Evolutionary History of the Hymenoptera

Highlights

- The most comprehensive dataset ever compiled for inferring the phylogeny of Hymenoptera
- A major radiation of primarily ectophytic sawflies (Eusymphyta) is hypothesized
- A major radiation of parasitoid wasps (Parasitoida) is identified
- The phylogenetic origins of wasp-waisted wasps, stinging wasps, and bees are resolved

Authors
Ralph S. Peters, Lars Krogmann, Christoph Mayer, ..., Jes Rust, Bernhard Misof, Oliver Niehuis

Correspondence
r.peters@leibniz-zfmk.de (R.S.P.), oliver.niehuis@biologie.uni-freiburg.de (O.N.)

In Brief
Peters et al. infer a time-calibrated and statistically solid phylogenetic tree of the mega-diverse insect order Hymenoptera (sawflies, wasps, ants, and bees) from the analysis of phylogenomic data. This sheds new light on the early history of this intriguing group, as well as on the origins and radiation of parasitoids, stinging wasps, and bees.
Evolutionary History of the Hymenoptera

Ralph S. Peters,1,24,25,* Lars Krogmann,2 Christoph Mayer,3 Alexander Donath,3 Simon Gunke,4 Karen Meusemann,3,5,6 Alexey Kozlov,7 Lars Podsiadlowski,3 Malte Petersen,3 Robert Lanfear,9,10 Patricia A. Diez,11 John Heraty,12 Karl M. Kjer,13 Seraina Klopfstein,14 Rudolf Meier,16 Carlo Polidori,16 Thomas Schmitt,17 Shianlin Liu,18,19,20 Xin Zhou,21,22 Torsten Wappler,4 Jes Rust,4 Bernhard Misof,3 and Oliver Niehuis3,5,23,24,*

1Center of Taxonomy and Evolutionary Research, Arthropoda Department, Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany
2Entomologie, Staatliches Museum für Naturkunde Stuttgart, 70191 Stuttgart, Germany
3Center for Molecular Biodiversity Research, Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany
4Steinmann Institut für Geologie, Mineralogie und Paläontologie, 53115 Bonn, Germany
5Department of Evolutionary Biology and Ecology, Institute for Biology I (Zoology), University of Freiburg, 79104 Freiburg (Brs.), Germany
6Australian National Insect Collection, CSIRO National Research Collections Australia (NRCA), Acton, ACT 2601, Australia
7Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany
8Institute of Evolutionary Biology and Ecology, University of Bonn, 53121 Bonn, Germany
9Ecology, Evolution and Genetics, Research School of Biology, Australian National University, Canberra, ACT 2601, Australia
10School of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia
11Centro de Investigaciones y Transferencia de Catamarca, CITCA-CONICET/UNCA, 4700 Catamarca, Argentina
12Department of Entomology, University of California, Riverside, Riverside, CA 92521, USA
13Department of Biological Sciences, Rutgers University, Newark, NJ 07102, USA
14Naturhistorisches Museum der Burgergemeinde Bern, 3005 Bern, Switzerland
15Department of Biological Sciences and Lee Kong Chian Natural History Museum, National University of Singapore, Singapore 117543, Singapore
16Instituto de Ciencias Ambientales (ICAM), Universidad de Castilla-La Mancha, 45071 Toledo, Spain
17Department of Animal Ecology and Tropical Biology, University of Würzburg, 97074 Würzburg, Germany
18China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen, Guangdong Province, 518083, People’s Republic of China
19BGI-Shenzhen, Shenzhen, Guangdong Province, 518083, People’s Republic of China
20Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark
21Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, People’s Republic of China
22Department of Entomology, China Agricultural University, Beijing 100193, People’s Republic of China
23School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
24These authors contributed equally
25Lead Contact
*Correspondence: r.peters@leibniz-zfmk.de (R.S.P.), oliver.niehuis@biologie.uni-freiburg.de (O.N.)
http://dx.doi.org/10.1016/j.cub.2017.01.027

SUMMARY

Hymenoptera (sawflies, wasps, ants, and bees) are one of four mega-diverse insect orders, comprising more than 153,000 described and possibly up to one million undescribed extant species [1, 2]. As parasitoids, predators, and pollinators, Hymenoptera play a fundamental role in virtually all terrestrial ecosystems and are of substantial economic importance [1, 3]. To understand the diversification and key evolutionary transitions of Hymenoptera, most notably from phytophagy to parasitoidism and predation (and vice versa) and from solitary to eusocial life, we inferred the phylogeny and divergence times of all major lineages of Hymenoptera by analyzing 3,256 protein-coding genes in 173 insect species. Our analyses suggest that extant Hymenoptera started to diversify around 281 million years ago (mya). The primarily ectophytophagous sawflies are found to be monophyletic. The species-rich lineages of parasitoid wasps constitute a monophyletic group as well. The little-known, species-poor Trigonaloidea are identified as the sister group of the stinging wasps (Aculeata). Finally, we located the evolutionary root of bees within the apoid wasp family “Crabronidae.” Our results reveal that the extant sawfly diversity is largely the result of a previously unrecognized major radiation of phytophagous Hymenoptera that did not lead to wood-dwelling and parasitoidism. They also confirm that all primarily parasitoid wasps are descendants of a single endophytic parasitoid ancestor that lived around 247 mya. Our findings provide the basis for a natural classification of Hymenoptera and allow for future comparative analyses of Hymenoptera, including their genomes, morphology, venoms, and parasitoid and eusocial life styles.

RESULTS AND DISCUSSION

We sequenced whole-body transcriptomes of 167 species of Hymenoptera and selected outgroups and supplemented our
dataset with sequenced and annotated genomes of five hymenopterans and a beetle (for details, see Supplemental Experimental Procedures and Data S1A–S1D). Our study includes 54 families of Hymenoptera, representing all major superfamilies. The phylogenetic inferences are based on the analysis of 1.5 million amino acid and 3.0 million nucleotide positions, respectively, derived from 3,256 single-copy protein-coding genes (Data S1E) and inferred by using a combination of domain-, gene-, and codon position-based data partition schemes to improve the fitting of the applied substitution models. Considering the taxonomic and molecular sampling, this is the most comprehensive dataset ever generated for investigating phylogenetic relationships within Hymenoptera or any other insect group. The dataset was furthermore used to estimate divergence times with an independent-rates as well as with a correlated-rates molecular clock approach (Data S1H) and a validated set of 14 fossils (Data S1F).

The inferred phylogenetic relationships and divergence time estimates were used to assess where in the phylogeny of Hymenoptera, when in their geological history, and how often major evolutionary transitions took place. Specifically, we studied the switch from feeding on plants to feeding on an insect host (parasitoidism), the formation of a wasp waist, the evolution of a venomous stinger to subdue mobile hosts, the evolution of eusociality, and the switch from hunting prey to collecting pollen. These evolutionary transitions are partially reflected by the historical classification of Hymenoptera: sawflies (“Symphyta”) are those Hymenoptera that lack the wasp waist that characterizes all remaining Hymenoptera (Apocrita), “Parasitica” encompasses the primarily parasitoid Apocrita that lack a stinger, and Aculeata comprises the stinging wasps, ants, and bees (Anthophila) [1]. Yet, how many major lineages each of these groups encompasses has been controversial for decades [4–11].

The results of our phylogenomic study received strong support in all analyses, unless stated otherwise, and alter previous ideas regarding the evolutionary history of Hymenoptera (Figure 1B; for full results and detailed experimental procedures, see Figure S1, Supplemental Experimental Procedures, and additional figures deposited at Mendeley Data, http://dx.doi.org/10.17632/s5j2f6z3d.2). According to our analyses, extant Hymenoptera started to diversify between the Carboniferous and the Triassic (95% confidence interval [CI]: 329–239 million years ago [mya]; mean: 281 mya; node 1 [n.1] in Figure 1B), with the oldest currently known Hymenoptera fossils being from the Triassic, ~224 million years old [8]. Previous studies suggested this divergence to have occurred between the sawfly lineage Xyeloidea and the remaining Hymenoptera [5, 7–11], whereas our analysis identified a much more inclusive clade of sawflies (Eusymphyta; n.2) that also contains Pamphilioida and Tenthredinoidea as closest relatives of all remaining Hymenoptera (Unicalcarida). These superfamilies had been thought to form a paraphyletic grade [5, 7, 9, 11]. Instead, they represent an unexpected and previously unrecognized major radiation of primarily ectophytophagous insects that comprises more than 7,000 described species [1]. We estimate the first diversification of the extant eusymphytan lineages to have occurred 276–157 mya (mean 212 mya). Note that Eusymphyta were corroborated as the sister group of all remaining Hymenoptera when additionally scrutinizing the analyzed molecular data for conflicting phylogenetic signal (Supplemental Experimental Procedures). Given the novelty and importance of our finding, we anticipate that it will significantly influence future research on Hymenoptera relationships, and we encourage researchers to further assess this particular phylogenetic hypothesis in future studies, for example by extending the taxon sampling within Eusymphyta and the outgroup.

A clade Eusymphyta representing the extant sister lineage of all remaining Hymenoptera (Unicalcarida) has profound consequences for inferring ground-plan characters of Hymenoptera. For example, Hymenoptera were previously thought to have been ancestrally ectophytophagous, based on the assumption that eusymphytans form a paraphyletic assemblage. Considering that the sister group of Hymenoptera (Aparaglossata) was ancestrally likely predacious [12], the inferred relationship between Eusymphyta and Unicalcarida implies that the most recent common ancestor of Hymenoptera could have been ecto- or endophytophagous. A sister group relationship between Eusymphyta and Unicalcarida furthermore implies that the remarkable ability of male Hymenoptera to restore diploidy in their muscle cells was already present in the last common ancestor of all Hymenoptera (with a secondary loss in Xyelidae), or that this feature evolved at least twice (in Unicalcarida and Tenthredinoidea) [13]. Finally, the unexpected finding that the turnip sawfly, Athalia rosae (Tenthredinoidea), whose genome has recently been sequenced by the i5K initiative [14], is a representative of the sister lineage of all remaining Hymenoptera will improve our understanding of the genetic composition of the most recent common ancestor of Hymenoptera: genomic features shared between the turnip sawfly and species of Unicalcarida with sequenced genomes (e.g., Nasonia parasitoid wasps, ants, bees) were likely inherited from their common ancestor.

In agreement with earlier studies [9, 10], we found a single origin of the endophytic sawfly lineages (i.e., Cephoidea, Orussoidea, Siricoidea, and Xiphydrioida; n.3), which form a paraphyletic grade, in which Orussoidea (parasitoid woodwasps) represent the closest relatives of Apocrita (n.4). Morphological data have suggested a sister group relationship of Orussoidea and Apocrita (Vespina) [6, 15], but results from analyzing molecular data have been inconsistent [7, 9]. Our analyses provide strong support for the monophyly of Vespina and of Apocrita (n.5) and imply that the bulk of primarily parasitoid wasps are descendants of a single endophytic parasitoid ancestor that lived in the Permian or in the Triassic (CI: 289–211 mya; mean: 247 mya). Contrary to earlier hypotheses of sawfly relationships (see [10]), we identified Cephoidea, and not Siricoidea and/or Xiphydrioida, as the closest extant relatives of Vespina (n.6), a result only recently suggested [7].

The evolution of the wasp waist, a constriction between the first and the second abdominal segment greatly improving the maneuverability of the abdomen’s rear section, including the ovipositor, was a major innovation in the evolution of Hymenoptera that undoubtedly contributed to the rapid diversification of Apocrita (n.5) [6]. Our analysis is the first to persuasively demonstrate that the most diverse parasitoid wasp lineages (i.e., Ceraphronoidea, Ichneumonoidea, and Proctotrupomorpha) constitute a natural group (Parasitoida; n.7) whose astonishing radiation was likely triggered by further optimization of the parasitoid lifestyle and related traits (e.g., endoparasitoidism,
Figure 1. Evolutionary History of the Hymenoptera

(A) Representatives of sawflies, wasps, ants, and bees. Scale bars represent 5 mm.

(B) Phylogenetic relationships and divergence time estimates of Hymenoptera. Key evolutionary events are indicated at the respective clades (note that only the major eusocial lineages are considered). The tree was inferred under the maximum-likelihood optimality criterion, analyzing 1,505,514 amino acid sites and applying a combination of protein domain- and gene-specific substitution models. Divergence times were estimated with an independent-rates molecular clock approach and considering 14 validated fossils. Triangular branches cover multiple species (number of species in parentheses) whose relationships are shown in detail in Figure S1. Nodes with circled numbers are referred to in the main text.
miniaturization), which allowed for successfully attacking a variety of new hosts. We estimate the beginning of the group’s radiation at 266–195 mya (mean: 228 mya), only a few million years after Parasitoida separated from the remaining Apocrita (CI: 276–203 mya; mean: 236 mya). The early radiation of Parasitoida thus falls within a time period when the parasitoids’ major host lineages (e.g., Hemiptera, Holometabola) also started to diversify [16].

We identified the enigmatic Trigonaloidea as the closest extant relatives of Aculeata with strong node support (n.8), a hypothesis only recently put forth [7, 9]. Evanioidea, which had also been discussed as a possible sister group of Aculeata [5, 10, 17, 18], cluster with Stephanioidea (n.9). Node support for this relationship is low, however, and it needs to be investigated further in future studies that include additional types of characters and samples of Megaloryoidea, a lineage that we were unable to sequence. Note that in contrast to Aculeata, the Evanioidea, Stephanioidea, and Trigonaloidea have all remained species-poor. The identification of the closest relatives of Aculeata will be important for better understanding which traits (e.g., venoms) fostered the diversification of the stinging wasps.

Our analysis sheds new light on the phylogeny of Aculeata (n.10), whose early diversification occurred 224–160 mya (mean: 190 mya). Chrysidoidea are confirmed as the sister group of all remaining Aculeata [19]. We corroborate the artificial nature of the former superfamiliy “Vespoidae” (i.e., all Aculeata except Apoidea and Chrysidoidea) [5], which comprises four major lineages that are paraphyletic with respect to Apoidea [20]. The potter, honey, and social wasps (Vespoidea sensu Pilgrim et al. [20]: Vespidae; n.11) were identified as the sister lineage of all remaining non-chrysoidoid Aculeata. However, the phylogenetic position of the species-poor Rhopalosomatidae (Vespoidea sensu Pilgrim et al. [20]), an aculeate wasp family that we were unable to sequence and possible sister lineage of Vespidae, remains controversial [9, 10, 20]. The inferred phylogenetic relationships within Vespidae suggest two independent origins of eusociality, a previously fiercely contested hypothesis [21, 22]. In agreement with an earlier phylogenomic study [23], we inferred ants (Formicoidea) as being the closest extant relatives of Apoidea (n.12) in all of our analyses, except when applying a Bayesian approach, which suggested ants plus scoliid wasps (Scolioidea, possibly including also the family Bradynobaenidae [20], which we were unable to sequence) as being sister to Apoidea (figure deposited at Mendeley Data, http://dx.doi.org/10.17632/e5j2f62z3d.2). We estimate the last common ancestor of ants and Apoidea to have lived in the Jurassic or the Cretaceous (CI: 192–136 mya; mean: 162 mya).

We located the phylogenetic origin of bees (Anthophila) within the apoid wasp family “Crabronidae” (n.13), which our study shows to be an artificial construct comprising five major lineages. The crabronid wasp lineage in our study most closely related to bees is the species-poor tribe Psenini. This result substantiates the idea that the switch from a predatory to a herbivorous lifestyle was a key to the tremendous diversification of bees [24]. We estimate the origin of bees to have been in the Cretaceous (CIs: 147–93 mya; means: 124 and 111 mya), a result that is consistent with a close temporal link between the diversifications of bees and angiosperms [24]. Melittid bees were identified as the sister lineage of all remaining Anthophila (n.14), which implies that short-tongued bees do not represent a natural group. In contrast, we confirmed long-tongued bees (i.e., Apidae and Megachilidae) to constitute a natural entity (n.15) [24]. We also found the eusocial apid bee lineages to be monophyletic, corroborating the hypothesis that eusociality has evolved once, not twice, in corbiculate (pollen basket) bees (n.16) [25].

Our study confirms the power of phylogenomic approaches for deciphering difficult-to-resolve arthropod phylogenetic relationships [12, 16, 26, 27] by yielding well-supported answers to some of the most pressing questions regarding the evolutionary history of the sawflies, wasps, ants, and bees. We provide strong evidence for understanding the phylogenetic relationships among all major lineages of Hymenoptera, and we were able to date the individual divergence events, both paramount for deciphering the tempo and mode of diversification of ecologically, economically, sociobiologically, and/or pharmaceutically relevant traits of interest (e.g., gene repertoires, haplodiploidy and sex determination, eusociality, chemosensation, and venoms). Finally, our study offers the basis for establishing a natural classification of the insect order Hymenoptera.

**EXPERIMENTAL PROCEDURES**

We sequenced the transcriptomes of 134 species of Hymenoptera using Illumina HiSeq 2000 sequencing technology (Data S1A–S1C). We complemented our dataset by including previously published transcriptomes of 29 Hymenoptera and four Neuropteroida [16, 28]. Finally, we considered the official gene sets of five Hymenoptera and the flour beetle Tribolium castaneum (Data S1D). All paired-end reads were assembled with SOAPdenovo-Trans (version 1.01) [29]. The assembled transcripts were filtered for possible contaminants, and the raw reads and filtered assemblies were submitted to the NCBI SRA and TSD archives. We searched the assemblies with the software Orthograph (version beta) [28] for transcripts of 3,260 protein-coding genes that the OrthoDB v7 database [30] suggested to be single-copy in Hymenoptera and Neuropteroidea (outgroup) by applying the best reciprocal hit criterion. Orthologous transcripts were aligned with MAFFT (version 7.017) [31] at the translational (amino acid) level. All multiple sequence alignments (MSAs) were quality assessed and, if necessary, improved and masked using the procedure outlined by Misof et al. [16]. The resulting MSAs were concatenated to a supermatrix that we simultaneously partitioned based on a combination of Pfam protein domains and genes [16]. The phylogenetic information content of each partition was assessed with MARE (version 0.1.2-rc) [32], and all uninformative partitions were removed. We subsequently used PartitionFinder (developer versions 2.0.0-pre2, 2.0.0-pre9, and 2.0.0-pre10) [33] to simultaneously infer a partition scheme and proper amino acid substitution models for analyzing each partition with the rcluster algorithm. We applied the same partition scheme when analyzing the corresponding supermatrix at the transcriptional (nucleotide) level, except that we modeled genetic trees were reconstructed with ExaML (versions 3.0.15 and 3.0.17) [34], conducting 50 independent tree searches per supermatrix. Node support was inferred with the bootstrap method [35]. Decisive datasets were used for testing the possible impact of missing data at the partition level on the inferred phylogenetic tree [36], and four-cluster likelihood mapping was used for assessing the phylogenetic signal for alternative phylogenetic relationships [37]. Permutation tests allowed assessing the impact of heterogeneous amino acid sequence composition, non-stationarity of substitution processes, and non-random distribution of missing data on the inferred phylogenetic tree [16]. We additionally conducted phylogenetic inferences in a Bayesian framework, using ExaBayes [38] with its default settings, enabling automatic substitution model detection and applying the same data partitioning scheme that we used in analyses under the maximum-likelihood optimality criterion. We analyzed three independent runs with four coupled Markov chain Monte Carlo
chains and 200,000 generations each. The consense tool (part of the ExaBayes software package) was used to obtain a consensus tree based on the extended majority rule method (MRE), discarding the first 25% of the sampled topologies as burn-in. Divergence times were calibrated using 14 fossils (Data S1F), selected following best-practice recommendations [39] and representing extant lineages distributed across the entire Hymenoptera Tree of Life. Divergence times were estimated with mcmtree in conjunction with codeml (both part of the PAML software package, version 4.9) [53]. We analyzed a subset of the amino acid and of the nucleotide supermatrix, both comprising only sites that had amino acids or nucleotides present in at least 95% of the species, both with an independent-rates model and with a correlated-rates model (Figure 1B; Data S1H) and sampling parameters previously assessed for convergence of results.

**Data Resources**

Data reported in this paper have been published in Mendeley Data and are available at http://dx.doi.org/10.17832/tbj4zm2n2.2 (inferred matrices and statistics) and http://dx.doi.org/10.17832/s5j26z2zd.2 (figures). All sequencing data are available at NCBI via the Umbrella BioProject accession number PRJNA183205 (“The 1KITE project: evolution of insects”).

**SUPPLEMENTAL INFORMATION**

Supplemental Information includes one figure, Supplemental Experimental Procedures, and one dataset and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.01.027.

**AUTHOR CONTRIBUTIONS**


**ACKNOWLEDGMENTS**

The presented data are the result of the collaborative efforts of the 1KITE consortium. The sequencing and assembly of the 1KITE transcriptomes were funded by BGI through support to the China National GeneBank. We thank L.K., O.N., and R.S.P. for help with identification of samples and R. Allemand, E. Altenhofer, S. Blank, A. Dorchin, J. Gusenleitner, V. Mauss, C. Schmid-Egger, and M. Ziesmann (Data S1F), selected following best-practice recommendations [39] and as blueprints for illustrating our main figure. U. Vaartjes kindly helped with illustrating the main figure. A.D., B.M., C.M., J.R., L.P., M.P., O.N., R.S.P., and S.G. were supported by the Leibniz Graduate School for Genomic Biodiversity Research. C.P. was supported by the Universidad de Castilla-La Mancha and the European Social Fund (ESF). O.N. and T.S. acknowledge the German Research Foundation (DFG) for supporting parts of this study (NI 1387/1-1; SCHM 2645/2-1).

Received: September 24, 2016
Revised: December 13, 2016
Accepted: January 16, 2017
Published: March 23, 2017

**REFERENCES**

Supplemental Information

Evolutionary History of the Hymenoptera

Figure S1. Inferred Ultrametric and Time-Calibrated Tree with Node Numbers, Related to Figure 1

The phylogenetic tree is identical to the one shown in Figure 1, but additionally shows the estimated divergence times in million years ago (mya; black numbers behind nodes) and 95% confidence intervals (green bars). The consecutively numbered nodes (blue numbers before nodes) refer to Data S1H, in which all estimated divergence times and 95% confidence intervals are listed.
Supplemental Experimental Procedures

1. Sampling of transcriptomes
We sampled RNA of 134 species of Hymenoptera for transcriptome sequencing (Data S1A). Whenever possible, we deposited voucher specimens (paragenophore or synngenophore) [S1] at the Zoological Research Museum Alexander Koenig (Bonn, Germany). Collected samples were ground and preserved in RNAlater (Qiagen, Hilden, Germany) and stored at +4 °C or -80 °C until further processing. One sample (Ceraphron sp.) was first stored in 99% ethanol for < 10 min. The sample was subsequently air-dried and ground in RNAlater. In all species, we sampled the RNA from the entire body of adult specimens. Detailed information (e., number of specimens and their sex) are summarized in Data S1A. Of one species (Dendrocerus carpenteri), we sequenced two different samples to ensure good data representation for this important taxon. Please note that we included in our study the raw reads of 33 already sequenced whole body transcriptomes (29 Hymenoptera and four outgroup species; marked with asterisks in Data S1A) published by us in two preceding investigations [S2, S3]. The raw reads of the nine extra samples were processed exactly as the other samples. Our sampling of transcriptomes thus comprised 164 samples of Hymenoptera, referring to 163 different species, and four outgroup taxa.

2. Transcriptome sequencing
RNA extraction, next generation sequencing (NGS) library preparation and sequencing of the prepared libraries on Illumina HiSeq 2000 sequencers (Illumina, San Diego, CA, USA) followed the protocol given by Misof et al. [S2]. However, in five species (i.e., Aleiodes testaceus, Ceraphron sp., Cosmocomoidea morrilli, Dolichurus corniculatus, Inostemma sp.), from which a total RNA yield < 3 µg was obtained, we constructed the NGS library using the TruSeq mRNA Library Prep Kit (Illumina). We sheared the purified mRNA into fragments of 160–170 bp in length using divalent cations at 98 °C. Fragment sizes and concentrations were determined with the aid of an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA) and a StepOnePlus Real-Time PCR thermocycler (Applied Biosystems, Waltham, MA, USA). All NGS libraries were paired-end sequenced on a HiSeq 2000 (Illumina) with 150 bp (all libraries except those prepared using the TruSeq kit) and 90 bp (libraries prepared with the TruSeq kit) read length. Per library, we collected approximately 2.5 Gb of raw data.

3. De novo assembly of transcriptomes
Transcript raw reads were assembled using the assembler SOAPdenovo-Trans-31kmer (version 1.01) [S4]. All raw reads were quality checked and trimmed. Specifically, we discarded (i) reads with adapter contamination (minimum length of the alignment: 15 bp; at most 3 mismatches), (ii) reads that included > 10Ns, and (iii) reads that included > 50 base pairs of low quality (i.e., Phred quality score: 2, ASCII 66 "B", Illumina 1.5+ Phred+64). All remaining reads were used for de novo assembly. For details on the assembly process, see Xie et al. [S4] and Misof et al. [S2]. In one step of the assembly, our settings differed from those of Misof et al. [S2]. In the contig forming step, linear k-mers (i.e., k-mers with a single out-degree) were merged to form edges and different edges were linked by arcs. Arcs with an abundance < 5% of the total out-degrees or < 2% of the total in-degrees were excluded. Subsequently, edges with an average abundance >= 3 were reported as contigs.

4. Identification and removal of contaminating sequences
All transcriptome assemblies were first searched for vector and linker/adapter contamination using a local installation of VecScreen (http://www.ncbi.nlm.nih.gov/tools/vecscreen/) and the UniVec database build 7.1 (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/). We removed both terminal and internal contaminations. The removal of internal contaminations resulted in a split of contigs/scaffolds. We next searched the assembled transcriptomes for cross-library contamination, as it can occur when pooling single index-tagged NGS libraries on the same Illumina NGS sequencer lane, using the search strategy outlined by Mayer et al. [S5]. Specifically, we compared each transcriptome assembly with all other assemblies sequenced in context of the 1KITE project using BLASTN of the BLAST+ (version 2.2.29) program suite [S6]. We explicitly refrained from restricting the contamination search to only those transcriptomes, which were sequenced on the same lane, in order to be able to detect also contamination that may have occurred in pre-sequencing steps. In those instances, in which BLASTN identified transcripts that shared over a length of at least 180 bp a nucleotide sequence identity of at least 98 %, we used the coverage depth to
determine, which transcript is likely the original sequence (by being more abundant) and which one likely represents the contamination (by being less abundant). We used as coverage depth of a given transcript the average k-mer coverage statistic provided by the assembly software SOAPdenovo-Trans-31kmer [S4]. Having identified transcripts sharing a high sequence identity, we applied the following procedure: (i) If two transcripts differed more than 2-fold in their relative coverage, we removed the transcript with the lower relative coverage from the corresponding assembly. (ii) If the coverage of the two transcripts in question differed 2-fold or less, we conservatively removed both of them from the two corresponding assemblies. This procedure helped with removing putative foreign contaminations (e. g., from third party libraries sequenced on the same lane, but not present in our analyses). It also meant that in case of multiple highly similar sequences, we retained only the single transcript with the highest relative coverage, given that its coverage was more than 2-fold higher than the coverage of the second-best matching transcript. We detected in samples sequenced on two specific Illumina lanes contaminations originating from the moth \textit{Plutella xylostella} (Insecta: Lepidoptera), a species that was not included in our dataset, and thus we decided to specifically search the 1KITE transcript assemblies generated from sequencing these two lanes for sequences of this moth. We retrieved all protein-coding sequences of the \textit{P. xylostella} official gene set (version 1.0) from the diamondback moth genome project database (http://iae.fafu.edu.cn/DBM/) and searched the coding sequences with BLASTN against the respective 1KITE transcriptome assemblies. Hits that exhibited 98\% sequence identity over a length of at least 180 bp were removed. We additionally removed transcripts from the assembled transcriptomes that NCBI identified as possibly foreign contaminations when submitting the assemblies to the NCBI Transcriptome Shotgun Assembly (TSA) database (see below). Information on how many transcripts were removed from each transcript library is summarized in Data S1B.

We removed per assembled transcript library between 14,258 (0.05\% of the unfiltered assembled transcriptome) and 9,925,188 (26.99\% of the unfiltered assembled transcriptome) nucleotides, as they likely or possibly represented contamination (Data S1B). However, the comparatively high number of removed nucleotides in some samples (i. e., congeneric species in the genera \textit{Eucera} and \textit{Tetraloniella}) is more likely explained by the chosen conservative approach for identifying possible contaminants than by actual excessive instances of contamination. Please note that we found a relatively small number of single-copy (target) genes in the species of these two genera. Since the two genera are deeply nested within Anthophila (bees), the small number of identified genes (Data S1E) is, however, insignificant for inferring the phylogenetic relationships of the major lineages of Hymenoptera.

5. Identification of orthologous transcripts of single-copy protein-coding genes

We used the program Orthograph (version beta4) [S3] to search the transcript libraries for transcripts that are orthologous to a set of target genes. Specifically, we selected target genes that the OrthoDB v7 database [S7] suggested to occur in single-copy across Hymenoptera and Coleoptera (beetles). Note that we included the latter to enable outgroup comparisons (i. e., rooting of the inferred tree). Specifically, target genes were selected by searching the OrthoDB v7 database [S7] for genes that are single copy (copy number = 1) in the sequenced and annotated genome of each of the following six reference taxa: \textit{Acromyrmex echinatior} (official gene set, OGS, 3.8) [S8], \textit{Apis mellifera} (OGS 3.2) [S9, S10], \textit{Camponotus floridanus} (OGS 3.3) [S11], \textit{Harpegnathos saltator} (OGS 3.3) [S11], \textit{Nasonia vitripennis} (OGS 2) [S12], and \textit{Tribolium castaneum} (OGS 3.0) [S13]. The hierarchical level for clustering orthologous genes in the OrthoDB query was set to the node “Holometabola” (= Endopterygota). The six reference taxa were selected, because (i) their genomes are well sequenced and annotated, (ii) their official genes set were publicly released and freely available (in respect of the Ft. Lauderdale agreement) when we started our analyses, and (iii) they covered all those major target lineages from which genomes of species had been sequenced. We additionally demanded that the target genes have not been recorded as being duplicated (but they can be missing) in other species of Hymenoptera. In addition to the six reference species, we exploited the publicly available gene orthology information stored in OrthoDB v7 [S7] to other species when inferring a set of target genes (ortholog set). Our custom query (kindly conducted and provided by Robert Waterhouse on August 11, 2014) explicitly demanded that single-copy genes of the above six reference taxa must not occur in multi-copy (i. e., their copy number must be <= 1) in any of the following additional species: \textit{Atta cephalotes} (OGS 1.2), \textit{Apis florea} (OGS 1.1), \textit{Bombus impatiens} (OGS 1.2), \textit{Bombus terrestris} (OGS 1.3), \textit{Cephus cinctus} (OGS from April 2013), \textit{Dufourea novaculae} (OGS 1.1), \textit{Eufriesea mexicana} (OGS 1.1), \textit{Habropoda labiata} (OGS 1.2), \textit{Lasiochlorus albipes} (OGS 5.4), \textit{Linepithema humile} (OGS 1.2), \textit{Melipona quadridascia} (OGS 1.1), \textit{Megachile rotundata} (OGS 1.1), \textit{Pogononyrmex barbatus} (OGS 1.2),
and Solenopsis invicta (OGS 2.2.3). Note that we intentionally did not restrict the copy number to 1 in order to cope with the problem that official gene sets of poorly sequenced and annotated genomes typically comprise a relatively low number of genes. We acknowledge that assembly and annotation artefacts can also result in artificially inflated copy numbers. We think, however, that our query represents a reasonable compromise between being able to search for as many promising target genes as possible in the 1KITE transcriptomes (even if these genes actually got secondarily lost in some lineages) and making sure that genes that indeed occur in multiple copies are omitted.

Orthograph requires the user to provide a tabulator-delimited file containing the information which genes in the reference species are orthologous. We obtained this information from Robert Waterhouse, who extracted it from OrthoDB v7 [S7]. Orthograph also requires the official gene sets (i.e., the predicted coding and the predicted amino acid sequences) to the sequenced genomes of these reference species if one intends to conduct analyses also on the nucleotide level. These were obtained from the sources specified in Data S1D.

Since the OrthoDB v7 file that stores information about which genes are orthologous in the six reference taxa also included information about all other taxa that were part of the OrthoDB query, we first removed with a custom Perl script from the tabulator-delimited OrthoDB v7 output file all entries that did not refer to the above six reference species. After this filtering, the file contained 3,275 entries per species, thus 19,650 entries in total. Given that the entries in the gene identifier column of this file need to perfectly correspond with the heading information of the corresponding genes in the official gene sets, we reformatted the headers in the OGS files with custom Perl scripts. In this context, we discovered that the OGS 2.0 of N. vitripennis is partially corrupt. Specifically, we found multiple entries for the following genes and proteins (the number of copies is given in parenthesis):

Nasvi2EG037307t1 (2), Nasvi2EG037281t1 (4), Nasvi2EG037267t1 (3), Nasvi2EG037334t1 (2), Nasvi2EG037012t1 (2), Nasvi2EG037232t1 (2), Nasvi2EG037311t1 (3), Nasvi2EG037344t1 (2), Nasvi2EG037308t1 (3), Nasvi2EG037337t1 (4), Nasvi2EG037324t1 (2), Nasvi2EG037295t1 (2), Nasvi2EG037340t1 (2), Nasvi2EG037310t1 (3), Nasvi2EG037309t1 (2), Nasvi2EG037306t1 (3), Nasvi2EG037352t1 (2), Nasvi2EG037352t1 (3), Nasvi2EG037335t1 (3), Nasvi2EG037349t1 (2), Nasvi2EG037338t1 (2), Nasvi2EG037315t1 (3), Nasvi2EG037318t1 (2), Nasvi2EG037318t1 (2), Nasvi2EG03720t1 (5), Nasvi2EG037351t1 (2), Nasvi2EG037317t1 (2), Nasvi2EG037342t1 (3), Nasvi2EG037322t1 (4), Nasvi2EG037294t1 (3), Nasvi2EG037343t1 (5), Nasvi2EG037339t1 (2), Nasvi2EG037231t1 (3), Nasvi2EG037313t1 (2), Nasvi2EG026831t1 (2), Nasvi2EG037316t1 (4), Nasvi2EG037325t1 (2), Nasvi2EG037230t1 (3), Nasvi2EG037326t1 (2).

Since it remained unclear why different genes in the Nasonia OGS 2.0 had the same identifier (Don Gilbert, personal communication; August 11, 2014), we removed these genes from the official gene set. A total of 15 of them (EOG7JTJMP, EOG783ZJ4, EOG77W4C2X, EOG78Q60C, EOG74RCGQ, EOG7B3CRR, EOG7W1GSS, EOG7B0H4C, EOG700M13, EOG7K459C, EOG77MBBJ, EOG72GBX3, EOG7DG821, EOG78Q602, EOG764JRC) referred to genes in our OrthoDB v7 output files. We conservatively removed the corresponding entries to all six reference taxa also included information about all other taxa that were part of the OrthoDB query, we first removed with a custom Perl script from the tabulator-delimited OrthoDB v7 output file all entries that did not refer to the above six reference species. After this filtering, the file contained 3,260 entries per species, 19,560 entries in total. Given that the entries in the gene identifier column of this file need to perfectly correspond with the header information of the corresponding genes in the official gene sets, we reformatted the headers in the OGS files with custom Perl scripts. In this context, we discovered that the OGS 2.0 of N. vitripennis is partially corrupt. Specifically, we found multiple entries for the following genes and proteins (the number of copies is given in parenthesis):
to the globally best matching ortholog group (OG). For this purpose, it creates a profile hidden Markov model (pHMM) from the amino acid sequences of each target gene (i.e., ortholog group (OG), containing the orthologous amino acid sequences of the six reference species). This pHMM is used to search (at the translational level) for ortholog candidates in transcript libraries. Each candidate transcript sequence is extracted and used as query in a protein BLAST search against a database of all amino acid sequences from all the reference OGS. The results from both the pHMM search and the protein BLAST search are stored in a database for evaluation and orthology delineation. The orthology delineation procedure pulls all pHMM search results from the database and sorts them by descending bit score. For each pHMM hit transcript, the corresponding BLAST result is checked whether the best hit sequence belongs to the OG that the pHMM is based on. If this is the case, the BRH criterion is fulfilled and the OG is extended with the candidate transcript. Subsequently, Exonerate [S14] is employed to infer the open reading frame (ORF) in the transcript sequence by calculating a pairwise alignment with a reference amino acid sequence (i.e., the amino acid sequence of the reference species that was retrieved in the reciprocal search as best hit; see above). This step also produces corresponding sequences on the transcriptional level and corrects frame shift errors. If possible, the ORF is extended beyond the pHMM alignment region, retaining the orthologous region.

We used Orthograph (version beta4) [S3] to search for the 3,260 target genes. Orthograph depends on the following bioinformatics software packages, of which we used the hereafter specified version numbers: BLAST+ (version 2.2.26) [S6], Exonerate (version 2.2) [S14], and MAFFT (version 7.123) [S15]. All potential orthologous transcripts found by a given pHMM were sorted by their degree of similarity to the pHMM and then reciprocally searched against the proteins of the official gene sets (from six different species; Data S1D combined). We applied a non-strict reciprocal search. Thus, the BRH criterion was fulfilled if the first reciprocal hit was a protein that was also part of the pHMM, irrespective of the species to which the protein belonged to. If the BRH criterion could not be established for a given transcript, all transcripts with a smaller degree of similarity to a given pHMM were discarded (soft-threshold = 0). The minimum length of transcripts (on the translational level) was set to 30 amino acids. We allowed frame shift-corrected transcriptional extension of each transcript beyond the region of the transcript for which the BRH criterion was established as long as this region was not longer than the region, for which the BRH was established (see above). All other Orthograph parameters were left at the default values. When summarizing the results, we removed terminal and masked internal stop codons with X and NNN, respectively, in all amino acid and corresponding nucleotide sequences.

Search for transcripts orthologous to the 3,260 target genes in all 168 analyzed transcript libraries revealed on average transcripts to 2,403 target genes (median: 2,450; minimum: 1,456; maximum 2,760; respectively, in all amino acid and corresponding nucleotide sequences. The remaining 518 sequences, referring to a total of 230 OGs, remained outliers (1,540), Eucera plumigera (1,456), Ampulex compressa (1,540), Ampulex fasciata (1,542), Eucera syriaca (1,560), Tetraloniella sp. (1,662), and Acantholyda hieroglyphica (1,668). Overall, 3,256 of the 3,260 target genes were represented in the 168 transcriptomes.

6. Inference of multiple sequence alignments
We constructed multiple sequence alignments (MSA) with MAFFT (version 7.017) [S15], using the L-INS-i algorithm. We subsequently checked the resulting MSAs on the translational level and, if necessary, refined the MSAs as outlined by Misof et al. [S2]. We inferred the MSA of each OG also on the transcriptional (nucleotide) level using a modified version of the software Pal2Nal [S16] (version 14; see Misof et al. [S2] for details on the software modification). All MSAs have been deposited in and are available from Mendeley Data (doi: 10.17632/trbj94zm2n.2).

Assessing the quality of the inferred Multiple Sequence Alignments (MSAs) revealed putatively misaligned amino acid transcripts (hereafter referred to as outliers) in 256 of the 3,256 investigated ortholog groups (OGs). Attempts to refine the alignment of the 619 outlier sequences succeeded in 101 instances (i.e., sequences). The remaining 518 sequences, referring to a total of 230 OGs, remained outliers and were consequently removed from the MSAs. We also removed the corresponding nucleotide sequences from the respective nucleotide MSAs.
7. Protein domain identification
To facilitate a protein domain-based sequence data partitioning scheme, we applied the procedure described by Misof et al. [S2]. Protein domains and protein domain clans (i.e., evolutionary related protein domains) [S17] were identified (on the amino acid level) in each MSA by using pHMMs from the Pfam database (release 27.0) [S18] and the software pfam_scan.pl (version 1.5, release date: October 15, 2013) [S18] in conjunction with HMMER (version 3.1b2) [S19]. For additional details, see Misof et al. [S2].

Search for protein domains in the refined amino acid sequence alignments of the 3,256 single-copy genes assigned 32.0% of the alignment sites to domains of Pfam-A and 6.0% of the alignment sites to domains of Pfam-B. A total of 62.0% of the alignment sites remained unannotated. Based on the domain identification results, we split the 3,256 multiple sequence alignments and rearranged their sites into 5,922 different data blocks. Of these, 312 belong to Pfam-A domains and were pooled according to their clan annotation [S17]. After pooling domains without clan annotation but having the same name, we obtained 1,243 data blocks belonging to different Pfam-A domains and 1,111 belonging to different Pfam-B domains. Finally, we pooled unannotated regions (voids) according to their gene origin. This resulted in 3,256 data blocks corresponding to the void regions of the 3,256 analyzed genes.

8. Multiple sequence alignment masking
We used a modified version of Aliscore (version 1.2) [S20–S22] to identify blocks of putative alignment ambiguities or randomized MSA sections in the MSA of each OG. Aliscore was executed with its default sliding window size, with requesting the maximum number of pairwise sequence comparisons, and with the option –e (i.e., indicating gappy amino acid sequence data). Blocks of putative alignment ambiguities or randomized MSA sections were subsequently removed from each domain data block at the amino acid and corresponding nucleotide level using custom scripts and concatenated to supermatrices containing protein domain- and gene-based data blocks (the latter comprising regions not annotated as protein domain). Gap symbols (-) at the beginning and at the end of the resulting MSAs were replaced with 'X' (translational level) and 'N' (transcriptional level), respectively.

9. Removal of phylogenetically non-informative data blocks
We assessed the phylogenetic signal of each data block at the translational level with the software MARE (version 0.1.2-rc) [S23]. Data blocks whose phylogenetic information was zero were removed from the supermatrix on the translational and on the transcriptional level using custom Perl scripts. Note that the phylogenetic information content of all retained data blocks also differed from zero when being assessed on the transcriptional level.

After (i) removing ambiguous alignment sites in each data block resulting from steps outlined above, (ii) deleting data blocks that contained no phylogenetic information (792 in total), and (iii) concatenating all remaining data blocks, the resulting supermatrices consisted of 1,505,722 amino acid and 3,011,444 nucleotide sites (1st and 2nd codon position only), respectively. Each of the two supermatrices contained information from 3,256 different single-copy protein-coding genes and comprised 4,818 data blocks (301 clan data blocks from pooled Pfam-A protein domains with clan association, 1,103 data blocks from pooled Pfam-B protein domains without clan association, 778 data blocks from Pfam-B protein domains, 2,636 data blocks from void regions).

The amino acid supermatrix was further processed for testing specific phylogenetic hypotheses via analysis of decisive datasets (i.e., when demanding the presence of at least one species from a given taxonomic group in each data block of the supermatrix). The statistics of the three inferred datasets were as follows: ‘backbone dataset’ (800,393 amino acid sites, comprising 2,142 data blocks), ‘Aculeata dataset’ (709,055 amino acid sites, comprising 1,712 data blocks), ‘Apoidea dataset’ (527,548 amino acid sites, comprising 841 data blocks).

All above specified supermatrices as well as the data block information files have been deposited at Mendeley Data (doi: 10.17632/trbj94zm2n.2).

10. Partitioning and substitution model selection
We used PartitionFinder (developer versions 2.0.0-pre2, 2.0.0-pre9, and 2.0.0-pre10) [S24] to identify combinations of data blocks that can be modeled with the same substitution model and which therefore might better be combined to partitions. The corrected Akaike information criterion (AICc) [S25] was used to assess whether or not to combine data blocks to partitions. PartitionFinder was run with the following
settings: branchlengths [linked], models [LG+G, LG+G+F, WAG+G, WAG+G+F, BLOSUM62+G, BLOSUM62+G+F, DCMUT+G, DCMUT+G+F, JTT+G, JTT+G+F, LG4X], model_selection [aicc], search [rcluster], weights [rate = 1.0, base = 1.0, model = 0.0, alpha = 1.0], rcluster-percent [100.0], rcluster-max [10,000]. PartitionFinder was started with the following additional command line arguments: ‘-narxml -all-states -min-subset-size 50’. We applied the same partition scheme that we inferred with PartitionFinder from analyzing protein domain-based data blocks at the translational (amino acid) level when analyzing the data at the transcriptional (nucleotide) level. However, we discarded the 3rd codon position, as it was saturated and typically exhibits lineage specific compositional biases [S26, S27]. We furthermore modeled the nucleotide substitutions of each codon position within a given partition separately. The applied nucleotide substitution model was GTR+G.

PartitionFinder suggested merging data blocks in each of the supermatrices outlined above to partitions. Specifically, PartitionFinder suggested analyzing the primary supermatrix (i.e., the non-decisive dataset) at the amino acid level by considering 2,058 partitions. We applied the same partition scheme when analyzing the corresponding supermatrix containing nucleotide sequences, except that we modeled the 1st and 2nd codon position of each partition separately (the 3rd codon position was not considered in our analyses). The resulting primary supermatrix comprised accordingly on the nucleotide level 4,116 partitions. PartitionFinder suggested considering the following number of partitions for testing specific phylogenetic hypotheses via analyses of decisive datasets: 1,025 (‘backbone dataset’), 853 (‘Aculeata dataset’), and 461 (‘Apoidea dataset’). The final data sets and files containing the inferred partition schemes and model selection results are deposited at Mendeley Data (doi: 10.17632/trbj94zm2n.2).

11. Phylogenetic tree inference and bootstrap analysis

We applied the maximum likelihood optimality criterion as implemented in ExaML (versions 3.0.15 and 3.0.17) [S28] for the phylogenetic inferences. We selected the tree with the best log-likelihood score found in 50 independent tree searches (25 randomized stepwise addition parsimony starting trees, 25 completely random starting trees) per dataset. All starting trees were inferred with RAxML (version 8.0.26) [S29] We applied the GTR substitution model when analyzing the nucleotide sequence data and applied the partition-specific substitution models suggested by PartitionFinder when analyzing the amino acid sequences in ExaML. Rate heterogeneity was approximated with a gamma distribution, using the median and four discrete rate categories. Node support was estimated via non-parametric bootstrapping [S30]. Bootstrap replicate alignments and random start trees were generated with RAxML and a custom shell script. Subsequently, ExaML was used to infer one ML tree per bootstrap replicate, applying the original partitioning scheme suggested by PartitionFinder for the respective supermatrix. In order to determine the minimum number of replicates needed for reliable estimation of bootstrap support values, we used the “autoMRE” bootstrap convergence criterion [S31], as implemented in RAxML, with the default threshold setting of 0.03. In brief, this method works as follows: first, all replicate trees are randomly divided into two subsets of equal size; then, a majority rule extended (MRE) consensus tree for each subset is built; finally, the relative weighted Robinson-Foulds distance (WRF) between the consensus trees is computed. Convergence is reached if WRF is below the threshold in at least 99% of 1,000 tree set permutations. Convergence was evaluated after analyzing every batch of 50 bootstrap replicates. All inferred trees were rooted at the branch that connects Hymenoptera with the outgroups.

The majority rule extended (MRE) bootstrap convergence criterion [S31] indicated that 50 bootstrap replicates are sufficient to accurately assess node support irrespective of whether we analyzed the primary supermatrix containing amino acid sequences or the primary supermatrix containing the corresponding nucleotide sequences. All node support values provided in the present investigation are consequently based on 50 non-parametric bootstrap replicates each.

Please note that RAxML and ExaML remove completely undetermined sites (i.e., sites containing only “-” and/or “X” and/or “N”) prior to the phylogenetic inference. The total size of the supermatrices in each of the phylogenetic inferences was accordingly 1,505,514 sites (primary supermatrix, amino acids), 3,011,099 sites (primary supermatrix, nucleotides), 800,252 sites (“backbone dataset”, amino acids), 708,924 sites (“Aculeata dataset”, amino acids), and 527,471 sites (“Apoidea dataset”, amino acids).

We used ExaBayes [S32] to conduct phylogenetic inferences also in Bayesian framework and analyzed again the primary amino acids supermatrix (see above). We completed three independent ExaBayes runs with four coupled MCMC chains each. After 200,000 generations, we assessed convergence of the results by computing the average deviation of split frequencies (ASDSF criterion). We obtained an ASDSF value of 2.51%, which is generally considered an acceptable convergence according to the ExaBayes manual.
We tested for the presence of rogue taxa with RogueNaRok (version 1.0) [S33] when analyzing the amino acid and the nucleotide supermatrix using five distinct settings: (i) bipartition support drawn onto the best ML tree (referred to as best), (ii) majority rule consensus (50% threshold, referred to as mr), (iii) greedily refined extended majority rule consensus (referred to as mre), (iv) strict consensus (100% threshold, referred to as strict), (v) the 75% threshold consensus – the criterion for pruning rogue taxa is to improve the number of edges that have at least 75% bootstrap support (referred to as t75). We used a dropset size of 1 when using the settings best and mre, a dropset size of 1+2 when using the setting mr and a dropset size of 1+4 when using the settings strict and t75.

We identified a single species that exhibited rogue taxon behavior: Nitela sp., a representative of the polyphyletic apoid wasp family “Crabronidae”. However, Nitela sp. was identified as a rogue taxon only when analyzing the amino acid primary supermatrix and only under one specific rogue taxon identification setting: 75% threshold consensus with dropset sizes 1+4. Given the fact that Nitela is deeply nested in a subordinated monophyletic lineage of apoid wasps and that its position within this lineage was not of primary interest in context of the present investigation, we refrained from excluding Nitela from our phylogenetic analyses. The results from the rogue taxon inference and bootstrap analysis have been deposited at Mendeley Data (doi: 10.17632/trbj94zm2n.2).

13. Testing specific phylogenetic hypotheses via analysis of decisive datasets

Following Dell’Ampio et al. [S34] and Misof et al. [S2], we generated decisive datasets, which contained at least one species from a given taxonomic group in each partition of the analyzed supermatrices, to assess whether a bias due to a possible lack of taxonomic sequence representation at the partition level has an impact on the inferred tree topology. We generated three different decisive datasets: (i) one for addressing the major lineages along the Hymenoptera tree (‘backbone dataset’), (ii) one for addressing the phylogenetic relationships of the major lineages of Vespoidea and Apoidea (‘Aculeata dataset’), and (iii) one for addressing the phylogenetic relationships within Apoidea specifically (‘Apoidea dataset’). Specific information on which groups we distinguished and which species are part of a given group has been deposited at Mendeley Data (doi: 10.17632/s5j2f62z3d.2). Given that the generation of decisive datasets resulted in some partitions being excluded from subsequent analyses, we inferred new partition schemes and corresponding partition-specific substitution models with PartitionFinder for each decisive dataset. Please note that in the phylogenetic analyses of the decisive datasets we considered only amino acid sequences. We conducted ten tree searches with ExaML (versions 3.0.15 and 3.0.17) with random starting trees per decisive dataset.

Phylogenetic analysis of all three decisive datasets (i.e., ‘backbone dataset’, ‘Aculeata dataset’, ‘Apoidea dataset’) resulted in topologies identical to the one inferred from analyzing the primary supermatrix (i.e., the non-decisive dataset) at the amino acid level when considering only those major lineages whose phylogenetic relationships to each other were meant to be assessed by the respective
decisive dataset. Since the decisive data sets are restricted to partitions with a required taxon coverage, the results indicate that the distribution of missing data did not produce a significant phylogenetic signal, rendering the probability of a missing data bias in our phylogenetic relationships unlikely. The results from analyzing the decisive datasets have been deposited at Mendeley Data (doi: 10.17632/s5j2f62z3d.2).

14. Four-cluster Likelihood Quartet Mapping (FclM)

We applied FclM to assess the phylogenetic support for conflicting hypotheses [S2, S34, S35, S36], focusing on five major phylogenetic hypotheses (see below). We used the PTHREADS implementation of RAxML (versions 8.2.6 and 8.2.8) [S29] to infer the likelihood scores of quartets, using the partition scheme and substitution models inferred before when analyzing the complete supermatrix at the translational (amino acid) level. We assessed the following phylogenetic hypotheses: (i) Eusymphyta represent the sister group of Unicalcarida. This hypothesis was assessed by testing the phylogenetic position of Xyeloidea (iA) relative to the outgroup, to Pamphilioidea+Tenthredinoidea and to Unicalcarida, (iB) relative to the outgroup, to Tenthredinoidea and to Pamphilioidea+Unicalcarida, and (iC) relative to the outgroup, to Pamphilioidea and to Tenthredinoidea+Unicalcarida. Please note that we did not consider representatives of Apocrita in the three tests. (ii) Cephoidea represent the sister group of Vespina; (iii) Trigonaloidea represent the sister group of Auleata; (iv) Stephanoidea represent the sister group of Evanioidea; (v) Formicidae represent the sister group of Apoidea. Specific information about which species were part of a specific group is given in Data S1G.

The results from testing major phylogenetic hypotheses (i.e., Eusymphyta as the sister group of Unicalcarida, Cephoidea are the sister group of Vespina, Trigonaloidea are the sister group of Auleata, Stephanoidea are the sister group of Evanioidea, Formicidae are the sister group of Apoidea) via Four-cluster Likelihood Mapping (FclM) as well as all analyzed amino acid data matrices have been deposited along with partition information files at Mendeley Data (doi: 10.17632/s5j2f62z3d.2 and 10.17632/trbj94zm2n.2).

Given that we found in all tree inferences strong support for monophyletic Eusymphyta, a result incompatible with the classical hypothesis of Xyeloidea being the sister group of all remaining Hymenoptera (supported by morphological data and data on muscle cell ploidy levels) [S37, S38], we additionally assessed the support for our and for alternative phylogenetic hypotheses regarding the phylogenetic position of Xyeloidea via FclM. FclM of the original amino acid sequence data consistently revealed the lowest phylogenetic signal (0–3%) for a sister group relationship of Xyeloidea and the remaining Hymenoptera, irrespective of the phylogenetic position of Pamphilioidea (please note that FclM does not assume any specific phylogenetic relationships among the species of a given group, e.g., the outgroup). While it cannot be excluded that consideration of additional outgroup lineages (e.g., non-holometabolous insects) could result in a change of the topology, doing so would require design of a very different ortholog set than the one used in the present investigation, comprising significantly fewer genes and consequently likely containing substantially less phylogenetic signal.

The results from FclM of the original amino acid sequence data revealed no phylogenetic signal for topologies that are incompatible with the hypothesis of Cephoidea being the sister group of Vespina and with the hypothesis of Trigonaloidea being the sister group of Auleata. However, FclM revealed phylogenetic signal in the dataset that is in conflict with the other two hypotheses. Specifically, FclM indicated besides strong support for a sister group relationship of Formicidae and Apoidea (66%) also noteworthy support a sister group relationship of Scoliidae and Apoidea (32%). This signal has the potential of misleading phylogenetic analyses, in particular when considering only a small taxonomic sampling. Authors should keep this in mind when inferring a sister group relationship of scoliid and apoid wasps in future analyses. A sister group relationship of Scoliidae and Formicidae, as it was inferred in the Bayesian analyses, was supported only by 2% of the quartets. Considering all evidence, a sister group relationship of Formicidae and Apoidea appears currently the most likely phylogenetic scenario. FclM also indicated strong support for a sister group relationship of Stephanoidea and Trigonaloidea (88%) and for a sister group relationship of Stephanoidea and Parasitoida (11%). In fact, the phylogenetic hypothesis suggested by the simultaneous phylogenetic analysis of all taxa (Figure 1; i.e., a sister group relationship of Stephanoidea and Evanioidea) is only supported by 1% of the evaluated quartets. This sister group relationship also received only a comparatively low bootstrap support (i.e., 90%), considering the size of the analyzed data matrix (Figure 1) and it was not found in the Bayesian analyses either (Fig. S4).

We consequently consider the question of the phylogenetic position of Stephanoidea relative to Parasitoida, Evanioidea, and Trigonaloidea as still open.
15. Testing for non-phylogenetic signal biases via permutation tests

To assess the possible impact of a heterogeneous composition of amino acid sequences, of a non-stationarity of substitution processes and of a non-random distribution of missing data [S39, S40] on our phylogenetic inferences, we adopted the permutation test strategy suggested by Misof et al. [S2]. For more information on the applied permutation schemes and their explanatory power, see Misof et al. [S2]. The phylogenetic signal for the conflicting hypotheses in the permuted supermatrices was assessed using FcLM (see above). FcLM was conducted on the translational level and using the same partition scheme as before, but using the WAG substitution model across all partitions. Permutation tests were done with RAxML (versions 8.2.6 and 8.2.8) [S29] and ExaML (version 3.0.17) [S28].

The results from testing major phylogenetic hypotheses (i. e., Eusymphyta are the sister group of Unicalcarida, Cephoidea are the sister group of Vespina, Trigonalioidea are the sister group of Aculeata, Stephanioidea are the sister group of Evanioidae, Formicidae are the sister group of Apoidea) via permutation tests in conjunction with FcLM have been deposited at Mendeley Data (doi: 10.17632/s5j2f62z3d.2). In none of the permutation tests did any of the three possible topologies receive more than 18% support when assessing the phylogenetic support for different topology by the permuted dataset via FcLM. The only exception were the results from testing the position of Xyeloidea and in which some topologies received up to 41% support. However, artificial phylogenetic signal supported almost exclusively topologies incompatible with monophyletic Eusymphyta, whereas artificial support for Eusymphyta remained low. We accordingly interpret the impact of artificial phylogenetic signal on the inferred phylogeny shown in Figure 1 as insignificant.

The results from the various permutation tests have been deposited at Mendeley Data (doi: 10.17632/s5j2f62z3d.2).

16. Divergence time estimation

We time-calibrated the inferred phylogenetic tree using information on 14 carefully selected fossils. We followed the best-practice recommendations put forth by Parham et al. [S41] for justifying fossil calibrations. We considered the following fossils to calibrate the inferred phylogeny, with the lineage each fossil was assigned to given in parentheses: Triassoxyela foveolata (Hymenoptera) [S42] Libanophron astarte (Ceraphronoidea) [S43], Rhetinorhyssalus morticinus (Braconidae) [S44], Protimaspi costalis (Cymipoidea) [S45], Archaeopelecinus jinzhouensis (Pelecinidae) [S46], an undescribed species (Chalcidoidea) deposited in the Natural History Museum of the Lebanese University (Faculty of Sciences II, Fanar, Lebanon, female, Coll. No. 874A, specimen represents an undescribed family, that can be placed in Chalcidoidea based on unambiguous synapomorphic character, i. e., presence of multipartite plate sensillae on antennae), Protoparevania lourothi (Evaniiidae) [S47], Lancepyris opertus (Chrysidioidea) [S48], Architithia rasnitsyni (Tiphidae) [S49], Paleogenia wahisi (Pepsinae) [S50], Apodolichurus sphaerocephalus (Ampulicidae) [S51], Palaeomacropis eocenicus (Macropidineae) [S52], Paleohabropoda oudardi (Anthophorini) [S53], Bombus (Bombus) randezekensis (Bombini) [S54]. Detailed information about the fossils (e. g., characters considered as autapomorphies of specific lineages, their estimate minimum geological age) is given in Data S1F. If multiple fossils of a given group were available for calibration, we selected the geologically oldest one. Please note that we considered Paleogenia wahisi (Pepsinae), deposited in Baltic Amber, despite the well-known problem of accurately estimating the geological age of Baltic Amber. However, the currently widely accepted conservative minimum age of Baltic Amber (i. e., 34 mya) [S55] proved to be informative for improving calibration of the phylogenetic tree. We used mcmctree, in conjunction with codeml (both part of the PAML software package, version 4.9) [S56] to estimate node ages. We applied the approximate likelihood method [S57] by generating a single Hessian matrix with codeml using standard parameters and the JTT model from a condensed version of our supermatrix of amino acids. Specifically, the supermatrix contained only sites at which at least 95% of the samples had a non-ambiguous amino acid symbol. The condensation of the supermatrix became necessary to overcome computational limitations when estimating node ages resulting from the large size of the dataset. Please note that we tested whether or not our data allowed for the generation and subsequent concatenation of partition-specific Hessian matrices. This approach would in principle allow different models to be applied to different partitions. However, we found that most partitions contained at least one taxon with no amino acid sequence information, rendering this approach unfeasible. Prior to the node age analyses with fossil calibration data, we started various test runs, using an arbitrary calibration and the Hessian matrix, to assess whether our sampling parameters would likely lead to convergence. These tests
revealed that a burn-in value of 40,000 and a number of samples of 400,000 typically led to convergence. We introduced the fossil calibrations as soft minima using default settings (i.e., as truncated Cauchy distributions with an offset of 0.1, a scale parameter of 1 and a left tail probability of 0.025) [S58]. The root age was specified at 412 mya in the control file, providing a very conservative estimate [S2]. We ran four dating analyses, using independent-rates clock and standard parameters, apart from the burn-in and sample size values, which we set to the values found in the above test runs and using the Hessian matrix generated from the non-partitioned dataset. All four runs provided very similar node age estimates, with two of the runs also agreeing in the boundaries of the 95% confidence intervals. The other two runs suggested slightly larger confidence limits at a small number of nodes, mostly close to the root of the tree. It is likely that in these two runs the MCMC chain took longer to converge to the final node ages and that thus a larger number of outliers was sampled. Since two of the runs agreed in their node age estimates and errors, we consider the results from these two runs to very likely represent the more realistic node ages and confidence intervals (shown in Figures 1 and S1 and specified in Data S1H). In addition, we ran the same dataset with the same settings again twice, this time using the correlated-rates clock. Finally, we used a condensed version of our supermatrix of nucleotides, analogous to the amino acid set including only sites at which at least 95% of the samples had a non-ambiguous base symbol, and ran this, with the same settings as outlined above for the amino acid set, twice with the independent and the correlated-rates clock, respectively. As model we specified HKY85. The node ages and confidence intervals are shown in Data S1H. The two replicate runs for these additional analyses converged (Figure deposited at Mendeley Data, doi: 10.17632/s5j2f62z3d.2). In general, the correlated rates runs and the runs using the nucleotide supermatrix delivered results that show older ages for the included nodes, with the effect of using correlated-rates slightly lower. However, we deem the correlated-rates clock as less suitable for the data and taxon sampling of our study, following [S59]. While correlated-rates models have been advocated for in past studies (e.g., [S60]), the basic autocorrelation method implemented in MCMCtree can result in unrealistically high or low rates, particularly if the number of taxa included in the study is high and if the root age is old (also see [S61, S62]).

Supplemental References


