Phylogenomic Analysis of Ants, Bees and Stinging Wasps: Improved Taxon Sampling Enhances Understanding of Hymenopteran Evolution

Short title: Phylogenomic Analysis of Aculeata

Authors:

Michael G. Branstetter[a,b], Bryan N. Danforth[c], James P. Pitts[d], Brant C. Faircloth[e], Philip S. Ward[f], Matthew L. Buffington[g], Michael W. Gates[g], Robert R. Kula[g], Seán G. Brady[b]

Author Affiliations:

[a]Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA

[b]Department of Entomology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, 10th & Constitution Aves. NW, Washington, D.C., 20560, USA

[c]Department of Entomology, 3119 Comstock Hall, Cornell University, Ithaca, NY 14853, USA

[d]Utah State University, Department of Biology, 5305 Old Main Hill, Logan, UT 84322-5305, USA

[e]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA

[f]Department of Entomology and Nematology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

[g]Systematic Entomology Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, U.S. Department of Agriculture, *c/o* Department of Entomology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, 10th & Constitution Ave. NW, Washington, D.C., 20560, USA

Corresponding Author:

Michael G. Branstetter; Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA; Phone: 801-581-6609; Email: mgbranstetter@gmail.com

Abstract

The importance of taxon sampling in phylogenetic accuracy is a topic of active debate. We investigated the role of taxon sampling in causing incongruent results between two recent phylogenomic studies of stinging wasps (Hymenoptera: Aculeata), a diverse lineage that includes ants, bees and the majority of eusocial insects. Using target enrichment of ultraconserved element (UCE) loci, we assembled the largest aculeate phylogenomic data set to date, sampling 854 loci from 187 taxa, including 30 out of 31 aculeate families, and a diversity of parasitoid outgroups. We analyzed the complete matrix using multiple analytical approaches, and also performed a series of taxon inclusion/exclusion experiments, in which we analyzed taxon sets identical to and slightly modified from the previous phylogenomic studies. Our results provide a highly supported phylogeny for virtually all aculeate lineages sampled, supporting ants as sister to Apoidea (bees+apoid wasps), bees as sister to Philanthinae+Pemphredoninae (lineages within a paraphyletic Crabronidae), Melittidae as sister to remaining bees, and paraphyly of cuckoo wasps (Chrysidoidea). Our divergence dating analyses estimate ages for aculeate lineages in close concordance with the fossil record. Our analyses also demonstrate that outgroup choice and taxon evenness can fundamentally impact topology and clade support in phylogenomic inference.

Keywords: ultraconserved elements, phylogenomics, Hymenoptera, Aculeata, next-generation sequencing, taxon sampling

Introduction

The role of taxon sampling in improving phylogenetic accuracy is a topic of long-term controversy (1–9). Rosenberg & Kumar (8) argued that increasing the number of characters sampled is a better investment of resources compared to adding taxa. However, this conclusion has received much criticism and many subsequent studies have argued the opposite point (4,6), including some recent investigations that have employed genome-scale data (7,10,11). In the current age of phylogenomics, in which it is now possible to generate data sets with hundreds to thousands of loci (12,13), the argument over the relative importance of taxon versus character sampling has become largely irrelevant, with the more important question being: does improved taxon sampling increase phylogenetic accuracy? Here, we examine this question using a genome-scale data set that focuses on relationships within a major clade of insects.

Encompassing over 120,000 described species and having an estimated richness that might exceed two million species, the insect order Hymenoptera represents one of four insect megaradiations, (14–16). This extreme diversity includes many important lineages (*e.g.* sawflies, wood wasps, parasitic wasps), with arguably the most well known taxa belonging to the stinging wasps (Aculeata). The aculeates have attracted much attention because they include all eusocial Hymenoptera, most notably the ecologically and economically important ants and bees (17,18), and also the eusocial wasps (*e.g.*, paper wasps, hornets, and yellow jackets) (19). Eusociality has in fact evolved independently at least 6–8 times within the clade, making the group a model for studying the evolution of

sociality (20–25). Outside of eusocial lineages, the group exhibits a wide range of life history strategies, with most species tending to be solitary or subsocial predators, specializing on a wide variety of arthropod prey (26,27). A number of taxa have also evolved endoparasitic or even herbivorous feeding strategies (*e.g.*, pollen and nectar) (14,15).

Given their diversity and importance, establishing a robust phylogeny and classification of the aculeates is of broad interest. Currently, the Aculeata includes over 70,000 described species and is divided into 9 superfamilies and 31 families (28). This classification is based upon a molecular study that found several morphologically circumscribed superfamilies and families to be non-monophyletic, most notably the Vespoidea, Bradynobaenidae, and Tiphiidae (28). More recent molecular studies have also provided new hypotheses for the phylogenetic positions of bees (29) and ants (30,31). Despite these improvements, considerable uncertainty exists as to the relationships among superfamilies and families within Aculeata.

To date, most molecular studies of Hymenoptera have used traditional Sanger sequencing methods, resulting in data sets with decent taxon sampling, but few loci and often low clade support (28,29,32–34). Several recent studies have instead employed next-generation sequencing approaches, but so far these have suffered from including few taxa (30,31,35). Two phylogenomic studies in particular produced conflicting relationships with regard to the phylogenetic position of ants. In the study of Johnson *et al.* (30) the authors used transcriptome data to resolve relationships among aculeate superfamilies

and found ants to be sister to apoid wasps and bees (Apoidea), a novel and biologically attractive result. Conversely, Faircloth *et al.* (31), using hundreds of ultraconserved element loci (UCEs), recovered ants as sister to all other aculeate superfamilies (minus Chrysidoidea, which was not represented). Despite both studies employing genome-scale data, each produced highly supported but conflicting results. One potential problem for both studies was sparse taxon sampling, with Johnson *et al.* (30) including all superfamilies, but only 19 taxa, and Faircloth *et al.* (31) including 44 taxa, spanning six out of seven superfamilies, but with sampling biased towards the ants and missing a key outgroup (Chrysidoidea).

To test the hypothesis that taxon sampling caused the incongruent results between these phylogenomic studies, and to address important remaining uncertainties within Aculeata at the family level, we have generated the largest phylogenomic data set to date for the Aculeata. Building upon the study of Faircloth *et al.* (31), we have assembled a UCE data set comprising 187 taxa that includes all aculeate superfamilies, 30 out of 31 aculeate families (missing only Scolebythidae), and a diversity of outgroup superfamilies from across Hymenoptera. We analyzed the complete, 187-taxon matrix using multiple analytical approaches and recovered a highly supported phylogeny for virtually all aculeate lineages sampled. We also focused our sensitivity analyses on the placement of ants and bees within the Aculeata and found that taxon sampling can have a major impact on results even with genome-scale data.

Results

126 *Sequencing Results*

127

128 To generate our phylogenomic data set we used a recently developed approach that
129 combines the targeted enrichment of ultraconserved element loci (UCEs) with
130 multiplexed next-generation sequencing (36). We followed published lab protocols
131 (31,36; see also materials and methods below) and used a Hymenoptera-specific probe
132 set that targets 1,510 UCE loci from across the entire order. Using this approach we
133 sequenced new molecular data for 139 taxa, and we combined these data with 16 taxa
134 from Faircloth et al. (31) and 32 taxa from available genomes, resulting in a final data set
135 that included 187 taxa (see electronic supporting information S1, Tables 1 and 2).

136

137 Within our taxon set we included 136 samples from within the Aculeata, representing 30
138 out of 31 recognized aculeate families (missing only Scolebythidae). Sampling within the
139 Apoidea was particularly dense with 53 species sampled from 23 out of 25 recognized
140 bee subfamilies, and 16 species from outside bees including the phylogenetically
141 enigmatic families Ampulicidae and Heterogynaidae. We also included 14 species from
142 four out of eight subfamilies within the paraphyletic family Crabronidae (29). For
143 outgroup taxa, we sampled all superfamilies from within the sawfly grade ("Symphyta"),
144 and 8 out of 12 non-aculeate superfamilies from within the Apocrita ("Parasitica"),
145 including Trigonaloidea, Evanioidea, Ichneumonoidea, and Ceraphronoidea. Those taxa
146 have been hypothesized in previous analyses as lineages closely related to Aculeata
147 (15,32–34,37). To better compare results between the Johnson *et al.* (30) transcriptome

148 study and our UCE study, we sampled DNA from 7 out of 12 of the same specimen series
149 that were sampled in Johnson *et al*. (30).
150
151 After sequencing of enriched samples, we used the PHYLUCE v1.5 software package
152 (38) to clean and assemble raw reads; extract, align and trim UCE loci (for sequenced and
153 genome-enabled taxa); filter loci for taxon completeness, and generate DNA matrices
154 ready for phylogenetic analysis (see materials and methods for details). For all taxa that
155 we enriched and sequenced, we recovered an average of 966 UCE contigs per sample,
156 with a mean contig length of 801 bp and an average coverage per UCE contig of 80X (for
157 complete assembly stats see supporting information S1, Table 4). For genome-enabled
158 taxa, we recovered an average of 1,036 UCE loci. Using our set of UCE alignments for
159 all taxa, we evaluated the effects of filtering alignments for various levels of taxon
160 occupancy (% of taxa required to be present in a given locus) and selected the 75%
161 filtered locus set ("*Hym-187T-F75*") as the primary locus set for analysis. The *Hym-*
162 *187T-F75* locus set included 854 loci and had an average locus length of 238 bp resulting
163 in a concatenated data matrix of 203,095 bp of which 143,608 sites were phylogenetically
164 informative (for all matrix stats see supporting information S1, table 5).
165
166 *Phylogeny of Aculeata*
167
168 After filtering for taxon completeness, we carried out maximum likelihood (ML) and
169 Bayesian (BI) analyses on the concatenated *Hym-187T-F75* matrix using RAXML v8.0.3
170 and EXABAYES v1.4.1 (39), respectively. For both approaches we partitioned the data

171 set using the kmeans algorithm available in a development version of
172 PARTITIONFINDER (PF) (40), and for the ML searches we analyzed the matrix in
173 several additional ways: (1) unpartitioned, (2) partitioned by locus, and (3) partitioned by
174 the hcluster algorithm in PF v1.1.1 (data pre-partitioned by locus). We also ran three
175 analyses using the summary method implemented in ASTRAL v4.8.0 for species tree
176 estimation (41). For input into ASTRAL we generated bootstrapped gene trees for all loci
177 using RAxML (200 reps). In the first analysis we used all individual gene trees and
178 accompanying bootstrap trees as input into ASTRAL (854 loci total). In the second
179 analysis we calculated and sorted loci by average bootstrap score (=informativeness)
180 using R v3.2.2 (42) and we selected the 500 loci that had the highest scores for input into
181 ASTRAL. We did this to reduce possible error/bias introduced by including
182 uninformative loci, a problem that has been observed in other studies (43–45). For the
183 third analysis we used all loci; however, to reduce error from loci with low information
184 content we employed weighted statistical binning, which bins loci together based on
185 shared statistical properties and then weights bins by the number of included loci (46)
186 (details in electronic supporting material). We ran all species-tree analyses with 100
187 multi-locus bootstrap replicates (47).
188
189 To investigate other potential biases in our data, we carried out several additional
190 analyses. In particular we wanted to address the observation that G+C variance can be a
191 problem for reconstructing phylogeny in aculeate Hymenoptera (21). First, using
192 PHYLUCE, we converted the complete, concatenated matrix to RY coding and we
193 performed a best tree plus rapid bootstrapping analysis (100 bootstrap replicates) in

194 RAxML using the BINGAMMA model of sequence evolution. Second, we filtered loci
195 for various parameters calculated in R (scripts modified from (48)) and PHYLUCE:
196 average bootstrap score, % invariant sites (= rate of evolution), and G+C variance. We
197 then removed the 10% of loci that had the highest values for GC variance, and the top
198 10% of loci that had the lowest values for bootstrap score and % invariant sites.
199 Following removal of outlier loci we retained 636 alignments ("best636"), and we
200 concatenated these into a single matrix and analyzed the matrix unpartitioned in RAXML
201 (best tree searches with 100 rapid bootstrap replicates). We did not partition the data
202 because partitioning had little effect in the analysis of the complete matrix.
203
204 Across analyses we recovered a robust phylogeny of the Aculeata (Fig 1 and electronic
205 supporting information S2, figures 1-14), with the topology being identical for all ML
206 and BI analyses of the complete, non-RY-coded data, and nearly identical for the ST
207 analyses and the ML analysis of the complete, RY-coded data (we recovered several
208 differences within Chrysidoidea, noted below).
209
210 We recovered the superfamily Trigonaloidea as sister to Aculeata in all analyses, and
211 with maximum support except in the unbinned species tree analyses (97-98% support).
212 Although we are missing several parasitoid superfamilies in our data set, this result is
213 congruent with results from several recent molecular analyses (32,34,37), but is
214 incongruent with results from (33). We did not recover the Ichneumonoidea, which has
215 been a long-standing candidate as the sister group to the Aculeata (15), to be the sister
216 group in any analysis. Within Aculeata, we recovered part of Chrysidoidea (cuckoo

217 wasps and relatives) as sister to the remaining superfamilies, with Chrysidoidea itself
218 paraphyletic, forming a grade of two (ML and BI analyses), three (binned ST analysis),
219 or four (unbinned ST analyses) clades, depending on the analysis. In the ML and BI
220 analyses of the non-RY-coded data, the first clade included
221 [Chrysdidae+[Plumariidae+Bethylidae]] and the second clade included
222 [Sclerogibbidae+[Embolemidae+Dryinidae]]. The placement of the second clade as sister
223 to the remaining Aculeata, and the placement of Sclerogibbidae within the clade, received
224 less than maximum bootstrap support in the ML analysis. In the analysis of the RY-coded
225 data, we recovered a paraphyletic Chrysidoidea, but with only Sclerogibbidae falling
226 outside of the superfamily. Results varied among the ST analyses, with the binned result
227 being the same as in the non-RY-coded analyses except that Sclerogibbidae was placed
228 outside of clade two and as sister to all remaining Aculeata. In the unbinned ST analyses
229 the taxon *Plumarius* (Plumariidae) was moved out of the first clade mentioned above and
230 placed as sister to Sclerogibbidae plus all other aculeates.
231
232 The remaining aculeate subfamilies separated into two major clades that were highly
233 supported in all analyses. The first clade includes the superfamilies Vespoidea,
234 Tiphioidea, Thynnoidea, and Pompiloidea, as well as the family Sierolomorphidae
235 (currently in Tiphioidea). The monophyly of this group received maximum or nearly
236 maximum support in all ML and BI analyses (≥ 98%), and slightly reduced support in the
237 ST analyses (≥ 93%). Within the clade, we recovered a consistent topology across all
238 analyses, with Vespoidea (includes Rhopalosomatidae and Vespidae) sister to the
239 remaining superfamilies, and the phylogenetically enigmatic family Sierolomorphidae

sister to [Pompiloidea+[Tiphioidea+Thynnoidea]]. Relationships among superfamilies received maximum support across analyses, except the monophyly of Vespoidea received less than maximum support in the ML analysis of the *best636* data set (98%) and the unbinned ST analyses (≥ 84%). Within Pompiloidea we recovered Pompilidae as sister to [Sapygidae+[Myrmosidae+Mutilidae]], but support for the position of Myrmosidae was less than maximum in all analyses except BI (≥ 57%), and support for the position of Sapygidae was reduced in a few analyses (≥ 74%), suggesting uncertainty. The second major clade contained the remaining aculeate superfamilies, with Scolioidea recovered as sister to Formicoidea+Apoidea in all analyses. This result received maximum support in all concatenated analyses. However, Scolioidea sister to Formicoidea+Apoidea received somewhat lower support in ST analyses (≥ 96%), and Formicoidea+Apoidea received 90% support in the binned ST analysis and only 43% and 7% support in the 500 best and all loci ST analyses. Overall, relationships among superfamilies largely agree with the recent Johnson *et al*. transcriptome study (30), except for the placement of Vespoidea.

Within Apoidea (bees and apoid wasps), our results are completely consistent across analyses and largely agree with Debevec *et al.* (29). We recovered Ampulicidae as sister to remaining taxa, and we found Crabronidae to be paraphyletic with respect to Sphecidae and bees. The remaining taxa formed a grade in the following order: [Heterogynaidae+[Crabroninae+Sphecidae], Bembicini, Phemphredoninae+Philanthinae, and the bees (Anthophila). The position of the enigmatic family Heterogynaidae as sister to Crabroninae+Sphecidae is a novel result, receiving less than maximum support only in the ST analyses (98% binned and ≥ 32% unbinned). The position of the bees as sister to

263 the Pemphredoninae+Philanthinae was first reported in Debevec *et al.* (29) and was also
264 recovered here with maximum support in all analyses except unbinned ST analyses (≥
265 89%).
266
267 Within bees, we recovered Melittidae to be sister to all remaining families, with
268 maximum support in concatenated analyses (≥ 47% in ST analyses), as found in several
269 previous studies (22). The remaining families were divided into two major clades:
270 [Megachilidae+Apidae] (i.e., "long-tongued" bees *sensu* Michener (18)), and
271 [Andrenidae+[[Stenotritidae+Colletidae]+Halictidae]]. Relationships of subfamilies
272 within all families are largely congruent with previous studies of bee higher-level
273 relationships (49). Within Apidae we recovered a monophyletic "cleptoparasitic clade"
274 (50), monophyly of Anthophorini, Xylocopinae (51), and a sister-group relationship
275 between Centridini and corbiculates. Relationships within corbiculates were notable
276 because we recovered monophyly of the eusocial corbiculate tribes
277 (Apini+Bombini+Meliponini) in all analyses except the ST analyses, which placed Apini
278 as sister to [Euglossini+[Bombini+Meliponini]] with less than maximum support.
279
280 *Taxon Sampling Experiments*
281
282 To test the effects of taxon sampling on phylogenetic inference and to examine the
283 incongruent placement of ants between previous phylogenomic studies (30,31), we
284 created and analyzed a series of alternative taxon sets, which can be divided into three
285 categories (Fig 2): (1) variations of Johnson *et al.* (30), (2) variations of Faircloth *et al.*

(31), and (3) variations of the current taxon set. For the first category, we generated two data sets, one with exactly the same taxon sampling as (30) ("*Johnson-19T*"), and one with the chrysidoid *Argochrysis armilla* removed ("*Johnson-18T*"). This particular manipulation was done because the major difference between the two phylogenomic studies was the presence/absence of Chrysidoidea, which is the sister taxon to the rest of Aculeata.

For the Faircloth *et al.* (31) manipulations we recreated the original 45 taxon matrix ("*Faircloth-45T*") and then created several alternative taxon sets. First we added a single chrysidoid ("*Faircloth-46T*"), and then continued to add additional aculeates to balance the data set ("*Faircloth-52T*", "*Faircloth-56T*" and "*Faircloth-61T*"). We also tried balancing the data set by removing most ant taxa from the original data set ("*Faircloth-26T*") and adding in a chrysidoid ("*Faircloth-27T*").

Finally, for the third category of taxon sampling experiments, we generated a data set with most outgroups removed ("*Hym-147T*"), leaving *Nasonia* as the earliest diverging outgroup and *Megaspilus* (Ceraphronoidea), Evanioidea, and Trigonaloidea as more recently diverging outgroups. From this taxon set, we removed chrysidoids ("*Hym-133T*") and chrysidoids plus trigonaloids ("*Hym-131T*"). We also attempted to create the most balanced data set we could by removing excessive ant, bee and wasp taxa ("*Hym-100T*"). By removing distantly related outgroups, we not only reduced the number of taxa, but we potentially increased the average length of alignments. This is because UCE loci become more variable away from the central, core region (36) and alignment

309 trimming (see materials and methods) removes poorly aligned regions. Thus, by
310 removing more distant outgroups, alignments should be improve at the flanks of loci and
311 less data should be trimmed.
312
313 In our description of the results we focus on the placement of ants (Formicoidea) among
314 the other major lineages (superfamilies, etc.) of Aculeata. Among taxon sets, we
315 recovered three alternative topologies (Fig 2, Table 1, and electronic supporting
316 information S2, Figs 18-30): (A) ants sister to Apoidea, (B) ants sister to all other groups,
317 minus Chrysidoidea, and (C) ants sister to Apoidea plus Scoliodea. In both of the Johnson
318 *et al.* matrices, we recovered topology A. However, when we removed the chrysidoid,
319 bootstrap support values for the relationships among ants, Apoidea, Scoliodea,
320 Vespoidea, and Tiphiodea+Pompiloidea were reduced from maximum to 89%. We found
321 a similar result in the analyses of the *Hym-147T* matrix and variants. All three matrices
322 (*Hym-147T, Hym-133T,* and *Hym-131T*) produced topology A, but when chrysidoids and
323 trigonaloids were removed (*Hym-131T*), support for the positions of ants as sister to
324 Scolioidea+Apoidea was lowered to 90%.
325
326 Analysis of the original Faircloth *et al.* (31) taxon set (*Faircloth-45T*) produced topology
327 B, as in the original study. Adding a chrysidoid to the taxon set (*Faircloth-46T*) did not
328 change the topology, but did reduce support for the position of ants slightly (99%).
329 In the *Faircloth-52T* and *Faircloth-56T* analyses, we also recovered topology B.
330 However, in the *Faircloth-61T* analyses the topology shifted to C, placing ants as sister to
331 Scolioidea plus Apoidea. The difference between *Faircloth-56T* and *Faircloth-61T* was

the addition of several chrysidoids (Embolemidae and Dryinidae), Rhopalosomatidae (Vespoidea), and Ampulicidae (Apoidea), with the latter two taxa breaking long branches. Reducing and balancing the taxa of *Faircloth-45T* also altered the resulting topology. By reducing the number of ant taxa from 22 in *Faircloth-45T* to 3 taxa in *Faircloth-26T* the topology changed to A, but with only moderate support for Formicoidea+Apoidea (88%). Adding in a chrysidoid (*Faircloth-27T*) also resulted in topology A, and with nearly maximum bootstrap support for Formicoidea+Apoidea (97%).

Lastly, for the *Hym-100T* matrix, in which we reduced the number of ant and bee taxa to balance the larger taxon set, we recovered topology A, with the Formicoidea+Apoidea clade receiving maximum bootstrap support. All other relationships among superfamilies and within Apoidea were the same as those in the ML analysis of the *Hym-147T* and *Hym-187T* matrices.

Divergence Dating

To generate a time tree for the evolution of the stinging wasps we estimated divergence dates for the complete 187 taxon matrix using the program BEAST v1.8.2 (52). We calibrated the analysis using 36 fossils representing taxa from across Hymenoptera and one secondary calibration taken from (53) for the root node (electronic supporting information S1, Table 3). For fossil ages we used midpoint dates taken from date ranges provided on the Fossilworks website (54) (http://fossilworks.org/). Due to computational

355 challenges with BEAST, arising from having both a large number of taxa and a large
356 amount of sequence data, we made the analysis feasible by inputting a starting tree (all
357 nodes constrained), turning off tree-search operators, and using only a subset of the
358 sequence data set rather than the entire concatenated matrix (details in electronic
359 supporting material). We performed three separate analyses to compare the effects of
360 different sets of loci on the final, dated results: (1) 25 loci that had the highest gene-tree
361 bootstrap scores, (2) 50 loci that had the highest gene-tree bootstrap scores, and (3) 50
362 randomly selected loci.
363
364 The analysis of the three different locus sets (25 best loci, 50 best loci, 50 random loci)
365 returned completely congruent dates (Table 2 and electronic supporting information S2,
366 Figs 15-17). Consequently, we report here just the dates from the analysis of 50 random
367 loci (Fig 2 and electronic supporting information S2, Fig 17). We estimated an age of 257
368 Ma (240-274 Ma 95% HPD) for crown Hymenoptera and 200 Ma (187-216 Ma) for
369 Euhymenoptera (Orussoidea+Apocrita). The Apocrita arose 194 Ma (181-208), followed
370 by the Aculeata at 161 Ma (154-169 Ma). Within Aculeata all of the superfamilies
371 originated between 161 Ma to 100 Ma. The ants, minus the earliest diverging subfamilies
372 Leptanillinae and Martialinae (not sampled in the current study), arose at least 118 Ma
373 (108-128 Ma; Amblyoponinae+formicoid clade). The Apoidea arose 131 Ma (121-141
374 Ma), followed by the bees at 100 Ma (92-107 Ma).
375
376 **Discussion**
377

The coupling of next-generation sequencing with reduced representation phylogenomics has driven a revolution in molecular systematics, making it possible to generate genome-scale data sets for hundreds of taxa at a fraction of the cost of traditional methods (12,13,55). Here, we further applied one of the most promising approaches, the target enrichment of ultraconserved elements (UCEs) (36), to the megadiverse insect order Hymenoptera, greatly extending a previous study which first introduced this approach in insects (31). We focused on family-level relationships of the stinging wasps (Aculeata) and produced a robust backbone phylogeny that confirms the utility of the UCE approach in Hymenoptera. In addition, by carrying out a series of taxon sampling experiments, we have demonstrated that even in the era of phylogenomics, careful taxon sampling and the use of taxon inclusion/exclusion experiments can be of critical importance.

Our phylogenomic results for Aculeata are largely consistent with and significantly amplify two previous molecular studies that employed traditional Sanger sequencing methods (28,29), and one recent transcriptome-based study (30). Compared to the two Sanger-based efforts, which both included a more limited number of taxa, our results agree in terms of the composition of superfamilies and families, with the only differences among studies being our finding that Chrysidoidea is paraphyletic and that the enigmatic family Sierolomorphidae is sister to [Pompiloidea+[Tiphioidea+Thynnoidea]]. The latter result supports resurrecting the superfamily Sierolomorphoidea, originally proposed by (56). Relationships among superfamilies, however, are quite different among these two studies, and our results mostly agree with those reported in the transcriptome study by Johnson *et al.* (30). An exception is the placement of Vespoidea in our study as sister to

401 [Sierolomorphidae+[[Tiphioidea+Thynnoidea]+Pompiloidea]]]. This novel result is
402 possibly due to our more extensive taxon sampling within Vespoidea (inclusion of
403 Rhopalosomatidae) as compared to (30).
404
405 Our results strengthen previous findings of relationships within the Apoidea (29) and
406 within the bees (Anthophila) (49). We confirm placement of Ampulicidae as sister to
407 remaining Apoidea and the bees as sister to the crabronid subfamilies
408 Philanthinae+Pemphredoninae. However, future studies should include an even broader
409 sampling of Pemphredoninae and Philanthinae to confirm this hypothesis. Novel to our
410 study is the placement of Heterogynaidae as sister to Crabroninae+Sphecidae. This taxon
411 was previously placed as either sister to Apoidea (inside Ampulicidae) (29), sister to
412 Astatinae+Bembicini (29), or sister to Philanthinae+Anthophila (57).
413
414 Within bees, our results provide further confirmation that the family Melittidae,
415 previously thought to be sister to the long-tongued bees based on morphology (58), is
416 monophyletic and sister to remaining bee families. It is also notable that most of our
417 analyses recovered the eusocial corbiculate bees as monophyletic and sister to the weakly
418 social Euglossini, thus favoring a single origin of eusociality within the group.
419 Relationships among these taxa have been controversial, but our result agrees with a
420 recent phylogenomic study that found that controlling for base-compositional
421 heterogeneity, specifically GC variance among taxa, favored monophyly of eusocial
422 corbiculates (21). The fact that we recovered this result without controlling for base

423 compositional bias suggests that our UCE loci are robust to this problem, as was also
424 suggested in another study of mammalian relationships (59).
425
426 Of major importance for understanding the evolution of eusociality, is our strongly
427 supported result that Formicoidea (the ants) is sister to Apoidea (apoid wasps and bees).
428 While disagreeing with the previous UCE study (31), it is in full agreement with the
429 transcriptome study (30). The reason for the earlier conflict between these sources of
430 phylogenomic data appears to be due to taxon sampling, with the earlier UCE study
431 missing a key outgroup (Chrysidoidea) and having an excessive number of ant taxa (note
432 that these were included intentionally to test the UCE method at resolving deep and
433 shallow divergences), making the data set unbalanced. By conducting a series of taxon
434 sampling experiments we demonstrated that excluding Chrysidoidea (or Chrysidoidea
435 and Trigonaloidea) reduced bootstrap support for ants being sister to Apoidea. We also
436 found that by either removing the disproportionate numbers of ant taxa, or adding
437 additional taxa to the Faircloth *et al.* (31) taxon set, we were able to infer a topology
438 consistent with both the transcriptome study and our more comprehensive taxon set
439 presented here. Although the placement of ants as sister to Apoidea should still receive
440 further investigation, we believe this result is the preferred one given its robustness across
441 all of our analyses (ML, BI, and ST). Moreover, as discussed in Johnson *et al.* (30), the
442 result is biologically attractive given that Apoidea includes the greatest number of
443 eusocial Hymenoptera and all ants are eusocial. Furthermore, the finding that both
444 bootstrap support and topology were affected by taxon sampling, provides additional
445 evidence that taxon sampling in phylogenetics should still be a major concern, even in the

446 age of phylogenomics, when data are no longer a limiting variable. Overcoming this
447 challenge will require expanded and informed taxon selection as well as improved
448 models and computational methods that can handle genome-scale data sets.
449
450 **Materials and Methods**
451
452 *UCE Sequencing Pipeline*
453
454 For all newly sampled taxa, we extracted DNA using Qiagen DNeasy Blood and Tissue
455 kits (Qiagen Inc., Valencia, CA) and we fragmented up to 500 ng of input DNA to an
456 average fragment distribution of 400-600 bp using a Qsonica Q800R sonicator (Qsonica
457 LLC, Newton, CT). Following sonication, we constructed sequencing libraries using
458 Kapa library preparation kits (Kapa Biosystems Inc., Wilmington, MA) and custom
459 sample barcodes (60). We assessed success of library preparation following PCR
460 amplification by measuring DNA concentration and visualizing libraries on an agarose
461 gel. We purified reactions following PCR using 0.8 to 1.0X AMPure substitute (61).
462
463 For UCE enrichment we pooled 6–10 libraries together at equimolar concentrations and
464 adjusted pool concentrations to 147 ng/µl. For each enrichment we used a total of 500 ng
465 of DNA (3.4 µl each pool), and we performed enrichments using a custom RNA bait
466 library developed for Hymenoptera (31) and synthesized by MYcroarray (MYcroarray,
467 Ann Arbor, MI). The probe set includes 2,749 probes, targeting 1,510 UCE loci. We
468 hybridized RNA bait libraries to sequencing libraries at 65ºC for a period of 24 hours,

469 and we enriched each pool following a standardized protocol (version 1.5; protocol
470 available from http://ultraconserved.org).
471
472 We verified enrichment success with qPCR (ViiA 7, Applied Biosystems, Waltham MA)
473 by comparing amplification profiles of unenriched to enriched pools using PCR primers
474 designed from several UCE loci. After verification, we used qPCR to measure the DNA
475 concentration of each pool, and we combined all pools together at equimolar ratios to
476 produce a final pool-of-pools. To remove overly large and small fragments, we size-
477 selected the final pools to a range of 300–800 bp using a Blue Pippin size selection
478 instrument (Sage Science, Beverly, MA). We mailed size-selected pools to either the
479 UCLA Neuroscience Genomics Core or the Cornell University Biotechnology Resource
480 Center (http://www.biotech.cornell.edu/brc/genomics-facility), where the samples were
481 quality checked on a Bioanalyzer (Agilent Technologies, Santa Clara, CA), quantified
482 with qPCR, and sequenced on an Illumina HiSeq 2500 (2x150 Rapid Run; Ilumina Inc,
483 San Diego, CA).
484
485 *Matrix Assembly*

486 The sequencing facilities demultiplexed and converted raw data from BCl to FASTQ
487 format using either BASESPACE or BCL2FASTQ (available at http:// support. illumina.
488 com/ downloads/ bcl2fastq_ conversion_ software_ 184. html). Using these files, we
489 cleaned and trimmed raw reads using ILLUMIPROCESSOR (62), which is a wrapper
490 program around TRIMMOMATIC (63,64). We performed all initial bioinformatics steps,
491 including read cleaning, assembly, and alignment, using the software package

PHYLUCE v1.5. For sequenced samples, we assembled reads *de novo* using a wrapper script around TRINITY v2013-02-25 (65). After assembly, we used PHYLUCE to identify individual UCE loci from the bulk of assembled contigs while removing potential paralogs. We then used PHYLUCE to combine the UCE contigs from the sequenced taxa with the contigs from the 32 genome-enabled taxa into a single FASTA file. We aligned all loci individually using a wrapper around MAFFT v7.130b (66), and we trimmed the alignments using a wrapper around GBLOCKS v0.91b (67,68), which we ran with reduced stringency settings (0.5, 0.5, 12, and 7 for b1–4 settings, respectively).

To extract an equivalent set of UCE loci from 32 genome-enabled taxa, we downloaded Hymenoptera genomes from NCBI and the Hymenoptera Genome Database (69). The genome of *Apterognya* za01 was provided by the authors of Johnson *et al.* (30). Using the software package PHYLUCE v1.5 (36,38), we aligned our UCE probe sequences to each genome and then sliced out matching sequence along with 400 bp of flanking DNA on either side (*i.e.*, 180 bp target plus 800 bp total flanking sequence). We then used the resulting UCE "contigs" for input into the downstream bioinformatics and matrix assembly steps.

Analytical Details for Phylogenomic Inference

We investigated the tradeoff between taxon occupancy and locus occupancy (=missing data) in order to select a set of loci to be used for all remaining analyses. Using PHYLUCE, we filtered the entire set of trimmed alignments for different amounts of

515 taxon completeness (% of taxa that must be included in a given alignment for it to be
516 retained). This resulted in six locus sets filtered at a taxon threshold of 0, 25, 50, 75, 90,
517 and 95% taxon completeness. To evaluate these locus sets we generated concatenated
518 matrices and inferred maximum likelihood trees in RAXML v8.0.3 (70) (best tree search
519 plus 100 rapid bootstrap replicates, GTR+Γ model of sequence evolution). We selected
520 the best locus set by considering matrix completeness (more complete is better),
521 topological consistency, and bootstrap support values (higher support is better). Using
522 these criteria, we selected the 75% filtered set of alignments as the primary locus set for
523 all subsequent analyses (electronic supporting information S1, Table 5; and S2, Figs 7-
524 11).
525
526 All maximum likelihood (ML) analyses were performed using the best-tree plus rapid
527 bootstrapping search ("-f a" option) in RAXML with 200 bootstrap reps for the kmeans
528 analysis and 100 for all others. We used the GTR+Γ model of sequence evolution for all
529 analyses (best tree and bootstrap searches). For the partitioned-Bayesian inference (BI)
530 search, we executed two independent runs, each with four coupled chains (one cold and
531 three heated chains). We linked branch lengths across partitions, and we ran each
532 partitioned search for one million generations. We assessed burn-in, convergence among
533 runs, and run performance by examining parameter files with the program TRACER
534 v1.6.0 (71). We computed consensus trees using the *consense* utility, which comes as part
535 of EXABAYES.
536

537 To carry out the weighted statistical binning ASTRAL analysis, we input all gene trees
538 into the statistical binning pipeline using a support threshold of 75 (recommend for data
539 sets with < 1000 loci). This grouped genes into 103 bins, comprising 73 bins of 8 loci and
540 30 bins of 9 loci. After binning we concatenated the genes into supergenes and used
541 RAXML to infer supergene trees with bootstrap support (200 reps). We then input the
542 resulting best trees, weighted by gene number, and the bootstrap trees, into ASTRAL and
543 conducted a species tree analysis with 100 multi-locus bootstrap replicates (47).
544
545 For each taxon sampling experiment, we realigned the data after removing taxa, filtered
546 alignments with GBLOCKS, filtered alignments for taxon completeness (using a 75%
547 threshold), and generated a new concatenated matrix. We then analyzed each matrix in
548 RAxML using a best tree plus rapid bootstrap search (100 replicates) with GTR+ Γ as the
549 model of sequence evolution.
550
551 As the input topology for the BEAST analyses, we used the best tree generated from the
552 kmeans partitioned RAXML search of all loci. For each analysis, we concatenated the
553 loci and analyzed the matrix without partitioning. We performed a total of four
554 independent runs per analysis in BEAST, with each run progressing for 200 million
555 generations, sampling every 1,000 generations. We also performed one search with the
556 data removed so that the MCMC sampled from the prior distribution only. For the clock
557 and substitution models, we selected uncorrelated lognormal and GTR+Γ, respectively.
558 For the tree prior, we used a birth-death model, and for the ucld.mean prior, we used an

559 exponential distribution with the mean set to 1.0 and the initial value set to 0.003
560 (determined empirically from preliminary runs).
561
562 **Acknowledgements**
563
564 We would like to thank Dave Smith for donating specimens. We thank Jeffrey Sosa-
565 Calvo, Ana Jesovnik, and Mike Lloyd for assistance with lab work. For sequencing we
566 thank Joe DeYoung at the UCLA Neurosciences Genomics Core and Peter Schweitzer at
567 the Cornell Genomics Facility. Lab work for this study was conducted at the Smithsonian
568 NMNH Laboratory of Analytical Biology (LAB) and phylogenetic analyses were
569 performed using the Smithsonian's High-Performance Computer Cluster (Hydra) and the
570 CIPRES Science Gateway. We thank X anonymous reviewers for helpful suggestions to
571 the manuscript. Mention of trade names or commercial products in this publication is
572 solely for the purpose of providing specific information and does not imply
573 recommendation or endorsement by the USDA. USDA is an equal opportunity provider
574 and employer.
575
576 **References**
577
578 1. Nabhan AR, Sarkar IN. The impact of taxon sampling on phylogenetic inference: a
579 review of two decades of controversy. Brief Bioinform. 2012;13(1):122–134. doi:
580 10.1093/bib/bbr014
581 2. Townsend JP, Lopez-Giraldez F. Optimal selection of gene and ingroup taxon

582 sampling for resolving phylogenetic relationships. Syst Biol. 2010;59(4):446–457.
583 doi: 10.1093/sysbio/syq025
584 3. Heath TA, Hedtke SM, Hillis DM. Taxon sampling and the accuracy of
585 phylogenetic analyses. J Syst Evol. 2008;46(3):239–257. doi:
586 10.3724/SP.J.1002.2008.08016
587 4. Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic
588 error. Syst Biol. 2002;51(4):588–98. doi: 10.1080/10635150290102339
589 5. Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. Is sparse taxon sampling a
590 problem for phylogenetic inference? Syst Biol. 2003;52(1):124-126. doi:
591 10.1080/10635150390132911
592 6. Pollock DD, Zwickl DJ, Mcguire JA, Hillis DM. Increased taxon sampling is
593 advantageous for phylogenetic inference. Syst Biol. 2002;51(4):664–671. doi:
594 10.1080/10635150290102357
595 7. Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, et al.
596 Improved phylogenomic taxon sampling noticeably affects nonbilaterian
597 relationships. Mol Biol Evol. 2010;27(9):1983–1987. doi:
598 10.1093/molbev/msq089
599 8. Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for
600 phylogenetic inference. Proc Natl Acad Sci. 2001;98(19):10751–10756. doi:
601 10.1073/pnas.191248498
602 9. Graybeal A. Is it better to add taxa or characters to a difficult phylogenetic
603 problem? Syst Biol. 1998;47(1):9–17. doi: 10.1080/106351598260996
604 10. Jansen RK, Kaittanis C, Saski C, Lee S-B, Tomkins J, Alverson AJ, et al.

Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol. 2006;6:32. doi: 10.1186/1471-2148-6-32

11. Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, et al. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. 2015;526:569-573. doi: 10.1038/nature15697

12. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2013;66(2):526–538. doi: 10.1016/j.ympev.2011.12.007

13. Lemmon EM, Lemmon AR. High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Evol Syst. 2013;44(1):19.1–19.23. doi: 10.1146/annurev-ecolsys-110512-135822

14. Grimaldi D, Engel MS. Evolution of the Insects. New York: Cambridge University Press; 2005.

15. Sharkey MJ. Phylogeny and classification of Hymenoptera. Zootaxa. 2007;548:521–48.

16. Stork NE. Measuring global biodiversity and its decline. In: Reaka-Kudla ML, Wilson DE, Wilson EO, editors. Biodiversity II: Understanding and Protecting Our Biological Resources. Washington, D.C.: Joseph Henry Press; 1996. p. 41–68.

17. Hölldobler B, Wilson EO. The Ants. Cambridge: Belknap Press; 1990.

18. Michener CD. The Bees of the World. 2nd ed. Baltimore: The Johns Hopkins University Press; 2007.

628 19. Hunt JH. The Evolution of Social Wasps. New York: Oxford University Press;
629 2007.
630 20. Bradley TJ, Briscoe AD, Brady SG, Contreras HL, Danforth BN, Dudley R, et al.
631 Episodes in insect evolution. Integr Comp Biol. 2009;49(5):590–606. doi:
632 10.1093/icb/icp043
633 21. Romiguier J, Cameron SA, Woodard SH, Fischman BJ, Keller L, Praz CJ.
634 Phylogenomics controlling for base compositional bias reveals a single origin of
635 eusociality in corbiculate bees. Mol Biol Evol. 2015;33(3):670–678.
636 22. Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D. The impact of
637 molecular data on our understanding of bee phylogeny and evolution. Annu Rev
638 Entomol. 2013;58(1):57–78. doi: 10.1093/icb/icp043
639 23. Schwarz MP, Bull NJ, Cooper SJB. Molecular phylogenetics of allodapine bees,
640 with implications for the evolution of sociality and progressive rearing. Syst Biol.
641 2003;52(1):1–14. doi: 10.1080/10635150390132632
642 24. Gibbs J, Brady SG, Kanda K, Danforth BN. Phylogeny of halictine bees supports a
643 shared origin of eusociality for *Halictus* and *Lasioglossum* (Apoidea: Anthophila:
644 Halictidae). Mol Phylogenet Evol. 2012;65(3):926–39. doi:
645 10.1016/j.ympev.2012.08.013
646 25. Hines HM, Hunt JH, O'Connor TK, Gillespie JJ, Cameron SA. Multigene
647 phylogeny reveals eusociality evolved twice in vespid wasps. Proc Natl Acad Sci.
648 2007;104(9):3295–3299. doi: 10.1073/pnas.0610140104
649 26. O'Neil KM. Solitary Wasps: Behavior and Natural History. Ithaca: Cornell
650 University Press; 2001.

651 27. Evans HE. Predatory wasps. Sci Am. 1963;208(4):144–55.

652 28. Pilgrim EM, von Dohlen CD, Pitts JP. Molecular phylogenetics of Vespoidea
653 indicate paraphyly of the superfamily and novel relationships of its component
654 families and subfamilies. Zool Scr. 2008;37(5):539–60. doi: 10.1111/j.1463-
655 6409.2008.00340.x

656 29. Debevec AH, Cardinal S, Danforth BN. Identifying the sister group to the bees: a
657 molecular phylogeny of Aculeata with an emphasis on the superfamily Apoidea.
658 Zool Scr. 2012;41(5):527–535. doi: 10.1111/j.1463-6409.2012.00549.x

659 30. Johnson BR, Borowiec ML, Chiu JC, Lee EK, Atallah J, Ward PS. Phylogenomics
660 resolves evolutionary relationships among ants, bees, and wasps. Curr Biol.
661 2013;23:1–5. doi: 10.1016/j.cub.2013.08.050

662 31. Faircloth BC, Branstetter MG, White ND, Brady SG. Target enrichment of
663 ultraconserved elements from arthropods provides a genomic perspective on
664 relationships among Hymenoptera. Mol Ecol Resour. 2015;15:489–501. doi:
665 10.1111/1755-0998.12328

666 32. Heraty J, Ronquist F, Carpenter JM, Hawks D, Schulmeister S, Dowling AP, et al.
667 Evolution of the hymenopteran megaradiation. Mol Phylogenet Evol.
668 2011;60(1):73–88. doi: 10.1016/j.ympev.2011.04.003

669 33. Sharkey MJ, Carpenter JM, Vilhelmsen L, Heraty J, Liljeblad J, Dowling APG, et
670 al. Phylogenetic relationships among superfamilies of Hymenoptera. Cladistics.
671 2012;28(1):80–112. doi: 10.1111/j.1096-0031.2011.00366.x

672 34. Klopfstein S, Vilhelmsen L, Heraty JM, Sharkey M, Ronquist F. The
673 hymenopteran tree of life: evidence from protein-coding genes and objectively

aligned ribosomal data. PLoS One. 2013;8(8):e69344. doi: 10.1371/journal.pone.0069344

35. Mao M, Gibson T, Dowton M. Higher-level phylogeny of the Hymenoptera inferred from mitochondrial genomes. Mol Phylogenet Evol. 2015;84:34–43. doi: 10.1016/j.ympev.2014.12.009

36. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 2012;61(5):717-726. doi: 10.1093/sysbio/sys004

37. Castro LR, Dowton M. Molecular analyses of the Apocrita (Insecta: Hymenoptera) suggest that the Chalcidoidea are sister to the diaprioid complex. Invertebr Syst. 2006;20(5):603–14. doi: 10.1071/IS06002

38. Faircloth BC. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics. 2015:Advance Access:1–3. doi: 10.1093/bioinformatics/btv646

39. Aberer AJ, Kobert K, Stamatakis A. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol Biol Evol. 2014;31(10):2553–2556. doi: 10.1093/molbev/msu236

40. Frandsen PB, Calcott B, Mayer C, Lanfear R. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol Biol. 2015;15:13. doi: 10.1186/s12862-015-0283-7

41. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics.

2014;30:i541–i548. doi: 10.1093/bioinformatics/btu462

42. Team RC. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015. Available from: https://www.r-project.org/

43. Meiklejohn KA, Faircloth BC, Glenn, Travis C, Kimball RT, Braun EL. Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): evidence for a bias in some multispecies coalescent methods. Syst Biol. 2016;65(4):612-627. doi: 10.1093/sysbio/syw014

44. Hosner PA, Faircloth BC, Glenn TC, Braun EL, Kimball RT. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol Biol Evol. 2015;33(4):1110–1125. doi: doi:10.1093/molbev/msv347

45. Manthey JD. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: *Piranga*). Syst Biol. 2016;65(4):640–650. doi: 10.1017/CBO9781107415324.004

46. Bayzid MS, Mirarab S, Boussau B, Warnow T. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS One. 2015;10(6):e0129183. doi: 10.1371/journal.pone.0129183

47. Seo TK. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol Biol Evol. 2008;25(5):960–971. doi: 10.1093/molbev/msn043

48. Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. 2015;16:987. doi: 10.1186/s12864-015-2146-4

720 49. Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D. The impact of
721 molecular data on our understanding of bee phylogeny and evolution. Annu Rev
722 Entomol. 2013;58:57–78. doi: 10.1146/annurev-ento-120811-153633
723 50. Cardinal S, Straka J, Danforth BN. Comprehensive phylogeny of apid bees reveals
724 the evolutionary origins and antiquity of cleptoparasitism. Proc Natl Acad Sci.
725 2010;107(37):16207–16211. doi: 10.1073/pnas.1006299107
726 51. Rehan SM, Leys R, Schwarz MP. First evidence for a massive extinction event
727 affecting bees close to the KT boundary. PLoS One. 2013;8(10):e76683. doi:
728 10.1371/journal.pone.0076683
729 52. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with
730 BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73. doi:
731 10.1093/molbev/mss075
732 53. Zhang C, Stadler T, Klopfstein S, Heath TA, Ronquist F. Total-evidence dating
733 under the fossilized birth-death process. Syst Biol. 2015;65(2)228-249. doi:
734 10.1093/sysbio/syv080
735 54. Alroy J. Fossilworks. Gateway to the paleobiology database. 2015. Available:
736 www.fossilworks.org.
737 55. McCormack JE, Faircloth BC. Next-generation phylogenetics takes root. Mol
738 Ecol. 2013;22:19–21. doi: 10.1111/mec.12050
739 56. Brothres DJ, Carpenter JM. Phylogeny of Aculeata: Chrysidoidea and Vespoidea
740 (Hymenoptera). J Hymenopt Res. 1993;2(1):227–304.
741 57. Ohl M, Bleidorn C. The phylogenetic position of the enigmatic wasp family
742 Heterogynaidae based on molecular data, with description of a new, nocturnal

743 species (Hymenoptera: Apoidea). Syst Entomol. 2005;31(2):321–337. doi:
744 10.1111/j.1365-3113.2005.00313.x
745 58. Roig-Alsina A, Michener CD. Studies of the phylogeny and classification of long-
746 tounged bees (Hymenoptera: Apoidea). Univ Kansas Sci Bull. 1993;55(13):124–
747 62.
748 59. Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. Less is more in
749 mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the
750 root of placental mammals. Mol Biol Evol. 2013;30(9):2134–44. doi:
751 10.1093/molbev/mst116
752 60. Faircloth BC, Glenn TC. Not all sequence tags are created equal: designing and
753 validating sequence identification tags robust to indels. PLoS One.
754 2012;7(8):e42543. doi: 10.1371/journal.pone.0042543
755 61. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries
756 for multiplexed target capture. Genome Res. 2012;22(5):939–946. doi:
757 10.1101/gr.128124.111
758 62. Faircloth BC. Illumiprocessor: a trimmomatic wrapper for parallel adapter and
759 quality trimming. 2013. Available: http://dx.doi.org/10.6079/J9ILL.
760 63. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, et al. RobiNA: A
761 user-friendly, integrated software solution for RNA-Seq-based transcriptomics.
762 Nucleic Acids Res. 2012;40(W1):W622–W627. doi: 10.1093/nar/gks540
763 64. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of
764 read trimming effects on illumina NGS data analysis. PLoS One.
765 2013;8(12):e85024. doi: 1. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM.

An extensive evaluation of read trimming effects on illumina NGS data analysis. PLoS One. 2013;8(12):e85024.

65. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–652. doi: 10.1038/nbt.1883

66. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–3066. doi: 10.1093/nar/gkf436

67. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–552. doi: 10.1093/oxfordjournals.molbev.a026334

68. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56(4):564–77. doi: 10.1080/10635150701472164

69. Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, et al. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Res. 2011;39(Database issue):D658–D662. doi: 10.1093/nar/gkq1145

70. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–1313. doi: 10.1093/bioinformatics/btu033

71. Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1.6, Available: http://beast.bio.ed.ac.uk/Tracer. 2014.

790 **Figure Captions**

791 **Fig 1.** Dated phylogeny of Hymenoptera. We inferred the topology by analyzing the
792 *Hym-187T-F75* matrix in RAxML (partitioned by kmeans algorithm; 854 loci; 203,095
793 bp of sequence data) and estimated the dates in BEAST (50 random loci; fixed topology;
794 38 calibration points). Black dots indicate nodes that received < 100% bootstrap support
795 in the ML analysis.

796

797 **Fig 2.** Alternative hypotheses for relationships among aculeate superfamilies. (A)
798 Topology from Johnson *et al.* (30). (B) Topology from Faircloth *et al.* (31). (C) Topology
799 from the *Faircltoh-61T* matrix analyzed in this study. (D) Preferred topology inferred in
800 this study (includes Sierolomorphoidea). Topologies correspond to those reported in
801 Table 1, except that topologies A and D are equivalent in terms of ants being sister to
802 Apoidea.

Tables

Table 1. Results of the taxon inclusion/exclusion experiments as evidenced by topological and bootstrap support differences. The results suggest that both outgroup choice (chrysidoid presence/absence) and taxon evenness are important. The matrix name indicates whether the taxon set is a version of Johnson *et al.*(30), Faircloth *et al.* (31), or this study ("Hym"). Three different topologies were recovered: (A) ants sister to Apoidea; (B) ants sister to all other aculeate superfamilies, except Chrysidoidea; and (C) ants sister to Apoidea+Scoliodea. Bootstrap support indicates support for the clade that includes ants plus its sister group. Topologies correspond to those shown in Figure 1A-C, with regard to the position of ants.

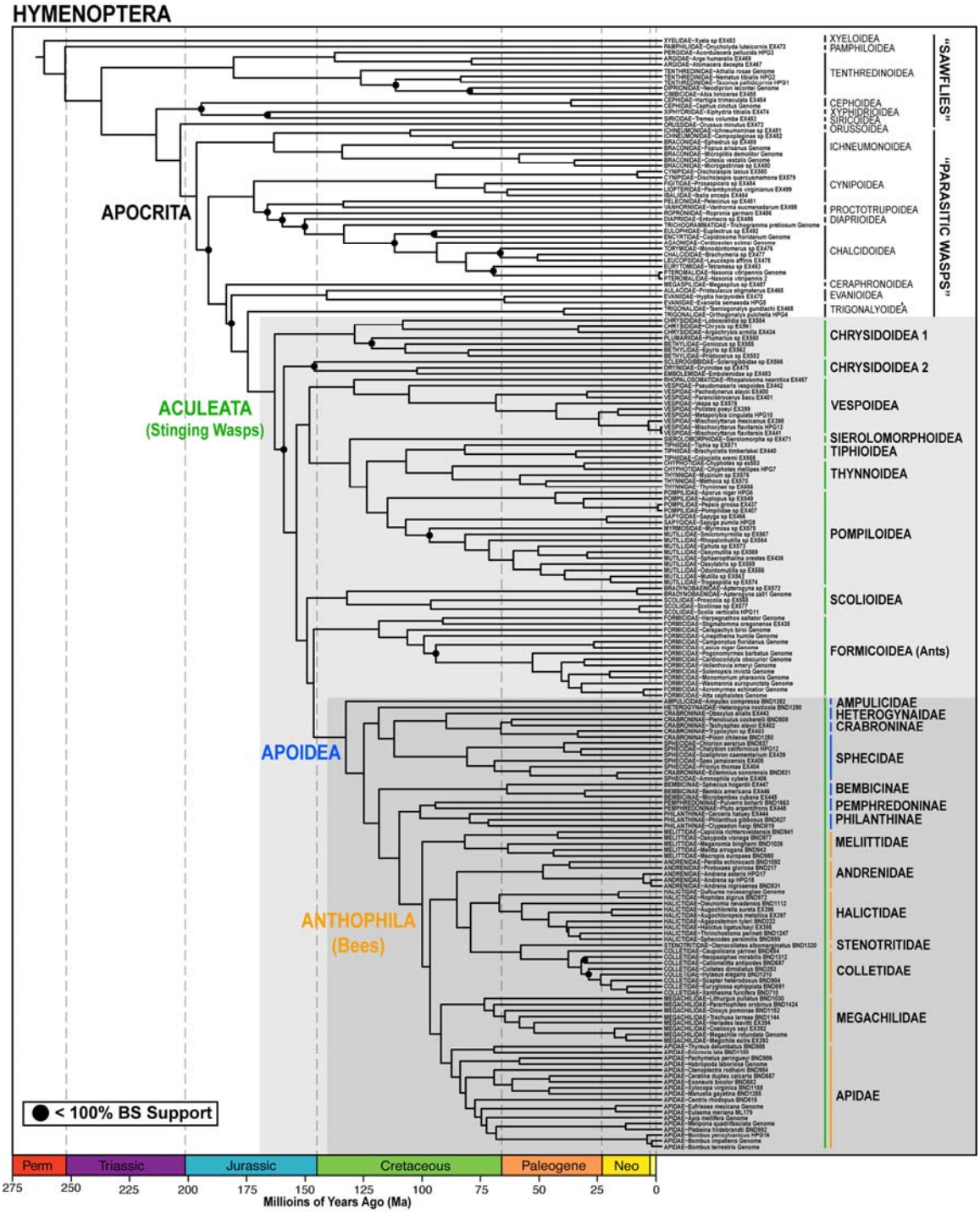
Matrix Name	Topology	BS Support (Ants+Sister Group)	Outgroup	Note
Johnson-18T	A	89	No chrysidoid	
Johnson-19T	A	100		Same taxon set as in (30).
Faircloth-26T	A	88	No chrysidoid	
Faircloth-27T	A	97		
Faircloth-45T	B	100	No chrysidoid	Same taxon set as in (31).
Faircloth-46T	B	99		
Faircloth-52T	B	98		
Faircloth-56T	B	100		
Faircloth-61T	C	100		
Hym-100T	A	100		Most balanced taxon set.
Hym-131T	A	90	No chrysidoid or trigonaloid	
Hym-133T	A	100	No chrysidoid	
Hym-147T	A	100		
Hym-187T-F75	A	100		This study.

Table 2. Divergence dates for key nodes (estimated with BEAST) comparing the 25 and 50 best loci (best equals loci with highest average gene-tree support values), and 50 randomly selected loci. Dates are given as median ages in millions of years ago (Ma), with the 95% highest posterior density given in parentheses.

Select Clades (Crown Group)	25 Best Loci (Ma)	50 Best Loci (Ma)	50 Random Loci (Ma)
Hymenoptera	255 (238-272)	256 (239-273)	257 (240-274)
Euhymenoptera	200 (187-215)	198 (186-213)	200 (187-216)
Apocrita	193 (180-206)	192 (180-205)	194 (181-208)
Aculeata (stinging Hymenoptera)	162 (155-170)	162 (155-169)	161 (154-169)
Apoidea+Formicoidea	144 (143-148)	145 (143-148)	145 (143-148)
Formicidae (ants, w/o Leptanillinae/Martialinae)	119 (109-128)	118 (110-126)	118 (108-128)
Apoidea (apoid wasps+bees)	136 (127-144)	134 (123-142)	131 (121-141)
Anthophila (bees)	102 (95-111)	102 (94-111)	100 (92-107)

Figures

Fig 1.



820 **Fig 2.**

821