Systematic Entomology

Systematic Entomology (2016), DOI: 10.1111/syen.12217

Phylotranscriptomics resolves ancient divergences in the Lepidoptera

ADAM L. BAZINET[1,*], KIM T. MITTER[2], DONALD R. DAVIS[3], ERIK J. VAN NIEUKERKEN[4], MICHAEL P. CUMMINGS[1] and CHARLES MITTER[2]

[1]Laboratory of Molecular Evolution, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, U.S.A., [2]Department of Entomology, University of Maryland, College Park, MD, U.S.A., [3]Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC, U.S.A. and [4]Naturalis Biodiversity Center, Leiden, the Netherlands

Abstract. Classic morphological studies of the oldest, so-called nonditrysian lineages of Lepidoptera yielded a well-resolved phylogeny, supported by the stepwise origin of the traits characterizing the clade Ditrysia, which contains over 98% of extant lepidopterans. Subsequent polymerase chain reaction (PCR)-based molecular studies have robustly supported many aspects of the morphological hypothesis and strongly contradicted others, while leaving some relationships unsettled. Here we bring the greatly expanded gene sampling of RNA-Seq to bear on nonditrysian phylogeny, especially those aspects that were not conclusively resolved by the combination of morphology and previous PCR-based multi-gene studies. We analysed up to 2212 genes in each of 28 species representing all 12 superfamilies and 15 of 21 families of nonditrysians, plus trichopteran outgroups and representative Ditrysia. Our maximum likelihood phylogeny estimates used both nonsynonymous changes only (degen1 coding) and all nucleotides (nt123) partitioned by codon position, recovering a novel hypothesis for early glossatan relationships that is the most strongly supported to date. We find strong support for Micropterigidae alone as the sister group to all other Lepidoptera, in agreement with morphology and early molecular evidence, but in contrast to recent PCR-based studies. Also very strongly supported are the previously recognized clades Angiospermivora, Heteroneura, Eulepidoptera and Euheteroneura. Finally, we find strong support for paraphyly of the southern hemisphere family Palaephatidae, with the South American genus *Palaephatus* Butler forming the previously undetermined sister group to Ditrysia. The remaining palaephatids, Australian and South American, form the sister group to Tischeriidae.

Introduction

Morphological studies of the oldest divergences within the insect order Lepidoptera, giving rise to the so-called nonditrysian lineages, comprised a landmark early application of Hennigian phylogenetics (Hennig, 1953; Kristensen, 1984; Davis, 1986; Kobayashi & Ando, 1988; Nielsen & Kristensen, 1996; reviews in Kristensen & Skalski, 1998; Kristensen *et al*., 2007). The well-resolved hypothesis of nonditrysian relationships that resulted, as summarized by Kristensen (2003), is shown in (Fig. 1A). It is supported by a series of synapomorphies through which the traits characterizing the great majority of modern species are hypothesized to arise in stepwise fashion.

Subsequent molecular studies have robustly supported many aspects of the morphological hypothesis, markedly contradicted

Correspondence: Charles Mitter, Department of Entomology, University of Maryland, College Park, Maryland 20742, U.S.A. E-mail: cmitter@umd.edu

*Present address: National Biodefense Analysis and Countermeasures Center, Fort Detrick, MD, U.S.A.

No conflicts of interest were discovered.

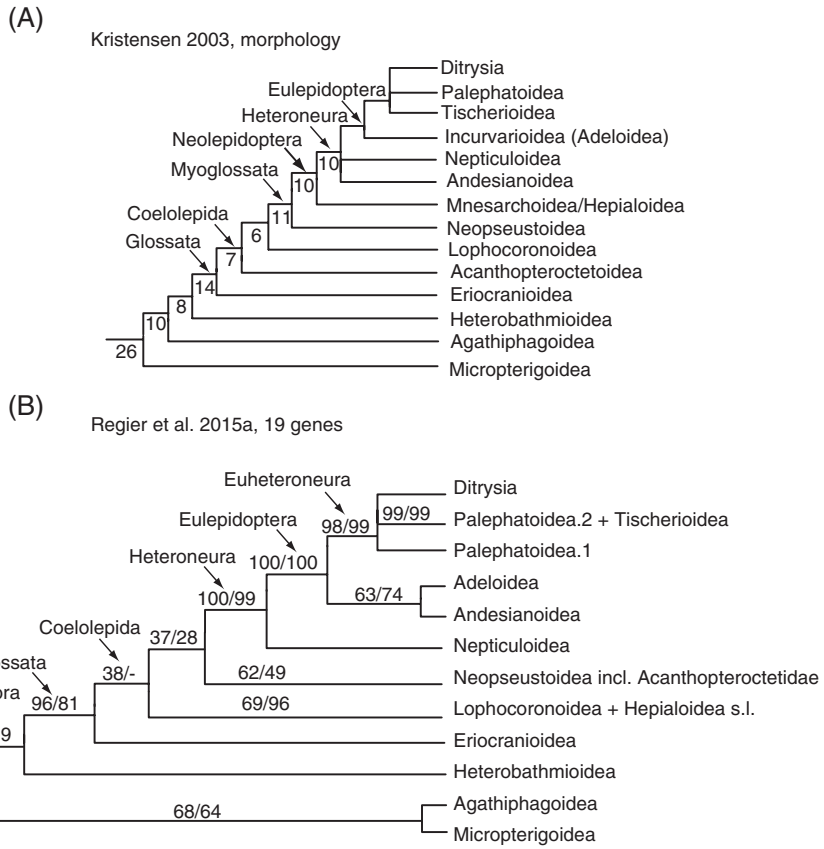

Fig. 1. Previous hypotheses on relationships among nonditrysian lepidopteran lineages. (A) Synopsis of relationships inferred from morphology, redrawn from Kristensen (2003). Numbers below branches are numbers of synapomorphies hypothesized by Kristensen (1984). (B) Summary of relationships among nonditrysian superfamilies found by Regier *et al*. (2015a). Bootstrap values above branches: degen1/nt123. The same topology was found by Kristensen *et al*. (2015).

a few, and left several ambiguous (Wiegmann *et al*., 2000; Mutanen *et al*., 2010; Regier *et al*., 2013, 2015a; Kristensen *et al*., 2015). Figure 1B summarizes the concordant results of recent analyses by Regier *et al*. (2015a) and Kristensen *et al*. (2015). Among the clades now corroborated by both morphology and strong molecular evidence are: Angiospermivora, characterized by larvae feeding predominantly on living angiosperms; Glossata, defined by haustellate adult mouthparts; Heteroneura, defined by differing hindwing versus forewing venation, frenular wing coupling and associated traits; Eulepidoptera, defined by, among other traits, origin of the pilifers and of an advanced locking mechanism in the proboscis; and Euheteroneura. The evidence remains less conclusive on several other nodes. Micropterigidae have long been thought to be the sister group to remaining Lepidoptera, a placement corroborated by early molecular evidence (Wiegmann *et al*., 2000). In two recent analyses based in part on the 19-gene dataset of Regier *et al*. (2013), however, Micropterigidae are the sister group to Agathiphagidae (Fig. 1B), with weak to strong support depending on the analysis (Kristensen *et al*., 2015; Regier *et al*., 2015a). Within Glossata, Lophocoronidae are strongly supported as the sister group to Hepialoidea, contradicting the morphological clades Myoglossata, defined by possession of a proboscis with intrinsic musculature, and Neolepidoptera, defined by, among other traits, musculate, crochet-bearing larval abdominal prolegs. The basal divergences among glossatans are otherwise weakly resolved. The position of Andesianoidea, discovered subsequent to the early morphological work, has not been established conclusively by molecular evidence, though it is consistently grouped with Adeloidea, sometimes with strong support (Mutanen *et al*., 2010). Finally, the basal divergence within Euheteroneura, and hence the sister group to the enormous clade Ditrysia (98% of the Lepidoptera), has not been discernible from morphology, while the molecular evidence has been contradictory (Regier *et al*., 2015a).

In this paper we bring the greatly expanded gene sampling of RNA-Seq to bear on nonditrysian phylogeny, particularly those aspects that were not conclusively resolved by the combination of morphology and previous PCR-based multi-gene studies. We analysed up to 2212 genes in each of 28 species representing all 12 superfamilies of nonditrysians (Regier *et al*., 2015a) plus outgroups and representative Ditrysia, obtaining strong bootstrap support at every node, and compare the results against previous hypotheses.

Materials and methods

Taxon sampling

The primary aim of this study was to re-examine the relationships of the nonditrysian superfamilies with each other and with Ditrysia. We sampled a total of 21 nonditrysian species representing all 12 superfamilies and 15 of 21 families recognized by Regier *et al*. (2015a), plus four species representing four early-diverging lineages of Ditrysia (Regier *et al*., 2015b). Exemplars of three divergent superfamilies of Trichoptera were used as outgroups. Of the 28 total transcriptomes analysed here, 18 were generated de novo for this study, three represent reanalyses of sequence data previously reported by us (Bazinet *et al*., 2013), and seven represent de novo reassemblies of sequence reads reported by others (Kawahara & Breinholt, 2014; Misof *et al*., 2014; Peters *et al*., 2014). The exemplars included in this study, and the sources of the data for each, are listed in Table S1.

The 21 specimens from which we generated RNA-Seq data were obtained by our collecting and with the gracious assistance of collectors around the world (see Acknowledgements). They were stored in 100% ethanol at −85°C, as a part of the ATOLep frozen tissue collection at the University of Maryland, College Park, U.S.A. Some had been stored for over 20 years. The species sequenced and specimen accession numbers are listed in Table S1. Most of the specimens we prepared (15) were adults, while six were larvae (Table S1). Nucleic acid extraction used only the head and thorax for most adult specimens, leaving the abdomen and genitalia as a voucher, although the entire specimen was consumed for small adults and most larvae. For very small moths (*Tischeria* Zeller, *Tineola* Herrich-Schäffer), multiple conspecific individuals were extracted together. DNA 'barcodes' were generated for all taxa, either by us, using standard primer sequences with M13 tails (Regier & Shi, 2005), or, more typically, by the All-Leps Barcode of Life project (http://www.lepbarcoding.org). COI DNA barcodes were checked against the Barcode of Life Data system reference library (Ratnasingham & Hebert, 2007) to confirm specimen identifications and also to facilitate future identification of specimens whose identity is still pending, i.e. species listed as 'sp.' or 'unidentified' in this report.

De novo RNA-Seq data generation for 18 taxa

About half of the extracts used in this study (see Table S1) had been prepared for previous reverse transcription PCR-based studies (e.g., Regier *et al*., 2015a,b) and stored at −80°C for 5 years or more. The rest were prepared de novo, using kits specifically designed for retrieval of low-quantity RNA. These kits made it possible to obtain RNA-Seq data from specimens that had been stored in 100% ethanol at −80°C for up to 20 years. A few of these specimens had even been dried before they were placed in ethanol.

Nucleic acids were extracted using Promega SV total RNA isolation mini-kits either with (five taxa) or without DNase digestion (18 taxa) (Promega, Fitchburg, WI, U.S.A.). Following DNase digestion, the RNA-only preps (five taxa) were subjected to poly-A selection and indexed library construction for sequencing on an Illumina (San Diego, CA, U.S.A.) HiSeq 1000 in the University of Maryland-Institute for Bioscience and Biotechnology Research Sequencing Core. The remaining extracts of total nucleic acids (18 taxa) were used to produce cDNAs with low-input Clontech kits for either poly-dT priming (SMARTer Ultra Low input RNA kit –v3, #634849) or universal priming (#634940) (Clontech, Woburn, MA, U.S.A.). Following shearing to 200 bp size with a Covaris instrument (Mountain View, CA, U.S.A.) cDNA fragments were used for indexed library construction (Clontech kit #634947 for low input). Following Hittinger *et al*. (2010), libraries were left unnormalized so as to favour highly expressed genes likely to be present in most species and all life stages. Libraries were run eight per lane, yielding an average of approximately 58 million 100 bp paired-end reads per taxon in high-output mode (15 taxa; Table S2) or approximately 33 million reads in rapid-run mode (nine taxa). For three taxa, two sequencing runs were made, once in each mode, and the data combined. Previously published transcriptome libraries obtained by other investigators, which were either 100 bp paired-end (*Micropterix* Hübner and *Philopotamus* Stephens) or 150 bp paired-end (five others), averaged about 37 million reads per taxon (Table S2). The Illumina reads for the 18 newly sequenced taxa are available in the NCBI Sequence Read Archive as part of BioProject PRJNA222254.

Sequence quality control and transcript assembly

Quality control of sequence reads and transcript assembly had previously been performed for four of our taxa (*Dryadaula* Meyrick, *Palaephatus* (*Palaephatus*), *Phymatopus* Wallengren, and *Thyridopteryx* Haworth in Table S1; Bazinet *et al*., 2013) and was not repeated here. We performed quality control and assembly for the remaining newly generated transcriptomes and seven previously published transcriptomes with the updated methods described here.

We used the default Illumina HiSeq 1000 quality filter, which ensured that at least 24 of the first 25 template cycles had a 'Chastity' value greater than 0.6. The Chastity value is a ratio between the highest intensity and the sum of the two highest intensities. We discarded reads that did not pass the Chastity quality filter (≈6% per sample; Table S2). We then used AUTOADAPT (AUTOADAPT, 2014), which in turn calls FASTQC (FastQC, 2014) and CUTADAPT (Martin, 2011) with default settings to detect and remove overrepresented sequences, as well as to trim and remove low-quality reads. To look for possible cross-sample contamination, J. Breinholt (personal communication) kindly screened our samples using an unpublished procedure. The estimated average frequency of contaminants was less than 2%, a level we considered unlikely to affect our phylogenetic analyses. We assumed that other contaminants, if present, would be removed by our orthology determination and paralogy filter workflow (see later).

De novo transcriptome assembly was initially performed using both TRINITY [versions 2.0.6, r2014-07-17, r2014-04-13, and r2013-02-25; (Grabherr *et al*., 2011)] and TRANS-ABYSS [version 1.4.4; ABySS version 1.5.2; (Birol *et al*., 2009; Robertson *et al*., 2010)]. Assembly statistics, including numbers and length of transcripts and N50 (the length N for which 50% of all bases are contained in contigs of length $L < N$), are given in Table S2. A typical TRINITY assembly required greater than 100 GB RAM and finished in 24–96 h using 16 processing cores. A typical TRANS-ABYSS run required less than 4 GB RAM and a single processor, finishing in 1–2 h. The same was true for each constituent ABYSS run, of which there were 23 per sample (k ranged from 52 to 96 in steps of two). In general, TRINITY used more RAM and produced fewer transcripts than TRANS-ABYSS, but it produced longer transcripts (Table S2). Combining the TRINITY and TRANS-ABYSS assemblies yielded a slightly more complete data matrix than using either assembly by itself, so for ten taxa analysed early in this study [*Dryadaula*, *Eudarcia* Clemens, *Palaephatus* (*Palaephatus*), *Phymatopus*, *Ptyssoptera* Turner, *Thyridopteryx*, *Tineola*, *Tischeria*, *Micropterix*, and *Philopotamus*] we used the combined assembly throughout the workflow. For the remaining taxa, however, we used only TRINITY. Comparative assembly statistics for *Micropterix* and *Philopotamus*, given in Table S3, show that our TRINITY assemblies yield more and larger contigs than the NEWBLER assemblies used by Peters *et al*. (2014). Similar trends were seen in comparisons of our assemblies to the SOAPDEVNOVO-TRANS (Xie *et al*., 2014) assembly of *Nemophora* Illiger & Hoffmannsegg by Kawahara & Breinholt (2014). Our assemblies are deposited in the NCBI Transcriptome Shotgun Assembly Database (https://www.ncbi.nlm.nih.gov/genbank/tsa/) as part of BioProject PRJNA222254, with contigs <200 bp eliminated as required by the database. The unfiltered assemblies are available in Dryad (doi:10.5061/dryad.hj278).

Orthology determination: constructing a Lepidoptera-specific orthologue database

In a previous study (Bazinet *et al*., 2013) we conducted orthology determinations using a database of 'known' orthologues assembled from a broad taxon sampling across the insect orders, other arthropods and related phyla. We hypothesized that the number and reliability of orthologues identified within Lepidoptera would have been greater if we had started from a database built using only lepidopteran genomes (see also Kawahara & Breinholt, 2014). In building a Lepidoptera-specific nuclear gene database for this study, we first downloaded peptide and coding sequences for *Bombyx mori* L., *Heliconius melpomene* L., and *Danaus plexippus* L. from Ensembl Metazoa, release 22 (Cunningham *et al*., 2014; Flicek *et al*., 2014). Providing all the peptide or gene sequence identifiers as input, we built up orthologous groups using the one2one, one2many, many2many, within_species_paralog, putative_gene_split, and contiguous_gene_split homology relationships defined in Ensembl that involved any two of these three taxa (Cunningham *et al*., 2014). We required an orthologous group to contain a *Bombyx* L. sequence and a minimum of one butterfly sequence (either *Danaus* or *Heliconius*), which resulted in 7042 orthologous groups. From Ensembl we retrieved the 'genetree alignment' corresponding to each orthologous group; from each genetree alignment we extracted only the sequences belonging to the three Lepidoptera species of interest, removed gaps, and realigned the amino acid sequences using the linsi algorithm in MAFFT (Katoh & Frith, 2012) and our custom LEP62 substitution matrix (see later). We built a preliminary moth + min-one-butterfly database for use with HAMSTR (version 13.2.2; [Ebersberger *et al*., 2009]) consisting of 7042 profile hidden Markov models (pHMMs) derived from the MAFFT alignments, and a BLAST database containing the complete proteome of *B. mori*, our designated reference taxon as required by HAMSTR.

Upon visual inspection, some of the amino acid alignments in the moth + min-one-butterfly database appeared to be suboptimal. To avoid including such alignments, we first used T-COFFEE (Notredame *et al*., 2000) to calculate a similarity score for each alignment in the database, finding a median alignment similarity score of 81.6%. We then removed alignments (i.e. orthologous groups) with a similarity score less than 70%, which roughly corresponded to the lowest quartile of alignment similarity scores. This left 5283 orthologous groups in the moth + min-one-butterfly database.

As we were performing this study, two additional Lepidoptera genomes became available [*Plutella xylostella* L. (Yponomeutoidea) and *Manduca sexta* L. (Bombycoidea)], although not through Ensembl. We incorporated these new taxa into our nuclear gene database. The *Manduca sexta* genome data was obtained from Manduca Base (http://agripestbase.org/manduca; retrieved late January 2014), and consisted of 27 633 transcripts (CDS regions extracted from the original genome/gff3 file using gffread). In the case of *Plutella xylostella*, two groups were sequencing the genome independently. The Japanese group made their genome sequence available through KONAGAbase (Jouraku *et al*., 2013), and the Chinese group made theirs available through DBM-DB (Tang *et al*., 2014). The data from KONAGAbase consisted of a putative gene set that was the result of combining their genome and transcriptome gene annotations (32 800 sequences) with a putative 'unknown' gene set (39 781 sequences). The data from DBM-DB consisted of the coding sequence associated with their genome-based gene predictions (18 073 sequences), together with all 'unigenes' from their transcriptome data (171 262 sequences). To choose between these alternatives, we combined the sequences from each data source (72 581 sequences for KONAGAbase and 189 335 sequences for DBM-DB) and ran each set of sequences against the moth + min-one-butterfly HAMSTR database. We found that the 'representative' sequences (i.e. the sequences that were the best match to each orthologous group in the database) were longer, on average, in the DBM-DB data than in the KONAGAbase data, and also slightly more numerous. Therefore, we used only the DBM-DB *Plutella* data in our analyses.

To add *Plutella* Schrank and *Manduca* Hübner to the moth + min-one-butterfly database we used HAMSTR, setting both the HMM search and the BLAST E-value cutoffs

to 1e−10. This yielded 9739 hits in the *Plutella* data (4809 unique orthologous groups) and 5593 hits in the *Manduca* data (4576 unique orthologous groups). We stipulated that in order to add a *Plutella* or *Manduca* hit sequence to an existing moth + min-one-butterfly orthologous group, the sequence needed to be at least half the length of the shortest sequence in the existing moth + min-one-butterfly orthologous group. Both the relatively stringent E-value and this minimum length criterion were an attempt to keep short, potentially spuriously matching sequences out of the database.

After addition of the *Plutella* and *Manduca* sequences, the orthologous groups in the HAMSTR database were realigned de novo using MAFFT as before. Following this, we used the T-COFFEE similarity statistic to evaluate the new alignments. The median alignment similarity score was 86.2%; once again, we removed alignments with a similarity score less than 70% (131 alignments), leaving 5152 orthologous groups in the moth + min-one-butterfly database. We did no further realignments after this point. The 5152-gene moth + min-one-butterfly database, together with gene identifiers, available annotations, and pHMMs for use in orthologue search (see below), is available in Dryad (doi:10.5061/dryad.hj278).

Orthology determination: identifying orthologues in our assemblies

To infer orthology, we used HAMSTR (version 13.2.2; [Ebersberger *et al*., 2009]), which in turn used BLASTP (Altschul *et al*., 1990), GENEWISE (Birney *et al*., 2004) and HMMER (Eddy, 2011) to search our assembled transcriptome data for translated sequences that matched a set of previously constructed amino acid gene models specific to Lepidoptera (the moth + min-one-butterfly database of 5152 nuclear genes).

In the first step of the HAMSTR procedure, substrings of assembled transcripts (translated nucleotide sequences) that matched one of the gene models in the database were provisionally assigned to the matching orthologous group. To reduce the number of highly divergent, potentially paralogous sequences returned by this initial search, we set the E-value cutoff defining a 'hit' to 1e−05 (the HAMSTR default was 1.0), and retained only the top-scoring quartile of hits. In the second HAMSTR step, the provisional hits from the HMM search were compared with a reference taxon (*B. mori*), for which both a genome and a transcriptome are available (Mita *et al*., 2004; Xia *et al*., 2009; Li *et al*., 2012), and retained only if they survived a reciprocal best BLAST hit test with *Bombyx*. In our implementation, we substituted FASTA (specifically, the FASTY program; Pearson & Lipman, 1988) for BLAST, and substituted a custom LEP62 substitution matrix (see later) for the more usual BLOSUM62 (Henikoff & Henikoff, 1992). We set the E-value cutoff for the FASTA search to 1e−05 (the HAMSTR default was 10.0). Amino acid sequences from our transcripts, once assigned to orthologous groups, were aligned using the addfragments option in MAFFT (Katoh & Frith, 2012) and our custom LEP62 substitution matrix, in which procedure the *Bombyx* sequences were considered the reference alignment to which the transcript fragments were added. The resulting amino acid alignments were then converted to the corresponding nucleotide alignments using a custom Perl script that substituted for each amino acid the proper codon from the original coding sequence.

Orthology determination: creating the LEP62 custom amino acid substitution matrix

A recent study showed the utility of using clade-specific amino acid substitution matrices in de novo orthology prediction for mollicute genomes (Lemaitre *et al*., 2011). As we perform amino acid alignments at several points in our own phylogenomic workflow, and these rely on a well-calibrated amino acid substitution matrix (usually BLOSUM62 by default), we hypothesized that these alignments would be improved if we used a substitution matrix derived from Lepidoptera-specific protein alignments.

We had initially constructed 7042 orthologous groups from Ensembl genome data. As part of our initial investigation, we calculated [using T-COFFEE (Notredame *et al*., 2000) and custom Perl scripts] that the average sequence identity of the aligned orthologous groups was 61.997%. To build the LEP62 matrix, we ran the scripts of Lemaitre *et al*. (2011); this package also included the BLOSUM program (Henikoff & Henikoff, 1992). We found that 86/200 entries differed between the LEP62 and BLOSUM62 matrices. More details can be found in Table S6.

In seeking to test the utility of the LEP62 matrix in similarity searches, we, like Lemaitre *et al*. (2011), could find no straightforward way to have BLAST use a custom substitution matrix. Instead we used the FASTA package (Pearson & Lipman, 1988), which readily accepted custom substitution matrices. Using a sample protein sequence from our transcriptome data, we performed five searches against the NCBI NR database with different combinations of alignment program and substitution matrix: BLAST + BLOSUM62; FASTA + BLOSUM62; FASTA + LEP62; SSEARCH + BLOSUM62; and SSEARCH + LEP62 (Table S7). We found that (i) the top two hits were the same in each search (*Bombyx* and *Danaus* Kluk sequences, respectively); (ii) SSEARCH produced better E-values than FASTA; and (iii) LEP62 produced better E-values than BLOSUM62. Here, 'better' is defined as providing more discrimination.

We then performed the same five searches using the *Bombyx* proteome as the database (Table S7). The top hit was the same in each search, and had a much lower E-value than any other hit. In this case the top hit was probably the only 'good' hit in the database. Once again, we found that our discriminatory power was highest with ssearch and the LEP62 matrix. After conducting these tests, we felt reasonably confident that using programs from the FASTA package in conjunction with the LEP62 matrix had the potential to improve workflow performance.

To test the efficacy of the LEP62 matrix in our workflow, we made incremental modifications to HAMSTR (version 13.1; [Ebersberger *et al*., 2009]) and ran our *Antaeotricha schlaegeri* (Zeller) ('Ant') RNA-Seq sample against the 7042-gene moth + min-one-butterfly database after each modification.

These modifications and the corresponding statistics generated from each HAMSTR run can be found in Table S8. With HAMSTR we used the FASTY program from the FASTA package instead of SSEARCH, despite the fact that in our previous tests SSEARCH performed best. This is because SSEARCH only supports DNA:DNA or protein:protein comparisons, whereas we needed to search a protein database with a translated nucleotide query. Overall, there was a slight increase in the number of hits when LEP62 was used, as we would expect. The only modification we made to HAMSTR that does not strictly pertain to the LEP62 matrix involved running pseg (Wootton & Federhen, 1996) on the *Bombyx* BLAST database to mask low-complexity regions. This procedure resulted in a substantial decrease in the number of hits (Table S8); presumably, we lost hits to low-complexity regions that were not desirable in the first place. We later discovered another place to use LEP62, namely, in the call HAMSTR makes to GENEWISE (Birney *et al*., 2004). This change, however, had only a minor effect on HAMSTR search statistics.

Ultimately we wanted to characterize the impact of using the LEP62 matrix throughout the workflow on the outcome of phylogenetic analyses. Previously we had analysed a 16-taxon, 2884-gene data matrix to test an early version of the moth + min-one-butterfly database, using GARLI 2.0. (Zwickl, 2006). This matrix was constructed with an older version of HAMSTR (version 9; Ebersberger *et al*., 2009) that had none of the modifications just described. We rebuilt the matrix using the modified version of HAMSTR and repeated the phylogenetic analyses. The new 16-taxon, 2884-gene matrix had about 10% fewer residues than the previous one and was slightly more complete. This correlated with the statistics for the 'Ant' sample, for which the final total number of sequences was about 10% less than the starting total (Table S8). The new phylogenetic analysis, also computed with GARLI (110 best tree search replicates; 279 bootstrap replicates; five search replicates per bootstrap replicate), yielded the same topology as the previous one, with comparable bootstrap support. While use of the LEP62 matrix did not noticeably improve the phylogenetic results, neither did it worsen them. Given that using an amino acid substitution matrix specific to the study group makes sense a priori, we used the LEP62 matrix for all analyses in this study. The LEP62 substitution matrix can be found in Dryad (doi:10.5061/dryad.hj278).

Paralogy filtering and data matrix construction

To screen for possible paralogues remaining among the 5152 nuclear genes in our dataset, we constructed a maximum likelihood (ML) gene tree, based on all nucleotides unpartitioned, for each orthologous group that was represented in at least 21 of our 28 taxa (75%). We used all sequences within each taxon that were assigned to that orthologous group. Initial orthology assignment often yields multiple sequences for individual taxon/locus combinations. This intraspecies variation can reflect the presence of multiple orthologues, heterozygosity, alternatively spliced transcripts, paralogy (including inparalogs; Sonnhammer & Koonin, 2002), and sequencing errors, among other possibilities. Gene tree construction used all of these variants. Each gene tree was a 50% majority-rule consensus of 100 bootstrap replicates.

We then provided the gene trees as input to PHYLOTREEPRUNER (Kocot *et al*., 2013). If the sequences from any one taxon formed a polyphyletic group supported by bootstrap of 80% or more, the program pruned that gene tree to the maximal subtree in which the nonpolyphyly criterion was met for all taxa. Gene trees were constructed for only 3264 of the 5152 orthologous groups (≈63%), as the others had fewer than 21 taxa. PHYLOTREEPRUNER pruned 1369 of the 3264 gene trees (≈42%) to some extent. Pruned gene trees were then eliminated if they contained fewer than 21 taxa, which was the case for 1052 trees.

At this point in the workflow there are still multiple sequences per taxon/gene combination. These need to be reduced to a single sequence for phylogenetic analysis. We previously evaluated two different approaches, 'representative' (Ebersberger *et al*., 2009) and 'consensus' (Bazinet *et al*., 2013), for reducing this variation to a single sequence, as required for phylogenetic analysis. As 'consensus' slightly outperformed 'representative' in a previous study (Bazinet *et al*., 2013), in this study we used only the consensus procedure, which uses degeneracy coding where necessary to combine information from all variant sequences into a single sequence for inclusion in the phylogenetic data matrix.

Following application of the paralogy filter and the consensus procedure, the 2212 surviving putative orthologue alignments were concatenated, adding gaps for missing data as necessary using a custom Perl script. The numbers of genes obtained for each taxon and their mean sequence length are shown in Table S4. The data were analysed both under the degen1 coding of Regier *et al*. (2010; version 1.4), which degenerates all synonymous differences using ambiguity coding (degen1), and with all nucleotides included unaltered and partitioned by codon position (nt123 partitioned). For these analyses, we removed sites not represented by sequence data in at least four taxa. Size and completeness statistics for the two paralogy-filtered matrices of 2212 nuclear genes are given in Table S5. The individual gene alignments and gene identifiers/available annotations for the 28-taxon, 2212 OG analyses, and the concatenated data matrices, can be found in Dryad (doi:10.5061/dryad.hj278).

Phylogenetic analysis

Our phylogenetic analyses used the ML criterion as implemented in RAXML version 8.2.3 (Stamatakis, 2014). We used a general time-reversible substitution model (GTR; Tavaré, 1986) with a rate heterogeneity model with a proportion of invariant sites (+I; Hasegawa *et al*., 1985) and the remainder with a gamma distribution (+G; Yang, 1993), and RAXML default settings, including the default rapid hill-climbing algorithm and parsimony starting trees. Each analysis consisted of ten search replicates plus 100 bootstrap (BP) replicates with one search replicate each. We used DENDROPY (Sukumaran & Holder, 2010) to plot BP values onto the best tree. The phylogenetic analyses were performed using the computing resources available at the

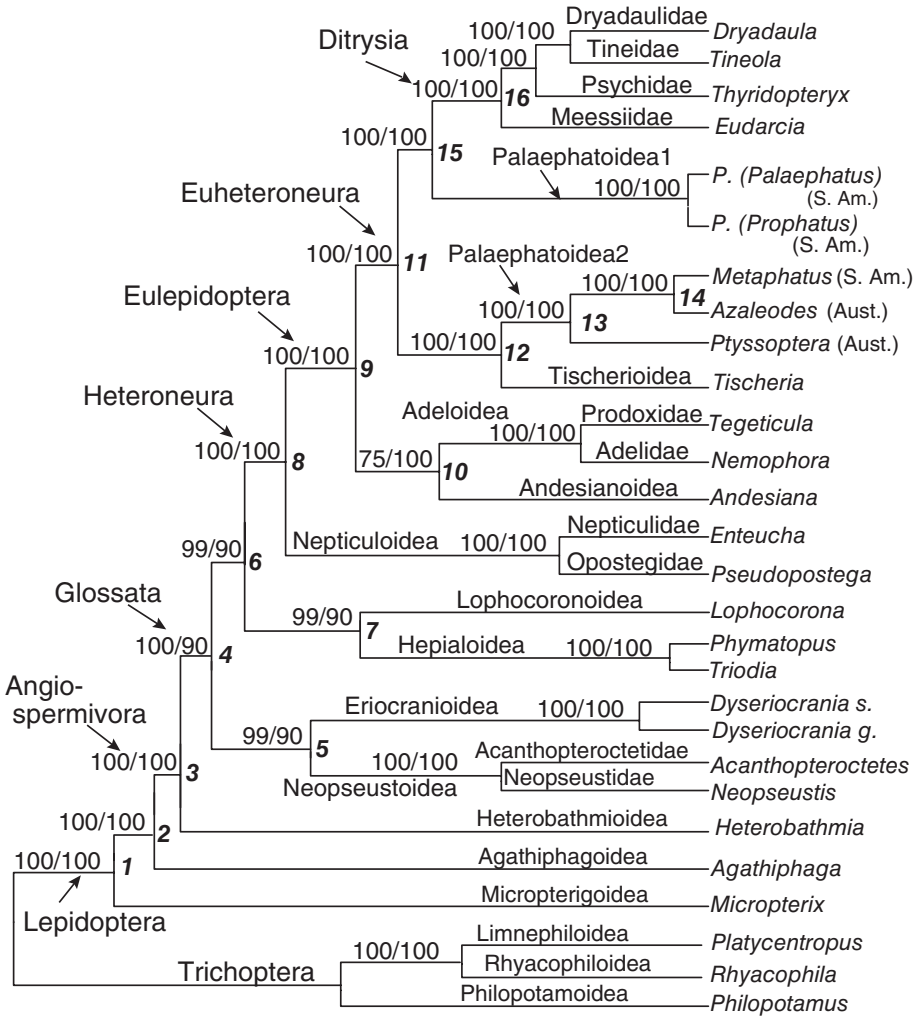

Fig. 2. Maximum likelihood estimate of phylogenetic relationships among nonditrysian Lepidoptera from 2212 gene sequences obtained by RNA-Seq. Best tree obtained from ten RAXML searches under a GTR + I + G model for degen1 and nt123 partitioned by codon position, both of which found the same topology. Bootstrap values are presented above branches. Nodes within Lepidoptera are numbered (to the right of the node) for purposes of discussion. Nomenclature follows Regier *et al*. (2015a,b). P. (Palaephatus) = *Palaephatus* subgenus *Palaephatus*; P. (Prophatus) = *Palaephatus* subgenus *Prophatus*.

University of Maryland, College Park. Each search replicate ran on a single node using 20 cores and 128 GB RAM, and required several hours of runtime.

From only the description we present here, it might appear that our data exploration and tree search effort were rather limited. The workflow and phylogenetic results just described, however, are only the final step in an extensive series of experiments that explored the behaviour of the data under a wide variety of analytical conditions (see Discussion). The preliminary analyses and results are described in File S1.

Results and discussion

Figure 2 shows the maximum likelihood tree inferred under degen1 coding (Regier *et al*., 2010), together with bootstrap values for both degen1 and nt123 partitioned. All bootstrap values for degen1 are 100% except for three that are 99% and one that is 75%. The same topology holds under nt123 partitioned, with all 100% bootstrap values except that four nodes have BP = 90%. It appears that both synonymous and nonsynonymous change strongly support this topology, which is shown for degen1 as a phylogram in Figure S1. Figure 3 shows relationships among the subfamilies, simplified from Fig. 2, together with representative images for all of the subfamilies. Further illustrations of all the nonditrysian families can be found in Regier *et al*. (2015a).

We now compare the present results with those from other recent molecular studies, and review their implications for our understanding of relationships among the nonditrysian superfamilies. Our treatment proceeds from the bottom to the top of the tree in Fig. 2, referring to the node numbers therein.

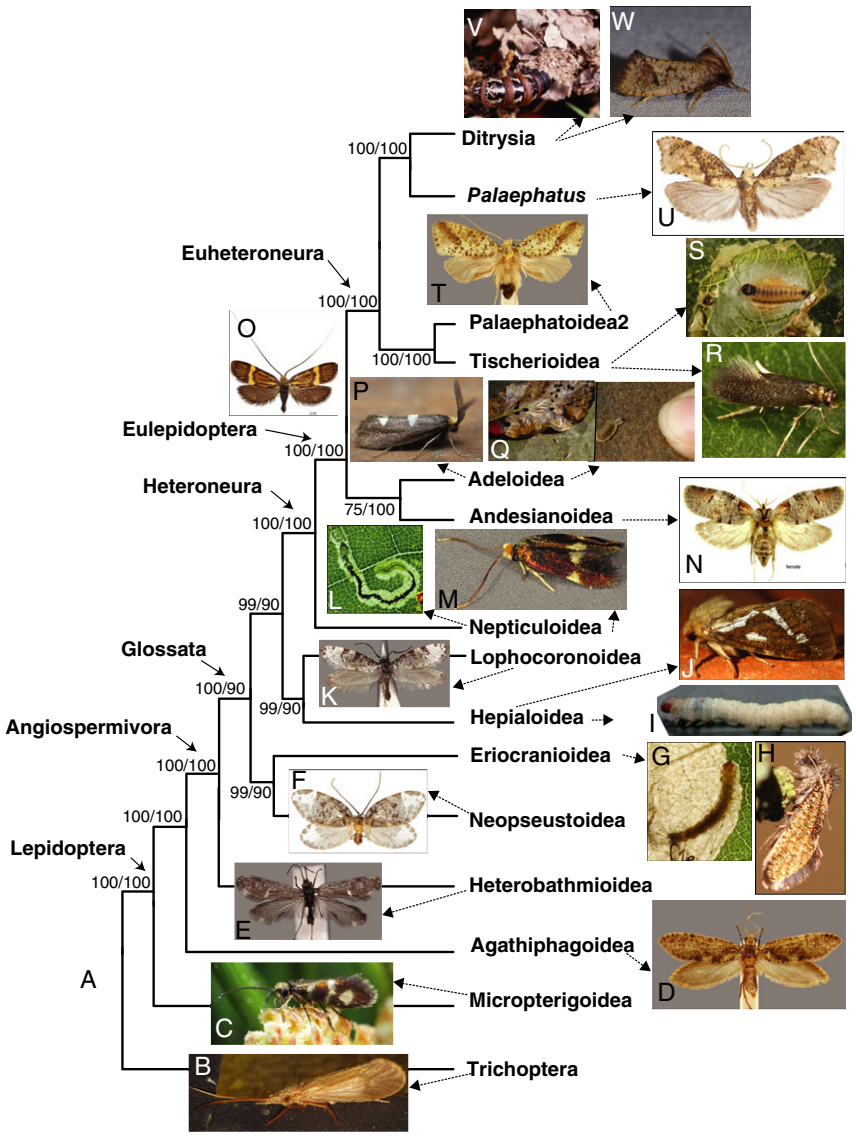

Fig. 3. Summary estimate for relationships among nonditrysian superfamilies, with bootstrap support, simplified from Fig. 2. Inset panels provide representative images for each superfamily. Format for panel legends: superfamily; family (if superfamily not monobasic); genus, species (if known); approximate (forewing) length if available; image author and/or source (for images taken from the web); licence code and link to source. Key to source and licence codes: CC, creative commons; A, attribution (only); x.x, version number for creative commons licence; G, generic; WK, Wikipedia; PD, public domain. Specifications for all of the creative commons licences can be found at https://creativecommons.org/licenses. (A) The phylogeny; (B) Trichoptera; D. Hobern; WK; CCA2.0G; (C) Micropterigoidea; *Micropterix aureoviridella* (Höfner); 3.5 mm; M. Kurz; WK; CCA2.0G; (D) Agathiphagoidea; *Agathiphaga vitiensis* Dumbleton; 10 mm; T. J. Simonsen; (E) Heterobathmioidea; *Heterobathmia pseuderiocrania* Kristensen & Nielsen; 4 mm; N. P. Kristensen; (F) Neopseustoidea; Neopseustidae: *Neopseustis meyricki* Hering; 10 mm; (G) Eriocranioidea; *Dyseriocrania subpurpurella* (Haworth); 5 mm, Svdmolen; WK; CCA2.5G; (H) Eriocranioidea; *Eriocrania semipurpurella* (Stephens); <5 mm; Charley Eiseman; (I) Hepialoidea; Hepialidae: *Korscheltellus lupulinus* (L.); 30 mm; Jeffdelonge; WK; ©entomart; (J) Hepialoidea; Hepialidae: *Korscheltellus lupulinus* (L.); 30 mm; Jeffdelonge; WK; CCA2.0G; (K) Lophocoronoidea; *Lophocorona pediasia* Common; 5 mm; N.P. Kristensen; (L) Nepticuloidea; Nepticulidae: *Stigmella aceris* (Frey), larva in mine; Gyorgy Csoka; Hungary Forest Research Institute, Bugwood.org; CCA3.0; (M) Nepticuloidea; Nepticulidae: *Bohemannia quadrimaculella*; 3 mm; Janet Graham; WK; CCA2.0G; (N) Andesianoidea; *Andesiana lamellata* Gentili; (O) Adeloidea; Adelidae: *Nemophora bellela* (Walker); 9 mm; (P) Adeloidea; Incurvariidae: *Incurvaria masculella* (Denis & Schiffermüller); 6 mm; D. Hobern; WK; CCA2.0G; (Q) Adeloidea; Heliozelidae; *Antispila nysaefoliella* Clemens, last-instar larvae in oval cases cut from *Nyssa* leaf; 3 mm; (R) Tischerioidea; *Coptotriche angusticollela* (Duponchel); 4 mm; Gyorgy Csoka, Hungary Forest Research Institute, Bugwood.org; CCA3.0; (S) Tischerioidea; *Tischeria ekebladella* (Bjerkander), larva; Gyorgy Csoka, Hungary Forest Research Institute, Bugwood.org; CCA3.0; (T) Palaephatoidea2; *Azaleodes micronipha* Turner; 10 mm; T. J. Simonsen; (U) Palaephatoidea1; *Palaephatus falsus* Butler; 11 mm; (V) Tineoidea; Psychidae: *Thyridopteryx ephemeraeformis* (Haworth); 50 mm; Gerald J. Lenhard, Louisiana State University, Bugwood.org; CCA3.0; (W) Tineoidea; Tineidae: *Acrolophus texanella* Chambers; 10 mm; A. Reago & C. McClarrene; CCA2.0G. [Colour figure can be viewed at wileyonlinelibrary.com].

Earliest divergences in the Lepidoptera

The basal phylogenetic split within Lepidoptera has been a notable point of uncertainty. In the morphology-based hypothesis of Kristensen (2003; Fig. 1A), Micropterigidae and then Agathiphagidae branch off successively from the remaining lepidopterans. This hypothesis was strongly supported by an early molecular study (Wiegmann *et al*., 2000) and a combined analysis of 18S rDNA data and morphology (Wiegmann *et al*., 2002). In contrast, the grouping of Micropterigidae + Agathiphagidae was favoured with weak to moderate support in 19-gene studies by Regier *et al*. (2013, 2015a), and with moderate to strong support by Kristensen *et al*. (2015), who combined the nonditrysian data from Regier *et al*. (2013) and Mutanen *et al*. (2010). While this history demonstrates the existence of inter-gene conflict, the present result suggests that the preponderance of molecular signal strongly favours the morphological hypothesis, under which Micropterigidae alone are sister group to the remaining Lepidoptera (Fig. 2, node 2). This finding was constant across all analyses of the present dataset (see later).

Corroborating previous molecular studies, our results strongly and invariably support the clade termed Angiospermivora by Regier *et al*. (2015a; Fig. 2, node 3), for which morphological evidence has also been strong (Fig. 1A, B; review in Regier *et al*., 2015a). We also corroborate (but see caveat later) the clade Glossata (Fig. 2, node 4), defined by origin of the sucking proboscis found in the vast majority of extant Lepidoptera, for which there is extensive previous support.

Morphological and molecular forms of evidence have been in conflict, however, on basal relationships within the Glossata (see Regier *et al*., 2015a for detailed review of the morphological evidence). In the morphological hypothesis (Fig. 1A), the first split separates Eriocranioidea from the clade Coelolepida. The basal divergence within Coelolepida then separates Acanthopteroctetidae from the rest, followed by the divergence of Lophocoronidae from the clade Myoglossata, defined by the origin of intrinsic muscles within the proboscis. Myoglossata are then divided basally into Neopseustoidae and the clade Neolepidoptera. Finally, the Neolepidoptera divides basally into the clade Mnesarchioidea + Hepialoidea (=Hepialoidea sensu lato of Regier *et al*., 2015a) versus the clade Heteroneura, which is strongly supported in all previous molecular and morphological studies.

Previous molecular phylogenies have departed strongly from this morphological arrangement while also leaving multiple nodes weakly supported. Regier *et al*. (2013, 2015a) (Fig. 1) reported monophyly, though with weak support, for a definition of Neopseustoidea that includes Acanthopteroctetidae (Fig. 1B), while Kristensen *et al*. (2015) found very strong support for monophyly of Neopseustoidea in this sense if the newly discovered family Aenigmatineidae is also included. This grouping renders the clade Myoglossata polyphyletic (Fig. 1A). Further, Regier *et al*. (2013, 2015a) and Kristensen *et al*. (2015) found very strong support (Fig. 1B) for the grouping of Lophocoronidae with the expanded Hepialoidea of Regier *et al*. (2015a). This pairing renders Neolepidoptera polyphyletic. Previous molecular evidence, in sum, divides Glossata into four strongly supported clades: (i) Eriocranioidea; (ii) Neopseustoidea sensu lato; (iii) Lophocoronoidea + Hepialoidea sensu lato; and (iv) Heteroneura. However, relationships among these four are very weakly supported in all previous molecular studies.

The current study samples one or more families from each of the four main glossatan clades, and finds strong support for yet another arrangement of these (Fig. 2). The first lineage to branch off from the rest in our tree consists of Neopseustoidea sensu lato (strongly corroborated here) plus Eriocranioidea (Fig. 2, node 5). The pairing of these superfamilies has not been previously proposed, and there are no obvious candidate synapomorphies. The second main split within Glossata in our tree (Fig. 2, node 6) separates Heteroneura from a clade consisting of Lophocoronoidea + Hepialoidea. The grouping of these superfamilies was also strongly supported in Regier *et al*. (2015a). This new hypothesis contradicts monophyly for Coelolepida as well as Myoglossata and Neolepidoptera, and possible morphological support for it has not been explored. On the other hand, it contradicts no strong previous molecular grouping and is the only arrangement so far with strong bootstrap support.

While we thus present a reasonable working hypothesis for basal glossatan relationships, we regard this region of our tree, especially nodes 5–7, to be distinctly less reliable than the rest, for several reasons. These are the only nodes in the tree to have less than 100% bootstrap support in both degen1 and nt123 partitioned analyses despite the massive number of loci. In addition, two of these nodes (5 and 6) have no prior support from either molecular or morphological data. (In contrast, node 7, Lophocoronidae + Hepialoidea, is strongly supported in Regier *et al*., 2015a.)

Further evidence comes from the numerous preliminary analyses we conducted before arriving at the workflow presented here. Those explorations, described in File S1, treated, among other variables, methods of alignment filtering, choice of phylogenetic software, taxon sampling, data matrix completeness, different implementations of PHYLOTREEPRUNER (Kocot *et al*., 2013), phylogenetic gene selection (Chen *et al*., 2015) and fraction of taxa represented for each gene. The current workflow was ultimately chosen because it gave the most consistent and strongly supported results that did not reject monophyly for Glossata, one of the most securely established clades in all of Lepidoptera (Kristensen & Skalski, 1998; Regier *et al*., 2013, 2015a). Across these analyses, most of the tree remained constant with 100% bootstrap support. Most notably, the monophyly and internal phylogeny of Heteroneura were invariant (with one minor exception), with bootstrap support nearly always 100%. In sharp contrast, topology and bootstrap support among the early-diverging lineages of Glossata (and even monophyly of Glossata itself) varied greatly, sometimes showing 100% bootstrap support for one or more groupings contradicting those in Fig. 2. Conflicting results of this kind, which probably reflect conflicts among genes, reinforce the argument (e.g. Salichos & Rokas, 2013) that, in phylogenomics, high bootstrap support is necessary but not sufficient evidence for drawing strong conclusions. Thorough exploration of tree and character space is needed.

Heteroneuran relationships and the position of Ditrysia

In contrast to those among homoneurous glossatans, the relationships among major lineages of Heteroneura (and monophyly of the latter; node 8) were invariable in our present results, with nearly always 100% bootstrap support. We corroborate previous strong evidence for successive origin of the clades Eulepidoptera (node 9) and Euheteroneura (node 11). We also find 100% bootstrap support (though under nt123 partitioned only) for a sister-group relationship of Andesianoidea to Adeloidea (node 10), a consistent grouping in molecular studies.

Within the Euheteroneura, our results very strongly resolve a position for the enormous clade Ditrysia (>98% of the Lepidoptera), a hitherto incompletely solved problem (Regier *et al*., 2015a). We strongly corroborate, with increased gene and taxon sampling, a previous finding of paraphyly for Palaephatidae (Regier *et al*., 2013, 2015a). The two Australian palaephatid genera, now joined by the South American *Metaphatus* Davis, are invariably grouped with Tischeriidae (Fig. 2, node 12). A subset of the South American palaephatids, now represented by both subgenera of *Palaephatus* Butler, are very strongly and invariably supported as sister group to the Ditrysia. Conflicting previous molecular findings on this question (Regier *et al*., 2015a) demonstrate the presence of disagreement among genes, but it now appears that the great preponderance of molecular evidence places *Palaephatus* and close relatives as sister group to the Ditrysia, to the exclusion of the remaining palaephatids plus Tischeriidae. This finding should be useful in future attempts to reconstruct the ground plan of Ditrysia in detail, in search for clues as to the causes of the spectacular ditrysian radiation.

Our findings are consistent with Nielsen's (1987) proposal of a basal split within palaephatids between an entirely South American clade containing *Palaephatus* and *Prophatus* (which he treated as a separate genus) and a clade containing both South American and Australian genera, including *Metaphatus* (South America) and *Azaleodes* Turner (Australia), both sampled here. Nielsen, however, did not consider the possibility that Palaephatidae might not be monophyletic. In the future we will re-examine both the morphological and molecular evidence on all the taxa now placed in Palaephatidae, including those not sampled here. The goal of that study, which is beyond the scope of the present work, will be to determine where each taxon falls in the phylogenetic dichotomy found here, and whether and how the family classification of nonditrysian Euheteroneura needs to be revised.

In summary, we believe that Fig. 2 presents the most credible, most strongly supported hypothesis of nonditrysian relationships to date. That hypothesis, summarized at the superfamily level in Fig. 3, is almost entirely consistent with strong previous molecular evidence and is at least as consistent with morphological evidence as any previous molecular hypothesis. Most of the tree, including the position of Ditrysia, has 100% bootstrap support that is stable to wide variation in the details of data matrix assembly and analysis. The chief remaining uncertainty about this dataset concerns the basal splits within Glossata, which have been consistently problematic in molecular studies. Our hypothesis for early glossatan divergences implicitly assumes monophyly for Glossata (for which prior evidence seems conclusive), because we rejected informatic workflows that did not lead to its recovery. It is possible that other analyses not explored here could more definitively support or reject the lower glossatan relationships we propose, which we regard as the most plausible to date but still provisional.

Supporting Information

Additional Supporting Information may be found in the online version of this article under the DOI reference: 10.1111/syen.12217

Figure S1. ML topology of Fig. 2 for the degen1 matrix, shown as a phylogram.

Table S1. Specimens sequenced and their classification.

Table S2. Summary statistics for RNA-Seq reads and assemblies.

Table S3. Comparative assembly statistics for Micropterix and Philopotamus.

Table S4. Per-taxon orthologous group and sequence inclusion statistics.

Table S5. Size and completeness of aligned data matrices from RNA-Seq.

Table S6. Substitution matrix statistics.

Table S7. Database search results.

Table S8. Modifications made to HaMStR.

File S1. Summary of preliminary analyses.

Acknowledgements

We are greatly indebted to the following generous colleagues for providing specimens used in this study: Glenn Cocking, Timothy P. Friedlander, Terry Harrison, Akito Y. Kawahara, Ebbe S. Nielsen, Luis Peña and David L. Wagner. Andreas Zwick, Niklas Wahlberg and an anonymous reviewer provided comments that led us to (we hope) greatly improve the manuscript. We also thank Andreas Zwick for follow-up discussion. We are grateful to Jesse Breinholt for performing the screen for cross-sample contamination. We much appreciate the provision of images by Charley Eiseman, (the late) Niels Kristensen, Jadranka Rota and Thomas Simonsen. This study builds on our long-term collaboration with Jerome Regier, who produced the bank of extracts and prior Sanger sequence data on which our taxon sampling plan and most of our sequencing were based. Financial support was provided by U.S. National Science Foundation awards DBI-0755048, DEB1355028 and DEB-1355023; the Hatch funds of the Maryland Agricultural Experiment Station; and a seed grant from the Department of Entomology, University of Maryland.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

AUTOADAPT (2014) [WWW document]. URL https://github.com/optimuscoprime/autoadapt [accessed on 1 April 2014].

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

Birney, E., Clamp, M. & Durbin, R. (2004) Genewise and genomewise. *Genome Research*, **14**, 988–995.

Birol, I., Jackman, S.D., Nielsen, C.B. *et al.* (2009) De novo transcriptome assembly with ABySS. *Bioinformatics*, **25**, 2872–2877.

Chen, M.Y., Liang, D. & Zhang, P. (2015) Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. *Systematic Biology*, **64**, 1104–1120.

Cunningham, C., Amode, M.R. & Barrell, D. (2014) Ensembl 2015. *Nucleic Acids Research*, **43**, D662–D669.

Davis, D.R. (1986) A new family of monotrysian moths from austral South America (Lepidoptera: Palaephatidae), with a phylogenetic review of the Monotrysia. *Smithsonian Contributions to Zoology*, **434**, 1–202.

Ebersberger, I., Strauss, S. & von Haeseler, A. (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, **9**, 157.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

FastQC (2014) [WWW document]. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [accessed on 1 April 2014].

Flicek, P.F., Amode, M.R., Barrell, D. *et al.* (2014) Ensembl 2014. *Nucleic Acids Research*, **42**, D749–D755.

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

Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Henikoff, S. & Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915–10919.

Hennig, W. (1953) Kritische bemerkungen zum phylogenetischen system der insekten. *Beiträge zur Entomologie. Sonderheft*, **3**, 1–85.

Hittinger, C.T., Johnston, M., Tossberg, J.T. *et al.* (2010) Leveraging skewed transcript abundance by RNA-seq to increase the genomic depth of the tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 1476–1481.

Jouraku, A., Yamamoto, K., Kuwazaki, S. *et al.* (2013) KONAGAbase: a genomic and transcriptomic database for the diamondback moth, *Plutella xylostella*. *Biomed Central Genomics*, **14**, 464.

Katoh, K. & Frith, M.C. (2012) Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*, **28**, 3144–3146.

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

Kobayashi, Y. & Ando, H. (1988) Phylogenetic relationships among the lepidopteran and trichopteran suborders (Insecta) from the embryological standpoint. *Zeitschrift für Zoologische Systematik und Evolutionsforschung*, **26**, 186–210.

Kocot, K.M., Citarella, M.R., Moroz, L.L. *et al.* (2013) PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics Online*, **9**, 429.

Kristensen, N.P. (1984) Studies on the morphology and systematics of primitive Lepidoptera (Insecta). *Steenstrupia*, **10**, 141–191.

Kristensen, N.P. (2003) Appendices. *Lepidoptera: Moths and Butterflies 2. Handbuch der Zoologie/Handbook of Zoology*, Vol. IV/36 (ed. by N.P. Kristensen), pp. 545–554. De Gruyter, Berlin and New York, New York.

Kristensen, N.P. & Skalski, A.W. (1998) Phylogeny and palaeontology. *Lepidoptera, Moths and Butterflies*, *Vol. 1: Evolution, Systematics and Biogeography*, *Handbuch der Zoologie/Handbook of Zoology*, Vol. 4 (ed. by N.P. Kristensen), pp. 7–25. De Gruyter, Berlin and New York, New York.

Kristensen, N.P., Scoble, M.J. & Karsholt, O. (2007) Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. *Zootaxa*, **1668** [Linnaeus Tercentenary Special Volume], 699–747.

Kristensen, N.P., Hilton, D.J., Kallies, A. *et al.* (2015) A new extant moth family from Kangaroo Island and its significance for understanding early Lepidoptera evolution (Insecta). *Systematic Entomology*, **40**, 5–16.

Lemaitre, C., Barre, A., Citti, C. *et al.* (2011) A novel substitution matrix fitted to the compositional bias in mollicutes improves the prediction of homologous relationships. *Biomed Central Bioinformatics*, **12**, 457.

Li, Y., Wang, G., Tian, J. *et al.* (2012) Transcriptome analysis of the silkworm (*Bombyx mori*) by high-throughput RNA sequencing. *PLoS ONE*, **7**, e43713.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.

Misof, B., Liu, S., Meusemann, K. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.

Mita, K., Kasahara, M., Sasaki, S. *et al.* (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Research*, **11**, 27–35.

Mutanen, M., Wahlberg, K. & Kaila, L. (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 2839–2849.

Nielsen, E.S. (1987) The recently discovered primitive (non-ditrysian) family Palaephatidae (Lepidoptera) in Australia. *Invertebrate Taxonomy*, **1**, 201–229.

Nielsen, E.S. & Kristensen, N.P. (1996) The Australian moth family Lophocoronidae and the basal phylogeny of the Lepidoptera-Glossata. *Invertebrate Taxonomy*, **10**, 1199–1302.

Notredame, C., Higgins, D.G. & Heringa, J. (2000) T-Coffee: A novel algorithm for multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–217.

Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Peters, R.S., Meusemann, K., Petersen, M. *et al.* (2014) The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. *Biomed Central Evolutionary Biology*, **14**, 52.

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: the barcode of life system (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Regier, J.C. & Shi, D. (2005) Increased yield of PCR product from degenerate primers with nondegenerate, nonhomologous 5' tails. *BioTechniques*, **38**, 34–38.

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

Regier, J.C., Mitter, C., Zwick, A. *et al.* (2013) A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PLoS ONE*, **8**, e58568.

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

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

Robertson, G., Schein, J., Chiu, R. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nature Methods*, **7**, 909–912.

Salichos, L. & Rokas, A. (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, **497**, 327–331.

Sonnhammer, E.L.L. & Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, **18**, 619–620.

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

Sukumaran, J. & Holder, M.T. (2010) DendroPy: a python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.

Tang, W., Yu, L., He, W. *et al.* (2014) DBM-DB: the diamondback moth genome database. *Database*, **2014**, bat087.

Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Wiegmann, B.M., Mitter, C., Regier, J.C. *et al.* (2000) Nuclear genes resolve Mesozoic-aged divergences in the insect order Lepidoptera. *Molecular Phylogenetics and Evolution*, **15**, 242–259.

Wiegmann, B.M., Regier, J.C. & Mitter, C. (2002) Combined molecular and morphological evidence on phylogeny of the earliest lepidopteran lineages. *Zoologica Scripta*, **31**, 67–81.

Wootton, J.C. & Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology*, **266**, 554–571.

Xia, Q.Y., Guo, Y.R., Zhang, Z. *et al.* (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science*, **326**, 433–436.

Xie, Y., Wu, G., Tang, J. *et al.* (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.

Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, **10**, 1396–1401.

Zwickl, D.J. (2006) *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD Thesis, The University of Texas at Austin.

Accepted 1 July 2016