Phylogenomics Resolves Evolutionary Relationships among Ants, Bees, and Wasps

Brian R. Johnson,1 Marek L. Borowiec,1 Joanna C. Chiu,1 Ernest K. Lee,2 Joel Atalah,3 and Philip S. Ward1,4
1Department of Entomology and Nematology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA
2Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

Summary

Eusocial behavior has arisen in few animal groups, most notably in the aculeate Hymenoptera, a clade comprising ants, bees, and stinging wasps [1–4]. Phylogeny is crucial to understanding the evolution of the salient features of these insects, including eusociality [5]. Yet the phylogenetic relationships among the major lineages of aculeate Hymenoptera remain contentious [6–12]. We address this problem here by generating and analyzing genomic data for a representative series of taxa. We obtain a single well-resolved and strongly supported tree, robust to multiple methods of phylogenetic inference. Apoidea (spheciform wasps and bees) and ants are sister groups, a novel finding that contradicts earlier views that ants are closer to ectoparasitoid wasps. Vespid wasps (paper wasps, yellow jackets, and relatives) are sister to all other aculeates except chrysidooids. Thus, all eusocial species of Hymenoptera are contained within two major groups, characterized by transport of larval provisions and nest construction, likely prerequisites for the evolution of eusociality. These two lineages are interpolated within two major groups, characterized by transport of larval provisions and nest construction, likely prerequisites for the evolution of eusociality. These two lineages are interpolated among three other clades of wasps whose species are predominantly ectoparasitoids on concealed hosts, the inferred ancestral condition for aculeates [2]. This phylogeny provides a new framework for exploring the evolution of nesting, feeding, and social behavior within the stinging Hymenoptera.

Results and Discussion

Aculeate (stinging) Hymenoptera are behaviorally diverse, encompassing both solitary and eusocial species and exhibiting a variety of life history strategies including parasitoidism, predation, omnivory, and pollenivory [2, 13]. Multiple lines of evidence provide strong support for the monophyly of the Aculeata [9, 10], but relationships among the major lineages within this group have been a matter of continued uncertainty [7–9, 11, 12]. The position of ants, the most species-rich and ecologically dominant of all eusocial insects, has been particularly problematic [7–9, 11, 12] (Figure 1).

Advances in next-generation sequencing have unleashed the potential of genomic data to clarify many previously intractable parts of the Tree of Life [14–16]. Here we addressed the problem of aculeate Hymenoptera phylogeny by generating transcriptome data for ten representative species in nine families and genomic data for one key taxon (Apterogyna) for which RNA was unavailable (Table 1; see also Table S1 available online). We then combined these data with the published genome sequences of three bee species and three ant species, a transcriptome from one additional bee species, and genomic data from Nasonia vitripennis, a nonaculeate hymenopteran used as an outgroup. Orthology identification and matrix assembly was accomplished with the OrthologID pipeline [17]. This yielded multiple partitioned amino acid matrices, with different levels of gene representation across the 18 ingroup taxa, ranging from a 5,214-gene matrix (3,001,657 amino acid sites) to a stringently filtered matrix of 308 genes (175,404 sites) and only 14.98% missing data (see Experimental Procedures and Supplemental Experimental Procedures).

Table 1 shows the summary statistics for all the transcriptomes and for the genomic assembly. For the transcriptomes, we identified a range of protein sequence numbers, with the scoliid wasp Cricoscelis alcione having the largest transcriptome size and the sweat bee Lasioglossum albipes the smallest. Whether variation in the transcriptome sizes represents actual variation in the number of genes present in these species or whether it represents variation in the quality of the assemblies is uncertain. The genomic library of Apterogyna AZ01 (family Brachynobaenidae) had only 1,717 genes, with a median amino acid length of 450. This reflects the relatively low coverage of sequencing (18×) and the apparent degradation of this older sample. Nevertheless, the Apterogyna sequence data were sufficient to reliably place this taxon within the hymenopteran tree.

Phylogenetic analyses produced a fully resolved tree of the aculeate Hymenoptera with robust support at all nodes (Figure 2). The same tree topology and relative branch lengths were obtained under a variety of analytical procedures, including partitioned maximum likelihood (ML) analyses and Bayesian analyses of concatenated data sets, as well as species tree estimates (Figures 2 and S1). All nodes in the topology have ML bootstrap support of 100% and Bayesian posterior probabilities of 1.0. Under species tree analyses, most nodes are also strongly supported, although support values drop for some of the deeper nodes in the tree (Figures 2 and S1). Most procedures employed the 308-gene matrix, but we also ran ML analyses of three other matrices of varying size and completeness (525, 3,018, and 5,214 genes, respectively), with the same results (Figures S1A–S1C).

As expected [6] we found that the cuckoo wasp (Argochrysis) is sister to all other aculeates, and that ants, bees, apoids, and vespid wasps are all monophyletic. We recovered the vespid wasps (represented by a nonsocial pollen wasp, Pseudomasaris, and a eusocial paper wasp, Mischocyttarus) as sister to all aculeates except the cuckoo wasp, a result that is in agreement with some other recent molecular studies [8, 11] although in strong conflict with morphology-based trees [7] in which vespids are nested well within the aculeate phylogeny, as sister to scoliid wasps (Figure 1A).

Of particular interest is the finding that ants are sister to Apoidea, a novel result that emphasizes a greater affinity of ants to the predatory wasps that characterize the earliest branching lineages of Apoidea than to scoliids, brachynobaenids, tiphiids, and other ectoparasitoid wasps with which they have been associated previously [1, 7, 9, 12]. This result
different branching order, such that the scolioid lineage was identified as sister to Apoidea. Our new results motivate a search for features in common between ants and apoids—not shared with scolioid wasps—that might predispose this group toward the evolution of sociality. The most obvious behavioral commonality is the collection and transport of resources (arthropod prey or pollen) to a constructed nest, a trait also shared with vespid wasps, the other group containing eusocial species. Scolioid wasps, in contrast, are ectoparasitoids on concealed hosts [2, 27]. (The life history of Bradynobaenidae remains to be elucidated.) It has long been argued that nest construction and provisioning are key prerequisites for the evolution of eusociality [1, 2]. The finding that ants and apoids are sister taxa suggests that this favorable combination of traits arose only once in their common ancestor, rather than separately from ectoparasitoid predecessors in the ant and apoid lineages, emphasizing that the preconditions for eusociality are rare and contingent.

Our phylogeny reveals a well-supported clade of tiphioid and pompiloid wasps, to the exclusion of the scolioids. This comports with an earlier molecular study [8], except that we find that the tiphiid wasp (*Brachycistis*) and the chyphotine wasp (*Chyphotes*) are sister taxa, to the exclusion of pompiloids (spider wasps and velvet ants), a result inconsistent with previous findings. Further taxon sampling is needed within this clade to test the monophyly and placement of tiphioid wasps (Tiphidae) and velvet ants (Mutillidae), but our results confirm an earlier inference [8, 11] that the family Bradynobaenidae is not monophyletic, with true bradynobaenids (represented in our data set by *Apterogyna*) being sister to Scoliidae, whereas the subfamily Chyphotineae (represented here by *Chyphotes*) is part of the tiphioid complex. Bradynobaenid-like wasps share a number of morphological features [7], some unique, and these must be interpreted as examples of convergence between two distantly related clades, perhaps generated in part by the independent loss of wings in females of both groups. It should be noted that scoliid females are winged, as are some members of the tiphioid-pompiloid clade, so winged females are the ancestral condition for both clades.

The phylogenetic results presented here support the following scenario of behavioral evolution in aculeate Hymenoptera (Figure 3). The ancestral aculeate wasp was an ectoparasitoid, attacking and paralyzing concealed hosts and leaving its offspring in or near the host cavity [2]. In two major lineages (ants + Apoidea, and Vespidae), this behavior became modified as wasps adopted a more active predatory lifestyle, with increased importance of prey transport, nest construction, and parental care. More specialized feeding habits (pollenivory) were acquired later. Eusocial behavior evolved multiple times within both of these lineages [4, 28, 29]. The three remaining clades of aculeates (chrysidoids, scolioids, and the tiphioid-pompiloid clade) have largely retained ectoparasitoid habits, except for pompilids [30], and no examples of eusociality are known in these groups.

This is the first comprehensive phylogenomic analysis of aculeate Hymenoptera. It demonstrates the utility and feasibility of employing transcriptome data to resolve outstanding problems in insect phylogeny. The new tree provides a robust framework for investigating the evolution of nesting, feeding, and social behavior within the stinging Hymenoptera, and for exploring genomic signatures of changes in these characteristics.

![Figure 1. Previous Hypotheses of Phylogenetic Relationships among Ants, Bees, and Stinging Wasps](image-url)
Lasioglossum albipes
Pseudomasaris vespoides
Mischocyttarus flavitarsis
Crioscolia alcione

downloaded from the NCBI Sequence Read Archive (SRR578269). MB, megabases; aa, amino acids. See also Tables S1 and S2.

sequence data generated in this study, we also assembled a transcriptome for the sweat bee Apterogyna The assembly for Brachycistis timberlakei Sceliphron caementarium Pepsis grossa Argochrysis armilla Sphaeropthalma orestes (Muttillidae) scoliid wasp 150.3 27,549 703 Crisocolia alcione (Scoliidae) scoloid wasp 219.2 45,155 689 Sceliphron caementarium (Scoliidae) mud dauber wasp 161.3 33,499 659 Brachycistis timberlakei (Tiphidae) tiphid wasp 84.3 15,036 627 Mischocyttarus flavitarsis (Vespidae) paper wasp 76.8 15,907 676 Pseudomasaris vespoides (Vespidae) pollen wasp 103.7 21,543 678 Lasiosglossum albipes (Halictidae) sweat bee 46.7 9,417 585

The assembly for Apterogyna is a partial genome assembly; all others are transcriptome assemblies. In addition to the assemblies derived from new sequence data generated in this study, we also assembled a transcriptome for the sweat bee Lasiosglossum albipes based on raw paired-end short reads downloaded from the NCBI Sequence Read Archive (SRR578269). MB, megabases; aa, amino acids. See also Tables S1 and S2.

Experimental Procedures

Taxon Sampling

Eleven species from key families across the aculeate Hymenoptera were chosen for the generation of new phylogenomic data (Tables 1 and S1). These represent all the major lineages of stinging Hymenoptera that have been considered in previous hypotheses of phylogenetic relationships [6–12]. Ten species were collected in the field, while one rare species, Apterogyna ZA01, was available only as preserved specimens in ethanol. We included this second representative of Bradynobaenidae because of the instability in the phylogenetic position of bradynobaenid wasps in previous studies [7, 8, 11, 12]. We supplemented our data on these eleven species with transcriptome data on one bee species (Lasiosglossum albipes) from the NCBI Sequence Read Archive and published genome assemblies of three other bee species (Megachile rotundata, Bombus terrestris, and Apis mellifera) and three ant species (Harpegnathos saltator, Linepithema humile, and Pogonomymex barbatus).

Sequencing and Assembly

For the fresh collected samples, cDNA libraries were prepared, while for the ethanol-preserved sample, a DNA library was prepared (further details in Supplemental Information, including Table S2). Samples were pooled and sequenced on an Illumina HiSeq 2000 (100 bp paired-end). Transcriptomes were assembled using the Trinity software package [31], while ABySS was used for the genome assembly [32]. After translation of contigs into amino acid sequences, orthology was evaluated using a prerrelease version 2.0 of the OrthologID pipeline [17]. OrthologID takes complete gene sets from all taxa and assigns them into gene clusters. It then generates a parsimony tree for each gene cluster and extracts one or more sets of orthologous genes. Orthologous sets of genes were then assembled into multiple partitioned matrices with different levels of taxon representation per gene, including (1) a 5,214-gene matrix with 3,001,657 amino acid sites and at least 9 ingroup taxa represented per gene partition, (2) a 3,018-gene matrix with 1,653,740 sites and at least 16 ingroup taxa represented per partition, and (3) a matrix that has every gene partition represented across all ingroup taxa with the additional requirement that they be single-copy in five publicly available ingroup genomes (525 partitions and 298,968 sites). For Bayesian inference and species tree estimation, a fourth, smaller 308-gene matrix with 175,404 amino acid sites was used. Of the 3,332,676 cells (19 taxa × 175,404 sites) in this 308-gene matrix, 73.42% are coded as amino acids, 11.60% are gaps, and 14.98% are missing.

Phylogenetic Analysis

For all four matrices, we performed partitioned (by gene) ML analyses with I-distributed rate heterogeneity over sites using RAxML v7.4.2 [18, 19]. The best protein substitution model for each gene partition was selected individually using the “ProteinModelSelection.pl” script [19] over 36 different models. For Bayesian inference, we used PhyloBayes MPI v1.3b [20, 21].

Figure 2. Maximum-Likelihood Tree of Aculeate Hymenoptera Derived from a 308-Gene Matrix

This tree was estimated with RAxML [18, 19] based on a 308-gene matrix with 175,404 amino acid sites. Support values are given in the following order: (1) posterior probabilities from a separate Bayesian analysis with PhyloBayes [20, 21], (2) RAxML bootstrap percentages based on 1,000 replicates, and (3) bootstrap percentages from a separate species tree analysis with STAR [22]. Unlabeled nodes have maximum support values (1/100/100). Scale bar indicates number of substitutions per site. From the top of the tree downward, species are as follows: N. vitripennis (Pteromalidae), A. armilla (Chrysididae), P. vespoides (Vespidae), M. flavitarsus (Vespidae), B. terrestris (Tiphidae), B. timberlakei (Tiphidae), C. mellipes (Bradynobaenidae), P. grossa (Pompilidae), S. orestes (Muttillidae), Apterogyna ZA01 (Bradynobaenidae), C. alcione (Scoliidae), P. barbatis (Formicidae), L. humile (Formicidae), S. oregonense (Formicidae), H. sattator (Formicidae), S. caementarium (Scoliidae), L. albipes (Halictidae), M. rotundata (Megalachidae), A. mellifera (Apidae), and B. terrestris (Apidae). Major lineages are color coded using the same scheme as in Figure 1. See also Figure S1.

Table 1. Species of Hymenoptera Sampled, and Summary Statistics for the Transcriptome Assemblies and the Genome Assembly

<table>
<thead>
<tr>
<th>Species</th>
<th>Common Name</th>
<th>Assembly Size (MB)</th>
<th>Number of Sequences (&gt;300 aa)</th>
<th>Mean Protein Length (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apterogyna ZA01 (Bradynobaenidae)</td>
<td>bradynobenid wasp</td>
<td>284.2</td>
<td>1,717</td>
<td>450</td>
</tr>
<tr>
<td>Chryphotes mellites (Bradynobaenidae)</td>
<td>bradynobenid wasp</td>
<td>137.8</td>
<td>26,513</td>
<td>698</td>
</tr>
<tr>
<td>Argochrysis armilla (Chrysididae)</td>
<td>cuckoo wasp</td>
<td>91.6</td>
<td>18,777</td>
<td>690</td>
</tr>
<tr>
<td>Stigmatomma oregeonense (Formicidae)</td>
<td>dracula ant</td>
<td>74.8</td>
<td>13,867</td>
<td>721</td>
</tr>
<tr>
<td>Sphaerophalma orestes (Muttillidae)</td>
<td>velvet ant (wasp)</td>
<td>84.3</td>
<td>15,820</td>
<td>707</td>
</tr>
<tr>
<td>Pepsis grossa (Pomphiliidae)</td>
<td>spider wasp</td>
<td>150.3</td>
<td>27,549</td>
<td>703</td>
</tr>
<tr>
<td>Crisocolia alcione (Scoliidae)</td>
<td>scoloid wasp</td>
<td>219.2</td>
<td>45,155</td>
<td>689</td>
</tr>
<tr>
<td>Sceliphron caementarium (Scoliidae)</td>
<td>mud dauber wasp</td>
<td>161.3</td>
<td>33,499</td>
<td>659</td>
</tr>
<tr>
<td>Brachycistis timberlakei (Tiphidae)</td>
<td>tiphid wasp</td>
<td>84.3</td>
<td>15,036</td>
<td>627</td>
</tr>
<tr>
<td>Mischocyttarus flavitarsis (Vespidae)</td>
<td>paper wasp</td>
<td>76.8</td>
<td>15,907</td>
<td>676</td>
</tr>
<tr>
<td>Pseudomasaris vespoides (Vespidae)</td>
<td>pollen wasp</td>
<td>103.7</td>
<td>21,543</td>
<td>678</td>
</tr>
<tr>
<td>Lasiosglossum albipes (Halictidae)</td>
<td>sweat bee</td>
<td>46.7</td>
<td>9,417</td>
<td>585</td>
</tr>
</tbody>
</table>

The assembly for Apterogyna is a partial genome assembly; all others are transcriptome assemblies. In addition to the assemblies derived from new sequence data generated in this study, we also assembled a transcriptome for the sweat bee Lasiosglossum albipes based on raw paired-end short reads downloaded from the NCBI Sequence Read Archive (SRR578269). MB, megabases; aa, amino acids. See also Tables S1 and S2.
with CAT-GTR as the amino acid replacement model, in an unpartitioned analysis of the 308-gene matrix. A species tree was estimated on the basis of average ranks of gene coalescence events, as calculated in STAR [22]. We also inferred a tree with CAT-GTR as the amino acid replacement model, in an unpartitioned analysis of the 308-gene matrix. A species tree was estimated on the basis of minimizing deep coalescences [34]. We used 308 input trees with bootstrap support values generated in RAxML. To evaluate alternate phylogenetic hypotheses against our best-scoring ML tree, we employed the Shimodaira-Hasegawa test [23]. Five constraints were considered (Table S3), and separate constrained partitioned analyses were conducted using RAxML on the same 308-gene matrix used to generate our ML tree (Figure 2). The five best trees satisfying the respective constraints were then subjected to the Shimodaira-Hasegawa test in RAxML.

Accession Numbers

Illumina reads have been deposited in the NCBI Sequence Read Archive with the accession number SRP020476. The matrices, partition files, and gene trees have been deposited in Dryad (http://doi.org/10.5061/dryad.j340).

Supplemental Information

Supplemental Information includes one figure, three tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.08.050.

Acknowledgments

This work was funded by the University of California, Davis. We thank James Pitts for provision of the Apterogyna (Bradyrobaenidae) specimens and three anonymous reviewers for helpful comments that improved the manuscript.

References


Supplemental Information

Phylogenomics Resolves Evolutionary Relationships among Ants, Bees, and Wasps

Brian R. Johnson, Marek L. Borowiec, Joanna C. Chiu, Ernest K. Lee, Joel Atallah, and Philip S. Ward

Supplemental Inventory

Figure S1. Aculeate Hymenoptera phylogeny inferred under different phylogenetic methods, related to Figure 2

Table S1. Species of aculeate Hymenoptera sampled: collection data and composition of pooled material from which RNA or DNA was extracted

Table S2. Number of raw and quality-controlled reads for each library

Table S3. Evaluation of alternative tree topologies using the SH-test as implemented in RAxML

Supplemental Experimental Procedures

Supplemental References
Figure S1.
Figure S1. Aculeate Hymenoptera Phylogeny Inferred under Different Phylogenetic Methods, Related to Figure 2

(A-C) Maximum likelihood (ML) trees derived from 5,214-gene matrix, 3,018-gene matrix, and 525-gene matrix, respectively, estimated with RAxML. All nodes have maximum (100%) bootstrap support. Scale bar indicates substitutions per site.

(D) Majority rule consensus tree from Bayesian analysis of 308-gene matrix, estimated with PhyloBayes. All nodes have 1.00 posterior probability. Scale bar indicates substitutions per site.

(E) Species tree resulting from STAR analysis. Only bootstrap percentages less than 100% are indicated. “Branch lengths” are average ranks of coalescences, not estimated substitutions per site.

(F) Species tree (cladogram) resulting from PhyloNet analysis. Only bootstrap percentages less than 100% are indicated.

Table S1. Species of Aculeate Hymenoptera Sampled: Collection Data and Composition of Pooled Material from which RNA Was Extracted (DNA for Apterogyna)

<table>
<thead>
<tr>
<th>Species</th>
<th>Voucher Specimen Code</th>
<th>Date of Collection</th>
<th>GPS Coordinates</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apterogyna ZA01</td>
<td>CASENT0106304</td>
<td>Oct 2004</td>
<td>-33.43 22.25</td>
<td>2 males</td>
</tr>
<tr>
<td>Chyphotes mellipes</td>
<td>CASENT0106316</td>
<td>3-Jul-2012</td>
<td>39.29020 -118.41550</td>
<td>2 males</td>
</tr>
<tr>
<td>Argochrysium armilla</td>
<td>CASENT0106319</td>
<td>14-Jul-2012</td>
<td>39.42435 -120.23679</td>
<td>3 females</td>
</tr>
<tr>
<td>Stigmatomma oregonense</td>
<td>CASENT0106313</td>
<td>21-Jun-2012</td>
<td>39.99652 -120.99261</td>
<td>2 workers, 6 larvae</td>
</tr>
<tr>
<td>Sphaeropthalma orestes</td>
<td>CASENT0106318</td>
<td>10-Jul-2012</td>
<td>39.77403 -120.07391</td>
<td>2 males</td>
</tr>
<tr>
<td>Pepsis grossa</td>
<td>CASENT0106314</td>
<td>2-Jul-2012</td>
<td>39.28784 -119.27530</td>
<td>1 female, 1 male</td>
</tr>
<tr>
<td>Crioscilia alcine</td>
<td>CASENT0106317</td>
<td>8-Jul-2012</td>
<td>39.84368 -119.43436</td>
<td>1 male</td>
</tr>
<tr>
<td>Sceliphron caementarium</td>
<td>CASENT0106321</td>
<td>26-Jul-2012</td>
<td>38.54030 -121.75628</td>
<td>1 female, 1 male</td>
</tr>
<tr>
<td>Brachycistis timberlakei</td>
<td>CASENT0106315</td>
<td>2-Jul-2012</td>
<td>39.28899 -119.27316</td>
<td>3 males</td>
</tr>
<tr>
<td>Mischocyttarus flavitarsis</td>
<td>CASENT0106320</td>
<td>26-Jul-2012</td>
<td>38.54030 -121.75628</td>
<td>1 female, 1 male</td>
</tr>
<tr>
<td>Pseudomasaris vespoide</td>
<td>CASENT0106312</td>
<td>19-Jun-2012</td>
<td>39.31914 -120.65916</td>
<td>1 female, 1 male</td>
</tr>
</tbody>
</table>

Each voucher specimen code is linked to more detailed collection data on AntWeb (www.antweb.org). Voucher specimens have been deposited in the Bohart Museum of Entomology (UCDC).
Table S2. Number of Raw and Quality-Controlled Reads for Each Library

<table>
<thead>
<tr>
<th>Species</th>
<th>Raw Reads</th>
<th>Filtered Reads 1</th>
<th>Filtered Reads 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apterogyna ZA01</td>
<td>27,243,036</td>
<td>25,293,951</td>
<td>25,293,951</td>
</tr>
<tr>
<td>Chyphotes mellipes</td>
<td>14,529,464</td>
<td>14,362,151</td>
<td>13,744,283</td>
</tr>
<tr>
<td>Argochrysis armilla</td>
<td>17,264,088</td>
<td>17,074,979</td>
<td>16,346,892</td>
</tr>
<tr>
<td>Stigmatomma oregonense</td>
<td>13,648,445</td>
<td>13,457,982</td>
<td>12,773,283</td>
</tr>
<tr>
<td>Sphaerophalma orestes</td>
<td>14,493,142</td>
<td>14,307,149</td>
<td>13,714,095</td>
</tr>
<tr>
<td>Pepsis grossa</td>
<td>18,098,742</td>
<td>17,860,027</td>
<td>17,023,108</td>
</tr>
<tr>
<td>Crioscilia alcione</td>
<td>14,585,832</td>
<td>14,388,055</td>
<td>13,707,012</td>
</tr>
<tr>
<td>Sceliphron caementarium</td>
<td>18,861,014</td>
<td>18,594,250</td>
<td>17,615,417</td>
</tr>
<tr>
<td>Brachycistis timberlakei</td>
<td>13,463,088</td>
<td>13,295,450</td>
<td>12,714,617</td>
</tr>
<tr>
<td>Mischocyttarus flavitarsis</td>
<td>16,910,898</td>
<td>16,742,131</td>
<td>16,148,909</td>
</tr>
<tr>
<td>Pseudomasaris vespoides</td>
<td>16,675,940</td>
<td>16,502,109</td>
<td>15,858,504</td>
</tr>
<tr>
<td>Lasioglossum albipes</td>
<td>19,412,393</td>
<td>19,388,645</td>
<td>19,388,645</td>
</tr>
</tbody>
</table>

“Filtered reads” refers to the number of reads remaining after quality control using fast-toolkit and cutadapt. Reads 1 and 2 refer to forward and reverse reads, respectively.

Table S3. Evaluation of Alternative Tree Topologies Using the SH Test as Implemented in RAxML

<table>
<thead>
<tr>
<th>Constraint Tree</th>
<th>Likelihood$^1$</th>
<th>D(LH)</th>
<th>Significantly Worse (p &lt; 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ants, all aculeates except <em>Argochrysis</em>)</td>
<td>-2515334.7</td>
<td>-753.3 ± 97.1</td>
<td>Yes</td>
</tr>
<tr>
<td>(ants, scolioids + apoids)</td>
<td>-2514722.3</td>
<td>-140.9 ± 47.3</td>
<td>Yes</td>
</tr>
<tr>
<td>(ants, scolioids)</td>
<td>-2514758.4</td>
<td>-176.9 ± 46.9</td>
<td>Yes</td>
</tr>
<tr>
<td>(ants, tiphioid-pompiloid wasps)</td>
<td>-2515486.6</td>
<td>-905.2 ± 83.1</td>
<td>Yes</td>
</tr>
<tr>
<td>(ants, vespids)</td>
<td>-2515424.9</td>
<td>-843.4 ± 97.2</td>
<td>Yes</td>
</tr>
</tbody>
</table>

D(LH) is the difference in log likelihood units between the best constrained tree and the best unconstrained tree.

$^1$Likelihood Score of best ML tree (Figure 2) = -2514581.4.
Supplemental Experimental Procedures

Sampling of Taxa
Samples of ten species of aculeate Hymenoptera were collected at field sites in California and Nevada, United States, in June and July 2012 (Table S1). Taxa were targeted to cover all major families of Aculeata. After collection, live insects were immobilized by freezing, cut into several pieces, briefly dipped in 95% ethanol to reduce the hydrophobic quality of the cuticle, and then immersed in RNAlater. Specimens were kept at 4°C for 24 hours to encourage perfusion of RNAlater, and then stored at -20°C until extraction. The sample of Apterogyna ZA01 (Bradynobaenidae) represents a special case, since fresh collection of this rare African insect was not possible. We therefore extracted DNA from two male specimens collected in 2004 and preserved in 95% ethanol.

Library Construction, Sequencing, and Quality Control
For all samples, one to several pooled specimens were used for RNA or DNA extraction. Table S1 shows the number and sex of each individual from each species used in the study. For samples for which transcriptomes were to be generated, total RNA was extracted using Trizol according to the manufacturer’s instructions. RNA quality and integrity were checked with the Nanodrop 1000 and Bioanalyzer 2100. Total RNA was used for library prep using the Illumina TruSeq v2 kit according to the manufacturer’s instructions. For the Apterogyna ZA01 (Bradynobaenidae) DNA library, DNA was extracted using the ChargeSwitch® gDNA Mini Tissue Kit according to the manufacturer’s instructions. The additional RNAase step was conducted to remove RNA. DNA concentration was determined using Qubit, and an Illumina genomic library was then made using the ChargeSwitch® gDNA Mini Tissue Kit according to the manufacturer’s instructions. The additional RNAase step was conducted to remove RNA. DNA concentration was determined using Qubit, and an Illumina genomic library was then made using the ChargeSwitch® gDNA Mini Tissue Kit according to the manufacturer’s instructions. Both cDNA and DNA libraries were checked for integrity with the Bioanalyzer 2100 before being pooled into one lane on the Illumina HiSeq 2000. Each library had a unique Illumina barcode. Because genome assembly requires more reads than transcriptome assembly, we biased the pooling such that we obtained more reads from the genomic library. 100 bp paired-end sequencing was conducted. Low quality bases and adapter contamination were removed from reads using the fastx toolkit and the cutadapt software package [S1].

Table S2 shows the number of raw and processed reads for each library. In addition to the new genomic and transcriptomic data sets generated, raw Illumina paired-end short reads from whole body extractions were downloaded from the NCBI SRA archive (SRR578269) for the sweat bee, Lasiglossum albipes (family Halictidae). This raw data was quality controlled using the same procedure as the newly generated short reads and was treated identically with respect to transcriptome assembly. Protein sequences of the leafcutter bee Megachile rotundata (primarily from the Megachile genome study) were downloaded from NCBI.

Transcriptome Assembly
De novo transcriptomes were generated with the Trinity short read assembler [S2]. For each assembly, the “transcripts_to_best_scoring_ORFs.pl” script, included with the Trinity software, was used to identify the longest open reading frame in each contig and translate it into a protein sequence. Further identification of unigenes was not necessary as the OrthologID pipeline (see below) automatically estimates the single best ortholog for each gene and discards other sequences.
Genome Assembly: *Apterogyna*

The rationale for this portion of the study was that a complete genome of *Apterogyna AZ01* was unnecessary for the phylogenetic problem at hand. Rather we required a sufficiently comprehensive and phylogenetically informative assembly of protein coding genes. Hence, while short reads alone are typically insufficient for a complete genome assembly, we hypothesized that they would be sufficient to assemble enough protein coding sequence to allow for a rigorous phylogenomic placement of this key taxon.

The genome assembly was performed with the ABYSS software package v1.3.5 [S3]. The resulting assembly was 284.2 MB with an N50 of 1,065 bp and a maximum contig length of 9,143. The assembly is highly fragmented due to low coverage (for a short read dataset) and the likely degradation of the sample, which was collected in ethyl alcohol eight years ago. The result is essentially a “gene assembly” more than a complete genome. For this reason, we used the “transcripts_to_best_scoring_ORFs.pl” script (from the Trinity assembly package) to identify the longest ORF in each contig and translate it into a protein sequence. A more formal genome bioinformatics process was not pursued, as our downstream bioinformatics analyses are designed to identify orthologs from each species and discard chimeric genes.

Orthology Estimation and Data Matrix Assembly

Gene orthology was evaluated using a pre-release version 2.0 of the OrthologID pipeline. Similar to the original version [S4], the latest version of OrthologID takes complete gene sets from all ingroup and outgroup taxa as input and assigns them into gene clusters. OrthologID then generates a parsimony tree for each gene cluster and extracts one or more sets of orthologous genes according to the gene tree topology. In addition to improved execution pipeline on Sun Grid Engine clusters, this version of OrthologID uses the MCL algorithm [S5, S6] for improved clustering, and includes automated extraction of orthologs from gene trees into a partitioned matrix in a single package. Using gene sets from our 19 species as input, OrthologID recovered 19,746 sets of orthologs from 11,119 gene clusters with at least one gene from 5 publicly available genomes (*Apis mellifera*, *Bombus terrestris*, *Harpegnathos saltator*, *Linepithema humile*, and *Pogonomyrmex barbatus*) in each cluster. For phylogenetic analysis, these orthologous sets of genes or gene fragments were then assembled by OrthologID into multiple partitioned matrices with different levels of taxon representation per gene, including (i) a 5,214-gene matrix with 3,001,657 amino acid sites and at least 9 ingroup taxa represented per gene partition, (ii) a 3,018-gene matrix with 1,653,740 sites and at least 16 ingroup taxa represented per partition, and (iii) a matrix that has every gene partition represented across all ingroup taxa with the additional requirement that they are single-copy in the 5 publicly available ingroup genomes. This matrix has a total of 525 partitions and 298,968 sites. To further reduce this matrix to allow for the most computationally intensive analysis, i.e. Bayesian inference, we extracted a submatrix by examining the maximum likelihood tree estimate of each of the 525 partitions, calculating the total tree length and branch length variance for each partition and choosing only the gene partitions that lay within the 68th percentile in both categories. We used the 68th percentile because this reduced the matrix to a manageable size (~300 genes) for Bayesian analysis. The resultant 308-gene matrix has 175,404 sites. Of the 3,332,676 cells (19 taxa × 175,404 sites) in this 308-gene matrix, 73.42% are coded as amino acids, 11.60% are gaps, and 14.98% are missing. Equivalent percentages of missing data in the 5,214-gene, 3,018-gene and 525-gene matrices are 36.77%, 21.53% and 15.16%, respectively.
**Maximum Likelihood Analysis**

We performed maximum likelihood (ML) analysis on the different matrices described above. Partitioned analyses with Γ-distributed rate heterogeneity over sites were performed using RAxML v7.4.2 [S7, S8]. For each matrix, the best protein substitution model for each gene partition was selected individually using the “ProteinModelSelection.pl” script [S8] over 36 different models. We used the MPI-AVX version of RAxML to perform rapid bootstrap on each of the matrices, and the PTHREADS-AVX version to search for the best scoring trees. At least 250 bootstrap replicates were computed for each matrix. For the 308-gene matrix, 1,000 bootstrap replicates were performed. Identical topologies with bootstrap value of 100 at every node were recovered for all 4 matrices with very small differences in relative branch lengths (Figures 2, S1A-C), except for *Apterogyna* ZA01, which has a shorter relative branch length for the smaller but more completely represented matrices.

**Bayesian Inference**

Bayesian inference was performed on the 308-gene matrix using PhyloBayes-MPI v1.3b [S9, S10]. CAT-GTR is recommended in Lartillot *et al.* [S9] as the best overall model implemented in PhyloBayes for large data sets and it was therefore chosen as the amino acid replacement model for our analysis. We executed two independent chains of at least 7,000 cycles on 128 CPU cores each. Convergence was determined by visually examining the trace plots of the PhyloBayes summary statistics using the mcmcplots R library [S11]. Using the *bpcomp* program of PhyloBayes on the two chains with a burn-in of 1,500 returned maxdiff and meandiff of 0, indicating that no discrepancies were observed across all bipartitions after convergence. The consensus of all trees pooled across both chains after burn-in was identical to the ML tree in topology, with posterior probabilities of 1.0 across all nodes (Figure S1D).

**Species Tree Inference**

Although model-based methods of species tree estimation have been shown to outperform other approaches, their computational costs are prohibitive for genomic-scale data sets [S12, S13]. For that reason, we estimated the species tree with average ranks of gene coalescence events in STAR [S14]. As input we used 100 bootstrap replicate trees of each of 308 genes from the small matrix, built under maximum likelihood in RAxML. The analyses were carried out on the STRAW web server [S15].

In addition we inferred a species tree in PhyloNet [S16], which uses the parsimony-based criterion of minimizing deep coalescences [S17]. We used 308 input trees with bootstrap support values generated in RAxML. PhyloNet allows for accounting of uncertainty in gene trees and performs better when collapsing poorly supported splits [S16]. We thus set an arbitrary cutoff value of 70, contracting all branches with lower bootstrap support and ran the analysis over 10,000 replicates. The resulting cladogram is presented in Figure S1F.
Supplemental References


S11. Curtis, S. M. (2012) R package - mcmcplots v0.4.1


