception of *H. puera*, the DNA extracts used for library preparation were highly fragmented. *De novo* assembly resulted in highly fragmented assemblies (N50 range of 317–2078 bp). These assemblies are consistent with assemblies obtained in other low-coverage whole genome sequencing projects, such as that of the swallowtail butterflies (Allio *et al.*, 2020) and skipper butterflies (Li *et al.*, 2019). Despite the highly fragmented nature of the resulting assemblies, the overall gene recovery rate was between 64% and 90%. Studies of bird museum specimens using Anchored Hybrid Enrichment-based approaches had recovery rates of 30–92% (Tsai *et al.*, 2019) and 49–62% (McCormack *et al.*, 2016). One advantage of sequencing genomes over target enrichment approaches is that in the future one may go back to the original data or assembly and extract new sets of genes, rather than just being limited to the genes which were enriched for. The successful library preparation and high rate of gene recovery from the *H. puera* sample highlights the usefulness of existing DNA extracts (which were originally extracted for PCR based studies) for whole genome sequencing. The ability to sequence existing extracts that are sitting around in storage from previous studies represents an important resource for expanding not only our genetic datasets but our understanding of interesting taxa and questions, which may have been previously limited due to the inability to collect fresh specimens for library construction.

We found that generating ten times more sequence data for the *Whalleyana* specimen compared to the other four specimens did not lead to a better *de novo* assembled genome, or to a higher recovery of gene regions of interest. It appears that approximately 20× coverage of a genome is enough to extract useful phylogenetic information from highly fragmented material. Lepidoptera tend to have fairly small genomes, approximately 500 Mb in size (Triant *et al.*, 2018), thus making them amenable to pooling for sequencing on the Illumina platform, with about ten genomes possible on a HiSeqX machine, or 60 genomes on the current NovaSeq machine.

We targeted 331 genes for our work, which were a set of genes that have been manually screened for orthology and alignment in a previous study (Rota *et al.*, 2021). However, given that we have sequenced the whole genomes of our specimens, we would be able to bioinformatically extract much more information from them if necessary. Assessment of the genomes for the presence of single-copy core orthologues present in the Insecta BUSCO lineage set, with a blast approach found the majority to genes were present, at least in a fragmented form. In our study, we are confident that the 331 genes have correctly placed *Whalleyana* in the Lepidoptera Tree of Life as sister to the family Callidulidae and are reasonably confident that Hyblaeidae is sister to Pyraloidea.

The morphological context

In our tree, a well-supported clade is composed of the Thyrididae, Whalleyanidae and Callidulidae (bootstrap support value: 100). The sister group to this clade is unclear,

as discussed by Rota *et al.* (2021). Based on the amino acid dataset (Fig. 1), Gelechioidea appear as the sister group of Thyrididae + Whalleyanidae + Callidulidae, or the latter is sister to Papilionoidea based on the nucleotide dataset (Fig. 1). We did not find significant morphological evidence supporting either hypothesis (although Papilionoidea share three reductions/losses with Thyrididae + Whalleyanidae + Callidulidae, namely, the above-mentioned apomorphies (i), (iii) and (iv) [see section Results, Morphological synapomorphies and autapomorphies]). All published phylogenomic analyses have been mainly based on amino acid data and have placed Callidulidae and/or Thyrididae close to Gelechioidea (Bazinet *et al.*, 2013; Kawahara & Breinholt, 2014; Kawahara *et al.*, 2019). Given the instability of the relationship of the Callidulidae/Thyrididae clade, the above-mentioned interpretation of morphological characters has taken into account the morphology of Gelechioidea and that of three other superfamilies, which have been associated with Thyrididae and/or Callidulidae in previous works (Mutanen *et al.*, 2010; Kaila *et al.*, 2013; Regier *et al.*, 2013; Wahlberg *et al.*, 2013; Heikkilä *et al.*, 2015, etc.), namely the Papilionoidea, Hyblaeoidea and Pyraloidea.

The monotypic family Prodidactidae was convincingly assigned to the superfamily Hyblaeoidea by Kaila *et al.* (2013), notably on the basis of an unusual apomorphy found in the male hindcoxa (namely, a variously developed process arising from the coxal membrane and present in both *Prodidactis* and Hyblaeidae). These authors also found a previously unnoticed larval apomorphy (modified apex of the spinneret) in the two hyblaeoid families and also in the Thyrididae. Accordingly they regarded the spinneret modification as a possible synapomorphy of Hyblaeoidea and Thyridoidea. However, they did not find clear molecular evidence supporting a sister-group relationship between these two superfamilies. It should be noted that Hyblaeidae and Thyrididae also share a possible forewing synapomorphy, namely a well-defined bunch of piliform scales arising (dorsally) from the base of vein A1 (Fig. 2H, arrow). We found this apomorphic trait in the two hyblaeid genera (*Erythrochrus*; *Hyblaea*) and in all thyridid subfamilies, but it does not exist in *Prodidactis* (A. Solis, personal communication), so it may correspond to a parallel evolution between Hyblaeidae and Thyrididae. Our molecular analyses tend to establish (like that of Heikkilä *et al.*, 2015) a well-supported sister-group relationship of Hyblaeoidea and Pyraloidea. Nevertheless, we found only two possible synapomorphies for these superfamilies, namely, the triangular shape, in lateral view, of the maxillary palps (due to the presence of a tuft of elongate scales: see for example, Janse's (1964) plate 21 for *Prodidactis*) and the closeness of the bases of M2 and M3 in the forewing venation (this apomorphy has probably arisen independently in Hyblaeoidea + Pyraloidea and Thyridoidea + Calliduloidea: cf. apomorphy (ii)). The maxillary palp apomorphy may be significant: it occurs in several groups of Pyralidae (e.g. *Synaphe* Hübner, 1825) and Crambidae (Scopariinae, Heliothelinae, Crambinae, etc.) but must have been secondarily lost in many taxa (replaced with filiform or reduced maxillary palps).

Conclusion

The results we present here show that good levels of gene recovery can be obtained from low-coverage whole genome sequencing of even highly fragmented museum samples. Our study highlights the usefulness of genome sequencing of museum specimens for which we have very little prior knowledge and lack the ability to collect fresh specimens. Additionally, we highlight that existing DNA extracts that were originally extracted for PCR are suitable for next-generation sequencing library preparation methods, and thereby represent a valuable untapped resource for expanding our datasets. This opens up possibilities for targeted taxon sampling for taxa that have been difficult to collect in recent years, often due to habitat destruction and extinction of populations. Such possibilities allow us to address vexing problems in phylogenetic studies, by giving us the possibility to strategically increase both the number of taxa and the amount of data used in analyses. The latter possibility makes the whole genome approach all the more attractive compared to genome reduction methods, as one can always return to the raw sequencing data to find more genes/regions of the genome for further analyses.

Author contributions

Niklas Wahlberg and Christopher W. Wheat conceived the study. Victoria G. Twort carried out the molecular work and processed data. Christopher W. Wheat provided advice on bioinformatic analysis. Niklas Wahlberg did the phylogenetic analysis. Joël Minet carried out dissections and morphological analysis. Victoria G. Twort, Joël Minet and Niklas Wahlberg wrote the first version of the manuscript. All authors reviewed the manuscript and contributed to the final version.

Supporting Information

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

Table S1. Library Preparation Reagents, Supplier and product code.

Table S2. Complete list of taxa included in the analysis. Genes included in the analysis for each sample are shown. X represents presence within the dataset. Gene codes correspond to those used in Rota *et al.* (2021).

Table S3. Complete summary of the 331 genes recovered for specimens sequenced in this study. Presence of the gene in the dataset is represented by an X. Gene codes corresponds to the codes used with Rota *et al.* (2021).

Acknowledgements

The authors wish to thank Nicolas Dussex for advice on library preparation protocols, Marko Mutanen for providing

the *H. puera* DNA extract, and also Alma Solis and David C. Lees for helpful information (on *Prodidactis* and *Whalleyana*, respectively). They also thank David C. Lees and an anonymous reviewer for helpful comments on a previous version of this manuscript. They acknowledge the support from the National Genomics Infrastructure in Genomics Production Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council. The computations were enabled by resources in project SNIC2018-8-345 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by Swedish Research Council through Grant Agreement No. 2018-05973 and the use of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme (URL https://www.nesi.org.nz). This work was funded Swedish Research Council (Grant No. 2015-04441). The authors declare that they have no competing interest.

Data availability statement

The laboratory protocol is available at Figshare DOI: https://doi.org/10.6084/m9.figshare.12927500. The raw genome data are registered under the BioProject PRJNA631866. Genome assemblies and orthologue alignments are available from Zenodo, DOI: https://doi.org/10.5281/zenodo.3629334.

References

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

Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data [Online]*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 15 July 2020).

Bailey, S.E., Mao, X., Struebig, 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.

Bankevich, A., Nurk, S., Antipov, D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**, 455–477.

Bazinet, A.L., Cummings, M.P., Mitter, K.T. & Mitter, C. (2013) Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. *PLoS One*, **8**, e82615.

Besnard, G., Christin, P.A., Malé, P.J.G., Lhuillier, E., Lauzeral, C., Coissac, E. & Vorontsova, M.S. (2014) From museums to genomics: old herbarium specimens shed light on a C3 to C4 transition. *Journal of Experimental Botany*, **65**, 6711–6721.

Besnard, G., Bertrand, J.A.M., Delahaie, B., Bourgeois, Y.X.C., Lhuillier, E. & Thébaud, C. (2016) Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biological Journal of the Linnean Society*, **117**, 71–82.

Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C.C. (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, **22**, 6018–6032.

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

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

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.

Call, E., Mayer, C., Twort, V., Dietz, L., Wahlberg, N. & Espeland, M. (2021) Museomics: phylogenomics of the moth family Epicopeiidae (Lepidoptera) using target enrichment. *Insect Systematics and Diversity*, **5**, 6. https://doi.org/10.1093/isd/ixaa021.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Carter, D.J. & Kristensen, N.P. (1998) Classification and keys to higher taxa. *Lepidoptera, Moths and Butterflies, Vol. 1: Evolution, Systematics, and Biogeography* (ed. by N.P. Kristensen), pp. 27–40. Walter de Gruyter, Berlin.

Chang, D., Knapp, M., Enk, J. *et al.* (2017) The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Scientific Reports*, **7**, 1–10.

Chistyakov, Y., Belyaev, E. & Omelko, M. (1994) Some peculiarities of the biology and morphology of *Pterodecta felderi* Brem. and systematic position of the family Callidulidae (Lepidoptera). *Entomological Review*, **72**, 16–27.

Common, I.F.B. (1990) *Moths of Australia*. Melbourne University Press, Carlton.

Cooper, A., Mourer-Chauvire, C., Chambers, G.K., von Haeseler, A., Wilson, A.C. & Paabo, S. (2006) Independent origins of New Zealand moas and kiwis. *Proceedings of the National Academy of Sciences*, **89**, 8741–8744.

Epstein, M.E. & Brown, J.W. (2003) Early stages of the enigmatic Prodidactis mystica (Meyrick) with comments on its new family assignment (Lepidoptera: Prodidactidae). *Zootaxa*, **247**, 1–16.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C.C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Green, R.E., Krause, J., Briggs, A.W. *et al.* (2010) A draft sequence of the neandertal genome. *Science*, **328**, 710–722.

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

Heintzman, P.D., Elias, S.A., Moore, K., Paszkiewicz, K. & Barnes, I. (2014) Characterizing DNA preservation in degraded specimens of *Amara alpina* (Carabidae: Coleoptera). *Molecular Ecology Resources*, **14**, 606–615.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2017) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, **35**, 518–522.

Hofreiter, M., Paijmans, J.L.A., Goodchild, H. *et al.* (2015) The future of ancient DNA: technical advances and conceptual shifts. *BioEssays*, **37**, 284–293.

Holloway, J.D. (1998) The moths of Borneo: families Castniidae, Callidulidae, Drepanidae and Uraniidae. *Malayan Nature Journal*, **52**, 1–155.

Houde, P. & Braun, M.J. (1988) Museum collections as a source of DNA for studies of avian phylogeny. *Auk*, **105**, 773–776.

Janse, A.J.T. (1964) Limacodidae. *The Moths of South Africa*. Pretoria: EP & Commercial Printing.

Kaila, L., Epstein, M.E., Heikkilä, M. & Mutanen, M. (2013) The assignment of Prodidactidae to Hyblaeoidea, with remarks on Thyridoidea (Lepidoptera). *Zootaxa*, **3682**, 485–494.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.

Kanda, K., Pflug, J.M., Sproul, J.S., Dasenko, M.A. & Maddison, D.R. (2015) Successful recovery of nuclear protein-coding genes from small insects in museums using Illumina sequencing. *PLoS One*, **10**, e0143929.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kawahara, A.Y. & Breinholt, J.W. (2014) Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20140970.

Kawahara, A.Y., Plotkin, D., Espeland, M. *et al.* (2019) Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences*, **116**, 22657–22663.

Li, W., Cong, Q., Shen, J., Zhang, J., Hallwachs, W., Janzen, D.H. & Grishin, N.V. (2019) Genomes of skipper butterflies reveal extensive convergence of wing patterns. *Proceedings of the National Academy of Sciences*, **116**, 6232–6237.

Maddison, D.R. & Cooper, K.W. (2014) Species delimitation in the ground beetle subgenus Liocosmius (Coleoptera: Carabidae: Bembidion), including standard and next-generation sequencing of museum specimens. *Zoological Journal of the Linnean Society*, **172**, 741–770.

Mayer, C., Dietz, L., Call, E., Kukowka, S., Martin, S. & Espeland, M. (2021) Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens. *Systematic Entomology*, **46**, 649–671.

McCormack, J.E., Tsai, W.L.E. & Faircloth, B.C. (2016) Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, **16**, 1189–1203.

Meyer, M. & Kircher, M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, **2010**. https://doi.org/10.1101/pdb.prot5448.

Meyer, M., Kircher, M., Gansauge, M.T. *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science*, **338**, 222–226.

Miller, M.A., Pfeiffer, W., & Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop (GCE)*. 2010, pp. 1–8.

Minet, J. (1983) Etude morphologique et phylogénétique des organes tympaniques des Pyraloidea. 1. Généralités et homologies (Lep. Glossata). *Annales De La Societe Entomologique De France*, **19**, 175–207.

Minet, J. (1990) Nouvelles frontières, géographiques et taxonomiques, pour la famille des Callidulidae (Lepidoptera, Calliduloidea). *Nouvelle revue d'entomologie*, **6**, 351–368.

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

Minet, J. (1998) Chapter 15. The Axioidea and Calliduloidea. *Lepidoptera, Moths and Butterflies, Vol. 1: Evolution, Systematics, and Biogeography* (ed. by N.P. Kristensen), pp. 257–261. Walter de Gruyter, Berlin.

Mosher, E. (1916) A classification of the Lepidoptera based on characters of the pupa. *Bulletin of the Illinois State Laboratory of Natural History*, **12**, 14–159.

Mutanen, M., Wahlberg, N. & Kaila, L. (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 2839–2848.

Nakamura, M. (2011) Pupae of Japanese Callidulidae (Lepidoptera). *Transactions of the Lepidopterological Society of Japan*, **62**, 98–101.

Neethiraj, R., Hornett, E.A., Hill, J.A. & Wheat, C.W. (2017) Investigating the genomic basis of discrete phenotypes using a Pool-Seq-only approach: new insights into the genetics underlying colour variation in diverse taxa. *Molecular Ecology*, **26**, 4990–5002.

Neunzing, H. (1987) Pyralidae (Pyraloidea). *Immature Insects* (ed. by F. Stehr), pp. 462–494. Kendall/Hunt Publishing Company, Dubuque, Iowa.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2014) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.

Paijmans, J.L.A., Gilbert, M.T.P. & Hofreiter, M. (2013) Mitogenomic analyses from ancient DNA. *Molecular Phylogenetics and Evolution*, **69**, 404–416.

Peña, C. & Malm, T. (2012) VoSeq: a voucher and DNA sequence web application. *PLoS One*, **7**, e39071.

Poinar, H., Schwarz, C., Qi, J. *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.

Prosser, S.W.J., Dewaard, J.R., Miller, S.E. & Hebert, P.D.N. (2016) DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources*, **16**, 487–497.

Rajaei, H., Greve, C., Letsch, H., Stüning, D., Wahlberg, N., Minet, J. & Misof, B. (2015) Advances in Geometroidea phylogeny, with characterization of a new family based on *Pseudobiston pinratanai* (Lepidoptera, Glossata). *Zoologica Scripta*, **44**, 418–436.

Regier, J.C., Shultz, J.W., Zwick, A. *et al.* (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, **463**, 1079–1083.

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**, e58568.

Rizzi, E., Lari, M., Gigli, E., De Bellis, G. & Caramelli, D. (2012) Ancient DNA studies: new perspectives on old samples. *Genetics, Selection, Evolution*, **44**, 21.

Robinson, G.S. (1976) The preparation of slides of Lepidoptera genitalia with special reference to the Microlepidoptera. *Entomologist's Gazette*, **27**, 127–132.

Rota, J., Twort, V., Chiocchio, A., Peña, C., Wheat, C.W., Kaila, L. & Wahlberg, N. (2021) The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): assessing the potential causes and consequences. *bioRxiv*. https://doi.org/10.1101/2021.04.09.439156.

Rowe, K.C., Singhal, S., Macmanes, M.D. *et al.* (2011) Museum genomics: low-cost and high-accuracy genetic data from historical specimens. *Molecular Ecology Resources*, **11**, 1082–1092.

Schmieder, R. & Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Shaffer, H.B., Fisher, R.N. & Davidson, C. (1998) The role of natural history collections in documenting species declines. *Trends in Ecology & Evolution*, **13**, 27–30.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Soltis, P.S. & Soltis, D.E. (1993) Ancient DNA: prospects and limitations. *New Zealand Journal of Botany*, **31**, 203–209.

Sproul, J.S. & Maddison, D.R. (2017) Sequencing historical specimens: successful preparation of small specimens with low amounts of degraded DNA. *Molecular Ecology Resources*, **17**, 1183–1201.

Staats, M., Erkens, R.H.J., van de Vossenberg, B. *et al.* (2013) Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS One*, **8**, e69189.

Suarez, A.V. & Tsutsui, N.D. (2004) The value of museum collections for research and society. *Bioscience*, **54**, 66.

Thessen, A.E., Patterson, D.J. & Murray, S.A. (2012) The taxonomic significance of species that have only been observed once: the genus *Gymnodinium* (Dinoflagellata) as an example. *PLoS One*, **7**, e44015.

Thomas, W.K., Pääbo, S., Villablanca, F.X. & Wilson, A.C. (1990) Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens. *Journal of Molecular Evolution*, **31**, 101–112.

Tin, M.M.Y., Economo, E.P. & Mikheyev, A.S. (2014) Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. *PLoS One*, **9**, e96793.

Triant, D.A., Cinel, S.D. & Kawahara, A.Y. (2018) Lepidoptera genomes: current knowledge, gaps and future directions. *Current Opinion in Insect Science*, **25**, 99–105.

Tsai, W.L.E., Mota-Vargas, C., Rojas-Soto, O. *et al.* (2019) Museum genomics reveals the speciation history of *Dendrortyx* wood-partridges in the Mesoamerican highlands. *Molecular Phylogenetics and Evolution*, **136**, 29–34.

Viette, P. (1977) Un nouveau genre et deux espèces nouvelles de Lépidoptères Thyrididae malgaches. *Bulletin Mensuel de la Société Linnéenne de Lyon*, **46**, 246–250.

Wahlberg, N., Wheat, C.W. & Peña, C. (2013) Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). *PLoS One*, **8**, e80875.

Wandeler, P., Hoeck, P.E.A. & Keller, L.F. (2007) Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution*, **22**, 634–642.

Wang, H.Y. (1993) *Illustrations of Day-Flying Moths in Taiwan*. Taiwan Museum, Taipei.

Wiley, A.E., Ostrom, P.H., Welch, A.J. *et al.* (2013) Millennial-scale isotope records from a wide-ranging predator show evidence of recent human impact to oceanic food webs. *Proceedings of the National Academy of Sciences*, **110**, 8972–8977.

Willerslev, E. & Cooper, A. (2005) Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, **272**, 3–16.

Yeates, D.K., Zwick, A. & Mikheyev, A.S. (2016) Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections. *Current Opinion in Insect Science*, **18**, 83–88.

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 16 June 2021
First published online 8 July 2021