The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data

Silas Bossert a,⇑, Elizabeth A. Murray a, Bonnie B. Blaimer b, Bryan N. Danforth a

a Department of Entomology, Cornell University, Ithaca, New York, USA
b Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

Article history:
Received 1 February 2017
Revised 6 March 2017
Accepted 24 March 2017

Keywords:
Gene tree incongruence
Ultraconserved elements
GC biased gene conversion
Gene trees
Sociality
Corbiculates

Abstract
The field of sequence based phylogenetic analyses is currently being transformed by novel hybrid-based targeted enrichment methods, such as the use of ultraconserved elements (UCEs). Rather than analyzing relationships among organisms using a small number of genes, these methods now allow us to evaluate relationships with many hundreds to thousands of individual gene loci. However, the inclusion of thousands of loci does not necessarily overcome the long-standing challenge of incongruence among phylogenetic trees derived from different genes or gene regions. One factor that impacts the level of incongruence in phylogenomic data sets is the level of GC bias. GC rich gene regions are prone to higher recombination rates than AT rich regions, driven by a process referred to as “GC biased gene conversion”. As a result, high GC content can be negatively associated with phylogenetic accuracy, but the extent to which this impacts incongruence among UCEs is currently unstudied. We investigated the impact of GC content on phylogeny reconstruction using in silico captured UCE data for the corbiculate bees (Hymenoptera: Apidae). The phylogeny of this group has been the subject of extensive study, and incongruence among gene trees is thought to be a source of phylogenetic error. We conducted coalescent- and concatenation-based analyses of 810 individual gene loci from all 13 currently available bee genomes, including 8 corbiculate taxa. Both coalescent- and concatenation-based methods converged on a single topology for the corbiculate tribes. In contrast to concatenation, the coalescent-based methods revealed significant topological conflict at nodes involving the orchid bees (Euglossini) and honeybees (Apini). Partitioning the loci by GC content reveals decreasing support for the inferred topology with increasing GC bias. Based on the results of this study, we report the first evidence that GC biased gene conversion may contribute to topological incongruence in studies based on ultraconserved elements.

1. Introduction
The field of phylogenetic analysis is changing rapidly due to the increasing accessibility of genome-scale data (Kjer et al., 2016; Lemmon and Lemmon, 2013; Wheeler and Giribet, 2016). Rather than analyzing phylogenetic relationships among organisms using a small number of genes obtained via PCR and Sanger sequencing, we can now evaluate phylogenetic relationships with many hundreds to thousands of individual loci. This change in scale transforms the tools that we use to resolve the tree of life, and opens new avenues to investigate deep evolutionary history. The inclusion of hundreds to thousands of individual gene loci in phylogenetic analysis, however, remains a challenge because of incongruence among phylogenetic trees of different genes or loci (Galtier and Daubin, 2008; Jeffroy et al., 2006; Nater et al., 2015), and the unequal distribution of topological conflict among clades (Wang et al., 2015). These issues of gene-tree/gene-tree and gene-tree/species-tree incongruence represent a persistent problem that does not necessarily disappear with genome-level datasets. In fact, phylogenomic datasets risk yielding wrong species trees with high confidence if significant incongruence is present, especially with concatenation based methods (Jeffroy et al., 2006; Kubatko and Degnan, 2007).

Topological discordance among gene and species trees can result from either methodological or biological factors (Rokas et al., 2003). Methodological factors are primarily conceptual or analytical limitations, such as the inadequacy of our current substitution models to accurately characterize sequence evolution (e.g., Kolaczkowski and Thornton, 2004; Roure and Philippe, 2011). In contrast, biological factors, such as incomplete coalescence in ancestral populations (incomplete lineage sorting, ILS) (Degnan et al., 2007).
These four tribes are united by the possession of a highly modified (bumblebees), Meliponini (stingless bees), and Apini (honeybees). Ehrlich et al. (2016) placed in four tribes: Euglossini (orchid bees), Bombini (corbiculate bees), with Euglossini at the base of the tree (Fig. 1C; Cardinal and Packer, 2007; Plant and Paulus, 2016), suggesting a single origin of eusociality in corbiculate bees. However, numerous molecular studies (e.g., Cardinal and Danforth, 2011; Kawakita et al., 2008; Martins et al., 2014) have supported a very different view: that Bombini and Meliponini form a well-supported monophyletic group which is sister to Apini + Euglossini (Fig. 1A), thereby implying two separate origins of advanced eusociality. A recent phylogenomic nivation accounted for GC bias in coding sequences obtained yet another topology. Romiguier et al. (2015) obtained a tree in which Apini is sister to the clade of Bombini + Meliponini (Fig. 1B), with strong support when utilizing the non-homogeneous GC98 substitution model (Galtier and Gouy, 1998).

Besides the high degree of topological conflict, two aspects render the corbiculate bees an ideal group for investigating the impact of GC-bias on incongruence among ultraconserved elements. First, this group has been the focus of research on the genetic basis of sociality in bees and therefore there is an unusually rich data set of published and annotated whole genomes, with 10 newly published bee genomes in the last three years (Kapheim et al., 2015; Kocher et al., 2013; Park et al., 2015; Rehan et al., 2016; Sadd et al., 2015). These genomes can be effectively used for extracting in silico UCE data using the phyloinformatic pipeline Phyluce (Faircloth, 2016). Second, the main phylogenetic controversy in the phylogeny of the corbiculate bees has been the subject of considerable study and establishing the relationships among the corbiculate tribes has been highly controversial. Previous morphological approaches supported a tree in which the advanced eusocial Meliponini and Apini formed a monophyletic group sister to Bombini, with Euglossini at the base of the tree (Fig. 1C; Cardinal and Packer, 2007; Plant and Paulus, 2016), suggesting a single origin of eusociality in corbiculate bees. However, numerous molecular studies (e.g., Cardinal and Danforth, 2011; Kawakita et al., 2008; Martins et al., 2014) have supported a very different view: that Bombini and Meliponini form a well-supported monophyletic group which is sister to Apini + Euglossini (Fig. 1A), thereby implying two separate origins of advanced eusociality. A recent phylogenomic nivation accounted for GC bias in coding sequences obtained yet another topology. Romiguier et al. (2015) obtained a tree in which Apini is sister to the clade of Bombini + Meliponini (Fig. 1B), with strong support when utilizing the non-homogeneous GC98 substitution model (Galtier and Gouy, 1998).

Besides the high degree of topological conflict, two aspects render the corbiculate bees an ideal group for investigating the impact of GC-bias on incongruence among ultraconserved elements. First, this group has been the focus of research on the genetic basis of sociality in bees and therefore there is an unusually rich data set of published and annotated whole genomes, with 10 newly published bee genomes in the last three years (Kapheim et al., 2015; Kocher et al., 2013; Park et al., 2015; Rehan et al., 2016; Sadd et al., 2015). These genomes can be effectively used for extracting in silico UCE data using the phyloinformatic pipeline Phyluce (Faircloth, 2016). Second, the main phylogenetic controversy in the phylogeny of the corbiculate bees has been the subject of considerable study and establishing the relationships among the corbiculate tribes has been highly controversial. Previous morphological approaches supported a tree in which the advanced eusocial Meliponini and Apini formed a monophyletic group sister to Bombini, with Euglossini at the base of the tree (Fig. 1C; Cardinal and Packer, 2007; Plant and Paulus, 2016), suggesting a single origin of eusociality in corbiculate bees. However, numerous molecular studies (e.g., Cardinal and Danforth, 2011; Kawakita et al., 2008; Martins et al., 2014) have supported a very different view: that Bombini and Meliponini form a well-supported monophyletic group which is sister to Apini + Euglossini (Fig. 1A), thereby implying two separate origins of advanced eusociality. A recent phylogenomic nivation accounted for GC bias in coding sequences obtained yet another topology. Romiguier et al. (2015) obtained a tree in which Apini is sister to the clade of Bombini + Meliponini (Fig. 1B), with strong support when utilizing the non-homogeneous GC98 substitution model (Galtier and Gouy, 1998).

Besides the high degree of topological conflict, two aspects render the corbiculate bees an ideal group for investigating the impact of GC-bias on incongruence among ultraconserved elements. First, this group has been the focus of research on the genetic basis of sociality in bees and therefore there is an unusually rich data set of published and annotated whole genomes, with 10 newly published bee genomes in the last three years (Kapheim et al., 2015; Kocher et al., 2013; Park et al., 2015; Rehan et al., 2016; Sadd et al., 2015). These genomes can be effectively used for extracting in silico UCE data using the phyloinformatic pipeline Phyluce (Faircloth, 2016). Second, the main phylogenetic controversy in
this group revolves primarily around a three-taxon statement involving Apini, Euglossini, and Bombini + Meliponini. Hence, we can apply quartet tree based evaluation methods, as the conflict is restricted to a single rooted triplet with three possible topologies (Fig. 1A, B, D). This study uses this publicly available data to gain first insights into base compositional effects on the performance of UCE data for phylogenomic inferences. We further utilize this new class of phylogenomic markers to explore the evolutionary relationships among the corbiculate bees.

2. Material and methods

2.1. In-silico sequence capture and processing

We used the Hymenoptera probe set designed by Faircloth et al. (2015) to extract UCEs from 13 publicly available bee genomes and UCE contigs of three additional bee taxa that were captured using this probe set (Table 1). Captured loci were extracted as fasta files with total up- and downstream flanking regions of 850 bp length. We called Mafft (Katoh et al., 2009) as implemented in Phyluce (Faircloth, 2016) to align the sequence data on the individual loci separately. We applied a 75% cutoff for data set completeness, meaning that each locus is represented by at least 75% of all taxa (i.e., a minimum of 12 taxa in total per locus). Capture statistics were calculated using Phyluce and Amas (Borowiec, 2016). Nucleotide variables of captured sequences in respect to their position on the UCE core were inferred with the Phyluce smilogram script and plotted with R 3.2.5 (R Core Team, 2016), utilizing the ggplot2 package (Wickham, 2009). An extensive description of the in silico capture methods and the subsequent processing are provided in the supplementary online information, including a step by step listing of the commands used (Supplementary data 1).

2.2. Locus selection and alignment filtering

Two alignments for concatenation-based phylogenetic methods were created. First, the loci of the 75% completeness matrix were concatenated without further processing (‘Raw75p’). For the second dataset, we filtered the Raw75p dataset for the best performing loci based on the average bootstrap support of their respective gene trees. This filtering approach is primarily a workaround to bypass the extraordinary computational demands of analyzing extremely large, phylogenetic datasets. The same approach was used in recent UCE based studies (Blaimer et al., 2015; Branstetter et al., 2016). Here we extract the 100 highest scoring loci and compare their performance and base composition against the unprocessed dataset. Therefore we called RAxML 8.2.8 to calculate the best-scoring ML gene trees on each unpartitioned individual locus of the Raw75p dataset. A rapid bootstrapping analyses with 200 bootstrap replicates (-f a) was conducted, while specifying GTR+F as the substitution model for the bootstrap search. Utilizing the gene.stats R script used in Borowiec et al. (2015), we then inferred the individual average bootstrap support values for each gene tree. We subsequently screened for loci that yielded the 100 highest scoring ML gene trees based on their average bootstrap support and discarded all other loci (‘Best100’).

To account for base compositional effects among genes with different GC composition, two separate subsets of loci with different GC characteristics were created. We utilized the 810 loci of the Raw75p dataset and excluded the 10 worst performing loci based on their average bootstrap support to reduce the data set to a manageable 800 loci. First, we calculated the GC content for each locus individually. All loci were ranked by their overall GC content and divided into 20 groups of 40 loci, thereby letting each group represent 5% of all loci (‘subset GC content’). This means that locus group 1 consists of loci with very low GC content, whereas locus group 20 is rich in GC. For the second subset, we calculated the variance of the GC content among taxa for each locus, as a measure for heterogeneity using a custom python script. The loci were ranked by increasing variance and again divided into 20 groups of 40 loci (‘subset GC heterogeneity’).

2.3. Phylogenetic inferences

To identify conflicting topologies among the individually processed datasets, species trees were inferred using both (1) concatenation and (2) gene tree methods.

2.3.1. Concatenation methods

For the concatenated multi-locus alignments, each locus was designated as a unique partition. We first executed best-scoring maximum likelihood (ML) tree searches and rapid bootstrap analyses with 1000 replicates for both concatenated datasets (Raw75p,
Best 100) using RAxML. We applied a GTR+\Gamma model for both the tree searches and the bootstrapping phases.

Bayesian tree inferences (BI) for each dataset were carried out by running four independent replicates of each four Metropolis-coupled MCMC using ExaBayes 1.4.2 (Aberer et al., 2014). The chains ran for 1,000,000 generations and were sampled every 25 generations, yielding 160,000 trees per alignment. Branch lengths were linked among partitions and a relative burn-in of 0.2 was used. Convergence was evaluated by ensuring effective sample size values (ESS) of >200 for each parameter, potential scale reduction factors (PSRF) ranging close to one and average standard deviations of split frequencies (ASDSF) of <0.5%. Posterior distributions of tree topologies of the individual runs were condensed using the implemented consense command and parameter files were saved with postProcParam.

2.3.2. Gene tree methods

Species trees using the multi-species coalescent model were inferred with Astral-2 4.10.6 (Mirarab et al., 2014; Mirarab and Warnow, 2015), as the software can properly handle unrooted gene trees with missing taxa and polytomies. Since summary methods such as Astral estimate species trees based on the topologies of the underlying input gene trees, the species trees are prone to be biased by poorly resolved gene trees of uninformative loci. We therefore in turn revised all individual gene trees and collapsed nodes with bootstrap support values of ≤50 using TreeCollapserCL 4 (Hodcroft, 2016). We subsequently summarized the 810 trees of the Raw75p dataset using the heuristic algorithm and inferred local posterior probabilities (PP; Sayyari and Mirarab, 2016) to measure topological support, as this measure was shown to be of superior accuracy than multi-locus bootstrapping sensu Seo (2008). Additionally, the quartet support for the individual quartet trees was calculated using the score option.

2.4. Tracing base compositional bias

We adopted and modified the topological incongruence measure of Romiguier et al. (2015) to explore trends of gene tree incongruences based on base composition. We therefore summarized the ML trees of the individual loci of both subsets ('GC content', 'GC heterogeneity') for each group separately using Astral-2, and scored the resulting species trees against each individual gene tree of their respective group. Instead of the quartet distance measure (cf., Romiguier et al., 2015), we extracted the quartet support (as defined in Sayyari and Mirarab, 2016) for the tested topologies (q1 in Astral-2) of every scored quartet and plotted their distribution as a function of the GC content and heterogeneity groups, respectively. In contrast to normalized quartet support of entire species trees, this measure is more sensitive to incongruence among single quartet trees, as it reflects the relative support of all branches and not the total of all satisfied quartets. It should therefore be superior in detecting effects of gene tree heterogeneity in actual biological data sets, as incongruence is unlikely to be equally distributed among entire phylogenies (cf., Wang et al., 2015).

As the sister group relationship of bumblebees (Bombini) and stingless bees (Meliponini) is well established, the investigated topological conflict is restricted to a single quartet tree with three possible hypotheses (Fig. 1A, B, D). We therefore scored the collapsed 40 ML trees of each locus group of both subsets ('GC content', 'GC heterogeneity') against the inferred topology of the concatenation based methods (Fig. 2; corresponds to topology B in Fig. 1). Subsequently, the local PP of the corresponding quartet tree for each of the three possible hypotheses was extracted for every locus group individually. We plotted the support for each hypothesis as a function of increasing GC content or GC heterogeneity, respectively, and measured the association with simple linear regressions. We tested for discordance of H₀ = 0 with simple linear regression models.

3. Results

3.1. Sequence capture and locus filtering

Matching the probes to the extracted genome slices and contig assemblies, and subsequent trimming yielded in a total of 1132 shared UCE loci with a mean alignment length of 518 bp. The level

![Fig. 2. Bayesian phylogram of the ‘Best100’ dataset. The scale bar indicates nucleotide substitutions per site. All nodes revealed Bayesian posterior probabilities and ML bootstrap support of 1 and 100, respectively. The branch values indicate the local posterior probability inferred by Astral-2 and the actual quartet score. Asterisks correspond to 1.](image)
of sequence variation of the UCE nucleotide positions in respect to their distance to the core revealed the expected ‘smilogram’-pattern with highly conserved anchor regions and increasingly variable flanking regions (Fig. 3). An average of 894 loci was captured per taxon, with all taxa represented by at least 840 loci except for *Apis florea* with 613. Filtering the captured loci for a 75% completeness matrix reduced the number of loci for the subsequent alignment to 810, with 12.02% representing gaps or missing information. The amount of missing data of the Best100 alignment is comparable, whereas the average Best100 matrix is relatively longer, with a higher percentage of parsimony informative sites and is 4% lower in the overall GC content than the average Raw75p locus. The higher variation of the average GC content among taxa further reveals that the loci of the Best100 dataset are more heterogeneous in their GC composition (Table 2).

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Raw75p alignment</th>
<th>Best 100 alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of loci</td>
<td>810</td>
<td>100</td>
</tr>
<tr>
<td>Alignment length (bp)</td>
<td>419,260</td>
<td>73,434</td>
</tr>
<tr>
<td>Average locus length (bp)</td>
<td>518</td>
<td>734.34</td>
</tr>
<tr>
<td>Undetermined characters (%)</td>
<td>12.02</td>
<td>12.68</td>
</tr>
<tr>
<td>Parsimony informative sites</td>
<td>110,513</td>
<td>20,198</td>
</tr>
<tr>
<td>Overall GC content (%)</td>
<td>46.2</td>
<td>42.24</td>
</tr>
<tr>
<td>Mean GC variance among taxa and loci</td>
<td>5.48</td>
<td>9.04</td>
</tr>
</tbody>
</table>

Overall, the GC content is roughly normally distributed among the captured loci, with the majority showing 45–55% GC after trimming. The homogeneity among taxa is less equally distributed with a large positive skew (Supplementary data 2 and 3). GC content and GC heterogeneity are strongly correlated ($R^2 = 0.167$, $p < 0.0001$), however, the few very heterogeneous loci are not necessarily rich in GC and vice versa (Supplementary data 4). Interestingly, the GC content of the most heterogeneous loci is unequally distributed among the included taxa. There is a trend that the corbiculate lineages are generally lower in GC content than other taxa. This is particularly apparent among the most heterogeneous loci (Fig. 4). Even if this trend is not shared among all loci, several of the most heterogeneous loci showed a clustering of the corbiculate lineages with lower GC content than all other taxa. Furthermore, several of these most heterogeneous loci produced well-supported species trees: The Best100 dataset includes 20 of the 10% most heterogeneous loci of the entire data set, whereas only 1 of the 10% GC richest loci falls into this category.

### 3.2. Coalescent vs. concatenated species trees and the inferred phylogeny

The concatenated species tree reconstructions of both the ML and BI approaches produced congruent topologies and maximum supported trees for both the 810 loci alignment ('Raw75p') and the Best100 locus set (Fig. 2). In contrast, the quartet supported gene tree recovered the same topology but reveals considerably lower confidence at the base of the corbiculate phylogeny, indicated by the fairly low local PP of 64. Only 14.4% of all induced quartet trees at this branch are actually congruent with the depicted topology, followed by 13.8% favoring the best alternative, which corresponds to topology D in Fig. 1. Notably, these values dropped from 36.3% and 36.2% (!), respectively, after collapsing nodes with bootstrap support values of ≤50. In contrast, all other quartet trees strongly support the concatenation-based trees. Besides the corbiculate taxa, the inferred phylogeny is in line with our current understanding of the higher-level relationships among the bees (Danforth et al., 2012). Moreover, it strongly supports the generic consensus tree of *Apis* sensu Koeniger et al. (2011), which was previously inferred using morphology and smaller scale molecular data analyses. It is further noteworthy that the phylogenetic position of the Xylocopinae (*Ceratina calcarata*) renders the remaining Apinae as paraphyletic (as expected; Danforth et al., 2012).

### 3.3. Effects of different locus selection

The GC sorted subsets of locus groups ('GC content', ‘GC heterogeneity') had opposing effects on the general congruence among
the phylogenetic inferences. Whereas the general congruence among the gene trees decreases with increasing GC content, the summary trees of the heterogeneous loci become slightly less conflicting with higher variance of GC among taxa. In other words, increasing GC content had a negative impact on the general phylogenetic accuracy (Fig. 5). Increasing GC heterogeneity had a slightly positive but insignificant impact on phylogenetic accuracy (Supplementary data 5). The trend caused by the varying GC contents on the general congruence is stronger ($R^2 = 0.025$, $p = 0.011$) and less likely due to randomness than the grouping by heterogeneous loci ($R^2 = 0.012$, $p = 0.077$). Nonetheless, both trends are fairly weak.

Most strikingly, the GC content had strong opposing effects on the support of the three tested corbiculate topologies (Fig. 6). Loci with low GC content favor the placement of Euglossini as sister group to the remaining corbiculates ($R^2 = 0.3181$, $p = 0.009$), whereas the GC rich groups strongly support the honeybees (Apini) at this position ($R^2 = 0.4457$, $p = 0.001$). The sister-group relationship between Euglossini and Apini, which has been obtained in previous studies of corbiculate relationships based on multilocus data sets (e.g., Cardinal and Danforth, 2011; Kawakita et al., 2008; Martins et al., 2014), is overall poorly supported across a wide range of GC content with no detectable correlation between quartet support and GC content ($R^2 = 0.0261$, $p = 0.497$). Scoring the three hypotheses against the loci groups of the GC heterogeneity subset did not reveal comparable trends (Supplementary data 6). Overall, our results demonstrate that increasing GC richness has a significant negative impact on the phylogenetic accuracy obtained by using UCE data in groups such as corbiculate bees, where significant incongruence can be expected.

4. Discussion

4.1. Incongruence and GC bias as a source of error in phylogenetic analysis

Our results indicate clearly that incongruence among gene trees has played a significant role in rendering the phylogeny of corbiculate bees so challenging. Whereas the concatenation-based methods provide very strong support for our “preferred” topology (Fig. 1B), examination of the underlying gene-trees, especially along the branches involving the Apini and Euglossini, reveals significant incongruence and limited support for this topology (Fig. 2). Partitioning of loci by GC content and GC heterogeneity reveals that both GC content and (to a lesser extent) GC heterogeneity are impacting phylogenetic accuracy and incongruence among loci: GC rich loci show increasing support for the “wrong” tree topology and increased topological incongruence among loci (Figs. 5 and 6); In turn, GC heterogeneous loci provide slightly less topological incongruence but no clear trend towards recovering the “right” tree (Supplementary data 5 and 6). These results support the link between GC bias and incongruence caused by GC biased gene conversion and provide indirect evidence that gBGC is present in UCEs.

We suspect that the topological incongruence that we have documented in corbiculate bees is due largely to the unusual feature of extraordinary recombination rates in honeybees. High recombination rates are known from several eusocial insect lineages (Sirvio et al., 2011), however, the recombination rates of A. mellifera are significantly higher than in any other insect species for which recombination rate data are available and are considered among the highest of all higher eukaryotes (Beye et al., 2006; Wilfert et al., 2007). The variation of recombination rates along the intra-genomic landscape of the honeybee genome is further
strongly correlated to GC content (Ross et al., 2015), indicating that gBGc is an important driver of the base composition (Wallberg et al., 2015). Previous studies showed that recombination rate is positively correlated with the effective population size Ne (Hellmann et al., 2005). Together with speciation time (i.e., internode branch length), Ne in turn modulates the rate of incomplete lineage sorting, which is a major driver of gene-tree/species-tree incongruences (e.g., Degnan and Rosenberg, 2009; Maddison, 1997). This means that accumulated GC in honeybee UCEs indicates elevated recombination rates, which in turn increases the probability for ILS and subsequently higher incongruence among gene trees of the corbiculates.

This high level of incongruence presents a major challenge to determining if Euglossini or Bombini + Meliponini, is the actual sister lineage to Apis. Honeybees are unusual bees in other ways as well. They represent the only bee group that uses symbolic language (the dance language) to communicate the location of food resources, and they differ from other corbiculate bees in the mode of new nest founding, mating, and many other behavioral traits (Hepburn and Radloff, 2011; Michener, 1974). Studies based on a small number of genes generated with PCR and Sanger sequencing may have been particularly vulnerable to the impact of gene-gene incongruence. Even with nearly 1000 genes we find that only a small majority of genes actually supports the “correct” tree (Fig. 2). Our results provide a cautionary tale for studies of taxa in which extremely rapid diversification (short internodes) and high recombination rates can lead to substantial gene-gene incongruence. Even with massive data sets with several thousands of genes, the underlying gene tree incongruence can be obscured when only concatenation methods are used. We recommend that future studies using phylogenomic data assess incongruence on a node-by-node basis in order to identify nodes that are problematic as we have done here.

4.2. The versatility of UCEs to detect topological conflict

Because phylogenomic data sets, including those developed using targeted enrichment methods, typically include information from thousands of independent loci scattered across the genome, these methods provide an extraordinarily powerful tool for examining the impact of incongruence in phylogenetic analysis. Our case study demonstrates the importance of examining both concatenation and gene-tree methods in order to assess the level of underlying incongruence in phylogenomic data sets. Conventional concatenation based methods are more prone to yielding “incorrect” phylogenetic trees if considerable discordance is present in the examined lineages (Mirarab et al., 2014), and the respective support measures can give misleading levels of confidence (Kubatko and Degnan, 2007). Species tree methods, such as the Astral-2 algorithm, are more robust against topological conflict due to ILS than concatenation, as they account for gene tree incongruences by implementing the multi-species coalescent model (MSC; Rannala and Yang, 2003). In contrast, concatenation approaches can provide greater accuracy when data sets have low levels of ILS and horizontal gene transfer (Chou et al., 2015; Mirarab and Warnow, 2015). As the persistently inconclusive relationship of the Apini and Euglossini shows strong phylogenetic conflict, it is apparent that the coalescent-related confidence measures reflect the fragility of the inferred topology among the corbiculates more accurately than the seemingly well-supported BI and ML trees. Moreover, the exaggerated robustness of the concatenation-based confidence measures depicts their stochastic limitations, as they basically show no uncertainty despite significant underlying incongruence. This does not at all imply that they are statistically inaccurate. It does, however, reveal the vulnerability of concatenation methods to methodological shortfalls such as model misspecification when sampling genome-scale data (cf., Kumar et al., 2012). In contrast, UCEs provide superior potential to address difficult and incongruent nodes using coalescent-based gene tree methods, since the large number of individual loci provides a sufficient sample size to overcome coalescent stochasticity.

4.3. Phylogeny of the corbiculate bees – have we finally gotten this right?

Resolving the phylogeny of corbiculate bees based on molecular data has been a challenge since Sydney Cameron published the first molecular study of corbiculate relationships (Cameron, 1993; reviewed in Cardinal and Packer 2007). Previous molecular studies consistently grouped Meliponini + Bombini, but the relationships between Apini and Euglossini were less consistent with alternating topologies (cf., Cardinal and Danforth, 2011; Hedtke et al., 2013). Many studies grouped Apini + Euglossini, some with high bootstrap support and high posterior probabilities (Cardinal and Danforth, 2011, 2013; Kawakita et al., 2008). However, from today’s perspective these studies included an insufficient number of independent genes or loci to overcome the challenge of reconstructing a difficult node in the presence of significant GC bias and incongruent gene tree topologies. This topology was subsequently challenged by a 20 gene study that depicted the Euglossini on a basal branch within the corbiculates (Hedtke et al., 2013). Furthermore, the grouping of Apini + Euglossini was viewed by most morphologists as highly unlikely (Cardinal and Packer, 2007; Noll, 2002; Plant and Pauus, 2016) and also difficult to reconcile with the distribution of sociality in these four tribes. Two phylogenomic studies (Romiquier et al., 2015 and the present study) have now supported a phylogeny that places the weakly social and solitary Euglossini as sister to the eusocial tribes (Apini, Bombini, Meliponini). This basal position is also strongly supported by behavioral characters, especially in respect to the distribution of sociality (Noll, 2002), and by the phylogenetic studies based on morphology (Cardinal and Packer, 2007; Plant and Pauus, 2016). Considering the congruent evidence of morphological, behavioral, and now phylogenomic data on the placement of the Euglossini, we are confident that the remaining eusocial corbiculates form a monophyletic group. However, the sister-group relationship of Bombini and Meliponini, which is essentially inferred by every nucleotide-based phylogeny (e.g., Cameron 1993; Cardinal and Danforth, 2013; Kawakita et al., 2008; Hedtke et al., 2013) and the present study, contrasts with the topology based on morphological data (Fig. 1C).

Both phylogenomic studies detected challenges reconstructing the placement of Apini, particularly because of gene-gene incongruence. However, both studies do converge on the same topology (Fig. 2), making this our current best estimate of corbiculate relations. Nonetheless, the gene tree analyses using the multi-species coalescent model revealed two very closely competing hypothesis, which indicates the fragility of our results. Coalescent theory (see Kingman, 1982) predicts stochastically sorted genes within descendant lineages when genes fail to coalesce in ancestral populations (ILS), and hence sufficient sample sizes of independently evolving loci directly lowers the probability of false results due to randomness. Maximizing the amount of orthologous loci for gene tree based species tree estimations will therefore not solve the nature of incongruences itself but increase the confidence that the pendulum swings to the side of the true species tree.

An evident weakness of both our and Romiguier et al.’s (2015) study concerns the sparse taxon sampling, as both studies include a fairly low number of corbiculate taxa and no closely related outgroup taxon, such as Centris (Hedtke et al., 2013; Martins et al., 2014). Outgroup choice can impact ingroup topology (e.g., Cameron et al., 2004; Philippe et al., 2005; Ware et al., 2008) and
can even alter levels of branch support (Kirchberger et al., 2014). Strikingly, phylogenetic distance from outgroup to ingroup taxon was shown to promote incongruence among loci (Rosenfeld et al., 2012). As the accuracy of gene tree methods directly depend on the quality of the underlying input gene trees, a primary objective of future studies should be to address the optimization of the individual tree inferences and should develop more thorough taxon sampling.

Acknowledgements

We are grateful to Jason Dombroskie from the Cornell University Insect Collection (http://cuic.entomology.cornell.edu/) for the access to the high quality imaging system used to prepare Fig. 2. We further thank two anonymous reviewers for their constructive comments on the manuscript. This study was funded by a U.S. National Science Foundation grant to B.N. Danforth, S.G. Brady, J.P. Pitts, and R. Ross [DEB-155905].

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2017.03.022.

References


Kirchberger, P.C., Seif, K.M., Sturmbauer, C., Kohmuller, S., 2014. Outgroup effects on root position and tree topology in the AFI phylogeny of a rapidly radiating