Editor's Choice Article

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

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.

© 2017 Elsevier Inc. All rights reserved.

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

and Salter, 2005; Maddison, 1997; Pamilo and Nei, 1988), gene loss or duplication events (Galtier and Daubin, 2008; Goodman et al., 1979; Maddison, 1997), and introgression (Fontaine et al., 2015; Kronforst, 2008) can yield gene trees that are incongruent with each other and with the actual species tree, even though the genealogy is correctly inferred.

One important factor that impacts the level of incongruence in phylogenomic data sets is the composition of GC content. GC rich gene regions are linked to higher recombination rates than AT rich alleles, substantially driven by GC biased gene conversion (gBGC, e.g., Figuet et al., 2015; Kent et al., 2012; Lartillot, 2013). Experimental evidence revealed gene conversion to be biased towards the fixation of GC nucleotides instead of AT nucleotides when alleles are heterozygous (Lesecque et al., 2013; Mancera et al., 2008). This effect leads to unequal base compositional distributions in genomes due to higher fixation rates of GC in frequently recombining regions, so-called "recombination hotspots" (Duret and Galtier, 2009; Mancera et al., 2008). In turn, elevated recombination rates enhance incongruence among gene trees through increased probabilities of ILS, an effect which can be theoretically substantiated by speciation time and the effective population size (N_e) of given loci (Maddison, 1997; Pamilo and Nei, 1988). In fact, recombination rate and overall congruence were shown to be negatively correlated in different systems (Pease and Hahn, 2013) and ILS was specifically shown to increase with recombination frequency (Hobolth et al., 2011).

Generally, two different measures are used to identify frequently recombining, GC biased loci: Overall GC content and GC heterogeneity. By "GC content", we mean the overall percentage of GC across a range of aligned nucleotide sites (i.e., gene or locus). "GC heterogeneity" is the degree to which individual taxa deviate in GC content from the overall background level of GC content in the alignment (i.e., the variance of GC content among taxa). The latter measure is less widely used but has been shown to impact topological incongruence on phylogenetic reconstruction of the corbiculate bees using transcriptome data (Romiguier et al., 2015).

Phylogenomic data sets provide a strong framework for evaluating the impact of GC bias on gene-tree incongruence. One increasingly popular approach to generating genomic-level data sets of thousands of loci is hybrid-based targeted enrichment (Faircloth et al., 2012; Lemmon et al., 2012). Targeted enrichment methods allow one to effectively shortcut the difficult and labor-intensive challenges of whole-genome assembly and orthology identification for non-model taxa. Faircloth et al.'s (2012) approach uses in-solution sequence capture (sensu Gnirke et al., 2009) of large quantities of so-called "ultraconserved elements" (UCEs). These universal markers were shown to be superior to traditional multi-locus sequencing (Blaimer et al., 2015) and provide great applicability for degraded material, such as pinned museum specimens (Blaimer et al., 2016). Interestingly, UCEs are generally more AT-rich than GC-rich (Faircloth et al., 2012, Supplementary Fig. 2), suggesting that they may be less prone to the kind of gene conversion that can increase gene-tree incongruence in phylogenomic data sets. Both UCEs and selectively filtered AT-rich protein-coding phylogenomic data sets performed comparably well in resolving inconclusive nodes in mammal phylogenies (McCormack et al., 2012 and Romiguier et al., 2013, respectively). However, the impact of GC composition on incongruence among UCE loci is currently unstudied.

In this paper, we investigate the impact of GC content and GC heterogeneity on phylogeny reconstruction using in silico captured UCE data for the corbiculate bees (Hymenoptera: Apidae). Corbiculate bees include a total of 1019 species (Ascher and Pickering, 2016) placed in four tribes: Euglossini (orchid bees), Bombini (bumblebees), Meliponini (stingless bees), and Apini (honeybees). These four tribes are united by the possession of a highly modified hind tibia, which forms a concave, shiny structure (the corbicula) for carrying pollen and plant resins (Martins et al., 2014). Euglossini are typically viewed as solitary or communal but a number of studies have indicated that some species of *Euglossa* exhibit features typical of a more social bee, including multifemale nests, overlap of generations, and temporary division of labor in which some females forage and others guard the nest (reviewed in Cardinal and Danforth, 2011). The vast majority of the 260 species of bumblebees are primitively eusocial. Both Apini and Meliponini are advanced eusocial with obligate swarm founding and morphologically distinct castes.

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 investigation 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 GG98 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
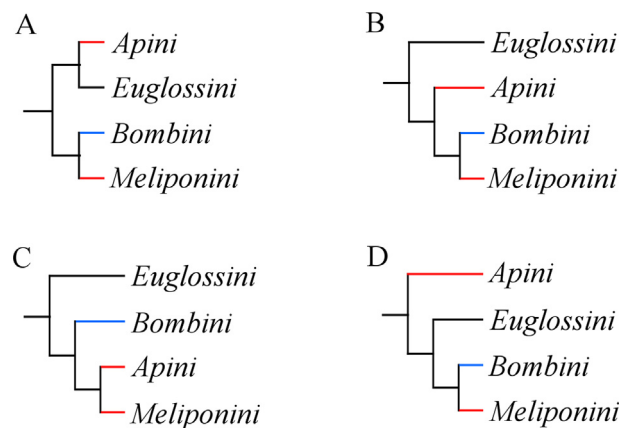

Fig. 1. Competing topologies of the corbiculate lineages. Topology A is favored by the majority of the single and multi-locus studies (e.g., Cardinal and Danforth, 2013; Kawakita et al., 2008), whereas topology B was revealed by a recent genomic study which accounted for GC bias (Romiguier et al., 2015). Topology C was inferred multiple times based on morphology (e.g., Cardinal and Packer, 2007; Plant and Paulus, 2016) but not by molecular data. Topology D is unlikely from the perspective of the evolution of eusociality but has been obtained using large data sets (Romiguier et al., 2015). Colored lines indicate eusociality of different degrees; red - advanced, blue - primitive. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)