Author's personal copy

Molecular Phylogenetics and Evolution 60 (2011) 295–304

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Rooting phylogenies using gene duplications: An empirical example from the bees (Apoidea)

Seán G. Brady [a,*], Jessica R. Litman [b], Bryan N. Danforth [b]

[a] Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, United States
[b] Department of Entomology, Cornell University, Comstock Hall, Ithaca, NY 14853, United States

A R T I C L E I N F O

Article history:
Received 12 July 2010
Revised 26 April 2011
Accepted 3 May 2011
Available online 12 May 2011

Keywords:
Apoidea
Hymenoptera
Elongation factor-1 alpha
Molecular phylogeny
Paralogy
Rooting

A B S T R A C T

The placement of the root node in a phylogeny is fundamental to characterizing evolutionary relationships. The root node of bee phylogeny remains unclear despite considerable previous attention. In order to test alternative hypotheses for the location of the root node in bees, we used the F1 and F2 paralogs of elongation factor 1-alpha (EF-1α) to compare the tree topologies that result when using outgroup versus paralogous rooting. Fifty-two taxa representing each of the seven bee families were sequenced for both copies of EF-1α. Two datasets were analyzed. In the first (the "concatenated" dataset), the F1 and F2 copies for each species were concatenated and the tree was rooted using appropriate outgroups (sphecid and crabronid wasps). In the second dataset (the "duplicated" dataset), the F1 and F2 copies were aligned to each another and each copy for all taxa were treated as separate terminals. In this dataset, the root was placed between the F1 and F2 copies (e.g., paralog rooting). Bayesian analyses demonstrate that the outgroup rooting approach outperforms paralog rooting, recovering deeper clades and showing stronger support for groups well established by both morphological and other molecular data. Sequence characteristics of the two copies were compared at the amino acid level, but little evidence was found to suggest that one copy is more functionally conserved. Although neither approach yields an unambiguous root to the tree, both approaches strongly indicate that the root of bee phylogeny does not fall near Colletidae, as has been previously proposed. We discuss paralog rooting as a general strategy and why this approach performs relatively poorly with our particular dataset.

Published by Elsevier Inc.

1. Introduction

Determining the root of an unrooted network is critical to many applications of phylogenetics to other areas of biology. While an unrooted network reveals information about relationships among taxa, it reveals nothing about the order in which taxa diverged from one another. A rooted tree is necessary to make inferences regarding ancestry and character evolution over time. While a number of methods have been proposed for rooting trees, including molecular clock (Hasegawa et al., 1985), non-reversible models of DNA substitution (Huelsenbeck et al., 2002; Yang, 1994), gene paralogs (Baldauf et al., 1996; Brinkmann and Philippe, 1999; Hashimoto and Hasegawa, 1996; Brown and Doolittle, 1995; Donoghue and Mathews, 1998; Doolittle and Brown, 1994; Gogarten et al., 1989; Iwabe et al., 1989; Mathews et al., 2010; Mathews and Donoghue, 1999, 2000; Zhaxybayeva et al., 2005), and indel polarization (Lake et al., 2007, 2008, 2009; Servin et al., 2008; Skophammer et al., 2007; Valas and Bourne, 2009), the most common technique employed by molecular systematists is outgroup rooting (Maddison et al., 1984; Nixon and Carpenter, 1993; Wheeler, 1990). Outgroup rooting assumes that one or more taxa fall outside the remaining taxa (the ingroup) such that the ingroup root is defined by the branch connecting the ingroup to the outgroup. Although outgroup rooting is the principal method currently used to root molecular trees, this method suffers from potential drawbacks under some conditions. Most notably, widely divergent outgroups and tree topologies that reflect rapid radiations (i.e. a combination of both short and long branches) may render outgroup rooting ineffective (Kodandaramaiah et al., 2010; Shavit et al., 2007).

The analysis of paralogs presents an alternative method to root phylogenies. In this method, duplicated genes found in each of the ingroup taxa are sequenced and aligned. The resulting alignment is analyzed and rooted at the branch connecting the two paralogs. The resulting phylogeny consists of two sub-trees; each sub-tree represents taxonomic relationships based on a single paralog. This method has been suggested for clades that are highly divergent from all living relatives, where potential outgroups are very distantly related to the ingroup.

In this paper, we examine this relatively novel method for rooting bee phylogeny using duplicate copies of the nuclear gene

* Corresponding author. Address: Department of Entomology, MRC-188, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, United States.

E-mail addresses: bradys@si.edu (S.G. Brady), jrl72@cornell.edu (J.R. Litman), bnd1@cornell.edu (B.N. Danforth).

1055-7903/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.ympev.2011.05.002

Author's personal copy

elongation factor 1-alpha (EF-1α), which encodes a protein involved with ribosomal binding of tRNA (Maroni, 1993). EF-1α is present as two paralogous copies in several insect orders including Hymenoptera (Danforth and Ji, 1998), Diptera (Hovemann et al., 1988), Coleoptera (Jordal, 2002; Normark et al., 1999), Hemiptera (Downie and Gullan, 2004), Thysanoptera (Morris et al., 2002), and Neuropterida (Haring and Aspock, 2004). Within bees and other Hymenoptera, these paralogs – denoted as the F1 and F2 copies – are easily discriminated due to well-documented variation in the exon–intron structure between the two copies (Brady and Danforth, 2004; Danforth and Ji, 1998). Consistent variation between EF-1α paralogs in intron distribution also occurs in other insect taxa such as Diptera and Coleoptera (Djernaes and Damgaard, 2006; Simon et al., 2010).

Establishing the root to bee phylogeny remains contentious. Bees represent the largest and arguably the most important lineage of angiosperm pollinators on earth (Michener, 2007). There are an estimated 20,000 described bee species worldwide (Ascher et al.: http://www.itis.gov/beechecklist.html), but the actual number could be much higher. Bees are presumed to have arisen 100–140 million years ago, roughly coincident with the early evolution of the angiosperms (Brady et al., 2009; Danforth, 2007) and have likely played an important role in angiosperm diversification (Grimaldi, 1999; Michener, 2007). While higher-level (family and subfamily level) bee phylogeny has been examined based on morphology (Alexander and Michener, 1995; Roig-Alsina and Michener, 1993), molecular data (Danforth et al., 2006a), and a combination of morphological and molecular data (Danforth et al., 2006b), questions remain about how to root the phylogeny of bees. Morphological studies (Alexander and Michener, 1995) have obtained conflicting results, depending on the weighting scheme used to analyze the data as well as the coding of certain key morphological characters such as the bifid glossa of Colletidae (reviewed in Danforth et al., 2006a). Based on morphology alone, the root of bee phylogeny could fall virtually anywhere; however, the most likely alternatives obtained by Alexander and Michener (1995) were between Colletidae and the rest of the bees, within Colletidae (with Euryglossinae sister to the rest of the bees), and between the clade LT bees (Apidae + Megachilidae) + Melittidae and the remaining short-tongued bees (Fig. 1a and b). Studies based on nuclear gene data (Danforth et al., 2006a,b) and combined analyses of morphological and molecular data (Danforth et al., 2006b) have strongly suggested a root within Melittidae, with Dasypodainae sister to the rest of the bees (Fig. 1c). This result seems to accord well with the temporal appearance of bees in the fossil record (Michez et al., 2007), as well as the biology of Melittidae (Danforth, 2007). The molecular data strongly reject a root node near or within the family Colletidae (Danforth et al., 2006b), which has traditionally been considered the bee family sister to all other bees (Engel, 2001; p. 156).

Here we use the F1 and F2 paralogs of elongation factor 1-alpha in bees to compare the tree topologies that result when using outgroup versus paralog rooting. Empirical examples comparing outgroup rooting with paralog rooting are few (but see Mathews and Donoghue, 2000) because paralog rooting has typically been used when outgroup rooting has been dismissed as a viable possibility. We compare these alternative approaches to rooting bee phylogeny, investigate possible sources of error between the two methods, and discuss the implications of our results toward inferring the early history of bee evolution.

2. Material and methods

We extracted DNA from individual specimens following a phenol–chloroform protocol detailed in Danforth (1999). We obtained sequence information from the F1 and F2 copies of EF-1α using

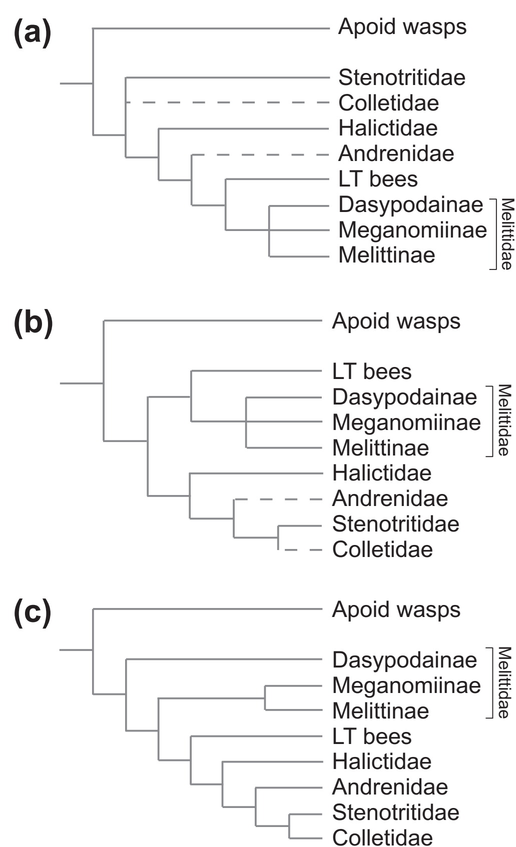

Fig. 1. Alternative rootings proposed for bee family-level phylogeny; (a and b): alternative rootings obtained by Alexander and Michener (1995); (c) topology obtained by Danforth et al. (2006b). "LT bees" refers to the "long-tongued" bee clade, including the families Apidae and Megachilidae. Long-tongued bees are consistently recovered as monophyletic based on both morphology and molecular data. Dashed lines indicate lineages that were not consistently recovered as monophyletic. Note that the Melittidae is not monophyletic in c.

standard PCR methods. The F2 sequences used in this study were included in a previous multi-gene bee phylogeny (Danforth et al., 2006b). This is the first application of F1 toward inferring bee phylogeny at a broad level, although this copy has been included in earlier studies of allodapine and xylocopine bees (Bull et al., 2003; Cronin, 2004; Leys et al., 2002; Schwarz et al., 2003) and is also commonly used in phylogenetic studies of ants (e.g., Brady et al., 2006; Schultz and Brady, 2008; Ward et al., 2010). We sequenced two F1 fragments using the following primers: For1deg (forward): 5′-GY ATC GAC AAR CGT ACS ATY G-3′ with Rev2 (reverse) 5′-YTC SAC YTT CCA TCC CTT GTA CC-3′ and For3 (forward) 5′-GGN GAC AAY GTT GGY TTC AAC G-3′ with Cho10 (reverse) 5′-AC RGC VAC KGT YTG HCK CAT GTC-3′. All PCR products were gel-purified in low-melting point agarose gels overnight at 4 °C. DNA was recovered from gel slices using the Promega Wizard PCR Preps DNA Purification kit. Automated sequencing of both strands was accomplished using the same amplification primers through the Cornell Life Sciences Core Laboratories (http://cores.lifesciences.cornell.edu/brcinfo/index.php). Sequence chromatograms were assembled and edited using Sequencher. Sequence alignments were generated with the Lasergene DNA Star software package using Clustal W (Thompson et al., 1994). Reading frames and intron/exon boundaries were determined by comparison with published sequences for the honeybee, *Apis mellifera* (Walldorf and Hovemann, 1990). All intron positions were removed from analyses.

Our sampling included 47 taxa spanning the diversity of bees (Table 1). We included multiple taxa from six of seven bee families (only

Author's personal copy

Table 1
Taxa used in this study with collection information and GenBank accession numbers.

Family	Subfamily	Genus and species	DNA code	F1 Copy	F2 Copy	Locality	Coll.
Crabronidae	Bembicinae	*Bicyrtes ventralis*	Bivn813	AY363036	AY585161	USA: NY: Oswego Co., Selkirk Shores SP, August 5 2000	Ascher
Crabronidae	Bembicinae	*Stizoides sp.*	Szsp817	JF806360	AY585164	USA: AZ: Cochise Co., Rucker Canyon, 9 mi W US80, August 17 2000	Danforth
Crabronidae	Crabroninae	*Anacrabro ocellatus*	Anoc803	AY363037	AY585160	USA: NY: Tompkins Co., Ithaca, July 22, 2000	Danforth
Crabronidae	Crabroninae	*Oxybelus sp.*	Oxsp805	JF806351	AY585169	USA: AZ: Cochise Co., Willcox, August 24, 2000	Danforth
Sphecidae	Sphecinae	*Sceliphron caementarium*	Spcm836	AY363042	AY585167	USA: NY: Tompkins Co., Ithaca, July 23, 2000	Danforth
Andrenidae	Alocandreninae	*Alocandrena porterae*	Alpo49	JF806339	AY585099	PERU: Lima Dept., St. Bartholome, October 21 1997	Rozen
Andrenidae	Andreninae	*Andrena* (*Callandrena*) *brooksi*	Ansp643	AY362984	AY230129	USA: AZ: Cochise Co., Animas, 20 mi S, September 17, 1999	Danforth
Andrenidae	Oxaeinae	*Protoxaea gloriosa*	Pxgl226	AY362987	AY585106	USA: AZ: Cochise Co., Portal	Danforth
Andrenidae	Panurginae	*Calliopsis* (*Nomadopsis*) *anthidia*	Caan597	AY362988	AY585100	USA: CA: Santa Clara Co., Mt Hamilton, May 28 1999	Ascher
Andrenidae	Panurginae	*Calliopsis* (*Calliopsis*) *pugionis*	Capu509	JF806342	AY585102	USA: CA: Riverside Co., San Jacinto WA, May 18, 1993	Danforth
Andrenidae	Panurginae	*Melitturga clavicornis*	Mtcl959	JF806348	AY585104	FRANCE: Herault, Causse de la Selle, June 17, 2002	Danforth
Andrenidae	Panurginae	*Panurgus calcaratus*	Pnca514	JF806352	AY585105	ITALY: Rome, June 7 1998	Ascher
Colletidae	Callomelittinae	*Callomelitta antipodes*	Cman687	AY362999, AY363000	AY585122	AUSTRALIA: NSW: Guyra, 74 km E, December 5 1999	Danforth
Colletidae	Colletinae	*Colletes inaequalis*	Coin451	JF806344	AY585123	USA: NY: Tompkins Co., Ithaca	Danforth
Colletidae	Scrapterinae	*Scrapter ruficornis*	Scrc938	JF806357	AY585138	SOUTH AFRICA: WCP Kunje Farm, nr Citrusdal, September 23, 2001	Danforth
Colletidae	Euryglossinae	*Xanthesma furcifera*	Xnfu709	AY363010	AY585140	AUSTRALIA: VIC: Patchewollock, December 10 1999	Danforth
Colletidae	Xeromelissinae	*Chilimelissa rozeni*	Chrz857	AY363013	AY585120	CHILE: Region III, Panamerican Hwy., km 1005 NE Chanaral	Packer
Halictidae	Halictinae	*Agapostemon tyleri*	Agty230	JF806338	AF140320	USA: AZ: Cochise Co., Portal	Danforth
Halictidae	Halictinae	*Augochlorella pomoniella*	Aupo592	JF806341	AF435373	USA: CA: Inyo Co., Big Pine, June 15 1999	Ascher
Halictidae	Halictinae	*Halictus* (*Halictus*) *rubicundus*	Haru32	JF806346	AF140335	USA: MT: Missoula Co., Missoula	Soucy
Halictidae	Nomiinae	*Dieunomia* (*Epinomia*) *nevadensis*	None207	JF806350	AF435397	USA: AZ: Cochise Co., Portal AZ	Danforth
Halictidae	Nomiinae	*Pseudapis unidentata*	Psun241	AY363018	AF435404	SPAIN: Almeria Prov., Laujar de Andarax, 20 August 1996	Danforth
Halictidae	Nomioidinae	*Nomioides facilis*	Nmsp243	AY363019	AF435394	SPAIN: Granada Prov., Trevelez, 20 August 1996	Danforth
Halictidae	Rophitinae	*Conanthalictus wilmattae*	Cowi351	JF806345	AF435378	USA: CA: Riverside Co., Palm Desert, 10 mi S. March 15 1997	Ascher
Halictidae	Rophitinae	*Dufourea mulleri*	Dumu233	AY363021	AF435383	locality unknown	Lincoln
Halictidae	Rophitinae	*Penapis penai*	Pnpe572	JF806353	AF435401	CHILE: Huasco Prov., N Vallenar (vial 1)	Rozen
Halictidae	Rophitinae	*Systropha curvicornis*	Sycu350	JF806359	AF435411	AUSTRIA: Vienna	Ayasse
Halictidae	Rophitinae	*Xeralictus bicuspidariae*	Xrbi566	JF806361	AF435413	USA: CA: San Diego Co., Anza Borrego Desert SP, February 29 1998	Snelling
Melittidae	Dasypodainae	*Dasypoda argentata*	Daar973	AY363029	AY585148	FRANCE: Gard, Generac 2 km E, June 22, 2002	Danforth
Melittidae	Dasypodainae	*Hesperapis* (*Panurgomia*) *larreae*	Hela488	AY363030	AY230131	USA: CA: Los Angeles Co., Palmdale, 13 June 1999	Ascher
Melittidae	Dasypodainae	*Hesperapis* (*Panurgomia*) *regularis*	Herg469	AY363031	AY585151	USA: CA: Santa Clara Co., Del Puerto Canyon, May 27 1999	Moeller
Melittidae	Dasypodainae	*Haplomelitta* (*Prosamba*) *griseonigra*	Hpgr939	AY363032	AY585153	SOUTH AFRICA: WCP 5 km S. Clanwilliam, September 7, 2001	Danforth
Melittidae	Melittinae	*Macropis* (*Macropis*) *europaea*	Maeu980	AY363033	AY585154	FRANCE: Herault, Portiragnes, along canal, June 27, 2002	Danforth
Melittidae	Melittinae	*Melitta* (*Melitta*) *eickworti*	Meew508	AY363034	AY585157	USA: NY: Tompkins Co., Ithaca, June 25 1997	Ascher
Melittidae	Melittinae	*Melitta* (*Melitta*) *leporina*	Mele981	JF806349	AY585158	FRANCE: Aude, Port la Nouvelle, June 17, 2002	Danforth
Melittidae	Melittinae	*Rediviva mcgregori*	Rvmc945	JF806355	AY585159	SOUTH AFRICA: NCP Dassiefontein, Kamieskroon, September 16, 2001	Danforth
Melittidae	Melittinae	*Redivivoides simulans*	Rvsi946	JF806356	AY585142	SOUTH AFRICA: WCP 5 km N. Clanwilliam, September 20, 2001	Danforth
Stenotritidae	Stenotritinae	*Stenotritus sp.*	Stsp1015	JF806358	DQ141115	AUSTRALIA: WA 23 km SW Coorow, November 17, 1997	Houston
Apidae	Apinae	*Anthophora montana*	Anmo633	JF806340	AY585107	USA: AZ: Cochise Co., Chiricahua Monument, September 14, 1999	Danforth
Apidae	Apinae	*Apis mellifera*	n/a	X52884	AF015267	Unknown	unknown
Apidae	Apinae	*Centris rhodopus*	Cnrh615	JF806343	AY585109	USA: AZ: Cochise Co., Rodeo, 2 mi N., September 8 1999	Danforth
Apidae	Apinae	*Ctenoplectra albolimbata*	Ctal983	AY362994	AY585118	SOUTH AFRICA: KZN, 20 km E Hluhluwe, March 9–12, 2002	Danforth
Apidae	Apinae	*Leiopodus singularis*	Lpsn651	AY362996	AY585113	USA: NM: Grant Co., Hachita, September 24, 1999	Danforth
Apidae	Apinae	*Pachymelus peringueyi*	Pmpe985	AY362992	AY585114	SOUTH AFRICA: NCP Dassiefontein, Kamieskroon, September, 16, 2001	Danforth
Apidae	Apinae	*Zacosmia maculata*	Zoma650	JF806337	AY585117	USA: NM: Grant Co., Hachita September 24, 1999	Danforth
Apidae	Nomadinae	*Holcopasites ruthae*	Horu511	JF806347	AY585112	USA: CA: Riverside Co., San Jacinto WA May 18, 1993	Danforth
Apidae	Nomadinae	*Paranomada velutina*	Pnve652	JF806354	AY585115	USA: AZ: Cochise Co., Apache, 2 mi E, September 10 1999	Danforth
Apidae	Nomadinae	*Triepeolus robustus*	Trrz635	AY362997	AY585116	USA: AZ: Cochise Co., Chiricahua Monument, September 14, 1999	Danforth
Apidae	Xylocopinae	*Ceratina calcarata*	Cedu656	AY362998	AY585108	USA: NY: Tompkins Co., Ithaca, August 4, 1999	Danforth
Megachilidae	Megachilinae	*Anthidium oblongatum*	Atob505	AY363025	AY585145	USA: NY: Tompkins Co., Ithaca, July 1 1999	Danforth
Megachilidae	Megachilinae	*Chelostoma fuliginosum*	Chfu496	AY363028	AY585146	USA: NY: Tompkins Co., Ithaca, June 27 1997	Danforth
Megachilidae	Megachilinae	*Megachile pugnata*	Mepg595	AY363027	AY585147	USA: NY: Schuyler Co., Valois gravel pit, June 14, 1999	Danforth

Author's personal copy

one species from the small family Stenotritidae was used), with representation from many subfamilies within each family. Although Melittidae is treated as a family here, it contains three well-defined lineages (Dasypodainae, Meganomiinae, and Melittinae s.s) that have been treated as families in some previous studies and Melittidae (*sensu lato*) may be a paraphyletic group (Alexander and Michener, 1995; Danforth et al., 2006b). Our study included multiple representatives from each of the three melittid subfamilies. We also gathered data from the closest relatives to bees, Crabronidae and Sphecidae, to serve as outgroups (Melo, 1999; Pilgrim et al., 2008).

We conducted phylogenetic analyses on two distinct arrangements of the data. Information from both paralogs was concatenated for each taxon, resulting in a dataset with 52 terminals and 2008 characters (857 from EF-1α F1; 1151 from EF-1α F2). We refer to this as the "concatenated" dataset, which represents the outgroup rooting strategy. Alternatively, the two paralogs for all ingroups were treated as separate taxa, producing a dataset that contained all sampled nucleotide sites that are shared between the two paralogs. This resulted in a dataset with 94 terminals and 828 characters, which we refer to as the "duplicated" dataset; this dataset represents the paralog rooting strategy. We also repeated analysis of this duplicated dataset after adding the Crabronidae and Sphecidae paralog fragments to investigate the influence of these non-bee sequences on the paralog rooting strategy.

Base pair frequencies and tests for base composition heterogeneity were calculated using PAUP* v.4.0b10 (Swofford, 2002). We used Bayesian phylogenetics to estimate tree topologies and branch lengths using MrBayes v.3.1.2 (Ronquist and Huelsenbeck, 2003) parallelized over eight Mac G5 processors. The data were partitioned by codon position and, in the case of the concatenated dataset, by gene as well. The best fitting model for each partition was selected using MrModelTest v.2.2 (Nylander, 2004) under the Akaike Information Criterion (AIC) (Posada and Buckley, 2004). In all cases, the selected model was GTR + I + Γ. All parameters were unlinked across partitions except branch lengths and topology. Partition-specific branch length rate multipliers (m) were unlinked and initially assigned a Dirichlet prior (*ratepr = variable*). Runs were initiated using random topologies with branch length priors of *brlenspr = Unconstrained:Exponential* (100) to avoid becoming trapped in parameter space with distorted branch lengths and partition rate multipliers (Marshall, 2010). Each run was distributed across four chains with a heating parameter of 0.05 (compared to the default setting of 0.2); this was done in order to increase mixing among chains. Runs were conducted for 50 million generations with a burnin of 10 million generations. Convergence between independent MCMC runs was verified using PSRF values, the average standard deviation of split frequencies, and by plotting likelihood values across generations using Tracer v1.4 (Rambaut and Drummond, 2007). To evaluate the performance of outgroup versus paralog rooting approaches, we used Bayesian posterior probabilities to facilitate direct comparison of nodal support between analyses, focusing on groups that are already well established from independent sources of evidence.

We also conducted supplementary phylogenetic analyses to explore the potential role of heterogeneity in causing systematic bias. First, we repeated our MrBayes analyses using RY-coding. This technique recodes nucleotides as purines or pyrimidines in order to reduce the influence of base pair compositional bias (Woese et al., 1991; Phillips et al., 2004). We used the program Mesquite v2.74 (Maddison and Maddison, 2010) to recode the data and conducted MrBayes analyses under the same conditions described above. Secondly, we used the program BayesPhylogenies to implement a mixture model without creating prior character partitions (Pagel and Meade, 2004). We implemented a mixture model with three independent rate matrices (*patterns* = 3) with a GTR + Γ model for each matrix; these conditions correspond to the 3Q + Γ model that Pagel and Meade (2004) preferred after analyzing a similar two gene data set. The BayesPhylogenies analyses were conducted twice for each data set, with each run consisting of 70 million generations. The first 20 million generations were removed from each run as burnin, and the post-burnin trees from both runs were combined to yield a final pool of 100 million generations of post-burnin trees.

We compared alternative rooting positions in the concatenated dataset using the Shimodaira Approximately Unbiased (AU) test (Shimodaira, 2002). We conducted this test using nine different rooting hypotheses to reflect the range of possible placements for the root. The established monophyly of long tongued bees and all short tongued families except perhaps Melittidae (Danforth et al., 2006b) drastically constrains plausible root positions. The alternative hypotheses we assessed correspond to the root at: (1) Andrenidae (e.g., a monophyletic Andrenidae as sister to the remaining extant bees); (2) Colletidae; (3) Colletidae + Stenotritidae (represented in Fig. 1A); (4) Melittidae; (5) Halictidae; (6) long-tongued bees (Apidae + Megachilidae); (7) long-tongued bees + Melittidae (Fig. 1B); (8) Dasypodainae (Fig. 1C); (9) *Panurgus* (part of Andrenidae). We inferred the maximum likelihood tree under each of the nine alternatives by conducting constraint analyses with GARLI v.1.0 (Zwickl, 2006) using the GTR + I + Γ model and 20 search replicates. The AU test was performed using these topologies in PAUP.

We evaluated protein functional divergence between the two paralogs using likelihood-ratio tests that examine which amino acid sites differ significantly between the two copies on a site-by-site basis (Knudsen and Miyamoto, 2001). An intuitive example of such divergence is a site that is highly variable in one copy but completely conserved in the other copy, which could be interpreted as possessing greater functional importance in the latter copy. This approach has proven successful at recovering true rates shifts in other gene family datasets (e.g., Blouin et al., 2005; Hyndman et al., 2009). The implementation of the Knudsen and Miyamoto (2001) test (Protein Rate Shift Analysis Server: www.daimi.au.dk/~compbio/rateshift) limits the number of sequences in each paralog to 45, so we excluded two species (*Melitta eickworti* and *Calliopsis anthidia*) with congeneric representatives in our dataset. This number of analyzed sequences is well over the 30 sequences recommended by the simulation work in Blouin et al. (2005) as the threshold needed to minimize sampling error in these types of tests. We converted DNA sequences into amino acids using MacClade (Maddison and Maddison, 2000), used the topology derived from the duplicate dataset phylogenetic analysis, and employed the JTT amino acid rate matrix (Jones et al., 1992).

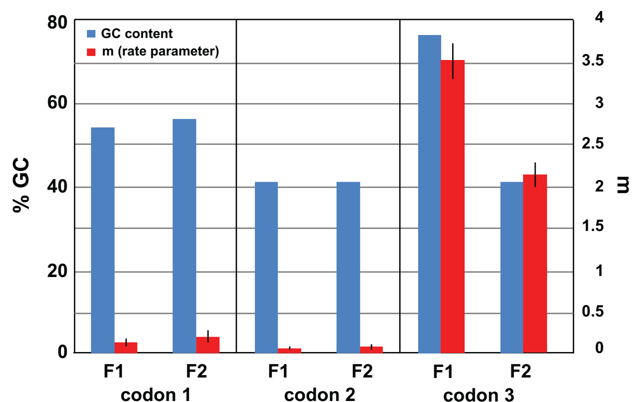

Fig. 2. Comparison between F1 and F2 paralogs of GC content (blue bars) and rate parameters (red bars, with 95% credibility intervals) for each codon position. The parameter m is the rate parameter estimated by MrBayes. Substantial differences between the two copies are limited to codon position three. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Author's personal copy

3. Results

We gathered sequence data from both F1 and F2 copies of EF-1α from 47 bee taxa and 5 closely-related outgroups. The concatenated matrix (F1 and F2 for each taxon) assembled from these data contained 2008 protein-coding nucleotide sites, 739 of which were parsimony-informative. GC content was virtually equal in both F1 and F2 copies for codon positions 1 and 2 (Fig. 2). However, substantial differences occurred in codon position 3, with F1 showing a much higher GC base composition (76%) than F2 (41%). The ordering of estimated partition-specific rate multipliers (m_i, where i denotes codon position) for both copies was $m_2 < m_1 < m_3$ as expected from general patterns of protein-coding gene evolution (Lin and Danforth, 2004). The values for m_1 and m_2 did not show significant differences between F1 and F2, while m_3 showed a significantly elevated estimate in F1 (mean = 3.5) compared to F2 (mean = 2.1) (Fig. 2).

Bayesian analysis of the concatenated dataset (Fig. 3) recovered monophyly with strong support for all bee families (except Stenotritidae which is represented by only a single species) (Table 2). Other higher-level relationships robustly inferred by this analysis included ((Colletidae + Stenotritidae) + Halictidae)) and Apidae + Megachilidae (i.e., long-tongued bees). The outgroups in this analysis rooted the tree essentially at a soft trichotomy (i.e., with PP < 0.5) involving Melittidae, Andrenidae, and all remaining bees. These results held under the RY-coding and mixture model analyses with the notable exception that support for a monophyletic Melittidae greatly eroded under RY-coding.

The Bayesian phylogenetic estimate from the duplicate dataset (828 protein-coding sites, 346 of which were parsimony-informative) was rooted using midpoint rooting, such that the initial split in the tree occurred between the F1 and F2 copies (Fig. 4). The topology within the F1 copy indicated a paraphyletic Melittidae as sister to all other extant bees (PP = 0.84), although this result did not hold up under RY-coding and mixture model analyses. However, under both standard and RY-coding part of Melittidae (subfamily Dasypodainae) still resolved as sister to all other bees with moderate support (standard coding, PP = 0.77; RY-coding, PP = 0.73). The F2 copy suggested a paraphyletic Andrenidae as sister to all other extant bees (PP = 0.68) with similar support under

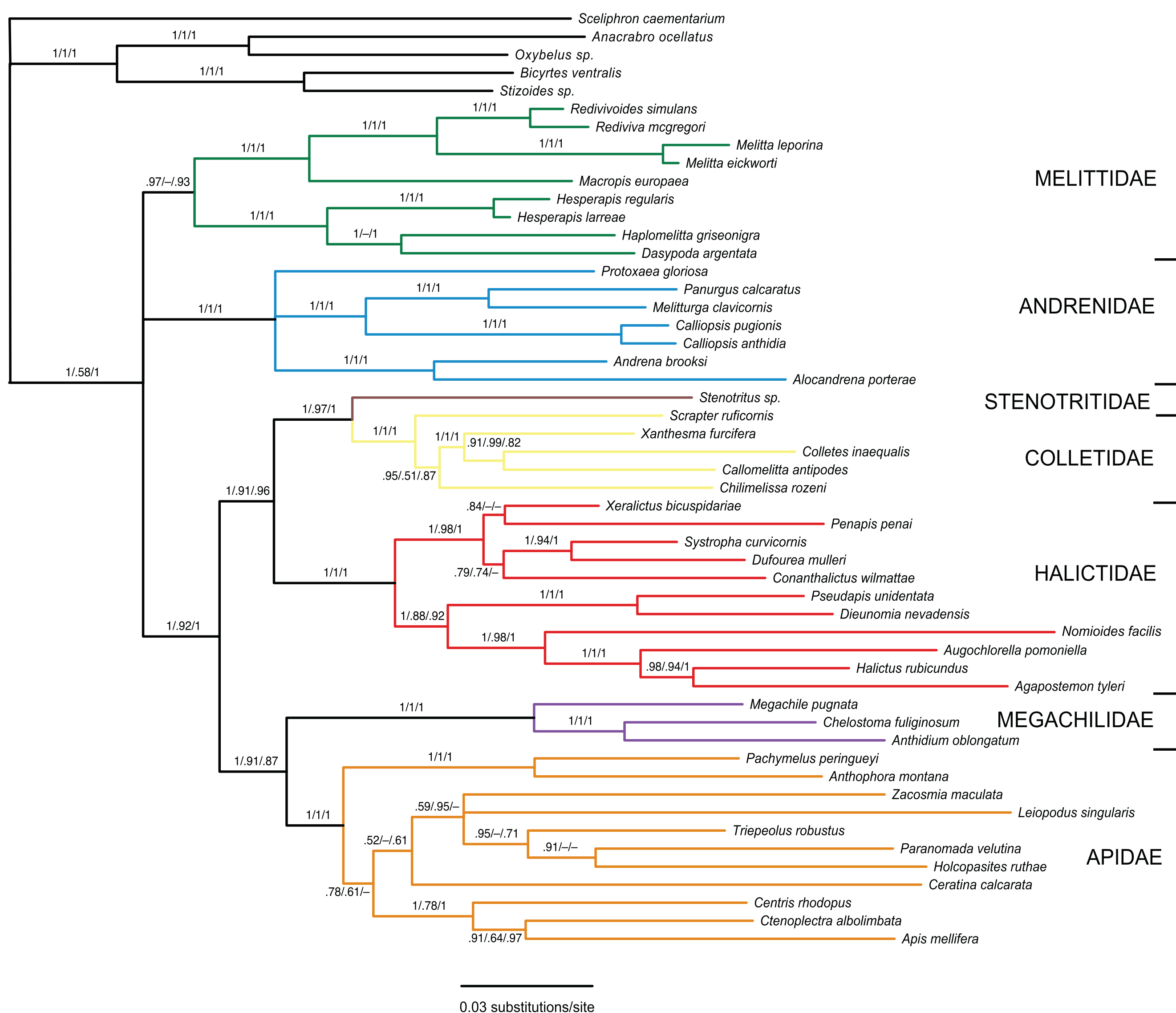

Fig. 3. Bayesian phylogram of the concatenated dataset with outgroup rooting. The tree is rooted at the branch connecting the ingroup to the five non-bee outgroups. Shown is the phylogram from the MrBayes run with nodes below 0.5 collapsed as polytomies. Branch labels indicate posterior probabilities above 0.5 from three analyses: MrBayes/MrBayes with RY-coding/BayesPhylogenies. Branch colors indicate different bee families.

Author's personal copy

Table 2
Comparison of posterior probabilities for bee families and other higher-level taxa between the concatenated gene and duplicate gene analyses. Support values are ordered according to method of analysis: MrBayes/MrBayes with RY-coding/BayesPhylogenies. LT, long tongue bees (Apidae + Megachilidae).

Taxon	Concatenated	Duplicate, F1 copy	Duplicate, F2 copy
Andrenidae	1/1/1	<0.5/<0.5/<0.5	<0.5/<0.5/<0.5
Apidae	1/1/1	<0.5/<0.5/<0.5	<0.5/0.74/<0.5
Colletidae	1/1/1	0.81/<0.5/0.55	1/0.81/0.97
Halictidae	1/1/1	1/1/1	1/0.99/1
Megachilidae	1/1/1	1/0.96/0.99	1/1/1
Melittidae	0.97/<0.5/0.93	<0.5/<0.5/<0.5	<0.5/<0.5/<0.5
LT bees	1/0.91/0.87	<0.5/0.60/<0.5	0.67/<0.5/1
Colletidae + Stenotritidae	1/0.97/1	1.0/<0.5/0.97	1/0.57/1
Halictidae + (Colletidae + Stenotritidae)	1/0.91/0.96	<0.5/<0.5/<0.5	0.96/0.73/0.87

RY-coding and stronger support under mixture model analysis. The pattern of Andrenidae paraphyly was also strongly supported under the mixture model but was much weaker under RY-coding. When analyses were repeated including taxa from Crabronidae and Sphecidae, these rooting patterns for bees were upheld (Figs. A.1 and A.2 in Supplementary material); Crabronidae and Sphecidae were weakly resolved to fall within the bees, a result contradicted by the overwhelming evidence for bee monophyly (Danforth et al., 2006b; Michener, 2007).

The topologies within both copies for the duplicate analyses showed much weaker support for some families and other higher-level bee taxa compared to the concatenated analyses (Table 2). Relationships along the spine of the tree within each copy were generally poorly supported under all analyses except for the F2 copy in the mixture model analysis.

The data were significantly worse fitting under the AU test when the root of the concatenated dataset was constrained to fall at Colletidae, Halictidae, or *Panurgus*, compared with the most likely root position at Andrenidae (Table 3). A root constrained at (Colletidae + Stenotritidae) actually showed a worse likelihood score than one constrained at Colletidae, although not significantly so ($p = 0.0773$). A root at Melittidae was the second most likely position, with a log likelihood ($-\ln L$) difference of only 0.18282 compared to a root at Andrenidae ($p = 0.6601$).

Results from the test for functional divergence between the two paralogs are summarized in Table 4. Of the 276 amino acid positions tested, 11 positions showed functional divergence at the $p = 0.05$ level or below, with five shifts occurring in the F1 copy and six shifts in the F2 copy. Two of the F1 sites are clustered together (sites 35 and 36) and three of the F2 sites likewise are in the same area (sites 109, 111, 115), suggesting regions of functional difference between the two copies. However, after application of a Bonferroni correction for multiple tests (as recommended by the developers of the test), only a single site, the highest ranked comparison, remained statistically significant.

4. Discussion

4.1. Sequence comparison between F1 and F2 paralogs

Comparison of the sequence characteristics within the F1 and F2 copies indicates substantially higher GC bias in the third codon position of F1. This GC bias in the F1 paralog is not limited to bees. A recently published phylogenetic dataset from dolichoderine ants that included 1074 bp of coding F1 sites and 516 of F2 (Ward et al., 2010) also showed inflation of third position GC content in F1 (F1: 79%; F2: 55%); similar patterns are evident in other ant datasets (Brady et al., 2006; LaPolla et al., 2010; Schultz and Brady, 2008). There is insufficient data from other lineages of Hymenoptera to determine if this phenomenon holds true for the entire order. In our analysis, only the third position of F1 showed a significantly higher rate of evolution compared to F2, which may indicate selection for increased GC content in these synonymous sites of F1.

Gene duplication may lead to functional divergence between the two copies, with one copy experiencing an increase in site-specific rate evolution as it evolves a new or modified function, so-called neofunctionalization (Dittmar and Liberles, 2010; Lynch, 2007; Ohno, 1970). It is well established that such site-specific rate shifts can negatively impact phylogenetic reconstruction (Philippe et al., 2000; Tuffley and Steel, 1998). Thus, we tested for strong and consistent functional differences between F1 and F2 at the amino acid level that might influence phylogenetic performance when the two copies are analyzed as paralogs. However, we found very little evidence suggesting that one copy is more functionally conserved than the other, as the few detectable differences were not statistically robust and were evenly distributed between the two copies (Table 4). These differences could perhaps influence protein function in some manner, but such evaluation would require future work on the crystal structure and *in situ* expression of these genes that is beyond the scope of the current study.

4.2. Comparison between paralog rooting and outgroup rooting

Outgroup rooting appears to outperform paralog rooting in our dataset, judging by several criteria. Although neither method yields an unambiguous root to the tree, outgroup rooting recovered deeper clades and also showed stronger support for clades that have been well established by morphological and other molecular data. Specific examples of these clades within bees are discussed in Section 4.4 below.

Several non-exclusive reasons may explain the poorer resolution seen in the analyses using paralog rooting compared to outgroup rooting. One factor may be that the paralog method is being rooted with a much longer branch compared to that of the outgroup method. Several studies have indicated that distantly related outgroups may confound reconstruction of ingroup relationships (Bergsten, 2005; Brady et al., 2006; Holland et al., 2003; Lin et al., 2002; Milinkovitch and Lyons-Weiler, 1998). In our paralog analysis, the branch connecting the two copies is quite long. Both paralogs are present in *Nasonia* (Niehuis et al., 2008, 2007), a parasitic wasp group that diverged from the aculeates (including bees and ants) perhaps 200 million years ago (Grimaldi and Engel, 2005). If the paralogs in other holometabolous insects such as Diptera and Coleoptera were the result of the same duplication event, this could push this age even farther back in time, as old as 350 million years (Wiegmann et al., 2009). However, much more information concerning the distribution of these paralogs in holometabolous insects, as well as the phylogeny of this group in general, will be needed to explore this further. In any event, the branch leading to the outgroup in the concatenated analysis (the branch separating bees and the wasp families Crabronidae and Sphecidae) is undoubtedly much shorter.

The poorer performance of the duplicated dataset may also be simply due to the paralog rooting analysis containing approximately half the characters distributed over twice the number of terminal taxa. Graybeal (1998) demonstrated that as the number of taxa in a dataset increases, phylogenetic accuracy also increases, provided that there are enough informative characters per taxon to

Author's personal copy

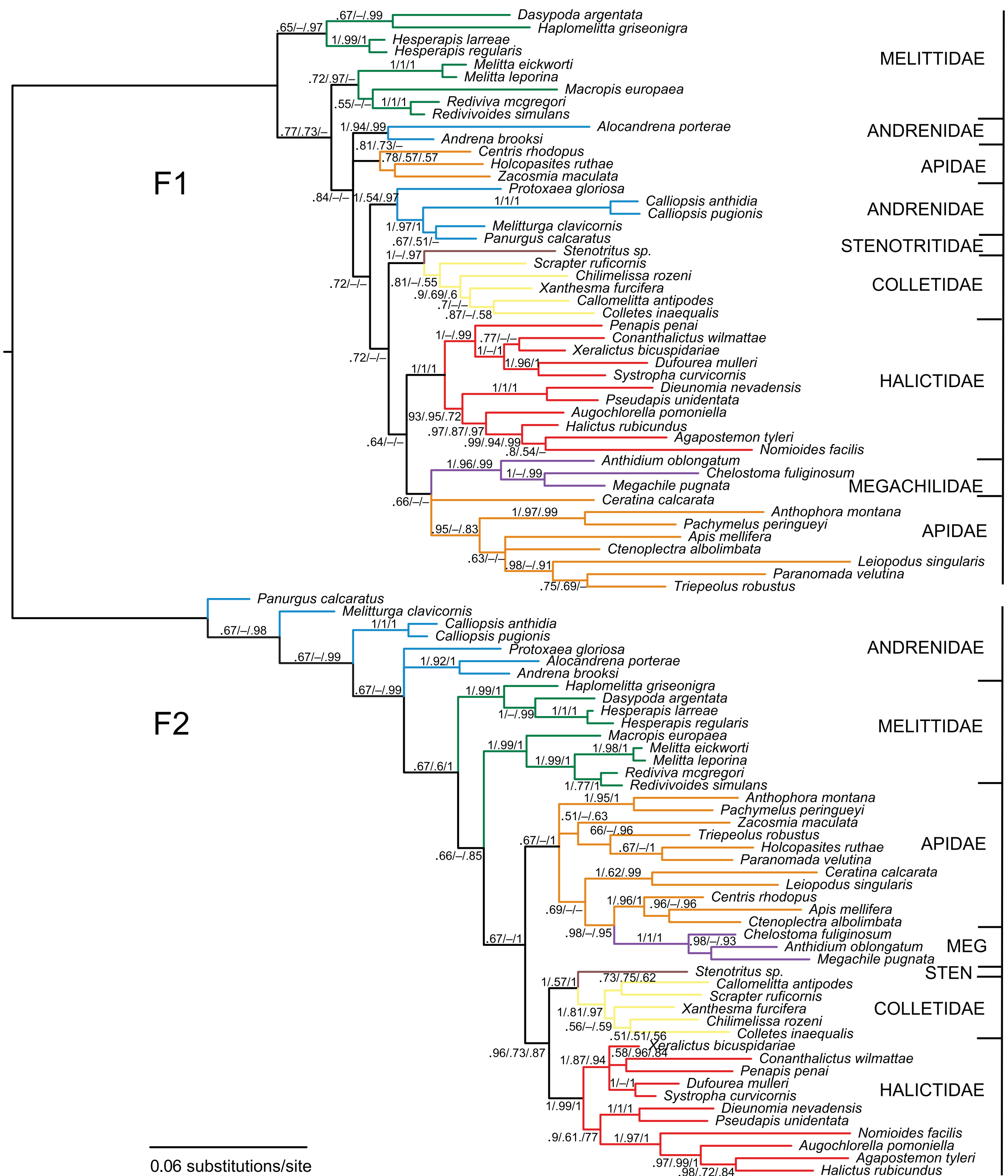

Fig. 4. Bayesian phylogram of the duplicate dataset with paralog rooting. The tree is rooted at the branch connecting the two copies of the paralogous gene copies (F1 and F2). Shown is the phylogram from the MrBayes run with nodes below 0.5 collapsed as polytomies. Branch labels indicate posterior probabilities above 0.5 from three analyses: MrBayes/MrBayes with RY-coding/BayesPhylogenies. Branch colors indicate different bee families. STEN, Stenotritidae; MEG, Megachilidae.

supply phylogenetic signal necessary to resolve relationships. In other words, there is a minimum number of characters necessary to characterize relationships in a dataset containing a certain number of taxa. While adding extra data beyond this minimum number does little to enhance tree recovery, using less may seriously impact recovery of the correct tree topology. Graybeal (1998; her Fig. 1) showed that for a dataset of 1000 base pairs, recovery of the correct tree topology is consistently low and drops off markedly for datasets containing more than eight taxa. By analyzing the F1 and F2 copies as paralogs instead of concatenating them,

Author's personal copy

Table 3
Comparison of alternative rooting hypotheses using the Shimodaira Approximately Unbiased (AU) test on the concatenated dataset. LT, long tongue bees (Apidae + Megachilidae).

Root position	−ln *L* value	−ln *L* difference	*P* value
Andrenidae	27956.55054	–	–
Melittidae	27956.73336	0.18282	0.6601
Dasypodainae	27959.24082	2.69028	0.4223
LT bees	27973.08175	16.53121	0.2434
LT + Melittidae	27974.63967	18.08914	0.1463
Halictidae	27982.84097	26.29043	0.0098
Colletidae + Stenotritidae	27983.02981	26.47928	0.0773
Colletidae	28003.53998	46.98945	0.0186
Panurgus	28058.09533	101.54480	0.0000

Table 4
Results of protein rate shift analysis showing the 11 amino acid sites with significant rate change between gene copies. Only the highest ranked site remains statistically significant after Bonferroni correction for multiple tests.

Rank	Alignment position	Significance	Slow copy
1	222	0.0001	F1
2	115	0.0043	F2
3	111	0.0058	F2
4	35	0.0074	F1
5	184	0.0077	F1
6	109	0.0135	F2
7	187	0.0146	F2
8	72	0.0266	F2
9	149	0.0355	F2
10	242	0.0361	F1
11	36	0.0486	F1

the length of our dataset is approximately halved (from 2008 characters to 828), while the number of terminals is nearly doubled (from 52 terminals to 94). According to Graybeal (1998), this is likely too few characters per taxon to resolve evolutionary relationships and may help explain the superior performance of the concatenated dataset over the duplicated. A further study by Wortley et al. (2005) showed that tree recovery improves as length of the dataset increases. Their Fig. 3 showed a sharp increase in tree recovery as a dataset with 40 terminal taxa increases from zero to 2000 base pairs; for datasets greater than 2000 base pairs, increase in tree recovery is less sharp. This suggests that while the concatenated dataset is expected to outperform the duplicated dataset, even the concatenated dataset may contain too few characters to adequately describe some bee relationships, especially with regard to the base of the tree.

Finally, one cannot entirely rule out the possibility that these gene copies have experienced different evolutionary histories that are being accurately reconstructed by the paralog rooted tree. It seems doubtful that this is a major factor here, since the paraphyly of Apidae, Melittidae and Andrenidae exhibited by the F1 and F2 genes in the paralog rooted tree is contradicted by the strong support for their monophyly in the outgroup rooted tree.

Considering these factors in combination, our study highlights several concerns when applying paralog rooting to an empirical dataset. There is no guarantee that the two duplicate copies have experienced identical histories, nor that those histories can be reconstructed without error. Given the relatively small amount of information afforded by a single gene, the prospect of completely accurate phylogenetic reconstruction may be unrealistic even when the gene copy histories are completely identical. This situation suggests as a remedy analyzing many paralogous genes in combination, which may be accessible by screening complete genomic data. Also, paralog rooting may not be as useful in cases where the duplication event occurred well before divergences in the clade of interest, and may be especially problematic for inferring the root when the two copies are connected by an extremely long branch. Again, future consideration of genome-scale data may allow the identification of more appropriately recent duplication events for analysis.

4.3. Comparison to other paralog rooting studies

The two most notable case studies that have explored rooting using paralogs are the phylogeny of angiosperms and the tree of life. These studies have generally yielded more robust results compared to our bee example, although these other studies do also experience some conflicting results.

Inferring the root of angiosperms on molecular phylogenies using outgroups has been notoriously difficult because the closest living relatives of the angiosperms are widely divergent (Mathews, 2009). The difficulty has been further compounded by a lack of congruence between molecular and morphological data. In order to avoid the problems associated with rooting using distantly related outgroups, duplicate phytochrome genes (*PHYA* and *PHYC*) have been used to infer a root to angiosperm phylogeny (Donoghue and Mathews, 1998; Mathews and Donoghue, 1999, 2000). Their results supported *Amborella* as the sister taxon to all other angiosperms (Mathews and Donoghue, 1999), although they allowed for the possibility of alternative rootings, namely one in which water lilies (Nymphaeales), either alone or together with *Amborella*, are sister to all other angiosperms (Mathews and Donoghue, 2000; Zanis et al., 2002). Subsequent analyses of duplicate phytochrome genes covering all seed plants indicate that cycads are sister to angiosperms (Mathews et al., 2010).

In the most extreme cases, such as inferring the root of the universal tree of life, outgroups simply do not exist. A number of gene paralogs have been used in an attempt to find the root of the tree of life: the alpha and beta subunits of ATP-ase (Gogarten et al., 1989; Iwabe et al., 1989), Tu and G subunits of elongation factor (Iwabe et al., 1989; Baldauf et al., 1996; Hashimoto and Hasegawa, 1996), signal recognition particle (SRP) and the receptor SRα (Brinkmann and Philippe, 1999), and three copies of aminoacyl-tRNA synthetase genes (Brown and Doolittle, 1995). Most studies recovered eubacteria as the sister lineage to a clade consisting of Archaebacteria and eukaryotes (Baldauf et al., 1996; Gogarten et al., 1989; Hashimoto and Hasegawa, 1996; Iwabe et al., 1989), although the study by Brinkmann and Philippe (1999) supported prokaryotes as a monophyletic group. A more recent study (Zhaxybayeva et al., 2005) isolated a series of gene duplicates from twelve bacterial and archaeal genomes and used them to explore the root of the tree of life; their results largely supported a rooting between archaea and bacteria. Another paralog-based rooting approach, top-down indel rooting, can be used to analyze incomplete datasets by considering indel gains and losses, as well as gene gains and losses (Lake et al., 2007, 2008, 2009; Servin et al., 2008; Skophammer et al., 2007). This method has been used to reassess the position of the root of the tree of life; these five studies, evaluated in tandem, ultimately converge on a rooting at the branch between the clade Actinobacteria and the double-membrane prokaryotes and the clade firmicutes and Archaebacteria. A further study (Valas and Bourne, 2009), using similar methods, contrasts the findings of Lake et al. (2008, 2009) and finds support for the tree of life rooting proposed by Cavalier-Smith (2006), which places the root near Chlorobacteria.

4.4. Implications for early bee evolution

Analyses based on our concatenated dataset infer strong support for many deep clades that are well established by

Author's personal copy

morphological data and previous molecular studies (Danforth et al., 2006a,b), including monophyly of bee families and relationships among them (Fig. 3 and Table 2). In contrast, analyses based on the duplicate dataset do not recover many of these clades. For example, although both the F1 and F2 copy in the paralog rooted tree support megachilid monophyly with a PP = 1, neither copy recovers the monophyly of the Apidae or Andrenidae. The F2 copy recovers Halictidae + (Colletidae + Stenotritidae), while the F1 copy clusters (Colletidae + Stenotritidae) in a clade with the long-tongued bees.

The monophyly of Melittidae remains tenuous, despite the fact that it is widely recognized as a family (e.g., Michener, 2007). Some studies based on morphology support the monophyly of Melittidae (Engel, 2001; Michener, 1981), while others do not (Alexander and Michener, 1995). In a combined analysis of five nuclear genes and morphology, Danforth et al. (2006b) found Melittidae to be paraphyletic. In the present analysis, both the F1 and F2 copies in the paralog rooted tree recover a paraphyletic Melittidae, with the subfamily Dasypodainae forming a clade apart from the remaining Melittinae. This result is concordant with that of Danforth et al. (2006b) and lends support for the recognition of these subfamilies as independent families. The outgroup rooted tree, however, results in a monophyletic Melittidae. This is congruent with a recent study of Melittidae (and relevant outgroups) based on nuclear protein-coding genes (Michez et al., 2009), although the monophyly of the melittids was not well-supported by this study.

Based on the outgroup rooted tree, the root of the bee phylogeny was only partially resolved, recovering a three-way split between Melittidae, Andrenidae, and the remaining bee families. Based on the paralog rooted tree, the root is either between Dasypodainae and the remainder of the bees (F1 copy; Fig. 4), congruent with previous multi-gene studies (Danforth et al., 2006a,b), or within a paraphyletic Andrenidae (F2 copy; Fig. 4). In neither case is there strong support for basal relationships but, as discussed in Section 4.2 above, this may be largely due to the small size of the duplicated gene data set. The F2 copy under a mixture model suggests that a paraphyletic Andrenidae is sister to all other extant bees. However, Andrenidae is widely viewed as a monophyletic group based both on molecular (Ascher, 2004) and morphological (Ascher, 2004; Michener, 2007) data, so rooting at this position is highly suspect. Also, the AU test conducted on the concatenated dataset strongly rejects this rooting position.

Our results add more evidence to support the view that the root node of bees does not fall within or near the family Colletidae. Previous molecular studies (Danforth et al. 2006b) have shown that rooting at Colletidae or Colletidae + Stenotritidae is significantly less likely than rooting at or within Melittidae, in agreement with the AU tests conducted on the present data set. Interestingly, the rooting of bee phylogeny displayed by the EF-1α F1 copy in our paralog rooted tree is exactly the same rooting obtained in previous multigene studies (Danforth et al., 2006a,b), in spite of the fact that these studies used the EF-1α F2 copy but not the F1 copy. Therefore, the results obtained here based on EF-1α F1 can be viewed as independent of those earlier studies, providing further evidence of a possible root node within or near Melittidae, and arguing strongly against a root position within or near Colletidae.

Acknowledgements

We thank colleagues listed in Table 1 who generously provided valuable specimens for this study. Comments from the editor and two anonymous reviewers substantially improved the manuscript. This project was supported by National Science Foundation Grants DEB-0211701 and DEB-0742998 to B.N.D and EF-0431330 to S.G.B.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.05.002.

References

Alexander, B.A., Michener, C.D., 1995. Phylogenetic studies of the families of short-tongued bees (Hymenoptera: Apoidae). Univ. Kansas Sci. Bull. 55, 377–424.
Ascher, J.S., 2004. Systematics of the Bee Family Andrenidae (Hymenoptera: Apoidea). Ph.D. Dissertation, Cornell University, Ithaca, New York.
Baldauf, S.L., Palmer, J.D., Doolittle, W.F., 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. 93, 7749–7754.
Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21, 163–193.
Blouin, C., Butt, D., Roger, A.J., 2005. Impact of taxon sampling on the estimation of rates of evolution at sites. Mol. Biol. Evol. 22, 784–791.
Brady, S.G., Danforth, B.N., 2004. Recent intron gain in elongation factor-1α of colletid bees (Hymenoptera: Colletidae). Mol. Biol. Evol. 21, 691–696.
Brady, S.G., Schultz, T.R., Fisher, B.L., Ward, P.S., 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc. Natl. Acad. Sci. USA 103, 18172–18177.
Brady, S.G., Larkin, L., Danforth, B.N., 2009. Bees, ants, and stinging wasps (Aculeata). In: Hedges, S.B., Kumar, S. (Eds.), The Timetree of Life. Oxford University Press, New York, pp. 264–269.
Brinkmann, H., Philippe, H., 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16, 817–825.
Brown, J.R., Doolittle, W.F., 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA 92, 2441–2445.
Bull, N.J., Schwarz, M.P., Cooper, S.J.B., 2003. Phylogenetic divergence of the Australian allodapine bees (Hymenoptera: Apidae). Mol. Phylogenet. Evol. 27, 212–222.
Cavalier-Smith, T., 2006. Rooting the tree of life by transition analyses. Biology Direct. 1, 19.
Cronin, A.L., 2004. A molecular phylogeny and social behavior of Japanese *Ceratina* (Hymenoptera, Apidae, Xylocopinae). Insect Syst. Evol. 35, 137–146.
Danforth, B.N., 1999. Phylogeny of the bee genus *Lasioglossum* (Hymenoptera: Halictidae) based on mitochondrial COI sequence data. Syst. Entomol. 24, 377–393.
Danforth, B.N., 2007. Bees – a primer. Curr. Biol. 17, R156–R161.
Danforth, B.N., Ji, S., 1998. Elongation factor-1a occurs as two copies in bees: implications for phylogenetic analysis of EF-1a sequences in insects. Mol. Biol. Evol. 15, 225–235.
Danforth, B.N., Fang, J., Sipes, S.D., 2006a. Analysis of family-level relationships in bees (Hymenoptera: Apiformes) using 28S and two previously unexplored nuclear genes: CAD and RNA polymerase II. Mol. Phylogenet. Evol. 39, 358–372.
Danforth, B.N., Sipes, S.D., Fang, J., Brady, S.G., 2006b. The history of early bee diversification based on five genes plus morphology. Proc. Natl. Acad. Sci. USA 103, 15118–15123.
Dittmar, K., Liberles, D., 2010. Evolution after Gene Duplication. Wiley-Blackwell, Hoboken.
Djernaes, M., Damgaard, J., 2006. Exon-intron structure, paralogy and sequenced regions of elongation factor-1 alpha in hexapoda. Arthropod Syst. Phylogeny 64, 45–52.
Donoghue, M.J., Mathews, S., 1998. Duplicate genes and the root of the angiosperms, with an example using phytochrome sequences. Mol. Phylogenet. Evol. 9, 489–500.
Doolittle, W.F., Brown, J.R., 1994. Tempo, mode, the erogenote and the universal root. Proc. Natl. Acad. Sci. USA 91, 6721–6728.
Downie, D.A., Gullan, P.J., 2004. Phylogenetic analysis of mealybugs (Hemiptera: Coccoidea: Pseudococcidae) based on DNA sequences from three nuclear genes, and a review of the higher classification. Syst. Entomol. 29, 238–259.
Engel, M.S., 2001. A monograph of the Baltic amber bees and the evolution of the Apoidea (Hymenoptera). Bull. Am. Mus. Nat. Hist. 259, 1–192.
Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., Konishi, J., Denda, K., Yoshida, M., 1989. Evolution of the vacuolar H+-atpase – implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA 86, 6661–6665.
Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47, 9–17.
Grimaldi, D., 1999. The co-radiations of pollinating insects and angiosperms in the Cretaceous. Ann. Mo. Bot. Gard. 86, 373–406.
Grimaldi, D., Engel, M.S., 2005. Evolution of the Insects. Cambridge University Press, Cambridge.
Haring, E., Aspock, U., 2004. Phylogeny of the Neuropterida: a first molecular approach. Syst. Entomol. 29, 415–430.
Hasegawa, M., Iida, Y., Yano, T., Takaiwa, F., Iwabuchi, M., 1985. Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. J. Mol. Evol. 22, 32–38.
Hashimoto, T., Hasegawa, M., 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. Adv. Biophys. 32, 73–120.

Author's personal copy

Holland, B.R., Penny, D., Hendy, M.D., 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. Syst. Biol. 52, 229–238.

Hovemann, B., Richter, S., Walldorf, U., Cziepluch, C., 1988. Two genes encode related cytoplasmic elongation factors 1a (EF-1a) in *Drosophila melanogaster* with continuous and stage specific expression. Nucleic Acids Res. 16, 3175–3194.

Huelsenbeck, J.P., Bollback, J.P., Levine, A.M., 2002. Inferring the root of a phylogenetic tree. Syst. Biol. 51, 32–43.

Hyndman, K.A., Miyamoto, M.M., Evans, D.H., 2009. Phylogeny, taxonomy, and evolution of the endothelin receptor gene family. Mol. Phylogenet. Evol. 52, 677–687.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T., 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA 86, 9355–9359.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci.: CABIOS 8, 275–282.

Jordal, B.H., 2002. Elongation factor 1α resolves the monophyly of the haplodiploid ambrosia beetles Xyleborini (Coleoptera: Curculionidae). Insect Mol. Biol. 11, 453–465.

Knudsen, B., Miyamoto, M.M., 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. USA 98, 14512–14517.

Kodandaramaiah, U., Pena, C., Braby, M.F., Grund, R., Muller, C.J., Nylin, S., Wahlberg, N., 2010. Phylogenetics of Coenonymphina (Nymphalidae: Satyrinae) and the problem of rooting rapid radiations. Mol. Phylogenet. Evol. 54, 386–394.

Lake, J.A., Herbold, C.W., Rivera, M.C., Servin, J.A., Skophammer, R.G., 2007. Rooting the tree of life using non-ubiquitous genes. Mol. Biol. Evol. 24, 130–136.

Lake, J.A., Servin, J.A., Herbold, C.W., Skophammer, R.G., 2008. Evidence for a new root of the tree of life. Syst. Biol. 57, 835–843.

Lake, J.A., Skophammer, R.G., Herbold, C.W., Servin, J.A., 2009. Genome beginnings: rooting the tree of life. Philos. Trans. Roy. Soc. B 364, 2177–2185.

LaPolla, J.S., Brady, S.G., Shattuck, S.O., 2010. Phylogeny and taxonomy of the *Prenolepis* genus-group of ants (Hymenoptera: Formicidae). Syst. Entomol. 35, 118–131.

Leys, R., Cooper, S.J.B., Schwarz, M.P., 2002. Molecular phylogeny and historical biogeography of the large carpenter bees, genus *Xylocopa* (Hymenoptera: Apidae). Biol. J. Linn. Soc. 77, 249–266.

Lin, C.P., Danforth, B.N., 2004. How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. Mol. Phylogenet. Evol. 30, 686–702.

Lin, Y.-H., McLenachan, P.A., Gore, A.R., Phillips, M.J., Ota, R., Hendy, M.D., Penny, D., 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol. Biol. Evol. 19, 2060–2070.

Lynch, V.J., 2007. Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. BMC Evol. Biol. 7, 2.

Maddison, D.R., Maddison, W.P., 2000. MacClade 4.0. Sinauer Associates, Sunderland, MA.

Maddison, W.P., Maddison, D.R., 2010. Mesquite: a Modular System for Evolutionary Analyses. Version 2.74 <http://mesquiteproject.org>.

Maddison, W.P., Donoghue, M.J., Maddison, D.R., 1984. Outgroup analysis and parsimony. Syst. Zool. 33, 83–103.

Maroni, G., 1993. An Atlas of *Drosophila* Genes. Oxford University Press, Oxford.

Marshall, D.C., 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. 59, 108–117.

Mathews, S., 2009. Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. Am. J. Bot. 96, 228–236.

Mathews, S., Donoghue, M.J., 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286, 947–950.

Mathews, S., Donoghue, M.J., 2000. Basal angiosperm phylogeny inferred from duplicate phytochromes A and C. Int. J. Plant Sci. 161, S41–S55.

Mathews, S., Clements, M.D., Beilstein, M.A., 2010. A duplicate gene rooting of seed plants and the phylogenetic position of flowering plants. Philos. Trans. Roy. Soc. B 365, 383–395.

Melo, G.A.R., 1999. Phylogenetic relationships and classification of the major lineages of Apoidea (Hymenoptera), with emphasis on the crabronid wasps. Scientific Papers. Univ. Kansas Nat. Hist. Mus. 14, 1–55.

Michener, C.D., 1981. Classification of the bee family Melittidae with a review of species of Meganomiinae. Contrib. Am. Entomol. Inst. 18, 1–135.

Michener, C.D., 2007. The Bees of the World. Johns Hopkins University Press, Baltimore, MD.

Michez, D., Nel, A., Menier, J.-J., Rasmont, P., 2007. The oldest fossil of a melittid bee (Hymenoptera: Apiformes) from the early Eocene of Oise (France). Zool. J. Linn. Soc. 150, 701–709.

Michez, D., Patiny, S., Danforth, B., 2009. Phylogeny of the bee family Melittidae (Hymenoptera: Anthophila) based on combined molecular and morphological data. Syst. Entomol. 34, 574–597.

Milinkovitch, M.C., Lyons-Weiler, J., 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. Mol. Phylogenet. Evol. 9, 348–357.

Morris, D.C., Schwarz, M.P., Cooper, S.J.B., Mound, L.A., 2002. Phylogenetics of Australian *Acacia* thrips: the evolution of behaviour and ecology. Mol. Phylogenet. Evol. 25, 278–292.

Niehuis, O., Judson, A.K., Werren, J.H., Hunter, W.B., Dang, P.M., Dowd, S.E., Grillenberger, B., Beukeboom, L.W., Gadau, J., 2007. Species-diagnostic single-nucleotide polymorphism and sequence-tagged site markers for the parasitic wasp genus *Nasonia* (Hymenoptera: Pteromalidae). J. Econ. Entomol. 100, 1033–1036.

Niehuis, O., Judson, A.K., Gadau, J., 2008. Cytonuclear genic incompatibilities cause increased mortality in male F2 hybrids of *Nasonia giraulti* and *N. vitripennis*. Genetics 178, 413–426.

Nixon, K.C., Carpenter, J.M., 1993. On outgroups. Cladistics 9, 413–426.

Normark, B.B., Jordal, B.H., Farrell, B.D., 1999. Origin of a haplodiploid beetle lineage. Proc. Roy. Soc. B 266, 2253–2259.

Nylander, J.A.A., 2004. MrModeltest v2. Program Distributed by Author. Evolutionary Biology Centre, Uppsala University.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin.

Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53, 571–581.

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Müller, M.H., Le Guyader, H., 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc. Roy. Soc. B 1449, 1213–1221.

Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21, 1455–1458.

Pilgrim, E.M., von Dohlen, C.D., Pitts, J.P., 2008. Molecular phylogenetics of Vespoidea indicate paraphyly of the superfamily and novel relationships of its component families and subfamilies. Zool. Scripta 37, 539–560.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.

Rambaut, A., Drummond, A.J., 2007. Tracer v1.4. <http://beast.bio.ed.ac.uk/Tracer>.

Roig-Alsina, A., Michener, C.D., 1993. Studies of the phylogeny and classification of long-tongued bees (Hymenoptera: Apoidea). Univ. Kansas Sci. Bull. 55, 123–173.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Schultz, T.R., Brady, S.G., 2008. Major evolutionary transitions in ant agriculture. Proc. Natl. Acad. Sci. USA 105, 5435–5440.

Schwarz, M.P., Bull, N.J., Cooper, S.J.B., 2003. Molecular phylogenetics of allodapine bees, with implications for the evolution of sociality. Syst. Biol. 52, 1–14.

Servin, J.A., Herbold, C.W., Skophammer, R.G., Lake, J.A., 2008. Evidence excluding the root of the tree of life from the Actinobacteria. Mol. Biol. Evol. 25, 1–4.

Shavit, L., Penny, D., Hendy, M.D., Holland, B.R., 2007. The problem of rooting rapid radiations. Mol. Biol. Evol. 24, 2400–2411.

Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51, 492–508.

Simon, S., Schierwater, B., Hadrys, H., 2010. On the value of elongation factor-1α for reconstructing pterygote insect phylogeny. Mol. Phylogenet. Evol. 54, 651–656.

Skophammer, R.G., Servin, J.A., Herbold, C.W., Lake, J.A., 2007. Evidence for a Gram-positive, eubacterial root of the tree of life. Mol. Biol. Evol. 24, 1761–1768.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis using Parsimony (*and other methods), v. 4. Sinauer Associates, Sunderland, MA.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147, 63–91.

Valas, R.E., Bourne, P.E., 2009. Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life. Biology Direct. 4, 30.

Walldorf, U., Hovemann, B.T., 1990. *Apis mellifera* cytoplasmic elongation factor 1-alpha (EF-1alpha) is closely related to *Drosophila melanogaster* EF-1alpha. FEBS Lett. 267, 245–249.

Ward, P.S., Brady, S.G., Fisher, B.L., Schultz, T.R., 2010. Phylogeny and biogeography of dolichoderine ants: effects of data partitioning and relict taxa on historical inference. Syst. Biol. 59, 342–362.

Wheeler, W.C., 1990. Nucleic acid sequence phylogeny and random outgroups. Cladistics 6, 363–368.

Wiegmann, B.M., Trautwein, M.D., Kim, J.W., Cassel, B.K., Bertone, M.A., Winterton, S.L., Yeates, D.K., 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7, 34.

Woese, C.R., Achenbach, L., Rouviere, P., Mandelco, L., 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14, 364–371.

Wortley, A.H., Rudall, P.J., Harris, D.J., Scotland, R.W., 2005. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. Syst. Biol. 54, 697–709.

Yang, Z., 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39, 105–111.

Zanis, M.J., Soltis, D.E., Soltis, P.S., Mathews, S., Donoghue, M.J., 2002. The root of the angiosperms revisited. Proc. Natl. Acad. Sci. USA 99, 6848–6853.

Zhaxybayeva, O., Lapierre, P., Gogarten, J.P., 2005. Ancient gene duplications and the root(s) of the tree of life. Protoplasma 227, 53–64.

Zwickl, D.J., 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion. Ph.D. Dissertation, The University of Texas at Austin.