Systematic Entomology

Systematic Entomology (2020), DOI: 10.1111/syen.12454

Assessing support for Blaberoidea phylogeny suggests optimal locus quality

DOMINIC EVANGELISTA[1,2,3], SABRINA SIMON[4], MEGAN M. WILSON[5], AKITO Y. KAWAHARA[6], MANPREET K. KOHLI[7], JESSICA L. WARE[7], BENJAMIN WIPFLER[8], OLIVIER BÉTHOUX[9], PHILIPPE GRANDCOLAS[1] and FRÉDÉRIC LEGENDRE[1]

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, UA, Paris, France, [2]Department of Ecology and Evolutionary Biology, The University of Tennessee, Knoxville, Tennessee, U.S.A., [3]Department of Biology, College of Arts and Sciences, Adelphi University, Garden City, New York, U.S.A., [4]Biosystematics Group, Wageningen University and Research, Wageningen, The Netherlands, [5]Federated Department of Biological Sciences, Rutgers, The State University of New Jersey and NJIT, Newark, New Jersey, U.S.A., [6]Florida Museum of Natural History, University of Florida, Gainesville, Florida, U.S.A., [7]Department of Invertebrate Zoology, American Museum of Natural History, New York, New York, U.S.A., [8]Center for Taxonomy and Evolutionary Research, Zoological Research Museum Alexander Koenig (ZFMK), Bonn, Germany and [9]CR2P (Centre de Recherche en Paléontologie – Paris), MNHN – CNRS – Sorbonne Université, Paris, France

Abstract. Phylogenomics seeks to use next-generation data to robustly infer an organism's evolutionary history. Yet, the practical caveats of phylogenomics motivate investigation of improved efficiency, particularly when quality of phylogenies are questionable. To achieve improvements, one goal is to maintain or enhance the quality of phylogenetic inference while severely reducing dataset size. We approach this by assessing which kinds of loci in phylogenomic alignments provide the majority of support for a phylogenetic inference of cockroaches in Blaberoidea. We examine locus substitution rate, saturation, evolutionary divergence, rate heterogeneity, stabilizing selection, and *a priori* information content as traits that may determine optimality. Our controlled experimental design is based on 265 loci for 102 blaberoidean taxa and 22 outgroup species. Loci with high substitution rate, low saturation, low sequence distance, low rate heterogeneity, and strong stabilizing selection derive more support for phylogenetic relationships. We found that some phylogenetic information content estimators may not be meaningful for assessing information content *a priori*. We use these findings to design concatenated datasets with an optimized subsample of 100 loci. The tree inferred from the optimized subsample alignment was largely identical to that inferred from all 265 loci but with less evidence of long branch attraction, improved statistical support, and potential 4-6x improvements to computation time. Supported by phylogenetic and morphological evidence, we erect three newly named clades (Anallactinae Evangelista & Wipfler *subfam. nov.*, Orkrasomeria *tax. nov.* Evangelista, Wipfler, & Béthoux and Hemithyrsocerini Evangelista *tribe nov.*) and propose other taxonomic modifications. The diagnosis of Pseudophyllodromiidae Grandcolas, 1996 is modified to accommodate Anallactinae and Pseudophyllodromiinae Vickery & Kevan, 1983. The diagnosis of Ectobiidae Brunner von Wattenwyl, 1865 is modified to add novel morphological characters.

Correspondence: Dominic A. Evangelista and Frédéric Legendre, Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, UA, 57 rue Cuvier, CP50, 75005 Paris, France. E-mail: dominicev@gmail.com (D. A. E.) and (E-mail: frederic.legendre@mnhn.fr (F. L.)

Introduction

The current 'postgenomic' era of phylogenetics is characterized by large datasets and increased efforts to maximize their usage (Bravo *et al.*, 2019). Yet, while more data increase potential phylogenetic information (Simon *et al.*, 2018), they can also increase data artefacts and bias (Breinholt & Kawahara, 2013; Dell'Ampio *et al.*, 2014; Brown & Thomson, 2017; Shen *et al.*, 2017; Platt *et al.*, 2018). In this study, we target genomic loci and assess their contribution to phylogenetic inference and topological support through phylogenetic subsampling. There are known features of loci that are thought to be more informative than others for inferring phylogenetic relationships (e.g., optimal evolutionary rate, saturation, long branch score, heterogeneity, entropy; Borowiec *et al.*, 2015; Brown & Thomson, 2017; Gilbert *et al.*, 2015; Klopfstein *et al.*, 2017; Molloy & Warnow, 2018; Reddy *et al.*, 2017; Tan *et al.*, 2015; Townsend, 2007). Phylogenetic subsampling allows assessment of the presence of confounding signal (Edwards, 2016; Simon *et al.*, 2018). This approach has been used to test how topologies change with taxon sampling (Evangelista *et al.*, 2018; Simon *et al.*, 2018), the utility of loci with different evolutionary rates (Narechania *et al.*, 2012), the stability of nodes of biological interest (Simon *et al.*, 2012), the effect of missing data (Xi *et al.*, 2016), and other phenomena (Edwards, 2016).

Determining the optimality of loci (Townsend, 2007; Reddy *et al.*, 2017), as opposed to characters, must be a practical priority given that DNA sequencing technologies read contiguous regions of genomes. This is especially true when deciding which loci to target using popular genome capture strategies (e.g., Lemmon *et al.*, 2012; Brandley *et al.*, 2015; Gilbert *et al.*, 2015). Although the features of individual characters are directly relevant to phylogenetic reconstruction (Dornburg *et al.*, 2018), removing characters in alignments may fail to improve phylogenetic accuracy (Tan *et al.*, 2015). Here, we use the term "locus" to refer to some contiguous segment of a genome more than a few nucleotides long.

The relationship between the features and phylogenetic utility of molecular loci has long been studied (e.g., see Blaxter, 2004). Perhaps the prime suspect in investigating a locus' quality for phylogenetic inference is its evolutionary rate (Klopfstein *et al.*, 2017). The optimal mean rate of a locus for resolving difficult tree shapes (Steel & Leuenberger, 2017; Dornburg *et al.*, 2018) is very conservative (Klopfstein *et al.*, 2017; Steel & Leuenberger, 2017). Yet, targeting optimal evolutionary rates does not necessarily improve phylogenetic inference (Narechania *et al.*, 2012; Chen *et al.*, 2015; Doyle *et al.*, 2015). High evolutionary rate can hinder phylogenomic studies when it leads to substitution saturation (Fong *et al.*, 2012; Breinholt & Kawahara, 2013). A locus' evolutionary divergence among a set of taxa may better indicate its phylogenetic utility. This measure loosely accounts for phylogeny and is related to other informative features such as: long branch score (Struck, 2014; Borowiec *et al.*, 2015), evolutionary rate (Chen *et al.*, 2015), saturation (Borowiec, 2019), and entropy (Bai *et al.*, 2013; Lewis *et al.*, 2016). Conforming to model assumptions is another feature that should improve phylogenetic trees (Doyle *et al.*, 2015; Reddy *et al.*, 2017; but see Dell'Ampio *et al.*, 2014). For instance, patterns of rate heterogeneity may be poorly estimated by the gamma distribution (Kjer & Honeycutt, 2007). Finally, stabilizing selective pressure is another desirable trait. Positive selection pressure may mislead phylogenetic inference through convergent evolution, which may further induce compositional bias (Cox *et al.*, 2014).

The most direct approach might be to measure phylogenetic information content of loci. Information content can be assessed several ways, including: (*i*) with reference to a phylogenetic tree (Townsend, 2007) or (*ii*) without one (Misof *et al.*, 2014b); (*iii*) in a Bayesian framework (Dornburg *et al.*, 2016; Lewis *et al.*, 2016); or (*iv*) in an empirical approach (Kück *et al.*, 2012; Misof *et al.*, 2014b). Using only high information content markers has sometimes improved tree congruence (Chen *et al.*, 2015) and given greater agreement with favoured morphological hypotheses (Borowiec *et al.*, 2015). Yet, assessing information content before all data are collected (*a priori*) is challenging because new taxa can influence how character information is interpreted (Venditti *et al.*, 2006; Hugall & Lee, 2007).

We examine node support for an empirical phylogeny derived from sets of loci of varying quality. We utilize data subsampling (Narechania *et al.*, 2012; Edwards *et al.*, 2017) to assess rate of convergence on reasonable tree topologies and certainty of nodes under various data subsamples (e.g., Soltis & Soltis, 2003; Simon *et al.*, 2012; Edwards, 2016; Evangelista *et al.*, 2018; Simmons *et al.*, 2018; Simon *et al.*, 2018). We test six features suspected to affect the accuracy of phylogenetic inference, including substitution rate, substitution saturation, pairwise sequence distance, among site rate heterogeneity, selective pressure, and *a priori* phylogenetic information content.

Our empirical aims are to recover a robust phylogeny of blaberoidean cockroaches. Blaberoidea is a clade ∼170 Myr-old (Evangelista *et al.*, 2019) or older (Djernæs *et al.*, 2015; Legendre *et al.*, 2015) comprising ∼3600 of the ∼7500 species of Blattodea (Beccaloni & Eggleton, 2013). The major relationships of Blaberoidea were recently investigated in a phylogenomic study (Evangelista *et al.*, 2019) but taxonomic sampling was limited. The major lineages of Blaberidae and many other splits occurring in the last 100 Myr were not sampled (Evangelista *et al.*, 2019). Studies that included denser taxon sampling recovered unstable topologies (Legendre *et al.*, 2015; Legendre *et al.*, 2017; Bourguignon *et al.*, 2018; Evangelista *et al.*, 2018). We combine previously published phylogenomic datasets with 265 genomic loci for 90 newly sequenced species of blaberoidean cockroaches. From the 265 loci dataset, we compose two optimized subsamples of 100 loci each. By comparing the full data tree and the two subsample trees, we derive conclusions about tree plausibility and the phylogeny of Blaberoidea.

Materials and methods

Taxon sampling

We added 90 newly sequenced Blaberoidea to the Blattodea sampled in Evangelista *et al.* (2019) and Wipfler *et al.* (2019).

Accession numbers and sample details are listed in supplemental data. These were chosen to cover biogeographical regions and known 'tribal-level' groupings as widely as possible (Supplemental Data). Live samples were obtained from breeders and remaining samples were museum specimens (see NCBI accession SRP155429 for full specimen information). Tissue was taken from the middle leg of all samples. Polyneopteran outgroups were chosen among published transcriptomes (Misof *et al.*, 2014a; Wipfler *et al.*, 2019).

Loci capture and probe design

Using alignments of orthologous loci from Misof *et al.* (2014a) we extracted sequences for *Blaberus atropos* and then used BLAST+ (Camacho *et al.*, 2009) to identify homologous sequences in transcriptomes of 17 cockroach species (*Anallacta methanoides*, *Asiablatta kyotensis*, *Blattella germanica*, *Cryptocercus wrighti*, *Diploptera punctata*, *Ectobius sylvestris*, *Ellipsidion* sp., *Gyna lurida*, *Ischnoptera deropeltiformis*, *Lamproblatta albipalpus*, *Lobopterella dimidiatipes*, *Nauphoeta cinerea*, *Panchlora nivea*, *Paratemnopteryx couloniana*, *Princisia vanwaerebecki*, *Sundablatta sexpunctata*, and *Symploce* sp.). These were added to the *B. atropos* sequences and aligned with MAFFT v. 7.3 (Katoh & Standley, 2013; -localpair - maxiterate 1000 - adjustdirection). Of the resulting 599 alignments, we chose 100 at random, 92 of which demonstrated high *a priori* information content, and 90 of which demonstrated tree likeness but not high information content. We define loci having high *a priori* phylogenetic information content as being able to recover a resolved topology (tree-likeness) and having high character support for non-conflicting bipartitions (see Supplement S1 for details). Probes for targeted enrichment were designed to target the final set of ∼280 alignments (untrimmed, nucleotide versions) using Baitfisher v. 1.2.5 (Mayer *et al.*, 2016). The resulting probe sequences were further quality-filtered and subsequently produced by Arbor Biosciences (formerly, MYcroarray).

Library preparation and sequencing

Genomic DNA was extracted from specimens belonging to 90 species (Supplementary Data) using the DNEasy kit (Qiagen USA), sheared using a timed Fragmentase (New England BioLabs Inc.) protocol and assessed using a Bioanalyzer (Agilent, Santa Clara, CA). Index-multiplexing strategies were planned and specimens were arranged on the plate to minimize the probability of cross-contaminating closely related samples. After NEBNext Ultra library preparation, indexed library pools were enriched with respect to targeted loci (MYbaits v.3.01) and then sequenced by GeneWiz USA using Illumina HiSeq's rapid-run paired-end-250 protocol.

Bioinformatics and phylogenetic inference

Demultiplexed FASTQ files had adapters removed and low-quality bases trimmed (Krueger, 2017; Trimgalore v. 0.4.3 options: -q 20; -stringency 3; -e 0.1; -length 30). Reads were assembled with Trinity v.2.0.6 (Grabherr *et al.*, 2011; Haas *et al.*, 2013). Orthologs from OrthoDB v. 7 (Waterhouse *et al.*, 2013) were identified in the 90 sequenced libraries and 37 additional Blattodea and Polyneoptera sequences (Evangelista *et al.*, 2019) using Orthograph v. 0.6 (Petersen *et al.*, 2017) with *Drosophila melanogaster*, *Pediculus humanus*, *Tribolium castaneum*, and *Zootermopsis nevadensis* as reference genomes [default options except – any symbol option called in MAFFT (Katoh & Standley, 2013), and blast-max-hits = 50]. Two hundred and sixty five targeted loci were extracted from the ortholog set. Quality filtered reads are available on NCBI GenBank (SRP155429).

Each locus was aligned in MAFFT v. 7.3 (Katoh & Standley, 2013; options: - retree 4 - maxiterate 10 - adjust direction) and then trimmed from the edges to eliminate leading or trailing sites missing >80% of data. A second, alignment was conducted in MAFFT v. 7.3 (-localpair - maxiterate 1000), which was then adjusted to maintain consistent reading-frame. Alignments were finalized with manual adjustment in AliView v. 1.18 (Larsson, 2014) to remove poor quality reads and correct misaligned sections. Custom scripts in Mathematica 10 (Wolfram Research, 2012) available in the package Phyloinformatica v. 0.9 (Evangelista, 2019) were used to manage sequence files, translate sequences, trim and concatenate alignments. We refer to the final concatenated alignment (127 taxa) as the" 265_Full" alignment.

We ran PartitionFinder 2 (Lanfear *et al.*, 2016) with the 265_Full alignment with blocks defined as codon positions per locus, possible models as GTR and GTR + G, branch lengths considered as unlinked, best model chosen with AICc and rcluster search scheme (percent = 10; max = 1000). Using the resulting codon partitioning scheme, we ran a preliminary tree search with 10 independent runs in RAxML v. 8.2 (Stamatakis, 2014), implemented on the CIPRES portal (Miller *et al.*, 2010). Assessment of the best preliminary tree showed that a few taxa (*Amazonina platystylata*, *Doradoblatta* sp., *Ischnoptera galibi*, *Lanxoblatta* sp., *Panchlora stolata*, *Pycnoscelus femapterus*, *Pycnoscelus striata*) had exceptionally long branches. The same taxa were among those with the largest proportion of missing data (Supplementary Data). After reassessing the alignments in which these species were present, we removed *Pycnoscelus striata*, *P. femapterous*, *Ischnoptera galibi*, *Amazonina platystylata* and *Doradoblatta* sp. from the analysis under the grounds that (i) their data were low quality (short reads ambiguously aligned and with many nucleotide differences) and (ii) the pattern of data presence would not allow for testing of their hypothesized taxonomic assignment (see Supplementary Data). When running the tree searches, there were no exceedingly long branch lengths, and Blaberidae was monophyletic.

Trees were inferred for three alignments: (1) the full alignment ('265_Full'), (2) using only the second codon positions (correcting for noise; '265_2nd'), (3) low missing data alignment (correcting for relationships inferred from missing data patterns; '265_Reduced'). The latter alignment was created by only retaining nucleotide positions having data for 51 or more taxa (Phyloinformatica function trimAlign2

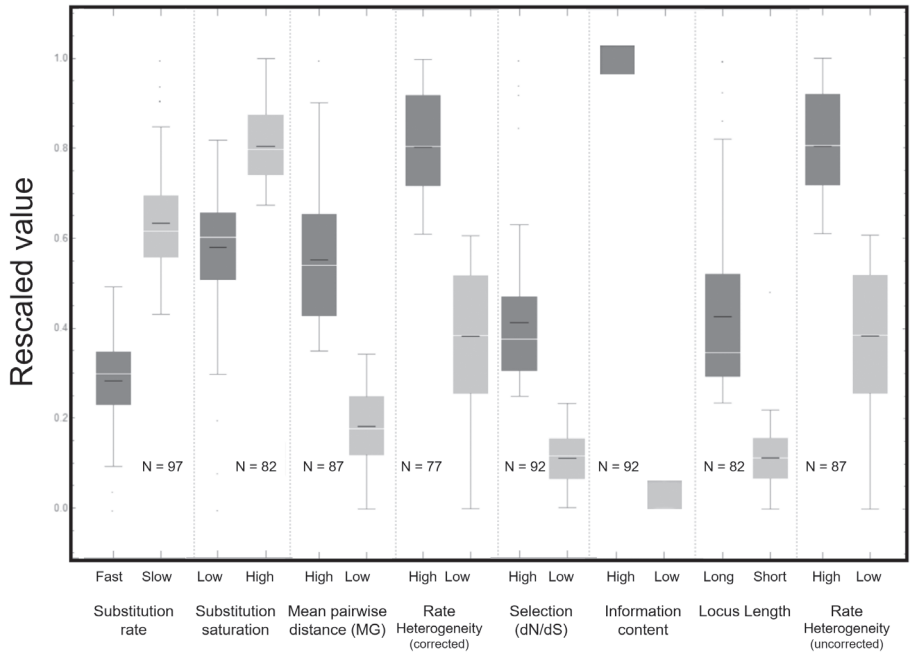

Fig. 1. Distributions of feature values tested for phylogenetic utility. Box-plots show the distribution of values (rescaled between 0 and 1) for both treatments of eight factors tested: substitution rate, substitution saturation, mean pairwise sequence distance, among site rate heterogeneity corrected for locus length, selection (dN/dS), *a prori* information content, locus length and total among site rate heterogeneity. Boxes represent middle 75% quantiles; whiskers represent the remaining quantiles with points being outliers. White lines are the median and black lines are the mean. N indicates how many loci are in each treatment.

missingProportion = 0.60; Evangelista, 2019)). The same partitioning, modelling and RAxML parameters as used above were applied to each analysis but with 100 independent tree runs (GTRGAMMA, -f d, -N 100). We inferred one more tree using the 265_Full alignment in IQTree (Nguyen *et al.*, 2015) using partitions determined by PartionFinder2, models determined by IQTree (Kalyaanamoorthy *et al.*, 2017) and the options: -ninit 200 -nbest 10 -allnni -ntop 40 -wbt -wsl -wsr. These four trees (265_Full, 265_2nd, 265_Reduced, 265_Full_IQ) were later used as a baseline for establishing subsample convergence on a reasonable topology. We assessed support for the three RAxML trees by bootstrap resampling using the auto-MRE stopping criteria (60, 300 and 108 for the first three trees respectively) and calculating bootstrap frequencies and node certainty scores (Kobert *et al.*, 2016).

Designing locus subsample sets

We designed experimental samples of loci based on calculations of the following six features for each of the 265 individual alignments: substitution rate, substitution saturation, mean pairwise sequence distance, substitution rate heterogeneity, level of selection, and *a priori* information content. We additionally calculated data completeness, and nucleotide compositional bias. Details on how each feature was calculated are given in Supplement S1. Each subsample was designed to be non-overlapping in one of our six features (e.g., all loci with high substitution rate or all loci with low substitution rate) (Fig. 1, Table S1.2), while independent to all other features (Fig. S1.1). The saturation and selection test subsamples could not be made sufficiently independent (Table S1.2), so additional controlling tests were done (see below and Fig. S4.1). A few selected loci were then added to ensure all taxa were represented in each set. *Panchlora stolata*, and *Lanxoblatta* sp., were too poorly represented among loci to include in all sets, so they were removed from all alignments and trees in our tests.

Random addition concatenation subsampling

We utilized the random addition concatenation of loci (RADICAL) described in Narechania *et al.* (2012) to infer phylogenies from each subsample set. This was implemented using the 'radicalRun' function in Phyloinformatica v. 0.93. Within RADICAL, the fast tree reconstruction method (options -fast, -ntop 10) in IQ-TREE (Nguyen *et al.*, 2015) was utilized. To overcome the low data completeness for some taxa, the first step of this implementation chose the three alignments with the most taxa represented and concatenate them with nine randomly drawn alignments. Subsequently, five loci were added at random for each additional step until all loci were sampled. A starting BIONJ tree was used in the first step (option -t BIONJ) and subsequent steps use the final tree from the previous step as an initial tree. The GTR + G model was used in an unpartitioned analysis parallelized over two cores. This was repeated for ∼100 iterations for each treatment.

To assess convergence of each subsample set, we measured the mean of all Robinson-Foulds (RF) distances (Robinson & Foulds, 1981) from the RADICAL tree from each step at each iteration to the four full-data trees (265_Full, 265_2nd, 265_Reduced, 265_Full_IQ). We plotted the means and fit with exponential models, which Narechania *et al.* (2012) showed was the shape of typical RADICAL curves.

To assess stability of topologies recovered from each set, a measure of phylogenetic precision, we plotted RF comparisons among all trees within a RADICAL step. We used a linear best-fit model for this data, as opposed to the exponential model, to accommodate the expectations that some RF values could equal 0.

We compared distributions of RF distances among the test subsampling treatments to assess the differences among subsamples. We compare each of the RADICAL tests at the 14th step (the latest step that is not the last step for any treatment) and the last step, which showed identical patterns of statistical significance. All statistical comparisons were done with a Z-Test, which is not sensitive to slight deviations from normality with sample size greater than 40 (Ghasemi & Zahediasl, 2012).

We also used each RADICAL tree set to assess their certainty (Kobert *et al.*, 2016) of relationships in the 265_Full tree. We calculated relative tree certainty (rTC) and relative tree certainty-all (rTCA) using the stochastic bipartition adjustment.

Final tree searches

The results from RADICAL identified five locus features that supported relationships more than others. To test the effect of applying this concept of phylogenetic information content to an analysis *a priori*, we selected one set of 100 loci demonstrating the optimal combination of these features (65 798 total nucleotides; '100_Full' alignment) and a second set of 100 that also considered taxon completeness (83 822 nucleotides long; 'C100_Full' alignment; see Supplement S1 and Table S1.3 for details).

Final tree searches and bootstrapping were performed on the 100_Full and C100_Full alignments in RAxML v. 8.2 using the same protocol, partitioning and node support strategy discussed above with 100 independent tree searches. Searches were again performed on three conformations of each alignment: 1) all 100 loci ('Full'), 2) only second codon positions ('second'), 3) only positions present for more than 50% of taxa ('Reduced').

Results

The 265_Full phylogenetic inference (Fig. 2) yields a number of deep relationships with strong node support: Ectobiinae (Ectobiidae) is recovered as sister to all other Blaberoidea; Anallactinae (*subfam. nov.*; Supplement S2) is recovered as sister to Pseudophyllodromiinae (together Pseudophyllodromiidae Supplemental Section S2); Blattellidae + Nyctiboridae (see Supplemental Section S2) inferred as sister to Blaberidae (together Orkrasomeria *tax. nov*. Supplemental Section S2) and Blaberidae is recovered as monophyletic. There is strong node support throughout Blattellidae + Nyctiboridae except for within Nyctiboridae, Pseudomopini and Hemithyrsocerini *tribe nov*. There is also strong support in most of Pseudophyllodromiidae with the exception being Plectopterini *s.l.* Node support is lacking throughout most of Blaberidae but two moderately well supported clades (the 'Peri-Atlantic' and 'Peri-Indian' clades) sort most taxa. *Rhabdoblatta stipata* as sister to the remaining Blaberidae, and Neotropical Epilamprinae as sister to (Peri-Atlantic + Peri-Indian Blaberidae) are strongly supported by our data.

Node support for the 265_Full tree (Fig. 2) was highest among bootstrap replicates inferred from the 265_Reduced alignment (mean bootstrap = 0.94; mean IC = 0.85), moderately high among bootstrap replicates from the 265_Full alignment (mean bootstrap = 0.93; mean IC = 0.83), and lowest among bootstrap replicates from the 265_second alignment (mean bootstrap = 0.85; mean IC = 0.66). Numerous relationships in the 265_Full and 265_Reduced trees were not recovered in the best tree inference with only second codon positions (Fig. 2).

Experimental comparisons of six features (Fig. 1) using the random-addition concatenation (RADICAL) process (Fig. 3) showed that fast substitution rate, low saturation, low mean pairwise sequence distance, low rate heterogeneity and strong stabilizing selection all contribute to faster convergence to the full data tree topologies (P values all $\ll 0.005$; Fig. 4A; also see, Fig. S4.2). Loci with high rate heterogeneity, which tend to be longer, improved tree recovery but only when not corrected for locus length (Fig. S4.4). Fast substitution rate, low mean pairwise sequence distance, low rate heterogeneity, strong stabilizing selection, and low information content all result in more precise estimation of trees (P values all $\ll 0.005$; Fig. 4B; also see, Fig. S4.3). Saturation had no statistically significant effect on tree precision.

Relative tree certainty (rTC) of the 265_Full topology as calculated from RADICAL subsample trees (Table 1) was highest among the subsamples for fast substitution rate (rTC = 0.58), high *a priori* information content (rTC = 0.58), and low mean pairwise sequence distance (rTC = 0.56). The lowest support is demonstrated by loci that are highly unsaturated (rTC = 0.462), high mean pairwise sequence distance (rTC = 0.49) and low rate heterogeneity (rTC = 0.49). These values can be interpreted as the sum of all internode certainties (IC) normalized by the number of non-tip edges. IC scores are the relative frequency of the recovered bipartition in relation to the two most frequently recovered bipartions (Salichos *et al.*, 2014; Kobert *et al.*, 2016).

The two trees inferred from the optimized subsamples of 100 loci were 48 (100_Full) and 20 (C100_Full) RF distance away from the 265_Full tree (Table 2). For context, the 265_2nd and 265_Reduced trees were 60 and 56 RFs different, respectively. C100_Full tree had the lowest distance to 265_Full tree (Table S5.1) and is in the 99.95th percentile of all the 23 740 comparisons to the baseline trees done in RADICAL (100_Full tree's distance was in the 88th percentile). The 100_Full and C100_Full trees also had slightly longer internal branch lengths compared to external branch lengths (i.e., decreased leafiness; Table 2). The mean bootstrap support and internode

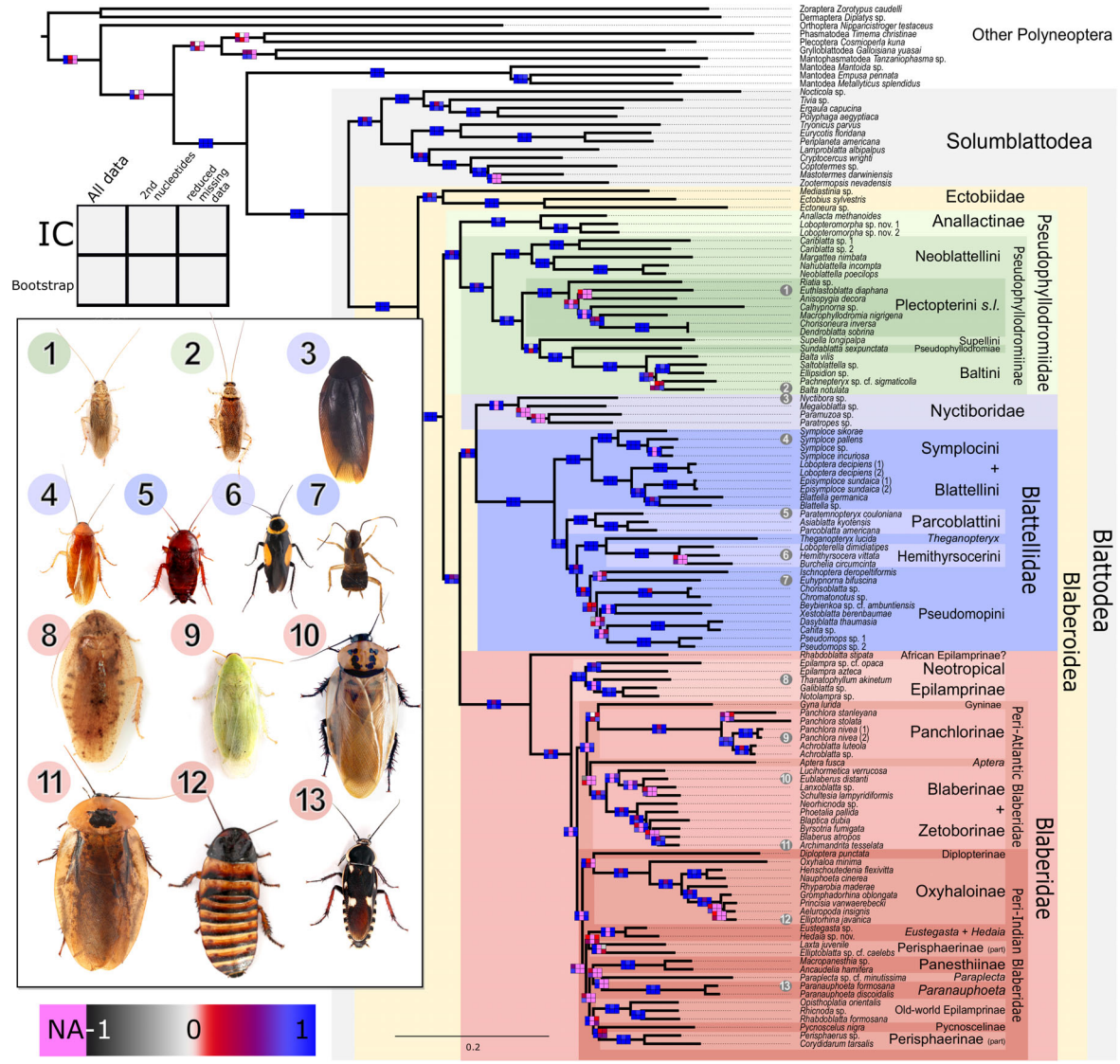

Fig. 2. Phylogeny of Blaberoidea (265_Full tree) as inferred from a partitioned RAxML tree inference from 265 loci. Support values in color-coded Navajo rugs are internode certainty and bootstrap frequency scores calculated from three trees, as described in the legend. 'NA'/pink indicates the bipartition does not appear in the specified tree. Branch lengths are proportional to substitutions. Numbers correspond to tip taxa depicted in photographs. Photographs by Dominic A. Evangelista. [Colour figure can be viewed at wileyonlinelibrary.com].

certainty of 100_Full and C100_Full were slightly lower than those in 265_Full (Table 2). The only trees deemed plausible given an alignment were 265_Full, 100_Full, C100_Full and C100_Reduced (Table 2). The C100_Full phylogeny was the only tree deemed plausible considering more than one alignment.

The C100_Full and 100_Full locus sets are composed of loci that demonstrated improved convergence (Figs 3, 4). Trees inferred from these sets display a lower ratio of total external branch length to internal branch length compared to the 265_Full phylogeny (Table 2). Among all the trees inferred, the C100_Full tree was estimated to be the most plausible given AU tests (Table 2). All the other trees were acceptable as plausible given their own alignment but not under any other alignment. C100_Full was the only tree accepted by more than one AU test (Table 2; also, see Tables S5.2, S5.3).

Discussion

Phylogeny of Blaberoidea

The relationships among the four clades of Blaberoidea (Fig. 2) largely conform to previous phylogenomic analysis (Evangelista *et al.*, 2019) but conflict with pre-genomic, morphological and combined-data phylogenetic studies. One of the

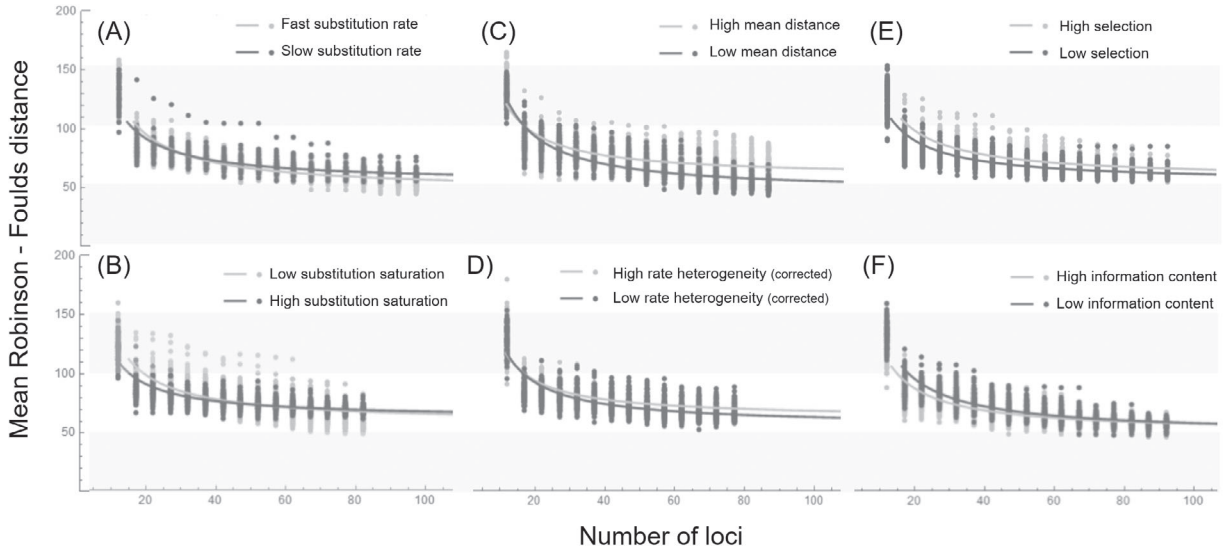

Fig. 3. RADICAL curves for six tests of phylogenetic utility. Dots represent the mean Robinson-Foulds distance to four baseline trees. Lines show best-fit exponential curves.

earliest cladistic analyses recovered ((Pseudophyllodromiinae + Blaberidae) + (Nyctiborinae + Ectobiinae + Blattellinae)) (Grandcolas, 1996; but see Klass, 2001). Evangelista *et al.* (2017) summarized topologies from phylogenetic studies published between 2003 and 2015 (Maekawa *et al.*, 2003; Klass & Meier, 2006; Inward *et al.*, 2007; Pellens *et al.*, 2007; Murienne, 2009; Djernæs *et al.*, 2012; Djernæs *et al.*, 2015; Legendre *et al.*, 2015) and underlined that there was little consensus about the backbone relationships of Blaberoidea. The most common topology was (Blattellinae + Nyctiborinae) + (Ectobiinae + (Pseudophyllodromiinae + Blaberidae)) but the poor support for these relationships with molecular data was highlighted in Evangelista *et al.* (2018). A recent mitogenomic study recovered ([Anallactinae + Ectobiinae] + Pseudophyllodromiinae) + (Nyctiborinae + [Blattellinae + Blaberidae]) but with <0.5 posterior probability (or < 50% bootstrap) on all but one node (Bourguignon *et al.*, 2018).

Our analysis included multiple terminal taxa that are important for clarifying some systematic issues. Robust placement of *Saltoblattella* within Baltini (Fig. 2) clarifies ambiguous prior affinities (Bohn *et al.*, 2010; Djernæs *et al.*, 2012, 2020). The sister-group relationship of *Lobopteromorpha* with *Anallacta* also helps clarify morphological synapomorphies for the Anallactinae *subfam. nov*. (Supplement S2). Our results support *Anisopygia* in Pseudophyllodromiinae (Djernæs *et al.*, 2020; see also Legendre *et al.*, 2015). Previous phylogenetic studies have recovered *Nahublattella* as outside of Pseudophyllodromiidae (Klass & Meier, 2006; Ware *et al.*, 2008; Djernæs *et al.*, 2020). The most recent analysis (Djernæs *et al.*, 2020) included two species of *Nahublattella* but had low genetic coverage and unstable placement. With coverage on 81 nuclear genomic loci (30% completeness) we recover *Nahublattella incompta* deep within Pseudophyllodromiinae with strong support, which agrees with its original morphology-based classification (Anisyutkin, 2009). *Calhypnorna* is listed as Blattellinae on the Cockroach Species File online database (Beccaloni, 2018) but this may be an error since Princis (1965) listed it in Anaplectidae and near *Plectoptera*. The latter agrees with our placement of *Calhypnorna* in Plectopterini *s.l.* Previous genetic barcode analyses indicated a close relationship of *Calhypnorna* with *Chorisoneura* (Evangelista *et al.*, 2014) also found here in Plectopterini *s.l.*. Support for *Mediastinia* and *Ectoneura* in Ectobiinae clarifies the position of the former taxa. *Mediastinia* was previously classified as Pseudophyllodromiinae, but only based on a superficial morphological description (Hebard, 1943). *Ectoneura* has already been recovered in Ectobiinae by Bourguignon *et al.* (2018) as expected from its morphology-based classification (Bohn *et al.*, 2010). The phylogenetic affinity of these two genera with *Ectobius* will assist in improved morphological revision of Ectobiinae (see Supplement S2).

The relationships among Blaberidae were highly volatile in our analyses, but some novel conclusions can be drawn (Fig. 2). Some recent studies have had strong taxon sampling of 'Peri-Atlantic Blaberidae' but had low support for the placement of most of the constituent lineages (Legendre *et al.*, 2017; Djernæs *et al.*, 2020). Contrary to these prior studies, we recover these taxa as monophyletic with strong support. Within the 'Peri-Atlantic Blaberidae' we find strong support for a monophyletic Panchlorinae. Panchlorinae is poorly defined, due to a simplification in morphology (Gurney & Roth, 1972). Before the present, no phylogenetic study had sampled African members of this predominantly neotropical genus. Dense sampling of Peri-Indian Blaberidae shows a greater extent of the rapid radiation that previous studies have only hinted at (Legendre *et al.*, 2017;Bourguignon *et al.*, 2018; Evangelista *et al.*, 2019). In agreement with Legendre *et al.* (2017),

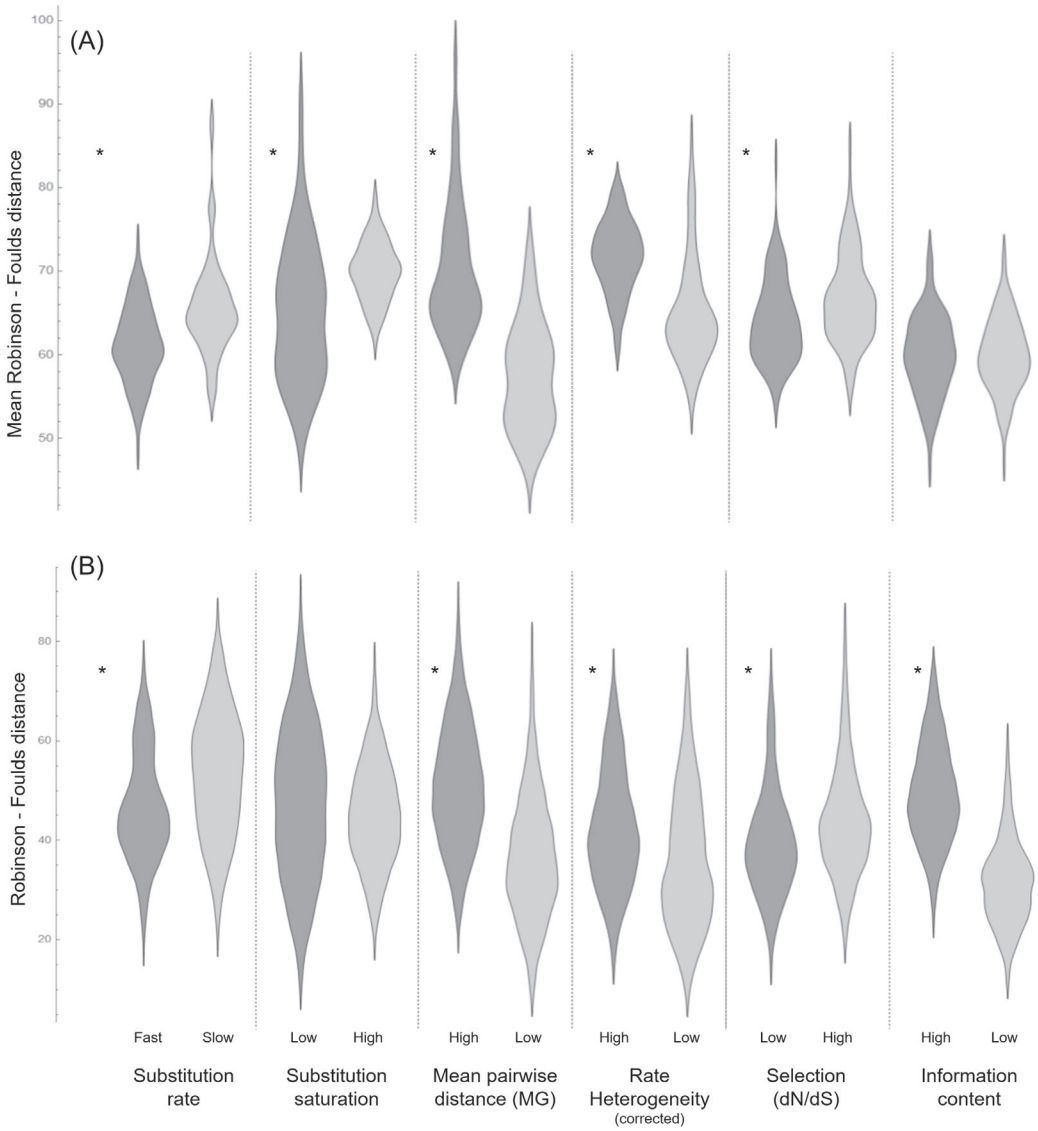

Fig. 4. Distribution plots showing (A) mean similarity (RF distance) to baseline trees and (B) similarity among tree inferences of two treatments for six factors. Each comparison is from the 14th concatenation step of each RADICAL run. Asterisks indicate a statistically significant difference, as determined by a Z-Test of mean comparisons (alpha <0.005).

we recover *Thanatophyllum akinetum* in Neotropical Epilamprinae, whereas it was previously considered to be a 'Zetoborinae' (Grandcolas, 1990; Djernæs *et al.*, 2020). The recovered topology of Blaberinae differs from that recovered by Djernæs *et al.* (2020), which also had a strongly supported topology. The position of the unusual genus *Aptera* is not strongly resolved, which was the case in other studies as well (Legendre *et al.*, 2015; Legendre *et al.*, 2017; Djernæs *et al.*, 2020). Novel approaches addressing the relationships among other lineages will be needed, as our inference gives little resolution.

There are also several systematic issues that can be clarified within Blattellidae + Nyctiboridae. The genus *Megaloblatta* has traditionally been recognized as Nyctiborinae but morphologically distinct from all others (Salazar & Malaver, 2012). Djernæs *et al.* (2020) listed this genus as *incertae sedis* due to the lack of evidence demonstrating relationship to other Nyctiboridae. Our best tree (Fig. 2) recovers *Megaloblatta* within Nyctiboridae with moderate support. The genera we recover in Parcoblattini (*Paratemnopteryx*, *Parcoblatta* and *Asiablatta*) were also recovered with strong support in Bourguignon *et al.* (2018). Earlier, Roth (1990) proposed this clade on morphological grounds, although he did not examine *Asiablatta*. *Theganopteryx* falls strongly in Blattellinae. This genus was previously thought to be in Ectobiinae but Bohn *et al.* (2010) discussed why this was not supported. While we recover *Chorisoblatta* as sister to *Chromatonotus*, we also doubt the accuracy of this relationship (see Supplement S3). *Beybienkoa* is recovered as sister to *Xestoblatta*. At various times, these genera have been thought

Table 1. Comparison of support for the 265_Full tree among RADICAL tree sets

Feature	Value	Relative tree certainty[a]	Relative-Tree certainty-All[b]	Tree certainty /alignment length	Tree certainty/# of loci
Substitution rate	Fast	0.581	0.599	5.65E−06	0.00599
	Slow	0.546	0.577	5.90E−06	0.00563
Substitution saturation	High	0.538	0.569	7.13E−06	0.00657
	Low	0.462	0.489	7.02E−06	0.00563
Pairwise sequence distance	High	0.491	0.523	5.83E−06	0.00564
	Low	0.562	0.567	7.75E−06	0.00646
Rate heterogeneity (corrected)	High	0.516	0.550	7.91E−06	0.00670
	Low	0.492	0.521	7.09E−06	0.00639
Stabilizing selection	Strong	0.540	0.570	6.15E−06	0.00587
	Weak	0.517	0.531	5.77E−06	0.00562
Aprior information content	High	0.578	0.598	6.89E−06	0.00628
	Low	0.538	0.547	6.45E−06	0.00584
Locus Length	Long	0.557	0.592	4.30E−06	0.00680
	Short	0.405	0.434	9.43E−06	0.00494

[a]Relative tree certainty (rTC) is the sum of all internode certainties (IC) normalized by the number of non-tip edges (Salichos *et al.*, 2014; Kobert *et al.*, 2016). IC scores are the relative frequency of the recovered bipartition in relation to the two most frequently recovered bipartions.
[b]Relative tree certainty-All is the same as rTC but calculated with IC-All. IC-All scores are the relative frequency of the recovered bipartition in relation to the all the other most frequently recovered bipartions (Salichos *et al.*, 2014; Kobert *et al.*, 2016).

Table 2. Comparison of tree quality and support among nine trees

Tree	Leafiness[a]	RF[b]	ln L[c]	Mean tree bootstrap support	Relative tree certainty	# of AU tests passed[d]
265_Full	0.171	0	−3 038 176.6	92.1	0.81	1
265_2nd	0.214	60	−433 293.3	73.7	0.51	0
265_Reduced	0.151	56	−793 673.2	87.4	0.75	0
100_Full	0.154	48	−919 296.0	87.9	0.72	1
100_2nd	0.179	98	−117 967.0	60.9	0.34	0
100_Reduced	0.151	66	−444 162.8	85.9	0.70	0
C100_Full	0.164	20	−1 207 800.6	89.7	0.74	2
C100_2nd	0.198	96	−159 517.2	65.4	0.41	0
C100_Reduced	0.157	44	−501 350.8	87.8	0.74	1

[a]The ratio of total tip branch lengths to internal branch lengths.
[b]Robinson-Foulds distance to the 265_Full tree.
[c]Log-likelihood of the tree given its alignment.
[d]how many times the tree was accepted ($P > 0.05$) by an AU test (maximum of three).

of as close relatives of *Ischnoptera* (Hebard, 1916; Legendre *et al.*, 2015; Bourguignon *et al.*, 2018), but this is the first time all three have been studied together. Further discussion of morphological support, and novelties in our phylogenetic inference are discussed in Supplement S3.

While strong support for Blattellidae + Nyctiboridae + Blaberidae has been demonstrated with both morphological and genomic data (Klass & Meier, 2006; Bourguignon *et al.*, 2018; Evangelista *et al.*, 2019) we have now extensively sampled taxa within the principal lineages of each clade. Given the unambiguous support for this relationship, we newly define it as Orkrasomeria *tax. nov*. (Supplement S2). Dense taxon sampling within Pseudophyllodromiidae and Orkrasomeria both imply a complex biogeographical history that should be given future study (but see Djernæs *et al.*, 2020). Recovery of an African 'Epilamprinae' as sister to the remaining Blaberidae (Fig. 2) implies that further sampling of these species are needed. The backbone topology of Blaberidae also suggests a complex biogeographical history (Djernæs *et al.*, 2020), likely due to dispersal given the age of this clade (Bourguignon *et al.*, 2018; Evangelista *et al.*, 2019).

Relationship between locus quality and phylogenetic support

We compared rate of phylogenetic convergence among loci of varying quality in a random concatenation test. In this context, *a priori* information content was not a predictor of convergence on the tree topology (Fig. 4A). Locus features that did result in improvements to convergence were high substitution rate, low saturation, low rate heterogeneity, strong stabilizing selection and low mean pairwise sequence distance (Fig. 4A).

We also compared support of the 265_Full topology with the RADICAL subsample trees from each of these respective

sets. In agreement with the test of convergence, loci with a high substitution rate and low mean pairwise sequence distance resulted in high tree certainty. However, there were conflicting results as well. Loci with *a priori* high information content demonstrated the highest tree certainty. High saturation, high rate heterogeneity both outperformed low saturation and low rate heterogeneity, respectively. We discuss these results below.

The lack of superior convergence of loci with *a priori* high information content is perhaps surprising but not unprecedented (Chen *et al.*, 2015). This may not be a failure of the methods (i.e., MARE and SAMS) to infer information content but rather an inappropriate application of the methods. For instance, the subset of taxa we applied these methods to have been insufficient since taxon sampling impacts inferred character evolution (Venditti *et al.*, 2006; Hugall & Lee, 2007). Additionally, MARE intentionally scores a locus as being uninformative if there is a high proportion of missing data, and missing data patterns change after new sequences are added. Thirdly, single genes may not contain enough characters to inform phylogenetic information estimators. Assignment of character state changes as autapomorphies, synapomorphies or symplesiomorphies depends on the number of available characters. Finally, the methods could be confounded by long branch attraction (LBA), mistaking saturated loci for highly tree-like loci and wrongly attributing homoplasy to split support. While we do not see exceedingly long branches, all the reconstructed trees do exhibit of features typical of long branch effects [e.g., abundant short internodes (Fig. 2 and Fig. S3.1) and a high leafiness (Table 2); Wägele & Mayer, 2007; Kück *et al.*, 2012).

Yet, the high support for the 265_Full tree among the trees inferred from the *a priori* high information content loci might suggest that there is more signal in these data. Filtering of loci with SAMS to determine information content resulted in discarding loci with strong signal for conflicting relationships. This relates directly to tree certainty, which is higher when the abundance of alternative topologies is low. We would note though, that the set of loci used for comparison (i.e., *a priori* non-high information content) also lent moderately high support for the 265_Full tree. Both analyses indicate that the relationships in the tree can be reliably inferred from either set of loci and thus true information content may be well distributed in either set of loci.

Fast evolving and low saturation loci converged rapidly towards the expected topology. This result contradicts the idea that saturation is a consequence of high evolutionary rate of a locus, two features that were not highly correlated ($R = -0.11$; Tables S1.1, S1.2) in our dataset. This seemingly contradictory finding results from how we calculated these values and how we framed our study on whole loci. First, we used tree-independent methods for approximating evolutionary rates and saturation, which were most appropriate for designing our experiment but do not provide the most sensitive estimates (see Supplement S1). Second, since we used the mean rate of the entire locus we are losing information about the distribution of rates within that locus. Evolutionary rate of an individual nucleotide position would correlate highly with true substitution saturation. However, when considering the mean rate of all positions in a locus, loci with a low or high mean rate can be equally saturated. For instance, consider two loci of equal length. All third codon positions may be saturated in both loci but in one the other codon positions may be informative and fast evolving while in the other they may be invariant. This suggests that a more informative value would be the distribution of site-specific rates for a given locus. This also contrasts with the assumption that slow evolutionary rates are assumed to be superior (Chen *et al.*, 2015). Although the superiority of slow rates are only supported when fast evolutionary rate results in saturation and the optimal rate changes depending upon the phylogenies' shape (Dornburg *et al.*, 2018). Unsaturated but fast evolving loci should have a higher density of phylogenetic information than slow loci.

Despite the overall low convergence of highly saturated loci, they lent relatively higher node support to the 265_Full tree. Convergence was calculated as the mean RF distance to four plausible species trees, and thus these tests are not directly analogous, but the comparison is interesting none-the-less. The wide variance in RF values among trees from low-saturation loci (Fig. 4A) suggests a diverse sample of trees, many of which are very different from the four baseline trees. Another potential reason for the high node support demonstrated by high saturation loci could be that some relationships in the 265_Full tree are falsely informed by highly saturated sites (i.e., long branch attraction).

Low rate-heterogeneity among loci also improved convergence in tree topology (Figs 3, 4). This is consistent with studies showing that rate heterogeneity confounds the assumptions of phylogenetic models (Galtier *et al.*, 2006; Frandsen *et al.*, 2015) and better adherence to those models improves phylogenetic inference (Kjer & Honeycutt, 2007; Doyle *et al.*, 2015; Reddy *et al.*, 2017). However, this effect is only seen if locus rate heterogeneity is corrected for locus length. Uncorrected high rate heterogeneity is not problematic (Fig. S4.4) since a higher variance in rates can be a result of loci being longer (Table S1.1). In reference to support of the 265_Full topology, while more heterogeneous loci resulted in higher tree certainty, the difference in certainty from low heterogeneity loci was negligible.

Strong stabilizing selection (Figs. 3, 4), as defined by low mean dN/dS, also improved tree convergence. This agrees with higher mean locus dN/dS corresponding to: more non-synonymous sites under positive selection, which are difficult to model (discussed in Beaulieu *et al.*, 2019); and purifying selection of synonymous substitution (Spielman & Wilke, 2015), which yield compositional bias and phylogenetic errors (Cox *et al.*, 2014).

Minimizing mean pairwise sequence distance among loci yielded the greatest improvement to tree convergence (Figs 3, 4) and increased node support (Table 1). Sequence distance is correlated to a suite of other features also related to phylogenetic utility (Bai *et al.*, 2013; Struck, 2014; Borowiec *et al.*, 2015; Chen *et al.*, 2015; Lewis *et al.*, 2016; Borowiec, 2019). In particular, decreasing mean pairwise sequence distance may be similar to minimizing saturated loci because they both reduce the probability of LBA (Struck, 2014). LBA can lead to rogue taxon placement, which our analysis is particularly sensitive to due to usage of the RF metric (Kuhner & Yamato, 2015).

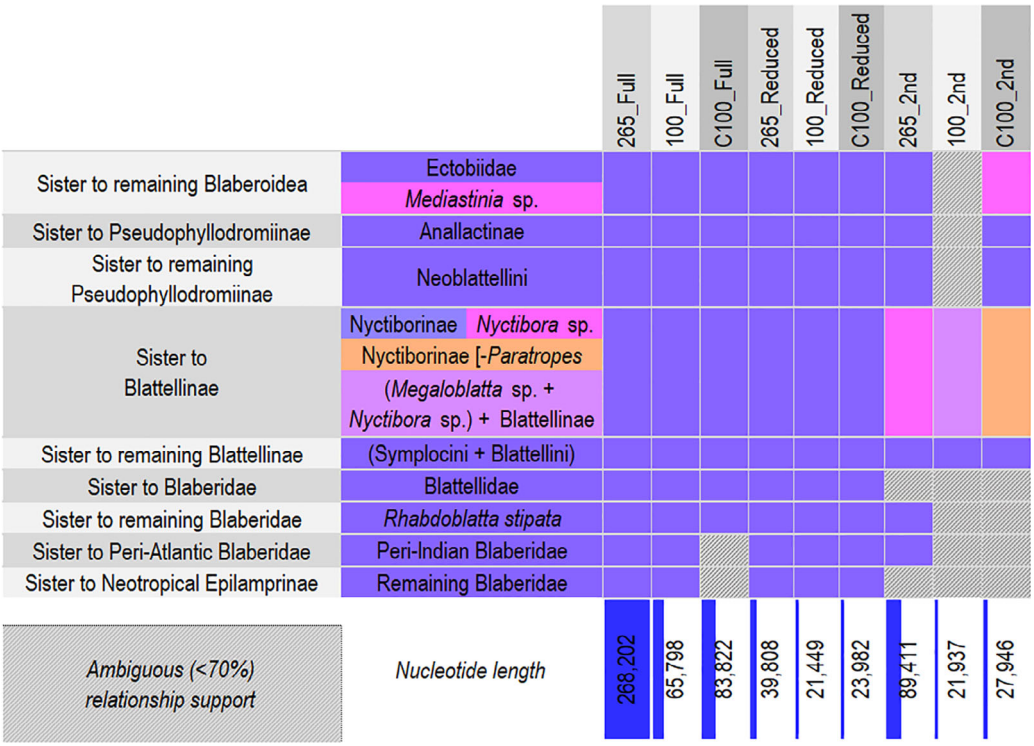

Fig. 5. Heat map of support for selected backbone relationships in Blaberoidea. Support for relationships are given for the nine trees indicated in column labels. Alternative proposed relationships are indicated by cell coloration. Only relationships with more than 70% support are considered unambiguous and coloured solidly [see Evangelista *et al.*, 2018 for the method of calculating support for relationships from bipartition support values]. Nucleotide length of each alignment is shown in the bottom row, and the cell is coloured proportional to the values. [Colour figure can be viewed at wileyonlinelibrary.com].

Our experimental design specifically attempts to control for extraneous features (see Figs. S1.1, S4.2, S4.4). We paid particularly close attention to locus length, which was moderately correlated with rate heterogeneity, mean pairwise sequence distance, and saturation (Table S1.2). Yet, these treatments with longer loci did not perform better (i.e., Fig. S4.4). Thus, more effectively controlling for locus length may yield a stronger effect-size.

Phylogenomic improvements

Our analyses demonstrate some clear relationships between features of genetic loci and the quality of downstream phylogenetic inference. These relationships could be used to improve future analyses. As a test of this, we inferred two additional phylogenetic trees, 100_Full and C100_Full, using collections of loci deemed to improve convergence on the species tree. Both of these trees demonstrate patterns consistent with higher phylogenetic signal (Table 2) (Dornburg *et al.*, 2018) and the C100_Full tree demonstrated the most plausibility given an alignment (Table 2). Regardless of if these represent true topological improvements, we have remarkably inferred two exceptionally similar trees (Table 2) with comparable node support after discarding a majority of the data (Fig. 5).

Reducing dataset size has practical implications for phylogenetic projects. Computation time for alignment estimation and topology testing (AU test) is reduced proportionally to the reduction in dataset size (Table 3). Gains in efficiency of phylogenetic inferences are 4–6 fold improvements (Table 3). These are significant in practice. For instance, the C100_Full tree only took 275 CPU hours to reach the optimization stop criteria, over a month faster than 265_Full (1025 CPU hours). Reducing heterogeneous loci provides the ability to simplify the evolutionary mode – avoiding over-parameterization (Sullivan & Joyce, 2005) and further improving computation time. We did not re-estimate optimal partition and model schemes for our dataset so we could consistently compare our results with the 265_Full phylogeny. Yet, doing so would be a qualitative improvement (e.g., Philippe & Roure, 2011).

Reducing computational effort of tree inference is important considering that phylogenetic studies rarely do such analyses only once. For instance, we conducted two preliminary tree inferences and two quality checking inferences, each of which involved analysing the 265_Full alignment. When adding together main tree searches, those of different alignment conformations and assessing support through resampling methods, computation time becomes a limiting resource even when having access to multiple world-class supercomputing clusters (e.g., Shen *et al.*, 2017 and this study). Improving the practicality of

Table 3. Comparison of various activities derived from three alignment conformations of 265 loci and two sets of 100 loci

Alignment	Alignment length	CPU time for best tree (Hrs)[a]	Bootstrap replicates to meet stop criteria	AU test CPU time (min)[b]	Time to align (min)[c]
265_Full	268 202	1025.4	60	96	132.5
265_2nd	89 411	175.6	300	-	NA
265_Reduced	39 808	109.8	108	-	NA
100_Full	65 798	166.0	156	37	40
100_2nd	21 937	27.2	408	-	NA
100_Reduced	21 449	37.6	108	-	NA
C100_Full	83 822	275.0	60	36	53
C100_2nd	27 946	38.4	360	-	NA
C100_Reduced	23 982	38.6	108	-	NA

[a] 100 starting trees in RAxML.
[b] 10 000 RELL bootstraps, unpartitioned analysis for nine trees.
[c] Extrapolated from alignment of ten unaligned FASTA files of varying lengths using MAFFT (options: - localpair - maxiterate 1000 - adjustdirection).

inferring trees and their support makes the application of downstream analyses (e.g., divergence date inference, diversification analyses, inferring trait evolution) more feasible as well. The important caveat, though, is one must have a large preliminary dataset from which to select the most optimal loci.

Conclusions

We inferred a robust phylogeny of Blaberoidea using 265 genomic loci and 126 taxa. This phylogeny provides congruence for previously hypothesized relationships and some novel classifications that can now be defined with advanced genomic and taxon sampling. When subsampling loci by quality, we find that the relationships in the tree are primarily supported by loci with high evolutionary rate, low saturation, low mean pairwise sequence distance, low rate heterogeneity, and strong stabilizing selection. Calculating *a priori* phylogenetic information content, as defined by split signal and tree-likeness, did not meaningfully provide additional support for relationships. These findings are consistent with past studies and phylogenetic theory (Townsend, 2007; Cox *et al.*, 2014; Doyle *et al.*, 2015; Dornburg *et al.*, 2018), and provide the opportunity to potentially target increasingly informative loci with locus capture methods (Lemmon *et al.*, 2012; Brandley *et al.*, 2015; Gilbert *et al.*, 2015). Subsamples reduced by two thirds of total data length and optimized under the above specifications exhibited recovery of a phylogeny exceptionally similar to the tree inferred from all the data. Thus, targeting maximally phylogenetically useful loci can potentially reduce the monetary cost and computational resources of projects drastically while maintaining the quality of the results.

Author contributions

DAE, FL and AYK obtained funding for the study. DAE and FL conceived the study and organized the taxon sampling design. DAE ran preliminary analyses and designed genomic sampling with assistance from SSi and BW. DAE, MMW and JLW designed the molecular baits. DAE extracted and enriched genomic DNA, with MMW, JLW and MKK providing guidance. DAE did all bioinformatics, wrote custom software, executed all analyses, with guidance from FL. DAE, OB and FL revised taxonomic descriptions of taxa with assistance from BW. DAE and FL wrote the paper with assistance from SSi, AYK and BW. DAE and BW composed the figures. SSi, BW, JLW, MKK, and OB, provided early access to transcriptome datasets and contributed in analyses and curation of those datasets as well as feedback to the manuscript.

Supporting Information

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

Supplementary S1: S1-S5. Supplementary results and discussion. Supplemental methods (S1), Taxonomic changes (S2), supplemental discussion of the Blaberoidea phylogeny (S3), additional tests of locus quality (S4), and evaluation of data reduction for phylogenetics (S5).

Supplementary data and results files: Data include alignments, PartitionFinder files, preliminary taxonomic hypotheses, loci features, and taxon list. Results include RADICAL output files and phylogenetic trees.

Acknowledgements

Thanks to the 1KITE consortium who supported this research with preliminary data and advice with software. Specifically, appreciation extends to Karen Meusemann, Alexander Donath, Bernhard Misof, Xin Zhou, Shanlin Liu, Ralph S. Peters, Lars Podsiadlowski, Ward Tollenaar, Mari Fujita, and Ryuichiro Machida. Huge thanks to all breeders (Nicolas Rousseaux, Tristan Shanahan, T.J. Ombrelle and Piotr Sterna), colleagues (Mike Picker), museums (MNHN, MFN, NHMUK and CAS) and curators (Jurgen Deckert, and George Beccaloni) who assisted in providing specimens. Great appreciation to New

England Biolabs, MycroArray (now Arbor Biosciences), Sara Ruane, Ciara-Mae Mendoza, Melissa Sanchez-Herrera, Steven Ramirez and Mihaela Glamoclija for providing assistance in the lab. Additional thanks to Brian O'Meara for guidance and advice. Great appreciation to the reviewers whose input helped us improve the manuscript greatly. This research could not have been completed without the support of NSF (award # 1608559), all other funding agencies, the MNHN - Paris, Rutgers University and the University of Tennessee - Knoxville. This work was support by the National Science Foundation (award number 1608559) to DAE, FL and AK. The authors declare no conflicting interests.

Data Availability

Transcriptome datasets were taken from Evangelista *et al.* (2019) and newly sequenced data are deposited on the NCBI Sequence Read Archive (SRP155429). Other analysed files are available on the Dryad digital repository (doi: https://doi.org/10.5061/dryad.9mf1pr7).

References

Anisyutkin, L.N. (2009) New representatives of the genus Nahublattella Bruijning, 1959 (Dictyoptera, Blattellidae) from central and South America. *Entomological Review*, **89**, 820–838.

Bai, F., Xu, J. & Liu, L. (2013) Weighted relative entropy for phylogenetic tree based on 2-step Markov model. *Mathematical Biosciences*, **246**, 8–13.

Beaulieu, J.M., O'Meara, B.C., Zaretzki, R., Landerer, C., Chai, J. & Gilchrist, M.A. (2019) Population genetics-based phylogenetics under stabilizing selection for an optimal amino acid sequence: a nested modelling approach. *Molecular Biology and Evolution*, **36**, 834–851.

Beccaloni, G. (2018) Cockroach Species File Online. Version 5.0/5.0, World Wide Web electronic publication.

Beccaloni, G. & Eggleton, P. (2013) Order: Blattodea. *Zootaxa*, **3703**, 46.

Blaxter, M.L. (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **359**, 669–679.

Bohn, H., Picker, M., Klass, K.D. & Colville, J. (2010) A jumping cockroach from South Africa, *Saltoblattella montistabularis*, gen. nov., spec. nov. (Blattodea: Blattellidae). *Arthropod Systematics & Phylogeny*, **68**, 53–69.

Borowiec, M.L. (2019) Convergent evolution of the army ant syndrome and congruence in big-data phylogenetics. *Systematic Biology*, **68**, 643–656.

Borowiec, M.L., Lee, E.K., Chiu, J.C. & Plachetzki, D.C. (2015) Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics*, **16**, 987.

Bourguignon, T., Tang, Q., Ho, S.Y.W. *et al.* (2018) Transoceanic dispersal and plate tectonics shaped global cockroach distributions: evidence from mitochondrial phylogenomics. *Molecular Biology and Evolution*, **35**, 1–14.

Brandley, M.C., Bragg, J.G., Singhal, S. *et al.* (2015) Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian Eugongylus group scincid lizards. *BMC Evolutionary Biology*, **15**, 62.

Bravo, G.A., Antonelli, A., Bacon, C.D. *et al.* (2019) Embracing heterogeneity: building the tree of life and the future of phylogenomics. *PeerJ*, **7**, e6399.

Breinholt, J.W. & Kawahara, A.Y. (2013) Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. *Genome Biology and Evolution*, **5**, 2082–2092.

Brown, J.M. & Thomson, R.C. (2017) Bayes factors unmask highly variable information content, bias, and extreme influence in Phylogenomic analyses. *Systematic Biology*, **66**, 517–530.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chen, M.-Y., Liang, D. & Zhang, P. (2015) Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. *Systematic Biology*, **64**, 1104–1120.

Cox, C.J., Li, B., Foster, P.G., Embley, T.M. & Civan, P. (2014) Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology*, **63**, 272–279.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U. *et al.* (2014) Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Molecular Biology and Evolution*, **31**, 239–249.

Djernæs, M., Klass, K.-D., Picker, M.D. & Damgaard, J. (2012) Phylogeny of cockroaches (Insecta, Dictyoptera, Blattodea), with placement of aberrant taxa and exploration of out-group sampling. *Systematic Entomology*, **37**, 65–83.

Djernæs, M., Klass, K.D. & Eggleton, P. (2015) Identifying possible sister groups of Cryptocercidae+Isoptera: a combined molecular and morphological phylogeny of Dictyoptera. *Molecular Phylogenetics and Evolution*, **84**, 284–303.

Djernæs, M., Varadínová, Z.K., Eulitz, U. & Klass, K.-D. (2020) Phylogeny and life history evolution of Blaberoidea (Blattodea). *Arthropod Systematics & Phylogeny*, **78**, 29–67.

Dornburg, A., Fisk, J.N., Tamagnan, J. & Townsend, J.P. (2016) PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. *BMC Evolutionary Biology*, **16**, 262.

Dornburg, A., Su, Z. & Townsend, J.P. (2018) Optimal rates for phylogenetic inference and experimental design in the era of genome-scale datasets. *Systematic Biology*, **68**, 145–156.

Doyle, V.P., Young, R.E., Naylor, G.J.P. & Brown, J.M. (2015) Can we identify genes with increased phylogenetic reliability? *Systematic Biology*, **64**, 824–837.

Edwards, S.V. (2016) Phylogenomic subsampling: a brief review. *Zoologica Scripta*, **45**, 63–74.

Edwards, S.V., Cloutier, A. & Baker, A.J. (2017) Conserved Nonexonic elements: a novel class of marker for Phylogenomics. *Systematic Biology*, **66**, 1028–1044.

Evangelista, D.A. (2019) Phyloinformatica, version 0.9, GitHub.

Evangelista, D.A., Bourne, G. & Ware, J.L. (2014) Species richness estimates of Blattodea s.s. (Insecta: Dictyoptera) from northern Guyana vary depending upon methods of species delimitation. *Systematic Entomology*, **39**, 150–158.

Evangelista, D.A., Djernæs, M. & Kohli, M.K. (2017) Fossil calibrations for the cockroach phylogeny (Insecta, Dictyoptera, Blattodea), comments on the use of wings for their identification, and a redescription of the oldest Blaberidae. *Palaeontologia Electronica*, **20**, 1–23.

Evangelista, D., Thouzé, F., Kohli, M.K., Lopez, P. & Legendre, F. (2018) Topological support and data quality can only be assessed through multiple tests in reviewing Blattodea phylogeny. *Molecular Phylogenetics and Evolution*, **128**, 112–122.

Evangelista, D.A., Wipfler, B., Béthoux, O. *et al.* (2019) An integrative phylogenomic approach illuminates the evolutionary history of cockroaches and termites (Blattodea). *Proceedings of the Royal Society B: Biological Sciences*, **286**, 1–9.

Fong, J.J., Brown, J.M., Fujita, M.K. & Boussau, B. (2012) A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic Lissamphibia. *PLoS One*, **7**, 1–14.

Frandsen, P.B., Calcott, B., Mayer, C. & Lanfear, R. (2015) Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evolutionary Biology*, **15**, 13.

Galtier, N., Enard, D., Radondy, Y., Bazin, E. & Belkhir, K. (2006) Mutation hot spots in mammalian mitochondrial DNA. *Genome Research*, **16**, 215–222.

Ghasemi, A. & Zahediasl, S. (2012) Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, **10**, 486–489.

Gilbert, P.S., Chang, J., Pan, C., Sobel, E.M., Sinsheimer, J.S., Faircloth, B.C. & Alfaro, M.E. (2015) Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. *Molecular Phylogenetics and Evolution*, **92**, 40–146.

Grabherr, M., BJ, H., Yassour, M. *et al.* (2011) Trinity reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, **29**, 644–652.

Grandcolas, P. (1990) Descriptions de nouvelles Zetoborinae guyanaises avec quelques remarques sur la sous-famille. *Bulleting of the Entomological Society of France*, **95**, 241–246.

Grandcolas, P. (1996) The phylogeny of cockroach families: a cladistic appraisal of morpho-anatomical data. *Canadian Journal of Zoology*, **74**, 508–527.

Gurney, A.B. & Roth, L.M. (1972) A generic review of the cockroaches of the subfamily Panchlorinae (Dictyoptera, Blattaria, Blaberidae). *Annals of the Entomological Society of America*, **65**, 521–532.

Haas, B., Papanicolaou, A., Yassour, M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-Seq reference generation and analysis with trinity. *Nature Protocols*, **8**, 1494–1512.

Hebard, M. (1916) Studies in the group Ischnopterites (Orthoptera, Blattidae, Pseudomopinae). *Transactions of the American Entomological Society*, **42**, 337–383.

Hebard, M. (1943) Australian Blattidae of the subfamilies Chorisoneurinae and Ectobiinae (Orthoptera). *The Academy of Natural Sciences of Philadelphia*, **14**, 1–129.

Hugall, A.F. & Lee, M.S.Y. (2007) The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution; International Journal of Organic Evolution*, **61**, 2293–2307.

Inward, D., Beccaloni, G. & Eggleton, P. (2007) Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. *Biology Letters*, **3**, 331–335.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Haeseler, A.v. & Jermiin, a.L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kjer, K.M. & Honeycutt, R.L. (2007) Site-specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evolutionary Biology*, **7**, 8.

Klass, K.-D. (2001) Morphological evidence on Blattarian phylogeny: "phylogenetic histories and stories" (Insecta, Dictyoptera). *Berliner Entomologische Zeitschrift*, **48**, 223–265.

Klass, K.-D. & Meier, R. (2006) A phylogenetic analysis of Dictyoptera (Insecta) based on morphological characters. *Entomologische Abhandlungen*, **63**, 3–50.

Klopfstein, S., Massingham, T. & Goldman, N. (2017) More on the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, **66**, 769–785.

Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. (2016) Computing the internode certainty and related measures from partial gene trees. *Molecular Biology and Evolution*, **33**, 1606–1617.

Krueger, F. (2017) TrimGalore v. 0.4.5, Babraham Bioinformatics.

Kück, P., Mayer, C., Wägele, J.W. & Misof, B. (2012) Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS One*, **7**, e36593.

Kuhner, M.K. & Yamato, J. (2015) Practical performance of tree comparison metrics. *Systematic Biology*, **64**, 205–214.

Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T. & Calcott, B. (2016) PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, **34**, 772–773.

Larsson, A. (2014) AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics*, **30**, 3276–3278.

Legendre, F., Nel, A., Svenson, G.J., Robillard, T., Pellens, R. & Grandcolas, P. (2015) Phylogeny of Dictyoptera: dating the origin of cockroaches, praying mantises and termites with molecular data and controlled fossil evidence. *PLoS One*, **10**, e0130127.

Legendre, F., Grandcolas, P. & Thouzé, F. (2017) Molecular phylogeny of Blaberidae (Dictyoptera, Blattodea) with implications for taxonomy and evolutionary studies. *European Journal of Taxonomy*, **291**, 1–13. http://dx.doi.org/10.5852/ejt.2017.291.

Lemmon, A.R., Emme, S.A. & Lemmon, E.M. (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Lewis, P.O., Chen, M.H., Kuo, L. *et al.* (2016) Estimating Bayesian phylogenetic information content. *Systematic Biology*, **65**, 1009–1023.

Maekawa, K., Lo, N., Rose, H.A. & Matsumoto, T. (2003) The evolution of soil-burrowing cockroaches (Blattaria: Blaberidae) from wood-burrowing ancestors following an invasion of the latter from Asia into Australia. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**, 1301–1307.

Mayer, C., Sann, M., Donath, A. *et al.* (2016) BaitFisher: a software package for multi-species target DNA enrichment probe design. *Molecular Biology and Evolution*, **33**, 1875–1886.

Miller, M.A., Pfeiffer, W. & Schwartz, T. (2010) Creating the CIPRES science gateway for inference of large phylogenetic trees.

Misof, B., Liu, S., Meusemann, K. *et al.* (2014a) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.

Misof, B., Meusemann, K., Reumont, B.M.v., Kück, P., Prohaska, S.J. & Stadler, P.F. (2014b) A priori assessment of data quality in molecular phylogenetics. *Algorithms for Molecular Biology*, **9**, 1–8.

Molloy, E.K. & Warnow, T. (2018) To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, **67**, 285–303.

Murienne, J. (2009) Molecular data confirm family status for the *Tryonicus–Lauraesilpha* group (Insecta: Blattodea: Tryonicidae). *Organisms Diversity & Evolution*, **9**, 44–51.

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

Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.

Pellens, R., D'Haese, C.A., Belles, X., Piulachs, M.D., Legendre, F., Wheeler, W.C. & Grandcolas, P. (2007) The evolutionary transition

from subsocial to eusocial behaviour in Dictyoptera: phylogenetic evidence for modification of the "shift-in-dependent-care" hypothesis with a new subsocial cockroach. *Molecular Phylogenetics and Evolution*, **43**, 616–626.

Petersen, M., Meusemann, K., Donath, A. *et al.* (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, **18**, 111.

Philippe, H. & Roure, B. (2011) Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biology*, **9**, 91.

Platt, R.N. 2nd, Faircloth, B.C., Sullivan, K.A.M. *et al.* (2018) Conflicting evolutionary histories of the mitochondrial and nuclear genomes in New World Myotis bats. *Systematic Biology*, **67**, 236–249.

Princis, K. (1965) Blattariae: Subordo Blaberoidea: Fam. Oxyhaloidae, Panesthiidae, Cryptocercidae, Chorisoneuridae, Oulopterygidae, Diplopteridae, Anaplectidae, Archiblattidae, Nothoblattidae: Orthopterorum Catalogs, v. Pars 7. 's-Gravenhage, W. Junk.

Reddy, S., Kimball, R.T., Pandey, A. *et al.* (2017) Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, **66**, 857–879.

Robinson, D.F. & Foulds, L.R. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–141.

Roth, L.M. (1990) A revision of the Australian Parcoblattini (Blattaria:Blattellidae:Blattellinae). *Memoirs of the Queensland Museum*, **28**, 531–596.

Salazar, J.A. & Malaver, C.R. (2012) Relation and illustration of some Nyctiborinae species from Colombia and Costa Rica (Insecta: Blattodea, Ectobiidae). *Boletín Científico Centro De Museos Museo De Historia Natural*, **16**, 185–197.

Salichos, L., Stamatakis, A. & Rokas, A. (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution*, **31**, 1261–1271.

Shen, X.X., Hittinger, C.T. & Rokas, A. (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution*, **1**, 126.

Simmons, M.P., Sloan, D.B., Springer, M.S. & Gatesy, J. (2018) Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses. *Molecular Phylogenetics and Evolution*, **131**, 80–92.

Simon, S., Narechania, A., Desalle, R. & Hadrys, H. (2012) Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. *Genome Biology and Evolution*, **4**, 1295–1309.

Simon, S., Blanke, A. & Meusemann, K. (2018) Reanalyzing the Palaeoptera problem – the origin of insect flight remains obscure. *Arthropod Structure & Development*, **47**, 328–338.

Soltis, P.S. & Soltis, D.E. (2003) Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, **18**, 256–267.

Spielman, S.J. & Wilke, C.O. (2015) The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, **32**, 1097–1108.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Steel, M. & Leuenberger, C. (2017) The optimal rate for resolving a near-polytomy in a phylogeny. *Journal of Theoretical Biology*, **420**, 174–179.

Struck, T.H. (2014) TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics Online*, **10**, 51–67.

Sullivan, J. & Joyce, P. (2005) Model selection in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 445–466.

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M. & Dessimoz, C. (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, **64**, 778–779.

Townsend, J.P. (2007) Profiling phylogenetic informativeness. *Systematic Biology*, **56**, 222–231.

Venditti, C., Meade, A. & Pagel, M. (2006) Detecting the node-density artifact in phylogeny reconstruction. *Systematic Biology*, **55**, 637–643.

Wägele, J.W. & Mayer, C. (2007) Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evolutionary Biology*, **7**, 1–24.

Ware, J.L., Litman, J., Klass, K.-D. & Spearman, L.A. (2008) Relationships among the major lineages of Dictyoptera: the effect of outgroup selection on Dictyopteran tree topology. *Systematic Entomology*, **33**, 429–450.

Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M. & Kriventseva, E.V. (2013) OrthoDB: a hierarchical catalogue of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, **41**, D358–D365.

Wipfler, B., Letsch, H., Frandsen, P.B. *et al.* (2019) Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects. *Proceedings of the National Academy of Sciences of the United States of America*, **116**, 3024–3029.

Xi, Z., Liu, L. & Davis, C.C. (2016) The impact of missing data on species tree estimation. *Molecular Biology and Evolution*, **33**, 838–860.

Accepted 20 August 2020

