


Reanalysis of the apoid wasp phylogeny with additional taxa and sequence data confirms the placement of Ammoplanidae as sister to bees

MANUELA SANN¹, KAREN MEUSEMANN^{1,2}, OLIVER NIEHUIS¹, HERMES E. ESCALONA^{2,†}, MIKHAIL MOKROUSOV^{3,†}, MICHAEL OHL^{4,†}, THOMAS PAULI^{1,†} and CHRISTIAN SCHMID-EGGER^{5,†}

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

²Australian National Insect Collection, National Research Collections Australia, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia, ³Institute of Biology and Biomedicine, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, ⁴Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany and ⁵Fischerstr. 1, Berlin, Germany

Abstract. Apoid wasps and bees (Apoidea) are an ecologically and morphologically diverse group of aculeate Hymenoptera (ants, bees and wasps). During the last decades, significant progress has been made in illuminating the phylogenetic relationships of the major Apoidea lineages. However, some uncertainties have remained. In this study, we present results from re-investigating the phylogeny of Apoidea by including genome skimming data of key taxa that were missing in previous investigations: a representative of Entomosericini (tribe of the former Pemphredoninae) and a representative of Eremiaspheciinae (subfamily of the former ‘Crabronidae’). We additionally skimmed the genomes of two *Heterogyna* species (Heterogynidae). Our results from applying concatenation and coalescence-based phylogenetic approaches confirm the previously suggested sister group relationship of Ammoplanidae and bees. They also corroborate most taxonomic changes published in 2018 granting eight lineages of the former family ‘Crabronidae’ family status. However, some of our analyses indicate that the families Pemphredonidae and Psenidae could be para- or polyphyletic. After carefully assessing topological discordance and data quality, the exact placements of *Heterogyna* and of the genera *Eremiasphecium* and *Entomosericus* in the apoid wasp phylogeny remain ambiguous. However, our analyses indicate that inclusion of *Entomosericus* and *Eremiasphecium* in any of the currently accepted apoid wasp families cannot be well justified, and we consequently suggest raising Entomosericinae and Eremiaspheciini to family rank, respectively, to acknowledge this situation in the apoid classification: Entomosericidae Dalla Torre, 1897 (stat. n.) and Eremiaspheciidae Menke, 1967 (stat. n.).

Correspondence: Manuela Sann, Institute of Biology I, Evolutionary Biology and Animal Ecology, Albert Ludwig University of Freiburg, Hauptstr. 1, 79104 Freiburg, Germany. E-mail: manuela.sann@biologie.uni-freiburg.de

†Authors are listed in alphabetical order.

Introduction

The superfamily Apoidea (apoid wasps and bees) represents a highly diverse group of stinging Hymenoptera with currently approximately 30 000 described species (Michener, 2000; Pulawski, 2019). Recently, significant progress has been made in understanding the phylogenetic relationships of the major lineages of apoid wasps and bees by utilising Next Generation Sequencing (NGS) technologies for compilation of phylogenomic datasets (Branstetter *et al.*, 2017; Peters *et al.*, 2017; Sann *et al.*, 2018). These efforts fostered the identification of the closest extant relatives of bees, exposing the polyphyletic nature of the former apoid wasp family 'Crabronidae' (*sensu lato*) (Branstetter *et al.*, 2017; Peters *et al.*, 2017; Sann *et al.*, 2018). According to Sann *et al.* (2018), Apoidea comprises 11 major clades: Anthophila (bees), Ammoplanidae, Astatidae, Bembicidae, Crabronidae, Heterogynidae, Mellinidae, Pemphredonidae, Philanthidae, Psenidae and Sphecidae. Of these, Ammoplanidae likely represent the closest extant relatives of bees (Sann *et al.*, 2018). Both diverged from each other during the Late Cretaceous, ca. 128 million years ago (Mya; Sann *et al.*, 2018). However, some apoid wasp phylogenetic relationships have remained subject of discussion (e.g. the relationships of Astatidae, Bembicidae, and Mellinidae and the phylogenetic position of the species-poor Heterogynidae; Branstetter *et al.*, 2017; Peters *et al.*, 2017; Sann *et al.*, 2018). Furthermore, previous studies lacked two phylogenetically critical lineages with unclear phylogenetic position: Entomosericini and Eremiaspheciinae.

Entomosericus Dahlbom, *Eremiasphecium* Kohl and *Heterogyna* Nagy are enigmatic, species-poor apoid wasp genera, whose placements have been controversially discussed. The biology of all three genera is hardly known or even unknown (*Heterogyna*). *Entomosericus* occurs from the eastern Mediterranean region to Central Asia. It has been placed in a monotypic subfamily by Bohart & Menke (1976), but was later considered as a tribe of the former Pemphredoninae within Crabronidae (Melo, 1999). *Eremiasphecium* is distributed from Northern Africa to Central Asia. It was previously considered as a subordinated taxon of the subfamily Philanthinae within Crabronidae. *Heterogyna* is an enigmatic genus with brachypterous females. The seven currently recognized species have been collected in the eastern Mediterranean region, Central Asia, eastern Africa and Madagascar (Ohl & Bleidorn, 2006). The placement of *Heterogyna* within Apoidea has been discussed controversially since the discovery of the genus. Even a phylogenetic placement outside Apoidea has been discussed. Recent phylogenetic analyses provided clear evidence that *Heterogyna* belongs to Apoidea, but the exact phylogenetic position of the genus is still unclear. We therefore decided to expand our taxonomic sampling of apoid wasps in a re-investigation of apoid wasp phylogenetic relationships by including a representative of Entomosericini, *Entomosericus concinnus* Dahlbom, a representative of the Eremiaspheciinae, *Eremiasphecium* sp., and two morphospecies of the genus *Heterogyna* (Heterogynidae).

After years of phylogenomic research across the tree of life, it is evident that evaluating the reliability of an inferred phylogenetic tree derived from large datasets is a complex endeavour (Kapli *et al.*, 2020). Despite enormous progress in generating large phylogenetic datasets due to advances in DNA sequencing technologies (Young & Gillung, 2019), incongruences between phylogenetic studies and uncertainties in the interpretation of phylogenetic results have remained ubiquitous (Rokas *et al.*, 2003; Degnan & Rosenberg, 2009; Smith *et al.*, 2015; Evangelista *et al.*, 2018; Betancur-R *et al.*, 2019). Improper modelling of biological phenomena by neglecting, for instance, orthology, compositional heterogeneity among sites and/or lineages, heterotachy, incomplete lineage sorting (ILS), rate heterogeneity and horizontal gene transfer (Romiguier *et al.*, 2016; Young & Gillung, 2019) might lead to unperceived error bias. When dealing with multi-gene datasets, it becomes more and more common to apply a multi-species coalescent-based (MSC) approach and/or a concatenation approach which both have their pros and cons (Springer & Gatesy, 2016; Zhang *et al.*, 2018; Young & Gillung, 2019; Kapli *et al.*, 2020; Simion *et al.*, 2020). Traditional measures indicating the reliability of phylogenetic inferences have been nonparametric bootstrapping (Efron, 1979; Felsenstein, 1985) and the Bayesian posterior probability (PP) estimation (Huelsenbeck & Ronquist, 2001). However, both approaches are sensitive to model misspecification (Young & Gillung, 2019). Fortunately, additional measures for evaluating competing phylogenetic hypotheses have been developed and are becoming more common. These include quartet puzzling (e.g. Four-cluster Likelihood Mapping, FcLM; Strimmer & von Haeseler, 1997) in combination with data permutation strategies (Misof *et al.*, 2014; Sann *et al.*, 2018) and quartet sampling (QS; Pease *et al.*, 2018).

To assess previous reported incongruencies in different phylogenetic inferences of apoid relationships, we re-analyzed a dataset consisting of transcriptomic and DNA target enrichment data published by Sann *et al.* (2018), extended with DNA sequence data of selected key taxa previously unavailable to us. We focus our study on previously ambiguously inferred phylogenetic relationships with potential implications for the systematics of Apoidea and for our understanding of the evolutionary history of this group. Specifically, to assess whether or not Ammoplanidae are indeed the closest extant relatives of bees, we collected nucleotide sequence data of two apoid wasp lineages missing in previous phylogenomic studies: Entomosericini (here represented by the species *Entomosericus concinnus*) and Eremiaspheciinae (here represented by an unidentified species of the genus *Eremiasphecium*). We further included nucleotide sequence data of two species of Heterogynidae, since the phylogenetic position of this family has remained elusive. All new nucleotide sequence data were obtained by applying a genome skimming approach. We analyzed the extended dataset in detail, using both a concatenation and an MSC tree inference approach, and we applied multiple tools to assess the presence of misleading and conflicting signals for alternative phylogenetic hypotheses and the impact of rogue on our phylogenetic inferences.

Material and methods

Taxon sampling

We studied 135 apoid wasp species, comprising representatives of all 14 subfamilies listed in the ‘Catalog of Sphecidae’ by W. J. Pulawski (<http://www.calacademy.org/scientists/projects/catalog-of-sphecidae>), and 42 bee species, comprising representatives of all seven described extant bee families (Michener, 2000). Furthermore, we included a total of nine outgroup species, comprising Formicidae (3), Mutillidae (1), Pompilidae (1), Sapygidae (1), Scoliidae (2) and Tiphidae (1).

The bulk of sequence data was mined from previously published studies. Specifically, we exploited the amino acid and nucleotide sequences identified with Orthograph version 0.5.6 (<https://github.com/mptrsen/Orthograph/>; Petersen *et al.*, 2017) as described by Sann *et al.* (2018). Thus, the bulk of identified sequence data comprises DNA target enrichment data referring to 92 species of apoid wasps and two species of bees published by Sann *et al.* (2018); Tables S1, S2) and RNAseq data referring to 39 species of apoid wasps and to 40 species of bees published by Peters *et al.* (2017); Tables S2, S3). We extended these datasets by applying genome skimming on samples of four apoid wasp species that were lacking or represented by only a small number of genes in the study by Sann *et al.* (2018); see below and Table S4).

Genome skimming

We shallowly sequenced the genomes of *Entomosericus concinnus*, *Eremiasphhecium* and of two species of *Heterogyna* (morphospecies ‘brown wings’, and morphospecies ‘pale abdomen’) (Table S4). Genomic DNA (gDNA) was extracted from whole specimens (one per species) using the QIAGEN DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) by following the manufacturers’ protocol and eluted the gDNA in 20 µL nuclease-free water. The quantity of extracted gDNA was assessed with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc. Waltham, MA, U.S.A.). Dual-indexed DNA libraries were paired-end sequenced on an Illumina NextSeq platform with a read length of 150 base pairs (bp) by StarSEQ (Mainz, Germany). We expected a 10× coverage from sequencing 35 million reads of a genome with a hypothesized size of 500 Mb.

De novo assembly of genome skimming generated data

The raw sequencing reads of each of taxon were assessed for data quality and adaptor presence with FastQC version 0.11.8 (<https://github.com/s-andrews/FastQC>). Two genome assemblers were used: SparseAssembler (Ye *et al.*, 2012) and Platanus version 1.2.4 (Kajitani *et al.*, 2014). Both programs were used with their default settings. The assemblies with better statistics, for instance a N50 value, were selected for consideration in subsequent analyses (Table S5). In all cases but one, we selected

the assemblies generated with the SparseAssembler; in case of *Eremiasphhecium*, we selected the Platanus assembly. The comparatively high coverage of the genomes of *Eremiasphhecium* sp. and *Heterogyna* (‘brown wings’) allowed further processing the assembled contigs (i.e. filtering, heterozygosity reduction, scaffolding and gap filling) with the pipeline Redundans version 0.13c (Pryszcz & Gabaldón, 2016) using the default software settings. Genome assembly statistics were obtained with the software Quast version 4.3 (Gurevich *et al.*, 2013; Table S5). All raw sequences have been submitted to the NCBI Sequence Read Archive and can be found in Table S4.

Orthology prediction and multiple sequence alignment

Orthologous nucleotide sequences were identified as described by Sann *et al.* (2018) using the software package Orthograph version 0.6.3 (Petersen *et al.*, 2017). Briefly, Orthograph searched for putative single-copy genes in the assembled genome skimming data using a custom ortholog set comprising 3260 single-copy genes identified in six representative Hymenoptera genomes (Peters *et al.*, 2017). However, we considered only those 195 genes that were also analyzed by Sann *et al.* (2018). Finally, we merged the target DNA sequences identified in the four genomes with those from the dataset published by Sann *et al.* (2018).

The amino acid sequences of the respective single-copy genes provided by Orthograph were aligned with the program MAFFT version 7.310 (Katoh & Standley, 2013), applying the L-INS-i algorithm. We identified and refined potentially misaligned sequences (outliers) in the MSAs at the amino acid level. After a second outlier check, we removed final outlier sequences from the amino acid MSAs and from the corresponding nucleotide sequence files as described by Misof *et al.* (2014). Subsequently, we removed the sequences of the reference species *Nasonia vitripennis* (Walker) and also removed all resulting gap-only sites from the amino acid MSAs. Nucleotide sequences were aligned using the corresponding amino acid MSAs as blue prints by applying a modified version of PAL2NAL version 14.1 (Suyama *et al.*, 2006; Misof *et al.*, 2014).

Alignment masking and supermatrix generation

To remove ambiguously aligned sections in the amino acid and in the nucleotide MSAs, we ran a modified version of the program Aliscore version 2.0 (Misof & Misof, 2009; Kück *et al.*, 2010) with the option -e, using a sliding window size, and allowing the maximal number of pairwise sequence comparisons. Sections and positions identified as ambiguously aligned were removed from the amino acid and from the nucleotide MSAs as described by Sann *et al.* (2018). Masked MSAs are available as Supplementary data at Dryad Repository.

We concatenated the masked MSA files to supermatrices and generated respective partition files using the gene boundaries as described by Sann *et al.* (2018) using the program FASconCAT-G version 1.02 (Kück & Longo, 2014). In total,

we generated three different concatenated supermatrices, each comprising 195 target loci: (1) on the amino acid level (sm-aa), (2) on the nucleotide level with first and second codon positions (sm-nt12), and (3) on the nucleotide level with all codon positions (sm-nt123).

The information content of the dataset on the amino acid level was calculated and visualized using the software MARE version 0.1.2-rc (Misof *et al.*, 2013). Please note that MARE found no gene partition with an information content of zero (IC), yet, one gene was discarded due to a reported bug in the MARE script. The total number of genes was thus 194 in all three supermatrices. Additionally, we examined the distribution of data completeness across the three superalignments using the software AliStat version 1.11 (Wong *et al.*, 2020). We generated heatmaps visualising completeness scores of sequence pairs in all the supermatrices. All MARE and AliStat metrics are available in the Electronic supplementary information and Figs S1 and S2.

Exploring stationary, reversible and homogeneous conditions

We tested whether nucleotide and amino acid sequences included in the three datasets had evolved under globally stationary, reversible and homogeneous (SRH) conditions using SymTest version 2.0.49 (<https://github.com/ottmi/symtest>) (Jermiin *et al.*, 2004; Ababneh *et al.*, 2006; Jermiin & Ott, 2017). SymTest uses matched-pairs tests of symmetry (Misof *et al.*, 2014). We applied the Bowker's test (Bowker, 1948) on the supermatrices sm-aa, sm-nt12 and sm-nt123, and we generated heatmaps based on the obtained *p*-values in order to determine which sequence pairs matched SRH conditions.

Phylogenetic tree inferences

Concatenated approach

Phylogenetic relationships were inferred using the maximum likelihood (ML) optimality criterion implemented in the software IQ-TREE version 1.6.12 (Nguyen *et al.*, 2014; Chernomor *et al.*, 2016) on the inferred supermatrices (i.e. sm-aa, sm-nt12, and sm-nt123). We chose the best-fitting substitution model for the amino acid sequences of each gene partition with ModelFinder (Kalyaanamoorthy *et al.*, 2017) implemented in IQ-TREE. Specifically, we tested available nuclear models plus the free rate models LG4X and LG4M (Le *et al.*, 2012). We opted for the edge-proportional partition model (–spp, Chernomor *et al.*, 2016), allowing partitions to have different evolutionary speeds and we choose the AICc (Hurvich & Tsai, 1989) criterion to select the best model for each partition. Further settings used in ModelFinder were: E, I, G, R for parameter optimisation excluding I + G as suggested by Yang (2004) and calculating the median for each GAMMA category (options: –spp –mrate E,I,G,R –msub nuclear –madd LG4X, LG4M –merit AICc –gmedian). We kept all other options at defaults. On the nucleotide level, we estimated the best substitution models out of all implemented nucleotide models, except for codon

models, applying the same options as described above for the two nucleotide supermatrices (i.e. sm-nt12 and sm-nt123).

For each of the three supermatrices, we conducted 50 independent ML tree searches (25 with random start trees and 25 with parsimony start trees) in IQ-TREE version 1.6.12. We used the software parameters –mrate E, I, G, R and –gmedian. We chose the best scoring ML tree for each supermatrix according to the best log-likelihood value. Branch support was assessed by conducting nonparametric bootstrapping (IQ-TREE version 1.6.12) with random start trees and 400 (sm-aa), 420 (sm-nt12) and 100 (sm-nt123) bootstrap replicates. Bootstrap convergence (Pattengale *et al.*, 2010) was assessed *a posteriori* ten times with random seeds using RaxML version 8.2.11 (Stamatakis, 2014) and applying the following parameters: –I autoMRE, –B 0.03, –m GTRGAMMA. Bootstrap support was mapped for each supermatrix onto the best ML tree with IQ-TREE version 1.6.12.

We used Unique Tree version 1.9 (Wong & Jermiin, available upon request) to assess how many unique tree topologies were obtained examining the 50 ML trees inferred from a given supermatrix separately. We checked all three datasets for whether or not the inferred topologies included rogue taxa with RogueNaRok version 1.0 (Aberer *et al.*, 2013), providing the software all bootstrap trees and the best ML tree.

Testing alternative topologies

AU test. We statistically assessed differences among the three tree topologies inferred from the concatenated datasets (sm-aa, sm-nt12 and sm-nt123) using approximately unbiased (AU) tests (Shimodaira, 2002) with IQ-TREE version 1.6.12. Specifically, we ran IQ-TREE with the AU test option by specifying 10 000 re-sampled estimated log-likelihood (RELL) bootstraps (Kishino *et al.*, 1990).

Quartet sampling. We applied the Quartet Sampling (QS) method described by Pease *et al.* (2018) version 1.3.1 to examine the inferred phylogenetic trees for discordance and poor support and for hidden phylogenetic signal that might be not visible in the ML trees derived from the concatenation approach. The QS method rapidly and simultaneously assesses confidence, consistency, and informativeness of internal tree relationships, as well as the reliability of each terminal branch by calculating the following quartet scores: Quartet Concordance (QC), Quartet Differential (QD), Quartet Informativeness (QI) and Quartet Fidelity (QF) (Pease *et al.*, 2018). We conducted QS on the concatenated datasets, providing the best model for each gene partition (1) with the best and the alternative ML tree topology (represented by the best and third best log-likelihood score; Table S6) of dataset sm-aa, and (2) with the best ML tree of dataset sm-nt12 (Table S7). We specified a maximum number of replicates per branch with reps 200 and evaluated the likelihood of the analyzed topologies for each quartet sample using RAXML version 8.2.10 (Stamatakis, 2014).

Multi-species coalescence based approach

Given that a species tree inferred from a concatenated supermatrix does not necessarily show the same branching as corresponding gene trees inferred separately for every gene partition due to incomplete lineage sorting, we additionally applied an MSC-based approach to account for gene tree discordance and heterogeneous signal across genes on (1) the amino acid dataset that corresponds to dataset sm-aa, hereinafter referred to as DS0-aa, and on (2) the nucleotide dataset with first and second codon positions included, hereinafter referred to as DS0-nt12, and being correspondent to the dataset sm-nt12. Due to a notable higher among-lineage heterogeneity identified on the data set sm-nt123, we refrained from applying the MSC approach on the dataset that included all three codon positions (Fig. S3).

We inferred phylogenetic trees from each of the 194 gene partitions using IQ-TREE version 1.6.12. The best-fit substitution model was estimated with ModelFinder (Kalyaanamoorthy *et al.*, 2017), applying the same parameter settings as described above on the amino acid and on the nucleotide level. For each gene, we conducted ten independent ML tree searches using neighbour joining trees (–t BIONJ) as start trees, random seeds, and 1000 ultrafast bootstrap replicates with optimized nearest neighbour interchange (NNI) based on the bootstrap alignments (options –bb 1000 and –bnni; Hoang *et al.*, 2018). The best ML tree per gene partition with statistical UF bootstrap support served as input to infer a species tree with the MSC approach as implemented in the Accurate Species TRee ALgorithm (ASTRAL) version 5.7.3 (Zhang *et al.*, 2018).

The MSC-based approach implemented in ASTRAL can have reduced accuracy when poorly resolved gene trees are provided and the gene tree error is high (Molloy & Warnow, 2018). Thus, we followed the suggestion by Barrow *et al.* (2018) and conducted additional analyses that are based on reduced gene sets: we only considered gene trees with an average bootstrap support of $\geq 60\%$ in the MSC analyses. Bootstrap support values were extracted for the generated gene tree files with Newick Utilities version 1.6 (Junier & Zdobnov, 2010), and the average BS value was calculated with a custom-made bash script. By doing so, the total number of gene trees dropped from 194 to 124 when conducting the analyses on the amino acid level (DS1-aa) and to 167 when conducting the analyses on the nucleotide level and considering first and second codon positions only (DS1-nt12). Finally, we evaluated the impact of collapsing low support branches ($\leq 10\%$) in all input gene trees to explore whether this would improve phylogenetic accuracy by reducing noise (Mirarab & Warnow, 2015). We conducted MSC analyses with ASTRAL on four sets of genes on the amino acid and the nucleotide level considering only first and second codon positions: (1) all 194 genes (datasets DS0-aa and DS0-nt12), (2) all 194 genes, but having collapsed branched with low bootstrap support ($\leq 10\%$; datasets DS0-aa-c10 and DS0-nt12-c10), (3) 124 (amino acid level) or 167 (nucleotide level) genes with an average bootstrap support $\geq 60\%$ (datasets DS1-aa and DS1-nt12) and (4) 124 (amino acid level) or 167 (nucleotide level) genes with

an average bootstrap support $\geq 60\%$, but collapsing branches with low bootstrap support ($\leq 10\%$; datasets DS1-aa-c10 and DS1-nt12-c10).

MSC local posterior probabilities inferred from quartet frequencies and quartet scores

To account for gene tree discordance in our datasets, we calculated (1) the three local posterior probabilities (pp1, pp2 and pp3) for each branch, since they are suggested to be more reliable than traditional branch support values like, for example, multi-locus nonparametric BS support (Sayyari & Mirarab, 2016), and (2) the three possible quartet scores for each branch as provided by ASTRAL version 5.7.3 (Mirarab *et al.*, 2014). The quartet scores display the support for any of three possible phylogenetic quartet arrangement around the internal split in a dataset (Sayyari & Mirarab, 2016). Local posterior probability (pp) values and quartet scores were calculated for the four inferred MSC species trees (options –q and –t 2): the two obtained when analysing the data on the amino acid level and the two obtained when analysing the data on the nucleotide level considering first and second codon positions only (see Multi-species coalescence based approach section). Quartet scores for each split were visualized using the ETE3 toolkit version 3.1.1 (Huerta-Cepas *et al.*, 2016).

Results

Our phylogenetic analyses are based on 194 single-copy protein-coding genes covering representatives of all currently accepted apoid wasp families and subfamilies, all bee families, and nine outgroup species (Peters *et al.*, 2017; Sann *et al.*, 2018). All results from genome sequencing and data processing, orthology predication, MSA alignment masking, supermatrix generation, compositional heterogeneity and rogue taxon analyses are given in the Electronic supplementary information. In the following paragraphs, we focus on presenting the results obtained from analysing the data on the amino acid level and on the nucleotide level considering first and second codon positions only.

Concatenation approach

We inferred largely identical topologies when analysing the different datasets with a concatenation approach, and these topologies are also largely congruent with those reported by Sann *et al.* (2018) (Figs 1, S4–S6; Table 1).

Monophyly of Apoidea is strongly supported in all inferred ML trees and in agreement with the results reported by Branstetter *et al.* (2017), Peters *et al.* (2017), and Sann *et al.* (2018) (Figs 1, S4–S6; Table 1). Our analyses consistently inferred Ampulicidae as the sister group of all remaining Apoidea (Figs 1, S4–S6; Sann *et al.*, 2018). The monophyly of the major Apoidea lineages recognized as monophyletic by Sann *et al.* (2018) is also confirmed: Ammoplanidae,

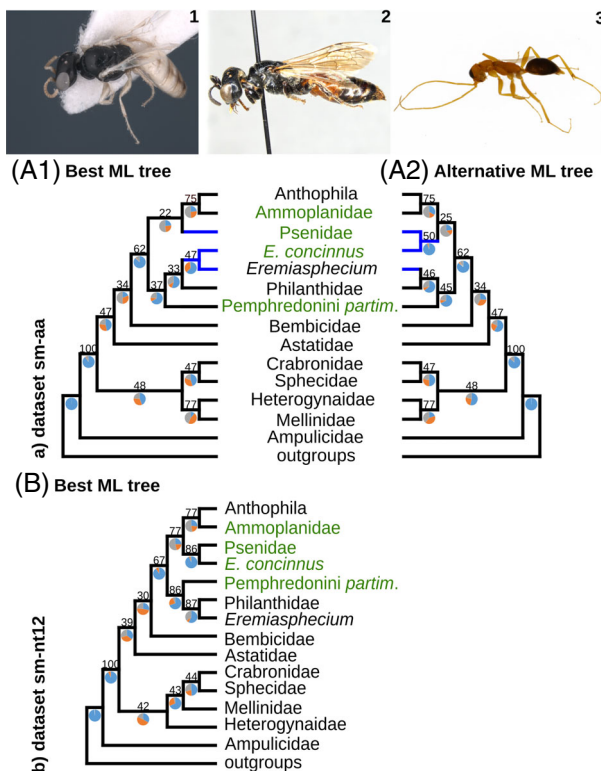


Fig. 1. Maximum likelihood (ML) phylogenetic trees inferred from datasets sm-aa (a) and sm-nt12 (b). From the 50 ML tree searches 16 similar topologies, including the best ML, tree were obtained when analysing dataset sm-aa (A1). The remaining 34 of the 50 ML trees represent the alternative tree topology (A2). Differences between the best (A1) and the alternative (A2) ML tree are indicated by blue lines. When analysing the data at the nucleotide level (dataset sm-nt12), the best ML tree is also the most frequent tree topology (B). Numbers along branches represent ML bootstrap values. Tripartitioned circles define the count of the number of QS replicates for the quartet arrangements (blue) being concordant with the ML tree. The number of QS replicates for the two discordant quartet arrangements are shown in orange and grey (Tables S8–S10). Taxa of the polyphyletic group Pempredoninae are coloured in green. The photographs show the following species: (1) *Eremiasphecium arabicum* Pulawski (female; photograph by C. Schmid-Egger), (2) *Entomosericus concinnus* (female; photograph by C. Schmid-Egger) and (3) *Heterogyna nocticola* (female; photograph by M. Ohl). [Colour figure can be viewed at wileyonlinelibrary.com].

Anthophila, Astatidae, Bembicidae, Crabronidae, Heterogynidae, Mellinidae, Pempredonidae, Philanthidae, Psenidae and Sphecidae (Figs 1, S6; Table 1). Consistent with the results reported by Sann *et al.* (2018), we find in all phylogenetic inferences strong signal for Ammoplanidae being the sister group of bees (Figs 1, S4–S6; Table 1). The phylogenetic placement of Astatidae, Bembicidae and Mellinidae remains uncertain due to their different placement in the trees inferred from analysing the datasets sm-aa, sm-nt12 and sm-nt123 (Figs 1, S4–S6).

The placement of Eremiasphecinae, here represented by *Eremiasphecium* sp. and *E. concinnus*, is ambiguous (Table 1). We find *Eremiasphecium* sp. either as sister group of Philanthidae (Fig. 1a; dataset sm-aa A2 and 1b dataset sm-nt12,

Fig. S7 dataset sm-nt123) or as sister group of *E. concinnus* (Fig. 1a dataset sm-aa A1). We further found *E. concinnus* either as sister group of Psenidae (Fig. 1a dataset sm-aa A2 and 1b dataset sm-nt12, Fig. S7 dataset sm-nt123) or as sister group of *Eremiasphecium* sp. (Fig. 1a dataset sm-aa A1).

In contrast to the results in our previous study (Sann *et al.*, 2018), we cannot confirm the genus *Heterogyna* to represent a subordinated lineage of the tribe Nyssonini (Table 1). Instead, we found a monophyletic *Heterogyna* to represent either the sister group of all remaining Apoidea except Ampulicidae (Fig. S7 dataset sm-nt123), or as sister group of Mellinidae (Fig. 1a dataset sm-aa), or as sister group of Mellinidae + (Crabronidae + Sphecidae) (Fig. 1b dataset sm-nt12).

Testing alternative topologies

AU test. The inferred topologies were not unambiguously supported by the data. Briefly, the best ML topology (logL: −2 882 589.491; Table S6) and the most frequently inferred topology (logL: −2 882 597.101; Table S6) of dataset sm-aa are both not rejected (Fig. 1; Table S6). Both topologies differ in the placement of *Eremiasphecium* sp. and *E. concinnus* (Fig. 1a). For the nucleotide dataset sm-nt12 the phylogenetic relationships with respect to clustered clades are identical (Fig. 1b; Table S7). More details are provided in the Electronic supplementary information Tables S13–S15.

Quartet sampling. Results from the QS analyses of the concatenated datasets sm-aa (best ML tree and alternative tree topology represented by the third best log likelihood score) and sm-nt12 are shown in Fig. 1. When clustering taxa into major clades (Table S8), the median Quartet Informativeness (QI) score indicates moderate phylogenetic information for all branches: for the best inferred ML tree of dataset sm-aa (Fig. 1a-A1; min: 26.5%; max: 99.5%; med: 51.5%; Table S8) as well as for the alternative topology of dataset sm-aa (Fig. 1a-A2; min: 25.5%; max: 95.0%; med: 44.0%; Table S9). The same holds for dataset sm-nt12 (min: 23.0%; max: 93.0%; med: 43.0%; Table S10). Among the tested datasets, major splits of interest are strongly supported with high quartet concordant (QC) scores ≥ 0.7 and low skew in discordant frequencies (quartet differential score) $QD \approx 1$, indicating that the majority of quartets support the input tree topology and not an alternative topology: (1) Ampulicidae as extant sister group to all remaining Apoidea (nearly full concordance quartet support; Fig. 1), (2) the node at which Crabronidae, Heterogynidae, Mellinidae and Sphecidae split from all remaining Apoidea (Fig. 1) and (3) the sister group relationship between *E. concinnus* and Psenidae (Figs. 1a-A2, 1b). We find the following splits to be well supported with $0.7 \geq QC > 0.3$ and with a low skew in discordant quartet frequencies $QD \geq 0.3$: (1) the clade comprising Mellinidae + (Crabronidae + Sphecidae) (Fig. 1b) and (2) the clade comprising ((Pempredonina, Spilomenina and Stigmata, hereafter referred to as Pempredonini partim) + (Philanthinae + *Eremiasphecium*)) + ((Anthophila + Ammoplanina) + (*E. concinnus* + [Psenini + Odontosphecini]))

Table 1. Comparison of Apoidea phylogenetic relationships recovered by an earlier and by this study.

Phylogenetic result	Sann <i>et al.</i> (2018) sm-nt123	This study: concatenation approach sm-aa/sm-nt12/sm-nt123	This study: MSC approach DS0-aa/DS1-aa/DS0-nt12/DS1-nt12
Monophyletic?			
Anthophila (bees)	Yes (100)	Yes (84/87/93)	Yes (100/95/100/100)
Ammoplanidae	Yes (100)	Yes (100/100/100)	Yes (100/100/100/100)
Ampulicidae	Yes (100)	Yes (100/100/100)	Yes (76/71/84/89)
Astatidae	Yes (100)	Yes (99/100/100)	Yes (100/100/100/100)
Bembicidae	Yes (80)	Yes (82/71/85)	Yes (100/100/100/100)
Crabronidae	Yes (100)	Yes (66/62/67)	Yes (100/100/100/100)
Heterogynaidae	No	Yes (100/100/100)	Yes (100/100/100/100)
Mellinidae	Yes	Yes	Yes
Sphecidae	Yes (100)	Yes (71/70/84)	Yes (100/100/100/100)
Pemphredonidae	Yes (100)	Yes (92/95/96)	Yes (94/84/NA/NA) and No
Philanthidae	Yes (100)	Yes (93/94/97)	Yes (99/100/100/100)
Psenidae	Yes (100)	Yes (92/95/96)	Yes (NA/NA/47/48) and No
Sister group to/phylogenetic position of:			
Anthophila (bees)	Sister group to Ammoplanidae (100)	Sister group to Ammoplanidae (75/77/90)	Sister group to Ammoplanidae (97/85/89/94)
Phylogenetic position of Heterogynaidae	Subordinated lineage of Nyssonini (86)	(a) Sister to Mellinidae (77/NA/NA) (b) Sister to Mellinidae + (Crabronidae + Sphecidae) (NA/42/NA) (c) Sister to all remaining Apoidea except Ampulicidae (NA/NA/99)	(d) Sister to Mellinidae + (Crabronidae + Sphecidae) (36/NA/NA/NA) (e) Sister to the group comprising Ammoplanidae, Anthophila, Astatidae, Bembicidae, <i>E. concinnus</i> , <i>Eremiasphecium</i> , Pemphredonini partim., Philanthidae, Psenidae and Odontosphecini (NA/67/NA/NA) (f) Sister to the group comprising Ammoplanidae, Anthophila, <i>E. concinnus</i> , <i>Eremiasphecium</i> , Pemphredonini partim., Philanthidae, Psenidae and Odontosphecini (NA/NA/73/77)
Phylogenetic position of <i>E. concinnus</i>	NA	(a) Sister to <i>Eremiasphecium</i> (47/NA/NA) (b) Sister to Psenidae (NA/86/100)	(a) Sister to the group comprising Ammoplanidae, Anthophila, <i>Eremiasphecium</i> , Pemphredonini partim., Philanthidae, Psenini and Odontosphecini (100/NA/100/100) (b) Sister to Psenini (NA/38/NA/NA)
Phylogenetic position of <i>Eremiasphecium</i>	NA	(a) Sister to <i>E. concinnus</i> (47/NA/NA) (b) Sister to Philanthidae (NA/87/97)	(a) Sister to Philanthidae (73/39/NA/NA) (b) Sister to <i>Spilomena beata</i> (NA/NA/29/NA) (c) Sister to Pemphredonini partim. (NA/NA/NA/36)

Given are results from (1) the concatenation approach published by Sann *et al.* (2018) and (2) the re-analyzed datasets of this study applying a concatenation approach, and (3) the re-analyzed datasets of this study applying an MSC approach. Numbers within parentheses represent branch support on the respective split, with bold numbers indicating the highest support value. NA: not applicable/not inferred.

(Figs. 1a–A2, b). A sister group relationship between Ammoplanidae and bees is equally represented by a low QC and high QD score in all three analyses.

Coalescence-based approach

MSC phylogenetic trees, irrespective of whether they were inferred from data on the amino acid or on the nucleotide level, strongly support a monophyly of Apoidea and confirm Ampulicidae as the sister group of all remaining Apoidea (Figs 2, 3). The

MSC approach confirms all previously described major lineages of Apoidea (Sann *et al.*, 2018): Ammoplanidae, Anthophila, Astatidae, Bembicidae, Crabronidae, Heterogynaidae, Mellinidae, Philanthidae and Sphecidae (Figs 2, 3; Table 1).

In contrast to the phylogenetic trees inferred with the concatenation approach, the placement and the relationship of Pemphredonidae and Psenidae is ambiguous (Figs 2, 3; Table 1). Pemphredonidae, comprising Pemphredonina, Spilomenina and Stigmina (Sann *et al.*, 2018), is also recovered when analysing the data on the amino acid level (Pemphredonini *partim*; Fig. 2). However, Pemphredonina is polyphyletic when the

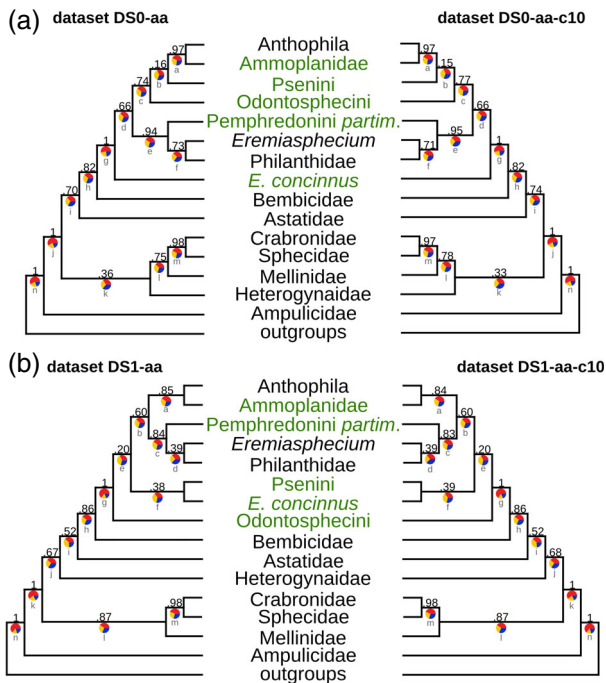


Fig. 2. Comparison of ASTRAL multi-species coalescent (MSC) phylogenetic trees. Shown are the results from analysing four datasets at the amino acid level: (a) DS0-aa (without splits collapsed) and DS0-aa-10c (with collapsed splits, bootstrap support <10) and (b) DS1-aa (without splits collapsed) and DS1-aa-10c (with collapsed splits, bootstrap support <10). Numbers along branches represent ASTRAL branch support values. Tripartitioned circles represent the alternative quartet topology scores defined as quartet topology one (red), two (yellow) and three (blue), with red representing the proportion of the MSC topology and blue and yellow the alternative quartet topologies. Taxa of the polyphyletic group Pemphredoninae are coloured in green. Letters a–n specify the phylogenetic splits for which the three local posterior probabilities and the three quartet topology scores are given in Table S11. [Colour figure can be viewed at wileyonlinelibrary.com].

supermatrix is analyzed at the nucleotide level, with *Spilomena beata* Blüthgen clustering either with *Eremiasphecium* sp. or as sister to *Eremiasphecium* sp. and Pemphredonini partim (Fig. 3). Psenidae, comprising Psenini and Odontosphecini (Sann *et al.*, 2018), is also recovered with moderate support when analysing the data on nucleotide level (Fig. 3; DS-nt12), but is paraphyletic when analysing the data on the amino acid level (Fig. 2; DS-aa).

The placement of *Eremiasphecium* sp. and *E. concinnus* is ambiguous (Figs 2, 3; Table 1). In three of the four topologies inferred by ASTRAL, *E. concinnus* is placed with strong support as sister group to a clade comprising Ammoplanidae, Anthophila, *Eremiasphecium* sp., Pemphredonini partim, Philanthidae, Psenini and Odontosphecini (Figs 2a, 3). We found *Eremiasphecium* sp. either as sister group of Philanthidae (Fig. 2) with high support, or with low support as sister group of *Spilomena beata* (Fig. 3a) or of Pemphredonini partim (Fig. 3b).

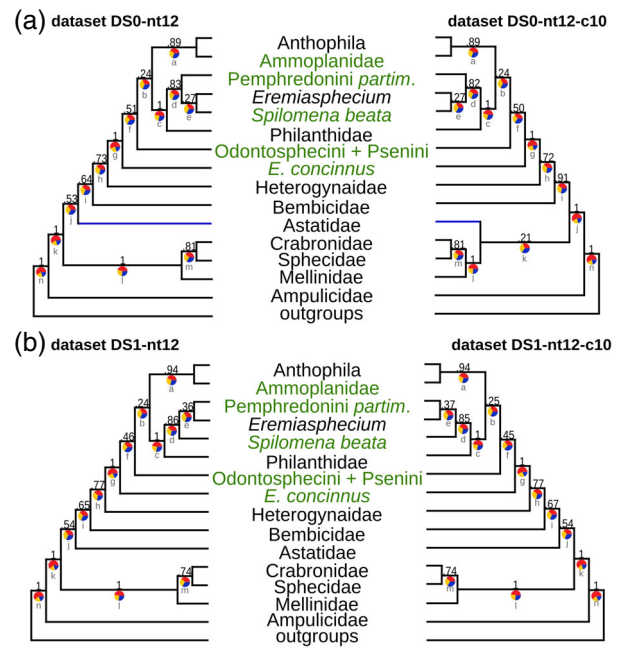


Fig. 3. Comparison of ASTRAL multi-species coalescent (MSC) phylogenetic trees. Shown are the results from analysing four datasets at the nucleotide level: (a) DS0-nt12 (without splits collapsed) and DS0-nt12-10c (with collapsed splits, bootstrap support <10) and (b) DS1-nt12 (without splits collapsed) and DS1-nt12-10c (with collapsed splits, bootstrap support <10). Species are merged according to their potential taxonomic grouping. Differences between the topologies are indicated by a blue line. Numbers along branches represent ASTRAL branch support values. Tripartitioned circles represent alternative quartet topology scores defined as quartet topology one (red), two (yellow) and three (blue) with red representing the proportion of the MSC topology and blue and yellow the alternative quartet topologies. Taxa of the polyphyletic group Pemphredoninae are coloured in green. Letters a–n specify the phylogenetic splits for which the three local posterior probabilities and the three quartet topology scores are given in Table S11. [Colour figure can be viewed at wileyonlinelibrary.com].

The phylogenetic position of Heterogynaidae remained ambiguous and received low to moderate support in all MSC-inferred phylogenetic trees (Figs 2, 3; Table 1). Detailed information on the results of the MSC data analyses are provided in the electronic supplementary information section 3.2.

MSC local posterior probabilities inferred from quartet frequencies and quartet scores

We evaluated the local posterior probabilities derived from quartet frequencies across all analyzed MSC datasets that revealed maximal to high support. Detailed information on all remaining local posterior probability values and support for alternative quartet topologies are provided in Table S11. In concordance with the results obtained from the concatenation approach, we found strong support for (1) Ampulicidae as sister group of all remaining Apoidea, (2) a monophyly of Mellinidae + (Crabronidae + Sphecidae), (3) *E. concinnus* as sister group

to a clade comprising Ammoplanidae, Anthophila, *Eremiasphecium* sp., Pemphredonidae, Philanthidae and Psenidae, and (4) Ammoplanidae as sister group of the bees (Figs 2, 3; Tables 1 and S11).

Discussion

Phylogenetic relationships of Apoidea

The phylogenetic relationships of Apoidea inferred in the present study are largely congruent to those presented by Sann *et al.* (2018), especially with respect to well-established clades (Figs 1–3). Specifically, our results are consistent with a monophyly of the families Ammoplanidae, Astatidae, Ampulicidae, Bembicidae, Crabronidae, Mellinidae and Sphecidae *sensu* Sann *et al.* (2018) (Figs 1–3). In contrast, the results of some of our analyses rendered the families Psenidae (Odontophecini and Psenini) and Pemphredonidae (Entomosericini, Pemphredonina, Spilomenina and Stigmina) para- or polyphyletic (Figs 2, 3). Despite the unclear phylogenetic placement of *Entomosericus* in the apoid wasp phylogeny, our analyses revealed that inclusion of this taxon in the family Pemphredonidae cannot be well justified. We therefore suggest raising the former Entomosericinae to family rank, Entomosericidae Dalla Torre, 1897 (stat. n.) to acknowledge this situation in apoid systematics. By doing so, the monophyly of the remaining Pemphredonidae (i.e. the former tribe Pemphredonini) is in most of our analyses re-established (the only exception being the MSC analyses of the nucleotide data, in which *Eremiasphecium* rendered Pemphredonini polyphyletic). For similar reasons, we suggest raising the former Eremiasphecini also to family rank, Eremiasphecidae Menke, 1967 (stat. n.). Although we cannot robustly infer the phylogenetic position of the genus *Eremiasphecium* in the apoid wasp phylogenetic tree, our analyses indicate that this taxon has no strong ties to Pemphredonidae or Psenidae, which would justify its inclusion in either of them. Since *Eremiasphecium* had been granted its own subfamily in the former Crabronidae, granting it now its own family is a logical consequence of splitting the former polyphyletic family ‘Crabronidae’ into multiple families.

Prentice (2000) and Hanson & Menke (2006) discussed a close phylogenetic relationship between Eremiasphecidae, or of a more comprehensive clade consisting of Eremiasphecidae, Pemphredonidae and Philanthidae, with bees based on phylogenetic analyses of morphological characters. However, our rigorous phylogenetic analysis strongly suggest that Eremiasphecidae are less closely related to bees than Ammoplanidae. The same holds true for Entomosericidae.

An unexpected result in our study has been the inability to infer the position of the enigmatic wasp family Heterogynaidae in the apoid wasp phylogeny. Despite the fact that we included a substantial larger amount of nucleotide sequence data than Sann *et al.* (2018), the phylogenetic inferences remained inconclusive. Thus, we can neither corroborate nor reject the hypothesis put forth by Sann *et al.* (2018) that Heterogynaidae represent a subordinated lineage of the tribe Nyssonini within the family Bembicidae (Table 1).

When comparing the topologies inferred from analysing different datasets and from applying different phylogeny inference methods, a striking pattern that we found is that topologies inferred with the MSC approach are largely incompatible among each other and with topologies inferred from applying the concatenation approach. One reason for this pattern could be that the phylogenetic signal contained in the MSA of individual genes is too limited to infer a reliably gene tree. We therefore tend to favour the results from the concatenation-based phylogenetic inferences, which recovered all of the currently apoid wasp families recognized by us (i.e. Ammoplanidae, Astatidae, Bembicidae, Crabronidae, Entomosericidae, Eremiasphecidae, Heterogynaidae, Mellinidae, Pemphredonidae, Philanthidae, Psenidae and Sphecidae) monophyletic. However, our inability to reliably infer the phylogenetic relationships of some of these families to each other indicate that classical phylogenomic approaches (i.e. those that analyse the primary nucleotide, or encoded amino acid, sequence information of genomes) have their limits. Thus, consideration of genomic meta-characters, such as gene content and near-intron pairs (see Niehuis *et al.*, 2012), are required to trace the evolutionary origin of some apoid wasp lineages, including that of the enigmatic family Heterogynaidae.

Author's contributions

Project idea: MS. Experimental design: KM, MS, ON, TP. Contributed materials and reagents: CSE KM, MM, MO, MS, ON. Molecular procedures: MS. Bioinformatics: HE, KM, MS, TP. Manuscript preparation: all authors contributed to the writing of the manuscript, with KM and MS taking the lead. All authors read and approved the final manuscript.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

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

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

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

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

Fig. S5. Best ML phylogenetic tree inferred from re-analysing the concatenated nucleotide dataset with first and second codon positions included (sm-nt12).

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

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

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 analyzed 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 first and second 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 of the UniqueTree analysis when examining the 50 ML trees of dataset sm-nt123.

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

Table S14. Statistical support of the 12 alternative topologies inferred from dataset sm-nt12.

Table S15. Statistical support of the 19 alternative topologies inferred from dataset sm-nt123.

Acknowledgements

We thank Ondrej Hlinka and the CSIRO HPC team (Australia) for granting us access and help with analyses on the CSIRO

HPC Cluster. We are grateful to Siavash Mirarab for the helpful discussions on MSC data analysis and data interpretation. Special thanks to Thomas Wong for the help with UniqueTree data analysis and Alexander Donath, Jan Philip Oyen, James B. Pease and Alexandros Vasilikopoulos for comments and help with the Quartet Sampling data analysis. Photos of *Eremiasphecium arabicum*, *Entomosericus concinnus* and *Heterogyna nocticola* were kindly provided by Michael Ohl and Christian Schmid-Egger. The authors declare that they have no competing interests.

Open Access funding enabled and organized by Projekt DEAL. WOA Institution: ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG Blended DEAL: Projekt DEAL.

Data availability statement

The data that support the findings of this study are openly available in Dryad Repository at <https://datadryad.org>, doi:10.5061/dryad.pc866t1nj. Supplementary data are available at Systematic Entomology online.

References

- Ababneh, F., Jermin, L.S., Ma, C. & Robinson, J. (2006) Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, **22**, 1225–1231.
- Aberer, A., Krompass, D. & Stamatakis, A. (2013) Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology*, **62**, 162–166.
- Barrow, L.N., Lemmon, A.R. & Lemmon, E.M. (2018) Targeted sampling and target capture: assessing phylogeographic concordance with genome-wide data. *Systematic Biology*, **67**, 979–996.
- Betancur-R, R., Arcila, D., Vari, R.P., Hughes, L.C., Oliveira, C., Sabaj, M.H. & Ortí, G. (2019) Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: the monophyly of characiform fishes. *Evolution*, **73**, 329–345.
- Bohart, R.M. & Menke, A.S. (1976) *Sphecid Wasps of the World*. University of California, Berkeley, California.
- Bowker, A.H. (1948) A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**, 572–574.
- Branstetter, M.G., Danforth, B.N., Pitts, J.P. et al. (2017) Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Current Biology*, **27**, 1019–1025.
- Chernomor, O., Von Haeseler, A. & Minh, B.Q. (2016) Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, **65**, 997–1008.
- Degnan, J.H. & Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, **24**, 332–340.
- Efron, B. (1979) Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*, **21**, 460–480.
- Evangelista, D., Thouzé, F., Kohli, M.K., Lopez, P. & Legendre, F. (2018) Topological support and data quality can only be assessed through multiple tests in reviewing Blattodea phylogeny. *Molecular Phylogenetics and Evolution*, **128**, 112–122.
- Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
- Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013) QUASt: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

- Hanson, P. & Menke, A.S. (2006) Capítulo 17. Las avispas apoideas: Ampulicidae, Sphecidae, Crabronidae. *Hymenoptera de la Región Neotropical*, Vol. 77 (ed. by P.E. Hanson and I.D. Gauld), pp. 694–733. Gainesville, Florida: Memoirs of the American Entomological Institute.
- Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, **35**, 518–522.
- Huelsenbeck, J.P. & Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
- Huerta-Cepas, J., Serra, F. & Bork, P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, **33**, 1635–1638.
- Hurvich, C.M. & Tsai, C.L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
- Jermiin, L.S., Ho, S.Y., Ababneh, F., Robinson, J. & Larkum, A.W. (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, **53**, 638–643.
- Jermiin L. & Ott M (2017). SymTest.
- Junier, T. & Zdobnov, E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, **26**, 1669–1670.
- Kajitani, R., Toshimoto, K., Noguchi, H. *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 1384–1395.
- Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.
- Kapli, P., Yang, Z. & Telford, M.J. (2020) Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, **21**, 1–17.
- Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
- Kishino, H., Miyata, T. & Hasegawa, M. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, **31**, 151–160.
- Kück, P. & Longo, G.C. (2014) FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*, **11**, 81.
- Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B.M., Wägele, J.W. & Misof, B. (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology*, **7**, 10.
- Le, S.Q., Dang, C.C. & Gascuel, O. (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution*, **29**, 2921–2936.
- Melo, G.A. (1999). *Phylogenetic relationships and classification of the major lineages of Apoidea (Hymenoptera), with emphasis on the crabronid wasps*. Scientific papers, Natural History Museum, The University of Kansas, No. 14
- Michener, C.D. (2000) *The Bees of the World*, Vol. 1. Johns Hopkins University Press, Baltimore, Maryland.
- Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. & Warnow, T. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.
- Mirarab, S. & Warnow, T. (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, **31**, 44–52.
- Misof, B., Liu, S., Meusemann, K. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.
- Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K. & Meusemann, K. (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics*, **14**, 348.
- Misof, B. & Misof, K. (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology*, **58**, 21–34.
- Molloy, E.K. & Warnow, T. (2018) To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, **67**, 285–303.
- Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2014) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.
- Niehuis, O., Hartig, G., Grath, S. *et al.* (2012) Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. *Current Biology*, **22**, 1309–1313.
- Ohl, M. & Bleidorn, C. (2006) The phylogenetic position of the enigmatic wasp family Heterogynaidae based on molecular data, with description of a new, nocturnal species (Hymenoptera: Apoidea). *Systematic Entomology*, **31**, 321–337.
- Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M. & Stamatakis, A. (2010) How many bootstrap replicates are necessary? *Journal of Computational Biology*, **17**, 337–354.
- Pease, J.B., Brown, J.W., Walker, J.F., Hinchliff, C.E. & Smith, S.A. (2018) Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany*, **105**, 385–403.
- Peters, R.S., Krogmann, L., Mayer, C. *et al.* (2017) Evolutionary history of the Hymenoptera. *Current Biology*, **27**, 1013–1018.
- Petersen, M., Meusemann, K., Donath, A. *et al.* (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, **8**, 111.
- Prentice MA (2000). *The Comparative Morphology and Phylogeny of Apoid Wasps (Hymenoptera: Apoidea)*. Dissertation Thesis: University of California, Berkeley.
- Pryszcz, L.P. & Gabaldón, T. (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, **44**, e113.
- Pulawski, W.J. (2019) *Catalog of Sphecidae sensu lato (= Apoidea Excluding Apidae)* [WWW document]. URL <http://www.calacademy.org/scientists/projects/catalog-of-sphecidae> [accessed 2019].
- Rokas, A., Williams, B.L., King, N. & Carroll, S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.
- Romiguier, J., Cameron, S.A., Woodard, S.H., Fischman, B.J., Keller, L. & Praz, C.J. (2016) Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. *Molecular Biology and Evolution*, **33**, 670–678.
- Sann, M., Niehuis, O., Peters, R.S. *et al.* (2018) Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. *BMC Evolutionary Biology*, **18**, 71.
- Sayyari, E. & Mirarab, S. (2016) Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, **33**, 1654–1668.
- Shimodaira, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, **51**, 492–508.
- Simion, P., Delsuc, F. & Philippe, H. (2020) To what extent current limits of phylogenomics can be overcome? *Phylogenetics in the Genomic Era* (ed. by C. Scornavacca, F. Delsuc and N. Galtier), pp. 2.1:1–2.1:34. CCSD, Villeurbanne.
- Smith, S.A., Moore, M.J., Brown, J.W. & Yang, Y. (2015) Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, **15**, 150.

- Springer, M.S. & Gatesy, J. (2016) The gene tree delusion. *Molecular Phylogenetics and Evolution*, **94**, 1–33.
- Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
- Strimmer, K. & Von Haeseler, A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences*, **94**, 6815–6819.
- Suyama, M., Torrents, D. & Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, W609–W612.
- Wong, T.K.K., Kalyaanamoorthy, S., Meusemann, K., Yeates, D., Misof, B. & Jermiin, L.S. (2020) A minimum reporting standard for multiple sequence alignments. *NAR Genomics and Bioinformatics*, **2**, 1–5.
- Yang, Z. (2004) *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford.
- Ye, C., Ma, Z.S., Cannon, C.H., Pop, M. & Douglas, W.Y. (2012) Exploiting sparseness in *de novo* genome assembly. *BMC Bioinformatics*, **13**, S1.
- Young, A.D. & Gillung, J.P. (2019) Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, **45**, 225–247.
- Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 153.

Accepted 1 March 2021
First published online 25 March 2021