PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Kawahara AY, Breinholt JW. 2014 Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proc. R. Soc. B* **281**: 20140970.
http://dx.doi.org/10.1098/rspb.2014.0970

Received: 25 April 2014
Accepted: 4 June 2014

Subject Areas:
evolution, taxonomy and systematics

Keywords:
butterfly, Lepidoptera, moth, orthologue, phylogeny, transcriptome

Authors for correspondence:
Akito Y. Kawahara
e-mail: kawahara@flmnh.ufl.edu
Jesse W. Breinholt
e-mail: jessebreinholt@gmail.com

†These authors contributed equally to this study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2014.0970 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing

Phylogenomics provides strong evidence for relationships of butterflies and moths

Akito Y. Kawahara† and Jesse W. Breinholt†

Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

Butterflies and moths constitute some of the most popular and charismatic insects. Lepidoptera include approximately 160 000 described species, many of which are important model organisms. Previous studies on the evolution of Lepidoptera did not confidently place butterflies, and many relationships among superfamilies in the megadiverse clade Ditrysia remain largely uncertain. We generated a molecular dataset with 46 taxa, combining 33 new transcriptomes with 13 available genomes, transcriptomes and expressed sequence tags (ESTs). Using HaMStR with a Lepidoptera-specific core-orthologue set of single copy loci, we identified 2696 genes for inclusion into the phylogenomic analysis. Nucleotides and amino acids of the all-gene, all-taxon dataset yielded nearly identical, well-supported trees. Monophyly of butterflies (Papilionoidea) was strongly supported, and the group included skippers (Hesperiidae) and the enigmatic butterfly–moths (Hedylidae). Butterflies were placed sister to the remaining obtectomeran Lepidoptera, and the latter was grouped with greater than or equal to 87% bootstrap support. Establishing confident relationships among the four most diverse macroheteroceran superfamilies was previously challenging, but we recovered 100% bootstrap support for the following relationships: ((Geometroidea, Noctuoidea), (Bombycoidea, Lasiocampoidea)). We present the first robust, transcriptome-based tree of Lepidoptera that strongly contradicts historical placement of butterflies, and provide an evolutionary framework for genomic, developmental and ecological studies on this diverse insect order.

1. Introduction

Butterflies and moths comprise the order Lepidoptera, which is one of the four insect super-radiations and includes approximately 160 000 described species, though the actual number of species might be as high as half a million [1,2]. These insects dominate the terrestrial landscape as butterflies during the day and moths at night [1]. Lepidoptera are predominantly herbivores and pollinators, and they are thought to have played a central role in the mega-radiation of angiosperms [1,3,4]. Their association with plants has led to numerous textbook examples of coevolution (reviewed in [5]). The feeding apparatus of Lepidoptera is thought to have transitioned from a mandibulate condition within the non-glossatan Lepidoptera to a coiled proboscis that is used to imbibe nectar from flowers in derived lineages such as the Ditrysia, a group that comprises more than 98% of species in the order [6]. Lepidoptera also include many model organisms as well as some of the most damaging agricultural pests (e.g. *Bombyx*, *Cydia*, *Helicoverpa*, *Manduca*, *Spodoptera* [7]). Although many lines of evidence provide strong support for the monophyly of Lepidoptera (summarized in [1,6,8]), relationships among superfamilies, especially those in the lower Ditrysia, remain largely uncertain.

One of the primary questions in lepidopteran phylogeny is the position and monophyly of butterflies, which remains unclear (figure 1) [9–12]. Morphological treatments considered butterflies as close to the inchworm moths and relatives (Geometroidea) [6,13,14], and recent studies based on up to 26 genes hinted that butterflies might belong within the lower ditrysian lineages,

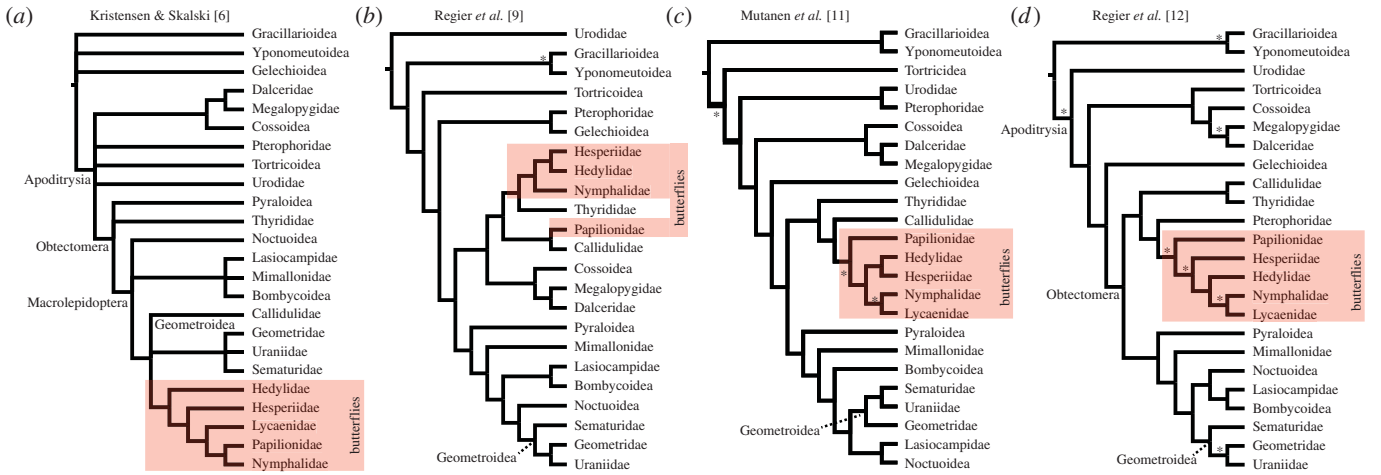

Figure 1. Phylogenetic trees of Lepidoptera showing the placement of butterflies. (*a*) Kristensen & Skalski [6], morphology. (*b*) Regier *et al.* [9], five nuclear loci. (*c*) Mutanen *et al.* [11], eight loci. (*d*) Regier *et al.* [12], 19 nuclear loci. Butterflies (Papilionoidea, *sensu* van Nieukerken *et al.* [2]) are highlighted in shaded boxes. For comparative purposes, these trees only show higher groups that were sampled in this study. Asterisks indicate branches with $\geq$70% bootstrap support. (Online version in colour.)

though these results were weakly supported [9–12]. Relationships among butterfly families are still being debated [15], and some studies based on a traditional Sanger-sequencing approach have suggested that butterflies might be a paraphyletic assemblage [10]. Assessing the placement and monophyly of butterflies has proven challenging because of the preponderance of low support and extremely short internal branches.

Recent advances in next-generation sequencing have resulted in novel of obtaining and analysing large amounts of data [16–19]. Two recent studies used next-generation data to examine relationships among lineages of Lepidoptera [20,21], but neither study included butterflies. In this study, we address the evolution of Lepidoptera and examine the problematic placement of butterflies by creating a dataset of 2696 putatively single-copy orthologous genes from 33 novel transcriptomes and publically available genomes, transcriptomes and expressed sequence tags (ESTs). We sampled at least one species from 19 major superfamilies of Lepidoptera and present one of the first phylogenomic analyses of butterflies and moths based on this many loci. To assess the influence of missing data on tree topology, we constructed two datasets: (i) a 'full dataset' with 2696 genes for 46 taxa and (ii) a 'reduced dataset' that included sequence data for all genes, totalling 465 loci for 26 taxa. Our conclusions challenge previous classifications and provide a solid evolutionary framework for future comparative studies on butterflies and moths.

2. Material and methods

(a) Transcriptome assembly, orthologue prediction and phylogenomic dataset construction

We followed the methods of Breinholt & Kawahara [21] for transcriptome assembly and orthologue prediction after sample collection, RNA extraction and library construction (see the electronic supplementary material for these methods). Paired-end sequences were merged with PEAR v. 0.8.1 (www.exelixis-lab.org/). Trim Galore! v. 0.3.2 (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to remove adapters and low-quality bases below a Phred score of 20. The additive multiple-k assembly method of Surget-Groba & Montoya-Burgos [22] was implemented in SOAPdenovo-Trans v. 1.01 [23] with five k-mer values (13, 23, 33, 43, 63). CD-HIT-EST [24] was used to combine redundant contigs from the multiple k-mer assemblies, and all sequences below 100 bp were removed with the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). OrthoDB v. 6 [25] was used to identify 6568 single-copy orthologous genes, using the single-copy profile for all species, from the reference genomes of *Bombyx mori* [26], *Danaus plexippus* [27], *Heliconius melpomene* [28] and *Manduca sexta* (Agricultural Pest Genomics Resource Database, http://www.agripestbase.org). Using these single-copy orthologous genes and reference genomes, we constructed a custom orthologue set (here called LEP1-COS) for use in HaMStR v. 8 [29] to extend the orthologue search to non-reference taxa (see the electronic supplementary material for details).

We built a data matrix from our 33 transcriptomes and available homologous genes from the four reference genomes and *Plutella xylostella* [30]. We combined these data with previously published transcriptomes that were assembled and processed from FASTQ files downloaded from the GenBank SRA database (see the electronic supplementary material). TranslatorX v. 9.03 [31] was used to convert nucleotides to amino acids, and back-translate amino acids to nucleotides after amino acid alignment in MAFFT v. 7.037 [32]. Both nucleotide and amino acid alignments were trimmed with ALISCORE v. 2.0 [33,34] and ALICUT v. 2.2 [35].

We created two separate datasets: a full dataset that included 46 taxa and 2696 genes, and a reduced dataset that included 26 species and 465 loci. The 26 taxa represented each major lineage in figure 2. We chose to create this smaller dataset to reduce the relative number of missing gene sequences, and to allow better partitioning of the data (the full dataset was restricted in partitioning options owing to its size and computational limitations; see discussion below).

(b) Phylogenomic analysis

We estimated phylogenies using nucleotide and amino acid data. Maximum-likelihood (ML) analyses were performed using RAxML v. 7.7.7 [36] for nucleotides and ExaML (https://github.com/stamatak/ExaML) for amino acids. We assessed the best partitioning strategy in PartitionFinder v. 1.1.1 [37]. Initial attempts to determine the best partitioning scheme by gene and codon position in PartitionFinder did not complete in a reasonable amount of time. The concatenated nucleotide matrix was partitioned by codon position, because PartitionFinder supported it as being better than an unpartitioned, single model. Despite large computational demands, we also managed to conduct a limited number of ML searches that were partitioned by gene. For the dataset partitioned by codon position, a GTRGAMMA

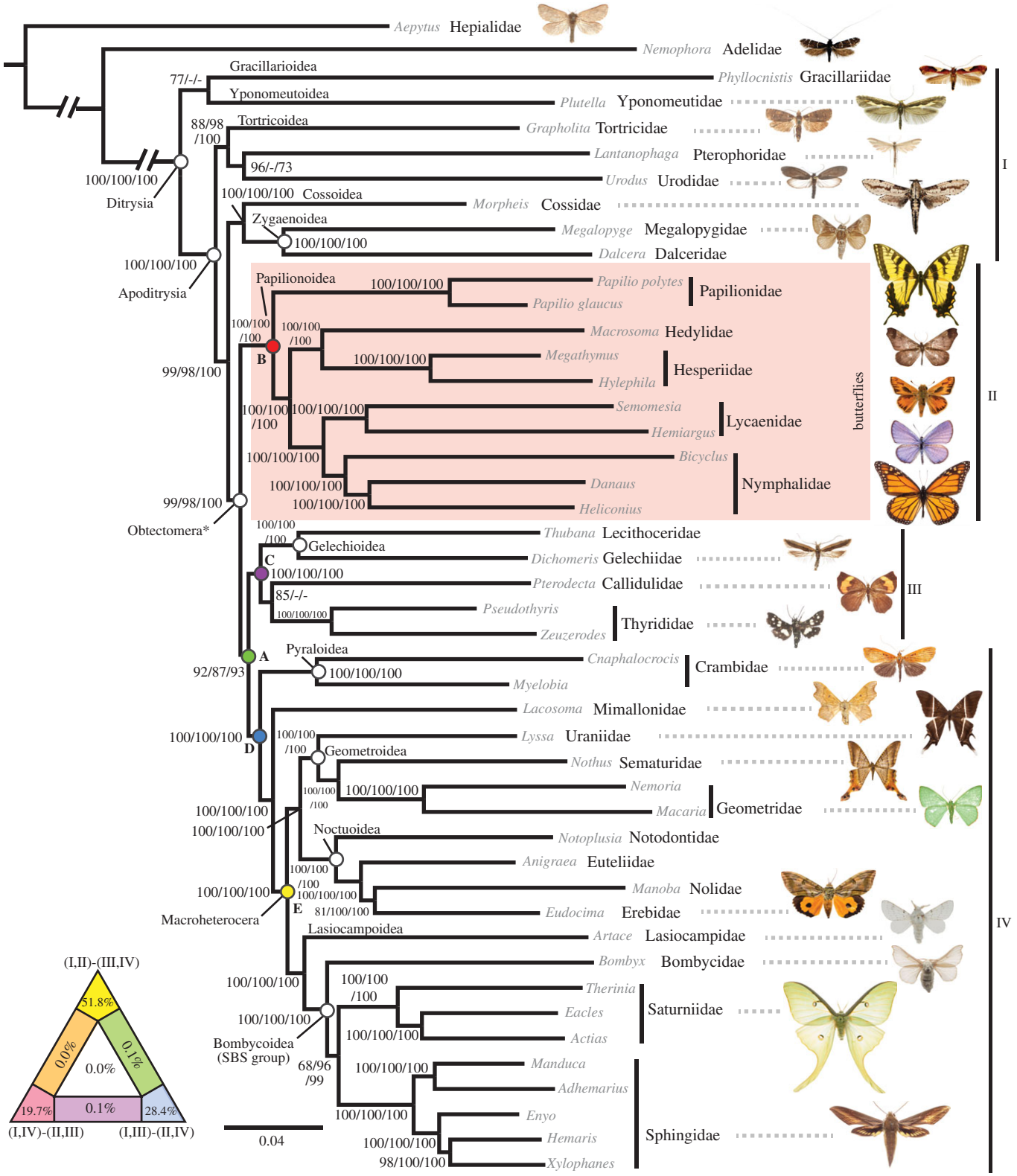

Figure 2. Maximum-likelihood tree estimated in RAxML from the 2696-gene nucleotide dataset. Values on branches are bootstrap values from the nucleotide, amino acid and SH-like amino acid analysis. The four taxon groups chosen for the TREE-PUZZLE mapping analyses are labelled (I–IV) to the right of each clade. The TREE-PUZZLE output (triangle) shows the probability among each possible topology for the defined groups. Clades A–E refer to groups that are discussed in the text. (Online version in colour.)

substitution model was applied to the nucleotide dataset for each codon position, and best ML tree searches were conducted using: (i) tree searches starting from a random topology using the '-f d' option (explained further in Kawahara *et al.* [38]) for 100 ML searches and (ii) tree searches using bootstrap trees as starting topologies for 50 ML searches.

For amino acids, a single partition was used owing to computational limitations. PartitionFinder identified the best model of evolution using the BIC score and the search was restricted to six possible protein models (JTT + G, JTT + G + F, LG + G, LG + G + F, WAG + G, WAG + G + F). We constructed 100 starting trees using the '-x' option in RAxML, which served as starting trees for 100 ExaML searches. To thoroughly search for the best tree, we also conducted 50 additional ExaML searches starting with likelihood trees that were constructed from ExaML bootstrap datasets. For both nucleotide and amino acid datasets, the number of bootstrap replicates was determined with the bootstrap stopping criterion [39] to determine a sufficient number of bootstrap replicates for each analysis. A total of 100 bootstrap replicates were run for the nucleotide dataset

and 150 for the amino acid dataset. For additional details, see the electronic supplementary material.

(c) Hypothesis testing and non-traditional branch support

We statistically compared our results with five previous phylogenetic hypotheses based on morphology [6,13] and molecules [9,11,12]. Tree searches were performed in RAxML with constraints that forced topologies that matched prior hypotheses. We also examined whether an alternative placement of butterflies that was found in some of the suboptimal trees was significantly worse than the best ML tree. In this alternative topology, Papilionoidea (clade B) was placed as the sister taxon to clade D (figure 2). For each of these tests, 50 ML tree searches were conducted in RAxML, and the SH test [40] was used to compare the tree with the highest likelihood from each alternative hypothesis with the best tree from the unconstrained search. Two additional methods were used to assess the placement of butterflies: SH-like branch support [41] in RAxML v. 7.7.7, and four-cluster likelihood mapping analysis [42] in TREE-PUZZLE v. 5.2 [43]. SH-like branch support was calculated because it can perform well even under conditions when models are misspecified [44]. For the four-cluster likelihood mapping analysis, we included one taxon per family to ease the computational demand on the analyses. We defined four taxon groups (labelled I–IV, figure 2), and the probability of each of the four arrangements was calculated.

3. Results and discussion

(a) Overview

We provide the first robust tree of Lepidoptera that strongly contradicts the traditional hypothesis that butterflies are a group nested within the macromoths. The Pyraloidea are confidently placed as the sister-group to Macroheterocera + Mimallonidae (100% BP), providing drastic improvement compared to recent Sanger-sequencing-based studies [9,11,12]. Butterflies, including the Hesperiidae and the enigmatic Hedylidae, constitute the Papilionoidea, which were placed at the base of the Obtectomera, as the sister group to clade A (figure 2), the clade including the Gelechioidea + Callidulidae + Thyrididae + Pyraloidea + Mimallonidae + Macroheterocera. The following results and discussion are focused on the full dataset (46 taxa, 2696 genes) unless otherwise noted.

(b) Transcriptome sequencing and bioinformatics

The 33 transcriptomes that were generated for this study had an average of 80 141 contigs that were above 100 bp, and the average N50 was 529 (electronic supplementary material, table S1). To identify orthologous genes across the entire dataset, we created a custom Lepidoptera core orthologue set (LEP1-COS) in HaMStR v. 8b [29], which identified an average of 4978 putative homologues per taxon after HMMER model prediction (electronic supplementary material, table S2). We used HaMStR to predict orthologous genes, but 282 genes could not be predicted in the four reference genomes that were used to make the core orthologue set. These genes, in addition to 3317 genes that had sequence data for 40 or fewer taxa, were removed. The final 'full' dataset included 2696 genes (2 525 742 nucleotides and 1 262 871 amino acid residues), and the 'reduced' dataset included 465 genes (492 800 nucleotides and 246 400 residues). For the full amino acid dataset, the BIC score from PartitionFinder supported LG + G as the best model of evolution. For the reduced dataset, PartitionFinder identified 178 partitions for nucleotides and 198 partitions for amino acids.

The full dataset had 94% gene coverage and 57–60% nucleotide completeness; the reduced dataset had 100% gene coverage and 65–69% nucleotide completeness (electronic supplementary material, tables S3–S6). Nucleotide completeness was estimated by removing gaps from sequences, summing the total number of nucleotides for each taxon, and dividing this quantity by the total number of nucleotides of two reference taxa, *B. mori* and *D. plexippus* (electronic supplementary material, table S6). We used two different reference taxa for this calculation owing to the inherent difference in the length of complete genes between these two taxa. Additional summary statistics for transcriptome assembly, orthologue prediction and phylogenomic dataset construction can be found in the electronic supplementary material, tables S1–S10.

(c) Phylogenomic analyses

Phylogenomic analyses resulted in a fully resolved tree with robust support for nearly all nodes (figure 2). Similar trees were obtained with nucleotides and amino acids, the only difference being the placement of *Phyllocnistis* (Gracillariidae), which was placed with weak support as the sister taxon to Yponomeutidae + Apoditrysia in the amino acid analysis. Butterflies (Papilionoidea) were placed sister to the remaining Obtectomera, the latter of which had robust branch support (92% BP (nucleotides), 87% BP (amino acids), 93% BP (SH-like amino acids); figure 2, clade A). Obtectomera *sensu* van Nieukerken *et al.* [2] was monophyletic, a result consistent with prior studies [12], albeit with stronger support here (greater than or equal to 98% BP support).

Both the 2696- and 465-gene datasets resulted in largely congruent trees. The 465-gene tree generally had lower branch support. The majority of nodes along the backbone of the 465-gene tree had weak (less than 50% BP) support, consistent with many previous studies based on smaller (less than 30) gene datasets [9–12]. Conflict between the 2696- and 465-gene tree topologies was minimal, but in the cases where it existed, there was strong support for a relationship in the 2696-gene tree, but support for a conflicting relationship in the 465-gene tree was weak. This could be due simply to limited phylogenetic signal in these genes, or misleading signal that is obscuring true signal.

Prior molecular phylogenetic work suggested that butterflies were more closely related to 'microlepidoptera' than to the large moths, but none of these studies was able to confidently come to this conclusion [9–12]. Regier *et al.* [9] tentatively placed Papilionoidea (including Hedyloidea + Callidulidae + Thyridoidea) sister to the Cossoidea + Zygaenoidea, but this relationship was not strongly supported. Mutanen *et al.* [11] recovered a monophyletic Papilionoidea (including Hedyloidea and Hesperioidea) + Callidulidae + Thyridoidea, but without strong branch support. Cho *et al.*'s [10] study, with greater gene sampling (26 loci) than the previous two studies, recovered a paraphyletic Papilionoidea with respect to several apoditrysian families (e.g. Callidulidae, Thyrididae). In analyses that excluded synonymous changes, Papilionoidea (including Hedyloidea

Table 1. Results from the SH test estimated from the full, 2696-gene dataset showing the likelihood (LH), difference in likelihood (*D*[LH]), standard deviation (s.d.) and *p*-value for each hypothesis of lepidopteran relationships, tested against the best tree from the RAxML analysis.

hypothesis	likelihood (LH)	*D*[LH]	s.d.	*p*-value
ML best tree	−22032784.662536	—	—	—
Kristensen & Skalski [6]	−22067315.274437	−34530.611901	456.735755	≤0.01
Regier *et al.* [9]	−22048197.710421	−15413.047885	316.163924	≤0.01
Mutanen *et al.* [11]	−22038389.679306	−5605.016770	219.595288	≤0.01
Regier *et al.* [12]	−22037635.791872	−4851.129336	232.101361	≤0.01
'macrolepidoptera' *sensu* Minet	−22035216.620019	−2431.957483	179.675611	≤0.01
alternative butterfly placement[a]	−22032927.703531	−143.040995	120.552791	>0.05

[a]((Gelechioidea, (Callidulidae,Thyrididae)), (Papilionoidea,(Pyraloidea,(Mimallonidae,Macroheterocera)))).

and Hesperioidea) were monophyletic with respect to the Pyraloidea + remaining Macrolepidoptera, but bootstrap support for this clade was low [10]. A recent study by Regier *et al*. [12] also recovered a monophyletic Papilionoidea (including Hedyloidea and Hesperioidea), which was placed as the sister group to the Pterophoroidea, but with bootstrap support that was not particularly high (72–83%). An SH test indicates that our ML tree is significantly ($p \leq 0.01$) more likely than trees constrained to previous morphological and molecular hypotheses (table 1).

The placement of butterflies in this study is contrary to Minet's [13] concept of a monophyletic Macrolepidoptera, a traditional group uniting butterflies with most of the large moths [45,46]. According to Minet [13], Papilionoidea are closely related to Geometroidea and Callidulidae (the lone family in Calliduloidea) within Macrolepidoptera. None of our analyses recovered a monophyletic Macrolepidoptera or a clade consisting of butterflies, Callidulidae and Geometroidea. To further test the monophyly of Macrolepidoptera, we constrained this clade and conducted an SH test. We conclude that a topology that enforces the monophyly of Macrolepidoptera is significantly worse ($p \leq 0.01$) than our most likely tree (table 1). These results are consistent with prior molecular studies that hinted at the possibility of a close relationship of butterflies with the Callidulidae, Gelechioidea and Thyrididae [9–12]. Thus, despite their close morphological resemblance to some of these moth families, butterflies are not closely related to macromoth families that are now grouped in the Macroheterocera.

Although our study places butterflies as the sister group to the remaining Obtectomera with strong bootstrap support, we chose to further examine its placement by conducting a four-cluster likelihood mapping analysis [42] in TREE-PUZZLE [43]. These results support the placement of butterflies in figure 2 (51.8% of quartets support this topology), compared with 28.4% for an alternative placement of butterflies, where the butterfly clade is sister to clade D (figure 2). The SH test confirms that this alternative relationship is less likely than our ML tree, though the results were not statistically significant (table 1).

We also tested to see whether the placement of butterflies in the ML tree was influenced by model misspecification. To do so, we used the amino acid dataset to calculate SH-like branch support [41], which can perform well even when models are misspecified [44]. SH-like support values were comparable with values obtained from standard bootstrap analyses, thus suggesting that the placement of butterflies is largely independent of the model chosen. Results from ML bootstraps, SH-like bootstraps and likelihood mapping all imply that there is a high probability that butterflies are placed sister to clade A, as in our ML topology (figure 2).

Our phylogeny supports the paraphyly of the historical definition of Papilionoidea [48] with regard to butterfly–moths (Hedylidae) and skippers (Hesperiidae), as well as strong support for the monophyly of Hedylidae + Hesperiidae (100% BP support). This result is consistent with results from recent studies that used a smaller number of genes [12,15] as well as some traditional morphological studies [49,50]. Our transcriptome-based topology provides strong support for Papilionidae as sister to the Hedylidae + Hesperiidae + Lycaenidae + Nymphalidae (electronic supplementary material, figures S1 and S2). Our study supports the classification of van Nieukerken *et al*. [2], that the division of butterflies, skippers and butterfly moths should not be split into three superfamilies [6,47].

One butterfly family, Pieridae, was excluded from the study because no adequate next-generation data were available. However, we do not believe that the overall topology recovered here will drastically change when Pieridae is included because of the very strong branch support (100% in all analyses, figure 2) found among the butterflies included in this study.

The phylogenetic position of the speciose clade Gelechioidea had been controversial [9,12]. Gelechioidea were historically excluded from the Apoditrysia and initially grouped with Gracillarioidea, Tineoidea and Yponomeutoidea [51]. However, recent molecular and morphological studies placed Gelechioidea within Apoditrysia [9,12,52]. The position of Gelechioidea appears to be largely influenced by the inclusion of synonymous signal, which, when removed, increases the probability that the group is placed near the root of Apoditrysia [10]. In our analyses, Gelechioidea appears sister to Callidulidae + Thyrididae with strong support, corroborating the inclusion of Gelechioidea within Apoditrysia.

Previous studies based on up to 26 genes had difficulty in confidently placing the Mimallonidae [9–12]. In this study, Mimallonidae is the sister group (100% BP support) to a monophyletic, well-supported Macroheterocera *sensu* van Nieukerken *et al*. [2] (also with 100% BP support). The position of Mimallonidae is corroborated by the fact that this family shares anatomical features with both Pyraloidea and macroheteroceran lineages [13]. In our reduced dataset, confidence in the placement of Mimallonidae is reduced; it becomes the

sister group to the Gelechioidea with weak, less than 50% bootstrap support (electronic supplementary material, figures S1 and S2). Bazinet *et al.*'s [20] recent transcriptome-based study also confidently places Mimallonidae as the sister-group to Macroheterocera (see discussion below), again hinting at the value of using a large number of genes to confidently estimate lepidopteran phylogeny.

Within the Macroheterocera, all nodes are supported by more than 80% bootstrap support except one: the node that unites the Bombycidae, Sphingidae and Saturniidae. This group, previously termed the 'SBS group' [53], has been difficult to estimate in many previous studies (summarized in [21]). In this study, the SBS group is supported by 68% bootstrap support when analysed with nucleotides, but support rises to 96% with amino acids (figure 2). Our results based on the full, 2696-gene dataset reveal the Bombycidae *sensu* Minet [54] as the sister group to Saturniidae + Sphingidae. This result is consistent with a previously published transcriptomic dataset [21]. However, when our dataset is reduced to 465 genes, relationships among these three families change (electronic supplementary material, figures S2 and S3). Saturniidae becomes the sister taxon to Bombycidae + Sphingidae, consistent with the results from previous studies based on fewer genes [9,55,56]. Breinholt & Kawahara [21] examined the interfamilial relationships among the SBS group using transcriptomic data and concluded that the strength of branch support among these three families appears to depend largely on the amount of data and how they are analysed. We suspect that weaker support in this study is owing to a difference in taxon sampling and alignment/pruning algorithms.

The phylogenetic relationships of Apoditrysia were the subject of a recently published transcriptome-based study by Bazinet *et al.* [20] that used 741 orthologous gene sequences (less than 28.5% of our genes). Although they did not include butterflies, their study revealed relationships consistent with those in our study. For instance, they resolved a well-supported Macroheterocera and Obtectomera. They also recovered Mimallonidae as the sister taxon of Macroheterocera, and the Pyraloidea as sister to that clade, with strong support. Their gene sampling had relatively little overlap with ours (11.9% overlapping genes between the two studies), suggesting that the overall congruence between the two studies reflects true phylogenetic signal.

One striking difference between these two studies is the relationship among the superfamilies Bombycoidea, Geometroidea, Lasiocampoidea and Noctuoidea, which altogether constitute approximately 73 000 described moth species [2]. Our tree resolves Bombycoidea + Lasiocampoidea as the sister group to Geometroidea + Noctuoidea with strong branch support (100% BP support for all nodes), whereas Bazinet *et al.*'s [20] study placed Geometroidea with Bombycoidea + Lasiocampoidea, and Noctuoidea as the sister group to this three-taxon clade (83% BP support). Relationships that were obtained here were also recovered in the Sanger-based trees of Regier *et al.* [9,12], albeit with weak support in their studies. Morphological characters, such as wing venation, also support the monophyly of Geometroidea plus Noctuoidea [57].

Our study design addressed two points that Bazinet *et al.* [20] suggested might further the recovery of a robust lepidopteran tree, namely the inclusion of more loci and the use of taxon-specific core orthologues. Our custom core orthologue set allowed for the collection of Lepidoptera-specific loci and an improvement in the number of genes that could be included in the phylogenomic analysis. The 2696 loci from our LEP1-COS produced a tree with 36 nodes (out of 43) with greater than or equal to 98% BP. Thus, as predicted by Bazinet *et al.* [20], it appears that using Lepidoptera-specific core orthologues and increasing the number of genes provides a more robust phylogeny of Lepidoptera.

(d) Perspectives

The phylogeny presented in this study offers insights into the evolutionary relationships of one of the largest diversifications of insects. By using transcriptomic data, we show that these data can resolve many difficult relationships among butterflies and moths that were previously based on a much smaller number of genes. The butterfly clade is well supported and is placed sister to the Callidulidae + Gelechioidea + Macroheterocera + Mimallonidae + Thyrididae. Hedylidae was the sister group to Hesperiidae, and this placement is consistent with recent molecular studies. Butterflies, traditionally thought to be close relatives of large moths based on anatomical features [6,13], now definitively appear to be placed outside of the Macroheterocera. Although previous molecular studies hinted at such a placement for butterflies, we suspect that because those analyses were based on a much smaller set of genes, they did not contain enough phylogenetic signal to resolve a rapid Mesozoic diversification [1,4]. Similar conclusions were reached for other major insect radiations that are thought to have diversified with flowering plants, such as flies [58,59] and beetles [60].

The proposed change in the phylogenetic position of Papilionoidea brings about the question of whether its new placement would substantially push back the age of Papilionoidea. Fossil Lepidoptera are quite sparse due to poor preservation and generally soft-bodied larval stages [1,6]. Many butterfly fossils are from the Miocene, and the oldest butterfly fossils are from the Late Palaeocene [61]. The new placement of butterflies remains consistent with the little that is known from the lepidopteran fossil record. A dated phylogeny of Lepidoptera based on next-generation data is clearly the next step.

Our robust phylogeny provides an initial framework necessary to understand life-history evolution in butterflies and moths. For instance, our results imply that a general shift in body size from small 'microlepidoptera' to large 'macrolepidoptera' is untenable, though there appears to be a general trend from ancestral moth lineages that feed on ferns to derived Lepidoptera that feed on advanced angiosperms [3,62]. Furthermore, our tree provides a baseline to test whether adult diurnal activity, a trait common to nearly all butterflies, evolved much earlier than previously thought. A shift to diurnal activity might have served as a means of escaping from nocturnal predators such as bats, which present significant pressure on lepidopteran prey [63–65]. Future work will involve understanding the causes of these transitions across the Lepidoptera tree of life. Although the trees presented here clarify our understanding of deep-level lepidopteran relationships, many lineages still need to be sampled. We expect that the many new phylogenomic initiatives for non-model Lepidoptera and relatives (e.g. 1KITE (http://www.1kite.org/), i5K (http://www.arthropodgenomes.org/), LepTree (http://entomology.umd.edu/mitterlab/leptree)) will provide a rich source of additional data that will further elucidate the evolution of one of the most charismatic and popular groups of insects.

Acknowledgements. L. Xiao prepared RNA-Seq libraries. G. Hill, C. Johns, L. Reeves and D. Plotkin assisted with figures, tables and text. J. R. Barber, J. E. Hayden, P. R. Houlihan, W. Hsu, B. Leavell, D. Matthews-Lott, and A. D. Warren helped obtain samples. We thank A. L. Bazinet, M. P. Cummings, C. Mitter, K. Mitter, J. C. Regier and A. Zwick for their continued support. We acknowledge the UF HPC for providing computational support and assistance. N. Wahlberg and two anonymous reviewers provided helpful suggestions.

Data accessibility.

— The LEP1 core orthologue set: Dryad data depository (http://datadryad.org) accession (doi:10.5061/dryad.qd27g).
— Assembled transcriptomes: Dryad data depository (http://datadryad.org) accession (doi:10.5061/dryad.qd27g).
— Raw Illumina reads: GenBank BioProject PRJNA248471, GenBank sequence read archive (SRA) database accession numbers: SRR1298384, SRR1299208–SRR1299214, SRR1299217, SRR1299267, SRR1299274, SRR1299296, SRR1299306, SRR1299316–SRR1299318, SRR1299347, SRR1299369, SRR1299394, SRR1299418, SRR1299435, SRR1299495, SRR1299746, SRR1299750-SRR1299752, SRR1299755, SRR1299769, SRR1299773, SRR1299782, SRR1300145, SRR1300148, SRR1300991.

Funding statement. This work was supported by the University of Florida, the Florida Museum of Natural History, and in part by NSF grants DEB-1354585, IOS-1121739 and National Geographic Society grant no. 9107-12 to A.Y.K.

References

1. Grimaldi D, Engel MS. 2005 *Evolution of the insects*, p. 772. Cambridge, UK: Cambridge University Press.
2. van Nieukerken EJ *et al.* 2011 Order Lepidoptera Linnaeus, 1758. In Z-Q Zhang (ed), animal biodiversity: an outline of higher-level classification and survey of taxonomic richness. *Zootaxa* **3148**, 212–221.
3. Powell JA, Mitter C, Farrell BD. 1998 Evolution of larval food preferences in Lepidoptera. In *Lepidoptera: moths and butterflies 1 Handbuch der Zoologie/handbook of zoology IV/35* (ed. NP Kristensen), pp. 403–422. Berlin, Germany: Walter de Gruyter.
4. 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)
5. Winkler IS, Mitter C. 2008 The phylogenetic dimension of insect/plant interactions: a summary of recent evidence. In *Specialization, speciation, and radiation: the evolutionary biology of herbivorous insects* (ed. KJ Tilmon), pp. 240–263. Berkeley, CA: University of California Press.
6. Kristensen NP, Skalski AW. 1998 Phylogeny and palaeontology. In *Lepidoptera: moths and butterflies 1 Handbuch der Zoologie/handbook of zoology IV/35* (ed. NP Kristensen), pp. 7–25. Berlin, Germany: Walter de Gruyter.
7. Roe AD *et al*. 2009 Evolutionary framework for Lepidoptera model systems. In *Genetics and molecular biology of Lepidoptera* (eds M Goldsmith & F Marec), pp. 1–24. Gainesville, FL: CRC Press.
8. Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. 2012 Advances in insect phylogeny at the dawn of the postgenomic era. *Annu. Rev. Entomol.* **57**, 449–468. (doi:10.1146/annurev-ento-120710-100538)
9. Regier JC *et al*. 2009 Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. *BMC Evol. Biol.* **9**, 280. (doi:10.1186/1471-2148-9-280)
10. Cho S *et al*. 2011 Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? *Syst. Biol.* **60**, 782–796. (doi:10.1093/sysbio/syr079)
11. Mutanen M, Wahlberg N, Kaila L. 2010 Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proc. R. Soc. B* **277**, 2839–2848. (doi:10.1098/rspb.2010.0392)
12. 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. (doi:10.1371/journal.pone.0058568)
13. Minet J. 1991 Tentative reconstruction of the ditrysian phylogeny (Lepidoptera, Glossata). *Entomol. Scand.* **22**, 69–95. (doi:10.1163/187631291X00327)
14. DeJong R, VaneWright RI, Ackery PR. 1996 The higher classification of butterflies (Lepidoptera): problems and prospects. *Entomol. Scand.* **27**, 65–101. (doi:10.1163/187631296X00205)
15. Heikkilä M, Kaila L, Mutanen M, Peña C, Wahlberg N. 2012 Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proc. R. Soc. B* **279**, 1093–1099. (doi:10.1098/rspb.2011.1430)
16. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013 Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* **66**, 526–538. (doi:10.1016/j.ympev.2011.12.007)
17. Metzker ML. 2010 Applications of next-generation sequencing technologies: the next generation. *Nat. Rev. Genet.* **11**, 31–46. (doi:10.1038/nrg2626)
18. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013 The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38. (doi:10.1016/j.cell.2013.09.006)
19. Li H, Homer N. 2010 A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11**, 473–483. (doi:10.1093/bib/bbq015)
20. Bazinet AL, Cummings MP, Mitter K, 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. (doi:10.1371/journal.pone.0082615)
21. Breinholt JW, Kawahara AY. 2013 Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. *Genome Biol. Evol.* **5**, 2082–2092. (doi:10.1093/gbe/evt157)
22. Surget-Groba Y, Montoya-Burgos JI. 2010 Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.* **20**, 1432–1440. (doi:10.1101/gr.103846.109)
23. Luo R *et al*. 2012 SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18. (doi:10.1186/2047-217X-1-18)
24. Li WZ, Godzik A. 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659. (doi:10.1093/bioinformatics/btl158)
25. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. 2013 OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41**, D358–D365. (doi:10.1093/nar/gks1116)
26. Xia QY *et al*. 2004 A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**, 1937–1940. (doi:10.1126/science.1102210)
27. Zhan S, Merlin C, Boore JL, Reppert SM. 2011 The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**, 1171–1185. (doi:10.1016/j.cell.2011.09.052)
28. Heliconius Genome Consortium. 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98. (doi:10.1038/nature11041)
29. Ebersberger I, Strauss S, von Haeseler A. 2009 HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9**, 157. (doi:10.1186/1471-2148-9-157)
30. You MS *et al*. 2013 A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220–225. (doi:10.1038/ng.2524)
31. Abascal F, Zardoya R, Telford MJ. 2010 TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13. (doi:10.1093/nar/gkq291)
32. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)

33. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wagele JW, Misof B. 2010 Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front. Zool.* **7**, 1–12. (doi:10.1186/1742-9994-7-10)
34. Misof B, Misof K. 2009 A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.* **58**, 21–34. (doi:10.1093/sysbio/syp006)
35. Kück P. 2009 ALICUT: a Perlscript which cuts ALISCORE identified RSS. Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version 2.3.
36. Stamatakis A. 2006 RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. (doi:10.1093/bioinformatics/btl446)
37. Lanfear R, Calcott B, Ho SY, Guindon S. 2012 PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701. (doi:10.1093/molbev/mss020)
38. Kawahara AY, Breinholt JB, Ponce FV, Haxaire J, Xiao L, Lamarre GPA, Rubinoff D, Kitching IJ. 2013 Evolution of *Manduca sexta* hornworms and relatives: biogeographical analysis reveals an ancestral diversification in Central America. *Mol. Phylogenet. Evol.* **68**, 381–386. (doi:10.1016/j.ympev.2013.04.017)
39. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. 2010 How many bootstrap replicates are necessary? *J. Comput. Biol.* **17**, 337–354. (doi:10.1089/cmb.2009.0179)
40. Shimodaira H, Hasegawa M. 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116. (doi:10.1093/oxfordjournals.molbev.a026201)
41. Anisimova M, Gascuel O. 2006 Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552. (doi:10.1080/10635150600755453)
42. Strimmer K, von Haeseler A. 1997 Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl Acad. Sci. USA* **94**, 6815–6819. (doi:10.1073/pnas.94.13.6815)
43. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002 TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504. (doi:10.1093/bioinformatics/18.3.502)
44. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011 Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699. (doi:10.1093/sysbio/syr041)
45. Scott JA. 1985 The phylogeny of butterflies (Papilionoidea and Hesperioidea). *J. Res. Lepid.* **23**, 241–281.
46. Scott JA, Wright DM. 1990 Butterfly phylogeny and fossils. In *Butterflies of Europe*, vol. 2 (ed. O Kudrna), pp. 152–208. Wiesbaden, Germany: Aula.
47. Scoble MJ. 1992 *The Lepidoptera. Form, function and diversity*. Oxford, UK: Oxford University Press.
48. Ackery PR, de Jong R, Vane-Wright RI. 1998 The butterflies: Hedyloidea, Hesperoidea and Papilionoidea. In *Lepidoptera: moths and butterflies 1 Handbuch der Zoologie/handbook of zoology IV/35* (ed. NP Kristensen), pp. 263–300. Berlin, Germany: Walter de Gruyter.
49. Weller SJ, Pashley DP. 1995 In search of butterfly origins. *Mol. Phylogenet. Evol.* **4**, 235–246. (doi:10.1006/mpev.1995.1022)
50. Scoble MJ. 1986 The structure and affinities of the Hedyloidea: a new concept of butterflies. *Bull. Brit. Mus. Nat. Hist.* (*Entomol.*) **53**, 251–286.
51. Minet J. 1983 Etude morphologique et phylogénétique des organes tympaniques des Pyraloidea. 1 - généralités et homologies. (Lep. Glossata). *Ann. Soc. Entomol. Fr.* **19**, 175–207.
52. Kaila L. 2004 Phylogeny of the superfamily Gelechioidea (Lepidoptera : Ditrysia): an exemplar approach. *Cladistics* **20**, 303–340. (doi:10.1111/j.1096-0031.2004.00027.x)
53. Zwick A, Regier JC, Mitter C, Cummings MP. 2011 Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera). *Syst. Entomol.* **36**, 31–43. (doi:10.1111/j.1365-3113.2010.00543.x)
54. Minet J. 1994 The Bombycoidea: phylogeny and higher classification (Lepidoptera, Glossata). *Entomol. Scand.* **25**, 63–88. (doi:10.1163/187631294X00045)
55. Huan-Na C, Yu-Zhou D, Bao-Ping Z. 2012 Characterization of the complete mitochondrial genomes of *Cnaphalocrocis medinalis* and *Chilo suppressalis* (Lepidoptera: Pyralidae). *Int. J. Biol. Sci.* **8**, 561–579. (doi:10.7150/ijbs.3540)
56. Kim MJ, Kang AR, Jeong HC, Kim KG, Kim I. 2011 Reconstructing intraordinal relationships in Lepidoptera using mitochondrial genome data with the description of two newly sequenced lycaenids, *Spindasis takanonis* and *Protantigius superans* (Lepidoptera: Lycaenidae). *Mol. Phylogenet. Evol.* **61**, 436–445. (doi:10.1016/j.ympev.2011.07.013)
57. Brock JP. 1971 Contribution towards an understanding of morphology and phylogeny of Ditrysian–Lepidoptera. *J. Nat. Hist.* **5**, 29–102. (doi:10.1080/00222937100770031)
58. Labandeira CC. 2005 Fossil history and evolutionary ecology of Diptera and their associations with plants. In *The evolutionary biology of flies* (eds D Yeates, BM Wiegmann), pp. 217–273. New York, NY: Columbia University Press.
59. Wiegmann BM *et al*. 2011 Episodic radiations in the fly tree of life. *Proc. Natl Acad. Sci. USA* **108**, 5690–5695. (doi:10.1073/pnas.1012675108)
60. Farrell BD. 1998 'Inordinate fondness' explained: why are there so many beetles? *Science* **281**, 555–559. (doi:10.1126/science.281.5376.555)
61. Sohn JC, Labandeira CC, Davis DR, Mitter C. 2012 An annotated catalog of fossil and subfossil Lepidoptera (Insecta: Holometabola) of the world. *Zootaxa* **3286**, 1–132.
62. Menken SBJ, Boomsma JJ, van Nieukerken EJ. 2010 Large-scale evolutionary patterns of host plant associations in the Lepidoptera. *Evolution* **64**, 1098–1119. (doi:10.1111/j.1558-5646.2009.00889.x)
63. Barber JR, Kawahara AY. 2013 Hawkmoths produce anti-bat ultrasound. *Biol. Lett.* **9**, 20130161. (doi:10.1098/rsbl.2013.0161)
64. Connor WE, Corcoran AJ. 2012 Sound strategies: the 65-million-year-old battle between bats and insects. *Annu. Rev. Entomol.* **57**, 21–39. (doi:10.1146/annurev-ento-121510-133537)
65. Corcoran AJ, Barber JR, Conner WE. 2009 Tiger moth jams bat sonar. *Science* **325**, 325–327. (doi:10.1126/science.1174096)