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

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 clocks (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 then compared against 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.
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/beecheklist.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 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 AAC GAT CGS AY-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 Prep DNA Purification Kit. Automated sequencing of both strands was accomplished using the same amplification primers through the Cornell Life Sciences Core Laboratories (http://cores.life sciences.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...
Table 1

<table>
<thead>
<tr>
<th>Family</th>
<th>Subfamily</th>
<th>Genus and species</th>
<th>DNA code</th>
<th>F1 Copy</th>
<th>F2 Copy</th>
<th>Locality</th>
<th>Coll.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crabronidae</td>
<td>Bembicinae</td>
<td>Bicyrtes ventralis</td>
<td>Ay63036</td>
<td>1</td>
<td>1</td>
<td>USA: NY: Oswego Co., Selkirk Shores SP, August 5 2000</td>
<td>Ascher</td>
</tr>
<tr>
<td>Crabronidae</td>
<td>Bembicinae</td>
<td>Stictodes sp.</td>
<td>JF006306</td>
<td>1</td>
<td>1</td>
<td>USA: AZ: Cochrine Co., Buckers Canyon, 9 mi W US80, August 17 2000</td>
<td>Danforth</td>
</tr>
<tr>
<td>Crabronidae</td>
<td>Crabroninae</td>
<td>Oxypeltis</td>
<td>JF006305</td>
<td>1</td>
<td>1</td>
<td>USA: NY: Tompkins Co., Ithaca, July 23 2000</td>
<td>Danforth</td>
</tr>
<tr>
<td>Sphicidae</td>
<td>Sphicininae</td>
<td>Sphicis corentumurum</td>
<td>Ay63034</td>
<td>1</td>
<td>1</td>
<td>USA: NY: Tompkins Co., Ithaca, July 23 2000</td>
<td>Danforth</td>
</tr>
<tr>
<td>Andrenidae</td>
<td>Andrenaceae</td>
<td>Andrenus spinosus</td>
<td>JF006309</td>
<td>1</td>
<td>1</td>
<td>USA: CA: Santa Clara Co., Mt Hamilton, May 28 1999</td>
<td>Ascher</td>
</tr>
<tr>
<td>Andrenidae</td>
<td>Xylocopa</td>
<td>Xylocopa spinosus</td>
<td>JF006312</td>
<td>1</td>
<td>1</td>
<td>USA: CA: Riverside Co., San Jacinto WA, May 18, 1993</td>
<td>Danforth</td>
</tr>
<tr>
<td>Megachilidae</td>
<td>Megachilinae</td>
<td>Megachile pugnata</td>
<td>Ay63025</td>
<td>1</td>
<td>1</td>
<td>USA: NY: Tompkins Co., Ithaca, July 27 1999</td>
<td>Danforth</td>
</tr>
</tbody>
</table>
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 (Dasypodinae, Meganominae, 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 melitid 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 10 terminals and 282 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-nee 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 brnsnr = 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 decrease tree space, a long chain recommended 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) Dasypodinae (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.)
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
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/Bayes-Phylogenies. LT, long tongued bees (Apidae + Megachilidae).

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Concatenated</th>
<th>F1 copy</th>
<th>Duplicate, F1 copy</th>
<th>F2 copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrenidae</td>
<td>1/1/1</td>
<td>&lt;0.5/0.5/0.5</td>
<td>0.5/0.5/0.5</td>
<td>&lt;0.5/0.5/0.5</td>
</tr>
<tr>
<td>Apidae</td>
<td>1/1/1</td>
<td>&lt;0.5/0.5/0.5/0.5/0.74</td>
<td>0.5/0.5</td>
<td></td>
</tr>
<tr>
<td>Colletidae</td>
<td>1/1/1</td>
<td>0.81/0.81/1.00/1.00/1.00/1.00</td>
<td>0.5/0.57/1</td>
<td></td>
</tr>
<tr>
<td>Halictidae</td>
<td>1/1/1</td>
<td>1/0.96/1</td>
<td>1.0/0.96/1</td>
<td>1.0/0.96/1</td>
</tr>
<tr>
<td>Megachilidae</td>
<td>1/1/1</td>
<td>0.99</td>
<td>1/0.99/1</td>
<td>1/0.99/1</td>
</tr>
<tr>
<td>Melitidae</td>
<td>0.97/0.5/0.5</td>
<td>&lt;0.5/0.5/0.5/0.5</td>
<td>&lt;0.5/0.5/0.5/0.5</td>
<td>&lt;0.5/0.5/0.5/0.5</td>
</tr>
<tr>
<td>LT bees</td>
<td>1/0.91/0.87/1/0.91/0.87</td>
<td>0.5/0.5/0.5/0.5/0.5/0.5</td>
<td>1/0.91/0.87/1/0.91/0.87</td>
<td></td>
</tr>
<tr>
<td>Colletidae + Stenotritidae</td>
<td>1/0.97/1</td>
<td>1.0/0.5/1</td>
<td>1.0/0.5/1</td>
<td>1.0/0.5/1</td>
</tr>
<tr>
<td>Halictidae + (Colletidae + Stenotritidae)</td>
<td>1/0.91/0.96</td>
<td>&lt;0.5/0.5/0.5/0.5/0.5/0.5/0.5</td>
<td>0.5/0.87</td>
<td></td>
</tr>
</tbody>
</table>

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 Melitidae 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 data-sets (Brady et al., 2008; LaPolla et al., 2010; Schultz and Brady, 2008). There is insufficient data from other lineages of Hymenop-tera to determine if this phenomenon holds true for the entire or-der. 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-spe-cific 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 con-served 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 fu-ture 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, judged 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 resolu-tion seen in the analyses using paralog rooting compared to out-group 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 re-lated outgroups may confound reconstruction of ingroup relation-ships (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 Dip-tera 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 holometab-olous 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 separat-ing 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 approxi-mately 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
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, recovery of the correct tree topology is improved. This is illustrated in Fig. 4, which shows the 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).
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 paralogy 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 (Nymphaeae), 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 SRa (Brinkmann and Philippe, 1999), and three copies of aminoacyl-tRNA synthetase genes (Brown and Doolittle, 1995). Most studies recovered euacteria as the sister lineage to a clade consisting of Archaea 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 Archaeabacteria. A further study (Valas et al., 2006) 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
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 Melittidae. 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 (Michener 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


