Phylogenomics Controlling for Base Compositional Bias Reveals a Single Origin of Eusociality in Corbiculate Bees

Jonathan Romiguier,*[1] Sydney A. Cameron,[2] S. Hollis Woodard,[3] Brielle J. Fischman,[4] Laurent Keller,*[1] and Christophe J. Praz*[5]

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[2]Department of Entomology, University of Illinois, Urbana

[3]Department of Entomology, University of California, Riverside

[4]Department of Biology, Hobart and William Smith Colleges, Geneva, NY

[5]Institute of Biology, University of Neuchatel, Neuchatel, Switzerland

***Corresponding author:** E-mail: christophe.praz@unine.ch, jonathan.romiguier@gmail.com, Laurent.Keller@unil.ch.

Associate editor: Andrew Roger

Abstract

As increasingly large molecular data sets are collected for phylogenomics, the conflicting phylogenetic signal among gene trees poses challenges to resolve some difficult nodes of the Tree of Life. Among these nodes, the phylogenetic position of the honey bees (Apini) within the corbiculate bee group remains controversial, despite its considerable importance for understanding the emergence and maintenance of eusociality. Here, we show that this controversy stems in part from pervasive phylogenetic conflicts among GC-rich gene trees. GC-rich genes typically have a high nucleotidic heterogeneity among species, which can induce topological conflicts among gene trees. When retaining only the most GC-homogeneous genes or using a nonhomogeneous model of sequence evolution, our analyses reveal a monophyletic group of the three lineages with a eusocial lifestyle (honey bees, bumble bees, and stingless bees). These phylogenetic relationships strongly suggest a single origin of eusociality in the corbiculate bees, with no reversal to solitary living in this group. To accurately reconstruct other important evolutionary steps across the Tree of Life, we suggest removing GC-rich and GC-heterogeneous genes from large phylogenomic data sets. Interpreted as a consequence of genome-wide variations in recombination rates, this GC effect can affect all taxa featuring GC-biased gene conversion, which is common in eukaryotes.

Key words: **phylogenomics, bees, eusociality, GC-biased gene conversion, base composition.**

Introduction

Reconstructing the evolutionary history of major adaptations across the Tree of Life is a central goal in evolutionary biology. Comparative methods used to examine the evolution of complex traits rely on a phylogenetic tree, either to pinpoint transitions in certain clades or to identify essential preadaptations. Nowadays, next-generation sequencing provides an enormous wealth of molecular markers for building well-resolved phylogenetic trees. However, despite this unprecedented amount of data, some nodes in the Tree of Life remain controversial, largely due to conflicting phylogenetic signals among loci (Gatesy and Springer 2014; Liu et al. 2015).

One particularly controversial phylogeny is that of the corbiculate bees, an ecologically and economically important group of bees that is also important for understanding social evolution. Corbiculate bees include the only bees exhibiting complex eusociality, which is characterized by large colonies, a perennial colony cycle, and morphologically and behaviorally specialized queen and worker castes. The corbiculate bees consist of four monophyletic tribes: the solitary orchid bees (Euglossini; hereafter E), the bumble bees (Bombini; hereafter B), which exhibit simple eusociality, and the complex eusocial honey bees (Apini; hereafter A), and stingless bees (Meliponini; hereafter M). Knowing the phylogenetic relationships among these four tribes is important for determining the number of origins of eusociality within this group and for evaluating the possibility of reversal from social to solitary life styles (Cardinal and Danforth 2011). Although the monophyly of each of the four tribes is well established (Cameron 1993; Koulianos et al. 1999; Mardulyn and Cameron 1999; Ascher et al. 2001; Cameron and Mardulyn 2001; Lockhart and Cameron 2001; Cameron and Mardulyn 2003; Michener 2007; Kawakita et al. 2008; Whitfield et al. 2008; Cardinal et al. 2010; Woodard et al. 2011; Hedtke et al. 2013), the phylogenetic relationships among tribes have remained controversial. For example, of the 15 possible unrooted phylogenies, nine have been supported by at least one study (reviewed in Cardinal and Packer 2007; Almeida and Porto 2014).

All molecular studies to date support a close phylogenetic relationship between Bombini and Meliponini but conflicts surround the position of Apini and Euglossini, reducing the conflicting topologies from nine to three (reviewed in Danforth et al. [2013]; table 1). One topology groups the three eusocial clades together (Apini, Bombini, and Meliponini), with the solitary Euglossini as the basal clade (hereafter the ABM topology, blue in fig. 1), whereas the alternative two topologies group the solitary clade (Euglossini)

Table 1. Summary of Previous Phylogenetic Studies of Corbiculate Relationships.

Study	Loci	Number of Aligned Sites (bp)	Most Supported Topology	Clade Support Values			
				BM	AE	ABM	EBM
Cameron (1993)	16S	536	EBM	98% (MP)	—	—	60% (MP)
Koulianos et al. (1999)	Cytb	520	AE[a]	91% (MP)	53% (MP)	—	—
Mardulyn and Cameron (1999)	LW-Rhodopsin	502	EBM[b]	71% (ML)	—	—	83% (ML)
Cameron and Mardulyn (2001)	16S, 28S, LW-Rhodopsin, Cytb	2,360	AE	99% (MP)	55% (MP)	—	—
Ascher et al. (2001)	LW-Rhodopsin	495	NA[b]	—	—	—	—
Kawakita et al. (2008)	12 nuclear genes	6,018	AE	100% (ML)	88% (ML)	—	—
Cardinal et al. (2010)	7 nuclear genes	5,844	AE	100% (ML)	77% (ML)	—	—
Hedtke et al. (2013)	up to 20 genes	up to 17,000	ABM	100% (ML)	—	74% (ML)	—
Woodard et al. (2011)	717 nuclear genes	69,461	ABM; AE[c]	1.0	1.0[c]	1.0	—

NOTE.—Support values indicate bootstrap support values in parsimony (MP) or maximum-likelihood (ML) analyses (when both types of analyses were conducted, values reported here refer to ML analyses) or posterior probability values in Bayesian analyses.

[a]Analyses of amino acid sequence recovered EBM topology.

[b]Corbiculate monophyly was not recovered.

[c]AE recovered in analyses of third codon positions.

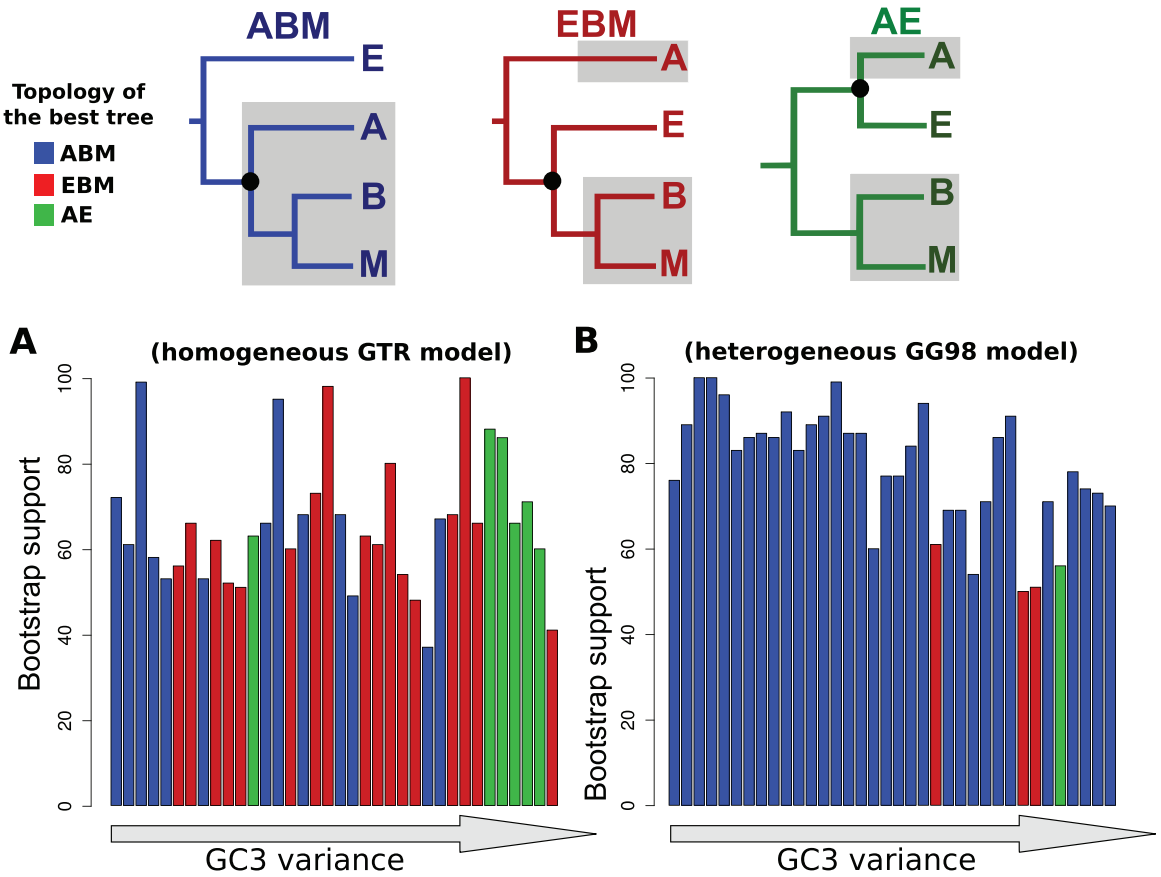

FIG. 1. Support values for the position of Apini according to GC3-heterogeneity derived from (*A*) homogeneous (GTR) and (*B*) nonhomogeneous (GG98) models of sequence evolution. Each bar represents a supermatrix of 100 genes grouped according to their GC3-heterogeneity. The color of the bars corresponds to the topology of the most supported tree (blue for ABM, red for EBM, and green for AE), the heights correspond to the node bootstrap value of the Apini position (node with a black dot on the corresponding topology). Clades with a eusocial lifestyle are surrounded by gray squares.

as the sister clade to either the eusocial Apini or the eusocial Bombini + Meliponini clade (respectively, the AE and the EBM topology; green and red in fig. 1). These conflicting topologies have different implications for eusocial evolution. The ABM topology suggests a single origin of eusociality from a solitary common ancestor, whereas the AE and EBM topologies suggest either two independent origins of eusociality from a solitary ancestor or a single origin of eusociality in the common ancestor, followed by a loss of eusociality in Euglossini. The AE topology has been favored in the literature recently and has been used as the true species tree in several comparative studies (Cardinal and Danforth 2011; Woodard et al. 2011; Kocher and Paxton 2014; Kapheim et al. 2015).

Discrepancies among species trees (table 1) stem from conflicting gene trees, with each gene representing its own evolutionary history, which may or may not correspond to the history of species (e.g., through incomplete lineage sorting, Pamilo and Nei 1988; Degnan and Rosenberg 2009). Given these discrepancies, it is important to identify which genes are reliable phylogenetic markers and which genes should be avoided for phylogenetic inference. Recently, it has been shown that GC-rich genes may produce false and inconsistent topologies in mammals (Romiguier et al. 2013). This so-called "GC effect" is thought to be due to genome-wide variations in recombination rate. Recombination is known to drive mammalian nucleotidic composition through GC-biased gene conversion, a DNA repair bias favoring GC alleles during meiotic recombination (Eyre-Walker 1993; Galtier et al. 2001; Romiguier et al. 2010). Consequently, recombination hotspots lead to increased GC content, elevated polymorphism, fast evolutionary rates, and, importantly, heterogeneity of base composition among taxa, which can bias phylogenetic reconstructions through incomplete lineage sorting (Hobolth et al. 2011), long-branch attraction (Bergsten 2005), or substitution model misspecifications (Boussau and Gouy 2006; Nabholz et al. 2011; Betancur et al. 2013). For all of these reasons, GC-rich genes with high GC heterogeneity among taxa have been avoided to resolve the trickiest nodes of the phylogeny of mammals (Romiguier et al. 2013) and birds (Jarvis et al. 2014), two taxa known to feature GC-biased gene conversion and heterogeneous GC content (Figuet et al. 2014).

GC biases may similarly affect phylogenetic reconstructions in insects, including the corbiculate bees. In particular, genomes of eusocial Hymenoptera feature both the highest recombination rates recorded in multicellular eukaryotes (Beye et al. 2006; Wilfert et al. 2007; Ross et al. 2015) and extremely heterogeneous GC contents (Jørgensen et al. 2007; Suen et al. 2011). Furthermore, the western honeybee (*Apis mellifera*) is the only known invertebrate where GC-biased gene conversion has been demonstrated (Kent et al. 2012), with fixation biases toward GC estimated to be up to 50 times stronger than in mammals (Wallberg et al. 2015).

Genome-wide gene sequence alignments of nine bee species (including two species per each corbiculate bee tribe, Woodard et al. 2011) provide an opportunity to test the influence of GC content on the honeybee phylogenetic position. Originally used to identify genes involved in convergent evolution of simple and complex eusocial bee lineages, these sequence alignments retrieved two conflicting topologies (ABM and AE; supplementary fig. S1 in Woodard et al. 2011). Based on phylogenetic studies including a high number of species but few genes (Kawakita et al. 2008; Cardinal et al. 2010), Woodard et al. (2011) favored the AE topology for their analyses, as has been the case in other recent comparative studies (Cardinal and Danforth 2011; Kocher and Paxton 2014; Kapheim et al. 2015). In this phylogenomic study, we examined how GC content impacts phylogeny reconstruction to specifically resolve the controversial relationships among honey bees, bumble bees, stingless bees, and orchid bees. Contrary to the prevalent hypothesis (AE topology), we found that the topology supporting a single origin of eusocial behavior with no reversal to solitary lifestyle (ABM topology) is the most likely.

Results

Higher Probability of Conflicting Gene Trees among GC-Rich and GC-Heterogeneous Genes

To test the impact of base composition on tree inference, we ranked 3,600 gene fragments examined in Woodard et al. (2011) according to their average GC content at the third codon position (hereafter, "average GC3") as well as their GC3% variance across species (hereafter, "GC3-heterogeneity"), then divided this ranked gene list into 36 groups of 100 genes. There was a strong correlation ($R = 0.85$, P value < 0.0001; fig. 2*A*) between the average GC3 of a gene group and the within-group topological conflict among gene trees (i.e., average quartet distance of gene trees compared to their group consensus tree, a supertree summarizing the mutual agreement of the 100 gene trees; see Material and Methods). Similarly, there was a strong positive correlation between interspecific GC3 heterogeneity and within-group topological conflict ($R = 0.92$, P value < 0.0001; fig. 2*B*).

GC-rich and GC-heterogeneous genes were also less likely than GC-poor and GC-homogeneous genes to group pairs of species of the same tribe as sister species (i.e., Apini: *A. mellifera* + *Apis florea*; Bombini: *Bombus terrestris* + *Bombus impatiens*; Meliponini: *Melipona quadrifasciata* + *Frieseomelitta varia*; and Euglossini: *Eulaema nigrita* + *Euglossa cordata*; hereafter, the intratribe nodes). Across the 36 gene groups and their corresponding supertrees, the average support for intratribe nodes was negatively correlated with both average GC3 ($R = -0.70$, P value < 0.0001; fig. 2*C*) and GC3 heterogeneity ($R = -0.93$, P value < 0.0001; fig. 2*D*). The most striking example was for the monophyly of Meliponini, which was supported by 94 of the 100 most GC-homogeneous genes but only 60 of the 100 most GC-heterogeneous genes (data not shown). In summary, gene groups with low average GC3 and low GC3-heterogeneity resulted in lower topological incongruence and higher support for tribal monophyly, suggesting that these genes are more reliable for phylogenetic inference than genes with high average GC3 and high GC3-heterogeneity. Interestingly, the same correlations were observed when gene trees were inferred from amino

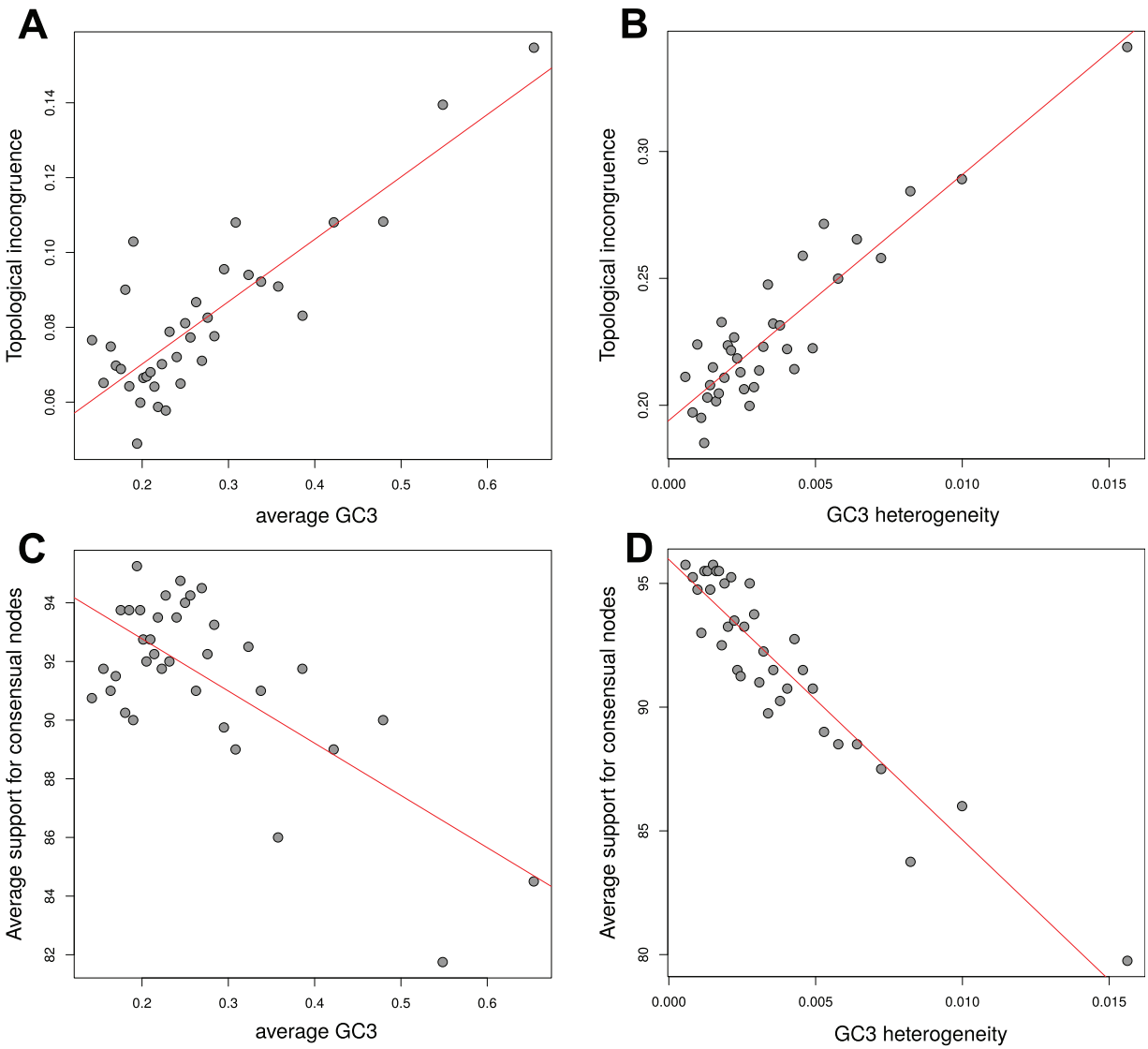

Fig. 2. GC effect on topological incongruence. Each dot represents one of the 36 groups of 100 genes (in total 3,600 alignments), ranked according to their average GC3 (*A* and *C*) or GC3-heterogeneity among species (*B* and *D*). Within-group topological conflict (used in *A* and *B*) measures gene tree incongruence in each gene group and is the average proportion of false quartets (quartet distance) of the 100 gene trees versus their own consensus tree (a supertree summarizing the 100 topologies). Support for consensual nodes (used in *C* and *D*) is the average support of a supertree for the uncontroversial nodes defining the monophyly of the four tribes of corbiculates bees (Apini, Bombini, Meliponini, and Euglossini).

acid sequences (supplementary fig. S1, Supplementary Material online), indicating that the GC effect could not be removed by translating the nucleotide data set into amino acids.

There was a high correlation between average GC3 and GC3-heterogeneity across the 36 gene groups ($R = 0.85$, P value < 0.0001, supplementary fig. S2, Supplementary Material online), as typically is the case in species exhibiting biased gene conversion (mammals and birds, Romiguier et al. 2010; Romiguier et al. 2013; Weber et al. 2014). Because GC3-heterogeneity was more highly correlated than average GC3 with both topological incongruence and support for tribe monophylies (fig. 2*B* and *D* vs. fig. 2*A* and *C*), we used GC3-heterogeneity (average for each group of genes) in further analyses investigating the role of GC content on phylogenetic reconstructions.

GC3-Homogenous Genes Support a Single Origin of Eusociality

Because GC3-heterogeneous genes were less reliable than GC3-homogeneous genes in retrieving consistent phylogenies, we investigated whether differences among GC-heterogeneity may be responsible for the conflicting topologies of corbiculate bees (table 1). We used the 36 groups of 100 genes ranked by GC3-heterogeneity (fig. 2*B* and *D*) to produce 36 supermatrices and performed maximum-likelihood analyses on each (RAxML, GTR + GAMMA model). In line with recent molecular studies (table 1), only three different topologies were retrieved from the resulting 36 trees: ABM, EBM, and AE (fig. 1). Overall, the EBM and ABM topologies (red and blue, fig. 1*A*) were the most supported across trees (respectively, 17/36 and 13/36 trees), whereas the AE topology (green, fig. 1*A*) was the least supported (6/36 trees).

Support for AE essentially came from the least reliable genes (or, most GC3-heterogeneous). In contrast, the five supermatrices containing the most reliable genes (most GC3-homogeneous) all supported the ABM topology (fig. 1*A*). Overall, the bootstrap support for ABM was negatively correlated with supermatrix GC3-heterogeneity ($R = -0.34$, P value = 0.04). GC3-heterogeneous supermatrices also produced the longest tree branches (positive correlation between log-transformed GC3-variance and branch length, $R = 0.95$, P value < 0.0001, supplementary fig. S3, Supplementary Material online), which can favor long-branch attraction artifacts.

To confirm the effect of GC3-heterogeneity on topology, we conducted another analysis where we concatenated the 3,600 genes into two large supermatrices (1,800 genes each) with low and high GC3-heterogeneity and repeated the maximum-likelihood analyses (RAxML, GTR + GAMMA model). Results of this analysis were consistent with our previous results, as the supermatrix comprising the more reliable genes (low GC3-heterogeneity) retrieved high bootstrap support for ABM (99%, fig. 3), whereas the supermatrix encompassing the less reliable genes (high GC3-heterogeneity) retrieved the AE topology with a middling bootstrap support (75%, fig. 3).

Nonhomogeneous Models of GC Evolution Strongly Support a Single Origin of Eusociality

Because our analyses showed a strong effect of GC3-heterogeneity on the inferred corbiculate bee phylogeny, we used a nonhomogeneous, nonstationary model of sequence evolution to reanalyze the data (GG98, Galtier and Gouy 1998; implemented in nhPhyml, Boussau and Gouy 2006). This nonhomogeneous model allows equilibrium base content to vary among lineages and should consequently provide more consistent topologies than the standard homogeneous model (GTR) used in the previous section.

The use of a nonhomogeneous model substantially increased the average bootstrap values of the inferred trees (79.1% vs. 65.25% in the previous analysis). Most supermatrices of 100 genes (32/36) supported the ABM topology (blue, fig. 1*B*), the topology also supported by the most GC3-homogeneous supermatrices in the previous analysis (fig. 1*A*). Three of the four remaining supermatrices recovered the EBM topology, whereas only one retrieved the AE topology (red and green, fig. 1*B*). Overall, the bootstrap support for ABM was negatively correlated with the GC3-heterogeneity of the supermatrices ($R = -0.39$, P value = 0.018).

To confirm these results, we reanalyzed the two large supermatrices previously created (each containing 1,800 genes with high or low GC3-heterogeneity) using two different nonhomogeneous models: the GG98 model in nhPhyml and the NCDH model implemented in P4 (Foster 2004). With the GC98 model, the supermatrix grouping the most GC3-homogeneous genes again retrieved the ABM topology with maximal bootstrap support (100%, fig. 3); the supermatrix containing the most GC3-heterogeneous genes also retrieved the ABM topology (bootstrap support of 59%, fig. 3), in contrast to the analyses done with a standard GTR homogeneous model (see previous section and fig. 3). With the NCDH model in P4, we used Bayes factors to compare the three alternate topologies (ABM, EBM, and AE) for these two supermatrices of 1,800 genes (see supplementary materials, Supplementary Material online, for details). In both cases, the ABM topology fit the data significantly better than either alternate topology (Bayes factors 164.39 and 102.36 for ABM over EBM and AE, respectively, for the 1,800 most GC3 homogenous genes; and 601.40 and 9.08 for the 1,800 most GC3 heterogeneous genes).

Discussion

Comparison with Previous Studies

In line with all previously published phylogenetic studies based on molecular data (table 1, see also Lockhart and Cameron [2001]), our analyses grouped bumble bees (Bombini) and stingless bees (Meliponini) as sister clades. This node was recovered in all supermatrix analyses with maximal bootstrap support for all models. Therefore, the clade grouping Bombini and Meliponini (hereafter; BM) appears to be unambiguous.

The relationship between the Bombini + Meliponini clade (BM) and the two other tribes (Apini and Euglossini) was less stable, with three different topologies: Apini as closest relative to BM and Euglossini in basal position (ABM topology), Euglossini as closest relative of BM and Apini in basal position (EBM topology), and Apini + Euglossini as sister to BM (AE topology) (fig. 1). Each of these topologies had been previously suggested in phylogenetic studies based on molecular data (table 1), although the AE topology is currently favored in the literature (Cardinal and Danforth 2011; Woodard et al. 2011; Almeida and Porto 2014; Kocher and Paxton 2014; Kapheim et al. 2015) mainly based on recent analyses (Kawakita et al. 2008; Cardinal et al. 2010). However, our study shows that the AE topology is only supported by the most GC3-heterogeneous genes, which have shown to be unreliable for phylogenetic inference (figs. 1*A* and 3), whereas the ABM topology is strongly supported when using more reliable GC3-homogeneous genes or nonhomogeneous models of sequence evolution (figs. 1*A*, 1*B*, and 3).

To test whether GC3-heterogeneity can also explain the topological discrepancies found in previous studies, we reanalyzed the data of two studies often cited as references to favor the AE topology (Kawakita et al. 2008; Cardinal et al. 2010). Both studies included relatively few loci (12 and 7, respectively) but greater taxon sampling in the corbiculates (11 and 26 species, respectively) than our data set. We found that several genetic markers in both studies (4 of 12 markers in Kawakita et al. 2008, and two of seven markers in Cardinal et al. 2010; supplementary table S1, Supplementary Material online) had significantly heterogeneous base composition at third codon positions across the included species (see supplementary material, Supplementary Material online, for details). The removal of these GC3-heterogenous genes from the data set for each study resulted in the ABM topology instead

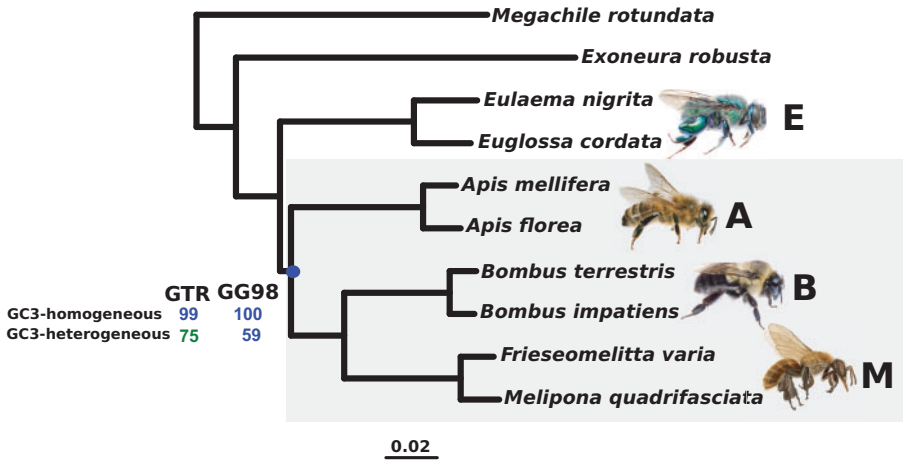

FIG. 3. Phylogenetic tree of the 10 species. The topology (ABM) is supported by the most GC-homogeneous genes. The bootstrap values of the position of Apini are displayed for maximum-likelihood analyses (GTR + gamma model using RAxML and nonhomogeneous GG98 model using nhPhyml) of two different supermatrices of equal size: 1) the group of 1,800 genes with the lowest GC3-heterogeneity and 2) the group of 1,800 genes with the highest GC3-heterogeneity. Blue bootstrap values are for the ABM topology (displayed on the tree) and green bootstrap values for the AE topology. The bootstrap values of all the other nodes are equal to 100 (all supermatrices, all models). Branch lengths correspond to the RAxML tree based on the 50% most GC3-homogeneous supermatrix. *E* stands for Euglossini (orchid bees), *A* for Apini (honeybees), *B* for Bombini (bumblebees), and *M* for Meliponini (stingless bees).

of AE (supplementary figs. S4 and S5, Supplementary Material online), although bootstrap support for ABM was weak (<50% in Kawakita et al. 2008 and 51% in Cardinal et al. 2010).

GC Effect on Tree Reconstruction

In line with what has been recently shown in mammals (Romiguier et al. 2013), GC-rich and GC-heterogeneous genes also appear to be unreliable genetic markers for phylogenetic inference in the corbiculate bees. Compared to GC-homogeneous genes, gene trees produced by GC3-heterogeneous genes presented a higher level of topological incongruence and were less likely to retrieve monophyly of corbiculate tribes, a clear effect supported by extremely significant and strong correlations (fig. 2). Interestingly, this effect is still very strong after an amino acid translation (supplementary fig. S1, Supplementary Material online), indicating that protein data are not immune to GC-biases.

Three nonmutually exclusive hypotheses can explain the origin of such a dramatic GC effect on phylogenetic reconstructions. First, GC-rich regions are characterized by high recombination rates (Duret and Arndt 2008; Kent et al. 2012), which can increase incomplete lineage sorting by increasing local levels of polymorphism (Hobolth et al. 2011).

Second, standard models of sequence evolution assume a homogeneous nucleotidic composition among species, which is violated in our ten bee species: GC3% average of the 3,600 genes ranged from 21% (*A. mellifera*) to 33% (*Exoneura robusta*). Using nonhomogeneous models of sequence evolution allowed us to retrieve a consistent topology and the highest bootstrap supports (figs. 1 and 3). It is worth noting that GC biases can also bias amino acid composition (Foster and Hickey 1999; Singer and Hickey 2000), which means that protein sequences are also vulnerable to this weakness of homogeneous models.

Third, biased gene conversion can favor homoplasy. Biased gene conversion is a bias that occurs in the process of DNA repair during meiotic recombination, promoting the fixation of G + C substitutions (Galtier et al. 2001). It has been reported that biased gene conversion can counteract natural selection by promoting the fixation of deleterious mutations toward GC (Galtier and Duret 2007; Necsulea et al. 2011), suggesting that GC deleterious substitutions are likely to be followed by AT compensatory substitutions to restore the protein function (Galtier et al. 2009). Such multiple substitutions at the same site are difficult to detect by substitution models, leading to homoplasy and long-branch attraction artifacts (Romiguier et al. 2013), which are well known to bias phylogenetic reconstructions (Bergsten 2005). This phenomenon could be particularly dramatic in bees because GC-biased gene conversion is estimated to be 50 times more important than in humans (Wallberg et al. 2015) and because the honeybee genome is characterized by an exceptional mutational bias toward AT (Kent et al. 2012; Wallberg et al. 2015). In line with previous studies linking biased gene conversion with accelerated evolution (Montoya-Burgos et al. 2003; Galtier and Duret 2007; Kostka et al. 2011; Kent et al. 2012), our analyses showed that the most GC3-heterogeneous genes are the fastest evolving (supplementary fig. S3, Supplementary Material online). Fast-evolving genes are predicted to be particularly prone to long-branch attraction artifacts, and consistent with this prediction, GC3-heterogeneous genes supported the topology grouping the two long branches of Apini and Euglossini (AE topology, figs. 1*A* and 3). Of note, the topology favored by GC-rich genes in mammals was also the topology grouping two long branches (Atlantogenata and Xenarthra, Romiguier et al. 2013), suggesting that GC-rich

markers are particularly prone to grouping distant taxa through long-branch attraction artifacts.

It is worth noting that incomplete lineage sorting, substitution model misspecification, and homoplasy are not mutually exclusive hypotheses and probably all contribute to the decreased phylogenetic reliability of GC-rich and GC-heterogeneous genes. Using nonhomogeneous models can solve the model misspecification issue, but these models do not correct for incomplete lineage sorting and homoplasy. Both incomplete lineage sorting and homoplasy can explain why, in spite of the use of nonhomogenous models, the most GC-heterogeneous genes retrieve the eusocial clade (ABM) with relatively low statistical support (59% bootstrap support, fig. 3; Bayes factor 9.08; see above and supplementary materials, Supplementary Material online).

Because GC-rich and GC-heterogeneous genes are fast evolving, their spurious phylogenetic signal can overcome the reliable signal of slow-evolving, GC-poor, and GC-homogeneous genes, which calls for gene filtering and the use of nonhomogeneous models in future phylogenomic studies.

Conclusion

Our analyses strongly suggest that the honey bees (Apini) are the closest relatives of the bumblebee + stingless bee clade. Given that these three lineages are all eusocial, this result supports a single origin of eusociality within the corbiculate bees, with no reversal to a solitary lifestyle. This finding has important implications for comparative genomic studies focusing on the evolution of social behavior (Woodard et al. 2011; Roux et al. 2014; Barribeau et al. 2015; Kapheim et al. 2015; Sadd et al. 2015; see Kent and Zayed 2015 for a review).

More broadly, our analyses reveal that the spurious phylogenetic signal of GC-rich and GC-heterogeneous genes are not restricted to mammals or birds but appears to be an important source of gene-tree conflicts across a wide variety of taxa. GC-rich gene tree conflicts can affect all taxa with GC-biased gene conversion, a mechanism known to be widespread across eukaryotes (Pessia et al. 2012) and are potentially one of the principal biases clouding the resolution of some of the more controversial nodes of the Tree of Life.

Materials and Methods

Alignments

We used the transcriptome data set from Woodard et al. (2011). This data set consists of highly curated alignments of fragments of coding sequence from 3,647 genes for ten species: two Apini (*A. mellifera* and *A. florea*), two Bombini (*B. impatiens* and *B. terrestris*, two Meliponini (*F. varia* and *M. quadrifasciata*), two Euglossini (*E. nigrita* and *Eug. cordata*), and the noncorbiculate species *Exoneura robusta* and *Megachile rotundata*. Details on the assembly and orthology prediction are given in Woodard et al. (2011).

Topology Incongruence in Gene Tree Groups

We divided the 3,647 genes into 36 ranked groups according to their average GC3-content (i.e., G + C content of third codon position) or their GC3-heterogeneity (i.e., variance in G + C at third codon position among species). To ensure that each group contained exactly 100 genes, we removed the 47 genes with GC3-mean or GC3-variance values closest to the median value.

Trees were inferred for each gene from both nucleotide and amino acid sequences using RAxML (Stamatakis 2006) with a GTR + GAMMA or LG + GAMMA model. For each of the 36 groups of 100 genes, we computed a supertree of the 100 gene trees using *SuperTriplet* (Ranwez et al. 2010). Then, we computed the average topological distance (quartet distance, Bansal et al. 2009) of the 100 gene trees versus their own supertree. This average quartet error was used as a measure of the topological incongruence of a gene tree set. For each supertree, we also computed the average support values for consensual nodes of the corbiculate bees (i.e., nodes defining the monophyly of the four tribes, Apini, Bombini, Meliponini, and Euglossini).

Supermatrices Topology Comparison

We concatenated the sequence of the genes in each of the 36 gene groups to generate 36 supermatrices of 100 genes and performed maximum-likelihood analyses on each of them, using RAxML (Stamatakis 2006) with a GTR + GAMMA model and 100 bootstrap replications. To control for heterogeneous base composition, we performed similar analyses using nhPhyml (Boussau and Gouy 2006). This software uses the nonstationary, nonhomogenous model GaltierGouy98 (GG98, Galtier and Gouy 1998) and takes into account heterogeneity in base composition when modelling the sequence evolution process. To avoid convergence problems arising when trying to optimize the G + C equilibrium frequency for each branch, we used the option *eqfreq* set to *lim* and the option *numeqfreq* set to 5, which means that each branch was limited by five different sets of G + C equilibrium frequencies. To be as conservative as possible, we used the topology the most supported in the literature and by GC3-heterogeneous genes (AE) as the starting tree. Because nhPhyml does not enable default bootstrap functionality, we sampled sites using a de novo script to perform 100 bootstrap replications for each supermatrix. Because the GG98 model is based on a T92 substitution model, while our homogenous analyses implemented a GTR model, we repeated the RAxML analyses with a HKY85 model (the T92 is not available in RAxML) to ensure that the differences were not due to the distinct substitution model. Results obtained with the HKY85 model (not shown) were virtually identical to those obtained with the GTR.

We also divided the 3,600 genes into two groups of 1,800 (high and low GC3-heterogeneity) based on GC3-heterogeneity rank. We concatenated the sequences of genes in each group into supermatrices and performed the same maximum-likelihood analyses described above. We applied a stringent cleaning on these large supermatrices, removing all positions with at least one gap or one ambiguous nucleotide.

Supplementary Material

Supplementary figures S1–S5 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Nicolas Salamin, Nils Arrigo, and three anonymous reviewers for comments on this manuscript. We also thank Peter G. Foster for his kind help and advices in using P4-phylogenetics. Some computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing (HPC) of the SIB Swiss Institute of Bioinformatics and the EPSRC-funded MidPlus HPC center. This work was supported by a Federation of European Biochemical Societies (FEBS) long-term fellowship to J.R., the Swiss NSF, and an ERC Advanced Grant.

References

Almeida EAB, Porto DS. 2014. Investigating eusociality in bees while trusting the uncertainty. *Sociobiology* 61:355–368.

Ascher J, Danforth B, Ji S. 2001. Phylogenetic utility of the major opsin in bees (Hymenoptera: Apoidea): a reassessment. *Mol Phylogenet Evol.* 19:76–93.

Bansal MS, Dong J, Fernández-Baca D. 2009. Comparing and aggregating partially resolved trees. *Science* 412:6634–6652.

Barribeau SM, Sadd BM, du Plessis L, Brown MJ, Buechel SD, Cappelle K, Carolan JC, Christiaens O, Colgan TJ, Erler S, et al. 2015. A depauperate immune repertoire precedes evolution of sociality in bees. *Genome Biol.* 16:83.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21:163–193.

Betancur RR, Li C, Munroe TA, Ballesteros JA, Ortí G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). *Syst Biol.* 62:763–785.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, et al. 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome Res.* 16:1339–1344.

Boussau B, Gouy M. 2006. Efficient likelihood computations with non-reversible models of evolution. *Syst Biol.* 55:756–768.

Cameron SA. 1993. Multiple origins of advanced eusociality in bees inferred from mitochondrial DNA sequences. *Proc Natl Acad Sci USA.* 90:8687–8691.

Cameron SA, Mardulyn P. 2001. Multiple molecular data sets suggest independent origins of highly eusocial behavior in bees (Hymenoptera:Apinae). *Syst Biol.* 50:194–214.

Cameron SA, Mardulyn P. 2003. The major opsin gene is useful for inferring higher level phylogenetic relationships of the corbiculate bees. *Mol Phylogenet Evol.* 28:610–613.

Cardinal S, Danforth BN. 2011. The antiquity and evolutionary history of social behavior in bees. *PLoS One* 6:e21086.

Cardinal S, Packer L. 2007. Phylogenetic analysis of the corbiculate Apinae based on morphology of the sting apparatus (Hymenoptera: Apidae). *Cladistics* 23:99–118.

Cardinal S, Straka J, Danforth BN. 2010. Comprehensive phylogeny of apid bees reveals the evolutionary origins and antiquity of cleptoparasitism. *Proc Natl Acad Sci USA.* 107:16207–16211.

Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D. 2013. The impact of molecular data on our understanding of bee phylogeny and evolution. *Annu Rev Entomol.* 58:57–78.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol.* 24:332–340.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4:e1000071.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc R Soc B Biol Sci.* 252:237–243.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53:485–495.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol.* 48:284–290.

Figuet E, Ballenghien M, Romiguier J, Galtier N. 2014. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol Evol.* 7:240–250.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23:273–277.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25:1–5.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.* 15:871–879.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159:907–911.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol.* 80:231–266.

Hedtke SM, Patiny S, Danforth BN. 2013. The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. *BMC Evol Biol.* 13:138.

Hobolth A, Dutheil J, Hawks J, Schierup M. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* 21:349–356.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

Jørgensen FG, Schierup MH, Clark AG. 2007. Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of *Apis mellifera*. *Mol Biol Evol.* 24:611–619.

Kapheim K, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. *Science* 348:1139–1143.

Kawakita A, Ascher JS, Sota T, Kato M, Roubik DW. 2008. Phylogenetic analysis of the corbiculate bee tribes based on 12 nuclear protein-coding genes (Hymenoptera: Apoidea: Apidae). *Apidologie* 39:163–175.

Kent C, Minaei S, Harpur BA, Zayed A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc Natl Acad Sci USA.* 109:18012–18017.

Kent CF, Zayed A. 2015. Chapter nine—population genomic and phylogenomic insights into the evolution of physiology and behaviour in social insects. In: Zayed A, Kent FK, editors. Advances in insect physiology. Waltham: Academic Press. p. 293–324.

Kocher SD, Paxton RJ. 2014. Comparative methods offer powerful insights into social evolution in bees. *Apidologie* 45:289–305.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2011. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol.* 29:1047–1057.

Koulianos S, Schmid-Hempel R, Roubik DW, Schmid-Hempel P. 1999. Phylogenetic relationships within the corbiculate Apinae (Hymenoptera) and the evolution of eusociality. *J Evol Biol.* 12:380–384.

Liu L, Xi Z, Wu S, Davis CC, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. *Ann N Y Acad Sci.*

Lockhart PJ, Cameron SA. 2001. Trees for bees. *Trends Ecol Evol.* 16:84–88.

Mardulyn P, Cameron S. 1999. The major opsin in bees (Insecta: Hymenoptera): a promising nuclear gene for higher level phylogenetics. *Mol Phylogenet Evol.* 12:168–176.

Michener CD. 2007. The bees of the world. Baltimore: The John Hopkins University Press.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet.* 19:128–130.

Nabholz B, Künstner A, Wang R, Jarvis E, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol Biol Evol.* 28:2197–2210.

Necsulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat.* 32:198–206.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol.* 5:568–583.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol.* 4:1–20.

Ranwez V, Criscuolo A, Douzery EJP. 2010. SuperTriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics* 26:115–123.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol Biol Evol.* 30:2134–2144.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20:1001–1009.

Ross CR, DeFelice DS, Hunt GJ, Ihle KE, Amdam GV, Rueppell O. 2015. Genomic correlates of recombination rate and its variability across eight recombination maps in the western honey bee (*Apis mellifera* L.). *BMC Genomics* 16:107.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol.* 31:1661–1685.

Sadd B, Barribeau S, Bloch G. 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.* 16:1–32.

Singer GAC, Hickey, DA. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol.* 17:1581–1588.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, et al. 2011. The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.* 7:e1002007.

Wallberg A, Glémin S, Webster MT. 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genet.* 11:e1005189.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15:1–16.

Whitfield JB, Cameron SA, Huson DH, Steel MA. 2008. Filtered Z-closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees. *Syst Biol.* 57:939–947.

Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* 98:189–197.

Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron SA, Clark AG, Robinson GE. 2011. Genes involved in convergent evolution of eusociality in bees. *Proc Natl Acad Sci USA.* 108:7472–7477.