Molecular Phylogenetics and Evolution 130 (2019) 297–303

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Insight from an ultraconserved element bait set designed for hemipteran phylogenetics integrated with genomic resources

Troy J. Kieran[a,*], Eric R.L. Gordon[b,c], Michael Forthman[d], Rochelle Hoey-Chamberlain[b], Rebecca T. Kimball[e], Brant C. Faircloth[f], Christiane Weirauch[b], Travis C. Glenn[a]

[a] Department of Environmental Health Science, College of Public Health, University of Georgia, Athens, GA, USA
[b] Department of Entomology, University of California, Riverside, CA, USA
[c] Department Ecology and Evolutionary Biology, University of Connecticut, Hartford, CT, USA
[d] Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA
[e] Department of Biology, University of Florida, Gainesville, FL, USA
[f] Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA, USA

ARTICLE INFO

*Keywords:*
Hemiptera
Heteroptera
Phylogenetics
Phylogenomics
Sequence capture
Target enrichment

ABSTRACT

Target enrichment of conserved genomic regions facilitates collecting sequences of many orthologous loci from non-model organisms to address phylogenetic, phylogeographic, population genetic, and molecular evolution questions. Bait sets for sequence capture can simultaneously target thousands of loci, which opens new avenues of research on speciose groups. Current phylogenetic hypotheses on the > 103,000 species of Hemiptera have failed to unambiguously resolve major nodes, suggesting that alternative datasets and more thorough taxon sampling may be required to resolve relationships. We use a recently designed ultraconserved element (UCE) bait set for Hemiptera, with a focus on the suborder Heteroptera, or the true bugs, to test previously proposed relationships. We present newly generated UCE data for 36 samples representing three suborders, all seven heteropteran infraorders, 23 families, and 34 genera of Hemiptera and one thysanopteran outgroup. To improve taxon sampling, we also mined additional UCE loci *in silico* from published hemipteran genomic and transcriptomic data. We obtained 2271 UCE loci for newly sequenced hemipteran taxa, ranging from 265 to 1696 (average 904) per sample. These were similar in number to the data mined from transcriptomes and genomes, but with fewer loci overall. The amount of missing data correlates with greater phylogenetic divergence from taxa used to design the baits. This bait set hybridizes to a wide range of hemipteran taxa and specimens of varying quality, including dried specimens as old as 1973. Our estimated phylogeny yielded topologies consistent with other studies for most nodes and was strongly-supported. We also demonstrate that UCE loci are almost exclusively from the transcribed portion of the genome, thus data can be successfully integrated with existing genomic and transcriptomic resources for more comprehensive phylogenetic sampling, an important feature in the era of phylogenomics. UCE approaches can be used by other researchers for additional studies on hemipteran evolution and other research that requires well resolved phylogenies.

## 1. Introduction

Hemiptera is the largest order of non-holometabolous insects with > 103,000 described species. They are characterized by distinctive piercing-sucking mouthparts, which allow exploitation of plant vascular tissue in many species, including some economically important agricultural pests. While some hemipteran species can be beneficial predators, one group includes vectors of human diseases, and others are nuisance pests. Hemiptera are thought to have diverged from their sister taxon Thysanoptera, the thrips, more than 300 mya (Misof et al., 2014). Although ancestral herbivorous feeding habits have been retained in several hemipteran lineages including aphids, whiteflies and relatives (Sternorrhyncha), cicadas and relatives (Auchenorrhyncha), and moss bugs (Coleorrhyncha), life history strategies diversified in a fourth lineage, the suborder Heteroptera, to include predacious, hematophagous, mycetophagous and mixed-feeding habits (Weirauch et al., 2018). Despite the diversity and economic importance of Hemiptera, phylogenetic relationships among and within major lineages, i.e.,

* Corresponding author at: Department of Environmental Health Science, University of Georgia, 206 Environmental Health Science Building, Athens, GA 30602, United States.
*E-mail address:* kierant@uga.edu (T.J. Kieran).

https://doi.org/10.1016/j.ympev.2018.10.026
Received 16 May 2018; Received in revised form 12 October 2018; Accepted 20 October 2018
Available online 22 October 2018

the suborders Sternorrhyncha, Auchenorrhyncha, Coleorrhyncha, and Heteroptera, have remained contentious (Cryan and Urban, 2012; Li et al., 2015; Song et al., 2016).

Based on increasingly extensive datasets with respect to taxon sampling and/or characters (few loci to complete mitochondrial genomes and transcriptomes), these analyses have converged on congruent topologies in certain parts of the tree (i.e., establishment of Auchenorrhyncha as more closely related to Heteroptera than to Sternorrhyncha). However, other major questions have remained unsolved, such as relationships among the early diverging lineages within Heteroptera (Li et al., 2012; Wang et al., 2016; Wang et al., 2017; Weirauch et al., 2018; Wheeler et al., 1993). Generating and analyzing comprehensive datasets with respect to both taxonomic and character sampling by using a large number of universal markers from throughout the genome has the potential to greatly advance our understanding of phylogenetic relationships across Hemiptera. Approaches using anchored hybrid enrichment (Lemmon et al., 2012) or ultraconserved elements (UCEs) (Faircloth et al., 2012) generate such data for relatively low costs, making them feasible for taxon-rich phylogenetic analyses.

UCEs are highly conserved across divergent taxa (Bejerano et al., 2004) which make them useful as anchors for target enrichment. They have been shown to be useful markers for comparison across diverse taxa in vertebrates (Alexander et al., 2017; Crawford et al., 2012; Crawford et al., 2015; Faircloth et al., 2012; Gilbert et al., 2015; McCormack et al., 2013; Moyle et al., 2016; Smith et al., 2014) and more recently in arthropods (Baca et al., 2017; Faircloth et al., 2015; Starrett et al., 2017; Van Dam et al., 2017). Sequence variability increases with distance from the conserved UCE (Faircloth et al., 2012), which allows for analyses at different phylogenetic scales, from deep divergence (Faircloth et al., 2013) to population level (Harvey et al., 2016; Manthey et al., 2016). While most UCE studies in arthropods have focused on the Hymenoptera (Blaimer et al., 2015; Blaimer et al., 2016; Bossert et al., 2017; Branstetter et al., 2017a; Branstetter et al., 2017b; Faircloth et al., 2015), bait sets designed for other groups have recently been used in empirical studies (Baca et al., 2017; Starrett et al., 2017; Van Dam et al., 2017), which show the promising utility and effectiveness of UCEs in non-vertebrates. Having a universal set of genetic markers for a diverse group like Hemiptera can help standardize the phylogenetic data available and make comparative studies across multiple projects, questions, and scales easier for researchers. Faircloth (2017) recently designed and *in silico* tested UCE bait sets for several arthropod orders including Hemiptera. Only two of the designed bait sets have been used to generate UCE loci from samples and evaluated [Arachnida; (Starrett et al., 2017) and Coleoptera; (Baca et al., 2017; Van Dam et al., 2017)]. Here, we expand such testing to the Hemiptera UCE bait set.

## 2. Methods

### 2.1. Bait and taxon sampling

Hemiptera UCE capture baits (Faircloth, 2017) consisting of 40,207 baits for 2731 loci were tested on 36 hemipteran samples representing three suborders and the seven heteropteran infraorders (Table 1, Supplementary Table S1). Taxa were chosen to include a mix of species closely related to the taxa used for bait design and more distantly related taxa to assess UCE efficacy across Hemiptera. We also included multiple individuals within two families, Coreidae (n = 9) and Reduviidae: Triatominae (n = 6), to assess the utility of the bait set for recovering shallower phylogenetic relationships.

### 2.2. DNA extraction and library preparation

For most specimens, genomic DNA was extracted from ethanol preserved and recently pinned specimens using a Qiagen DNeasy kit, and older pinned specimens using a Qiagen QIAquick PCR Clean Up kit. Specimens of Coreidae were extracted using a Puregene Solid Tissue kit (Supplementary Table S1). DNA concentration and quality were assessed on a Qubit, fragment analyzer, and a 1.5% agarose gel. Samples with higher molecular weight were fragmented on a Bioruptor UCD-300 sonication device (Diagenode) based on quality for 2–9 cycles of 30 s on/30 s off. Resulting fragments were in the range of 200–1000 bp.

Libraries were prepared with a KAPA Hyper Prep Kit (Kapa Biosystems) following manufacturer's protocol with a few modifications. Half volume reactions were performed on all samples. Universal TruSeq compatible adaptor stubs were ligated onto A-tailed DNA fragments. Adapter-ligated product was amplified using Illumina TruSeq compatible dual-indexed primers with modified 8 bp indexes (Glenn et al., 2016). PCR reactions were 25 μL consisting of 10 μL of adapter-ligated DNA, 12.5 μL 2X KAPA HiFi HotStart ReadyMix, and 2.5 μL each of the 5 μM dual-indexed primers. Thermocycler conditions were 98 °C for 45 s, followed by 14 cycles of 98 °C for 30 s, 60 °C for 30 s, and 72 °C for 30 s, and then a final extension of 72 °C for 1 min. All clean-up steps used Sera-Mag magnetic beads (Thermo-Scientific, Waltham, MA, USA). Post-PCR cleaned product was quantified on Qubit and equimolar amounts of 9–12 samples were combined into 500 ng pools.

### 2.3. UCE enrichment and sequencing

Enrichments of library pools were performed using the MYbaits kit (MYcroarray, now Arbor Biosciences) following the manufacturer's protocol v3.01. Hybridizations were performed at 65 °C for 24 h. After hybridization, library pools were bound to Dynabeads M-280 Streptavidin magnetic beads (Life Technologies) for enrichment. Post-hybridization enrichments were amplified in a 25 μL volume reaction consisting of 10 μL enriched DNA, 12.5 μL 2X KAPA HiFi HotStart ReadyMix, and 2.5 μL each of 5 μM of Illumina P5/P7 primers. Amplification conditions were 98 °C for 45 s, followed by 16 cycles of 98 °C for 20 s, 60 °C for 30 s, and 72 °C for 60 s, and then a final extension of 72 °C for five minutes. Enriched and amplified library pools were quantified on Qubit and pooled in equimolar ratios. Libraries were sequenced using paired-end 150 bp reads on an Illumina HiSeq 3000 (Oklahoma Medical Research Foundation).

### 2.4. Data processing and analysis

For clarification, we make use of the following terms to distinguish between the different data sets analyzed as part of this study: (1) empirical – newly generated UCE data for this study, (2) *in silico* – UCE data retrieved from Faircloth (2017), (3) transcriptome – UCE data retrieved from publicly available transcriptomes and protein-encoding portions of genomes.

Raw sequencing data were processed using PHYLUCE v1.5.0 (Faircloth, 2016) with associated software as incorporated in the pipeline. We used default values unless otherwise noted. Adaptors and low-quality bases were removed using Illumiprocessor (https://github.com/faircloth-lab/illumiprocessor). Reads were assembled using Trinity v2.0.6 (Grabherr et al., 2011). We aligned UCE loci with MAFFT (Katoh and Standley, 2013), changing the max divergence from 20% (for empirical + *in silico* dataset) to 40% (empirical dataset), and trimmed with GBLOCKS (Castresana, 2000; Talavera and Castresana, 2007). Data matrices that were 50% and 60% complete (i.e., single locus alignments contain at least this percentage of total taxa), were used for further maximum likelihood (ML) phylogenetic analysis using RAxML v8.1.20 (Stamatakis, 2014). To complement our taxon sampling and assess the capacity to integrate our data with existing genomic data, we reassembled transcriptomic data from nine paraneopteran taxa from Misof et al. (2014) (https://doi.org/10.5061/dryad.3c0f1) using Trinity v2.0.6 (Grabherr et al., 2011). We also downloaded the protein-coding sequences of the genome of *Bemisia tabaci* from GenBank (Xie

**Table 1**
Summary results of each sample in the empirical data set. Suborder names are abbreviated to the first three letters. Infraorder names are abbreviated by removing -morpha.

| Suborder | Infraorder | Family | Genus | Species | Reads Passed QC | | Contigs | UCE Loci | On-Target | Missing Data |
|---|---|---|---|---|---|---|---|---|---|---|
| Auc. | Cicado. | Cicadellidae | *Stephanolla* | *rufoapicata* | 930,754 | 91.23% | 4209 | 1059 | 25.16% | 41.22% |
| Het. | Cimico. | Anthocoridae | *Xylastocoris* | *sp.* | 800,230 | 90.14% | 2594 | 697 | 26.87% | 30.46% |
| Het. | Cimico. | Cimicidae | *Cimex* | *adjunctus* | 6,127,852 | 96.72% | 5587 | 1216 | 21.76% | 31.42% |
| Het. | Cimico. | Miridae | *nr. Sophianus* | *sp.* | 3,164,378 | 92.44% | 2643 | 305 | 11.54% | 61.73% |
| Het. | Cimico. | Nabidae | *Alloeorhynchus* | *sp.* | 2,590,442 | 89.28% | 2278 | 570 | 25.02% | 34.36% |
| Het. | Cimico. | Reduviidae | *Dipetalogaster* | *maximus* | 2,040,344 | 96.88% | 4317 | 913 | 21.15% | 47.59% |
| Het. | Cimico. | Reduviidae | *Panstrongylus* | *geniculatus* | 1,177,566 | 89.61% | 3026 | 1177 | 38.90% | 24.6% |
| Het. | Cimico. | Reduviidae | *Psammolestes* | *arthuri* | 8,484,296 | 97.83% | 4164 | 1588 | 38.14% | 21.5% |
| Het. | Cimico. | Reduviidae | *Rhodnius* | *robustus* | 3,195,092 | 89.93% | 3793 | 1508 | 39.76% | 13.69% |
| Het. | Cimico. | Reduviidae | *Triatoma* | *dimidiata* | 2,971,874 | 92.37% | 4768 | 1401 | 29.38% | 22.83% |
| Het. | Dipsocoro. | Ceratocombidae | *Trichotonannus* | *sp.* | 920,424 | 91.64% | 1389 | 265 | 19.08% | 58.74% |
| Het. | Dipsocoro. | Schizopteridae | *Hoplonannus* | *sp.* | 657,912 | 90.31% | 1085 | 271 | 24.98% | 62.46% |
| Het. | Enicocephalo. | Enicocephalidae | *Oncylocotis* | *sp.* | 1,099,202 | 90.89% | 1235 | 347 | 28.10% | 50.32% |
| Het. | Gerro. | Gerridae | *Gerris* | *sp.* | 1,721,476 | 91.89% | 4080 | 1290 | 31.62% | 34.35% |
| Het. | Gerro. | Hebridae | *Hebrus* | *ifellus* | 2,034,830 | 87.07% | 2453 | 481 | 19.61% | 45.38% |
| Het. | Leptopodo. | Leptopodidae | *Valleriola* | *sp.* | 1,956,488 | 90.98% | 2326 | 601 | 25.84% | 40.85% |
| Het. | Nepo. | Belostomatidae | *Abedus* | *indentatus* | 1,742,898 | 88.58% | 1908 | 577 | 30.24% | 38.01% |
| Het. | Nepo. | Corixidae | *Micronectus* | *sp.* | 2,470,014 | 91.99% | 2270 | 502 | 22.11% | 45.21% |
| Het. | Pentatomo. | Aradidae | *Mezira* | *sp.* | 1,199,180 | 93.37% | 1334 | 381 | 28.56% | 57.49% |
| Het. | Pentatomo. | Coreidae | *Anisoscelis* | *flavolineatus* | 1,344,146 | 89.15% | 5381 | 698 | 12.97% | 34.55% |
| Het. | Pentatomo. | Coreidae | *Anoplocnemis* | *sp.* | 1,433,772 | 92.32% | 6701 | 1035 | 15.45% | 15.16% |
| Het. | Pentatomo. | Coreidae | *Mozena* | *nr. lineolata* | 1,528,344 | 88.13% | 6001 | 967 | 16.11% | 22.92% |
| Het. | Pentatomo. | Coreidae | *Acanthocephala* | *thomasi* | 2,764,968 | 92.07% | 8911 | 1215 | 13.63% | 16.93% |
| Het. | Pentatomo. | Coreidae | *Acanthocephala* | *femorata* | 839,558 | 91.11% | 4037 | 814 | 20.16% | 25.95% |
| Het. | Pentatomo. | Coreidae | *Lycambes* | *sargi* | 1,081,110 | 91.00% | 4890 | 887 | 18.14% | 19.28% |
| Het. | Pentatomo. | Coreidae | *Mygdonia* | *tuberculosa* | 1,301,796 | 92.40% | 5855 | 1046 | 17.87% | 10.03% |
| Het. | Pentatomo. | Coreidae | *Stenoeurilla* | *nr. prolixa* | 722,064 | 91.30% | 3152 | 944 | 29.95% | 18.1% |
| Het. | Pentatomo. | Coreidae | *Thasus* | *neocalifornicus* | 1,779,776 | 90.48% | 5862 | 1163 | 19.84% | 14.71% |
| Het. | Pentatomo. | Cydnidae | | | 1,769,802 | 90.76% | 2135 | 946 | 44.31% | 26.49% |
| Het. | Pentatomo. | Lygaeidae | *Oncopeltus* | *sp.* | 1,267,356 | 91.94% | 4301 | 1696 | 39.43% | 21.45% |
| Het. | Pentatomo. | Pentatomidae | *Brochymena* | *sp.* | 1,764,158 | 92.78% | 4784 | 1460 | 30.52% | 26.68% |
| Het. | Pentatomo. | Pentatomidae | *Euschistus* | *latimarginatus* | 1,511,668 | 94.13% | 4665 | 1437 | 30.80% | 24.88% |
| Het. | Pentatomo. | Pachygronthidae | *Oedancala* | *sp.* | 2,149,304 | 91.70% | 2223 | 605 | 27.22% | 33.88% |
| Ste. | [S.F.] Psylloidea | Aphalaridae | *Glycaspis* | *brimblecombei* | 989,064 | 89.77% | 3436 | 776 | 22.58% | 35.76% |
| Ste. | [S.F.] Aphidoidea | Aphididae | *Aphis* | *fabae* | 3,858,732 | 92.57% | 3053 | 1,240 | 40.63% | 27.98% |
| Ste. | [S.F.] Psylloidea | Psyllidae | *Heteropsylla* | *texana* | 775,336 | 91.82% | 2967 | 479 | 16.14% | 55.69% |
| Thy. | | Phlaeothripidae | *Klambothrips* | *myopori* | 171,672 | 88.50% | 887 | 117 | 13.19% | 81.45% |
| | | | | **Averages** | **1,955,078** | **91.49%** | **3641** | **883** | **25.32%** | **34.44%** |

et al., 2017). Next, we used a custom pipeline (https://github.com/AlexKnyshov/main_repo) that uses tblastx to search for homologous loci in transcriptomes. We extracted the best matching amino-acid coding portion of sequences from transcriptomes that matched UCE loci with an e-value of 1e−10 or less, which allowed for inclusion of even short matching sequences in amino acid space. After excluding *Xenophysella greensladeae*, the taxon with the fewest UCE loci recovered and the worst assembly, we realigned the data using the MAFFT E-INS-i algorithm and trimmed the alignments from the end to include at least 80% taxon representation. We used the best of 20 ML trees, followed by 100 bootstrap replicates, using the GTRGAMMA model with genes partitioned by locus. Furthermore, for the four genomes with annotated coding sequences on GenBank that were used in bait design, we used blastn to assess whether the UCE loci corresponded with coding regions. We used an e-value cutoff of 1e−30 (equivalent to an exact match of a string of ~75 base pairs). We also conducted a cross-species check for the pair of the most closely related genome and transcriptome included in our analysis, using tblastx and an e-value cutoff of 1e−10 (as used in our analysis to find corresponding loci in transcriptomes). Analyses were conducted on the University of Georgia and the University of Connecticut high-performance computer clusters.

## 3. Results

### 3.1. UCE recovery

Summary results for newly generated UCE data are presented in Table 1. We produced 2,114,434 raw paired-end reads per sample on average, with an average of 1,955,078 (91.49%) passing filter. Assemblies resulted in an average of 3641 contigs per sample. We recovered a total of 2721 UCE loci across all taxa. Loci per sample ranged between 265 and 1696 (average = 904) for hemipteran taxa, with 117 loci from the thysanopteran outgroup for which new data was gathered. From the 50% and 60% complete data matrices, we recovered 532 and 220 UCE loci from the empirical data set, respectively, while the inclusion of *in silico* data increased recovery to 744 and 325 UCE loci, respectively. We found an average of 34.44% (33.13% for Hemiptera) missing data between UCE alignments, with a range of 10.03% to 62.46% within Hemiptera and 81.45% for the outgroup. We recovered more UCE loci than average for two dried specimens used in this study (Table 1, Supplementary Table S1). Summary results of empirically generated UCE data processed with *in silico* data are presented in Supplementary Table S2. Summary numbers for parsimony and invariant sites for each data matrix are reported in Supplementary Table S3.

Within Heteroptera, UCE loci numbers varied between the infraorders, e.g., from a low of 268 loci (Dipsocoromorpha) to a high of 1042 loci (Cimicomorpha), corresponding to the amount of missing data within each group (Supplementary Table S4). Overall, the number of loci recovered was less than expected when compared to the Faircloth (2017) *in silico* study. We found the average number of loci, compared to the *in silico* study, to be 73.33% within the same genus, 68.02% within the same family, 52.74% within the same suborder, 51.28% within the order Hemiptera, and 13.54% within the Thysanoptera (Supplementary Table S5).

**Table 2**
Summary results of UCE loci found in annotated protein-coding sequences of genomes (top), and UCE loci designed for *Gerris buenoi* that matched to the *Velia caprai* transcriptome (bottom).

| UCE loci vs. CDS of self | | | | |
|---|---|---|---|---|
| Species | # of target UCE loci | # of transcripts | # of blastn hits with 1e−30 cutoff | % match |
| *Acyrthosiphon pisum* | 2059 | 30,790 | 2050 | 99.56 |
| *Cimex lectularius* | 2283 | 26,626 | 2273 | 99.56 |
| *Diaphorina citri* | 1545 | 21,652 | 1364 | 88.28 |
| *Halyomorpha halys* | 2257 | 27,675 | 2233 | 98.94 |
| UCE loci vs. Velia transcriptome | | | | |
| Species | # of target UCE loci | # of transcripts | # of tblastx hits with 1e−10 cutoff | % match |
| *Gerris buenoi* | 2266 | 46,481 | 1599 | 70.56 |

### 3.2. UCEs from transcriptome data

The results of extracting UCE loci from 10 transcriptomes are shown in Supplementary Table S6. We recovered the most UCE loci from the coding sequences of the *Bemisia tabaci* genome, most likely due to its completeness compared to transcriptomes based only on cDNA sequencing (88%, or 287 of the 325 loci, used in recovering the tree). We excluded *Xenophysella greensladeae* due to the low N50 of the assembly and the few UCE loci we were able to recover. On average, we recovered 71.5% of the 325 loci used in reconstructing the phylogeny across the remaining eight transcriptomes. For the four annotated genomes, we found that an average of 96.5% UCE loci of the ~1500–2300 per taxon contained a match to an annotated protein-coding sequence (Table 2). With a cross-species check of a closely related genome and transcriptome pair, we found that 70.5% of the 2266 UCE loci designed for *Gerris buenoi* could be found in the transcriptome of *Velia caprai* (Table 2).

### 3.3. Phylogenetic trees and taxa relationships

The phylogenetic tree for the full set of empirical, *in silico*, and transcriptome data is shown in Fig. 1. Bootstrap values were 100% for all but two (*Glycaspis* + *Pachypsylla* and *Brochymena* + *Halyomorpha*) of the shallow evolutionary relationships (Fig. 1) with a trend of decreasing support values with increased evolutionary depth. Trees for each data set showed mostly consistent topologies with similar support, trending toward more support at deeper phylogenetic nodes when additional taxa are added (i.e., addition of *in silico* and transcriptome data). The inclusion of this additional data did help to recover the well-supported and uncontentious relationship of Sternorrhyncha as sister to Auchenorrhyncha + Heteroptera which was not recovered otherwise in analyzing the newly acquired data in combination with *in silico* data or by itself (Supplementary Figs. S1 and S2).

Our analysis recovered a monophyletic Hemiptera, with Sternorrhyncha highly supported as the sister group to Auchenorrhyncha + Heteroptera. Support for Auchenorrhyncha + Heteroptera was weak (68%). Inter-infraorder support within Heteroptera ranged from 77 to 100%. We recovered with high support a monophyletic Geoheteroptera (Leptopodomorpha + (Cimicomorpha + Pentatomomorpha)), the land bugs, which are sister to a clade comprising the four remaining heteropteran infraorders. A strongly supported clade comprising Enicocephalomorpha, Dipsocoromorpha, and Gerromorpha) (GED clade; 100%) was recovered and moderately supported (77%) as the sister group of Nepomorpha. No topological differences were observed between phylogenetic trees produced using 50% versus 60% data matrices.

The results of intra-familiar level sampling of Reduviidae that focused on the subfamily Triatominae strongly supported (both 100%) a *Rhodnius* + *Psammolestes* clade as sister to *Panstrongylus* + (*Dipetalogaster* + *Triatoma*). For Coreidae, our analysis, which includes two subfamilies and six tribes, recovered all relationships with 100% support. Both species of *Acanthocephala* (Acanthocephalini) were recovered as sister to one another. The genera *Mygdonia* and *Anoplocnemis* were also recovered as sister taxa, supporting a monophyletic Mictini, which is sister to *Anisoscelis* + (*Stenoeurilla* + *Acanthocephala*). The only sampled representative of the subfamily Meropachyinae, *Lycambes sargi*, was nested within the coreine tribe Nematopodini, which together formed a clade sister to all other sampled coreids.

## 4. Discussion

### 4.1. Study rationale

Ultraconserved elements have been widely used for phylogenetic research among vertebrate groups during the past several years, with arthropod UCEs being developed comparatively more recently. Research in the area of arthropod UCEs is still largely open and untested for the vast majority of taxonomic groups. While several UCE bait sets have been empirically evaluated (Baca et al., 2017; Faircloth et al., 2015; Starrett et al., 2017; Van Dam et al., 2017), the designs for Diptera, Hemiptera, and Lepidoptera (Faircloth, 2017) have yet to be similarly evaluated. We demonstrate the utility of the Hemiptera UCE bait set (Faircloth, 2017) across divergent hemipteran taxa.

### 4.2. Recovery of UCE loci

We obtained UCE loci from all taxa sequenced and the number of loci recovered for each sample was correlated with sequencing read depth, consistent with previous UCE studies. We were also able to recover a large number of loci from 25 + year old museum specimens, which will facilitate studies of heteropteran taxa in the future. The taxa for which we recovered the most UCE loci were frequently those with the closest relationship to the species used for bait design, e.g., *Oncopeltus* sp. and *Rhodnius robustus* are congeneric with *Oncopeltus fasciatus* and *Rhodnius prolixus*, respectively, and were the two taxa with the most UCE loci recovered. The amount of missing UCE data for taxa sampled is correlated (Pearson $r = 0.477$, $p = 0.003$) with greater phylogenetic divergence from taxa used to design baits (Supplementary Table S5), which along with some nodes at deeper evolutionary depths having lower support, highlights the importance of including phylogenetically diverse taxa when developing baits for lineages as old as Hemiptera (300 mya; (Misof et al., 2014). Despite the missing data, however, most nodes had 100% support, and all but four exceeded 70% (Fig. 1).

On average, we obtained about two-thirds the number of loci we expected when compared to the *in silico* study (Faircloth, 2017) and 7.4 times less sequencing data than the generated *in silico* data. The amount of UCE loci recovered positively trended (Pearson $r = 0.328$, $p = 0.054$) with the amount of coverage. With increased sequencing

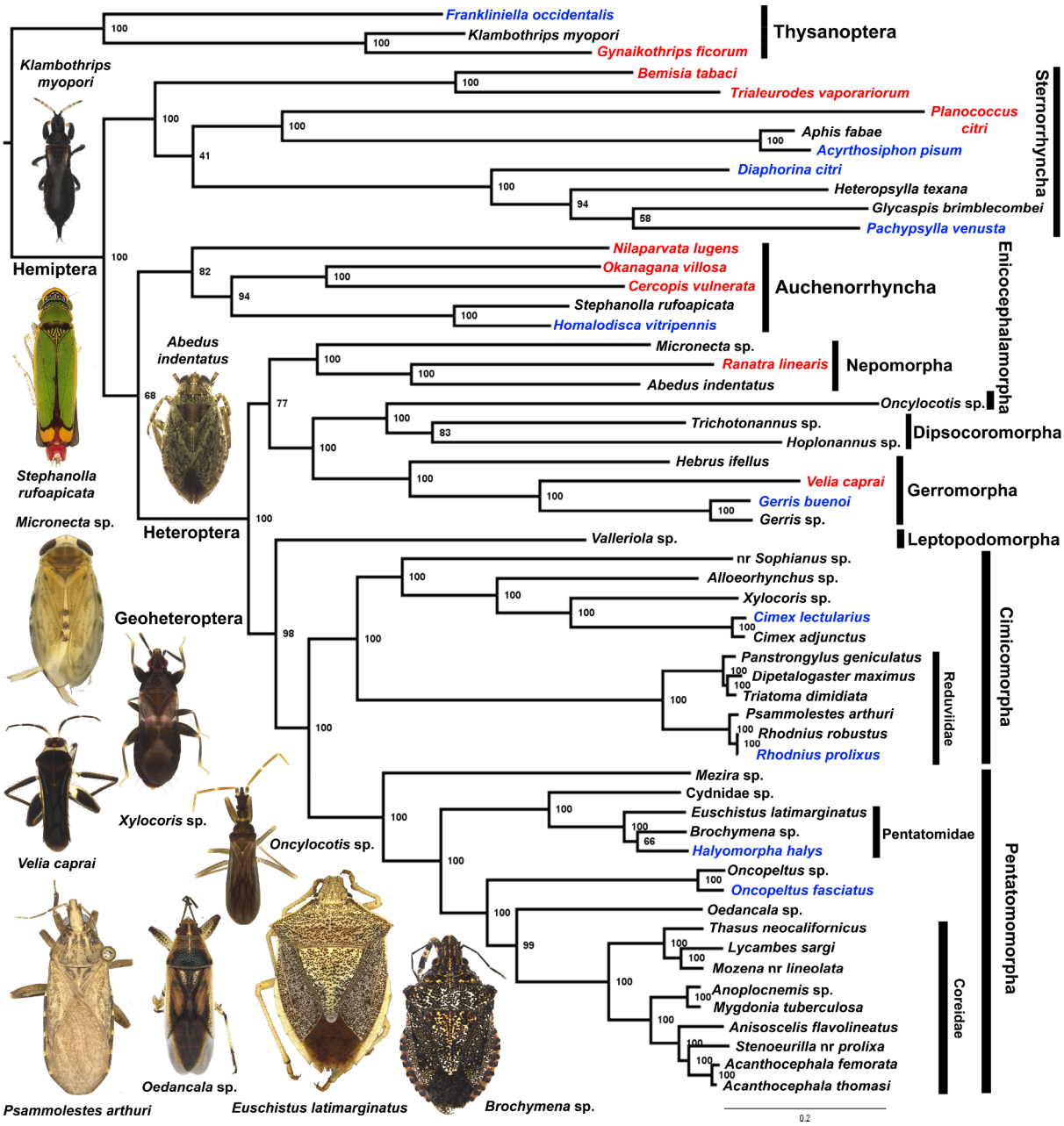


**Fig. 1.** Best Maximum Likelihood tree from a search of 20 trees with 100 bootstraps of the 80% data matrix of samples using empirical, *in silico* (blue), and transcriptome UCE data (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

depth for certain samples, it is likely we would obtain more unique UCE loci, and the gap between *in silico* expected and empirically obtained would diminish, though even at this level of coverage the data can resolve most relationships. Changes to the assembly methods and enrichment stringency may also increase the number of loci collected across Hemiptera.

We recovered matching sequences for about 90% of the 325 UCE loci from the empirical + *in silico* dataset for which we conducted a search on amino acid coding sequences of the *Bemisia tabaci* genome. We also found an average of 96.5% UCE loci matching annotated protein-coding sequences with the corresponding percentages in two of the four examined taxa as high as 99.6%. We further investigated the only nine loci with no matches in the annotations of the aphid genome, which was the genome with the fewest number of loci without matches to clarify the nature of these UCE loci. We found that seven of these nine loci matched proteins annotated in other aphid species and two loci corresponded with spliceosomal RNAs (U11 and U12). Thus for the pea aphid at least 2,057 of 2,059 UCE loci (99.9%) contain a protein-encoding core. We suspect that lower percentages found in some transcriptome and genome assemblies can be attributed to incomplete annotations, assemblies, or limited sequencing efforts. This reflects a fundamental difference of UCE loci in vertebrates where UCEs are primarily noncoding yet conserved elements versus invertebrates, where they are primarily protein-coding, as is being increasingly recognized (Bossert and Danforth, 2018).

### 4.3. Systematics of Hemiptera

Recent published studies have proposed several alternative hypotheses for relationships within Hemiptera, and particularly within the Heteroptera (Wang et al., 2017; Weirauch et al., 2018). Using a large molecular dataset of loci not previously employed that samples broadly

throughout the genome, our analyses test these relationships and corroborate some. For example, our phylogenetic hypothesis is congruent with many recently published topologies (Misof et al., 2014; Wang et al., 2017; Weirauch et al., 2018) in supporting a monophyletic Auchenorrhyncha as sister to the Heteroptera (Coleorrhyncha was not included in this analysis, so their position was not evaluated). Consistent with most analyses in recent decades (reviewed in Weirauch et al., 2018), relationships within the more densely sampled Heteroptera strongly support the monophyly of Geoheteroptera, the land bugs, which include the great majority of the extant species diversity in this suborder. Also congruent with some recently published phylogenetic hypotheses is the well-supported clade formed by the Gerromorpha, Enicocephalomorpha, and Dipsocoromorpha [GED clade; (Wang et al., 2017; Weirauch et al., 2018)], although relationships within this clade differ between analyses with either Dipsocoromorpha (this study; Wang et al., 2017) or Gerromorpha (Weirauch et al., 2018) being recovered as sister group to the two remaining infraorders. However, one of the most controversial issues that has significant impact on our understanding of character evolution within the Heteroptera remains unresolved: in contrast to recent phylogenies that either supported the GED clade (Wang et al., 2017) or the aquatic Nepomorpha (Weirauch et al., 2018) as the sister group to all remaining Heteroptera, the current analysis modestly (77%) supported Nepomorpha as sister to the GED clade, putting forward a third alternative hypothesis. Taxon sampling within the Geoheteroptera in our analysis is limited, but relationships conform with currently accepted hypotheses (e.g., Aradidae as sister to Trichophora, Pentatomoidea sister to all other Trichophora, Miroidea and Cimicoidea + Naboidea clade are sister taxa within Cimicomorpha [Weirauch et al., 2018]).

Intra-familial relationships within the reduviid subfamily Triatominae are consistent with previous phylogenetic analyses of the group based on fewer loci (Georgieva et al., 2017; Justi et al., 2014). For the Coreidae, relationships among and within the four subfamilies and 37 tribes have remained unresolved across morphological and single-gene phylogenetic studies (Li, 1997; Pan et al., 2007; Souza et al., 2016). We expected and recovered a sister group relationship between the two sampled species of *Acanthocephala* (Acanthocephalini). Phylogenetic analyses during the past couple of decades have also supported the monophyly of the Micitini, albeit with different taxon sampling compared to our study that included *Anoplocnemis* and *Mygdonia* (Li, 1997; Pan et al., 2007). Our analysis recovered a paraphyletic Coreinae with respect to *Lycambes* (Meropachyinae), a result that has been supported in some previous analyses (Li, 1996, 1997). However, in these previous studies, Meropachyinae was supported as the sister group to Chariesterini (not included in our analysis), whereas our study finds the subfamily to be nested within the Nematopodini (*Thasus* and *Mozena*).

Furthermore, our results show congruent topologies across data sets, indicating the usefulness of UCEs even with a relatively small number of samples. However, increasing the taxonomic representation with the addition of *in silico* and transcriptome data improved support for many deep phylogenetic nodes, which may improve with further additions. More importantly, the ability to sample UCE loci from other genomic resources expands possibilities of taxonomic representation and improves the utility of UCEs for phylogenomic studies.

### 4.4. Conclusion

Our study adds to the accumulating evidence that custom UCE bait sets can resolve most phylogenetic nodes with high bootstrap support, including baits designed for invertebrate groups. We also have shown the capability of integrating our invertebrate UCE loci with protein-coding data from transcriptomes. As phylogenomic datasets become more common and varied in structure, the capability of combining large genetic datasets with others from different sources will become more important to generate a complete tree of life, which represents another strength of this approach. Because the number of loci recovered empirically is significantly lower than expected, we recommend researchers explore various options to improve loci recovery as needed based on study objectives. For example, comparing different assembly methods, less stringent enrichment conditions, and incorporating additional taxa to improve phylogenetic relationships. Certain taxonomic clades within Hemiptera may also benefit from a designed subset of UCE baits as more genomic resources become available. However, we have shown that established relationships within Hemiptera can be recovered with relatively few loci, which provide broad application of the current bait set to researchers.

## Acknowledgements

Work was supported by the National Science Foundation [DEB-1136626 to B.C.F. and T.C.G. and IOS-1553100 awarded to Christine W. Miller]. We thank Arbor Biosciences for providing complimentary MYbaits kit for evaluations. We thank Alexander Knyshov for use of his scripts. Authors declare no conflicts of interest.

## Data accessibility

Raw sequencing reads are deposited in Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/SRP161492). All configuration files and scripts, milestone data used for analyses, and a workflow are available on Dryad Digital Repository (https://doi.org/10.5061/dryad.425vg4m). Scripts used for transcriptome analyses are on the github repository of Alexander Knyshov (https://github.com/AlexKnyshov/main_repo).

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2018.10.026.

## References

Alexander, A.M., Su, Y.C., Oliveros, C.H., Olson, K.V., Travers, S.L., Brown, R.M., 2017. Genomic data reveals potential for hybridization, introgression, and incomplete lineage sorting to confound phylogenetic relationships in an adaptive radiation of narrow-mouth frogs. Evolution 71, 475–488.
Baca, S.M., Alexander, A., Gustafson, G.T., Short, A.E.Z., 2017. Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of 'Hydradephaga'. Syst. Entomol. 42, 786–795.
Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D., 2004. Ultraconserved elements in the human genome. Science 304, 1321–1325.
Blaimer, B.B., Brady, S.G., Schultz, T.R., Lloyd, M.W., Fisher, B.L., Ward, P.S., 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. BMC Evol. Biol. 15, 271.
Blaimer, B.B., Lloyd, M.W., Guillory, W.X., Brady, S.G., 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS One 11, e0161531.
Bossert, S., Danforth, B.N., 2018. On the universality of target-enrichment baits for phylogenomic research. Methods Ecol. Evol.
Bossert, S., Murray, E.A., Blaimer, B.B., Danforth, B.N., 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol. Phylogenet. Evol. 111, 149–157.
Branstetter, M.G., Jesovnik, A., Sosa-Calvo, J., Lloyd, M.W., Faircloth, B.C., Brady, S.G., Schultz, T.R., 2017a. Dry habitats were crucibles of domestication in the evolution of agriculture in ants. Proc. Biol. Sci. 284.
Branstetter, M.G., Longino, J.T., Ward, P.S., Faircloth, B.C., Price, S., 2017b. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. Methods Ecol. Evol. 8, 768–776.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8, 783–786.
Crawford, N.G., Parham, J.F., Sellas, A.B., Faircloth, B.C., Glenn, T.C., Papenfuss, T.J., Henderson, J.B., Hansen, M.H., Simison, W.B., 2015. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83, 250–257.
Cryan, J.R., Urban, J.M., 2012. Higher-level phylogeny of the insect order Hemiptera: is Auchenorrhyncha really paraphyletic? Syst. Entomol. 37, 7–21.
Faircloth, B.C., 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32, 786–788.
Faircloth, B.C., 2017. Identifying conserved genomic elements and designing universal

bait sets to enrich them. Methods Ecol. Evol. 8, 1103–1112.
Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Resour. 15, 489–501.
Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.
Faircloth, B.C., Sorenson, L., Santini, F., Alfaro, M.E., 2013. A Phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One 8, e65923.
Georgieva, A.Y., Gordon, E.R.L., Weirauch, C., 2017. Sylvatic host associations of Triatominae and implications for Chagas disease reservoirs: a review and new host records based on archival specimens. PeerJ 5, e3826.
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. Mol. Phylogenet. Evol. 92, 140–146.
Glenn, T.C., Nilsen, R., Kieran, T.J., Finger, J.W., Pierson, T.W., Bentley, K.E., Hoffberg, S., Louha, S., Garcia-De-Leon, F.J., del Rio, Angel, Portilla, M., Reed, K., Anderson, J.L., Meece, J.K., Aggery, S., Rekaya, R., Alabady, M., Belanger, M., Winker, K., Faircloth, B.C., 2016. Adapterama I: Universal Stubs and Primers for Thousands of Dual-Indexed Illumina Libraries (iTru & iNext). BioRxiv.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652.
Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C., Brumfield, R.T., 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Syst. Biol. 65, 910–924.
Justi, S.A., Russo, C.A., Mallet, J.R., Obara, M.T., Galvao, C., 2014. Molecular phylogeny of Triatomini (Hemiptera: Reduviidae: Triatominae). Parasit Vectors 7, 149.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.
Li, H., Shao, R., Song, N., Song, F., Jiang, P., Li, Z., Cai, W., 2015. Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. Sci. Rep. 5, 8527.
Li, M., Tian, Y., Zhao, Y., Bu, W., 2012. Higher level phylogeny and the first divergence time estimation of Heteroptera (Insecta: Hemiptera) based on multiple genes. PLoS One 7, e32152.
Li, X., 1996. Cladistic analysis and higher classification of Coreoidea (Heteroptera). Entomologia Sinica 3, 283–292.
Li, X., 1997. Cladistic analysis of the phylogenetic relationships among the tribal rank taxa of Coreidae (Hemiptera-Heteroptera: Coreoidea). Acta Zootaxonomica Sinica 22, 60–68.
Manthey, J.D., Campillo, L.C., Burns, K.J., Moyle, R.G., 2016. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: Piranga). Syst. Biol. 65, 640–650.
McCormack, J.E., Harvey, M.G., Faircloth, B.C., Crawford, N.G., Glenn, T.C., Brumfield, R.T., 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8, e54848.
Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspock, U., Aspock, H., Bartel, D., Blanke, A., Berger, S., Bohm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B.M., Schutte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M., Zhou, X., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767.
Moyle, R.G., Oliveros, C.H., Andersen, M.J., Hosner, P.A., Benz, B.W., Manthey, J.D., Travers, S.L., Brown, R.M., Faircloth, B.C., 2016. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat. Commun. 7, 12709.
Pan, X.L., Guan, J., Su, F.K., 2007. Discussion on the phylogeny of partial species of Coreinae and Mictinae based on sequences of cytochrome b gene (Hemiptera: Coreidae). Sichuan J. Zool. 26, 516–519.
Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Brumfield, R.T., 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63, 83–95.
Song, N., Li, H., Cai, W., Yan, F., Wang, J., Song, F., 2016. Phylogenetic relationships of Hemiptera inferred from mitochondrial and nuclear genes. Mitochondrial DNA A DNA Mapp. Seq. Anal. 27, 4380–4389.
Souza, H.V., Marchesin, S.R., Itoyama, M.M., 2016. Analysis of the mitochondrial COI gene and its informative potential for evolutionary inferences in the families Coreidae and Pentatomidae (Heteroptera). Genet. Mol. Res. 15.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Starrett, J., Derkarabetian, S., Hedin, M., Bryson Jr., R.W., McCormack, J.E., Faircloth, B.C., 2017. High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. Mol. Ecol. Resour. 17, 812–823.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.
Van Dam, M.H., Lam, A.W., Sagata, K., Gewa, B., Laufa, R., Balke, M., Faircloth, B.C., Riedel, A., 2017. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-weevils. PLoS One 12, e0188044.
Wang, Y.-H., Cui, Y., Rédei, D., Baňař, P., Xie, Q., Štys, P., Damgaard, J., Chen, P.-P., Yi, W.-B., Wang, Y., Dang, K., Li, C.-R., Bu, W.-J., 2016. Phylogenetic divergences of the true bugs (Insecta: Hemiptera: Heteroptera), with emphasis on the aquatic lineages: the last piece of the aquatic insect jigsaw originated in the Late Permian/Early Triassic. Cladistics 32, 390–405.
Wang, Y.-H., Wu, H.-Y., Rédei, D., Xie, Q., Chen, Y., Chen, P.-P., Dong, Z.-E., Dang, K., Damgaard, J., Štys, P., Wu, Y.-Z., Luo, J.-Y., Sun, X.-Y., Hartung, V., Kuechler, S.M., Liu, Y., Liu, H.-X., Bu, W.-J., 2017. When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera-Heteroptera. Cladistics 1–25.
Weirauch, C., Schuh, R.T., Cassis, G., Wheeler, W.C., 2018. Revisiting habitat and lifestyle transitions in Heteroptera (Insecta: Hemiptera): insights from a combined morphological and molecular phylogeny. Cladistics.
Wheeler, W.C., Schuh, R.T., Bang, R., 1993. Cladistic relationships among higher groups of Heteroptera congruence between morphological and molecular data sets. Entomologica Scandinavica 24, 121–137.
Xie, W., Chen, C., Yang, Z., Guo, L., Yang, X., Wang, D., Chen, M., Huang, J., Wen, Y., Zeng, Y., Liu, Y., Xia, J., Tian, L., Cui, H., Wu, Q., Wang, S., Xu, B., Li, X., Tan, X., Ghanim, M., Qiu, B., Pan, H., Chu, D., Delatte, H., Maruthi, M.N., Ge, F., Zhou, X., Wang, X., Wan, F., Du, Y., Luo, C., Yan, F., Preisser, E.L., Jiao, X., Coates, B.S., Zhao, J., Gao, Q., Xia, J., Yin, Y., Liu, Y., Brown, J.K., Zhou, X.J., Zhang, Y., 2017. Genome sequencing of the sweetpotato whitefly Bemisia tabaci MED/Q. Gigascience 6, 1–7.