Received: 17 July 2018 | Revised: 19 October 2018 | Accepted: 11 November 2018

DOI: 10.1111/zsc.12332

ORIGINAL ARTICLE

WILEY Zoologica Scripta

Exploring gene tree incongruence at the origin of ants and bees (Hymenoptera)

Gabriela P. Camacho[1,2] | **Marcio R. Pie**[1,3] | **Rodrigo M. Feitosa**[1] | **Marcos S. Barbeitos**[3]

[1]Programa de Pós-Graduação em Entomologia, Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, Brazil

[2]Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, District of Columbia

[3]Programa de Pós-Graduação em Zoologia, Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, Brazil

Correspondence
Gabriela P. Camacho, Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC.
Email: gabipcamacho@ymail.com
Marcos S. Barbeitos, Programa de Pós-Graduação em Zoologia, Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, Brazil.
Email: msbarbeitos@gmail.com

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 301636/2016-8 and 302462/2016-3; Smithsonian Institution

Abstract

The fact that different phylogenomic data sets can lead to highly supported but inconsistent results suggest that conflict among gene trees in real data sets could be severe. We provide here a detailed exploration of gene tree space to investigate the relationships in Hymenoptera based on data obtained by Johnson et al. (*Current Biology*, 2013, 23, 2058), in which ants and Apoidea (bees and spheciform wasps) were recovered as sister groups, contradicting previous studies. We found high levels of topological variation among gene trees, several of them disagreeing with previously published hypotheses. To profile the dynamics of emerging support versus conflicting signal in combined analysis of data, we employed a novel method based on the incremental addition of randomized data to coalescence-based phylogenetic inference. Although the monophyly of Aculeata and of Formicidae were consistently recovered using as little as 6.5% of the 308 available markers, signal for the Formicidae + Apoidea clade prevailed only after more than 50% of the loci were sampled. Still, non-negligible support for alternative hypotheses remained until all genes were added to the analysis. Our results suggest that phylogenetic conflict is rather pervasive and not scattered as noise across individual gene trees because alternative topologies were recovered not from a specific subset, but from several random combinations of loci. Thus, even though phylogenetic signal recovered from full gene data sets was already dominant in much smaller ensembles, large amounts of data may be indeed necessary to overcome phylogenetic conflict.

KEYWORDS

Apidae, coalescent theory, emergent support, gene trees, RADICAL, species trees

1 | INTRODUCTION

One of the hopes of analyzing extensive multi-locus datasets, like the ones generated through next-generation sequencing, is the possibility of solving many vexing problems in phylogenetics, such as the higher relationships among bird lineages (McCormack et al., 2013), the phylogeny of placental mammals (McCormack et al., 2012), or the phylogenetic position of turtles in tetrapods (Crawford et al., 2012). These problems are stimulating particularly due to the potential incongruence between species trees and their underlying gene trees (Maddison, 1997). Several factors are said to be causing phylogenetic incongruence, such as phylogenetic estimation error, homoplasy, lateral gene transfer, incomplete lineage sorting and introgression (Mallet, Besansky, & Hahn, 2016). In fact, the assumption of a single evolutionary history for all loci that underlies the hitherto common practice of concatenation might cause the resulting inferences to be statistically inconsistent, particularly near relatively short internodes (Degnan & Rosenberg, 2006; Edwards, Liu, & Pearl, 2007). As a consequence, there has been a progressively stronger reliance on methods based on the multispecies coalescent model (Rannala & Yang, 2003; Takahata, Satta, & Klein, 1995), which often assumes that discrepancies between gene trees and the species tree are exclusively due to deep coalescence.

A tacit assumption of many phylogenomic efforts is that one could solve long-standing problems in molecular phylogenetics by "brute force." For instance, studies using only a handful of loci frequently show some nodes with relatively low support, which are followed by the author's suggestion that more data are necessary to resolve that particular issue. However, as more extensive data sets are increasingly common, computation time of many species tree methods based on the multispecies coalescent model such as BEST (Liu, 2008) and *BEAST (Heled & Drummond, 2010) is becoming increasingly impractical. A heuristic alternative is the use of a two-step process, that is first inferring gene trees for each locus using a robust method such as maximum likelihood and then using these trees as input for species tree inference. Several freely available software such as STAR (Liu, Yu, Pearl, & Edwards, 2009), STEAC (Liu et al., 2009), GLASS (Mossel & Roch, 2010), MP-EST (Liu, Yu, & Edwards, 2010), NJst (Liu & Yu, 2011) and ASTRAL-II (Mirarab & Warnow, 2015) implement this approach. Those methods are statistically consistent to the multispecies coalescent model if gene trees are known without error (Allman, Degnan, & Rhodes, 2013; Liu & Yu, 2011; Liu et al., 2009). Yet, the extent to which this assumption is violated in real data sets is poorly known.

The fact that different phylogenomic data sets can lead to highly supported, yet mutually inconsistent results is not only alarming, but also suggests that inconsistency among gene trees in real data sets could be severe. For instance, Johnson et al. (2013) used transcriptome data on 19 species from all superfamilies of aculeate Hymenoptera (wasps, ants and bees) and found that ants (Formicidae) and Apoidea (bees and spheciform wasps) were sister groups. This result is surprising given previous morphological (Brothers, 1999: phylogenetic analysis of 92 morphological characters), molecular (Heraty et al., 2011: one mitochondrial and three nuclear genes) or total-evidence hypotheses (Pilgrim, Dohlen, & Pitts, 2008: four nuclear genes plus the morphological matrix of Brothers (1999)) that, although contradicting each other, never recovered topologies similar to the study of Johnson and cols (see Supporting Information Figure S1). Alternatively, Faircloth, Branstetter, White, and Brady (2014) used another data set based on 638 ultraconserved elements (UCEs, Faircloth et al., 2012) loci to infer the relationships among 44 taxa from six aculeate superfamilies, which placed ants at the base of the aculeate tree, as sister to the remaining aculeate lineages. More recently, Branstetter et al. (2017) showed that taxon sampling may be the cause for inconsistency among those two previous phylogenomic studies by analyzing 854 UCE loci for 187 taxa, including a broader sampling of ant groups and representatives of different parasitoid wasps (Chrysidoidea). This result was corroborated by Peters et al. (2017) using a data set of 3,256 protein-coding genes in 173 insect species. However, although the sister group relationship between ants, bees and spheciform wasps was recovered by several different studies by now (Branstetter et al., 2017; Johnson et al., 2013; Peters et al., 2017), the cause and degree of those conflicting sources of phylogenetic signal is currently unknown. In the present study, we explore topological variation in gene trees to provide a detailed exploration of gene tree space in the data set of Johnson et al. (2013), to access the variation and inconsistency present in the data and how the final species tree recovers the position of ants as sister to apoid wasps and bees.

2 | MATERIAL AND METHODS

2.1 | Data set

We focused on the data generated by Johnson et al. (2013) as a test-case to assess incongruence among gene trees and their effects on phylogenetic inference. Although Johnson et al. (2013) analyzed four different data sets with varying levels of missing data, we chose to focus only on their most complete matrix, with 308 genes and 19 taxa (175,404 sites, of which 73.42% are coded as amino acids, 11.60% are gaps, and 14.98% are missing) to avoid conflating the effects of missing data and gene tree incongruence. The matrix was obtained directly from Dryad (http://doi.org/10.5061/dryad.jt440).

2.2 | Gene tree and species tree estimation

Each gene tree was inferred by maximum likelihood (ML) using RAxML's v8.2.x (Simmons, 2014) implementation at the Supercomputing Center of the Ohio State University (Ohio Supercomputer Center, 1987), USA. We used the same models of evolution for each gene chosen in the original study, selected after testing 36 possible protein models (see Johnson et al., 2013 for details). Branch support on each gene tree was determined by bootstraping, using stabilization of the majority-rule consensus tree as stopping criterion. Given that some of the genes reached convergence (i.e., bootstrapping was halted by RAxML) before the completion of 1,000 replicates, the consensus gene trees were obtained from 100 randomly selected bootstrap pseudo-replicates for each of the 308 genes. We used 100 pseudo-replicates because this is the number usually referred as representing sufficient replicates for phylogenomic data sets (see Blaimer, Lloyd, Guillory, & Brady, 2016; Branstetter et al., 2017; Johnson et al., 2013). Trees were rooted using *Nasonia vitripennis* Walker, 1836 (Pteromalidae) as the outgroup.

We estimated species trees based on two different summary statistics approaches. First, we used methods based on the average time of gene coalescence events (consensus method), as implemented in STAR (Liu et al., 2009) and ASTRAL-II v4.7.12 (Mirarab & Warnow, 2015). Second, we used MP-EST (Liu et al., 2010) to generate a maximum pseudo-likelihood estimate under the multispecies coalescent model (gene tree-based coalescent method). As input for the species tree analysis, we used all individual gene trees and accompanying bootstrap trees as input (100 replicates). Bootstrap support values were calculated based on multilocus bootstrapping method by Seo (2008) for all the species tree analyses.

2.3 | Evaluation of individual genes support

We used five metrics that could potentially indicate locus informativeness for gene tree inference, namely locus length (number of base pairs), the proportion of parsimony informative sites, mean bootstrap support across all nodes of the gene tree, number of bootstrap replicates needed for a stable consensus and the Robinson–Foulds (RF—Robinson & Foulds, 1981) distances between individual gene trees and the reference species tree obtained from the full data set. RF distance is a metric that determines the number of bipartitions that differ between a pair of trees to indicate the amount of topological discordance between them. To calculate the percentage of parsimony informative sites by gene, we used the package IPS 0.0-7 (Heibl, 2014) for R 3.1.3 (R Core Team, 2017). The level of correlation between the metrics was assessed by a linear model using the package Vegan 2.4–2 (Oksanen et al., 2018). In order to evaluate the emerging support in individual gene trees, we used custom Python scripts as wrappers for Dendropy's methods (Sukumaran & Holder, 2010), to count the number of bootstrap pseudo-replicates that were compatible with each one of the aforementioned hypotheses. We used the same approach to quantify individual gene tree support for each of the 18 nodes in the reference topology. Results were summarized as heatmaps using the packages STATS v3.4.0 and RCOLORBREWER v1.1-2 (Neuwirth, 2007).

Topological variation among all ML gene trees was assessed by computing pairwise RF distances using the package PHANGORN 2.0.3 (Schliep, 2011). A matrix of RF distances among gene trees was subjected to multidimensional scaling (MDS, see Hillis, Heath, & John, 2005). We focused on the first two ordination axes and coloured the obtained results according to their support for each of five alternative sister group relationships for ants, namely Brothers (1999), Faircloth et al (2014), Heraty et al. (2011), Johnson et al. (2013), and Pilgrim et al. (2008). In order to estimate the phylogenetic signal favouring these different hypotheses in individual gene trees, we counted the number of bootstrap pseudo-replicates obtained using RAxML that were compatible with the same five alternatives.

Finally, to better visualize the similarities between the gene trees topologies, we also employed MetaTree v2 (Nye, 2008). As the name indicates, this algorithm builds a "tree of trees", in which similar topologies cluster together, so that conflicting evolutionary histories within a set of trees become apparent as separate "clades." Internal nodes defining these "clades" correspond to the 50% majority-rule consensus trees computed from the descendants of the node and branch lengths are scaled according to RF distances. For our MetaTree analyses, we used as input the 308 ML gene trees generated by RAxML.

2.4 | Evaluation of emergent support

Random Addition Concatenation Analysis (RADICAL) is a method that evaluates the effects of partitioning and combining genes in genome-level analyses by assessing the effect of increasingly larger matrix sizes on phylogenetic inference (Narechania et al., 2012). Each matrix is randomly assembled along a set of concatenation steps that range from one to all genes in the data set. Thus, one set of steps makes up an "analysis path" that goes from a single gene tree to the species tree obtained from the full data set. In order to account for the stochastic effect of randomly concatenating genes at each step, analysis paths are replicated a number of times. RADICAL catalogs tree heterogeneity while allowing the visualization of emergent support through concatenation, that is, how the addition of new genes influences tree topology and support. Moreover, RADICAL monitors the dynamics of concatenation by calculating support statistics for the topologies generated at each step and comparing them to the whole library of trees (Narechania et al., 2012). Since our aim is to assess incongruence among individual gene trees and the resulting species trees, we developed a custom pipeline using the same logic, but employing the coalescent approach instead.

We started with 100 randomly chosen pairs of gene trees as first steps to "analysis paths" that were incrementally constructed by adding gene trees to each of these sets. In RADICAL, this is done by randomly concatenating genes (without replacement) and analyzing the new alignment using RAxML. Our approach estimated a new species tree at each step of the path using MP-EST. In order to decrease execution time, steps were constructed using 18-gene batches (approximately the square root of 308, so each path is also made up of 18 steps). Thus, the first step of a path was the species tree corresponding to 2 (randomly) chosen genes, the second was a species tree inferred from 20 genes (two genes from the first step plus 18 additional genes, also randomly sampled without replacement), the third was inferred from 38 genes, etc. The analysis ended when all 308 genes were added to each replicate. Thus, the last tree in each path was the "total-evidence" species tree.

With these data in hands, we employed the same approach used by the authors of RADICAL to evaluate the phylogenetic support for each node of that tree that is, calculating, at each step, the percentage of replicate species trees that contains a given node present in the total-evidence tree. If phylogenetic signal for any given node is strong in the majority of genes, it will be present in most species trees already at the first steps. If a node shows up in all (100%) species trees after a certain step, it has become "fixed" in the analysis. The strength of the phylogenetic signal supporting that node is thus inversely proportional to the number of steps required for fixation. In other words, the data provide much stronger evidence for a node that goes to fixation after, for instance, two steps in a path (when species tree were computed from random combinations of 20 genes) than after 10 steps, when much larger randomized data sets (164 genes) were used.

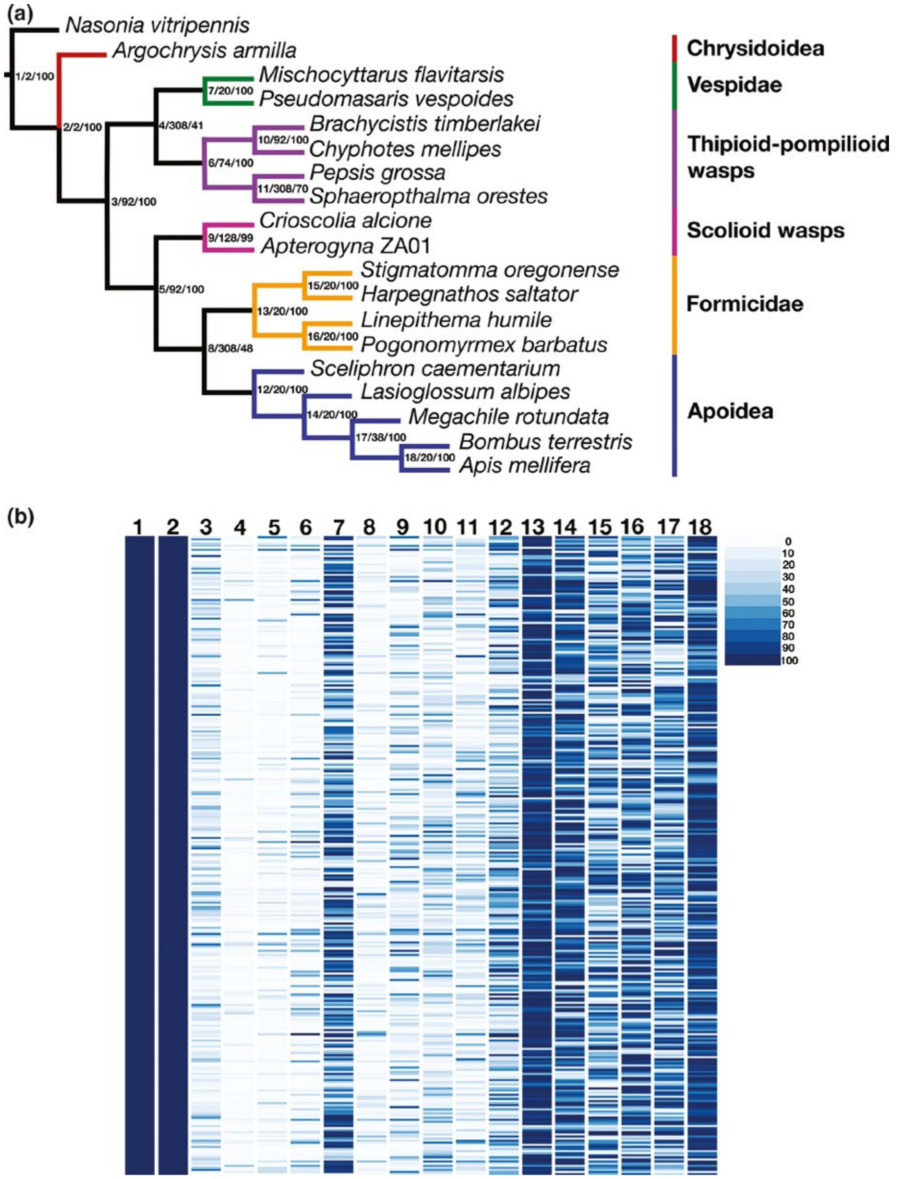

FIGURE 1 (a) Aculeate Hymenoptera species tree resulting from a MP-EST analysis of 308 individual gene trees from Johnson et al. (2013) data set. Annotations are node number/number of genes necessary for the clade to be present in 100 species tree replicates of the emergent support analysis/bootstrap values. (b) Heatmap depicting the percentage of times that each node of the reference topology (a) was found in 100 randomly selected bootstrap pseudo-replicates of the 308 gene trees. Node 1 was trivially recovered in all pseudo-replicates due to rooting. Node 2 was also recovered in all pseudo-replicates of all genes. Each line along the columns corresponds to one gene and the number of bootstrap replicates recovering the node is colour coded according to the scale on the right

We also calculated the normalized Consensus Fork Index (CFI) (Colless, 1980) by counting the number of identical nodes between the total-evidence species trees and each replicate species tree along the 100 paths, divided by the maximum number of nodes possible (18). Normalized CFIs vary between 0 and 1, where 0 indicates total-evidence and replicate species trees with no nodes in common while 1 means identical trees. The distribution of normalized CFIs at each step were represented by density kernels, using the Vioplot v0.2 package (Hintze & Nelson, 1998).

3 | RESULTS

3.1 | Species tree estimation

The species tree analyses recovered a phylogeny for the aculeate Hymenoptera supporting the hypothesis of ants as sister group to Apoidea, regardless of the method, as seen in Johnson et al. (2013). However, topologies and support values varied among the different methods of estimation. MP-EST recovered a tree with much lower bootstrap support for the relationship between Formicidae and Apoidea than the other methods (Supporting Information Figure S2). STAR and ASTRAL-II produced the same topology as Johnson et al. (2013), with slightly different bootstrap support values (Supporting Information Figure S2). Also, the MP-EST recovered a sister group relationship between Vespoidea and the tiphioid-pompilioid wasps with low support, a result that is congruent with the findings of Branstetter et al. (2017), but not found in the STAR analysis by Johnson et al. (2013) or in the concatenated analysis by Peters et al. (2017). We choose the MP-EST species tree as our working hypothesis, numbering each node of the reference topology for the assessment of individual gene support and emergent support (Figure 1a).

3.2 | Evaluation of individual gene support

Many of the 308 loci were short (ranging from 141 to 1,907 amino-acid sites, with a mean of 570 sites per locus), the number of parsimony informative sites within each locus was low (ranging from 2 to 115, with a mean of 28 per locus) and so was the mean bootstrap support recovered from each gene tree (between 21 and 79, with a grand mean of 48). We noticed that bootstrapping via the consensus stabilization method was not halted until 1,000 bootstrap replicates in most cases, suggesting that phylogenetic signal supporting any particular topology is also weak in these genes. The five metrics that could potentially indicate locus informativeness for gene tree inference showed very little statistical correlation to each other (Supporting Information Figure S3). The mean bootstrap support of each gene tree showed significant correlation with the number of replicates of bootstrap needed to achieve a consensus ($R^2 = 0.07$), with the locus length ($R^2 = 0.28$) and with the percentage of parsimony informative sites per loci ($R^2 = 0.60$). The locus length was also correlated with the number of bootstrap replicates ($R^2 = 0.05$). However, the *r*-squared values are low and show that the correlation between those metrics is not strong. It is also important to notice that the values that do show some correlation to each other are not independent, that is, correlations may be spurious. On the other hand, as shown in Figure 1b, several nodes that had full bootstrap support in the reference species tree (Figure 1a) had also high bootstrap support in individual gene trees (e.g., 13, 14 and 18). In contrast, nodes 5 and 7 had full support in the final species tree but varied significantly in individual tree support (Figure 1b).

Regarding the topological variation, we noticed that there was very little phylogenetic signal in favour of two of the hypothesized Formicidae placements (Brothers, 1999; Pilgrim et al., 2008) and not a single bootstrap pseudo-replicate was compatible with the other two reference topologies in Figure 1 (i.e., Heraty et al., 2011 and Faircloth et al., 2014). The constraint analysis using the RF distance between gene trees showed that only 39 out of 308 gene trees (12%) were consistent with the final species tree hypothesis that ants and Apoidea are sister groups (Supporting Information Figure S4a, in blue) and, on average, only 5.6% of the bootstrap pseudo-replicates supported this relationship in individual gene trees. Additionally, we noticed that these gene trees were not topologically similar, that is, the data were not biased in favour of this particular relationship. The 39 gene trees that are congruent with the hypothesis of Johnson et al. (2013) were not clustered neither in the NMDS space (Supporting Information Figure S4a) nor in the meta tree (Supporting Information Figure S4b, also in blue). Leaf branches in our meta tree are long, showing that the gene trees were quite dissimilar, and the internal vertices correspondingly consist of highly unresolved consensuses, as evidenced by their proximity to the root node (Supporting Information Figure S4b). Long edges radiating from a few central vertices are typical of data sets with a high degree of conflict (Nye, 2008).

3.3 | Evaluation of emergent support

The number of genes necessary for fixation varied several folds among non-trivial nodes from 2 (node 2) to 308 (nodes 4, 8 and 11, Figure 1a). In fact, several of the nodes on the species tree were recovered early during the addition process (Figure 2a), despite substantial amount of signal conflict among individual gene trees, as evidenced in the previous section. For instance, node 12 (Apoidea) appears in all (100) replicate species trees computed from randomized samples equal or larger than 20 genes, that is, fixation happened after 2 steps along the path (Figure 2a). Nodes 3 (Aculeata) and 10 (tiphioid complex) were fixed after 6 steps (92 genes), while node 9 (Scolioidea) required 8 steps (128 genes, Figure 2a).

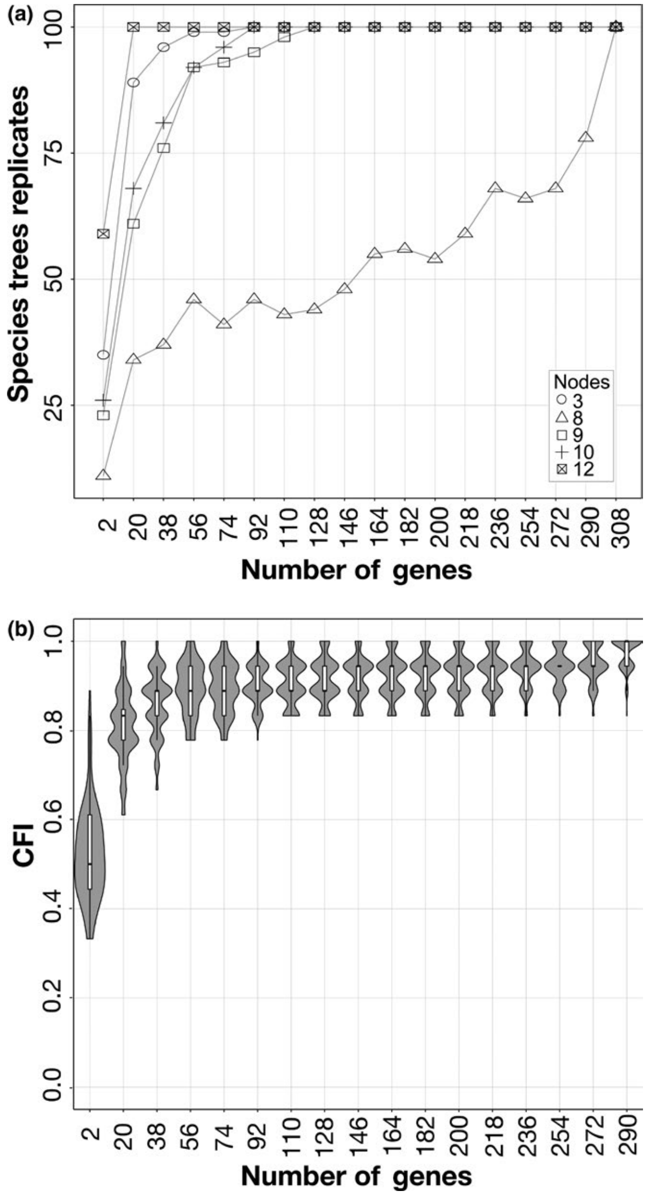

FIGURE 2 (a) Example of fixation paths for selected nodes with different levels of support. Nodes are numbered according to Figure 1a. (b) "Violin plot" of Consensus Fork Index (CFI) values by the gene addition steps. The width of each kernel corresponds to the number of replicate species trees, white dots within the kernels represent median CFI values, thick and thin lines are inter-quartile and min–max ranges, respectively. Normalized CFIs vary between 0 and 1, where 0 indicates trees with no nodes in common with the reference topology (Figure 1a) and 1 are tree with topology identical to the reference tree

In contrast, after 17 steps (290 genes), node 8 (Formicidae and Apoidea) is present in fewer than 80% of the replicate species trees (Figure 2a). While the fixation paths for other nodes showed clear asymptotic trends, increasing monotonically with the number of analyzed genes, node 8's path is much "wobblier" and, the percentage of replicate species trees containing this node decreased frequently among successive steps. Fixation of the sister group relationship between ants and apoids happened somewhere between steps 17 and 18, requiring at least 291 genes.

When averaged across all nodes, 50% of all the emergent support on the tree occurs by the time 20 genes have been added to the coalescence analysis, a data set size that comprises only 6.5% of the total gene space (Figure 2b). As more genes were added to the analysis, the CFI values increased steadily with the number of genes and showed clear convergence towards the reference topology (Figure 1a). It is interesting to note that, after step 8, the median CFI (white dot) was 0.94, meaning that half of the replicate species trees differed by 1 or 0 nodes from the reference topology (Figure 1a); the other half differed by 2 or 3 nodes. Median CFI became 1 somewhere between steps 16 and 17 (273–290 genes), indicating that, by then, half of the tree replicates were identical to the reference topology.

With the exception of node 4 (Vespidae + tiphioid-pompilioid wasps), 8 (Formicidae + Apoidea) and 11 (Pompilidae + Mutillidae), all nodes underwent fixation after step 8 (128 genes, Figure 1a). Because there were three descendant terminals from each one of these three nodes, the tree space after 128 genes consists of 27 (3^3) fully resolved topologies. In order to test how these alternatives were recovered throughout the fixation paths, we used the function resolveAllNodes from Phytools v0.6–20 (Revell, 2012) to enumerate all these topologies. We verified that they were not equally represented: Figure 3a shows each of the 27 topologies and their frequencies among the 100 replicates at each step. With 128 genes, three topologies had equal representation among the species trees, the black line being the one consistent with the reference tree. Signal in its support overwhelmed the alternatives after step 15 (254 genes or more than 80% of the data set, Figure 3a).

Scoring the tree space according to the position of Formicidae reduced it to three alternative topologies (Figure 3b) and only one was consistent with a published hypothesis (Johnson et al., represented with circles in Figure 3b). The sister group relationship between ants and a clade formed by Scolioidea + Apoidea (triangles in Figure 3b) was as frequent as the topology recovered by Johnson and coworkers up to step 2 (20 genes), gradually losing representation as more loci were added to the path. A similar pattern was found for the third possibility, that is, the sister group relationship between Apoidea and a clade formed by Formicidae + Scolioidea (squares in Figure 3b). In the parlance of Narechania et al. (2012), the signal supporting the two unpublished relationships "degraded" as more genes were added to the problem. After the second step of the fixation path (20 genes), signal favouring Johnson et al. became stronger than its alternatives and dominant somewhere between 9 and 10 steps (146–164 genes) when its frequency surpassed the sum of the other two possible placements (i.e., more than 50% of the replicate topologies). The unpublished hypotheses "fought for second place" between steps 3 and 10 (38–164) after which the signal in support of Formicidae + Scolioidea faded away (Figure 3b).

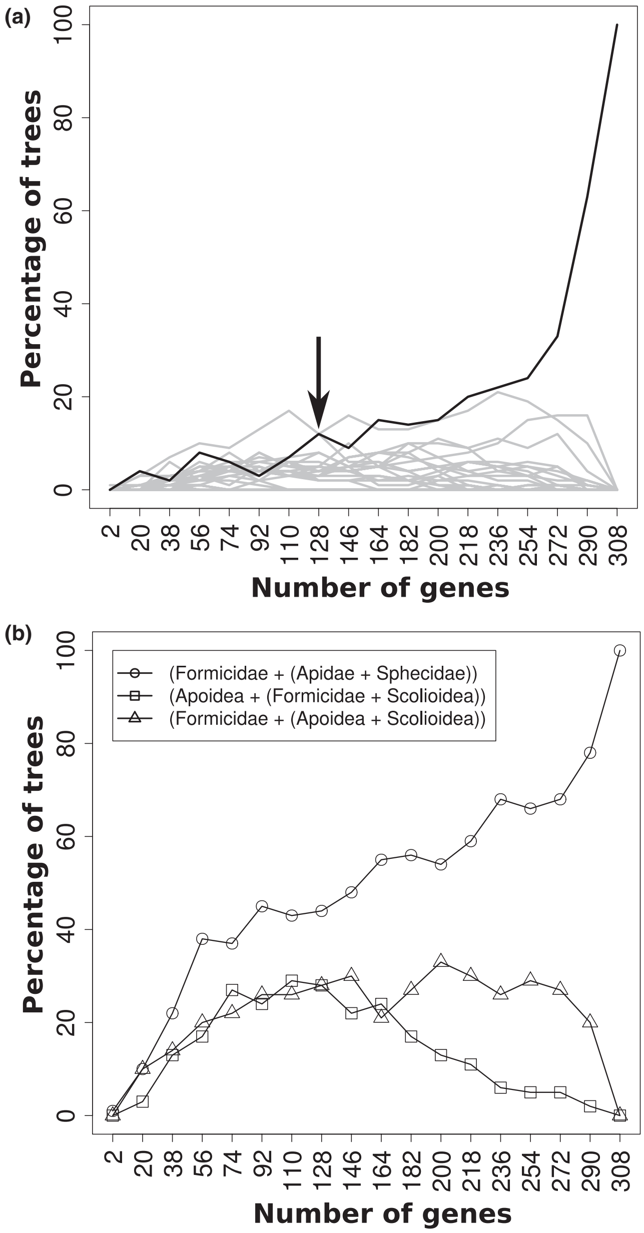

FIGURE 3 (a) Graphical illustration of the topology frequencies of the 27 possible fully dichotomous alternatives that make up the tree space after the fixation of 15 out of 18 nodes of the reference topology. The arrow shows step 8, where the frequency of three of those topologies are identical. (b) Graphical demonstration of the frequency among replicates for the three-main hypothesis recovered by this data set

4 | DISCUSSION

4.1 | Signal conflict

This study highlights the importance of using many loci when tackling difficult phylogenetic problems. Besides demonstrating the weakness of the phylogenetic signal in individual gene trees favouring either of the previously published hypothesis, our results suggest that the use of large data matrices might be necessary to overcome conflicting emergent phylogenetic signal arising from different gene sets. Improved sampling could reduce the probability that the signal of one of these sets will overcome competing information and generate species trees with biased topologies. For most of the nodes in Johnson et al. (2013) data, there was rapid convergence on well-accepted relationships, such as the monophyly of aculeates or ants. The use of a smaller data set might have obscured the general agreement between gene tree topologies, because emergent support for a given topology may not be as strong: the additional data probably enhanced phylogenetic signal through the accumulation of hidden support (Gatesy & Baker, 2005; Gatesy, O'Grady, & Baker, 1999). Those inferences can be made by the use of coalescence-based approaches, as shown by this study, but also by the use of concatenation methods, as demonstrated by Narechania et al. (2012) and recovered by Peters et al. (2017).

The large number of individual gene trees that did not recover Formicidae and Apoidea as sister groups (as found in the species tree), suggests that inconsistency among gene trees was severe, possibly due to incomplete lineage sorting in the data set. Xi Liu and Davis (2015) demonstrated this effect through simulations, but other causes like horizontal transfer, introgression and reticulated evolution could also contribute to swamp the phylogenetic signal in individual genes (Mallet et al., 2016). Explicitly assessing the causes of conflict is, however, beyond the scope of this study.

While some clades are strongly supported in most of the gene trees (i.e., node 13, Formicidae), others had very low support, like the sister group relationship between Vespidae and the tiphioid-pompilioid wasps (Figure 1b). This is the case for the relationship between Formicidae and Apoidea (node 8), indicating that most of the genes used in the analyses did not support this hypothesis. However, other hypotheses recovered in Brothers (1999), Pilgrim et al. (2008), Heraty et al. (2011), and Faircloth et al. (2014), were even more underrepresented, with only a few gene trees recovering the topologies reported in the first two studies, while none recovered results in the latter two. Additionally, despite the high level of discordance among gene trees, those that do agree with Johnson et al. (2013) were neither closely clustered with the final species tree (Supporting Information Figure S4a) nor with each other (Supporting Information Figure S4b). Thus, the data set is coherent and, despite many differences among the estimated gene trees, data were not biased towards any particular topology. Gene tree bias is one major problem in the construction of species trees based in the coalescent model, but this can be alleviated by sampling more genes. This is true even when these genes are minimally informative (Xi et al., 2015). Furthermore, the fact that very few gene trees recovered this clade shows that relying on a single locus or a few loci as a proxy for species trees could be a risky practice (Ruane, Raxworthy, Lemmon, Lemmon, & Burbrink, 2015).

4.2 | Locus quality

The "quality" of the loci (i.e., individual gene informativeness) used in the analysis is still poorly explored in the context of the application of gene tree-based coalescent methods to phylogenomic data. Xi et al. (2015) demonstrated that genes with minimal phylogenetic information can produce unreliable gene trees (i.e., high error in gene tree estimation), which may in turn reduce the accuracy of species tree estimation using gene-tree-based coalescent methods. However, the parameters used to measure gene informativeness (locus length, proportion of parsimony informative sites, mean bootstrap node support, number of pseudo-replicates needed to stabilize the consensus and RF distances between individual gene trees and the total-evidence species tree) were poorly correlated (Supporting Information Figure S3), suggesting that no single parameter is indicative of gene tree informativeness or the construction of reliable gene trees. In this sense, it is clear that narrowing the set of suitable markers for species tree estimation can be a very difficult task, but sampling a large number of genes can alleviate problems of low phylogenetic informativeness. Lack of phylogenetic signal is likely to be especially problematic for those clades with very short internal branches (Townsend, 2007). Mirarab and Warnow (2015) demonstrated that this is especially true in gene tree estimation error using short and thus less informative markers. Under these circumstances, gene trees estimated from alignments with minimal phylogenetic information may reduce the accuracy of gene tree-based coalescent methods (or any coalescent method for that matter), in the same way that uninformative data will reduce the accuracy of concatenation methods.

4.3 | Emergent support

Putative phylogenetic signal by itself does not shed light on questions of gene choice or the optimal number of loci necessary to recover reliable phylogenetic results. The high level of incongruence among data contrasts with the rapidly increasing emergent support observed for some clades as more genes are added to the analysis. Our results are consistent with Narechania et al. (2012), who found that emergent and independent node support are not correlated. This is the case of nodes 3 (Aculeata) and 5 (ants, bees and spheciform wasps), for example, both of which have low individual gene support in this data set (Figure 1b) but are recovered in all species trees after 92 genes (less than 30% of the data) are used (Figure 1a).

When looking at the number of genes necessary to recover different clades in the analysis (Figure 1a), we notice that the support does not increase monotonically. When combinations of 128 genes are used, 14 (or over 3/4) out of the 18 non-trivial internal nodes in the final species tree are recovered in all replicates (Figure 1a). This threshold may be actually smaller because we worked with batches of 18 genes, so the actual number may be anywhere in between 111 and 128 genes. In contrast, the three remaining nodes (4, 8 and 11) only underwent fixation when the number of individual gene trees used in species tree estimation exceeded 290 (or almost 95% of the data; again, because we used 18-genes steps, this result was only recovered for the full data set of 308 genes—Figure 1a).

It is important to emphasize that the upper limit of this fixation threshold (128 genes) is not achieved for a single set of "good quality" gene trees, but for all 100 pseudo-random combinations, sampled from the universe of 308 loci. We believe that these results are more robust and conservative than traditional bootstrapping, because we re-sampled 40% of our data matrix, with a cut-off value of 100%, while bootstrapping employs a cut-off value of 70%–75% (Zharkikh & Li, 1992). This rapid fixation of the majority of nodes reflects the presence of strong emergent phylogenetic signal, a situation in which the accumulation of nodal support is more rapid than would be predicted based on the levels of support on individual gene trees (Gatesy & Baker, 2005; Gatesy et al., 1999). In such cases, congruent phylogenetic signal is amplified as genes are combined during the stepwise addition, whereas divergent patterns of homoplasy specific to single genes or a small set of genes should cancel each other out (Narechania et al., 2012).

Still, the remaining 55%–60% of the loci are necessary to resolve the position of nodes 4 (Vespidae + tiphioid-pompilioid wasps), 8 (Formicidae + Apoidea) and 11 (Pompilidae + Mutillidae). The CFI values (Figure 2b) showed that, after 128 genes, half of the recovered topologies differs from the final species tree by 1 or 0 nodes, but their frequencies are highly variable. In other words, the one node that disagrees with the final species tree (Figure 1a) is not necessarily the same in all trees. At the 128-genes step, three topologies appear with equal representation, while all the other 24 possibilities appear with lower frequencies. Two of these three topologies were recovered at similar frequencies up to 236 genes (roughly 3/4 of the data); after this point, only the topology recovered by the final species tree (in black, Figure 3a) increased in frequency. The remaining topologies cannot be dismissed as negligible noise throughout most of the analysis paths due to their combined frequency being quite large (Figure 3a). These results suggest that there is gene tree incongruence among the data and also phylogenetic conflict which is not diffuse, but pervasive in a significant fraction of the randomly assembled data sets. If loci are discordant, it is expected that numerous additional markers are required to generate a credible species tree. Our study supports this view. The dynamics of emergent support demonstrated that the addition of a large number of loci to the analysis is more than a "brute force" approach to produce a definitive topology as a result of overwhelming data set size. The recovery of specific nodes is not simply the result of additional characters but reflects a disproportionate amplification of phylogenetic signal with the increase of the amount of data (Narechania et al., 2012).

4.4 | Other methods and data sets

Using the data set of Johnson et al. (2013) with a different method (MP-EST), we found an almost identical result as the original authors arrived at by using concatenated data (ML) and species tree analysis (STAR), with the only difference being the sister group relationship between Vespidae and tiphioid-pompilioid wasps. The same result was also recovered by Branstetter et al. (2017) with a different phylogenomic data approach (target enrichment of Ultra-Conserved Elements) using Maximum Likelihood, Bayesian Inference and Species tree analysis (Astral-II). A different study using transcriptomes recovered the same topology as that obtained by Johnson et al. (2013) using STAR, through a concatenated analysis approach (ML) (Peters et al., 2017). These differences in performance by the three methods (STAR/ASTRAL-II and MP-EST) are remarkable, but fairly common, especially while considering that those studies used different data sets, markers and taxon sampling. Still, the main reason for these divergent results may be due to differences in robustness of each algorithm to the influence of missing data on phylogenetic inference (Springer & Gatesy, 2015) and gene tree estimation error (Bayzid & Warnow, 2012). Hence, the choice of coalescence-based method does matter (Mirarab & Warnow, 2015). These divergent results among different data types and analytical approaches show that, although a large number of loci may be important for accurate phylogenetic inference, sufficient taxon sampling (Leebens-Mack et al., 2005; Philippe et al., 2011), choice of alignment method (Wong, Suchard, & Huelsenbeck, 2008), and rigorous tree search (Simmons & Goloboff, 2014) are extremely important for addressing systematic problems. Nevertheless, there is high congruence in the latest results regarding the position of ants among the Aculeata, with all studies recovering a sister group relationship among Formicidae and Apoidea, despite the diversity of previous morphological or single genes hypothesis on the position of ants.

The total-evidence final species tree recovered by MP-EST places Formicidae as the sister group of Apoidea, in a clade where the spheciform wasps represents the early branching lineage (Figure 1a). The support for alternative relationships for ants among the Aculeate appears to present this same pattern in the gene trees, being recovered by a few loci (Supporting Information Figure S4) in very low frequencies. However, in those cases, the conflict between the average gene support was so high that, with the addition of more data, phylogenetic signal favouring these relationships rapidly degraded. In fact, only Pilgrim et al. (2008) and the hypothesis of a sister group relationship between Apoidea and a clade formed by Formicidae + Scolioidea appear to have a relatively higher frequency after the addition of 128 genes. While their combined frequencies are relatively high if compared to the dominant hypothesis (Formicidae + Apoidea), their individual frequencies are much lower than the final species tree after 128 genes, falling fast after 15 steps, that is, after 254 genes were used (Figure 3b).

5 | CONCLUSIONS

The position of ants among the Aculeata appears to be mostly resolved and robustly supported by several phylogenomic studies to this date (Branstetter et al., 2017; Johnson et al., 2013; Peters et al., 2017). The present scenario provides a framework for investigating the evolution of important traits in Hymenoptera, such as nesting, feeding and social behaviour (Branstetter et al., 2017; Johnson et al., 2013), as well as the genomic signatures of changes in these characteristics (Johnson et al., 2013). However, the fact that the target relationship of ants as sister to the apoid wasps and bees is only recovered by the use of 95%–100% of the data suggests that internal conflicts persist. Although several studies, with even larger data sets, have also recovered the same placement for Formicidae, phylogenomic studies could broadly benefit from an exploration of the dynamics of these data sets apart from the traditional measures of support, in order to better assess the nature of any conflicts and the robustness of the emerging results.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian Council of Research and Scientific Development (CNPq) doctorate fellowship and by a Smithsonian Institution Peter Buck Pre-Doctoral Fellowship to GPC; by CNPq/MCT (grant number 301636/2016-8) to MRP; by CNPq/MCT (grant number 302462/2016-3) to RMF. Several analyses were performed using the Supercomputing Center of Ohio State University (OSU), USA. We would like to thank Fred Larabee for valuable comments on earlier versions of this manuscript and the members of the Powell Lab at George Washington University for helpful comments and discussion.

AUTHORS' CONTRIBUTIONS

All authors conceived the ideas, G.P.C., M.S.B. and M.R.P. designed the methodology, M.S.B. built the bioinformatics pipeline, G.P.C., M.S.B. and M.R.P. analyzed the data and G.P.C. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ORCID

Gabriela P. Camacho https://orcid.org/0000-0002-8792-7719
Rodrigo M. Feitosa https://orcid.org/0000-0001-9042-0129

REFERENCES

Allman, E. S., Degnan, J. H., & Rhodes, J. A. (2013). Species tree inference by the STAR method and its generalizations. *Journal of Computational Biology*, *20*(1), 50–61. https://doi.org/10.1089/cmb.2012.0101

Bayzid, M. S., & Warnow, T. (2012). Estimating optimal species trees from incomplete gene trees under deep coalescence. *Journal of Computational Biology*, *19*, 591–605. https://doi.org/10.1089/cmb.2012.0037

Blaimer, B. B., Lloyd, M. W., Guillory, W. X., & Brady, S. G. (2016). Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS ONE*, *11*, e0161531. https://doi.org/10.1371/journal.pone.0161531

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., … Brady, S. G. (2017). Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Current Biology*, *27*, 1019–1025. https://doi.org/10.1016/j.cub.2017.03.027

Brothers, D. J. (1999). Phylogeny and evolution of wasps, ants and bees (Hymenoptera, Chrysidoidea, Vespoidea and Apoidea). *Zoologica Scripta*, *28*, 233–250. https://doi.org/10.1046/j.1463-6409.1999.00003.x

Colless, D. H. (1980). Congruence between morphometric and alloyzme data for *Menidia* species: A reappraisal. *Systematic Zoology*, *29*, 288–299. https://doi.org/10.1093/sysbio/29.3.288

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., & Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, *8*, 783–786. https://doi.org/10.1098/rsbl.2012.0331

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, *2*(5), e68. https://doi.org/10.1371/journal.pgen.0020068

Edwards, S. V., Liu, L., & Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, *104*(14), 5936–5941. https://doi.org/10.1073/pnas.0607004104

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2014). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, *15*, 489–501. https://doi.org/10.1111/1755-0998.12328

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*, 717–726. https://doi.org/10.1093/sysbio/sys004

Gatesy, J., & Baker, R. H. (2005). Hidden likelihood support in genomic data: Can forty-five wrongs make a right? *Systematic Biology*, *54*, 483–492. https://doi.org/10.1080/10635150590945368

Gatesy, J., O'Grady, P., & Baker, R. H. (1999). Corroboration among data sets in simultaneous analysis: Hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*, *15*, 271–313. https://doi.org/10.1006/clad.1999.0111

Heibl, C. (2014). *ips: Interfaces to phylogenetic software in R*. CRAN. Available athttps://CRAN.R-project.org/package=ips

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, *27*, 570–580. https://doi.org/10.1093/molbev/msp274

Heraty, J., Ronquist, F., Carpenter, J. M., Hawks, D., Schulmeister, S., Dowling, A. P., … Sharkey, M. (2011). Evolution of the hymenopteran megaradiation. *Molecular Phylogenetics and Evolution*, *60*, 73–88. https://doi.org/10.1016/j.ympev.2011.04.003

Hillis, D. M., Heath, T. A., & John, K. S. (2005). Analysis and visualization of tree space. *Systematic Biology*, *54*(3), 471–482. https://doi.org/10.1080/10635150590946961

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*, 181–184. https://doi.org/10.2307/2685478

Johnson, B. R., Borowiec, M. L., Chiu, J. C., Lee, E. K., Atallah, J., & Ward, P. S. (2013). Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. *Current Biology*, *23*, 2058–2062. https://doi.org/10.1016/j.cub.2013.08.050

Leebens-Mack, J., Raubeson, L. A., Cui, L., Kuehl, J. V., Fourcase, M. H., Chumley, T. W., … dePamphilis, C. W. (2005). Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein Zone. *Molecular Biology and Evolution*, *22*, 1948–1963. https://doi.org/10.1093/molbev/msi191

Liu, L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, *24*, 2542–2543. https://doi.org/10.1093/bioinformatics/btn484

Liu, L., & Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, *60*, 661–667. https://doi.org/10.1093/sysbio/syr027

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*, 302. https://doi.org/10.1186/1471-2148-10-302

Liu, L., Yu, L., Pearl, D. K., & Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, *58*, 468–477. https://doi.org/10.1093/sysbio/syp031

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, *46*, 523–536.

Mallet, J., Besansky, N., & Hahn, M. W. (2016). How reticulated are species? *BioEssays*, *38*(2), 140–149. https://doi.org/10.1002/bies.201500149

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome Research*, *22*, 746–754. https://doi.org/10.1101/gr.125864.111

McCormack, J. E., Harvey, M. G., Faircloth, B. C., Crawford, N. G., Glenn, T. C., & Brumfield, R. T. (2013). A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS ONE*, *8*, e54848. https://doi.org/10.1371/journal.pone.0054848

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*, i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Mossel, E., & Roch, S. (2010). Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *Transactions on Computational Biology and Bioinformatics*, *7*, 166–171. https://doi.org/10.1109/tcbb.2008.66

Narechania, A., Baker, R. H., Sit, R., Kolokotronis, S. O., DeSalle, R., & Planet, P. J. (2012). Random addition concatenation analysis: A novel approach to the exploration of phylogenomic signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. *Genome Biology and Evolution*, *4*, 30–43. https://doi.org/10.1093/gbe/evr121

Neuwirth, E. (2007). *RColorBrewer: ColorBrewer palettes*. R package version 1.0-2.

Nye, T. M. W. (2008). Trees of trees: An approach to comparing multiple alternative phylogenies. *Systematic Biology*, *57*, 785–794. https://doi.org/10.1080/10635150802424072

Ohio Supercomputer Center. (1987). Columbus, OH. Retrieved from http://osc.edu/ark:/19495/f5s1ph73

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2018). *vegan: Community Ecology Package. R package version 2.5-2*. Retrieved from https://CRAN.R-project.org/package=vegan

Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., … Niehuis, O. (2017). Evolutionary history of the hymenoptera. *Current Biology*, *27*, 1013–1018.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, *9*(3), e1000602. https://doi.org/10.1371/journal.pbio.1000602

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. *Zoologica Scripta*, *37*, 539–560. https://doi.org/10.1111/j.1463-6409.2008.00340.x

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, *164*, 1645–1656.

Revell, L. J. (2012). phytools: A R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*, 217–223. https://doi.org/10.1111/j.2041-210x.2011.00169.x

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131–147.

Ruane, S., Raxworthy, C. J., Lemmon, A. R., Lemmon, E. M., & Burbrink, F. T. (2015). Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: An empirical study on Malagasy pseudoxyrhophiine snakes. *BMC Evolutionary Biology*, *15*, 221.

Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, *27*, 592–593. https://doi.org/10.1093/bioinformatics/btq706

Seo, T. K. (2008). Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, *25*, 960–971. https://doi.org/10.1093/molbev/msn043

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

Simmons, M. P., & Goloboff, P. A. (2014). Dubious resolution and support from published sparse supermatrices: The importance of thorough tree searches. *Molecular Phylogenetics and Evolution*, *78*, 334–348. https://doi.org/10.1016/j.ympev.2014.06.002

Springer, M. S., & Gatesy, J. (2015). The gene tree delusion. *Molecular Phylogenetics and Evolution*, *94*, 1–33. https://doi.org/10.1016/j.ympev.2015.07.018

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A python library for phylogenetic computing. *Bioinformatics*, *26*, 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Takahata, N., Satta, Y., & Klein, J. (1995). Divergence time and population size in the lineage leading to modern humans. *Theoretical Population Biology*, 198–221. https://doi.org/10.1006/tpbi.1995.1026

Townsend, J. P. (2007). Profiling phylogenetic Informativeness. *Systematic Biology*, *56*, 222–231. https://doi.org/10.1080/10635150701311362

Wong, K. M., Suchard, M. A., & Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, *319*, 473–476. https://doi.org/10.1126/science.1151532

Xi, Z., Liu, L., & Davis, C. C. (2015). Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution*, *92*, 63–67.

Zharkikh, A., & Li, W. H. (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution*, *9*(6), 1119–1147. https://doi.org/10.1093/oxfordjournals.molbev.a040782

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Camacho GP, Pie MR, Feitosa RM, Barbeitos MS. Exploring gene tree incongruence at the origin of ants and bees (Hymenoptera). *Zool Scr*. 2018;00:1–11. https://doi.org/10.1111/zsc.12332