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

Victoria G. Twort*[1,4], Joël Minet[2], Christopher W. Wheat[3] and Niklas Wahlberg*[1]

[1] Department of Biology, Lund University, Lund, Sweden

[2] Muséum National d'Histoire Naturelle, ISYEB, Paris, France

[3] Department of Zoology, Stockholm University, Stockholm, Sweden

[4] Current Address: The Finnish Museum of Natural History Luomus, Zoology Unit, University of Helsinki, Helsinki, Finland

*Corresponding authors: vtwort@gmail.com; niklas.wahlberg@biol.lu.se

Data availability

The raw genome data are registered under the BioProject PRJNA631866. Genome assemblies are available from Zenodo, DOI: 10.5281/zenodo.3629334

Abstract

Museomics is a valuable tool that utilises 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*, as well as the families Callidulidae and Hyblaeidae. 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 – 2,078 bp. Mining of a manually curated gene set of 332 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 (Yeates et al. 2016; Graham et al. 2004; Suarez & Tsutsui 2004; Shaffer et al. 1998). Many species present in these collections are more accessible than in their original habitats due to a variety of factors (Thessen et al. 2012; Wandeler et al. 2007), such as remote geographical distributions, rare or endangered taxa, taxa that have since gone extinct, or taxa that 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 has expanded their role into the realm of genetics, and many other fields (Paijmans et al. 2013; Wiley et al. 2013; Shaffer et al. 1998). 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 material 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 (for example: Bi et al., 2013; Besnard et al., 2014, 2016; Chang et al., 2017). Initial studies focused on using PCR and Sanger sequencing of short fragments of genes (Cooper et al. 2006; Thomas et al. 1990; Houde & Braun 1988), 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 (Wandeler et al. 2007; Soltis & Soltis 1993).

The development of high-throughput sequencing (HTS) technologies offers a promise of more efficient ways of sequencing DNA from museum specimens (Hofreiter et al. 2015; Rizzi et al. 2012). 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 (Meyer et al. 2012; Green et al. 2010) showed the promise of HTS technologies and 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 (UCEs), 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 (for example: Bi et al., 2013; Staats et al., 2013; Bailey et al., 2016; Blaimer, Lloyd, Guillory, & Brady, 2016; Prosser, Dewaard, Miller, & Hebert, 2016), 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.

An alternative approach to targeted sequence capture is whole-genome sequencing. Although initial studies focused on taxa for which a reference genome, or one closely related, already existed (for example: 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 HTS based approaches on old insect specimens (Kanda et al. 2015; Heintzman et al. 2014; Maddison & Cooper 2014; Tin et al. 2014; Staats et al. 2013). Initial studies focused on recovering specific regions, such as mitochondrial or ribosomal DNA, which are present in multiple copies per cell (Heintzman et al. 2014; Staats et al. 2013). 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 highlighted 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 and Maddison (2017) presented a non-destructive 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 combined 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 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). Very little is known about the two species (*Whalleyana vroni* Viette, 1977 and *W. toni* Viette, 1977) that make up this genus, including their phylogenetic placement within Lepidoptera. In addition to the lack of knowledge, neither species has been collected since the 1990's (David C. Lees, pers. comm. to JM). 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 does *Whalleyana* belong. We also sequence museum specimens or existing genomic DNA extracts of potentially related taxa belonging to the families Callidulidae and Hyblaeidae. 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 mainly from the adult stage. Indeed, 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: *Whalleyana vroni* (collected in 1969), *Helicomitra pulchra* Butler, 1878 (collected in 1974), *Griveaudia vieui* Viette, 1958 (collected in 1969), *Hyblaea madagascariensis* Viette, 1961 (collected in 1975), as well as a DNA extract of *Hyblaea puera* (Cramer, 1777) from Mutanen et al. (2010). *Helicomitra* and *Griveaudia* belong to the family Callidulidae, and *Hyblaea* to Hyblaeidae. Both families are potentially related to *Whalleyana*, and the latter has no genomic resources available 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 minutes 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 approx. 200 – 300 bp fragments using a Bioryptor® with the following settings: (M) medium power output, 30 sec ON/ 90 sec OFF pulses for 30 minutes 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 and 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: 1x 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: 1x 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 was 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 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 0.5X was used to remove long fragments that are likely to represent contamination from fresh DNA, libraries were selected with a bead concentration of 1.8X to size select the expected library range of 100 – 300 bp. The resulting libraries were quantified and quality checked with Quanti-iT^{TM} PicoGreen^{TM} 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 9 samples and sequenced over two runs, while the remaining specimens were pooled in equimolar concentrations in a pool of 6 samples and sequenced using the Illumina HiSeq2500 technology with 150 bp paired-end reads.

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 evaluation read quality with a sliding window approach. Quality was measured for sliding windows of 4 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), with a custom set of 332 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 200 taxa of Lepidoptera. The details of the vetting process are described in Rota et al. (in prep.). The resulting DNA sequences were aligned to pre-existing reference alignments (taken from Rota et al. in prep.) based on 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, screen 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: 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. in prep). The final dataset consisted of a total of 338 gene fragments, spanning 332 genes, across 169 taxa (164 taxa were taken from Rota et al. (in prep.), See Table S2 for 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 (Zwick et al. 2012; Regier et al. 2010). 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 dissected by one of us and belonging to the MNHN. These specimens had been chosen to represent most ditrysian families within the framework of several previous publications (e.g. Minet, 1991 and Rajaei et al., 2015). Among the families which are more precisely the focus of the present study, specimens that have been entirely dissected belong to the following genera: *Striglina* Guenée, 1877 (Thyrididae: Striglininae), *Marmax* Rafinesque, 1815 (Thyrididae: Charideinae), *Thyris* Laspeyres, 1803 (Thyridinae), *Chrysotypus* Butler, 1879 (Thyrididae: Siculodinae), *Rhodoneura* Guenée, 1858 (same subfamily), *Whalleyana* Viette, 1977 (Whalleyanidae), *Helicomitra* Butler, 1878 (Callidulidae: Pterothysaninae), *Griveaudia* Viette, 1958 (Callidulidae: Griveaudiinae), *Callidula* Hübner, 1819 (Callidulinae), and *Hyblaea* Fabricius, 1794 (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 X7.5 to X128. 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 (Kaila et al. 2013; Epstein & Brown 2003).

Results

a) The molecular approach

DNA extraction of the four museum specimens were successful with DNA fragments ranging from 70 - 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. 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 – 2,078 bp (Table 1). Assessment of the completeness of the resulting assemblies with BUSCO, highlighted the difficulty the program has in finding ortholgues in fragmented genomes (results not shown). However, the ortholgue set can still be used with a BLAST approach to assess presence of the conserved genes. The blast search for the 1, insecta orthologues showed 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% - 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 these relationships have 100% UFbootstrap support regardless of data form, except for nt12, where the sister relationship of *Hyblaea* and Pyraloidea receives only 86%.

b) 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 + Calliduloidea lineage.

Thyrididae, Whalleyanidae and Callidulidae share six imaginal synapomorphies: (1) on the head, the ocellus is either absent or devoid of a distinct lens (as also in all Papilionoidea and Hyblaeoidea); (2) 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, 1880 and *Donacostola* Meyrick, 1931); (3) 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, 1793, which retains a large spinarea (Common, 1990: Fig. 108), and in many Gelechioidea and Pyraloidea); (4) 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, 1820); 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*; (5) 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, 1921, only the base of Sc is approximated to the upper edge of the hindwing discal cell); (6) 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 e.g. Fig. 7 in Chistyakov et al. (1994)). While 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 (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: (7) 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, 1832), although often less distinctly; (8) on vertex, the chaetosemata are large and include minute scales between their setae; (9) 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)); (10) in the forewing, veins Rs3 and Rs4 run to the termen, reaching it below the apex (Minet 1998: Fig. 15.1, B and C; Viette 1977: Fig. 1) (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 apomorphy (10): (Common 1990:Fig. 109.1)); (11) 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); (12) 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, 1866: Common 1990: Fig. 109.4) and arises slightly closer to M1 than to M3 in the hyblaeid genus *Erythrochrus* Herrich-Schäffer, 1858); (13) 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 just variously developed anterolateral processes that do not reach the marginotergites); (14) 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: e.g. Fig. 12 in Minet (1983)); (15) the male genitalia lack a complete gnathos (retained in many Thyrididae); (16) 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 (7), (8), (11), and (16) 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 apomorphies (7)-(16)).

Within the thus redefined Calliduloidea, nine autapomorphies support the monophyly of the family Callidulidae (= Pterothysaninae + Griveaudiinae + Callidulinae), namely: (17) 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); (18) male forewing without a subcostal retinaculum (while *Whalleyana vroni* retains this retinaculum (Fig. 2D, short arrow); through parallel evolution, the male forewing of *Whalleyana toni* has lost this structure); (19) 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); (20) mesopleurosternum with the precoxal sulcus faintly indicated to wholly absent (unlike that observed in the two species of *Whalleyana*: Fig. 2E, arrow); (21) metascutellum less elongate, in posterior view (Fig. 2F, long arrow), than in *Whalleyana* (Fig. 2G, long arrow) and most moths; (22) 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); (23) in the male genitalia, juxta without “erect arms” (a loss, as mentioned above: see (6)); (24) male genitalia with a short, sclerotized bridge, which unites the sacculi ventrad of the juxta (Minet 1990: Fig. 23); (25) female genitalia with a characteristic – flat and quadrilobate – ovipositor (see also Figs 21-25 in Holloway 1998; Minet 1990: Figs 27-29). Since Pterothysaninae and Griveaudiinae appear as sister groups 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

a) 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 – 2,078 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 AHE 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 sequencing 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 10 times more sequence data for the *Whalleyana* specimen compared to the other four specimens did not lead to better *de novo* assembled genomes, or to higher recovery of gene regions of interest. It appears that approximately 20X 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 10 genomes possible on a HiSeqX machine, or 60 genomes on the current NovaSeq machine.

We targeted 332 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. in prep). 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 332 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.

b) 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. (in prep.). Based on the amino acid dataset, Gelechioidea appear as the sister group of Thyrididae + Whalleyanidae + Callidulidae, or the latter is sister to Papilionoidea based on the nucleotide dataset. We did not find significant morphological evidence supporting either hypothesis (although Papilionoidea share three reductions/losses with Thyrididae + Whalleyanidae + Callidulidae, viz. the above-mentioned apomorphies (1), (3) and (4) (see section Results, b). 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 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 (viz. 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 but 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* (Alma Solis, pers. comm.) so that it may correspond to a parallel evolution between Hyblaeidae and Thyrididae. Our molecular analysis 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 e.g.

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 (2)). 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).

c) Conclusion

In 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 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.

Acknowledgements

We would like 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). We 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, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure and the use of New Zealand eScience Infastructure (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 by the Swedish Research Council (grant no. 2015-04441)

References

Allio R et al. 2020. Whole Genome Shotgun Phylogenomics Resolves the Pattern and Timing of Swallowtail Butterfly Evolution. Syst. Biol. 69:38–60. doi: 10.1093/sysbio/syz030.

Andrews S. 2010. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Available at http://www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. doi: citeulike-article-id:11583827.

Bailey SE et al. 2016. The use of museum samples for large-scale sequence capture: A study of congeneric horseshoe bats (family Rhinolophidae). Biol. J. Linn. Soc. 117:58–70. doi: 10.1111/bij.12620.

Bankevich A et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455–77. doi: 10.1089/cmb.2012.0021.

Bazinet AL, Cummings MP, Mitter KT, 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. https://doi.org/10.1371/journal.pone.0082615.

Besnard G et al. 2014. From museums to genomics: Old herbarium specimens shed light on a C3 to C4 transition. J. Exp. Bot. 65:6711–6721. doi: 10.1093/jxb/eru395.

Besnard G et al. 2016. Valuing museum specimens: High-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). Biol. J. Linn. Soc. 117:71–82. doi: 10.1111/bij.12494.

Bi K et al. 2013. Unlocking the vault: Next-generation museum population genomics. Mol. Ecol. 22:6018–6032. doi: 10.1111/mec.12516.

Blaimer BB, Lloyd MW, Guillory WX, Brady SG. 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS One. 11:e0161531. doi: 10.1371/journal.pone.0161531.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120. doi: 10.1093/bioinformatics/btu170.

Brock JP. 1971. A contribution towards an understanding of the morphology and phylogeny of the ditrysian Lepidoptera. J. Nat. Hist. 5:29–102.

Camacho C et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421. doi: 10.1186/1471-2105-10-421.

Chang D et al. 2017. The evolutionary and phylogeographic history of woolly mammoths: A comprehensive mitogenomic analysis. Sci. Rep. 7:1–10. doi: 10.1038/srep44585.

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). Entomol. Rev. 72:16–27.

Common IFB. 1990. *Moths of Australia*. Melbourne University Press, Carlton.

Cooper A et al. 2006. Independent origins of New Zealand moas and kiwis. Proc. Natl. Acad. Sci. 89:8741–8744. doi: 10.1073/pnas.89.18.8741.

Epstein ME, Brown JW. 2003. Early stages of the enigmatic Prodidactis mystica (Meyrick) with comments on its new family assignment (Lepidoptera: Prodidactidae). Zootaxa. 247:1–16.

Graham CH, Ferrier S, Huettman F, Moritz CC, Peterson AT. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19:497–503. doi: 10.1016/j.tree.2004.07.006.

Green RE et al. 2010. A draft sequence of the neandertal genome. Science (80-.). 328:710–722. doi: 10.1126/science.1188021.

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 Evol. Biol. 15:260. doi: 10.1186/s12862-015-0520-0.

Heintzman PD, Elias SA, Moore K, Paszkiewicz K, Barnes I. 2014. Characterizing DNA preservation in degraded specimens of Amara alpina (Carabidae: Coleoptera). Mol. Ecol. Resour. 14:606–615. doi: 10.1111/1755-0998.12205.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518–522. doi: 10.1093/molbev/msx281.

Hofreiter M et al. 2015. The future of ancient DNA: Technical advances and conceptual shifts. BioEssays. 37:284–293. doi: 10.1002/bies.201400160.

Holloway JD. 1998. The moths of Borneo: families Castniidae, Callidulidae, Drepanidae and Uraniidae. Malayan Nat. J. 52:1–155.

Houde P, Braun MJ. 1988. Museum collections as a source of DNA for studies of avian phylogeny. Auk. 105:773–776. http://www.jstor.org/stable/4087394.

Janse AJT. 1964. *Limacodidae. The Moths of South Africa*. EP & Commercial Printing.

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

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587. https://doi.org/10.1038/nmeth.4285.

Kanda K, Pflug JM, Sproul JS, Dasenko MA, Maddison DR. 2015. Successful recovery of nuclear protein-coding genes from small insects in museums using Illumina sequencing. PLoS One. 10:e0143929. doi: 10.1371/journal.pone.0143929.

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 Res. 30:3059–66. doi: 10.1093/nar/gkf436.

Kawahara AY et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proc. Natl. Acad. Sci. 116:22657 LP – 22663. doi: 10.1073/pnas.1907847116.

Kawahara AY, Breinholt JW. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. Proc. R. Soc. B Biol. Sci. 281:20140970. doi: 10.1098/rspb.2014.0970.

Li W et al. 2019. Genomes of skipper butterflies reveal extensive convergence of wing patterns. Proc. Natl. Acad. Sci. 116:6232 LP – 6237. doi: 10.1073/pnas.1821304116.

Maddison DR, Cooper KW. 2014. Species delimitation in the ground beetle subgenus Liocosmius (Coleoptera: Carabidae: Bembidion), including standard and next-generation sequencing of museum specimens. Zool. J. Linn. Soc. 172:741–770. doi: 10.1111/zoj.12188.

McCormack JE, Tsai WLE, Faircloth BC. 2016. Sequence capture of ultraconserved elements from bird museum specimens. Mol. Ecol. Resour. 16:1189–1203. doi: 10.1111/1755-0998.12466.

Meyer M et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science (80-.). 338:222–226. doi: 10.1126/science.1224344.

Meyer M, Kircher M. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc. . 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gatew. Comput. Environ. Work. 1–8.

Minet J. 1998. Chapter 15. The Axioidea and Calliduloidea. In: Lepidoptera, moths and butterflies, vol. 1:

evolution, systematics, and biogeography. Kristensen, NP, editor. Walter de Gruyter, Berlin pp. 257–261.

Minet J. 1983. Etude morphologique et phylogénétique des organes tympaniques des Pyraloidea. 1. Généralités et homologies (Lep. Glossata). Ann. la Société Entomol. Fr. (NS) 19(2):175–207.

Minet J. 1990. Nouvelles frontières, géographiques et taxonomiques, pour la famille des Callidulidae (Lepidoptera, Calliduloidea). Nouv. Rev. d'Entomologie. 6(4):351–368.

Minet J. 1991. Tentative reconstruction of the ditrysian phylogeny (Lepidoptera: Glossata). Insect Syst. Evol. 22:69–95.

Mosher E. 1916. A classification of the Lepidoptera based on characters of the pupa. Bull. Illinois State Lab. Nat. Hist. 12:14–159.

Mutanen M, Wahlberg N, Kaila L. 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc. R. Soc. B Biol. Sci. 277:2839–2848. doi: 10.1098/rspb.2010.0392.

Nakamura M. 2011. Pupae of Japanese Callidulidae (Lepidoptera). Trans. Lepidopterol. Soc. Japan. 62:98–101. doi: 10.18984/lepid.62.2_98.

Neethiraj R, Hornett EA, Hill JA, Wheat CW. 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. Mol. Ecol. 26:4990–5002. doi: 10.1111/mec.14205.

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

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32:268–274. doi: 10.1093/molbev/msu300.

Paijmans JLA, Gilbert MTP, Hofreiter M. 2013. Mitogenomic analyses from ancient DNA. Mol. Phylogenet. Evol. 69:404–416. doi: 10.1016/j.ympev.2012.06.002.

Peña C, Malm T. 2012. VoSeq: A voucher and DNA sequence web application. PLoS One. 7:e39071–e39071. doi: 10.1371/journal.pone.0039071.

Poinar H et al. 2006. Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. Science (80-.). 311:392–394. doi: 10.1126/science.1123360.

Prosser SWJ, Dewaard JR, Miller SE, Hebert PDN. 2016. DNA barcodes from century-old type specimens using next-generation sequencing. Mol. Ecol. Resour. 16:487–497. doi: 10.1111/1755-0998.12474.

Rajaei H et al. 2015. Advances in Geometroidea phylogeny, with characterization of a new family based on Pseudobiston pinratanai (Lepidoptera, Glossata). Zool. Scr. 44:418–436. doi: 10.1111/zsc.12108.

Regier JC et al. 2013. A Large-Scale, Higher-Level, Molecular Phylogenetic Study of the Insect Order Lepidoptera (Moths and Butterflies). PLoS One. 8:e58568. https://doi.org/10.1371/journal.pone.0058568.

Regier JC et al. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature. 463:1079. https://doi.org/10.1038/nature08742.

Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D. 2012. Ancient DNA studies: New perspectives on old samples. Genet. Sel. Evol. 44:21. doi: 10.1186/1297-9686-44-21.

Robinson GS. 1976. The preparation of slides of Lepidoptera genitalia with special reference to the Microlepidoptera. Entomol. Gaz. 27(2). 127–132.

Rowe KC et al. 2011. Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. Mol. Ecol. Resour. 11:1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27:863–864. doi: 10.1093/bioinformatics/btr026.

Shaffer HB, Fisher RN, Davidson C. 1998. The role of natural history collections in documenting species declines. Trends Ecol. Evol. 13:27–30. doi: 10.1016/S0169-5347(97)01177-4.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

Soltis PS, Soltis DE. 1993. Ancient DNA: Prospects and limitations. New Zeal. J. Bot. 31:203–209. doi: 10.1080/0028825X.1993.10419497.

Sproul JS, Maddison DR. 2017. Sequencing historical specimens: successful preparation of small specimens with low amounts of degraded DNA. Mol. Ecol. Resour. 17:1183–1201. doi: 10.1111/1755-0998.12660.

Staats M et al. 2013. Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. PLoS One. 8:e69189. doi: 10.1371/journal.pone.0069189.

Suarez A V., Tsutsui ND. 2004. The Value of Museum Collections for Research and Society. Bioscience. 54:66.

doi: 10.1641/0006-3568(2004)054[0066:tvomcf]2.0.co;2.

Thessen AE, Patterson DJ, Murray SA. 2012. The Taxonomic Significance of Species That Have Only Been Observed Once: The Genus Gymnodinium (Dinoflagellata) as an Example. PLoS One. 7:e44015. doi: 10.1371/journal.pone.0044015.

Thomas WK, Pääbo S, Villablanca FX, Wilson AC. 1990. Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens. J. Mol. Evol. 31:101–112. doi: 10.1007/BF02109479.

Tin MMY, Economo EP, Mikheyev AS. 2014. Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. PLoS One. 9:e96793. doi: 10.1371/journal.pone.0096793.

Triant DA, Cinel SD, Kawahara AY. 2018. Lepidoptera genomes: current knowledge, gaps and future directions. Curr. Opin. insect Sci. 25:99—105. doi: 10.1016/j.cois.2017.12.004.

Tsai WLE et al. 2019. Museum genomics reveals the speciation history of Dendrortyx wood-partridges in the Mesoamerican highlands. Mol. Phylogenet. Evol. 136:29–34. doi: 10.1016/j.ympev.2019.03.017.

Viette P. 1977. Un nouveau genre et deux espèces nouvelles de Lépidoptères Thyrididae malgaches. Bull. Mens. la Société linnéenne Lyon. 46(7):246–250.

Wahlberg N, Wheat CW, Peña C. 2013. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). PLoS One. 8:e80875. doi: 10.1371/journal.pone.0080875.

Wandeler P, Hoeck PEA, Keller LF. 2007. Back to the future: museum specimens in population genetics. Trends Ecol. Evol. 22:634–642. doi: 10.1016/j.tree.2007.08.017.

Wang HY. 1993. *Illustrations of day-flying moths in Taiwan*. Taiwan Museum, Taipei.

Wiley AE et al. 2013. Millennial-scale isotope records from a wide-ranging predator show evidence of recent human impact to oceanic food webs. Proc. Natl. Acad. Sci. 110:8972–8977. doi: 10.1073/pnas.1300213110.

Willerslev E, Cooper A. 2005. Ancient DNA. Proc. R. Soc. Lond., B, Biol. Sci. 272:3–16.

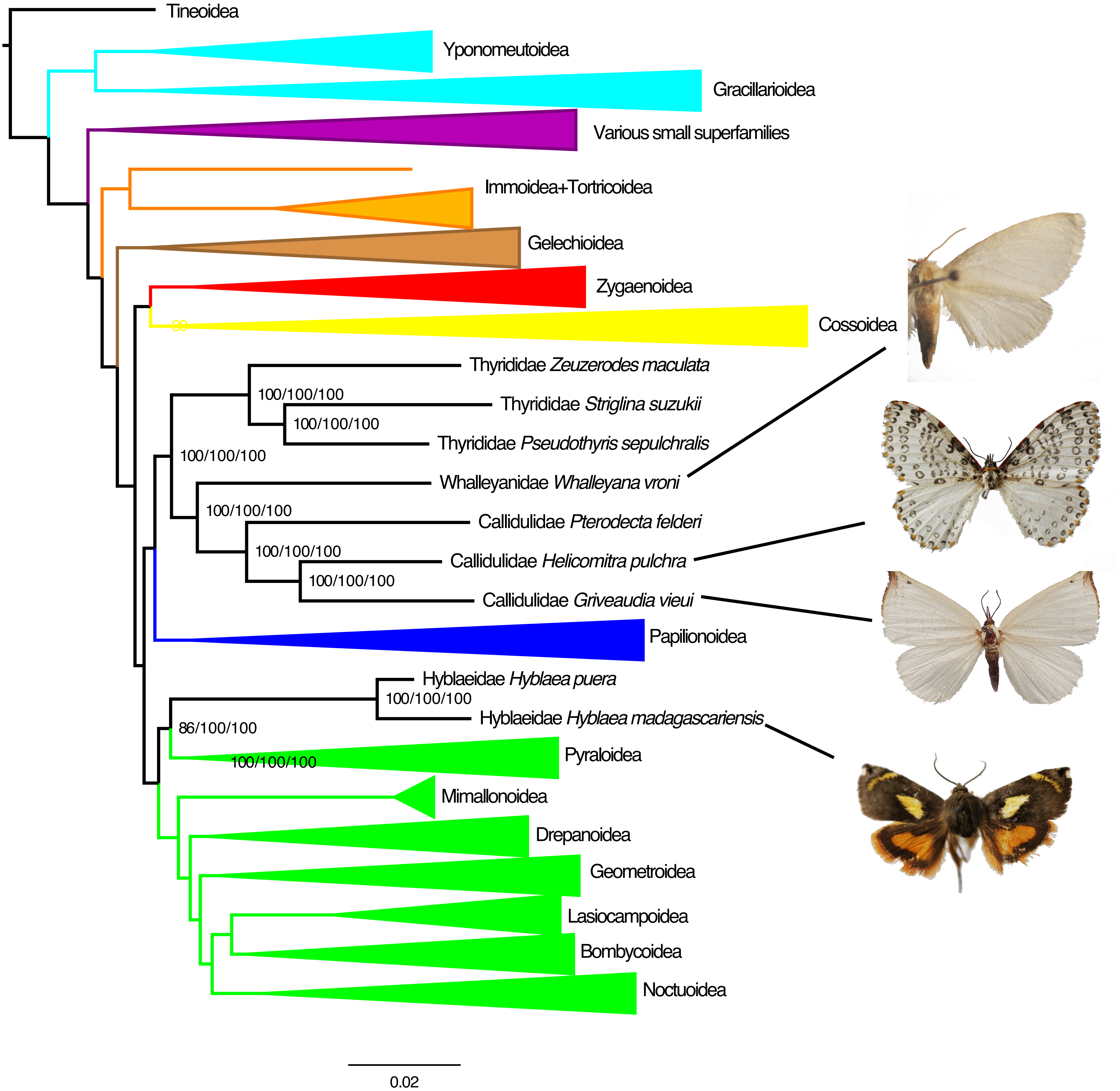
Yeates DK, Zwick A, Mikheyev AS. 2016. Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world’s biological collections. Curr. Opin. Insect Sci. 18:83–88. doi: 10.1016/j.cois.2016.09.009.

Zwick A, Regier JC, Zwickl DJ. 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. https://doi.org/10.1371/journal.pone.0047450.

Figure captions

Figure 1. Phylogenetic relationships of *Whalleyana*, *Helicomitra*, *Griveaudia* and *Hyblaea* based on 332 genes. Superfamilies that are not relevant to these taxa are shown as collapsed. Numbers to the right of each node give the UltraFastBootstraps for each dataset analysed: nt12/degen1/aa.

Figure 2. A. *Whalleyana vroni*, male genitalia (arrow: one of the two free, dorsally directed, arms of the juxta). B. *Whalleyana 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. *Whalleyana 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. *Whalleyana 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. *Whalleyana 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).


Tineoidea
Yponomeutoidea
Gracillarioidea
Various small superfamilies
Immoidea+Tortricoidea
Gelechioidea
Zygaenoidea
88
Cossoidea
Thyrididae *Zeuzerodes maculata*
100/100/100
Thyrididae *Striglina suzukii*
100/100/100
Thyrididae *Pseudothyris sepulchralis*
100/100/100
Whalleyanidae *Whalleyana vroni*
100/100/100
Callidulidae *Pterodecta felderi*
100/100/100
Callidulidae *Helicomitra pulchra*
100/100/100
Callidulidae *Griveaudia vieui*
Papilionoidea
Hyblaeidae *Hyblaea puera*
100/100/100
Hyblaeidae *Hyblaea madagascariensis*
86/100/100
Pyraloidea
100/100/100
Mimallonoidea
Drepanoidea
Geometroidea
Lasiocampoidea
Bombycoidea
Noctuoidea
0.02

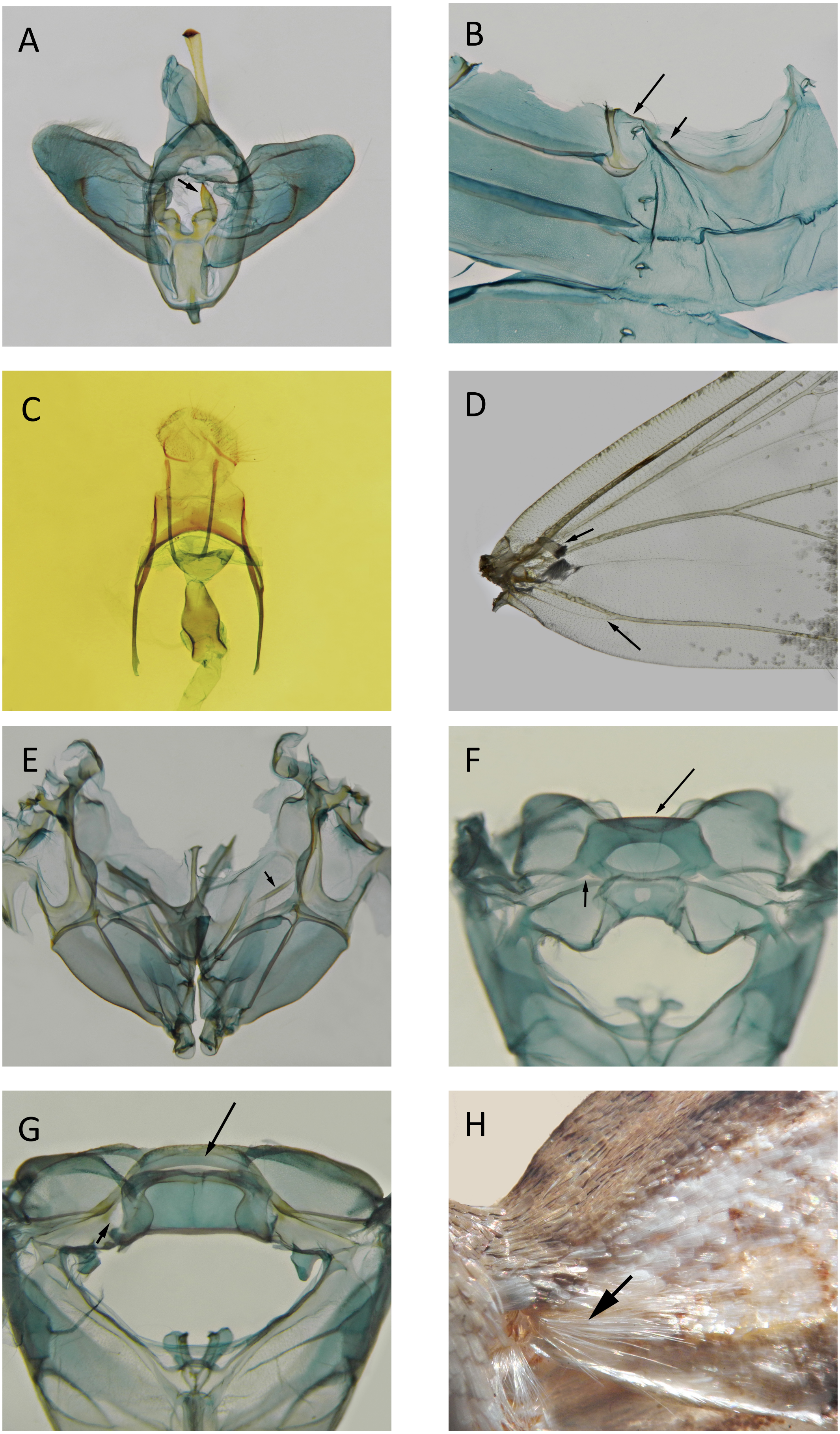

A
B
C
D
E
F
G
H

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	---	---	---	---	23863949
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	5,413	4,792	5,441	7,920	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	264.0	291.0	250.0	209.0	108.0
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	1,429	6,611	8,871	65,308	84,638
Contigs >= 10 Kbp	---	---	---	---	2,639
N50 value	317	367	361	478	2,078

Numer of Genes Recovered (/338)	215 (63%)	244 (72%)	244 (72%)	255 (75%)	304 (90%)
Number of BUSCO Insecta Genes Identified (/1,658)	1,201 (72%)	1,415 (85%)	1,396 (84%)	1,435 (87%)	1,400 (84%)