**Electronic supplementary information**

**Title**

Reanalysis of the Apoid Wasp Phylogeny With Additional Taxa and Sequence Data Confirms the Placement of Ammoplanidae as Sister to Bees

**Authors**

Manuela Sann1\*, Karen Meusemann1,2, Oliver Niehuis1,Hermes E. Escalona2, Mikhail Mokrousov3‡, Michael Ohl4‡, Thomas Pauli1‡, Christian Schmid-Egger5‡

**Affiliations**

1 Albert Ludwig University of Freiburg, Institute of Biology I, Evolutionary Biology and Animal Ecology, Hauptstr. 1, 79104 Freiburg, Germany

2 Australian National Insect Collection, National Research Collections Australia, Commonwealth Scientific and Industrial Research Organisation (CSIRO), ACT 2601, Clunies Ross Street, Canberra, Australia

3 Institute of Biology and Biomedicine of the Lobachevsky State University of Nizhny Novgorod, Gagarin Av., 23/1, Nizhny Novgorod 603950 Russia

4 Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstr. 43, 10115, Berlin, Germany

5 Fischerstr. 1, 10317 Berlin, Germany

‡ Authors are listed in alphabetical order

\*Corresponding author:

E-mail address: [manuela.sann@biologie.uni-freiburg.de](mailto:manuela.sann@biologie.uni-freiburg.de) (MS)

**SUPPLEMENTARY INVENTORY**

1. **Supplementary results**
2. **Supplementary Tables**

**Table S1**: Detailed list of species studied with the DNA target enrichment approach.

**Table S2**: Statistics of Orthograph results for all species of the data analysis.

**Table S3**: List of species whose transcriptomic data (1KITE) was embedded in the enriched dataset.

**Table S4**: List of apoid wasps whose genomes we skimmed.

**Table S5**: Assembly statistics and number of identified single-copy genes in the analysed genomes.

**Table S6**: Results of the UniqueTree analysis when examining the 50 ML trees inferred from dataset sm-aa.

**Table S7**: Results of the UniqueTree analysis when examining the 50 ML trees inferred from dataset sm-nt12.

**Table S8**: Quartet sampling (QS) scores (QC/QD/QI) from analysing the best ML tree inferred from analysing the amino acid dataset (sm-aa).

**Table S9**: Quartet sampling (QS) scores (QC/QD/QI) from analysing the alternative ML tree inferred from the amino acid dataset (sm-aa).

**Table S10**: Quartet sampling (QS) scores (QC/QD/QI) from analysing the best ML tree inferred from the nucleotide dataset with 1st and 2nd codon positions included (sm-nt12).

**Table S11**: Results on alternative quartet topologies estimated for the main topology inferred under the MSC approach.

**Table S12**: Results on alternative quartet topologies estimated for the main topology inferred under the MSC approach.

**Table S13**: Statistical support of the 20 alternative topologies inferred from dataset sm-aa.

**Table S14**: Statistical support of the twelve alternative topologies inferred from dataset sm-nt12

**Table S15**: Statistical support of the 19 alternative topologies inferrd from dataset sm-nt123.

1. **Supplemental Figures**

**Figure S1**: Heat map indicating information content inferred with MARE when analysing the amino acid dataset (sm-aa).

**Figure S2**: AliStat heat maps of pairwise sequence comparison when analysing the concatenated datasets (a) sm-aa, (b) sm-nt12, and (c) sm-nt123.

**Figure S3**: Heat maps indicating among-lineage compositional heterogeneity when analysing the concatenated datasets (a) sm-aa, (b) sm-nt12, and (c) sm-nt123.

**Figure S4**: Best ML phylogenetic tree inferred from re-analysing the concatenated amino acid dataset (sm-aa).

**Figure S5**: Best ML phylogenetic tree inferred from re-analysing the concatenated nucleotide dataset with 1st and 2nd codon positions included (sm-nt12).

**Figure S6**: Best ML phylogenetic tree inferred from re-analysing the concatenated nucleotide dataset with all codon positions included (sm-nt123).

**Figure S7**: Maximum likelihood (ML) phylogenetic trees inferred from analysing dataset sm-nt123 with all codon positions included.

1. **Supplementary results**
2. **Sequencing and data processing**

We shallowly sequenced the genome of four apoid wasp species comprising *Eremiasphecium* sp., *E. concinnus*, *Heterogyna* (“brown wings"), and *Heterogyna* (“pale abdomen"). Genome skimming resulted into 34–40 million raw reads per species (median: 36 M) which assembled after adaptor trimming into 0.1–5 M contigs (median: 2,4 M). Particularly, raw reads assembled as follows: *E. concinnus* (4,802,285 with N50 of 576 bp), *Eremiasphecium* sp. (102,65 with N50 of 3125 bp), *Heterogyna* (“brown wings", 570,87 with N50 of 632 bp), and *Heterogyna* (“pale abdomen", 5,170,999 with N50 of 773 bp). For detailed information please refer to Tables S1 and S2.

We successfully recovered more than 95% of our 194 target single-copy protein-coding genes for each genome skimmed species (Table S2 and S5). Briefly, we obtained 189 genes for *E. concinnus*, 191 genes for *Eremiasphecium* sp., 186 genes for *Heterogyna* (“brown wings"), and 188 genes for *Heterogyna* (“pale abdomen").

1. **Multiple sequence alignment, alignment masking and supermatrix generation**

A total of 209 outlier sequences scattered over 81 of the 195 genes were removed from the amino acid MSAs and corresponding nucleotide sequence files. We identified on average 27.4% of the sites in the 194 amino acid MSAs as ambiguously (min: 0.3%; mdn: 25.2%; max: 77.1%) and removed these sites from the alignments at both, the amino acid and nucleotide level. Subsequently, the masked and concatenated supermatrices span an alignment length of 96,110 amino acid sites and respectively 288,330 nucleotide sites (including all codon positions) encoding a total of 194 single-copy protein-coding genes. The information content (IC) of the concatenated amino acid dataset calculated with MARE is 0.65 (Fig. S1). The overall completeness alignment score inferred with AliStat (Ca) is 52.3% (Fig. S2). All three concatenated datasets exhibit varying degrees of among-lineage compositional heterogeneity (Fig. S3). Due to a notable higher among-lineage heterogeneity identified in the dataset sm-nt123, we further refrained to infer separate gene trees including all three codon positions for the MSC approach.

1. **Phylogenetic tree inferences**
   1. **Concatenated approach**

RAxML *a posteriori* bootstrap convergence assessment indicated that bootstrap convergence was achieved after 200 bootstrap replicates for dataset sm-aa and dataset sm-nt12, and after 100 bootstrap replicates for dataset sm-nt123 (Fig. S4–6).

We checked all three concatenated datasets for rogue taxa with RogueNaRok version 1.0. We identified five species that show a rogue taxon behaviour in dataset sm-aa, four in dataset sm-nt12, and two in dataset sm-nt123. Particularly, when analysing the dataset sm-aa the following five species of apoid wasps exhibited rogue taxon behaviour: *Alysson* sp., *Argogorytes fargeii*, *Entomosericus concinnus*, *Lindenius panzeri* and *Tanyoprymnus moneduloides*. When analysing the dataset sm-nt12 the following four species of apoid wasps exhibited rogue taxon behaviour: *Alysson* sp., *Argogorytes fargeii*, *Lindenius panzeri* and *Tanyoprymnus moneduloides*. When analysing the dataset sm-nt123 the following two species of apoid wasps exhibited rogue taxon behaviour: *Alysson* sp. and *Argogorytes fargeii.* Except for *Entomosericus concinnus* (represented in the supermatrices by the amino acid and nucleotide sequences of 189 different genes), these species are characterised in our analyses by a small number of identified single-copy genes, i.e. one (*Alysson* sp.), two (*Argogorytes fargeii*), 44 (*Lindenius panzeri*), and 13 (*Tanyoprymnus moneduloides*).

* + 1. **Testing alternative topologies**

The inferred topologies were not unambiguously supported by the data. Except when having had analysed dataset sm-nt12, the ML tree with the best log-likelihood score does not reflect the most frequently inferred tree topology (Fig. 1b; Table S7). Briefly, phylogenetic analyses of dataset sm-aa resulted in 20 different tree topologies, in which the main topology (species clustered according to their potential phylogenetic group) of the best ML tree was represented only six times (Table S6). After merging species according to their potential taxonomic grouping, the only difference among the obtained topologies was found to be in the placement of *Eremiasphecium* and *E. concinnus* (Fig. 1a). The phylogenetic analyses of dataset sm-nt12 revealed twelve different topologies with identical phylogenetic relationships with respect to clustered clades with the best ML tree occurring most frequently (Table S7). Further, ML tree searches of dataset sm-nt123 resulted in 19 different tree topologies with the best ML tree inferred only once (Table S12).

The AU tests revealed that (a) from the 20 different tree topologies found in dataset sm-aa, twelve are rejected (Table S13), that (b) from the twelve different tree topologies found in dataset sm-nt12, ten are rejected (Table S14), and that (c) from the 19 different tree topologies found in dataset sm-nt123, four are rejected (Table S15).

* 1. **Coalescence-based approach**

Due to a notable higher proportion rejecting SRH conditions identified in dataset sm-nt123, we refrained to analyse this dataset with the MSC approach.

The resulting MSC trees derived from the complete amino acid dataset (DS0-aa, 194 genes) revealed no differences among the obtained topologies irrespective of whether lowly supported branches were collapsed or not (Fig. 2a, DS0-aa vs. DS0-aa-c10). The same was found when analysing the reduced amino acid dataset comprising 124 gene trees (Fig. 2b, DS1-aa vs. DS1-aa-c10). When comparing the complete amino acid dataset DS0-aa and the reduced amino acid dataset DS1-aa, we found the following four taxonomic groups with different placement: (1) Heterogynaidae, (2) Psenini, (3) Odontosphecini, and (4) *E. concinnus* (Fig. 2a and b). On nucleotide level, we found a slightly different topology when comparing the complete datasets comprising 194 genes (DS0-nt12) and the collapsed dataset (DS0-nt12-c10) with respect to the position of Astatidae (Fig. 3a). No topological difference was found between the reduced and reduced collapsed nucleotide datasets comprising 167 genes (Fig. 3b, DS1-nt12 vs. DS1-nt12-c10). Differences were found when comparing the complete and reduced nucleotide datasets (DS0-nt12 vs. DS1-nt12), for the position of *Eremiasphecium* (Fig. 3).

We found a strong support for *E. concinnus* as sister to a group comprising Ammoplanidae, Anthophila, *Eremiasphecium*, Pemphredonini *partim*, Philanthidae, Psenini, and Odontosphecini in three of the four ASTRAL-inferred topologies (Fig. 2a and 3). The moderately supported placement of Astatidae, Bembicidae, and Heterogynaidae in the concatenated topologies received stronger support trees inferred with the MSC approach, indicating a sister group relation of (1) Astatidae to the group comprising Ammoplanidae, Anthophila, Bembicidae, *E. concinnus, Eremiasphecium*, Pemphredonini *partim.*, Philanthidae, Psenini, and Odontosphecini, (2) Bembicidae to the group comprising Ammoplanidae, Anthophila, *E. concinnus, Eremiasphecium*, Pemphredonini *partim*, Philanthidae, Psenini, and Odontosphecini, and (3) Heterogynaidae to the group comprising Ammoplanidae, Anthophila, *E. concinnus, Eremiasphecium*, Pemphredonini *partim*, Philanthidae, Psenini, and Odontosphecini (Figure 2 and 3).