The history of early bee diversification based on five genes plus morphology

Bryan N. Danforth†, Sedonia Sipes‡, Jennifer Fang§, and Seán G. Brady¶

Department of Entomology, 3119 Comstock Hall, Cornell University, Ithaca, NY 14853

Communicated by Charles D. Michener, University of Kansas, Lawrence, KS, May 16, 2006 (received for review February 2, 2006)

Bees, the largest (>16,000 species) and most important radiation of pollinating insects, originated in early to mid-Cretaceous, roughly in synchrony with the angiosperms (flowering plants). Understanding the diversification of the bees and the coevolutionary history of bees and angiosperms requires a well supported phylogeny of bees (as well as angiosperms). We reconstructed a robust phylogeny of bees at the family and subfamily levels using a data set of five genes (4,299 nucleotide sites) plus morphology (109 characters). The molecular data set included protein coding (elongation factor-1α, RNA polymerase II, and LW rhodopsin), as well as ribosomal (28S and 18S) nuclear gene data. Analyses of both the DNA data set and the DNA+morphology data set by parsimony and Bayesian methods yielded a single well supported family-level tree topology that places Melittidae as a paraphyletic group at the base of the phylogeny of bees. This topology (“Melittidae-LT basal”) is significantly better than a previously proposed alternative topology (“Colletidae basal”) based both on likelihood and Bayesian methods. Our results have important implications for understanding the early diversification, historical biogeography, host-plant evolution, and fossil record of bees. The earliest branches of bee phylogeny include lineages that are predominantly host-plant specialists, suggesting that host-plant specificity is an ancestral trait in bees. Our results suggest an African origin for bees, because the earliest branches of the tree include predominantly African lineages. These results also help explain the predominance of Melittidae, Apidae, and Megachilidae among the earliest fossil bees.

---

**Angiosperms** (flowering plants), with an estimated 250,000–260,000 species (1), represent the largest and most diverse lineage of vascular plants on earth. To Darwin, the rapid emergence and early diversification of the angiosperms was an “abominable mystery” (ref. 2 and refs. therein). Among the most important traits attributable to the explosive radiation of the angiosperms is animal-mediated pollination (3–7). Insects are by far the most important animal pollinators (~70% of angiosperm species are insect pollinated; ref. 8) and among insects, bees are the most specialized and important pollinator group. All of the >16,000 species of bees living today (9) rely virtually exclusively on angiosperm products, including pollen and nectar for adult and larval nutrition (10), floral oils for larval nutrition (11, 12), floral waxes and perfumes that serve as sexual attractants (13), and resins for nest construction (14). Bees are morphologically adapted to collecting, manipulating, carrying, and storing pollen and other plant products (15, 16), and many bee species are specialists on one or a few closely related host plants (10).

One step toward resolving Darwin’s “abominable mystery” is to develop a better understanding of the role that bees played in the evolutionary history and diversification of the angiosperms. A robust phylogeny of bees would allow us to infer attributes of the early bees and to reconstruct the types of interactions that existed between the earliest bees and their angiosperm hosts. Higher-level (family- and subfamily-level) bee phylogeny is poorly understood. Currently, bees are divided into seven extant families: the long-tongued (LT) bee families Megachilidae and Apidae and the short-tongued (ST) bee families Colletidae, Stenotritidae, Andrenidae, Halictidae, and Melittidae sensu lato (s.l.) (9). Colletidae is widely considered the most basal family of bees (i.e., the sister group to the rest of the bees), because all females and most males possess a glossa (tongue) with a bifid (forked) apex, much like the glossa of an apoid wasp (18–22).

However, several authors have questioned this interpretation (9, 23–27) and have hypothesized that the earliest branches of bee phylogeny may have been either Melittidae s.l., LT bees, or a monophyletic group consisting of both. The most recent morphological analysis of family-level phylogeny in bees (17) obtained two different tree topologies based on alternative coding of relatively few mouthpart characters. One tree topology places Colletidae as sister to the rest of the bees (“Colletidae basal”), whereas the other places Melittidae s.l. + LT bees as sister to the rest of the bees (“Melittidae-LT basal”). The major difference between the Colletidae basal and Melittidae-LT basal topologies involves the placement of the root node of bees (27). Placing the root between Colletidae and the rest of the bees yields the Colletidae basal topology, whereas placement of the root node near or within Melittidae s.l. yields the Melittidae-LT basal topology. The biological implications of these alternative topologies are radically different. The Colletidae basal topology implies an Australian and/or South American origin for bees and suggests the earliest bees were a mix of floral generalists and specialists. Melittidae-LT basal implies an African origin for bees and indicates that the earliest bees were likely to have been floral specialists. These alternative topologies also have implications for understanding the fossil record and antiquity of bees.

To resolve the root node of bees, we combined >4,000 bp of DNA sequence data with the previous morphological data set of Alexander and Michener (17). We report here the results of an analysis of the largest molecular and morphological study to date on bee phylogeny, bee evolution, molecular evolution, molecular systematics, and coevolution.

---

**Author contributions:** B.N.D. and S.S. designed research; B.N.D., S.S., J.F., and S.G.B. performed research; S.G.B. contributed new reagents/analytical tools; B.N.D. analyzed data; and B.N.D. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

**Abbreviations:** ST, short-tongued; LT, long-tongued; GTR, general time-reversible; I–G, gamma distribution plus a proportion of invariant sites; s.l., sensu lato; s.s., sensu stricto.

**Data deposition:** The sequences reported in this paper have been deposited in the GenBank database (accession nos. listed in Table 4, which is published as supporting information on the PNAS web site). The data matrix has been deposited in the TreeBASE database, www.treebase.org (accession nos. M2878 and S1599).

1†To whom correspondence should be addressed. E-mail: bnd1@cornell.edu.

2‡Present address: Department of Plant Biology, Life Science II 420, Mailcode 6509, Southern Illinois University, Carbondale, IL 62901-6509.

3§Present address: Department of Physiology, P.O. Box 245051, University of Arizona, Tucson, AZ 85724-5051.

4¶Present address: Department of Entomology and Laboratories of Analytical Biology, Smithsonian Institution, Sullivant, MD 20746.

5Melittidae in the sense of Michener (9) is a paraphyletic group based on our results. We refer to the three melittid subfamilies as families, following an earlier suggestion by Alexander and Michener (17). The three families are Melittidae (s.s.), Dasypodaidae, and Meganomiidae.

© 2006 by The National Academy of Sciences of the USA

15118–15123 | PNAS | October 10, 2006 | vol. 103 | no. 41 | www.pnas.org/cgi/doi/10.1073/pnas.0604033103
bee family-level phylogeny. Our analyses provide insights into the phylogeny, historical biogeography, host–plant associations, and fossil record of bees.

**Results**

Our data set consisted of a total of 4,299 total aligned nucleotide sites (1,648 parsimony informative sites; 1,141 total aligned nucleotide sites, 438 parsimony informative sites); RNA polymerase II (889 total aligned nucleotide sites, 300 parsimony informative sites); LW rhodopsin (781 total aligned nucleotide sites, 448 parsimony informative sites); 28S rDNA (772 total aligned nucleotide sites, 362 parsimony informative sites); 18S rDNA (716 total aligned nucleotide sites, 362 parsimony informative sites); and 18S rDNA (772 total aligned nucleotide sites, 109 morphological characters).
null
Figure 2. Bayesian analysis of five genes combined. Posterior probabilities for the GTR+SYM[18s]+I+G model (the preferred model) are shown above the branches, and posterior probabilities for the GTR+I+G model are shown below the branches. Bayesian posterior probabilities are based on the last 8,000 trees of each analysis. Families are color-coded. The clade united by the presence of a hemicryptic midcoxa is indicated by the black dot.

Discussion
This conclusion is consistent with the maximum likelihood results (above).

This study establishes phylogenetic relationships among bee families and subfamilies with high levels of support. Our results unambiguously support the Melittidae-LT basal topology (Figs. 1 and 2; see Fig. 3, which is published as supporting information on the PNAS web site). The Colletidae basal topology is more widely accepted because of the perception that the bifid glossa of Colletidae is a plesiomorphic trait shared with apoid wasps. This hypothesis appears in numerous publications but is rarely supported by characters other than the overall appearance of the glossa (18–22). Although a number of morphological studies have questioned the
Colletidae basal hypothesis (23–26), the morphological support for the Melittidae-LT basal hypothesis has been largely overlooked in the bee phylogenetic literature. Among the most convincing morphological characters that support the tree presented herein is the morphology of the midcoxa. Michener (32) discovered that in apoid wasps, Melittidae s.l. and LT bees, the midcoxa is exposed, whereas in the remaining ST bee families, the upper portion of the midcoxa is internal and hidden beneath the mesopleuron (a condition described as “hemicycrtic”). The hemicycrtic condition is a unique and unreversed character congruent with monophyly of Andrenidae, Halictidae, Colletidae, and Stenotritidae (Figs. 1 and 2), thus strongly supporting the Melittidae-LT basal topology.

Implications for Bee Historical Biogeography and Host–Plant Evolution. Melittidae-LT basal topology substantially alters prevailing hypotheses of bee phylogeny, biogeography, evolution, and early diversification. The hypothesis that Melittidae s.l. represents the earliest branch(es) of bee phylogeny suggests an African rather than an Australian or South American origin for the bees. Melittidae s.l. is absent from Australia and South America, and Africa is the only continent where all major lineages (e.g., families) of Melittidae s.l. occur (20). Meganomiidae are restricted to Africa, and for Dasy- podidae and Melittidae sensu stricto (s.s.), Africa is the continent that hosts the greatest number of genera and species. Based on this distribution, Michener (20) hypothesized an African origin for all families of Melittidae s.l. Given their placement in our phylogenetic trees, this would suggest an African origin for bees in general. Disjunct biogeographic distributions in some “melittid” genera, such as Hesperapis (in the Dasyopodidae; ref. 9), provide further support for the antiquity of this group.

The placement of a paraphyletic Melittidae s.l. at the base of the phylogeny also supports the view that the earliest bees were narrow host–plant specialists. Host–plant specialization is widespread among the melittids as well as among many basal lineages in other families, including Rophitinae (Halictidae), Andreninae and Pan- urginae (Andrenidae), Colletinae (Colletidae), and Fidellinae (Megachilidae). Most species in Melittidae s.s. and Dasyopodidae are well known to be host–plant specialists. Female Hesperapis oraria, for example, are monolectic and forage for pollen exclusively on Balduina angustifolia (Asteraceae; ref. 33). Female H. trochanterata forage exclusively on plants in the genus Nama (Boraginaceae) and have elongate slender heads with specialized hairs on the mouthparts for extracting pollen from the tubular flowers (34).

Other species of Hesperapis specialize on a small number of closely related host–plant species within diverse angiosperm families, including Polemoniaceae, Rosaceae, Zygophyllaceae, Onagraceae, Papaveraceae, Fabaceae, and Malvaceae. Host–plant specialization is widespread within Melittidae s.s., including Melitta, Rediviva, Redivivoides, and Macropis. All species of Macropis are narrow host–plant specialists on oil-producing plants in the genus Lysimachia (Primulaceae; ref. 35), and all species possess modified legs for collecting and manipulating viscous floral oils. Species of Rediviva are involved in an intense host–plant association with plants in the oil-producing genus Diascia (Scrophulariaceae), in which variation in floral spur length in the host plants is paralleled by variation and extreme exaggeration in foreleg length in bees (36, 37). Given the placement of Melittidae s.l. as a paraphyletic group at the base of the tree, our results indicate that host–plant specialization is the primitive state for bees.

Implications for Understanding the Bee Fossil Record. The Melittidae-LT basal hypothesis may help explain the chronological appearance of bee families in the fossil record. If the Colletidae basal topology were indeed correct, one of the most puzzling aspects of the bee fossil record would be the abundance of Melittidae s.l., Apidae, and Megachilidae in the oldest deposits, such as Eocene (Baltic) amber (22, 38) and Cretaceous amber from New Jersey (39–41). Among the bees in Baltic amber deposits, 15 of 18 described genera are LT bees (Apidae and Megachilidae; ref. 22). Melittid bees are also well represented in the Eocene both from Baltic amber (Eomacropis; ref. 22) and French Eocene amber (Paleomacropis; ref. 38). The oldest fossil bee, Cretotrigona priscia, is an apid bee closely related to extant stingless bees (Meliponini; refs. 39–41). In contrast, ST bee families, such as Halictidae, are much less well represented in the Eocene, and representatives of Andrenidae and Colletidae are completely absent in the fossil record up until the Miocene (42, 43). The high proportion of Melittidae s.l. and LT bees in the Eocene fossil deposits has generally been interpreted as an artifact because of the poor fossil record of bees and possibly a bias toward resin collecting bees, most of which are LT bees (44). However, if one accepts the Melittidae-LT basal hypothesis, Melittidae s.l., Megachilidae, and Apidae represent early branches in the phylogeny of the bees and are therefore relatively old compared with some families of ST bees, such as Halictidae, Colletidae, and Stenotritidae.

Materials and Methods

Data Sets Analyzed. Molecular data. We generated a data set based on five nuclear genes that have previously shown promise for resolving deep divergences in insects and other arthropods: elongation factor-1a (45), RNA polymerase II (27), LW rhodopsin (46), 28S rDNA (47), and 18S rDNA (48). PCR and sequencing protocols followed standard methods detailed in Danforth et al. (27, 49, 50). PCR products were gel-purified overnight on low-melting-point agarose gels, and bands were extracted by using the Promega Wizard PCR purification system (Promega, Madison, WI). All PCR products were sequenced in both directions. Sequencing was performed by using an Applied Biosystems (Surrey, U.K.) Automated 3730 DNA Analyzer. We used Big Dye Terminator chemistry and AmpliTag-FS DNA polymerase.

Morphological data. We obtained 109 morphological characters from a previous study of family-level phylogenetic relationships in bees (17). Characters were treated as unordered and of equal weight. Coding of characters followed the Series I codings of Alexander and Michener (17). This is the coding method that supports the Colletidae basal topology.

Phylogenetic Methods and Taxon Sampling. We included a total of 94 species (14 apoid wasp outgroups and 80 bee ingroups; Table 4, which is published as supporting information on the PNAS web site) representing all seven families of bees and all 21 of the currently recognized subfamilies (9). Our taxon sampling was extensive and included representatives of three previously recognized bee families (Oxaeidae, Ctenoplectridae, and Fidellidae). We focused particular attention on sampling within the five ST bee families and in particular in the two families previously considered to be potentially the most basal lineages of bees: Colletidae and Melittidae s.l. Sampling within Melittidae s.l. was considered particularly important, because this family is not clearly monophyletic (17). The only melittid tribe lacking from our data set is Promelitini.

Outgroups included representatives of two of the four apoid wasp families, Crabronidae and Sphecididae (51). Voucher specimens are deposited in the Cornell University Insect Collection. Complete locality data, GenBank accession nos., and our combined data set are available Table 4. Our data set is deposited in TreeBASE (www.treebase.org/treebase/index.html) as submission nos. M2878 and S1599.

Alignments for all genes were generated in the Lasergene DNASTar (Madison, WI) software package using Clustal W. LW rhodopsin presented particular problems for sequencing as well as alignment because of pronounced variation in the lengths of introns I and III within Colletidae. Long introns in some subfamilies (e.g., Euryglossininae, Hylaeinae, and Xeromelissininae) required manual alignments or exclusion of introns. Alignments for the 28S D2–D4 region were adjusted by eye, and some unalignable regions were excluded from the analysis. Reading frames and intron/exon
boundaries were determined by comparison with sequences obtained for the honeybee, *Apis mellifera*.

**Parsimony methods.** We performed maximum parsimony analyses using PAUP*. Initially, we performed equal-weight parsimony analyses on each of the six data sets separately and then combined the data sets into a single analysis. Branch support for the individual data sets as well as the combined data set was estimated by using bootstrap analysis (53). For parsimony searches, we performed 500 random sequence additions. For calculating bootstrap proportions, we performed 500 replicates with 10 random sequence additions per replicate.

**Bayesian methods.** Analysis of individual gene partitions by MrModelTest Ver. 2.2 (54) indicated that the GTR+I+G model was the most appropriate for EF-1a, 28S, opsin, and pol II, and that the SYM+I+G model was the most appropriate for the 18S data set. The only difference between the two models is that GTR allows empirical base frequencies, whereas SYM assumes equal base frequencies (55). The 18S data showed little evidence of base-compositional bias (Table 2). We thus applied a model in which GTR+I+G was applied to EF-1a, 28S, opsin, and pol II, whereas a SYM+I+G model was applied to the 18S data set (referred to as the GTR+SYM(18S)+I+G model). In addition, we explored alternative models to evaluate the robustness of the data set to model choice.

For the Bayesian analyses, we used MrBayes Ver. 3.1.2 (refs. 55 and 57). We analyzed the combined data set using a range of models including the Kimura two-parameter, Hasegawa–Kishino–Yano, SYM (55), and GTR models (58). Various among-site rate variation models were used to account for rate variation among genes and codon positions, including gamma distribution (G) as well as I+G. For all analyses, we treated the separate genes as “unlinked,” so that separate parameter estimates were obtained for each gene for all runs. We ran two simultaneous runs with four chains each for 1 × 10^6 generations and sampled trees every 100 generations. Plots of the −ln likelihood scores over generation time showed that stable values have been obtained for the null hypothesis (Colletidae basal). Twice the difference in log likelihood can be interpreted by using tables in refs. 31 and 60.

**Hypothesis testing.** To compare our results with the previous family-level phylogenies of bees, we used both maximum likelihood and Bayesian methods. Maximum likelihood were determined by Rossing the honeybee, *Apis mellifera*.

We performed maximum likelihood analyses using PAUP*. Initially, we performed equal-weight parsimony analyses on each of the six data sets separately and then combined the data sets into a single analysis. Branch support for the individual data sets as well as the combined data set was estimated by using bootstrap analysis (53). For parsimony searches, we performed 500 random sequence additions. For calculating bootstrap proportions, we performed 500 replicates with 10 random sequence additions per replicate.

**Bayesian methods.** Analysis of individual gene partitions by MrModelTest Ver. 2.2 (54) indicated that the GTR+I+G model was the most appropriate for EF-1a, 28S, opsin, and pol II, and that the SYM+I+G model was the most appropriate for the 18S data set. The only difference between the two models is that GTR allows empirical base frequencies, whereas SYM assumes equal base frequencies (55). The 18S data showed little evidence of base-compositional bias (Table 2). We thus applied a model in which GTR+I+G was applied to EF-1a, 28S, opsin, and pol II, whereas a SYM+I+G model was applied to the 18S data set (referred to as the GTR+SYM(18S)+I+G model). In addition, we explored alternative models to evaluate the robustness of the data set to model choice.

For the Bayesian analyses, we used MrBayes Ver. 3.1.2 (refs. 55 and 57). We analyzed the combined data set using a range of models including the Kimura two-parameter, Hasegawa–Kishino–Yano, SYM (55), and GTR models (58). Various among-site rate variation models were used to account for rate variation among genes and codon positions, including gamma distribution (G) as well as I+G. For all analyses, we treated the separate genes as “unlinked,” so that separate parameter estimates were obtained for each gene for all runs. We ran two simultaneous runs with four chains each for 1 × 10^6 generations and sampled trees every 100 generations. Plots of the −ln likelihood scores over generation time showed that stable values have been obtained for the null hypothesis (Colletidae basal). Twice the difference in log likelihood can be interpreted by using tables in refs. 31 and 60.

**Hypothesis testing.** To compare our results with the previous family-level phylogenies of bees, we used both maximum likelihood and Bayesian methods. Maximum likelihood were determined by Rossing the honeybee, *Apis mellifera*.


(100,941),(991,998)