Supporting Information
to

Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths

Akito Y. Kawahara, David Plotkin, Marianne Espeland, Karen Meusemann, Emmanuel F. A. Toussaint, Alexander Donath, France Gimnich, Paul B. Frandsen, Andreas Zwick, Mario dos Reis, Jesse R. Barber, Ralph S. Peters, Shanlin Liu, Xin Zhou, Christoph Mayer, Lars Podsiadlowski, Caroline Storer, Jayne E. Yack, Bernhard Misof, and Jesse W. Breinholt

Table of Contents

Supplementary Methods

1. Taxon sampling

Sequences from 186 species of Lepidoptera were used for our study, in addition to sequences of 17 outgroup species representing other holometabolous insect orders, totaling 203 species (Dataset S1). A total of 187 sequences are transcriptomes; previously-published genomes were used for the remaining 13 ingroup and three outgroup species were included. Some of these genomes were used to compile a reference ortholog set (see Section 5). Fifty-eight of the transcriptomes were newly sequenced by the 1KITE consortium, eleven were newly sequenced from specimens of the Florida Museum of Natural History, University of Florida (FLMNH, Gainesville, Florida, USA), and the remaining 118 were from previously-published studies (1-22). Published transcriptomes were added in order to increase taxon sampling across Lepidoptera. A full list of sequences with taxonomic information, sources for previously-published sequences, and GenBank Bioproject accession numbers, are provided in Dataset S1.

2. RNA extraction and sequencing

Nearly all specimens sequenced newly for this study were collected as adults, preserved in RNAlater (Qiagen, Maryland, USA), and stored at +4 °C or -80 °C until further processing (Dataset S2). Whenever possible, specimen vouchers for the newly-generated transcriptomes were kept at the FLMNH. For each specimen at the FLMNH, one pair of wings was removed and retained as a voucher, following the protocol of Cho *et al.* (23). For small specimens, the entire body was ground with a pestle and used for extraction of total RNA. For large-bodied specimens, only a piece of the thoracic tissue (usually flight muscles) was used. Detailed preservation methods for each specimen are provided in Dataset S2.

Total RNA extractions, mRNA isolation, fragmentation, and cDNA library construction for the 58 samples processed by the 1KITE consortium were performed using the protocols of Misof *et al.* (24) and Peters *et al.* (25). Eight of these samples were extracted using the TruSeq v2 kit and protocol (Dataset S1). Protocols of Kawahara and Breinholt (11) were used for extraction, fragmentation, and library construction of the eleven samples processed at the FLMNH.

Paired-end sequencing of the 58 1KITE transcriptome libraries was conducted through the Beijing Genome Institute (BGI, Shenzhen, Guangdong, China) on HiSeq 2000 or 2500 platforms (Illumina, San Diego, CA, USA) with read lengths of either 90 bp (for libraries prepared with the TruSeq kit) or 150 bp (for libraries not prepared with the TruSeq kit). The eleven FLMNH samples were single or paired-end sequenced on a HiSeq 2000 platform, with read lengths of either 100 or 150 bp, at the sequencing cores of the Florida State University (two samples), University of Missouri (two samples), and University of Wisconsin at Madison (seven samples). Approximately 2.5 gigabases of raw data per library were retained.

3. *De novo* transcriptome assemblies

The 58 1KITE transcriptomes were assembled following the protocol of Peters *et al.* (25). The eleven FLMNH transcriptomes were assembled following the protocol of Kawahara and Breinholt (11). Raw reads were assembled with SOAPdenovo-Trans-63mer (version 1.01) (26)., using five k-mer values (13, 23, 33, 43, 63) following the additive multiple-k assembly method of Surget-Groba and Montoya-Burgos (27). Redundant contigs from multiple k-mer assemblies were combined using CD-HIT-EST (28) and all sequences below 100 bp were removed with the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

The 58 assembled 1KITE transcriptomes are deposited under the 1KITE Umbrella Project (PRJNA183205) and at the NCBI Sequence Read Archive (SRA) and Transcriptome Shotgun Assembly Sequence Database (TSA). Raw reads of the eleven FLMNH transcriptomes are deposited under the project "Evolutionary history of butterflies and moths" (PRJNA522250) and at the NCBI SRA archive. Assemblies of the FLMNH transcriptomes can be found in Supplementary Archive 1.

The majority of previously published transcriptomes used in this study were provided as assemblies that had already been processed and screened by their respective authors (see Dataset S1 for sources of all previously published transcriptomes). RNASeq data for the additional samples from previous transcriptomic studies were either downloaded from the NCBI GenBank SRA database or had their corresponding RNASeq data generated following the protocol of Breinholt and Kawahara (5). These samples were assembled using the same methods as the FLMNH assemblies described above. Lastly, the transcripts of all protein coding genes from eleven published genomes deposited on LepBase (http://lepbase.org/) were added to the dataset. All assemblies derived from previously published transcriptomes can be found in Supplementary Archive 1.

4. Contaminant removal

The 1KITE transcriptome assemblies were checked for contamination (Dataset S3) following the protocol of Peters *et al.* (25). Partial sequences from 70 1KITE assemblies (58 new assemblies and 12 assemblies previously published in Pauli *et al.* (16)) were removed from the analysis after contamination filtering steps. Additionally, all transcripts that NCBI identified as possible foreign contaminants were removed when submitting the assemblies to the NCBI

Transcriptome Shotgun Assembly (TSA) database. Dataset S3 lists the number of nucleotides removed from each assembly, which ranged from 16,696 (0.04% of the unfiltered assembled transcriptome) to 853,228 (2.49% of the unfiltered assembled transcriptome).

We initially detected contamination in three samples sequenced on Illumina lanes. As a secondary means of detecting contamination, fast approximate Maximum Likelihood (ML) phylogenetic analyses were conducted with FastTree2 (29), and preliminary ML tree inferences and bootstrap replicates were obtained in IQ-TREE (30) using methods described later in Section 8, in order to assess the phylogenetic placement of particular taxa. Three contaminated samples were identified and their 1KITE internal library identifier names are: *Scythris scopolella* (Linnaeus) (Gelechioidea: Scythrididae; WHANIsrmTMAGRAAPEI-15), an undetermined species in the family Nolidae (RINSInITDWRAAPEI-55), and *Epipomponia nawai* (Dyar) (Zygaenoidea: Epipyropidae; WHANIsrmTMCHRAAPEI-56). The scythridid sample had 1.3% cross-contamination not explicitly from this species. The nolid had the full-length COI sequence from this taxon (pulled from the transcriptome), which was searched with BLASTN against all Lepidoptera in the BOLD database (31): the COI sequence for this sample did not match with any lepidopteran species with high confidence. Because the identification of this specimen was questionable, and the sample could not be confidently placed in any clade in the phylogenetic tree in preliminary analyses, it was removed from subsequent analyses (see below). A similar COI search was performed with BLASTN for *E. nawai*, and although some fragments had a high degree of similarity to Lepidoptera, other fragments had 99% similarity to the barcode of the hybosorid beetle genus *Hybosorus* MacLeay. Furthermore, a preliminary rogue taxon analysis in the program RogueNaRok (32) revealed that *E. nawai* should not be included in the dataset. *Epipomponia nawai* is a parasitic moth (33), so although the sample sequence may be correct,

the possibility of contamination cannot be excluded. In preliminary phylogenetic analyses, the three abovementioned species of concern were placed in clades that were very different from what is expected based on their classification. Consequently, all three contaminated samples were removed from all subsequent analyses.

5. Ortholog set generation and identification of orthologous transcripts

An ortholog set was compiled based on the OrthoDB v7 database (34), containing clusters of orthologous sequences such that orthologs had to be present and single-copy (i.e. copy number equal to 1) in the official gene sets (OGSs) of three reference species, *Bombyx mori* (Linnaeus) (OGS v2.0, see Dataset S4), *Danaus plexippus* (Linnaeus) (OGS v2.0; downloaded from http://monarchbase.umassmed.edu/resource.html) and *Tribolium castaneum* (Herbst) (OGS v3.0, see Dataset S4). The hierarchy was set to Aparaglossata (i.e. all Holometabola except Hymenoptera; see Peters *et al.* (25)), and *T. castaneum* was set as the outgroup. To exclude OGs that had more than one copy in other Lepidoptera, OGSs of four additional lepidopterans, as they were implemented in OrthoDB v7 (*Heliconius melpomene* (Linnaeus) (Nymphalidae), *Manduca sexta* (Linnaeus) (Sphingidae), *Melitaea cinxia* (Linnaeus) (Nymphalidae), *Plutella xylostella* (Linnaeus) (Plutellidae)), were screened, and the ortholog groups (OGs) that were either absent or only occurring in single-copy (i.e. copy number ≤ 1), were kept. This resulted in a set of 3,429 OGs in the final ortholog reference set that was subsequently used to identify OGs in our assembled transcriptomes. Finally, orthologs were downloaded from OrthoDB v7 as a tab-delimited table and fasta files. Additionally, the OGSs for the three reference species were downloaded from their respective genome databases (see above) as amino acids and corresponding nucleotides (CDS or transcript sequences). OGSs were modified with the Perl

script, *make-ogs-corresponding.pl*, which is part of the Orthograph package beta4 (35). Specifically, this script made the following modifications: (*i)* Make sequence headers that correspond to the public gene identifiers in the OrthoDB-tab-delimited file. (*ii*) Remove sequences with ambiguous or duplicate headers. (*iii*) Remove protein sequences with no equivalent nucleotide identification, and vice versa. (*iv*) Make header names corresponding to amino acids and nucleotide datasets. (*v*) Remove terminal stop codons from the protein OGSs of all species. The ortholog set (table of orthologs and full official gene sets used for the reciprocal BLAST step on proteins and CDS) used in Orthograph is provided as Supplementary Archive 2.

All transcriptome assemblies were searched for the 3,429 single-copy protein-coding genes using Orthograph beta4 (35) with the following settings: hmmsearch-score-threshold = 10, blast-score-threshold = 10, hmmsearch-evalue-threshold = 1e-05, blast-evalue-threshold = 1e-05, max-blast-searches = 100, blast-max-hits = 100, minimum-transcript-length = 30, orf-overlap-minimum = 0.5, extend-orf = 1, strict-search = 1 with *Bombyx mori* and *Danaus plexippus* as the reference sequence taxa, substitute-u-with = X. Orthograph results were summarized with a Perl script which is part of the Orthograph package. The three reference species were kept and putative internal stop codons and some amino acids were removed (i.e. the amino acid selenocysteine (U) was replaced with “X” or the equivalent coding nucleotide sequence replaced with “NNN". The number of average orthologs identified in Orthograph across the 200 taxa was 2,564 (median: 2,592; minimum: 602; maximum: 3,232). Of the 200 taxa, six had less than half of the 3,429 OGs and 74 taxa had 80% or more of the 3,429 OGs. Orthograph identified a total of 3,427 of the 3,429 orthologs across all included taxa (see Dataset S4).

6. Alignment, alignment masking, dataset generation

Amino acid transcripts were aligned in MAFFT v7.294 (36) with the L-INS-I algorithm. Amino acid alignments were screened for outliers and refined, with the remaining outliers removed from further downstream analyses, using the protocols and scripts provided in the supplementary material of Misof *et al.* (24). After outlier removal, alignment columns that only contained gaps were deleted (24). Corresponding nucleotide alignments were generated using a modified version of Pal2Nal v14 (37) with amino acid alignments as blueprints. Aliscore v1.2 (38, 39) was used to identify ambiguous and randomly similar aligned sections in the amino acid alignments. Aliscore was invoked with a custom –r 10^{27} option, in order to compare all sequence pairs in each sliding window (default size), with a special scoring approach for gap-filled amino acid sites (option -e). Apart from these options, default parameters were used. Using a custom Perl script, lists were generated for the nucleotide alignments with corresponding codons to be excluded. Alicut v2.0 (40) was used to delete ambiguously aligned amino acid and nucleotide sites that were identified by Aliscore. To reduce the overall amount of missing data, all loci with less than 70% taxon coverage were removed, resulting in 2,380 OGs. These OGs were further screened by estimating uncorrected p-Hamming average distance to identify fast-evolving genes, using the Python script p-distance_script.py (https://github.com/lteasdale). p-Hamming distances were sorted by size and the largest change (Δ) between consecutive estimated distances between 0.03 and 0.4 (max) was used to choose an upper limit (0.3214) for OGs included in downstream analyses. This resulted in 2,098 OGs, excluding particularly divergent OG alignments that might include non-orthologous sequences. Individual OG alignment files were retained for subsequent multispecies coalescent analyses, and supermatrices for both amino acid and nucleotide datasets were generated using FasConCat-G v1.0 (41), along with corresponding files containing

information on the gene boundaries of the masked multiple sequence alignments (Supplementary Archive 3).

7. Data partitioning and model selection

PartitionFinder2 v2.1.1 (42) and RAxML v8.2.11 (43) were used to merge OGs into an optimal partitioning scheme for the concatenated amino acid and nucleotide datasets. For the amino acid dataset, model selection was restricted to five amino acid substitution models available in RAxML. These were: BLOSUM62 (44), DCMUT (45), JTT (46), LG (47), WAG (48). We also included the protein mixture model, LG4X (49) that accounts for FreeRate heterogeneity, as this mixture model had the best fit in previous phylogenomic studies of insects (25, 50). We allowed for testing of rate heterogeneity type +G (51) both with empirical rates (+F) and without empirical rates, and with four rate categories (default setting in RAxML). The FasConCat-G output file with gene boundary information was used as input and we used the 'branchlengths = linked' option, the AICc for model selection (52), and the rcluster search algorithm (53). Additional options and parameters include: --rcluster-max 10000 --rcluster-percent 50 --weights 1,1,0,1 -q -p 100 --all-states --min-subset-size 100. The OGs were then merged into partitions.

The data partitioning analysis was conducted prior to IQ-TREE, allowing for the merging of OGs into partitions with linked branch lengths, which was done in ModelFinder (54). We subsequently re-estimated best-fit models using ModelFinder in IQ-TREE v1.5.5 (30, 55) with the protein mixture model LG4X and all available amino acid substitution models of nuclear coding genes. We used the following commands to re-estimate models: -st AA -nt 20 -m MF -msub nuclear -mrate E,I,G,I+G,R -cmin 2 -cmax 10 -madd LG4X -safe. We adjusted partition

boundaries and reordered the amino-acid superalignment with custom-made Perl scripts so that the OGs that had been merged into one partition were ordered consecutively, which was necessary for later analyses (see Section 9). The best partitioning scheme and models are provided in Supplementary Archive 3 and Supplementary Archive 4.

Best-fit models were also selected for the individual amino acid alignments for each gene using IQ-TREE v1.6.10 (30), for subsequent multispecies coalescent analyses. Model selection was constrained to the same six substitution models used for the PartitionFinder analysis of the concatenated dataset. Multiple models of rate variation, including the FreeRate model, were estimated for each substitution model. The selected models for individual amino acid alignments are provided in Supplementary Archive 5.

Synonymous signal was removed from the concatenated nucleotide dataset with Degen v1.4 (56, available from http://www.phylotools.com/ ptdegendownload.htm), which finds all sites with synonymous substitutions and replaces the corresponding nucleotides with IUPAC ambiguity codes. Elimination of synonymous signal has been shown to be advantageous when conducting phylogenetic analyses focused on estimating deep-level relationships (e.g. 5, 56-59). The "degen1" nucleotide dataset used the same partitioning scheme as the amino acid partitioning scheme, but the boundaries of the merged partitions were adjusted to match a nucleotide dataset (Supplementary Archive 6). Models were re-estimated using ModelFinder in IQ-TREE v1.5.5 with the following commands: -st DNA -nt 20 -m MF -mrate E,I,G,I+G,R -cmin 2 -cmax 10 -safe.

8. Tree inference and branch support

Maximum likelihood (ML) tree inferences were conducted on the amino acid dataset and nucleotide degen1 dataset (Supplementary Archives 4, 6). For the amino acid dataset, 69 ML tree searches were performed in IQ-TREE v1.6.1 (30) using the partitioning scheme and models obtained previously (see Section 7). Since the analyses required large computational resources, tree searches were done simultaneously on the following computer clusters: PEARCEY HPC (CSIRO, Australia; 41 searches), Smithsonian Institution HPC (Washington, DC, USA; 14 searches), Zoological Research Museum Alexander Koenig HPC (Bonn, Germany; 12 searches), HiPerGator HPC (University of Florida, Gainesville, FL, USA; 2 searches). Parsimony-derived start trees and edge-proportional partition models (option -spp) were used for each search, allowing partitions to have different evolutionary rates. The tree with the highest log-likelihood was selected as the best tree, and its topology was recovered in 38 of the 69 tree-searches (Fig. S2) determined using Unique Tree (v1.9), provided by Thomas Wong. The remaining 31 tree-searches converged to the same tree topology, which differed slightly from the best (Fig. S3).

Multiple metrics of support were calculated to assess the reliability of the best ML tree. Using random starting trees, 121 non-parametric slow bootstrap replicates were generated for the amino acid dataset in IQ-TREE v1.6.1. To confirm that a sufficient number of bootstrap replicates had been performed, the *a posteriori* bootstrap criterion of Pattengale *et al.* (60) was used to check for bootstrap convergence as implemented in RAxML v8.2.11 (43) with the following settings: autoMRE, -B 0.03, --bootstop-perms=1000. The bootstrap convergence test was performed 10 times with different random seeds, and all tests revealed convergence after 50 replicates. Two additional metrics of support were calculated: the SH-aLRT test (61), calculated in IQ-TREE with 10,000 replicates (Fig. S4), and TBE support values (62) (Fig. S5), which were

calculated in BOOSTER v0.1.0 (62) using the 121 bootstrap trees from the IQ-TREE amino acid analysis. For these analyses, we used the best tree from IQ-TREE as the start tree. Support values for major nodes of all analysis are provided in Dataset S5. After calculating support values, RogueNaRok v1.0 (32) was used to check for rogue taxa, with the 121 bootstrap trees and the best tree provided as input and otherwise default settings. No rogue taxa were identified.

The same approach was used for tree inference with the nucleotide degen1 dataset with IQ-TREE v1.5.5. Fifty ML searches and 100 non-parametric slow bootstrap replicates were conducted using parsimony starting trees. Bootstraps were mapped onto the best ML tree (Fig. S6). SH-aLRT and TBE statistical supports were also calculated (Figs. S7, S8). Bootstrap convergence check and identification of rogue taxa were performed with the same methods described above for the amino acid tree searches; bootstrap convergence occurred after 50 replicates, and no rogue taxa were identified. All analyses of the degen1 dataset were performed on the PEARCEY HPC which only had version 1.5.5 of IQ-TREE available at the time of this study. The ML tree in Newick format and selected models are provided in Supplementary Archive 6.

After completing phylogenetic analyses on the concatenated datasets, we also performed tree reconstruction using a multispecies coalescent (MSC) approach on the amino acid dataset (Supplementary Archives 3, 6). For each locus, we generated two gene trees in IQ-TREE v1.6.10 (30): The best ML tree, out of 25 ML tree searches, and a consensus tree (*.contree) derived from 1000 UFBoot2 replicates (63). The MSC analyses were carried out using ASTRAL-III v5.6.3 (64) on both sets of gene trees. ASTRAL analyses typically utilize a set of best ML trees, but we chose to also perform an analysis with a set of consensus trees because the resulting species tree could potentially better reflect uncertainty (65). Local posterior probabilities (PP)

calculated in ASTRAL were mapped onto the trees (Figs. S9, S10) and PP for select nodes are presented in Dataset S5.

9. Topology tests

Three different four-cluster likelihood mapping (FcLM) analyses were performed on the amino acid dataset to assess the placement of particular Lepidoptera clades that have been the subject of interest in previous phylogenetic studies (e.g. 6, 11, 66). For these three hypotheses, we defined four taxonomic groups of interest: three monophyletic ingroups and one outgroup. Lists of species present in each group for each hypothesis are provided in Datasets S6-S8, and additional input and output files are provided in Supplementary Archive 7.

We generated optimized data subsets for each of the three FcLM analyses by including only those partitions of the amino-acid supermatrix that contained sequences of at least one representative from the four groups specified for testing each hypothesis (Datasets S6-S8). Using this approach, we aimed to identify signal undetectable by traditional branch support metrics in our inferred ML trees. In addition, we assessed whether a non-phylogenetic, and possibly confounding, signal was present in our dataset that could have affected our ML tree inference and FcLM results on 'original' (i.e. non-permuted) data. Various sources of possible confounding signal that might violate globally stationary, reversible and homogeneous (SRH) conditions (67, 68) have been described elsewhere (e.g. 24, 25, 69, 70). These sources include heterogeneous composition of amino acid sequences (among-lineage heterogeneity) and non-randomly distributed (missing) data. Therefore, we applied three different FcLM permutation approaches. In permutation I, all phylogenetic signal was removed. In permutation II, compositional heterogeneity was removed by randomly drawing amino acids using the

frequencies of the LG substitution matrix, but the distribution of missing data was left untouched. Permutation III included the features of permutation II, but also had random distribution of missing data (i.e. no phylogenetic signal, homogeneous composition, and missing data randomly distributed). For additional information on this strategy and rationale, see Supplementary Information in Misof *et al.* (24).

We used IQ-TREE v1.6.7 (30) to infer the support for each quartet and to allow the drawing of all possible quartets (option -lmap ALL; all four groups represented in each quartet). Quartet log-likelihoods (support for each quartet) were parsed into separate output files (option -wql). For the FcLM analyses with the 'original' non-permuted data, we kept the partition boundaries and substitution models as our previous ML analyses. We chose the partitioned approach and allowed partitions to have different evolutionary speeds (option -spp). For all permutation approaches, we kept the partition boundaries, but for all partitions we used the LG model for analyses and -q (edge-equal partition model) to avoid program crashes due to highly saturated partitions. For each analysis, quartets were automatically mapped by IQ-TREE in a 2D simplex graph. Quartets mapped onto area T1 (area 1), T2 (area 2) and T3 (area 3) show unambiguous support for the respective topology; T12 (area 4), T13 (area 6), T23 (area 5) show partially resolved quartets; quartets mapped onto T* (area 7) have star-like topologies that remain unresolved (see Fig. S11, Dataset S9). The results of these analysis were in large part congruent with our ML and ASTRAL analyses, but in some cases were not. Differences are noted in each of the sections below.

Placement of plume moths and false plume moths (Pterophoroidea+Alucitoidea)

The first hypothesis tested (Fig. S11A, Dataset S5) assesses the placement of the clade containing plume moths (Pterophoroidea) and many plume moths (Alucitoidea). According to Kawahara and Breinholt (11), Pterophoroidea is not placed within Obtectomera, but is instead sister to a non-obtectomeran clade in Apoditrysia. Alucitoidea is sister to Pterophoroidea in our ML analysis, though Alucitoidea was not included in the analysis of Kawahara and Breinholt (11). Our ML analysis of the amino acid dataset suggests that the clade Alucitoidea + Pterophoroidea (Alu_Pte) is in a monophyletic group with the obtectomeran Calliduloidea + Thyridoidea (Cal_Thy) and Gelechioidea (Gel), and with Alu_Pte sister to Cal_Thy, albeit with caveats. Monophyly of Alu_Pte + Cal_Thy + Gel, which contradicts the results of Kawahara and Breinholt (11), has high SH-aLRT, and TBE support values (Figs. 1, S4, S5), but low non-parametric bootstrap support (Fig. S2). Furthermore, the sister relationship between Alu_Pte and Cal_Thy is weakly supported (Figs. 1, S2), and is absent from our ML analysis of the nucleotide degen1 dataset, in which Alu_Pte is sister to Gel (Fig. S6).

We assume that Alu_Pte, Cal_Thy, and Gel form a monophyletic group (as suggested by our ML analyses), with the remaining species selected as an outgroup (OUT). We examined signal for alternative topologies and possible confounding signal among the Alu_Pte, Cal_Thy, and Gel. The optimized dataset used for FcLM comprised of: (*i*) four species belonging to Alu_Pte (Group 1), (*ii*) five species belonging to Cal_Thy (Group 2), (*iii*) six species belonging to Gel (Group 3) and (*iv*) 188 outgroup representatives, for a total of 203 species 1,322 partitions, 749,791 sites, and 22,560 drawn quartets. The three possible unambiguous topologies were:

T1: Alu_Pte, Cal_Thy | Gel, OUT

T2: Alu_Pte, Gel | Cal_Thy, OUT

T3: Alu_Pte, OUT | Cal_Thy, Gel

Proportions of all drawn quartets are provided in Fig. S11A and Dataset S8. The majority of the quartets contradict our tree topology in Fig. 1 (T1: 13.9%), with both other topologies each being supported by over one-third of the quartets (T2: 44.0%, T3: 41.8%). Results from the permutation approaches imply that confounding signal might have a slight impact on the support for T2 but at least some signal was not confounding. In contrast, we cannot rule out that support for T1 (as inferred in Fig. 1) comes from confounding, non-phylogenetic signal. This might also be reflected by the low statistical support values in the amino acid ML tree (Figs. 1, S2, S5). Although the Alucitoidea and Pterophoroidea both appear to belong in Obtectomera, their precise placement within this taxon remains elusive.

Placement of butterflies (Papilionoidea)

The second hypothesis tested (Fig. S11B, Dataset S7) assesses the placement of butterflies (Papilionoidea). The superfamily Papilionoidea (Pap) is placed as sister to all other Obtectomera in the best ML trees from our amino acid and nucleotide degen1 analyses (Figs. 1, S2, S4-S8) and in the ML tree of Kawahara and Breinholt (11). In contrast, prior morphological analyses suggested that butterflies were more closely related to the pyraloids and 'macromoths' than to the gelechioid 'micromoths'; the gelechioids were consequently thought to be more closely related to the other non-obtectomerans (66). In our analysis, pyraloids and macromoths are represented by Macroheterocera + Mimallonoidea + Pyraloidea (MMP), and the clade containing gelechioids (as well as other superfamilies that form a clade in our analysis) is termed Gelechioidea + relatives (Gel_rel). Our topology indicates that Pap is sister to the monophyletic

Gel_rel + MMP, and the alternative, morphology-based topology indicates that Pap is sister to MMP, forming a clade that is sister to Gel_rel. We assume that Pap, Gel_rel, and MMP form a monophyletic group (i.e. Obtectomera *sensu* Kawahara and Breinholt (11)); a subset of non-Obtectomera was thus selected as an outgroup (OUT). We examined signal for alternative topologies and possible confounding signal among Pap, Gel_rel and MMP. The optimized dataset used for FcLM comprised of: (*i*) seven species belonging to Pap (Group 1), (*ii*) eight species belonging to Gel_rel (Group 2), (*iii*) 21 species belonging to MMP (Group 3) and (*iv*) 18 outgroup representatives for a total of 54 species 1,322 partitions, 749,791 sites, and 21,168 drawn quartets. The three possible unambiguous topologies were:

T1: Pap, Gel_rel | MMP, OUT

T2: Pap, MMP | Gel_rel, OUT

T3: Pap, OUT | Gel_rel, MMP

Proportions of all drawn quartets are provided in Fig. S11B and Dataset S8. A plurality of all drawn quartets from the optimized FcLM dataset (original, not-permuted) support our amino acid ML topology (T3: 43.2%), with butterflies sister to the remaining Obtectomera. However, the morphology-based hypothesis has some signal (T2: 23.9%), and so does the topology that places Papilionoidea sister to the Gelechioidea and their relatives (T1: 32.3%). Permutation results imply that signal for T1 and/or T2 might be confounding. Although the true phylogenetic signal that takes putative confounding signal into account is weak, it still provides stronger signal than the putative confounding signal alone. Therefore, we consider our topology to be reliable, although there is confounding signal inherent within the data.

Sister-group relationships of silk moths and relatives (Lasiocampoidea + Bombycoidea)

The third hypothesis tested (Fig. S11C, Dataset S8) examines the placement of the clade containing lappet moths (Lasiocampoidea) and hawkmoths, silk moths, and relatives (Bombycoidea). The clade Lasiocampoidea + Bombycoidea (Las_Bom) was believed to be sister to the clade containing the sister taxa Geometroidea (Geo) and Noctuoidea (Noc) (11). The best trees from our amino acid and nucleotide degen1 analyses suggest a sister-group relationship between Las_Bom and Geo (Figs. 1, S2, S4-S8), implying that Geo and Noc are not monophyletic. Since these three clades form a monophyletic group within Macroheterocera that is well supported, we selected Drepanoidea (which is also in Macroheterocera) as an outgroup (OUT).

We examined signal for alternative topologies and possible confounding signal (see above) among the clades Las_Bom, Geo, and Noc. The optimized dataset used for FcLM comprised of: (*i*) 28 species belonging to Las_Bom (Group 1), (*ii*) eleven species belonging to Geo (Group 2), (*iii*) 22 species belonging to Noc (Group 3) and (*iv*) seven outgroup representatives for a total of 68 species, 1,322 partitions, 749,791 sites, and 47,432 drawn quartets. The three possible unambiguous topologies were:

T1: Las_Bom, Geo | Noc, OUT

T2: Las_Bom, Noc | Geo, OUT

T3: Las_Bom, OUT | Geo, Noc

Proportions of all drawn quartets are provided in Fig. S11C and Dataset S9. A plurality of quartets supports a sister-group relationship between Las_Bom and Noc (T2: 49.5%), which contradicts both our tree topology in Fig. 1 (T1: 26.0%) and the topology of Kawahara and Breinholt (11) (T3: 24.0%). Although the sister-group relationship between Las_Bom and Geo

has sufficient support in our amino acid ML tree (BS: 93; SH-aLRT: 96.6; TBE: 0.9785; Figs. S2, S4-S5), the permutation results imply that our topology (T1) is highly biased by confounding signal mainly coming from unequal distribution of missing data (cf. permutation II and III). This implies that our ML tree might be a result of confounding signal that is stronger than genuine phylogenetic signal.

10. Divergence time estimation

Because our molecular data matrices are very large, it was not computationally feasible to perform divergence time estimations on the same datasets that were used for the ML analyses. Thus, a subsampled version of our amino acid supermatrix was used to estimate dates of divergence. All non-amphiesmenopteran were removed from the supermatrix and pruned from the input tree. Sequences of the 195 Amphiesmenoptera species contained only sites for which at least 80% of samples had unambiguous amino acids. AliStat v1.6. (71) was used to generate a subsampled dataset that contained 198,050 amino acids and had an alignment completeness score (Ca) of 0.8533%. Because our subsampled dataset was still too large to use in programs such as BEAST (72) and STARBEAST (73), we instead used MCMCTree and codeml (both part of the PAML software package, v4.9g (74)) to estimate dates of divergence, using the approximate likelihood method (75).

In order to conduct divergence time analyses in a computationally feasible manner, we first re-estimated models on the reduced dataset which was set to a single partition. In ModelFinder (part of IQ-TREE v1.5.5), we set models applicable to those that are available in MCMCTree. The six models were: Dayhoff (76), DCMut (45), JTT (46), JTTDCMut (45), LG (47), WAG (48), with a maximum of five rate categories (-cmax 5, and the default setting -mrate

E,I,G,I+G,R). The LG substitution model (47) was chosen as the best substitution model by all three information criteria (AIC, AICc, and BIC), and was subsequently used for dating analyses. The subsampled dataset, along with AliStat and model selection output files, are provided in Supplementary Archive 8.

The input tree was time-calibrated using 16 fossils, carefully selected following the best-practice recommendations by Parham *et al.* (77). Fossils chosen to calibrate the inferred phylogeny are listed in Dataset S10. Additional details about the fossils, as well as justifications for their inclusion, are provided in the supplementary appendix. Although all 16 fossils have diagnostic morphological characters that enable reasonably confident placement in particular clades, only three of these fossils possess true synapomorphies. In order to more strictly follow the guidelines of Parham *et al.* (77), an additional divergence-time estimation analysis was performed using only these three fossils. Divergence time estimates were inferred using the best ML topology as a fixed input tree (Figs. S12-S19). Node ages for select clades in both the 16-fossil and 3-fossil analyses are provided in Fig. S20 and Dataset S11.

In MCMCTree, the birth-death process used to calculate the time prior is conditioned on the age of the root (78), and therefore the root calibration must be a proper statistical distribution. Fossil calibrations were introduced as minimum bounds of uniform priors and with the maximum root age as a hard maximum bound. However, maximum calibration bounds (which must be applied to the root age) are notoriously difficult to specify, as they rely on the absence of evidence (79). We applied a conservative age constraint on the Amphiesmenoptera (Lepidoptera + Trichoptera) root between 201 Ma (based on stem Glossata scale fossils (80), see Dataset S5) and 314.4 Ma based on the absence of amphiesmenopteran fossils in the late Carboniferous. The oldest definitive amphiesmenopteran fossil is from the Permian/Triassic boundary at 252 Ma

(81), with putative occurrences in the range 273-280 Ma (82). Thus, our maximum age constraint is very conservative, follows the maximum age estimate of Amphiesmenoptera by Tong *et al.* (83), and is close to the age of the oldest known winged insect (84).

We used two strategies to convert fossil ages into calibrations on the tree nodes. In the conservative strategy, fossil calibrations are uniform distributions constrained between the corresponding fossil age (the minimum bound) and a hard maximum equal to the maximum on the root age. This strategy leads to diffuse calibrations, which may extend over several hundred million years for some nodes. In the informative strategy, the truncated-Cauchy distribution (85) was used to set the calibrations for internal nodes younger than 80 Ma: the fossil age provides the minimum constraint, and the Cauchy distribution provides a tail that decays as it extends back in time. The resulting calibration density has most of the probability mass accumulated near the fossil age (the minimum age constraint). This additional constraint can be useful when applied to younger fossils that are very far from the root, since it improves convergence during the MCMCTree analyses. After we chose to constrain the root age between 201 and 314.4 Mya, we decided that any fossils under 80 Mya were sufficiently young to justify using Cauchy priors to set the calibrations. In both strategies, a uniform distribution (between 201 and 314.4 Ma) was used for the root.

The Dirichlet-Gamma density (86) was used to set the prior on the molecular rate, with parameters $\alpha_\mu = 2$, $\beta_\mu = 6$, $\alpha = 1$, which produces a diffuse prior density close to the empirical rate in the tree. The birth-death process with parameters $\lambda = \mu_{BD} = 1$, $\rho = 0.1$, which produces a diffuse, uniform-like density, was used to set the time prior for nodes without fossil calibrations. We used the independent-rates (clock = 2) model (87) and the autocorrelated-rates model to estimate divergence times. In total, eight analyses were performed: 2 sets of fossil calibrations

(16 fossils and 3 fossils) x 2 fossil calibration strategies (uniform priors and Cauchy priors for nodes < 80 Ma) x 2 rate models (independent and autocorrelated). Input trees with calibrated fossils are provided in Supplementary Archive 8.

We used the LG (aaRatefile = LG.dat) substitution model with 5 rate categories. Hessian matrices were calculated according to the above specifications with codeml using empirical (+F, model = 3) base frequencies estimated from the respective dataset. Multiple test runs were carried out to determine the optimal values for the sampling parameters. For our first test runs, we used a burn-in of 100,000, a sample frequency of 10, and a sample number of 1,000,000 as our sampling parameters. These parameters were increased for subsequent test runs, until we settled on a burn-in of 200,000, sample frequency (samplefreq) of 4,000 and a sample size (nsample) of 10,000 (resulting in a total run of 40.2 million generations) for each of our four MCMCTree analyses.

Each analysis was repeated four times (i.e. four MCMC chains). All four runs were subsequently evaluated for convergence in Tracer v1.6 (88). We also checked each MCMC run for convergence by plotting node ages from the four runs (the posterior mean and the lower and upper bounds for the 95% credibility intervals) against each other in R (Figs. S21-S28), although for some analyses, two of the four runs failed to summarize (Figs. S27, S28). Converged runs were merged in order to reach a sufficient effective sample size (ESS > 200) for all parameters. The posterior from the merged runs was used to create a summary tree. Time priors under the two different fossil strategies (16-fossil and 3-fossil) were obtained by running MCMCTree without the sequence alignment (see density plots in Figs. S29-S32). Additional files used for these analyses are provided in Supplementary Archive 8.

To facilitate the discussion of correlations between the radiation of Lepidoptera and radiation of flowering plants, recent angiosperm phylogenetic studies were consulted (89-96). Multiple hypotheses for the mean age of the ancestral angiosperm were compiled; these ages are presented in Dataset S12 and incorporated into Fig. 2. The estimated time since divergence between angiosperms and gymnosperms is large, and it is possible that flowering plants existed long before the crown of angiosperms. Therefore, the interval between the mean age of the crown (node Angiosperm) and the mean age of the stem (node Angiosperm + Gymnosperm) from these studies is presented. The ingroup topology of extant angiosperms from one of these phylogenetic studies (figure S6 in Foster *et al.* (93)) was incorporated into Fig. 2, with the interval between mean crown age and mean stem age shown for select studies (Fig.1, Dataset S12).

11. Ancestral state reconstruction

Ancestral state reconstruction analyses were performed on the best tree from the ML analysis of the concatenated amino acid dataset (Fig. S1). Some parts of the best tree have small branch lengths that make it difficult to visualize the data, so results were subsequently mapped onto ultrametric versions of the topology (Figs. S33, S34). For the hearing organ ASR, species were assigned character states based on published papers of hearing in Lepidoptera (e.g. 97, 98, 99), and unpublished studies (100, 101). The presence of a hearing organ was treated as a binary character in one ASR analysis, and as a more conservative, seven-state character in a separate ASR analysis (see matrices A and B in Dataset S13). For the ASR analysis of diel activity, we coded characters from data in table S1 of Kawahara *et al.* (102). If a species in our dataset lacked behavioral observations from the literature, the species was scored using the diel-activity data for

a corresponding higher-level taxon, as found in Kawahara *et al.* (102). The coding scheme for diel activity can be found in Dataset S14.

All ASR analyses were conducted using stochastic character mapping, with the 'make.simmap' command in the R package Phytools v06-44 (103). Ten thousand stochastic maps were generated for each ASR analysis, and forward and reverse character state transitions were assumed to have equal rates.

Acknowledgments

The first author thanks C. Mitter for his ongoing support throughout this project. M. Heikkilä, K. Kjer and J.C. Regier provided significant advice that helped this project. K. Dexter, S. Epstein, and L. Xiao assisted with specimen preparation and vouchering. R. Holzenthal provided discussion on trichopteran fossils. D. Davis and T. Simonsen provided discussion on lepidopteran fossils. O. Niehuis and M. Petersen provided suggestions on Orthograph. T. Wong provided programs to assist with phylogenetic analyses. L. Teasdale provided average p-Hamming distances. L. Waidele provided R scripts to assist with divergence time estimation. G. Meng, C. Yang, and other colleagues at the BGI assisted with sample and data curation. The CSIRO PEARCEY cluster (facilitated by O. Hlinka), Smithsonian Institution High Performance Cluster (SI/HPC), UF HPC, ZFMK HPC, and BGI provided computational support and assistance. The following people provided specimens or assisted with collection of specimens: D. Bartsch, R. Blahnik, A. Blanke, A. Böhm, K. Fischer, C. Greve, R. Holzenthal, C. Johns, K. Kjer, M. Kubiak, J. Macko, C. Manchester, C. Mitter, L. Mound, S. Nekum, H. Riefenstahl, D. Ruiter, K. Schütte, J. Schwarzer, J. Truman, and R. Wisseman. L. Reeves and X. Zheng provided some images in Figure 1. The photo of the Lepidoptera scale in Figure 2 is attributed to T. van

 This project was financially supported by the U.S. National Science Foundation grants (NSF) DEB 1354585, 1541500, 1557007; IOS 1121739, 1121807, 1920895, 1920936; China National GeneBank and BGI-Shenzhen, China; German Research Foundation BE 1789/8-1, 1789/10-1, MI 649/6, 649/10, NI 1387/1-1, RE 345/1-2, STA 860/4, and the Heisenberg grant WA 1496/8-1.

Author contributions

Kawahara, Misof, and Peters conceived the study. Barber, Frandsen, Kawahara, Meusemann, Misof, Peters, Yack, and Zwick collected or provided samples. Breinholt, Liu, and Zhou sequenced and assembled data. Donath, Meusemann, and Podsiadlowski processed assembled data. Breinholt, Donath, Espeland, Frandsen, Gimnich, Kawahara, Mayer, Meusemann, Misof, Plotkin, and Zwick conducted phylogenetic analyses. Espeland, Kawahara, Meusemann, Plotkin, dos Reis, Storer, Toussaint and Yack helped with ancestral state

reconstructions, fossil verifications, and divergence time estimation analyses. All authors contributed to the writing of the manuscript, with Kawahara taking the lead.

Data and materials availability

All data are available in the main text, the Supplementary Information, or in gzipped data archives available at the Dryad Digital Repository (https://doi.org/10.5061/dryad.j477b40). A list of contents of these archives is provided after the supplementary dataset legends.

References

1. V. Ahola *et al.*, The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat. Commun.* **5**, 1-9 (2014).
2. A. L. Bazinet, M. P. Cummings, K. T. Mitter, C. W. Mitter, Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. *PLoS ONE* **8**, e82615 (2013).
3. A. L. Bazinet *et al.*, Phylotranscriptomics resolves ancient divergences in the Lepidoptera. *Syst. Entomol.* **42**, 305-316 (2017).
4. M. Berger *et al.*, Insecticide resistance mediated by an exon skipping event. *Mol. Ecol.* **25**, 5692-5704 (2016).
5. J. W. Breinholt, A. Y. Kawahara, Phylotranscriptomics: Saturated third codon positions radically influence the estimation of trees based on next-gen data. *Gen. Biol. Evol.* **5**, 2082-2092 (2013).
6. J. W. Breinholt *et al.*, Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Syst. Biol.* **67**, 78-93 (2018).
7. J. W. Davey *et al.*, Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3 (Bethesda)* **6**, 695-708 (2016).
8. J. Duan *et al.*, SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.* **38**, D453-456 (2010).
9. F. C. de Assis Fonseca *et al.*, Sugarcane giant borer transcriptome analysis and identification of genes related to digestion. *PLoS ONE* **10**, e0118231 (2015).
10. J. A. Galarza, K. Dhaygude, J. Mappes, *De novo* transcriptome assembly and its annotation for the aposematic wood tiger moth (*Parasemia plantaginis*). *Genomics Data* **12**, 71-73 (2017).
11. A. Y. Kawahara, J. W. Breinholt, Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proc. R. Soc. Lond., Ser. B: Biol. Sci.* **281**, 20140970 (2014).

12. H. S. Kim *et al.*, BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res.* **38**, D437-D442 (2010).
13. R. J. Challis, S. Kumar, K. K. K. Dasmahapatra, C. D. Jiggins, M. Blaxter, Lepbase: the Lepidopteran genome database. *BioRxiv* 10.1101/056994, 056994 (2016).
14. T. Lowe, P. Chan, Genomic tRNA Database: tRNAscan-SE Analysis of *Drosophila melanogaster*. (2013).
15. S. Y. Ma *et al.*, Increasing the yield of middle silk gland expression system through transgenic knock-down of endogenous sericin-1. *Mol. Genet. Genomics* **292**, 823-831 (2017).
16. T. Pauli *et al.*, Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects. *BMC Genomics* **17**, 861 (2016).
17. R. S. Peters *et al.*, The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. *BMC Evol. Biol.* **14**, 52 (2014).
18. J. Romiguier *et al.*, Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261-263 (2014).
19. M. V. Sharakhova *et al.*, Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biol.* **8**, R5 (2007).
20. M. You *et al.*, A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220-225 (2013).
21. S. Zhan, S. M. Reppert, MonarchBase: the monarch butterfly genome database. *Nucleic Acids Res.* **41**, D758-D763 (2012).
22. W. Zhang, K. Kunte, M. R. Kronforst, Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using de novo transcriptome assemblies. *Gen. Biol. Evol.* **5**, 1233-1245 (2013).
23. S. Cho *et al.*, Preserving and vouchering butterflies and moths for large-scale museum-based molecular research. *PeerJ* **4**, e2160 (2016).
24. B. Misof *et al.*, Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767 (2014).
25. R. S. Peters *et al.*, Evolutionary history of the Hymenoptera. *Curr. Biol.* **27**, 1013-1018 (2017).
26. Y. L. Xie *et al.*, SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660-1666 (2014).
27. Y. Surget-Groba, J. I. Montoya-Burgos, Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.* **20**, 1432-1440 (2010).
28. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
29. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
30. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268-274 (2015).
31. S. Ratnasingham, P. D. N. Hebert, BOLD: The barcode of life data system (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355-364 (2007).
32. A. J. Aberer, D. Krompass, A. Stamatakis, Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162-166 (2013).

33. R. Ohgushi, Ecological notes on *Epipomponia nawai* (Dyar), a parasite of cicada in Japan (Lepidoptera: Epipyropidae). *Transactions of the Shikoku Entomological Society* **3**, 185-191 (1953).
34. R. M. Waterhouse, F. Tegenfeldt, J. Li, E. M. Zdobnov, E. V. Kriventseva, OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41**, D358-D365 (2013).
35. M. Petersen *et al.*, Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* **18**, 111 (2017).
36. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
37. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-W612 (2006).
38. B. Misof, K. Misof, A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst. Biol.* **58**, 21-34 (2009).
39. P. Kück *et al.*, Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology* **7**, 1-12 (2010).
40. P. Kück (2009) ALICUT: a Perlscript which cuts ALISCORE identified RSS. (Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany).
41. P. Kück, G. C. Longo, FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* **11**, 81 (2014).
42. R. Lanfear, P. B. Frandsen, A. M. Wright, T. Senfeld, B. Calcott, PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772-773 (2017).
43. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
44. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A* **89**, 10915-10919 (1992).
45. C. Kosiol, N. Goldman, Different versions of the Dayhoff rate matrix. *Mol. Biol. Evol.* **22**, 193-199 (2004).
46. D. T. Jones, W. R. Taylor, J. M. Thornton, The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275-282 (1992).
47. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307-1320 (2008).
48. S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691-699 (2001).
49. S. Q. Le, C. C. Dang, O. Gascuel, Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921-2936 (2012).
50. B. Misof *et al.*, A priori assessment of data quality in molecular phylogenetics. *Algorithms for Molecular Biology* **9** (2014).
51. Z. Yang, Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306-314 (1994).

52. C. M. Hurvich, C.-L. Tsai, Regression and time series model selection in small samples. *Biometrika* **76**, 297-307 (1989).
53. R. Lanfear, B. Calcott, D. Kainer, C. Mayer, A. Stamatakis, Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* **14**, 82 (2014).
54. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587-589 (2017).
55. O. Chernomor, A. von Haeseler, B. Q. Minh, Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997-1008 (2016).
56. A. Zwick, J. C. Regier, D. J. Zwickl, Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: Differentiating serine codons in 21-amino-acid models. *PLoS ONE* **7** (2012).
57. J. C. Sohn *et al.*, Phylogeny and feeding trait evolution of the mega-diverse Gelechioidea (Lepidoptera: Obtectomera): new insight from 19 nuclear genes. *Syst. Entomol.* **41**, 112-132 (2016).
58. J. C. Regier *et al.*, A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PLoS ONE* **8**, e58568 (2013).
59. A. Y. Kawahara *et al.*, Increased gene sampling strengthens support for higher-level groups within leaf-mining moths and relatives (Lepidoptera: Gracillariidae). *BMC Evol. Biol.* **11**, 182 (2011).
60. N. D. Pattengale, M. Alipour, O. R. P. Bininda-Emonds, B. M. E. Moret, A. Stamatakis, "How Many Bootstrap Replicates Are Necessary?" in Research in Computational Molecular Biology, S. Batzoglou, Ed. (Springer Berlin Heidelberg, 2009), vol. 5541, chap. 13, pp. 184-200.
61. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
62. F. Lemoine *et al.*, Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
63. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518-522 (2018).
64. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
65. X.-X. Shen, C. T. Hittinger, A. Rokas, Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* **1**, 0126 (2017).
66. N. P. Kristensen, *Handbook of Zoology, Volume IV, Arthropoda: Insecta, Part 35. Lepidoptera, Moths and Butterflies, Volume 1: Evolution, Systematics, and Biogeography*. N. P. Kristensen, Ed. (Walter de Gruyter, Berlin, New York, 1998).
67. S. Y. W. Ho, L. S. Jermiin, Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.* **53**, 623-637 (2004).
68. L. Jermiin, S. Y. Ho, F. Ababneh, J. Robinson, A. W. Larkum, The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol.* **53**, 638-643 (2004).
69. K. P. Johnson *et al.*, Phylogenomics and the evolution of hemipteroid insects. *Proc. Natl. Acad. Sci. U.S.A* **115**, 12775-12780 (2018).
70. S. Simon, A. Blanke, K. Meusemann, Reanalyzing the Palaeoptera problem–The origin of insect flight remains obscure. *Arthropod Struct. Dev.* **47**, 328-338 (2018).

71. T. K. F. Wong *et al.*, AliStat version 1.3. v1. CSIRO. Software Collection. 10.4225/08/59309da8368e1 (2014).
72. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
73. J. Heled, A. J. Drummond, Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27**, 570-580 (2010).
74. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586 - 1591 (2007).
75. M. dos Reis, Z. H. Yang, Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161-2172 (2011).
76. M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, "A model of evolutionary change in proteins" in Atlas of protein sequence and structure. (National Biomedical Research Foundation Silver Spring, 1978), vol. 5, pp. 345-352.
77. J. F. Parham *et al.*, Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346-359 (2011).
78. Z. H. Yang, B. Rannala, Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**, 212-226 (2006).
79. M. J. Benton, P. C. Donoghue, Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24**, 26-53 (2007).
80. T. J. B. van Eldijk *et al.*, A Triassic-Jurassic window into the evolution of Lepidoptera. *Sci. Adv.* **4**, e1701568 (2018).
81. J. Minet, D.-Y. Huang, H. Wu, A. Nel, Early Mecopterida and the systematic position of the Microptysmatidae (Insecta: Endopterygota). *Ann. la Société Entomol. Fr.* **46**, 262-270 (2010).
82. D. S. Aristov, The fauna of grylloblattid insects (Grylloblattida) from the end of the late Permian to the first half of the Triassic. *Paleontological Journal* **38**, 514-521 (2004).
83. K. J. Tong, S. Duchene, S. Y. W. Ho, N. Lo, Comment on "Phylogenomics resolves the timing and pattern of insect evolution". *Science* **349**, 487 (2015).
84. C. Brauckmann, B. Brauckmann, E. Gröning, The stratigraphical position of the oldest known Pterygota (Insecta. Carboniferous, Namurian). *Annales de la Société géologique de Belgique* **117**, 47-56 (1994).
85. J. Inoue, P. C. J. Donoghue, Z. H. Yang, The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* **59**, 74-89 (2010).
86. M. dos Reis, T. Q. Zhu, Z. H. Yang, The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. *Syst. Biol.* **63**, 555-565 (2014).
87. B. Rannala, Z. H. Yang, Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453-466 (2007).
88. A. Rambaut, M. A. Suchard, D. Xie, A. J. Drummond, Tracer, version 1.6. http://beast.bio.ed.ac.uk/Tracer. (2014).
89. S. Magallón, Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. *Syst. Biol.* **59**, 384-399 (2010).
90. K. Salomo *et al.*, The emergence of earliest angiosperms may be earlier than fossil evidence indicates. *Syst. Bot.* **42**, 607-619 (2017).

91. J. T. Clarke, R. C. M. Warnock, P. C. J. Donoghue, Establishing a time−scale for plant evolution. *New Phytol.* **192**, 266-301 (2011).
92. S. Magallón, A. Castillo, Angiosperm Diversification through Time. *Am. J. Bot.* **96**, 349-365 (2009).
93. C. S. P. Foster *et al.*, Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst. Biol.* **66**, 338-351 (2017).
94. S. Magallón, K. W. Hilu, D. Quandt, Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.* **100**, 556-573 (2013).
95. C. S. P. Foster, S. Y. W. Ho, Strategies for partitioning clock models in phylogenomic dating: application to the angiosperm evolutionary timescale. *Gen. Biol. Evol.* **9**, 2752-2763 (2017).
96. H.-T. Li *et al.*, Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **5**, 461-470 (2019).
97. J. Minet, A. Surlykke, "Auditory and sound producing organs" in Handbook of Zoology, Volume IV, Arthropoda: Insecta, Part 36. Lepidoptera, Moths and Butterflies, Volume 2: Morphology, Physiology, and Development. (Walter de Gruyter, Berlin, New York, 2003), pp. 289-323.
98. A. Y. Kawahara, J. R. Barber, Tempo and mode of ultrasound and jamming in the diverse hawkmoth radiation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6407-6412 (2015).
99. P. Sun, N. Mhatre, A. C. Mason, J. E. Yack, In that vein: inflated wing veins contribute to butterfly hearing. *Biol. Lett.* **14**, 20180496 (2018).
100. S. J. Mahony (2006) Hearing in the speckled wood butterfly, *Pararge aegeria* (Nymphalidae: Satyrinae). in *Department of Biology* (Carleton University).
101. L. E. Hall (2014) Tympanal Ears in Nymphalidae Butterflies: Morphological Diversity and Tests on the Function of Hearing. in *Department of Biology* (Carleton University), p 173.
102. A. Y. Kawahara *et al.*, Diel behavior in moths and butterflies: a synthesis of data illuminates the evolution of temporal activity. *Org. Divers. Evol.* **18**, 13-27 (2018).
103. L. J. Revell, Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217-223 (2012).
104. P. A. Hochuli, S. Feist-Burkhardt, Angiosperm-like pollen and *Afropollis* from the Middle Triassic (Anisian) of the Germanic Basin (Northern Switzerland). *Frontiers Plant Sci.* **4**, 1-14 (2013).

Figure legends

Fig. S1. Hearing organs in adult Lepidoptera. Tineidae (second abdominal segment); Hedylidae (ventral base of anterior forewing); Nymphalidae (ventral base of anterior forewing); Thyrididae (ventral base of anterior forewing); Pyralidae (second abdominal segment); Cimeliidae (seventh abdominal segment); Drepanidae (first abdominal segment); Doidae (posterior metathorax); Noctuidae (posterior metathorax); Uraniidae (second abdominal segment); Geometridae (second abdominal segment); Sphingidae (palp-pilifer). All scale bars for upper images of each family are 1 mm (except for Sphingidae, which is 500 μm). Lower images are light micrographs of the respective ears. All scale bars for bottom (ear) images are 100 μm. Due to the limited taxon sampling in our dataset, some lepidopteran lineages that possess hearing organs were not represented in these analyses. Hearing organs are only found in three of the eight nymphalids (brush-footed butterflies) in our dataset, all in subfamily Satyrinae. However, there are five other nymphalid subfamilies that have at least some species with hearing organs (97). The moth family Dudgeoneidae, which contains less than ten species, all with ears, is also not represented in our dataset because representatives of this family were not obtainable for transcriptome sequencing.

Fig. S2. Maximum likelihood best tree from IQ-TREE amino acid analysis, with non-parametric bootstrap support values. lnL = -42341345.559.

Fig. S3. Maximum likelihood suboptimal tree topology from the IQ-TREE amino acid analysis.

Fig. S4. SH-aLRT support values mapped on the best maximum likelihood best tree from the IQ-TREE amino acid analysis.

Fig. S5. TBE support values mapped on the best maximum likelihood tree from the IQ-TREE amino acid analysis.

Fig. S6. Maximum likelihood best tree from IQ-TREE degen1 analysis. Support values are non-parametric bootstrap values. lnL = -45121944.618.

Fig. S7. SH-aLRT support values mapped on the maximum likelihood topology from the IQ-TREE degen1 analysis.

Fig. S8. TBE support values mapped on the maximum likelihood topology from the IQ-TREE degen1 analysis.

Fig. S9. ASTRAL species tree derived from the best ML amino acid gene trees.

Fig. S10. ASTRAL species tree derived from the amino acid consensus bootstrap trees.

Fig. S11. Results of the Four-cluster Likelihood Mapping analyses performed on the amino acid dataset, showing quartet support (in %) of all drawn quartets mapped onto 2D simplex graphs for possible topologies. 2D simplex graphs from left to right: original data, permutation I, II, III. **A)** Assessment of the placement of Alucitoidea+Pterophoroidea (Alu_Pte). Other groups are Gelechioidea (Gel), Calliduloidea+Thyridoidea (Cal_Thy), and an outgroup containing all other species in the dataset (OUT). **B)** Assessment of the placement of Papilionoidea (Pap). Other groups are Macroheterocera+Mimallonoidea+Pyraloidea (M_M_P), Gelechioidea and relatives (Gel_rel), and an outgroup containing non-obtectomeran Lepidoptera (OUT). **C)** Assessment of the placement of Lasiocampoidea+Bombycoidea (Las_Bom). Other groups are Geometroidea (Geo), Noctuoidea (Noc), and an outgroup containing Drepanoidea (OUT). T1, T2, T3 show unambiguous quartet topologies (area 1, 2, 3 in IQ-TREE), T12 (area 4 in IQ-TREE), T13 (area 6 in IQ-TREE) and T23 (area 5 in IQ-TREE) show partially resolved quartets, quartets in T* (area 7 in IQ-TREE) unresolved. $ indicates the topology supported in the best amino acid ML tree.

Fig. S12. Divergence time estimates from the 16-fossil MCMCTree analysis with uncorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in millions of years (Ma).

Fig. S13. Divergence time estimates from the 16-fossil MCMCTree analysis with uncorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S14. Divergence time estimates from the 16-fossil MCMCTree analysis with autocorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S15. Divergence time estimates from the 16-fossil MCMCTree analysis with autocorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S16. Divergence time estimates from the 3-fossil MCMCTree analysis with uncorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S17. Divergence time estimates from the 3-fossil MCMCTree analysis with uncorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S18. Divergence time estimates from the 3-fossil MCMCTree analysis with autocorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S19. Divergence time estimates from the 3-fossil MCMCTree analysis with autocorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

Fig. S20. Comparison of ages of major clades in Lepidoptera and Trichoptera.

Fig. S21. Convergence plots for the 16-fossil analysis with uncorrelated rates and uniform priors showing relationship between posterior means and confidence intervals among the four runs.

Fig. S22. Convergence plots for the 16-fossil analysis with uncorrelated rates and Cauchy priors showing relationship between posterior means and confidence intervals among the four runs.

Fig. S23. Convergence plots for the 16-fossil analysis with autocorrelated rates and uniform priors showing relationship between posterior means and confidence intervals among the four runs.

Fig. S24. Convergence plots for the 16-fossil analysis with autocorrelated rates and Cauchy priors showing relationship between posterior means and confidence intervals among the four runs.

Fig. S25. Convergence plots for the 3-fossil analysis with uncorrelated rates and uniform priors showing relationship between posterior means and confidence intervals among the four runs.

Fig. S26. Convergence plots for the 3-fossil analysis with uncorrelated rates and Cauchy priors showing relationship between posterior means and confidence intervals among the four runs.

Fig. S27. Convergence plots for the 3-fossil analysis with autocorrelated rates and uniform priors showing relationship between posterior means and confidence intervals among the two available runs.

Fig. S28. Convergence plots for the 3-fossil analysis with autocorrelated rates and Cauchy priors showing relationship between posterior means and confidence intervals among the two available runs.

Fig. S29. Density plots for the 16-fossil analysis with uncorrelated rates and uniform priors showing the relationship between sampling frequency and the age of a particular clade.

Fig. S30. Density plots for the 16-fossil analysis with uncorrelated rates and Cauchy priors showing the relationship between sampling frequency and the age of a particular clade.

Fig. S31. Density plots for the 3-fossil analysis with uncorrelated rates and uniform priors showing the relationship between sampling frequency and the age of a particular clade.

Fig. S32. Density plots for the 3-fossil analysis with uncorrelated rates and Cauchy priors showing the relationship between sampling frequency and the age of a particular clade.

Fig. S33. Ancestral state reconstructions of hearing organs on the amino acid ML tree topology. Left: binary (2-state) ancestral state reconstruction. Black = hearing organ absent; red = hearing organ present. Right: multi-state (7-state) ancestral state reconstruction. Black = hearing organs absent; red = tympana on the sternum of the second abdominal segment; green = tympana on the metathorax; dark blue = tympana beneath the forewing bases; cyan = tympana on the first abdominal segment; magenta = hearing organs near the spiracles of the seventh abdominal segment; yellow = hearing organs on palpi.

Fig. S34. Ancestral state reconstruction of adult diel activity on the amino acid ML tree topology. Black = nocturnal; Red = diurnal; Green = crepuscular; Blue = active at all times.

Dataset legends

Dataset S1. (provided as separate Excel file) List of the 203 species included in the present study, with species taxonomy, library identification codes, NCBI taxon identification codes, and accession numbers.

Dataset S2. (provided as separate Excel file) Preservation and storage methods for the specimens used to generate 69 new transcriptomes in the present study.

Dataset S3. (provided as separate Excel file) Sequences removed during contamination filtering steps.

Dataset S4. (provided as separate Excel file) Additional sequence statistics for the genomes and transcriptomes in the present study.

Dataset S5. (provided as separate Excel file) Support values for major nodes on the best tree found in each phylogenetic analysis. "N/A" indicates the clade was not recovered.

Dataset S6. (provided as separate Excel file) Phylogenetic hypotheses and groups for the Four-cluster Likelihood Mapping (FcLM) analyses examining the placement of Alucitoidea and Pterophoroidea.

Dataset S7. (provided as separate Excel file) Phylogenetic hypotheses and groups for the Four-cluster Likelihood Mapping (FcLM) analyses examining the placement of butterflies (Papilionoidea).

Dataset S8. (provided as separate Excel file) Phylogenetic hypotheses and groups for the Four-cluster Likelihood Mapping (FcLM) analyses examining the placement of Lasiocampoidea and Bombycoidea.

Dataset S9. (provided as separate Excel file) Percentages of quartets supporting all possible topologies for the phylogenetic hypotheses tested using Four-cluster Likelihood Mapping (FcLM) analyses.

Dataset S10. (provided as separate Excel file) Lepidoptera and Trichoptera fossils included in the 16-fossil MCMCTree dating analyses.

Dataset S11. (provided as separate Excel file) Summary of divergence time estimates in MCMCTree, with median ages and age ranges (95% CIs) provided for select clades (in Ma).

Dataset S12. (provided as separate Excel file) Hypothesized ages of angiosperms from recent phylogenetic studies.

Dataset S13. (provided as separate Excel file). Character matrices for the ancestral state reconstructions (ASRs) used to study the evolution of anti-bat hearing organs in Lepidoptera.

Dataset S14. (provided as separate Excel file). Character matrices for the ancestral state reconstruction (ASR) used to study the evolution of diel activity in Lepidoptera.

Supplementary Archive 1. (provided as separate data file)

Directory including all non-1KITE assemblies (FASTA format) and a list of respective information on file names and species (SA1_assembly_information.csv).

Supplementary Archive 2. (provided as separate data file)

- Text file (TBD_COS.table) listing all ortholog groups (OGs) of the three reference species compiled from OrthoDB v7, as described in Section 5 of Materials and Methods, which served as input for Orthograph when building the ortholog-set databases
- Directory (OGS/) including official gene sets (OGSs) of the three reference genomes used for the reciprocal BLAST step at the amino acid level (*AA*) and at the nucleotide level (*NUC*). Official gene sets were modified, as described in Section 5 of Materials and Methods, for usage within Orthograph.
    - BMORI → *Bombyx mori* (OGS v2.0): corresp-BMORI.AA.fa, corresp-BMORI.NUC.fa
    - DPLEX → *Danaus plexippus* (OGS v2.0): corresp-DPLEX.AA.fa, corresp-DPLEX.NUC.fa
    - TCAST → *Tribolium castaneum* (OGS v3.0): corresp-TCAST.AA.fa, corresp-TCAST.NUC.fa

Supplementary Archive 3. (provided as separate data file)

- *.zip files including all loci after Alicut, in FASTA format. Amino acid level: SA3_aa_2098.zip. Nucleotide level: SA3_nuc_2098.zip.

- Amino acid supermatrix (SA3_aminoacid_supermatrix.phy) and list of gene boundaries within the supermatrix (SA3_aa_gene_boundaries_for_PF.txt) used as input for merging partitions with PartitionFinder
- Nucleotide supermatrix (SA3_nucleotide_supermatrix_before_degen.fas), prior to running Degen v1.4, and list of gene boundaries converted from the amino acid boundary file to fit the nucleotide dataset (SA3_nucleotide_gene_boundaries.txt).

Supplementary Archive 4. (provided as separate data file)

- PartitionFinder input configuration file (partition_finder.cfg),
- Best partition scheme selected by PartitionFinder (SA4_best_partition_scheme.txt)
- Amino acid supermatrix (SA4_aminoacid_supermatrix_resorted_renamed.fas) used as input for IQ-TREE and FcLM analyses, resorted as described in Section 4 of Materials and Methods. Sequence names in this file have been modified to incorporate the correct taxon names discussed in Section 1 of Materials and Methods.
- Best partition scheme including boundaries of merged partitions. (SA4_aa_best_partition_scheme_merged.txt)
- Best partition scheme and best models selected by ModelFinder (SA4_aa_best_partition_scheme_Modelfinder.txt) for phylogenetic inference and FcLM analyses
- lnL scores for all ML tree searches performed using the amino acid supermatrix (SA4_aminoacid_lnL_scores.csv).
- Inferred trees in Newick format (*.tre files) used to generate Figs. S2-S5.

Supplementary Archive 5. (provided as separate data file)

- Best models selected by ModelFinder for phylogenetic inference using individual amino acid alignments.
- *.zip file containing best ML trees (*.treefile) from individual amino acid alignments (SA5_aa_ml_trees.zip).
- *.zip file containing consensus bootstrap trees (*.contree) from individual amino acid alignments (SA5_aa_bs_contrees.zip).
- Inferred ASTRAL trees in Newick format (*.tre files) used to generate Figs. S9-S10.

Supplementary Archive 6. (provided as separate data file)

- Nucleotide supermatrix (SA6_degen_nucleotide_supermatrix_renamed.fasta) after applying Degen v1.4, and respective partition information (SA6_degen_nucleotide_partition_scheme.txt) used as input for IQ-TREE.
- Best partition scheme and best models selected by ModelFinder for phylogenetic inference using the degenerated nucleotide supermatrix (SA6_nuc_best_partition_scheme_Modelfinder.txt).
- lnL scores for all ML tree searches performed using the degenerated nucleotide supermatrix (SA6_nuc_lnL_scores.csv).
- Inferred trees in Newick format (*.tre files) used to generate Figs. S6-S8.

Supplementary Archive 7. (provided as separate data file)

Directory including all data subsets used for FcLM analyses testing three hypotheses, as described in Section 9 of Materials and Methods. Data are organized in three subdirectories:

Pterophoroidea_Alucitoidea/, Papilionoidea/, and Lasiocampoidea_Bombycoidea/. Each subdirectory includes "decisive" data subsets for their respective hypothesis (FASTA files of original data and permuted data (permutations I-III)), partition files (plain text format), and a nexus file with four defined taxonomic groups (plain text format). Note: The Pterophoroidea_Alucitoidea/ does not contain a FASTA file of the original data because the entire supermatrix from Supplementary Archive 4 (SA4_aminoacid_supermatrix_resorted_renamed.fas) was used for this analysis.

Supplementary Archive 8. (provided as separate data file)

Directory including files used for divergence time estimation analyses.

- Appendix with justifications for selecting all fossils used in the analyses.
- Subsampled dataset used for divergence time estimation analyses (subsampled_dataset_80pct.fas). The dataset includes 195 samples and sites for which at least 80% of samples had non-ambiguous amino acids.
- Subsampled dataset statistics and heatmap generated with AliStat.
- IQ-TREE logfile from model selection on the unpartitioned dataset, restricted to models available in PAML.
- MCMCTree files
    - Four input trees
        - Uniform priors, 16 fossil calibrations (Input_uniform_16.tre)
        - Uniform priors, 3 fossil calibrations (Input_uniform_3.tre)
        - Uniform and Cauchy priors, 16 fossil calibrations (Input_cauchy_16.tre)
        - Uniform and Cauchy priors, 3 fossil calibrations (Input_cauchy_3.tre)

    - Example of config file for step 1.
    - Example of config file for step 2 (generation of the Hessian matrix with CODEML).
    - Example of config file for step 3 (final MCMCTree runs;)
    - Calibrated time trees resulting from four merged MCMCTree runs for each of the eight analyses, used to generate Figs. S12-S19.
- Comparisons of uncorrelated vs. autocorrelated runs (SA8_uncorr_vs_autocorr.pdf).

All Supplementary Archives are available at the Dryad Digital Repository (https://doi.org/10.5061/dryad.j477b40)

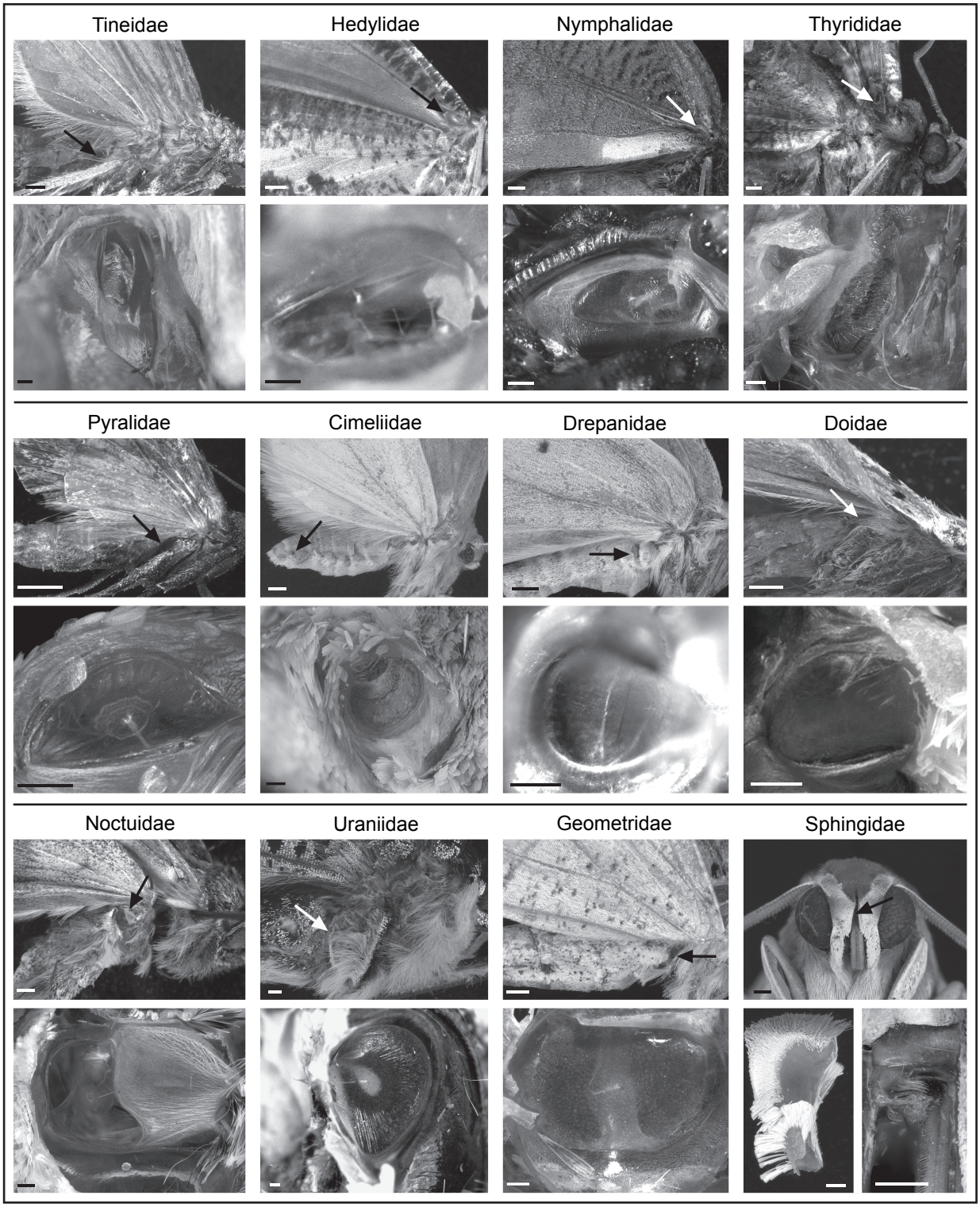

Fig. S1. Hearing organs in adult Lepidoptera. Tineidae (second abdominal segment); Hedylidae (ventral base of anterior forewing); Nymphalidae (ventral base of anterior forewing); Thyrididae (ventral base of anterior forewing); Pyralidae (second abdominal segment); Cimeliidae (seventh abdominal segment); Drepanidae (first abdominal segment); Doidae (posterior metathorax); Noctuidae (posterior metathorax); Uraniidae (second abdominal segment); Geometridae (second abdominal segment); Sphingidae (palp-pilifer). All scale bars for upper images of each family are 1 mm (except for Sphingidae, which is 500 μm). Lower images are light micrographs of the respective ears. All scale bars for bottom (ear) images are 100 μm. Due to the limited taxon sampling in our dataset, some lepidopteran lineages that possess hearing organs were not represented in these analyses. Hearing organs are only found in three of the eight nymphalids (brush-footed butterflies) in our dataset, all in subfamily Satyrinae. However, there are five other nymphalid subfamilies that have at least some species with hearing organs (97). The moth family Dudgeoneidae, which contains less than ten species, all with ears, is also not represented in our dataset because representatives of this family were not obtainable for transcriptome sequencing.

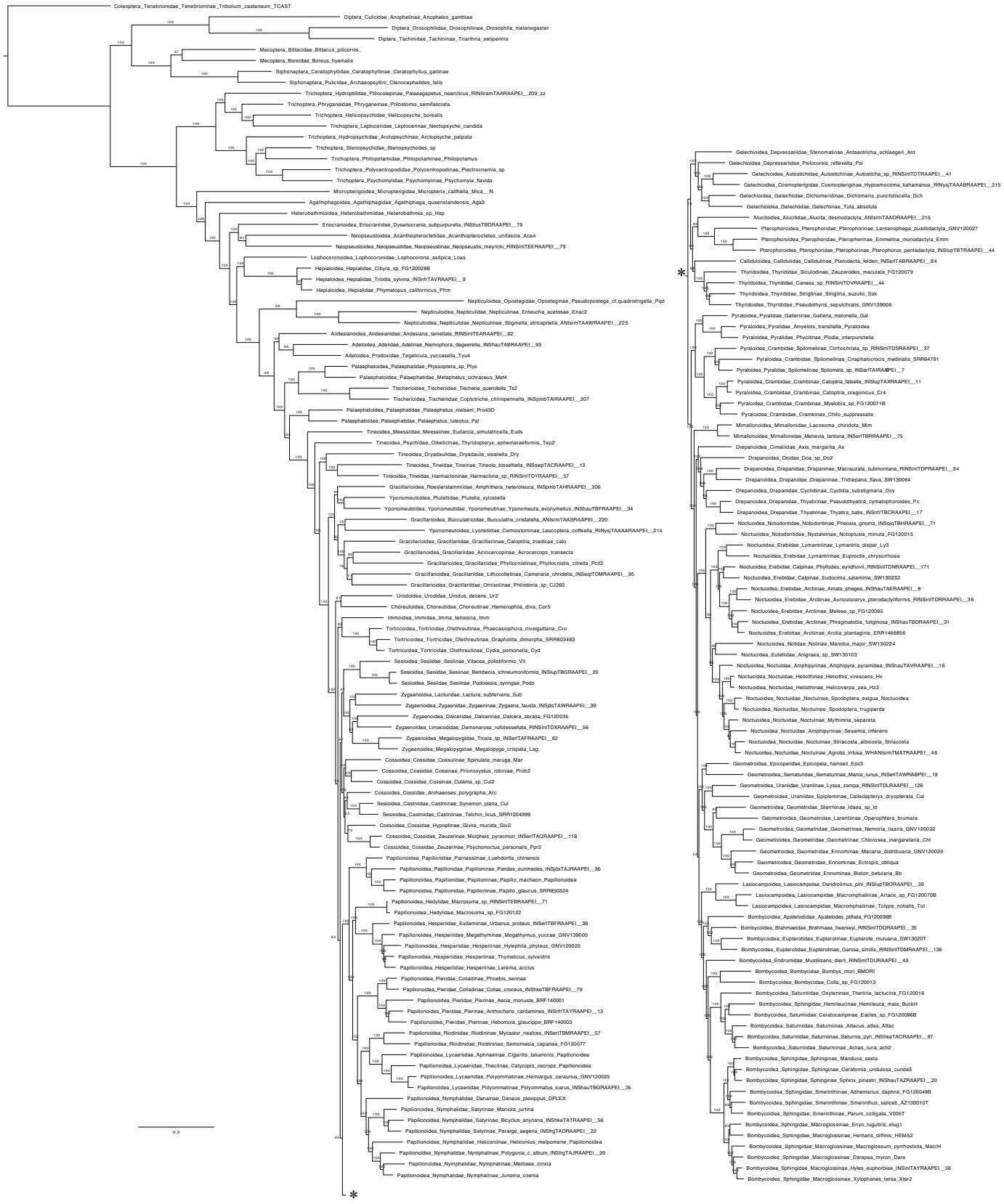

Figure S2. Maximum likelihood best tree from IQ-TREE amino acid analysis, with non-parametric bootstrap support values. lnL = -42341345.559.

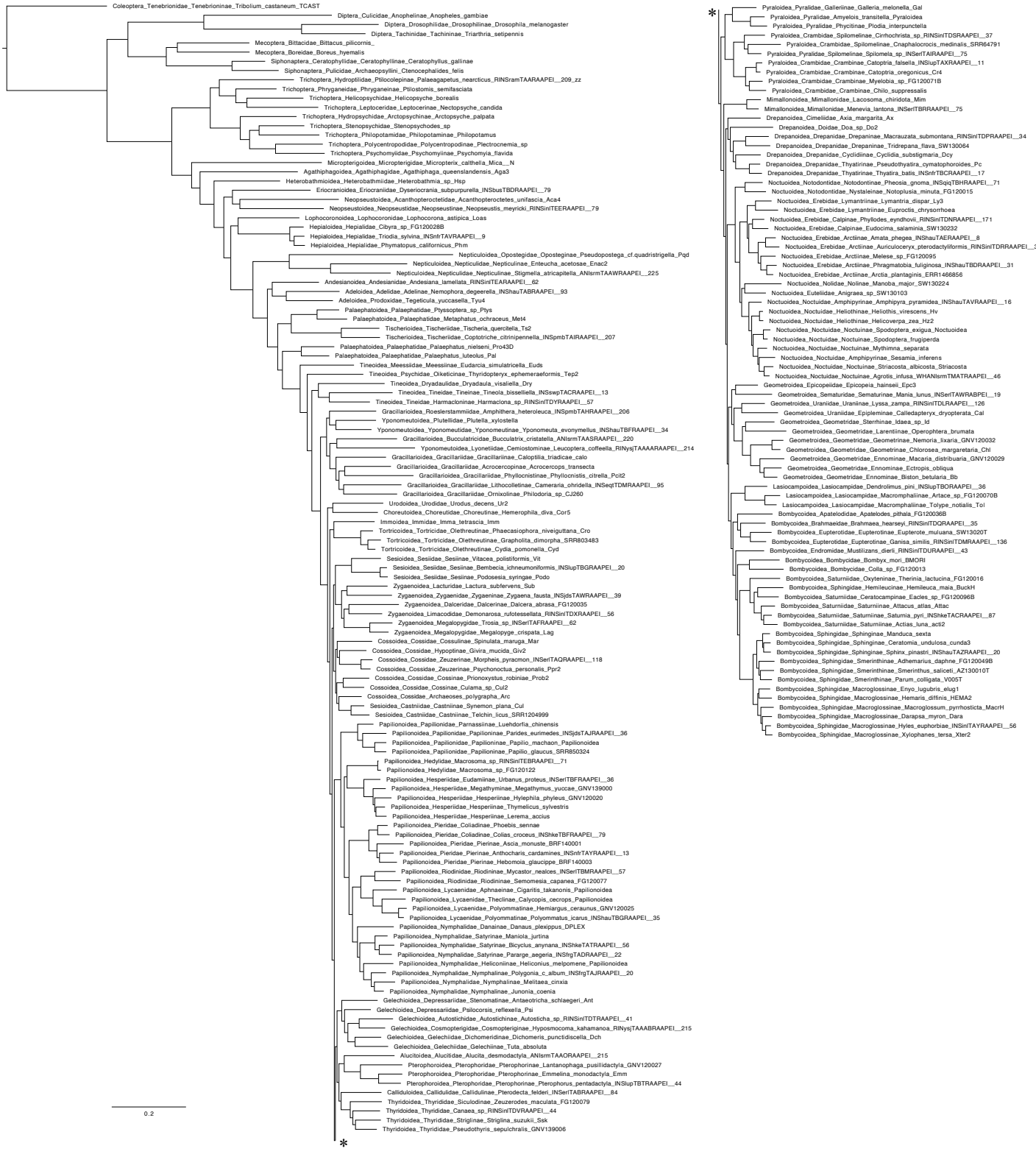

Figure S3. Maximum likelihood suboptimal tree topology from the IQ-TREE amino acid analysis.

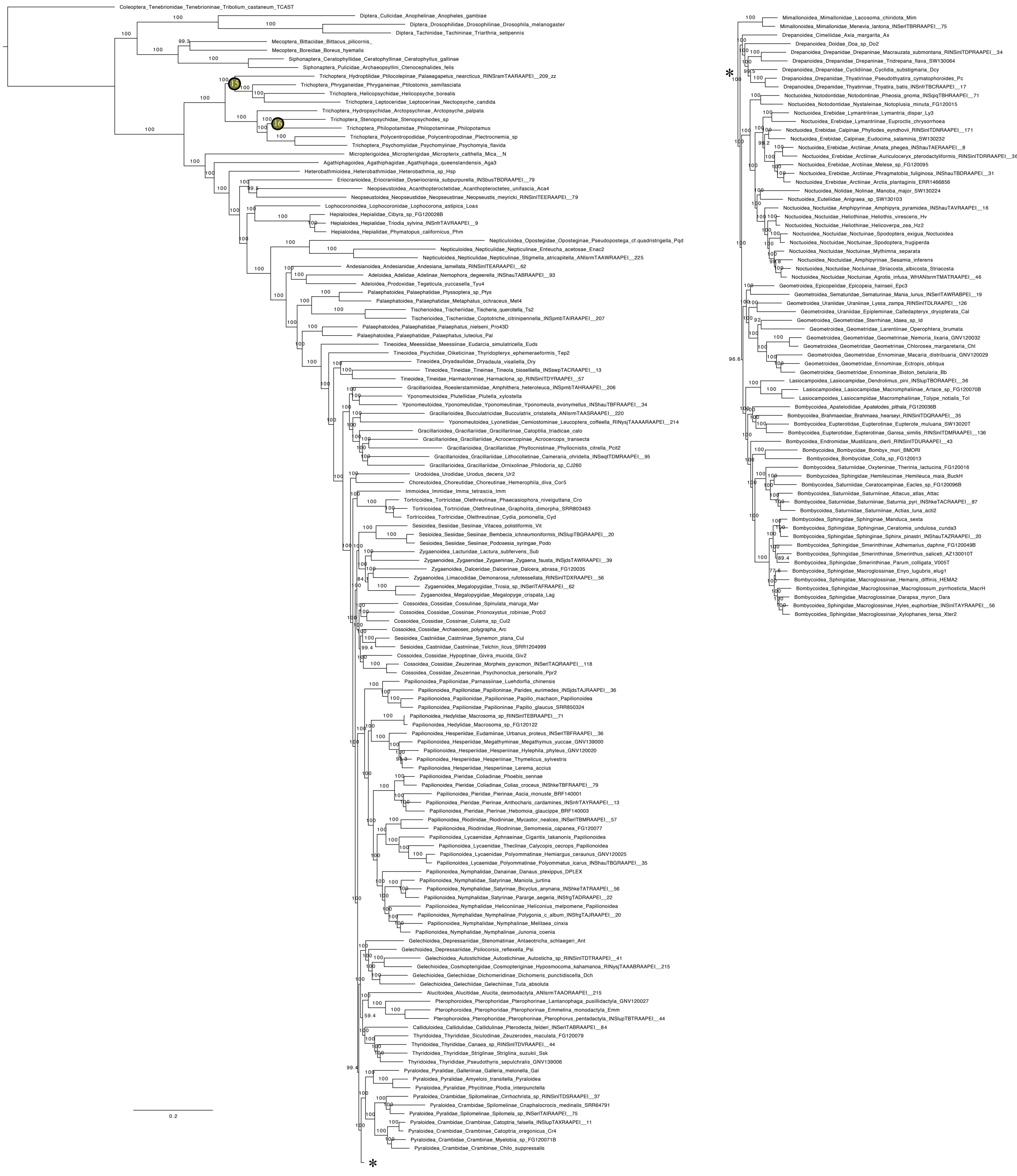

Figure S4. SH-aLRT support values mapped on the best maximum likelihood best tree from the IQ-TREE amino acid analysis.

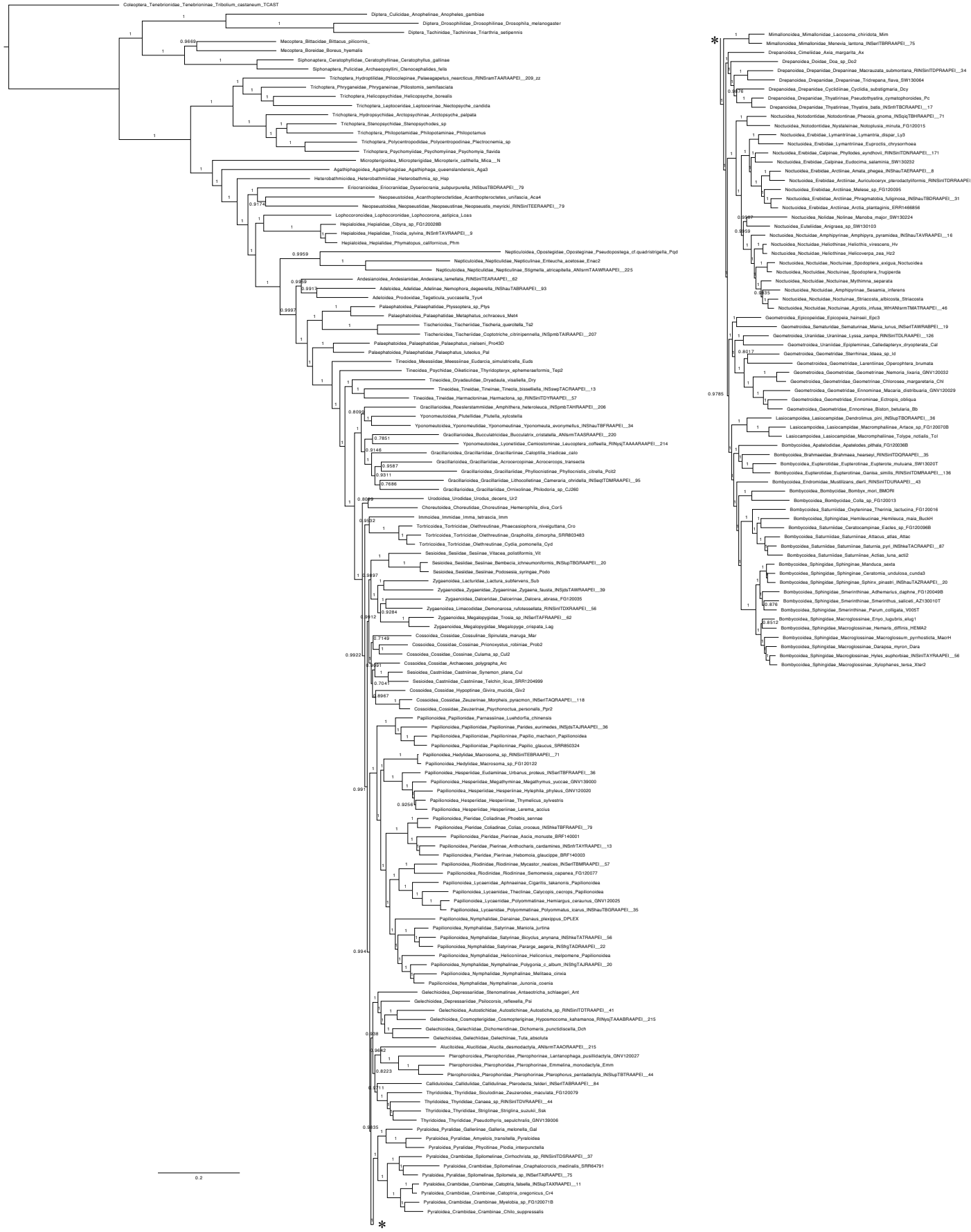

Figure S5. TBE support values mapped on the best maximum likelihood tree from the IQ-TREE amino acid analysis.

Figure S6. Maximum likelihood best tree from IQ-TREE degen1 analysis. Support values are non-parametric bootstrap values. lnL = -45121944.618.

Figure S7. SH-aLRT values mapped on the maximum likelihood topology from the IQ-TREE degen1 analysis.

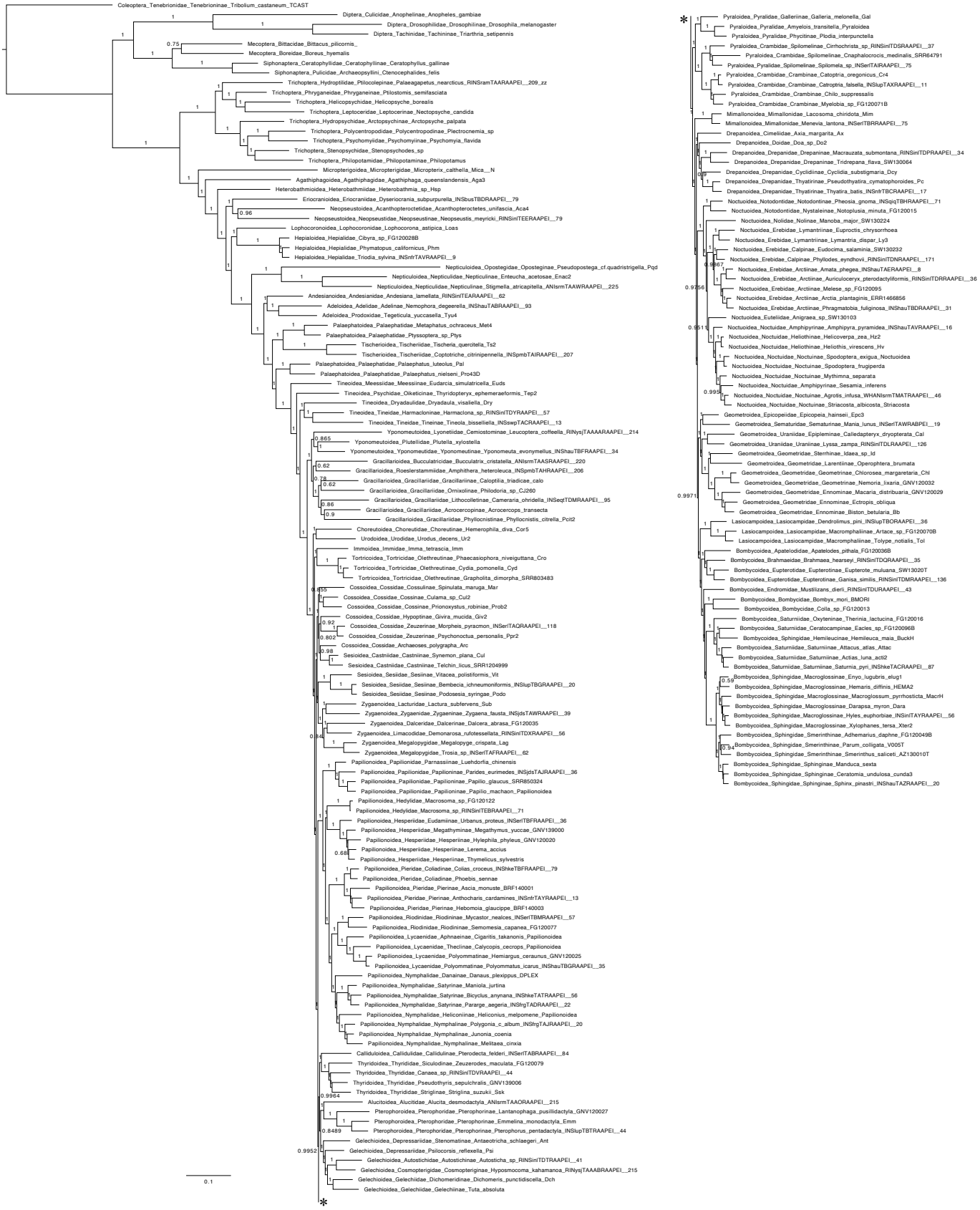

Figure S8. TBE support values mapped on the maximum likelihood topology from the IQ-TREE degen1 analysis.

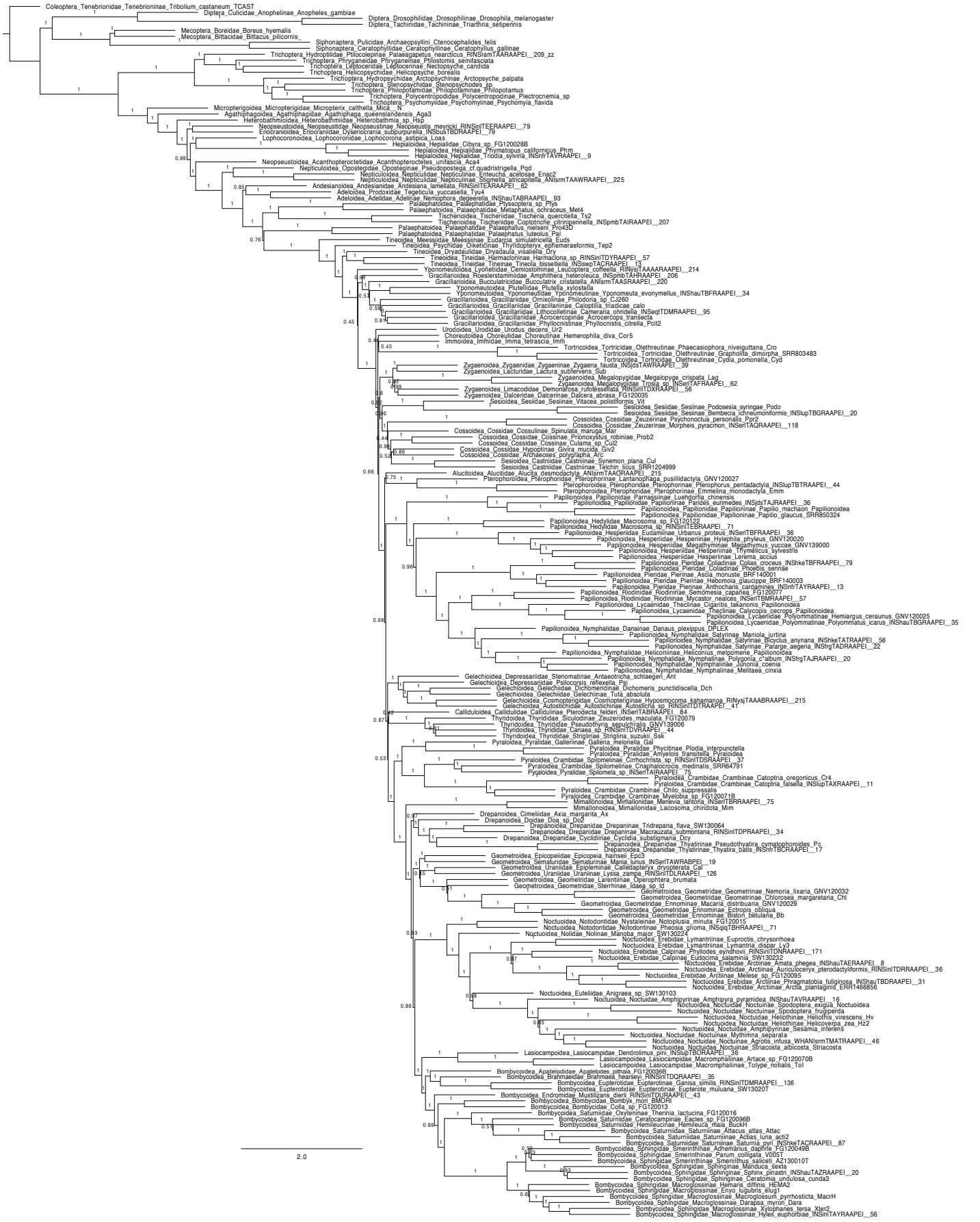

Figure S9. ASTRAL species tree derived from the best ML amino acid gene trees.

Figure S10. ASTRAL species tree derived from the amino acid consensus bootstrap trees.

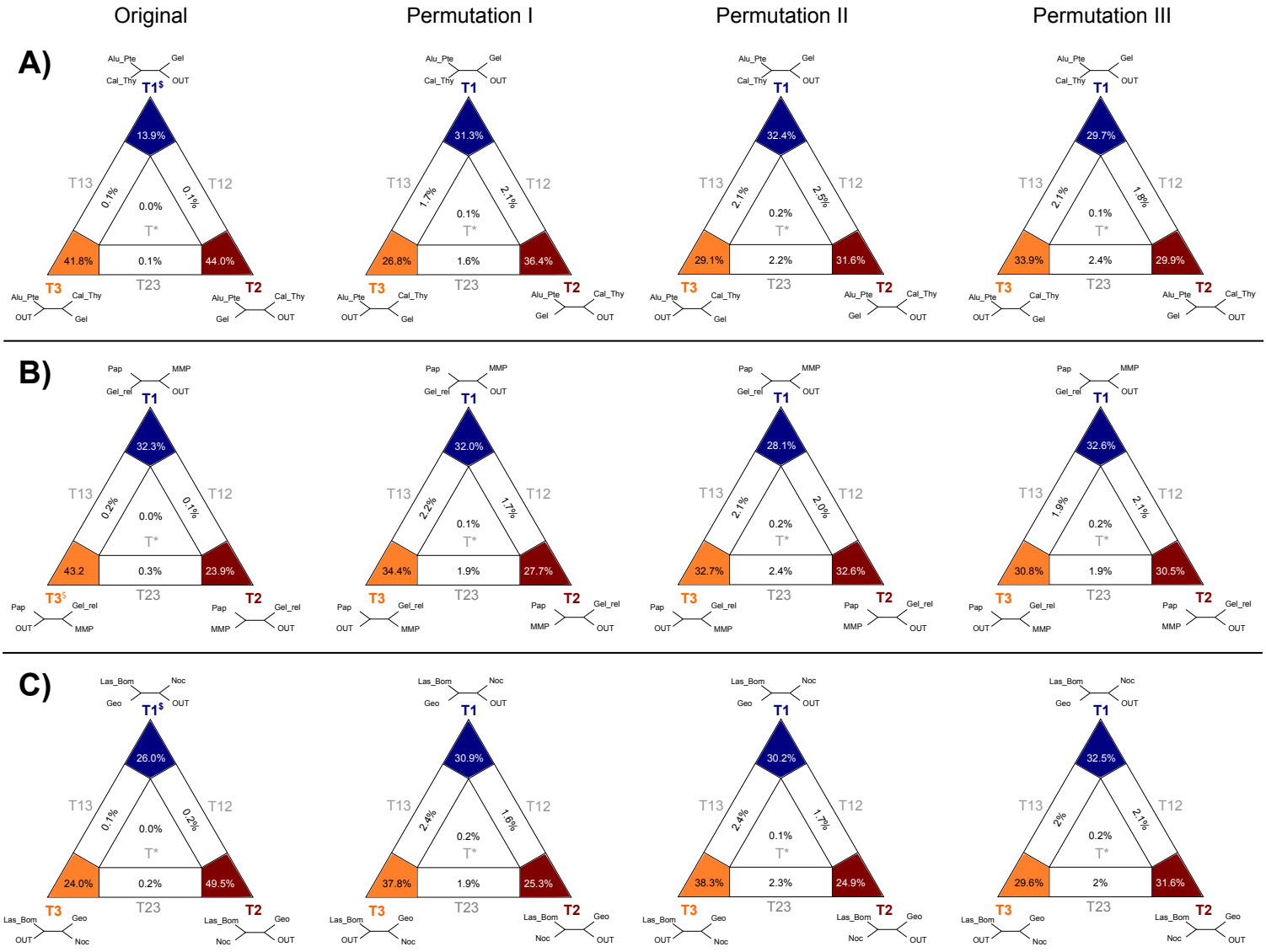

Figure S11. Results of the Four-cluster Likelihood Mapping analyses performed on the amino acid dataset, showing quartet support (in %) of all drawn quartets mapped onto 2D simplex graphs for possible topologies. 2D simplex graphs from left to right: original data, permutation I, II, III. A) Assessment of the placement of Alucitoidea+Pterophoroidea (Alu_Pte). Other groups are Gelechioidea (Gel), Calliduloidea+Thyridoidea (Cal_Thy), and an outgroup containing all other species in the dataset (OUT). B) Assessment of the placement of Papilionoidea (Pap). Other groups are Macroheterocera+Mimallonoidea +Pyraloidea (MMP), Gelechioidea and relatives (Gel_rel), and an outgroup containing non-obtectomeran Lepidoptera (OUT). C) Assessment of the placement of Lasiocampoidea+Bombycoidea (Las_Bom). Other groups are Geometroidea (Geo), Noctuoidea (Noc), and an outgroup containing Drepanoidea (OUT). T1, T2, T3 show unambiguous quartet topologies (area 1, 2, 3 in IQ-TREE), T12 (area 4 in IQ-TREE), T13 (area 6 in IQ-TREE) and T23 (area 5 in IQ-TREE) show partially resolved quartets, quartets in T* (area 7 in IQ-TREE) unresolved. $ indicates the topology supported in the best amino acid ML tree.

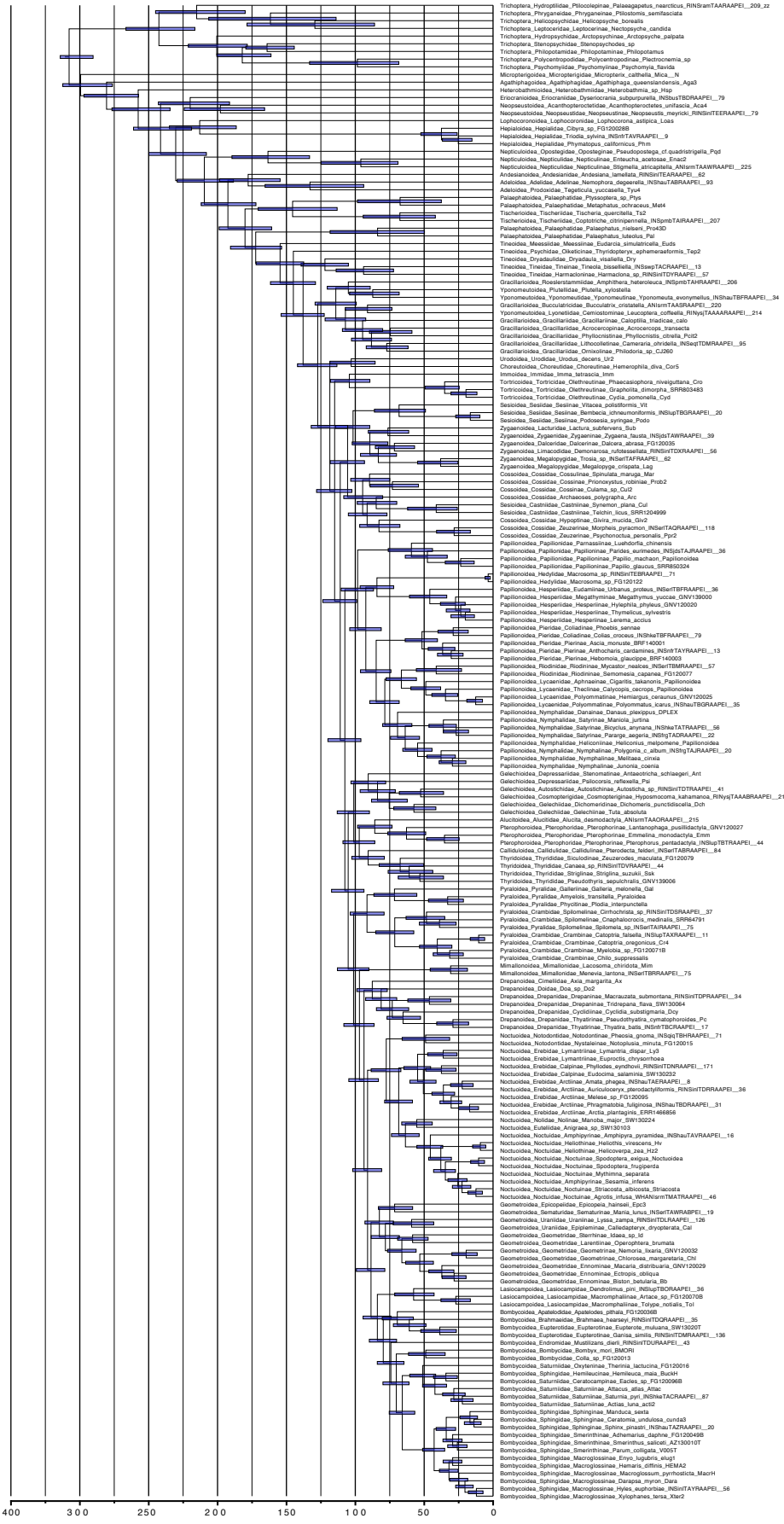

Figure S12. Divergence time estimates from the 16-fossil MCMCTree analysis with uncorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

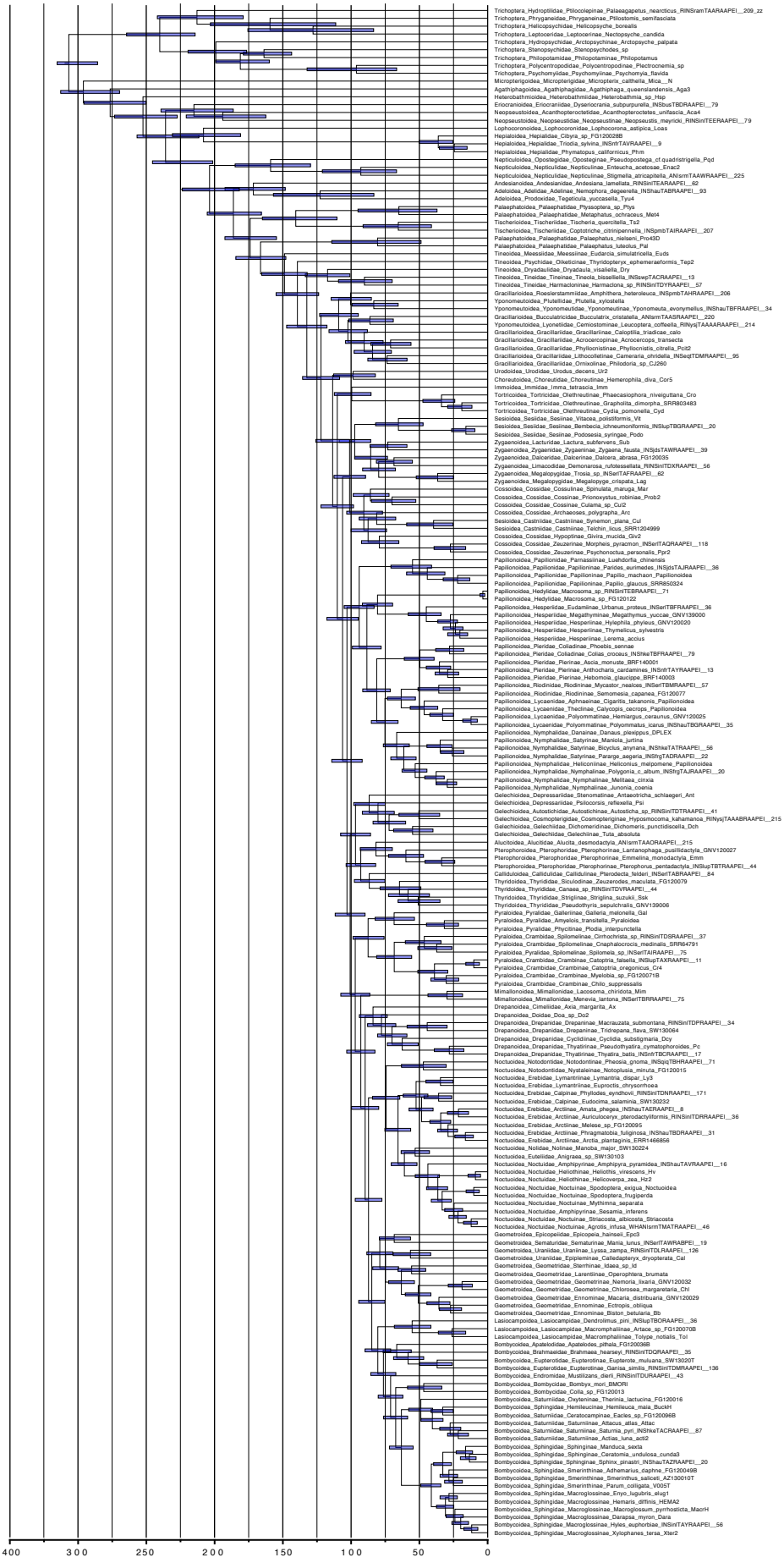

Figure S13. Divergence time estimates from the 16-fossil MCMCTree analysis with uncorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

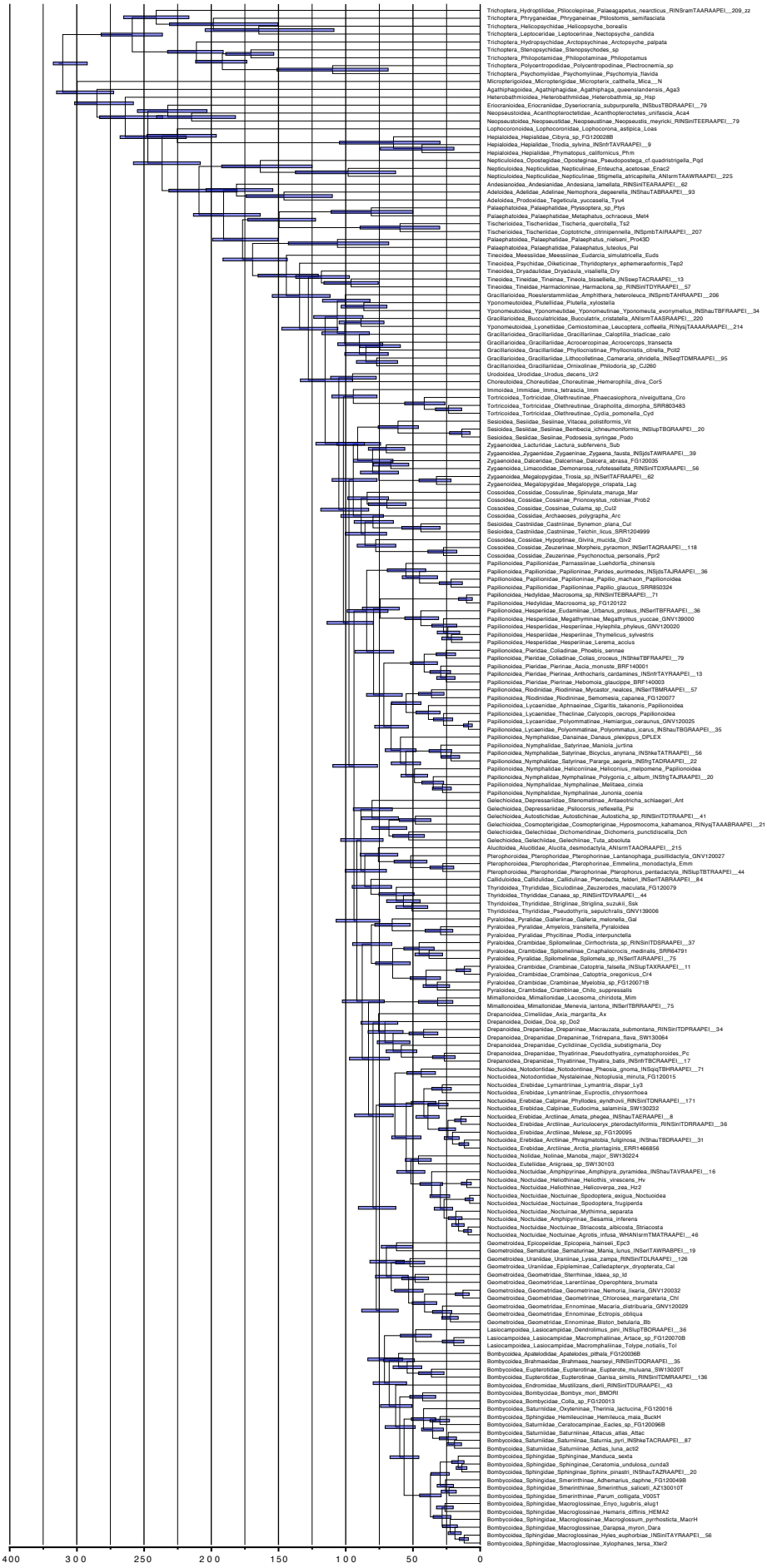

Figure S14. Divergence time estimates from the 16-fossil MCMCTree analysis with autocorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

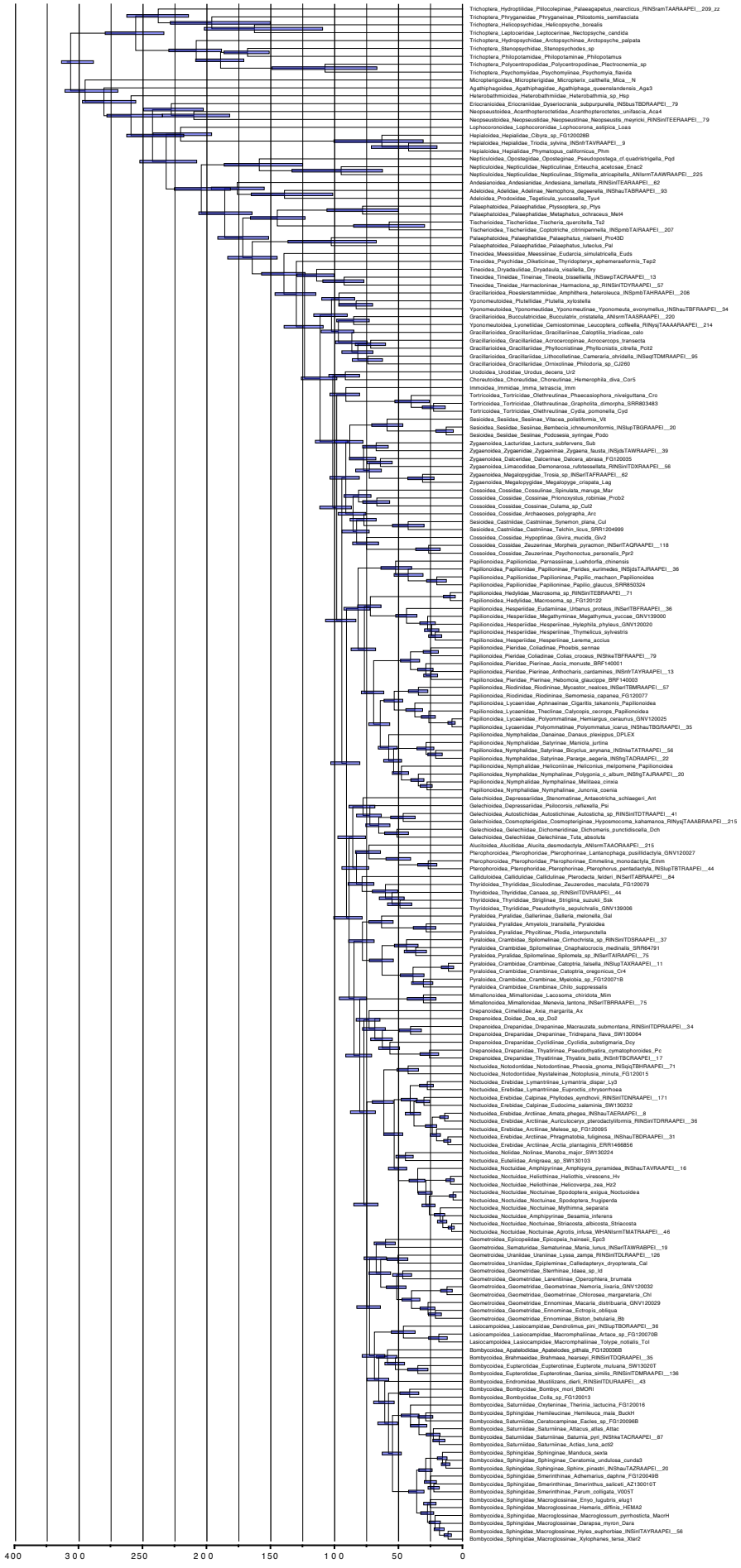

Figure S15. Divergence time estimates from the 16-fossil MCMCTree analysis with autocorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

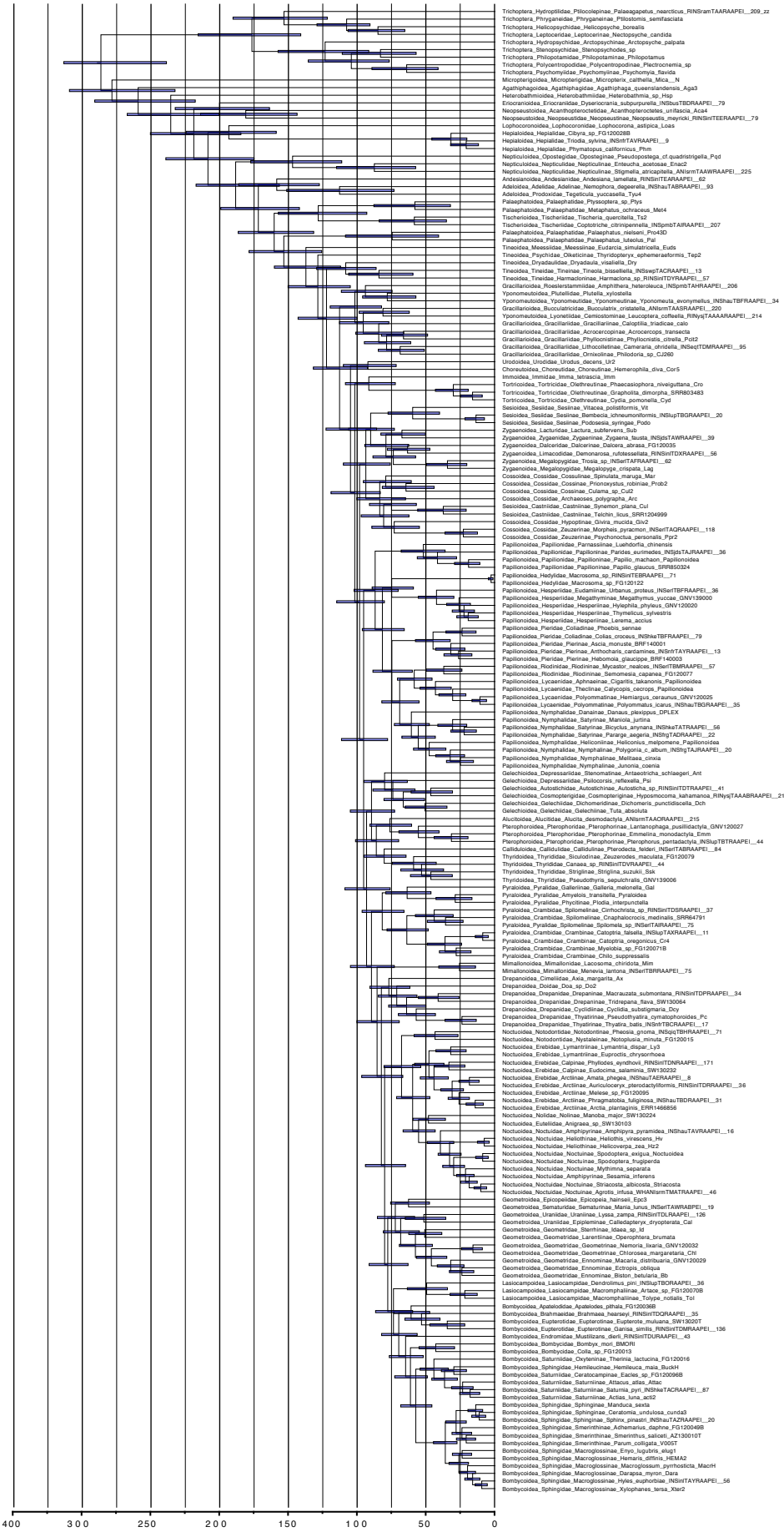

Figure S16. Divergence time estimates from the 3-fossil MCMCTree analysis with uncorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

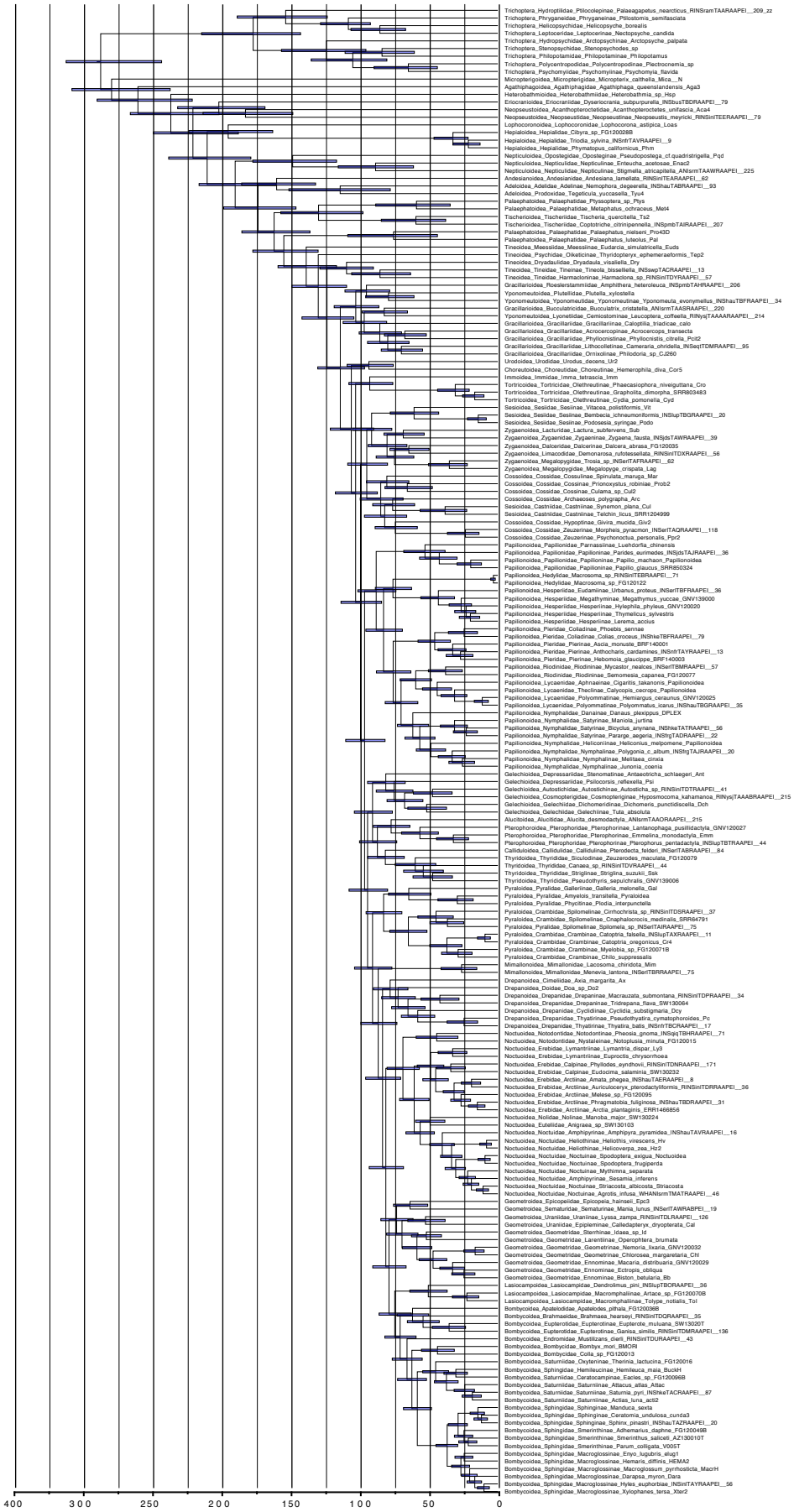

Figure S17. Divergence time estimates from the 3-fossil MCMCTree analysis with uncorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

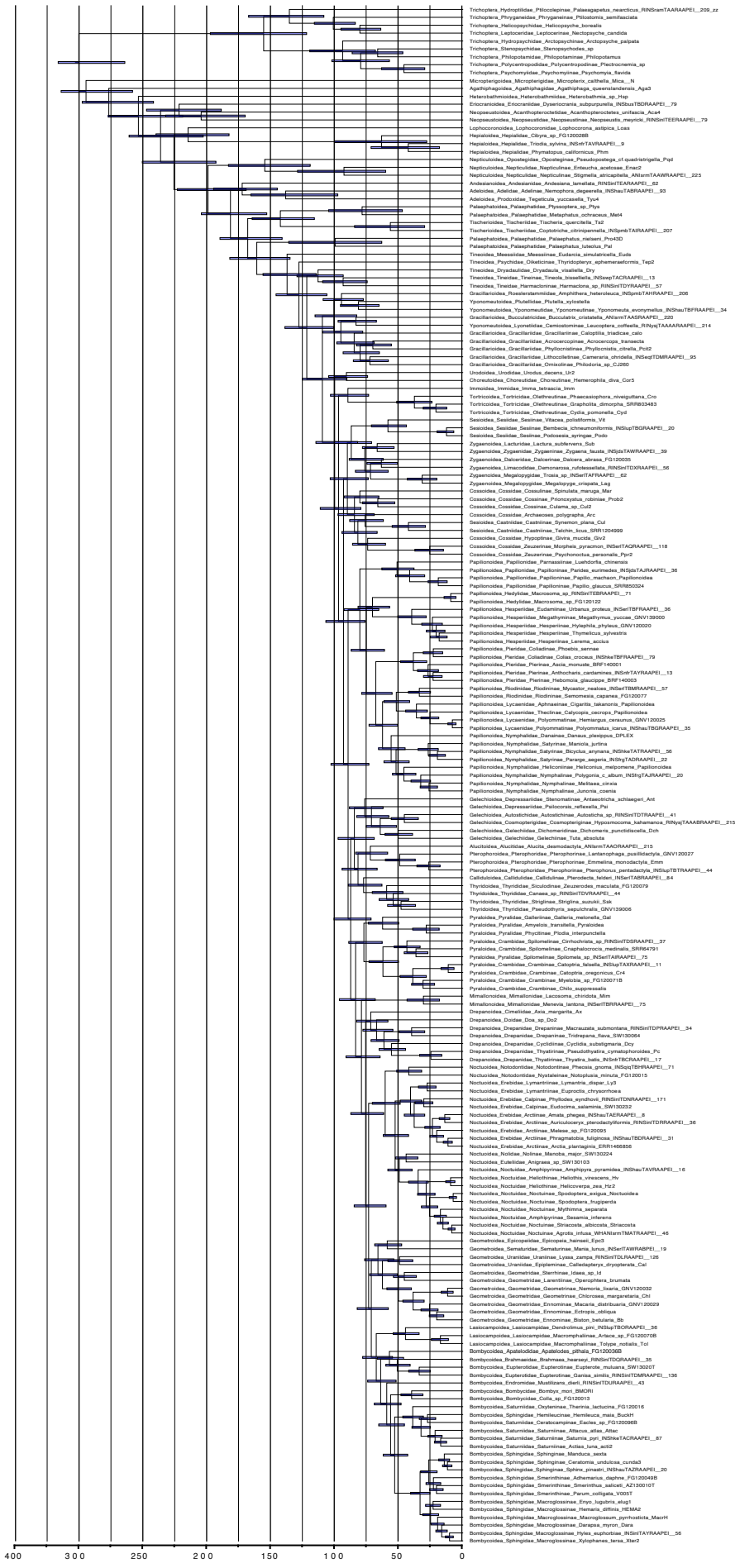

Figure S18. Divergence time estimates from the 3-fossil MCMCTree analysis with autocorrelated rates and uniform priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

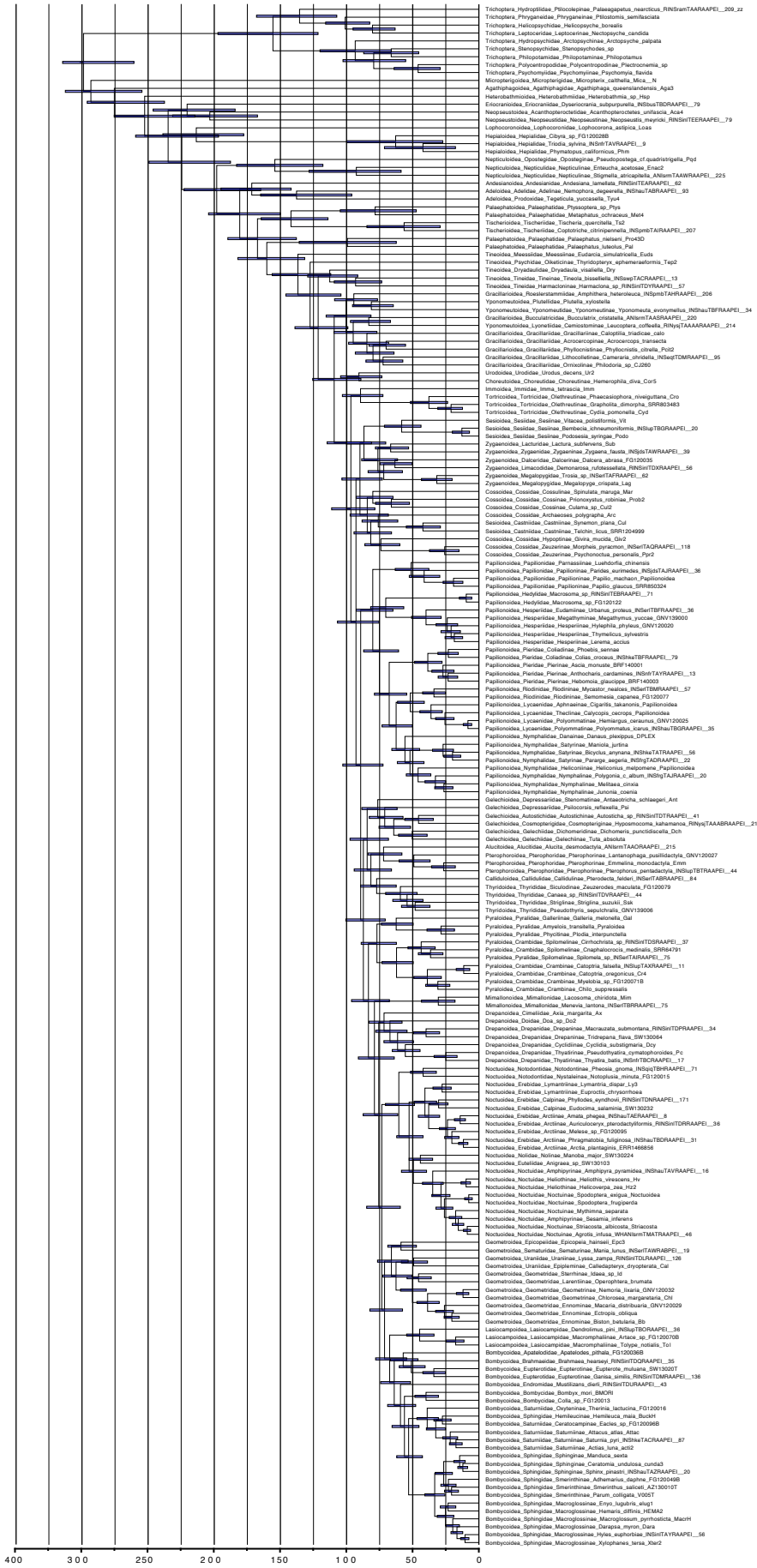

Figure S19. Divergence time estimates from the 3-fossil MCMCTree analysis with autocorrelated rates and Cauchy priors, based on the topology from the ML amino acid analysis. Node bars represent age ranges (95% credibility intervals) in Ma.

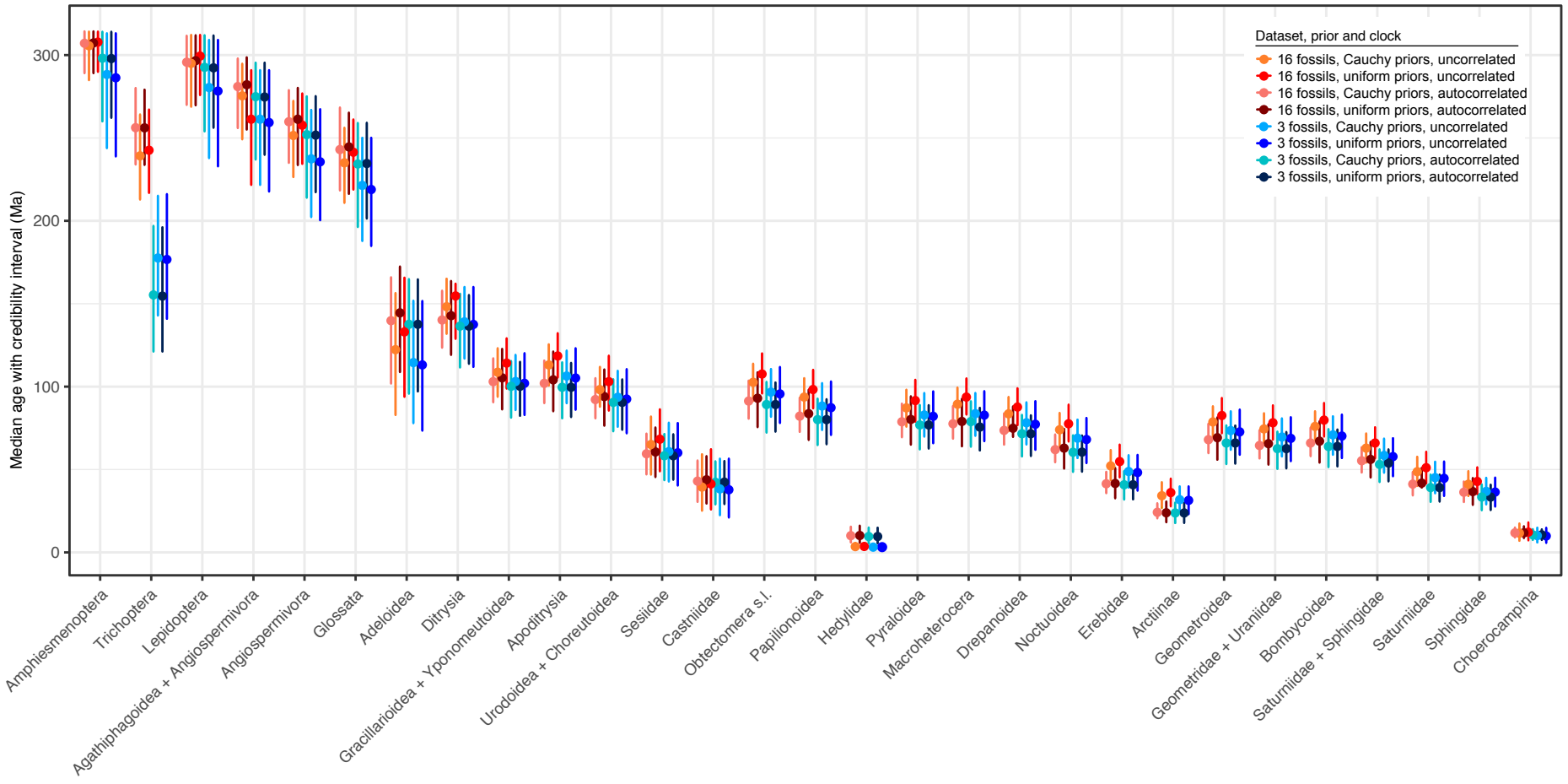

Figure S20. Comparison of ages of major clades in Lepidoptera and Trichoptera.

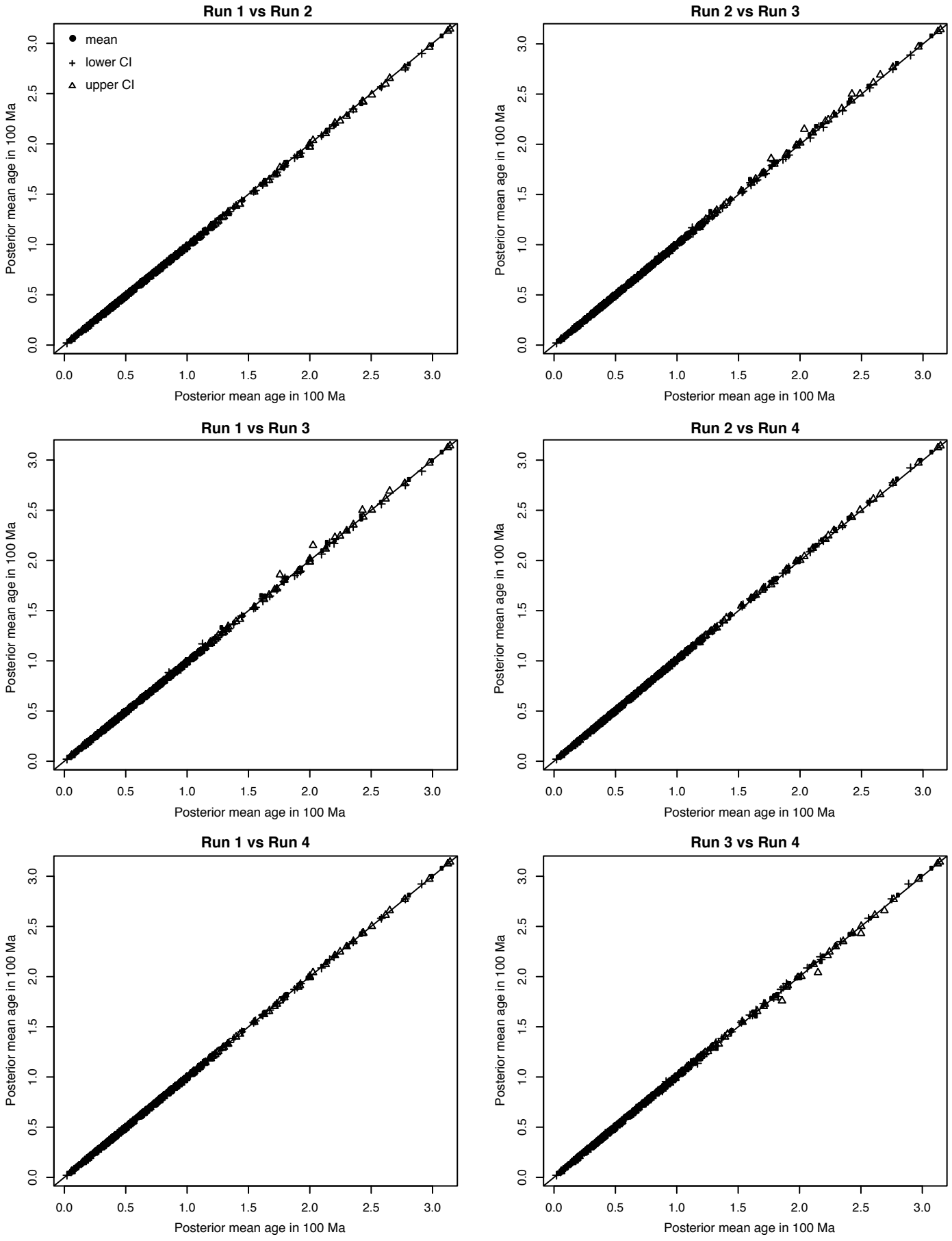

Figure S21. Convergence plots for the 16-fossil analysis with uncorrelated rates and uniform priors showing the relationship between posterior means and confidence intervals among the four runs.

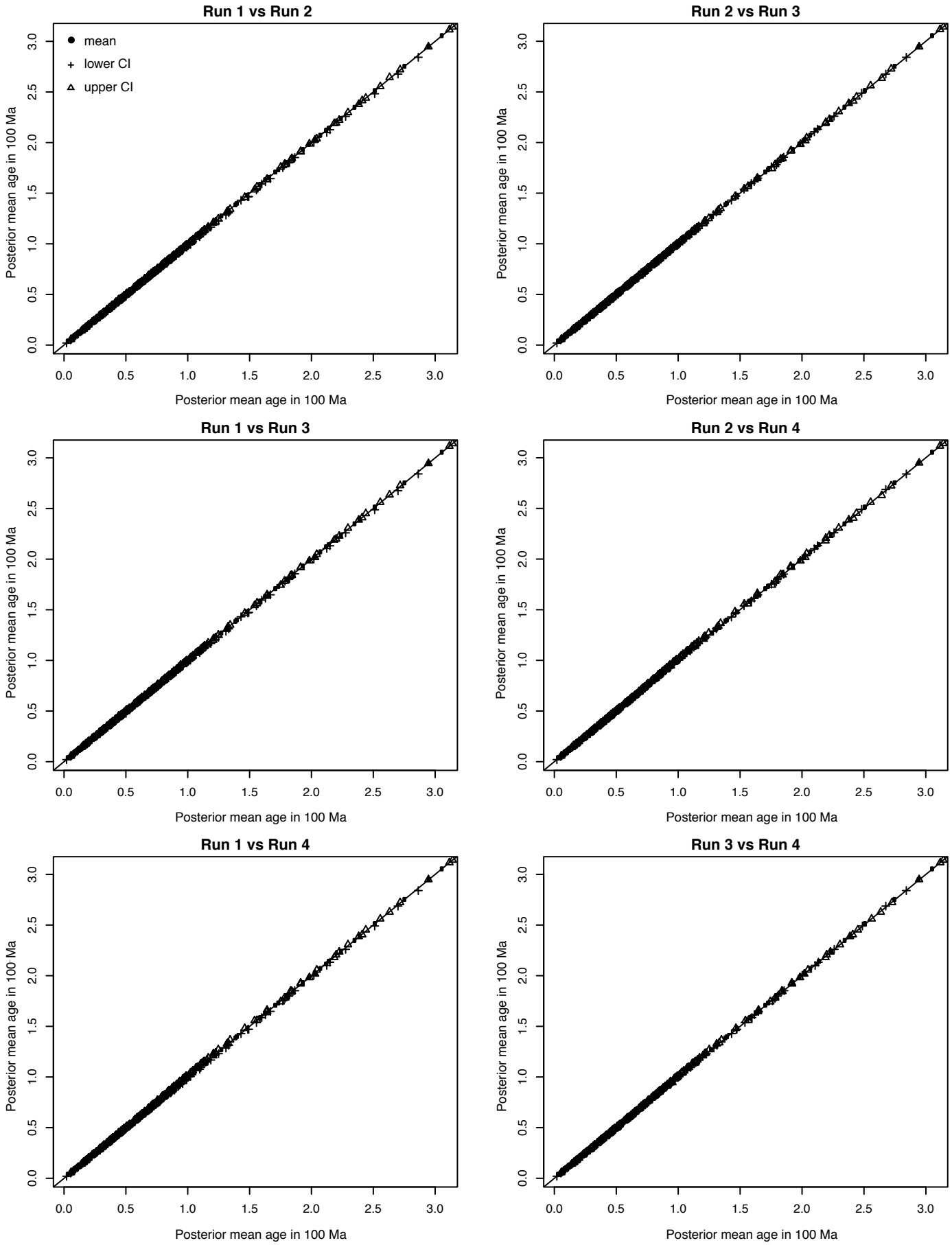

Figure S22. Convergence plots for the 16-fossil analysis with uncorrelated rates and Cauchy priors showing the relationship between posterior means and confidence intervals among the four runs.

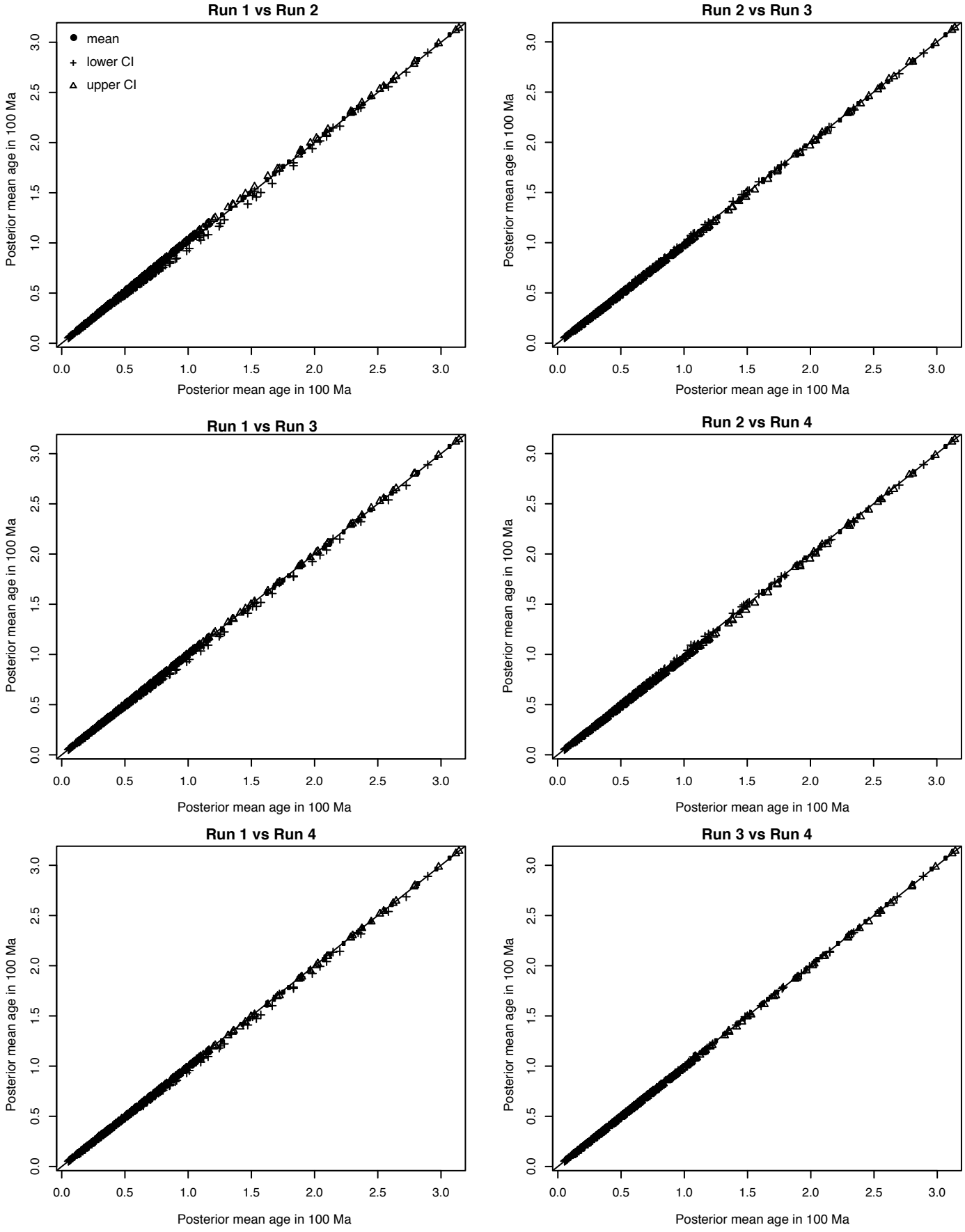

Figure S23. Convergence plots for the 16-fossil analysis with autocorrelated rates and uniform priors showing the relationship between posterior means and confidence intervals among the four runs.

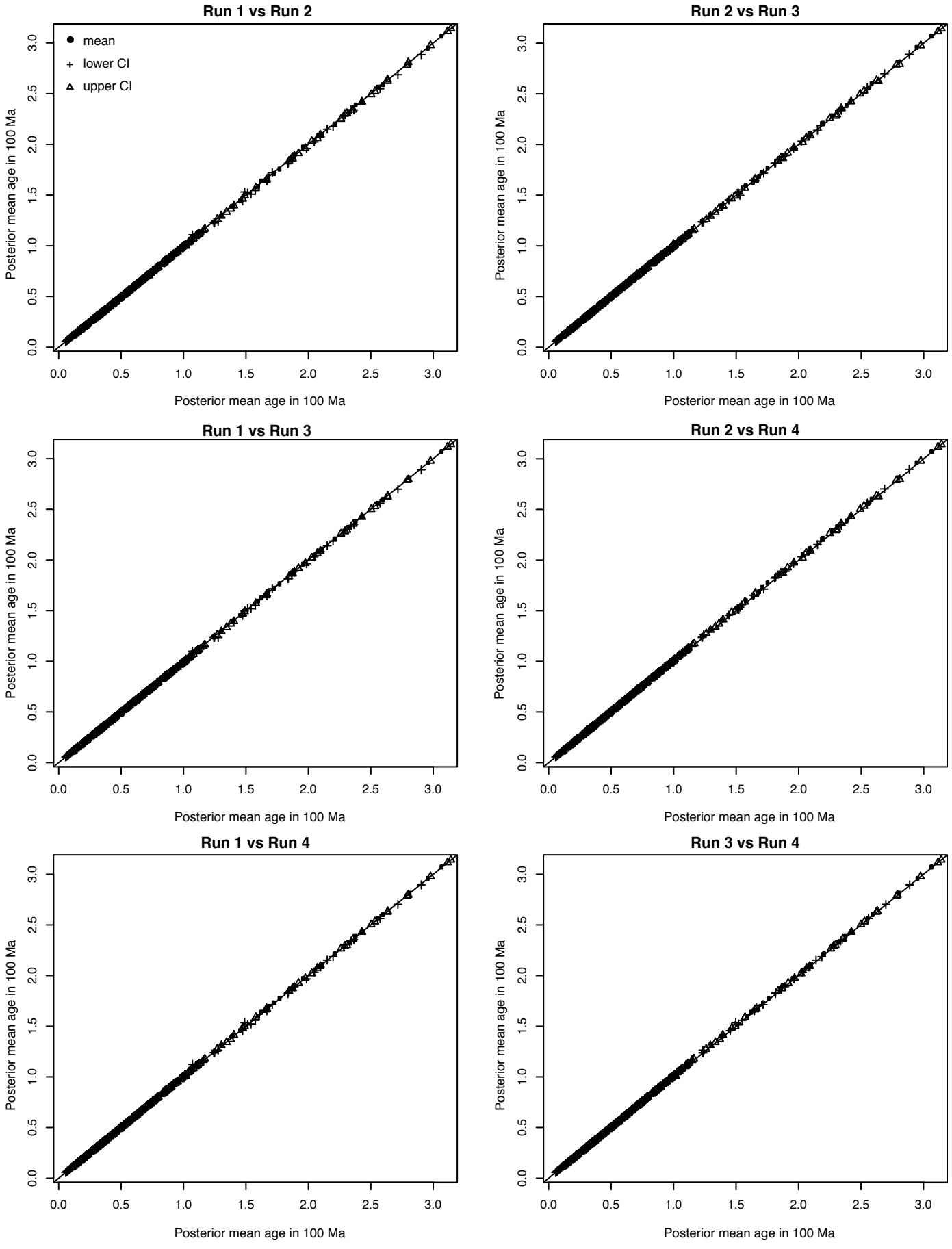

Figure S24. Convergence plots for the 16-fossil analysis with autocorrelated rates and Cauchy priors showing the relationship between posterior means and confidence intervals among the four runs.

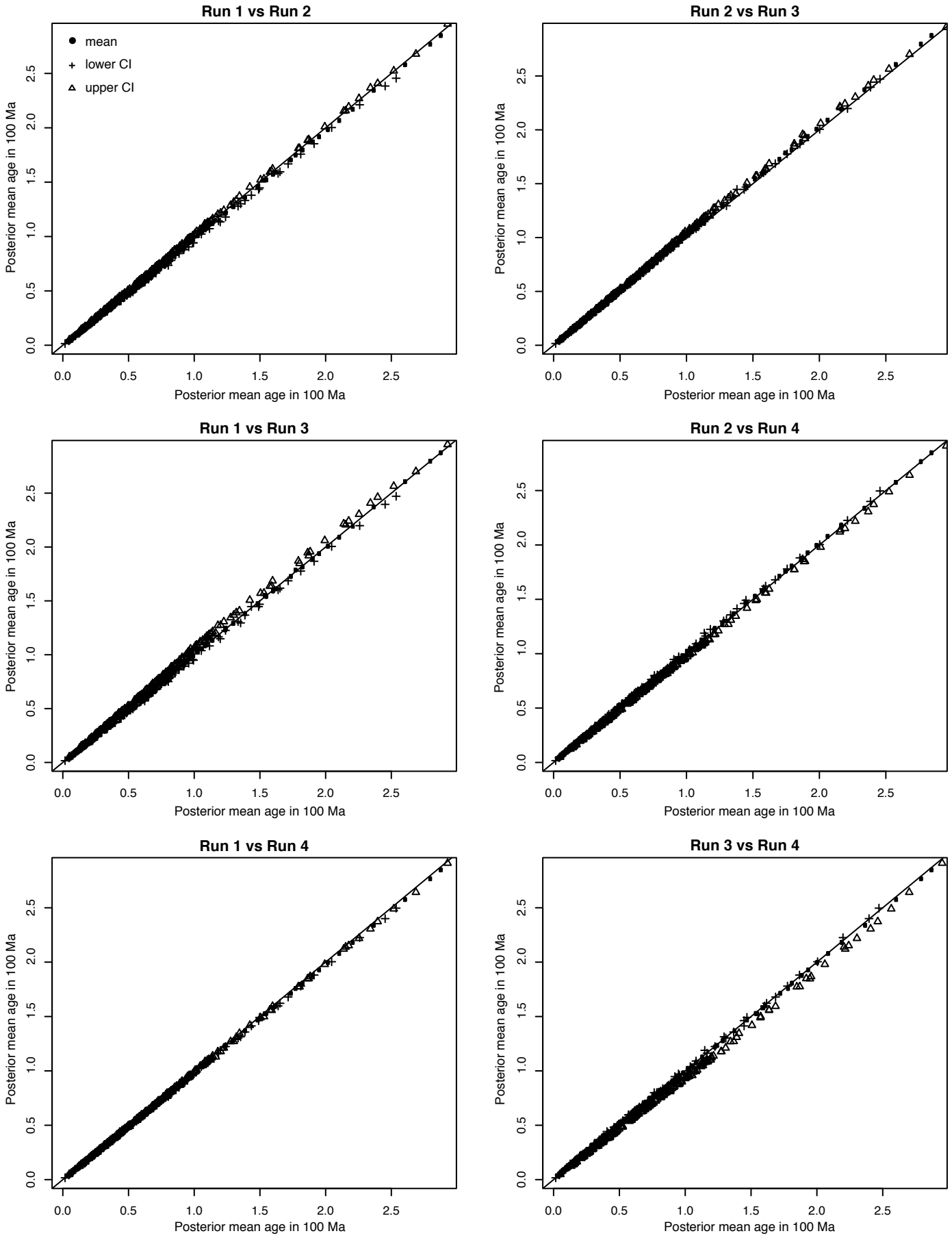

Figure S25. Convergence plots for the 3-fossil analysis with uncorrelated rates and uniform priors showing the relationship between posterior means and confidence intervals among the four runs.

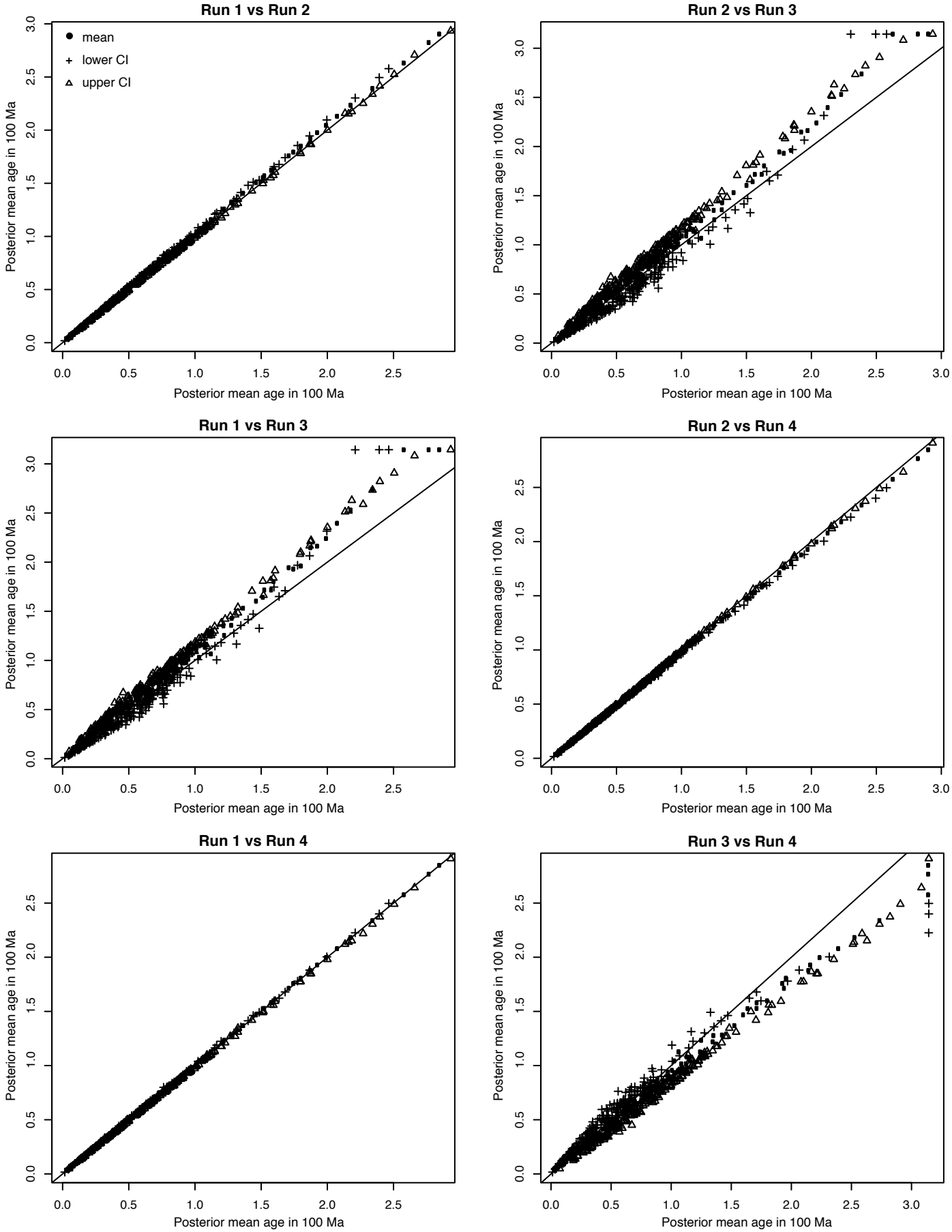

Figure S26. Convergence plots for the 3-fossil analysis with uncorrelated rates and Cauchy priors showing the relationship between posterior means and confidence intervals among the four runs.

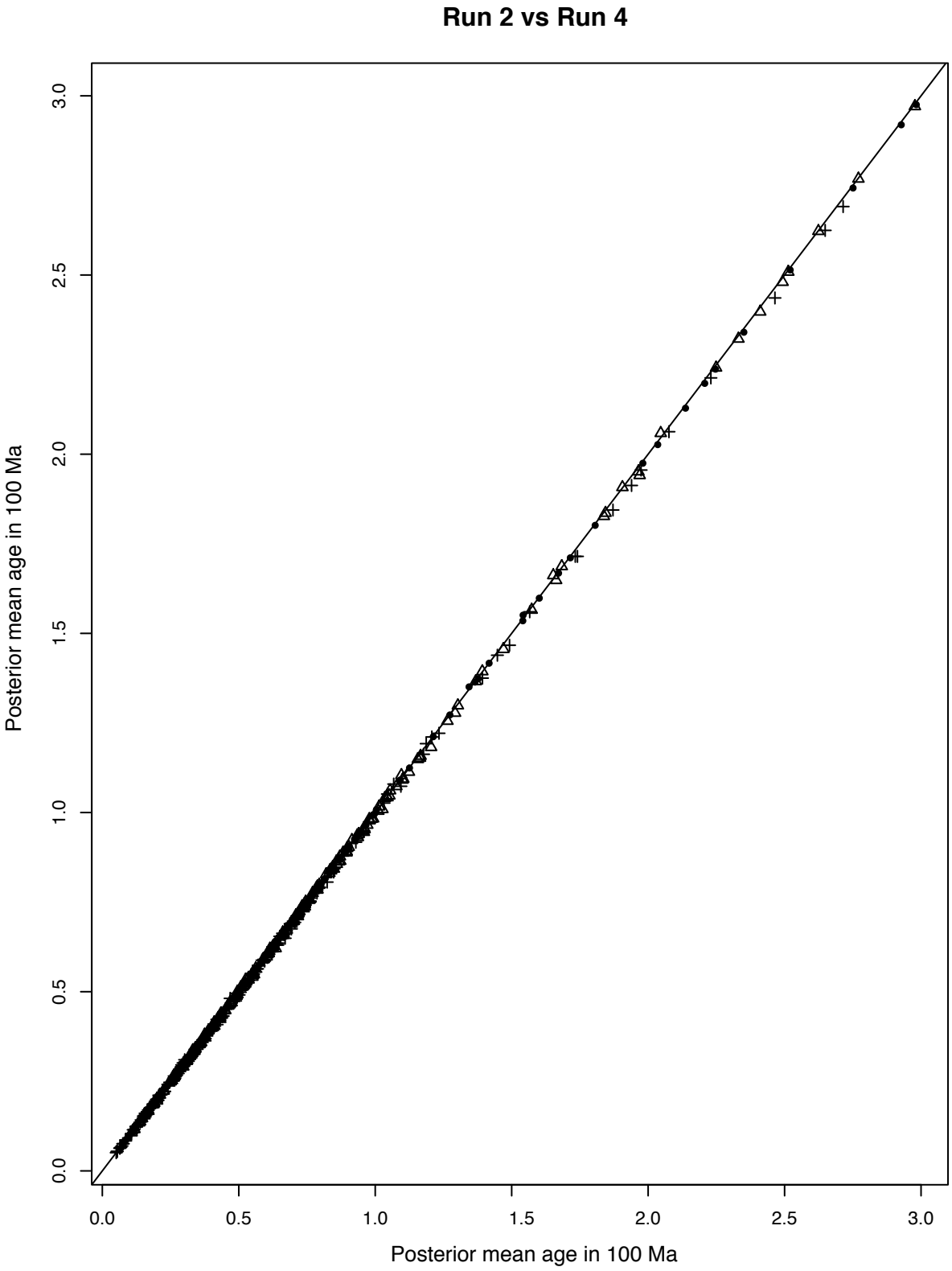

Figure S27. Convergence plot for the 3-fossil analysis with autocorrelated rates and uniform priors showing the relationship between posterior means and confidence intervals among the two available runs.

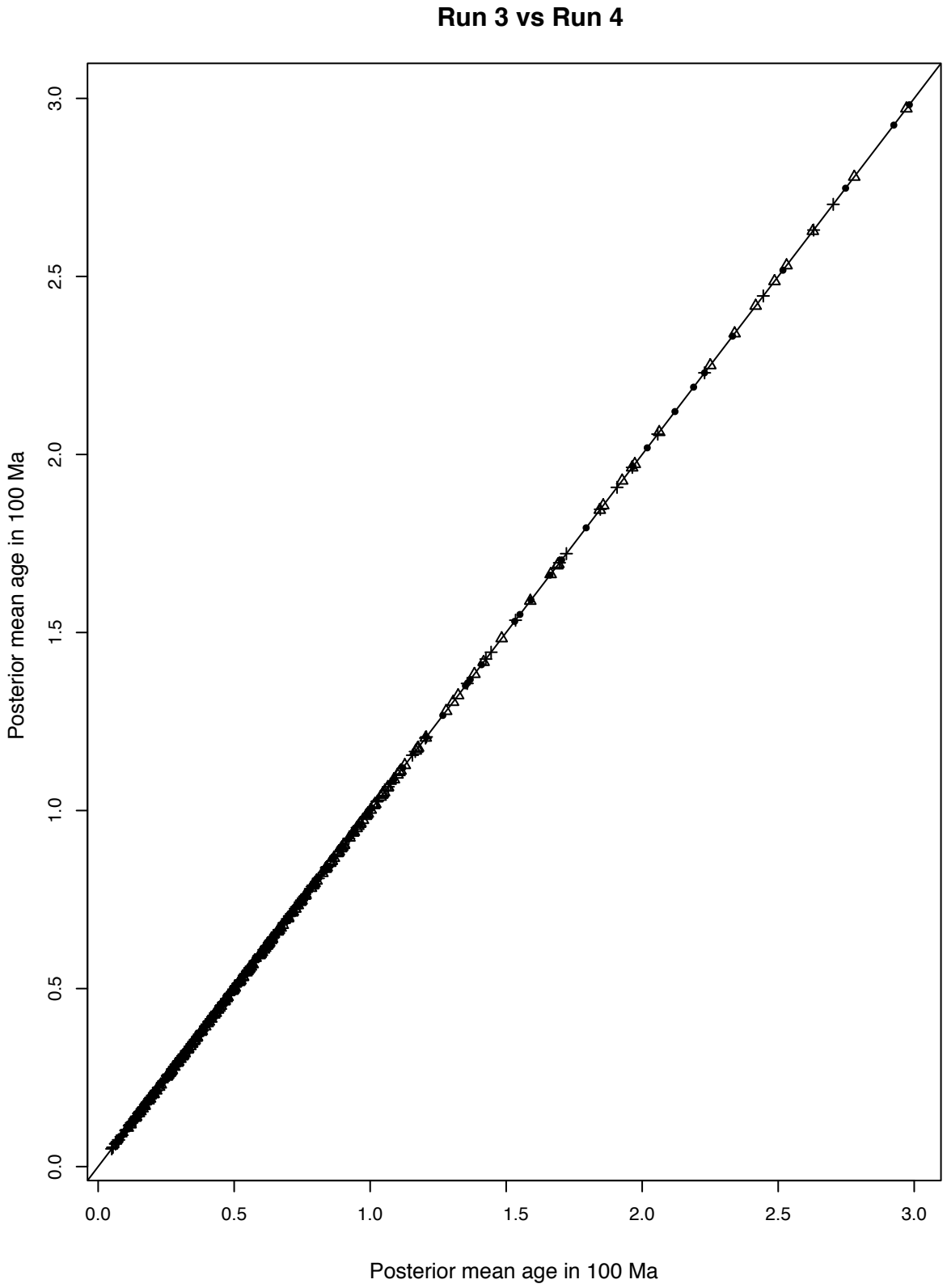

Figure S28. Convergence plot for the 3-fossil analysis with autocorrelated rates and Cauchy priors showing the relationship between posterior means and confidence intervals among the two available runs.

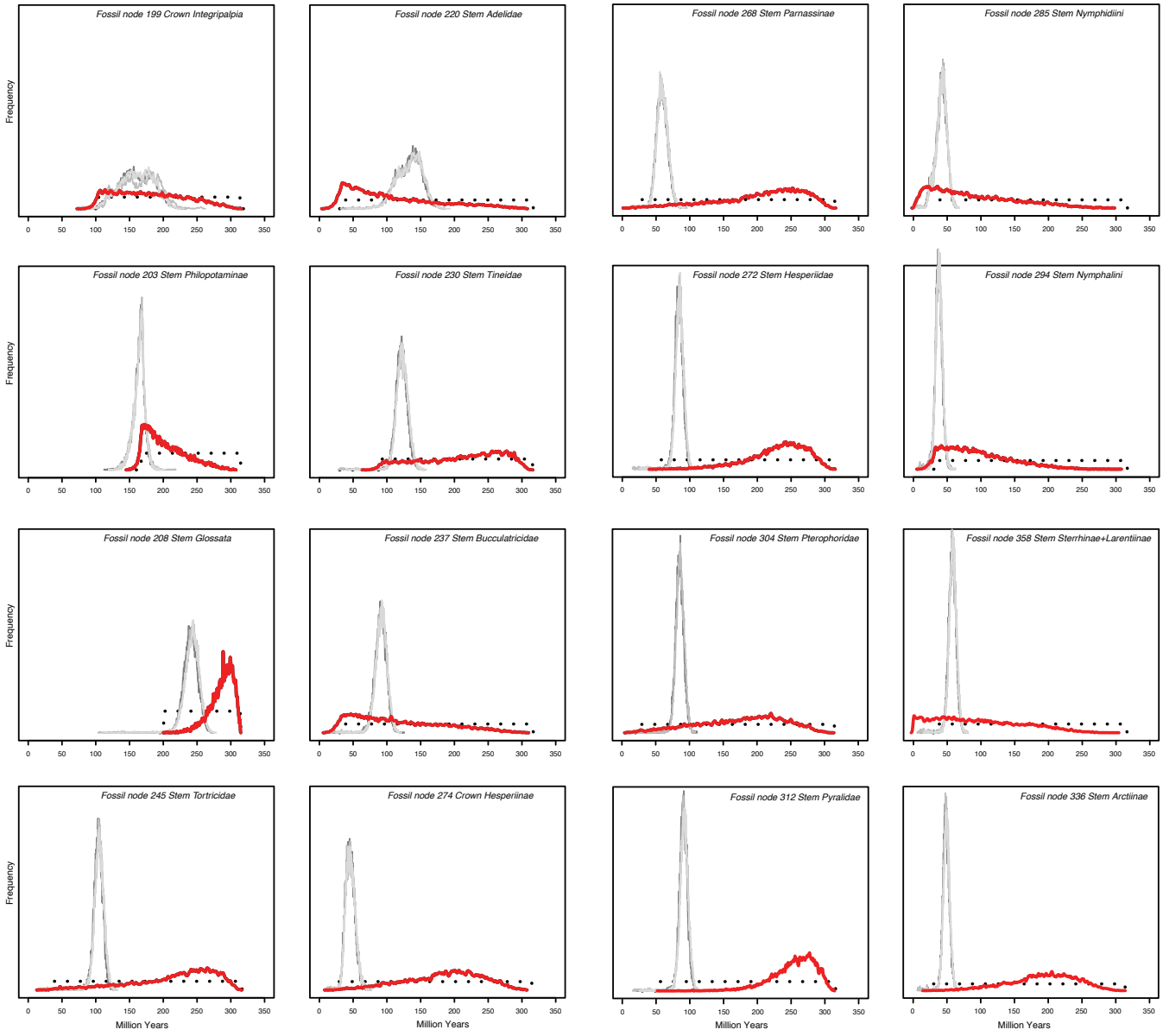

Figure S29. Density plots for the 16-fossil analysis with uncorrelated rates and uniform priors showing the relationship between sampling frequency and the age of a particular clade.

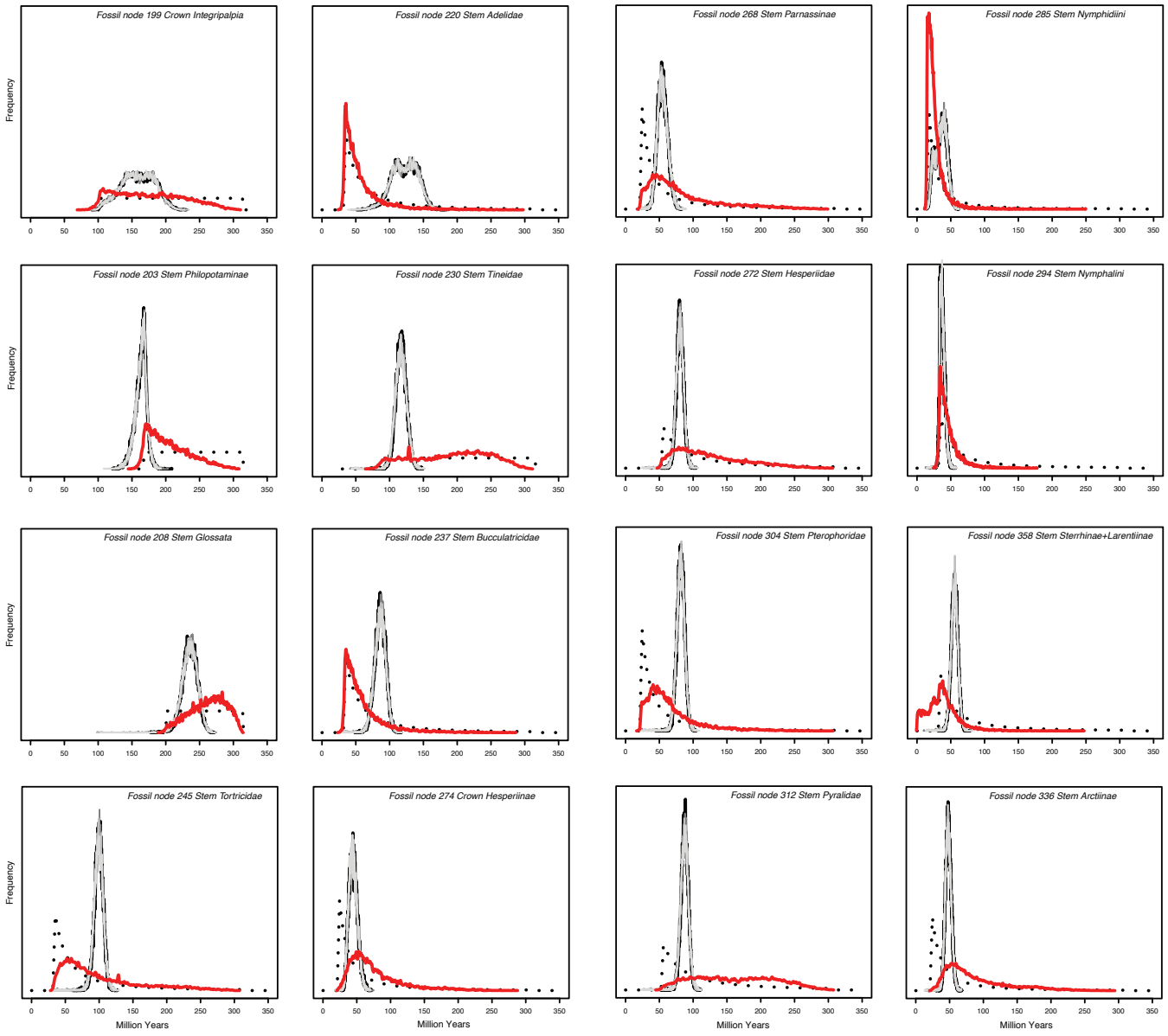

Figure S30. Density plots for the 16-fossil analysis with uncorrelated rates and Cauchy priors showing the relationship between sampling frequency and the age of a particular clade.

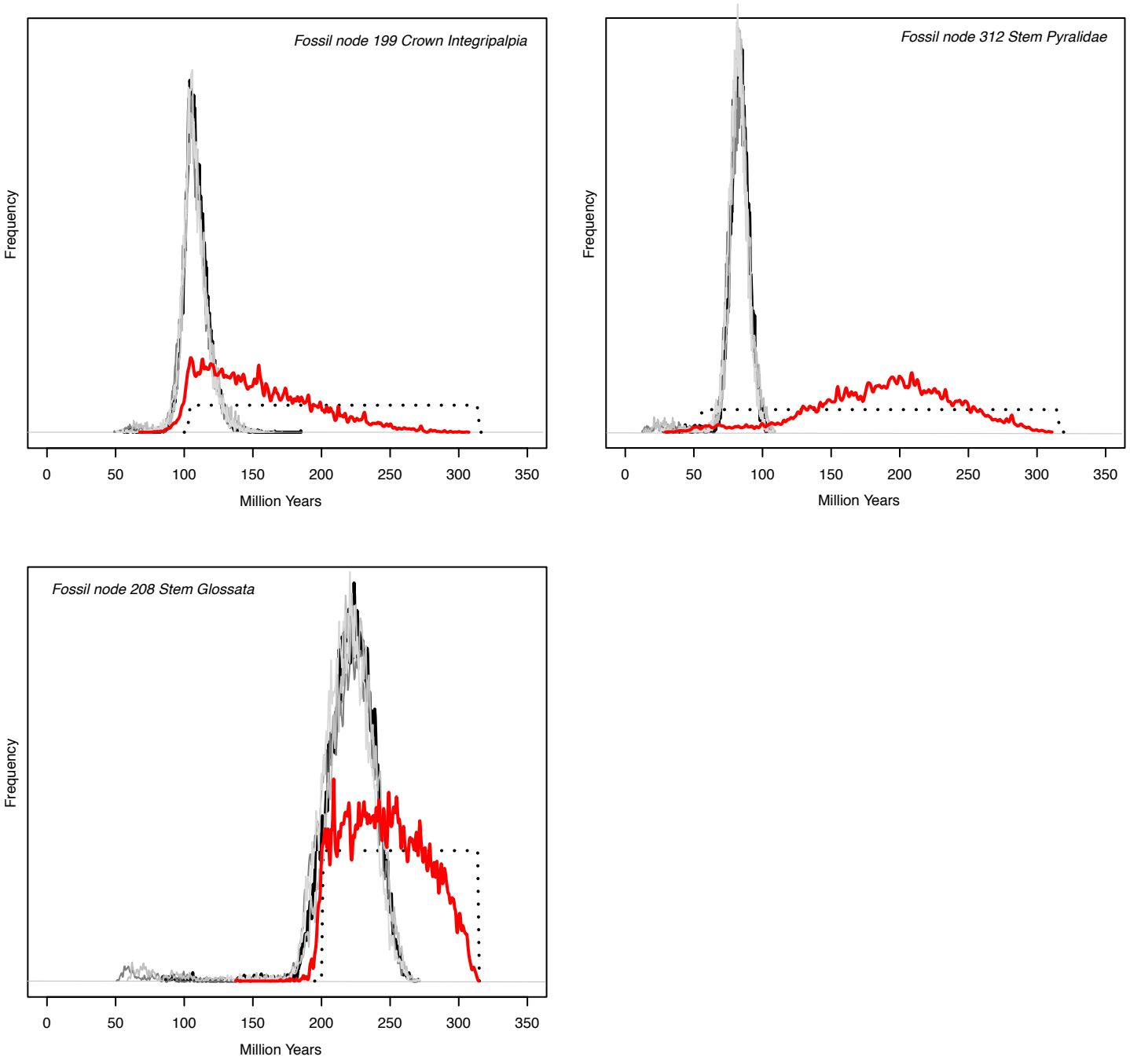

Figure S31. Density plots for the 3-fossil analysis with uncorrelated rates and uniform priors showing the relationship between sampling frequency and the age of a particular clade.

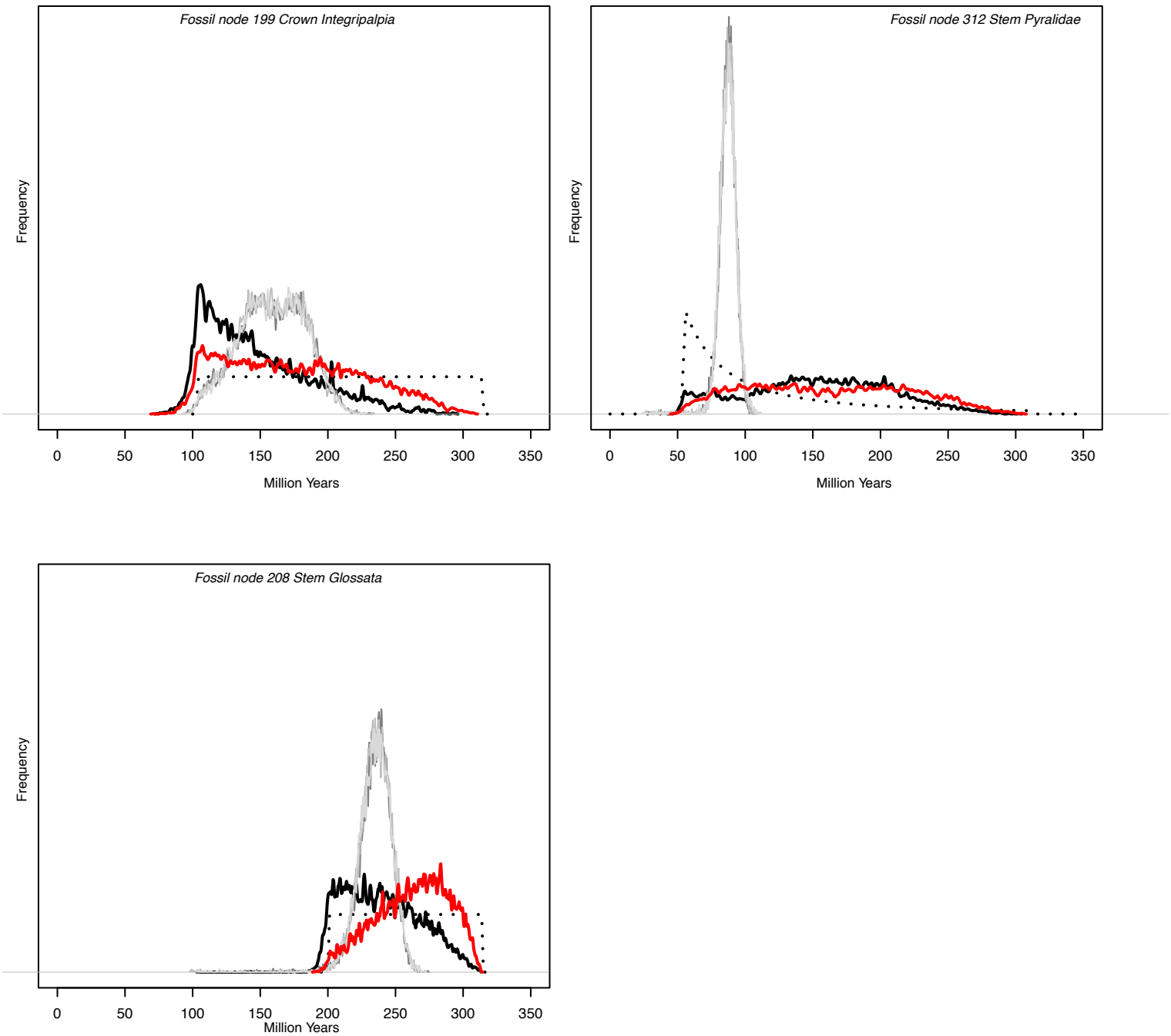

Figure S32. Density plots for the 3-fossil analysis with uncorrelated rates and Cauchy priors showing the relationship between sampling frequency and the age of a particular clade.

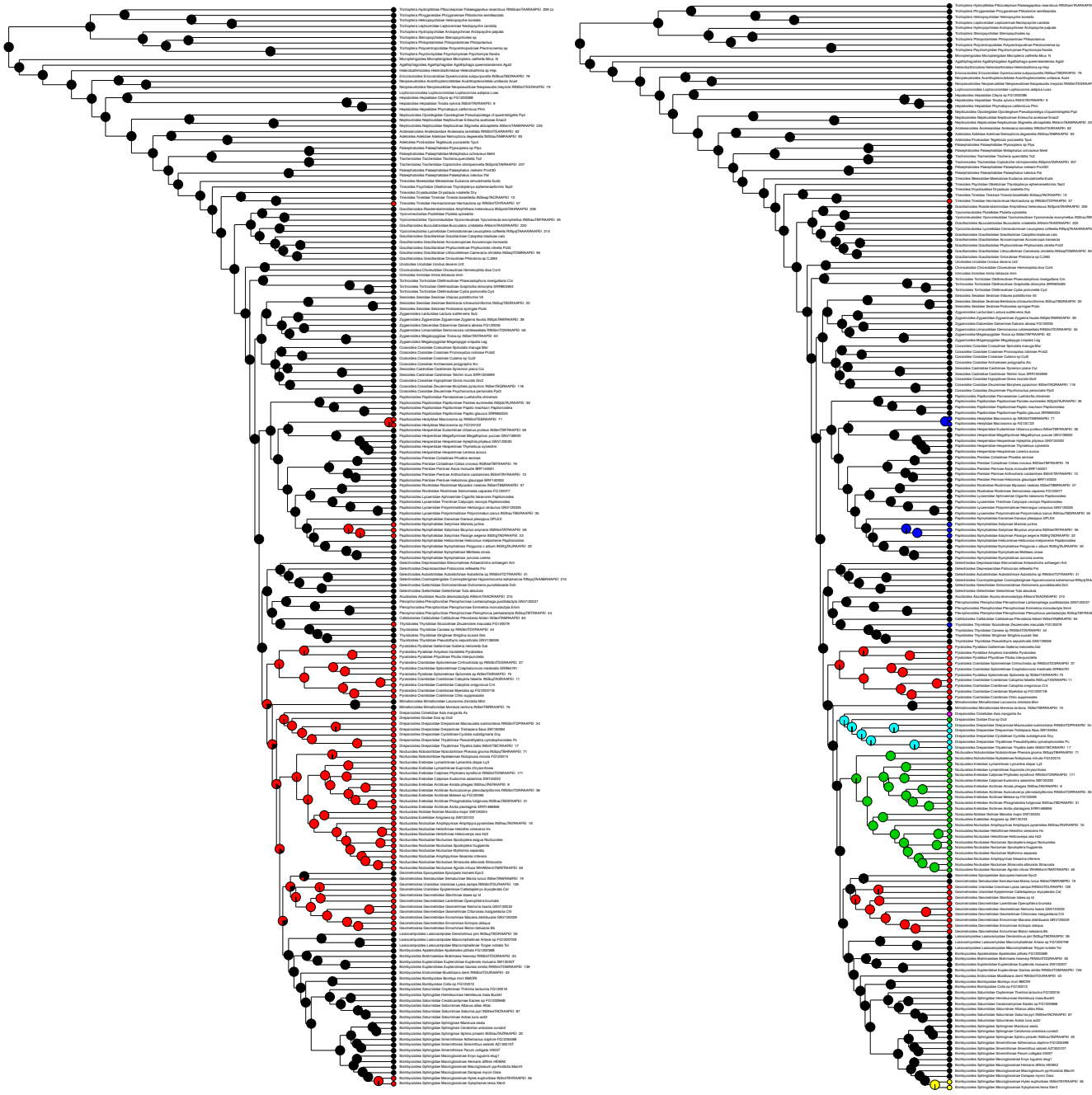

Figure S33. Ancestral state reconstructions of hearing organs on the amino acid ML tree topology. Left: Binary (2-state) ancestral state reconstruction. Black = hearing organ absent; Red = hearing organ present. Right: Multi-state (7-state) ancestral state reconstruction. Black = hearing organs absent; Red = tympana on the sternum of the second abdominal segment; Green = tympana on the metathorax; Dark blue = tympana beneath the forewing base; Cyan = tympana on the first abdominal segment; Magenta = hearing organs near the spiracles of the seventh abdominal segment; Yellow = hearing organs on palpi.

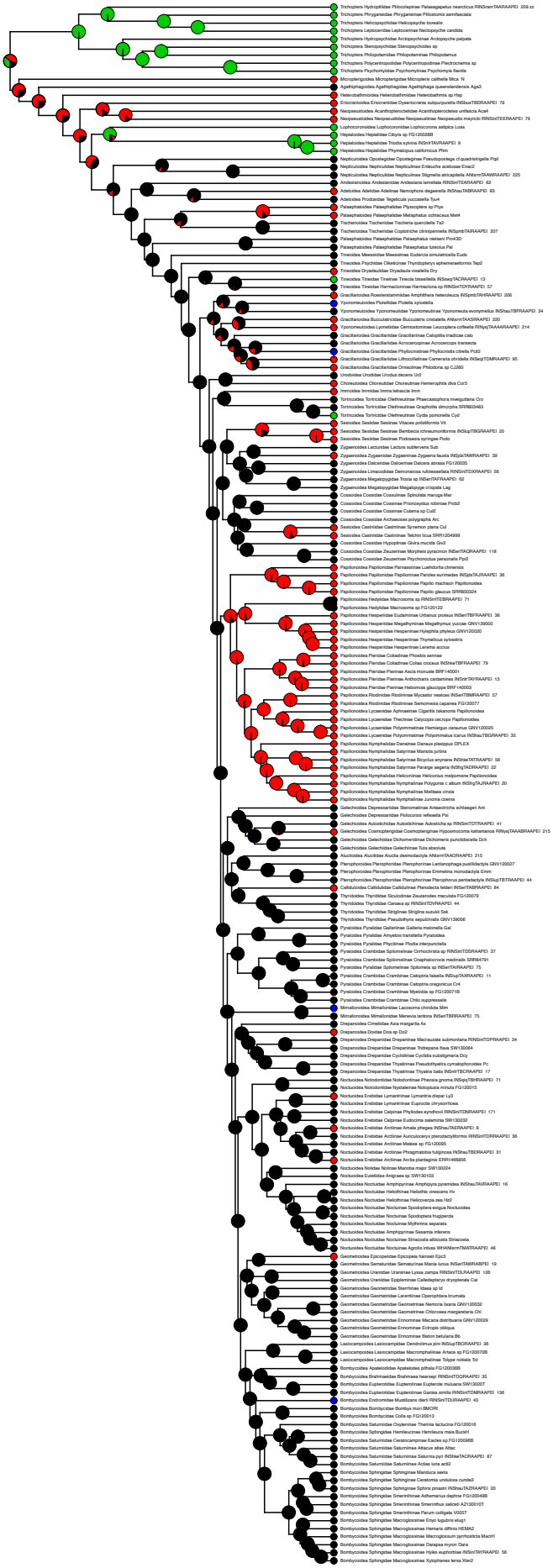

Figure S34. Ancestral state reconstruction of adult diel activity on the amino acid ML tree topology. Black = nocturnal; Red = diurnal; Green = crepuscular; Blue = active at all times.