Received: 3 May 2021 | Revised: 20 August 2021 | Accepted: 3 September 2021

DOI: 10.1111/zsc.12513

ORIGINAL ARTICLE

Zoologica Scripta | WILEY

Extracting 'legacy loci' from an invertebrate sequence capture data set

Caroline D. Miller[1] | **Michael Forthman**[1,2] | **Christine W. Miller**[1] | **Rebecca T. Kimball**[3]

[1]Department of Entomology & Nematology, University of Florida, Gainesville, FL, USA

[2]California State Collection of Arthropods, Plant Pest Diagnostics Branch, California Department of Food & Agriculture, Sacramento, CA, USA

[3]Department of Biology, University of Florida, Gainesville, FL, USA

Correspondence
Michael Forthman, California State Collection of Arthropods, Plant Pest Diagnostics Branch, California Department of Food & Agriculture, 3294 Meadowview Road, Sacramento, CA 95832, USA.
Email: michael.forthman@cdfa.ca.gov

Funding information
Division of Integrative Organismal Systems, Grant/Award Number: 1553100

Abstract

Sequence capture studies result in rich data sets comprising hundreds to thousands of targeted genomic regions that are superseding Sanger-based data sets comprised of a few well-known loci with historical uses in phylogenetics ('legacy loci'). However, integrating sequence capture and Sanger-based data sets is of interest as legacy loci can include different types of loci (e.g. mitochondrial and nuclear) across a potentially larger sample of species from past studies. Sequence capture data sets include nontargeted sequences, and there has been recent interest in extracting legacy loci from invertebrate data sets. Here, we use published legacy data from leaf-footed bugs (Hemiptera: Coreoidea) to recover 15 mitochondrial and seven nuclear legacy loci from off-target sequences in a sequence capture data set, explore approaches to improve legacy locus recovery, and combine these loci with sequence capture data for phylogenetic analysis. Two nuclear loci were determined to already be targeted by sequence capture baits. Most of the remaining loci were successfully recovered from off-target sequences, but this recovery varied greatly. Additionally, complementing complete mitogenomes with additional reference mitochondrial sequences from a genetic depository did not offer improvement for most of our taxa; however, supplementing these reference sequences with extracted legacy loci offered ≥6% improvement across taxa for a given mitochondrial locus (negligible improvement for nuclear loci). Phylogenetic analysis of legacy and sequence capture data produced a topology generally congruent with recent studies, but support was lower. Thus, future studies may employ the approaches used in this study to integrate legacy data with newly generated sequence capture data sets without added expenses.

KEYWORDS

legacy loci, next-generation sequencing, Sanger sequencing, sequence capture

1 | INTRODUCTION

The ability to sequence DNA revolutionized the field of systematics by establishing molecular phylogenetics. The shift from morphologically driven data sets to targeting a few Sanger-sequenced loci (herein referred to as legacy loci) allowed for independent data sources to infer the evolutionary history of several well-studied groups of organisms (Sanger, Air, et al., 1977; Sanger et al., 1977; Schuster, 2007). The contributions of early molecular phylogenetic studies greatly shaped our collective understanding of the Tree of Life. However, long-standing challenges in resolving phylogenetic nodes remain, partly due to the limited availability of well-sampled loci across model and nonmodel species.

Within the last decade, another critical transition occurred with the rise of next-generation sequencing (NGS), which facilitates the collection of large amounts of data. This technological advancement, coupled with approaches to subsample the genome (e.g. sequence capture), allows systematists to target hundreds to thousands of loci across model and nonmodel species (Schuster, 2007; Zhang, et al., 2019) and resolve historically challenging nodes in the Tree of Life (e.g. Allard et al., 2012; Hamilton et al., 2016; Parks et al., 2009). Sequence capture approaches, such as exon capture, ultraconserved elements (UCEs) and anchored hybrid enrichment (AHE), are being increasingly utilized in molecular phylogenomic studies, particularly due to their cost-effectiveness compared with whole-genome sequencing (Bi et al., 2013; Faircloth et al., 2012; Lemmon et al., 2012) and their scalability compared with Sanger sequencing (Peñalba et al., 2014). Furthermore, sequence capture approaches may allow for the inclusion of lineages that are relatively limited in genomic resources compared with model species (Bi et al., 2013; Faircloth et al., 2013; Lemmon & Lemmon, 2013; Zhang, Williams, et al., 2019).

Although NGS and sequence capture approaches continue to supplant Sanger-sequencing approaches based on one to a few loci (herein referred to as 'legacy loci'), there has been recent interest in exploring how legacy loci can complement sequence capture data (e.g. Blaimer et al., 2015; Branstetter et al., 2017, 2021; Derkarabetian et al., 2019; Hughes et al., 2021; Simon et al., 2019; Zhang et al., 2019). Specifically, the generation of more comprehensive data sets may increase resolution power for phylogenetic inference and allow for the inclusion of rare and vital species that may be difficult to sample repeatedly (Branstetter et al., 2017; Derkarabetian et al., 2019; Zhang, Williams, et al., 2019). The amount of legacy sequence data—as well as genomes and transcriptomes—has rapidly expanded over the past couple of decades and is readily available from public databases, such as the National Center for Biotechnology Information (NCBI) (Cameron, 2014; Łukasik et al., 2019). The wealth of available sequence data can be a useful resource when attempting to integrate legacy loci from Sanger-sequencing studies with sequence capture data.

One approach to integrating legacy loci with sequence capture data involves designing capture baits of legacy loci for inclusion in existing sequence capture bait kits, such as optimized UCE and exon-capture bait sets (Branstetter et al., 2017; Hughes et al., 2021; Simon et al., 2019). However, this approach may increase the cost of generating custom probe kits and may require more baits across more species due to higher rates of substitutions in some legacy loci (particularly mitochondrial DNA [mtDNA]). Additionally, some legacy loci, such as mtDNA loci, have a high copy number within each cell, which would require mtDNA baits to be designed in a separate kit and drastically diluted before combining with sequence capture kits targeting loci with lower copy numbers to prevent mtDNA dominance in capture data (Allio et al., 2020; Branstetter et al., 2021; Pierce et al., 2017; Ströher et al., 2016).

An alternative approach that has the potential to circumvent some of these issues includes the extraction of legacy loci from the by-catch of sequence capture studies (e.g. Amaral et al., 2015; Meiklejohn et al., 2014; Tamashiro et al., 2019; Wang et al., 2017; Zarza et al., 2018). Conceptually, sequence capture approaches should eliminate nontarget sequences, but off-target sequences are frequently recovered in sequence capture data sets and may include legacy loci (Caparroz et al., 2018; Derkarabetian et al., 2019; Simon et al., 2019). Although vertebrate sequence capture studies have extracted legacy loci from off-target sequences, to our knowledge, this has often not been performed in invertebrate sequence capture studies, although there is recent interest in doing so (e.g. Branstetter et al., 2017, 2021; Derkarabetian et al., 2019; Longino & Branstetter, 2020; Meza-Lázaro et al., 2018; Pierce et al., 2017; Simon et al., 2019). However, there may be additional challenges in extracting legacy loci (particularly fast-evolving loci) from off-target invertebrate sequence capture data relative to studies on vertebrates. First, old, divergent clades may have limited genetic data (i.e. reference sequence data) available in comparison with more well-studied vertebrate groups. This is a potential issue for these invertebrate groups as there may not be close relatives available to use as reference sequences. One other potential issue with this approach may arise when, for example baits target regions that include legacy loci, but these regions are then intentionally extracted from sequence capture data as 'off-targets.' This would allow some loci to be included twice in a data set, thus biasing phylogenomic results if not properly screened. As such, an additional step is required to verify that, for example, nuclear protein-coding legacy loci are not already targeted by baits targeting similar genomic regions in sequence capture studies, but we have not found indication that such a step is performed in some capture studies (e.g. Derkarabetian et al., 2019).

Here, we use published mitochondrial and nuclear sequence data deposited in NCBI to identify and extract legacy loci from a sequence capture data set comprised of

282 taxa of leaf-footed bugs and allies (Insecta: Hemiptera: Coreoidea). The Coreoidea is a morphologically and behaviourally diverse superfamily of insects with 3,106 species in five families comprised of numerous subfamilies and tribes (CoreoideaSF Team, 2021). This group of invertebrates is also a model in sexual selection studies and includes many agricultural pests. The objectives of this study were to (a) recover legacy loci from off-target sequences in a sequence capture data set, (b) explore approaches to increase legacy locus recovery, (c) determine whether any loci targeted by sequence capture baits correspond to known legacy loci used in past coreoid phylogenetic studies and (d) combine legacy locus data with sequence capture data to infer coreoid phylogeny.

2 | MATERIALS AND METHODS

2.1 | Molecular sampling of focal taxa

A total of 282 coreoid taxa were sampled for this analysis. Of these, we retrieved published contigs for 12 taxa from Kieran et al. (2019). We then retrieved raw sequence reads from 148 additional taxa published in recent sequence capture studies (Emberts et al., 2020; Forthman et al., 2019, 2020).

We generated new sequence data for the remaining 122 taxa (Table S1). Given the varying sizes of specimens and to sample similar amounts of tissues across samples, where possible, we used the legs, abdomen, thorax, head or whole body from specimens that were frozen, dried or preserved in ethanol or silica beads (our sampling primarily targeted freshly preserved material). Of these 122 taxa, 13 were subjected to DNA extraction, library construction and standard sequence capture protocols published in Forthman et al. (2019), while another 13 taxa were subjected to protocols in Forthman et al. (2020). The remaining taxa were subjected to the same DNA extraction and library construction protocols in Forthman et al. (2020), but the touchdown sequence capture protocol was modified in the following ways: (a) bait–target hybridization was performed at 65°C for 12 hr, followed by 62°C for 12 hr and then 60°C for 12 hr, (b) bait-target products were washed at 60°C, and (c) enriched libraries were amplified for 16–17 cycles before final pooling and sequenced at the University of Florida's Interdisciplinary Center for Biotechnology Research (ICBR) on a single Illumina HiSeq 3000 lane (2 × 100 run). Newly generated sequence reads were demultiplexed by ICBR.

For the newly generated sequence reads and those retrieved from Forthman et al. (2019, 2020) and Emberts et al. (2020), adapters were trimmed from sequence reads with illumiprocessor (Bolger et al., 2014; Faircloth et al., 2013). Duplicate reads were filtered using PRINSEQ-lite v0.20.4 (Schmieder & Edwards, 2011), with the remaining reads error-corrected in QuorUM v1.1.0 (Marçais et al., 2015). Quality reads were *de novo*-assembled into contigs with SPAdes v3.13.0 with the single-cell and auto coverage cut-off options invoked (Nurk et al., 2013).

2.2 | Mitochondrial DNA identification and extraction

We first retrieved 19 complete coreoid mitogenomes (i.e. includes all 13 protein-coding and two ribosomal mtDNA loci; see Table 1 for locus names and abbreviations) from NCBI as an initial set of reference sequences (mtDNA_ref1; Table S2). Here, we use the term 'loci' to define mtDNA protein-coding genes and rRNA to be comparable to nuclear loci, but we recognize that in most species, the mitochondrion is likely a single recombining locus. These genomes were used to determine the approximate full sequence length of each mtDNA locus targeted. We additionally retrieved individually accessioned coreoid sequences for each mtDNA locus from NCBI to supplement mitogenomes as a second set of reference sequences (mtDNA_ref2; see Table S2). Quality control of NCBI sequence data was necessary for accurate identification and extraction of legacy loci. The organism field in nine accessions listed these samples as coreoid species, but the accession descriptions and associated publication indicated these were from bacterial symbionts. As such, we excluded these accessions from our reference sequences. We additionally discovered that 54 sequences from *Leptoglossus occidentalis* Heidemann, 1910 were incorrectly listed as cytochrome oxidase subunit I (COX1) after comparing translated protein sequences against NCBI's BLAST+v2.10.1 nucleotide database; these sequences consistently had the best hits to cytochrome *b* (CYTB) sequences of other coreoid and insect taxa, except for a few hits to COX1 sequences from other *L. occidentalis* individuals published in the same study. We further confirmed this finding with the associated publication (Lesieur et al., 2019), which explicitly states that the CYTB locus was amplified. Thus, we corrected the gene ID for these *L. occidentalis* sequences from COX1 to CYTB. Lastly, we retained multiple sequences of the same mtDNA locus for accessions with the same species name given that species identity cannot be confirmed in the absence of available physical vouchers.

Mitochondrial legacy loci were identified and extracted from assembled contigs using the mtDNA_ref1 and mtDNA_ref2 references in MitoFinder v1.1 (Allio et al., 2020) using default settings. Because MitoFinder requires the annotated reference data to have standardized gene names, manual curation of our references was needed

TABLE 1 Locus taxonomy, classification and the number of reference sequences from unique taxa

Standard locus name	Locus abbreviation	Locus type	No. of sequences	No. of unique taxa
ATP synthase subunit 6	ATP6	mtDNA	22	20
ATP synthase subunit 8	ATP8	mtDNA	22	20
Cytochrome C oxidase subunit 1	COX1	mtDNA	1,681	286
Cytochrome C oxidase subunit 2	COX2	mtDNA	24	21
Cytochrome C oxidase subunit 3	COX3	mtDNA	58	21
Cytochrome b	CYTB	mtDNA	77	21
NADH-ubiquinone oxidoreductase chain 1	ND1	mtDNA	60	23
NADH-ubiquinone oxidoreductase chain 2	ND2	mtDNA	22	20
NADH-ubiquinone oxidoreductase chain 3	ND3	mtDNA	22	20
NADH-ubiquinone oxidoreductase chain 4	ND4	mtDNA	23	21
NADH-ubiquinone oxidoreductase chain 4L	ND4L	mtDNA	23	21
NADH-ubiquinone oxidoreductase chain 5	ND5	mtDNA	23	21
NADH-ubiquinone oxidoreductase chain 6	ND6	mtDNA	23	21
16S ribosomal RNA	rrnL	mtDNA	57	51
12S ribosomal RNA	rrnS	mtDNA	23	21
18S ribosomal RNA	18S rRNA	nucDNA	52	43
28S ribosomal RNA	28S rRNA	nucDNA	128	25
deformed (5′)	dfd5′	nucDNA	3	3
deformed (3′)	dfd3′	nucDNA	3	3
abdominal-A	abd-A	nucDNA	3	3
proboscipedia	pb	nucDNA	3	3
ultrabithorax	ubx	nucDNA	3	3
sex combs reduced	scr	nucDNA	3	3

Abbreviations: mtDNA, mitochondrial DNA; nucDNA, nuclear DNA.

to standardize the sequences (e.g. 'CO1,' 'COI' and 'cytochrome *c* oxidase subunit 1' were changed to COX1). We first extracted mtDNA legacy loci using the mtDNA_ref1 references. To see whether the inclusion of additional mtDNA sequences from a more diverse sampling of coreoids would improve mtDNA legacy locus recovery, we separately extracted loci using the mtDNA_ref2 references. The extracted sequences resulting from the MitoFinder analysis with mtDNA_ref1 and mtDNA_ref2 references were manually checked for contamination and errors using NCBI's complete nucleotide database; these checks included evaluation of *e*-values and bit scores (e.g. for ribosomal loci, an acceptable *e*-value was considered to be 1×10^{-100}) while taking into account relative sequence lengths (e.g. hits with lower scores might be a result of smaller sequence lengths) and determining whether sequences had multiple matches to the same gene (e.g. multiple matches to COX1) from closely related coreoids or other insects. We also performed a second quality check by aligning reference and extracted sequences for each locus using the FFT-NS-i iterative refinement algorithm in MAFFT v7 (Katoh et al., 2019) and visually inspecting them for premature stop codons and frameshift mutations at the 5′ or 3′ ends in Mesquite V3.61 (Maddison & Maddison, 2019). Any extracted sequences determined to be problematic after manual evaluation, as well as sequences less than 50 bp in length, were then excluded from the results.

Following searches with mtDNA_ref1 and mtDNA_ref2 references, some taxa still did not recover one or more legacy loci. Thus, we assessed whether the inclusion of newly extracted mtDNA legacy loci from our focal taxa (i.e. valid mtDNA_ref2 results) could improve recovery in those other taxa that remained unsuccessful for a given locus. To do this, we complemented our mtDNA_ref2 references with newly recovered mtDNA_ref2 sequences from our focal taxa (excluding problematic sequences; now referred to as mtDNA_ref3) and performed new searches only for the subset of taxa that had no recovery for a given locus in previous searches. In addition to excluding problematic sequences, some extracted mtDNA sequences were trimmed before adding them to the mtDNA_ref3 reference as these sequences exceeded the complete sequence range maximum (e.g. up to 425 bp more). Trimming was done using the BLAST output generated by MitoFinder to extract the NCBI accession for the best hit followed by a tblastx search with the new mtDNA_ref2 sequences. We then used the start and end coordinates from the best tblastx match to trim sequences using BEDTools v2.29.2 (Quinlan & Hall, 2010).

Because MitoFinder requires a GenBank-formatted reference sequence file, which was not available for the new extracted legacy sequences, we created a local BLAST nucleotide database for each mtDNA locus in the mtDNA_ref3 references, which were subsequently used to perform tblastx (for protein-coding loci) and blastn (for ribosomal loci) searches using the same settings as implemented in MitoFinder (i.e. *e*-value = 1×10^{-5}, per cent identity = 50, genetic code [for tblastx] set to invertebrate mitochondrial code). We then used BEDTools to extract the best-hitting sequence. Extracted sequences were manually checked as described above for the mtDNA_ref1 and mtDNA_ref2 results.

2.3 | Nuclear DNA identification and extraction

We retrieved 198 published coreoid sequences for seven nuclear DNA (nucDNA) loci from NCBI (nucDNA_ref1; see Table 1 for locus names and abbreviations; Table S2). To determine the complete sequence lengths of ribosomal nucDNA (i.e. 28S and 18S rRNA), literature searches were performed (DeLeo et al., 2018; Tautz et al., 1988; Xie et al., 2013). For most protein-coding nucDNA (i.e. ultrabithorax [ubx], proboscipedia [pb], sex combs reduced [scr], and abdominal-A [abd-A]), complete sequence lengths were determined from the complete, annotated genome of *Halyomorpha halys* Stål, 1855 (Hemiptera: Pentatomidae). For the protein-coding nucDNA locus deformed (dfd), Tian et al. (2011) targeted different regions of the locus (i.e. dfd5′ and dfd3′), and thus, only partial information on the dfd sequence length was available for coreoids.

A local BLAST nucleotide database was created for each nuclear locus (Table 1). These databases were subsequently used to search against the sequence capture data using tblastx for protein-coding loci or blastn for ribosomal loci. Nuclear ribosomal loci (i.e. 18S and 28S rRNA) were searched using an *e*-value cut-off of 1×10^{-50}, which is appropriate due to the conserved characteristics of this type of locus (Eickbush & Eickbush, 2007). Protein-coding loci were searched using an *e*-value cut-off of 1×10^{-30} (i.e. abd-A, dfd and scr) or 1×10^{-10} (i.e. pb and ubx). The *e*-value cut-off increase to 1×10^{-30} allowed for the increase of putative matches, while still maintaining high-quality hits; however, the protein-coding loci pb and ubx required a further increase to 1×10^{-10} to allow for the inclusion of a sufficient number of putative matches for evaluation (i.e. these loci had no recovery at 1×10^{-50} or 1×10^{-30}). To extract sequences with the best BLAST hit, BEDTools was used as outlined previously. Putative BLAST matches were evaluated following the manual assessments as described for the mtDNA loci. Any sequences deemed problematic were then excluded from results.

We assessed whether a second BLAST search using newly extracted sequences complementing our references (now nucDNA_ref2; excluding problematic sequences extracted from nucDNA_ref1 results) would increase nucDNA locus recovery in unsuccessful taxa. Using our expanded nucDNA reference sequences, we performed a similar search as for the nucDNA_ref1 BLAST search (i.e. using the same *e*-value cut-offs). Putative matches were similarly processed and verified as described above, retaining only those sequences that were not considered problematic.

2.4 | Screening legacy markers and targeted sequences

To confirm that targeted loci did not already consist of legacy loci of interest, we first extracted targeted loci from our sequence capture data using PHYLUCE v1.5.0

(Faircloth, 2016). We then performed a local tblastx (for protein-coding loci) or blastn (for ribosomal loci) search using extracted legacy loci and targeted loci from our focal taxa using an *e*-value cut-off of 1×10^{-50} for ribosomal loci, 1×10^{-30} for some nuclear protein-coding loci (i.e. abd-A, dfd and scr) and 1×10^{-10} for mitochondrial and some nuclear protein-coding loci (i.e. pb and ubx). In cases where putative matches were recovered, we queried both the legacy and target sequences against the entire NCBI nucleotide database to determine whether both had similar matches (with low *e*-value and high bit scores) to the same gene ID in multiple insect species. The extracted sequences, NCBI references and corresponding targeted sequences for each locus were also aligned and visually inspected as outlined previously.

2.5 | Phylogenetic analysis

An objective of extracting legacy loci from sequence capture studies is to obtain additional informative sites for phylogenetic inferences, as well as to include additional taxa for which only legacy loci are currently available (i.e. taxa not sampled in sequence capture studies). Here, we integrated Forthman et al.'s (2020) UCE sequence capture data from 124 taxa with legacy locus data obtained from two sources (i.e. those extracted from our focal taxa and those retrieved from NCBI) to test the most recent phylogenetic hypothesis of Coreidae based on UCEs (Forthman et al., 2020). In doing so, and to minimize computational time, we subsampled our legacy data set to include only the taxa analysed in Forthman et al. (2020), as well as NCBI reference taxa listed as species within the same genera as those in Forthman et al. (2020). Sequences from each legacy locus were individually aligned prior to concatenation with Forthman et al.'s (2020) '50p total evidence' UCE locus alignments (UCE loci containing 50% or more of the sampled taxa and not filtered by parsimony-informativeness; 1,000 loci) in PHYLUCE. We analysed this concatenated matrix of 194 taxa following Forthman et al. (2020). Specifically, we selected the best-fit partition scheme using PartitionFinder v2.1.1 (Lanfear et al., 2017), with the following settings: rcluster algorithm, unlinked branch lengths, all models under the 'raxml' option, individual loci treated as separate data blocks, and corrected Akaike information criterion (Hurvich & Tsai, 1989) for model selection. We then performed 20 partitioned maximum-likelihood optimal tree searches using the GTRGAMMA model of evolution and 500 bootstrap iterations in RAxML v8.2.10 (Stamatakis, 2014). Bootstrap support was summarized on the best tree using SumTrees v4.0.0 (Sukumaran & Holder, 2010).

Because our initial analysis resulted in unexpected and highly questionable phylogenetic results (specifically for the phylogenetic placement of many NCBI reference taxa; see Data S1 for phylogenetic tree), we subsequently filtered our legacy data set further to only include sequences from NCBI reference taxa that have been published in accessible theses (Banho, 2016; de Souza, 2013) or peer-reviewed journals so as to assess their taxon identity (e.g. species assignment likely given geographic source, accessioned sequences not from bacterial symbionts, or relatively low chances of contamination due to source of DNA material [such as DNA sequenced from gut contents or faeces of another organism]), sequence identity (e.g. *L. occidentalis* sequences accessioned as COX1, but the associated publication clearly states these are CYTB), and the approaches used to generate sequence data. However, we did not include 18S or COX1 sequences derived and published by Li et al. (2005) given potential issues concerning the quality and identity of some sequences (see Tian et al., 2011). We also found that some of our legacy loci were poorly sampled across our NCBI and focal taxa relative to others (e.g. nuclear protein-coding genes, ATP6, ND4L, etc.) resulting in little overlap in loci shared across both sets of taxa (which we expect to result in random placement of taxa when using these loci). As such, we restricted our legacy locus analysis to well-sampled loci across our taxon sampling (COX1, 18S and 28S) and excluded any NCBI reference taxon that did not have at least one sequence for these loci that was also recovered in a species of the same genus sampled by Forthman et al. (2020) (except three reference species of *Leptoglossus* that did not meet this criterion as these had legacy loci also found in other reference *Leptoglossus* species that did meet the criterion). This resulted in a data set with 124 focal taxa from Forthman et al. (2020) and 37 NCBI taxa (see Table S3 for accessions) for phylogenetic analysis.

Sequences from each of the three legacy loci were individually aligned, screened, and concatenated with Forthman et al.'s (2020) '50p total evidence' UCE alignments in PHYLUCE. We analysed our concatenated matrix of 161 taxa following the same procedures outlined above for our initial data set.

3 | RESULTS

3.1 | Some targeted sequences correspond to legacy loci

Overall, the localized BLAST searches resulted in no matches or matches with poor scores, with some exceptions. For most cases where there were acceptable local matches (i.e. low *e*-values and high bit scores), verification using NCBI's entire nucleotide database did not confirm the matches (Table S4). However, abd-A and scr legacy sequences did appear to be targeted by our sequence capture baits, with each recovered in over 50% of our focal taxa (Table S5). Nearly all recovered sequences for these two targeted loci were <50% of the length of the complete locus length (Table S5). Because these two nucDNA loci were targeted by sequence capture baits, we excluded them from the rest of our results presented below.

3.2 | Comparison of error rates across all reference files used

For mtDNA_ref1 results, 91 out of 1,117 (8%) sequences recovered across 14 mtDNA loci (out of 15; all ATP8 sequences were valid) were removed after manual evaluation (Table S6). From the results based on mtDNA_ref2 reference sequences, 80 out of 1,140 (7%) sequences recovered across 14 mtDNA loci (out of 15; all ND4L sequences were valid) were removed (Table S6). From the results based on mtDNA_ref3 references, 865 out of 2,309 (37%) sequences across all mtDNA loci were removed after evaluation (Table S6). For nucDNA_ref1 results, nine out of 490 (2%) sequences (18S rRNA, 28S rRNA and dfd3′ sequences) were considered problematic and excluded (Table S6). For nucDNA_ref2 results, 49 out of 533 (9%) sequences (18S rRNA, 28S rRNA and dfd3′ sequences) were removed after evaluation (Table S7). The results following this section have had these problematic sequences removed.

3.3 | Impact of mtDNA reference sequences on recovery of legacy loci

Our mtDNA reference dataset obtained from NCBI (i.e. mtDNA_ref2, which contained 19 mitogenomes + individual mtDNA sequences), included 22 (ATP6, ATP8, ND2 and ND3 each) to 1,681 (COX1) reference sequences (mean = 144 reference sequences per locus) from a cumulative total of 293 unique taxa, although the number of taxa available for a given locus varied (Table 1, Table S2; Figure 1). Of the total 2,160 mtDNA reference sequences compiled for mtDNA_ref2, 60.8% were <50% complete in length and 23.8% were ≥50% to <100% complete, while the remaining 15.4% attained or slightly surpassed (by up to 20 bp) the complete locus length range (Table S2).

When using mtDNA_ref1 or mtDNA_ref2 references to recover legacy loci, the locus recovery and recovered sequence lengths resulted in similar patterns of locus recovery (about four loci recovered on average, with ~230 taxa having at least one locus recovered), as well as the proportion of sequences that were <50% (~57% of sequences), ≥50% to <100% (~27%) and ≥100% (15%) complete in sequence length (Table 2, Tables S6, S8 and S9). For mtDNA_ref1 or mtDNA_ref2, about 44 taxa recovered at least 50% of the mitogenomes (i.e. eight or more loci), and about seven taxa recovered complete mitogenomes (i.e. all 15 loci) (Table 2, Table S6 and S8). With mtDNA_ref1, 68 taxa, on average, had sequences recovered for a given locus (20 [rrnL]–119 [COX1] taxa; Table S9; Figure 2). Similarly, on average, 71 taxa had sequences recovered

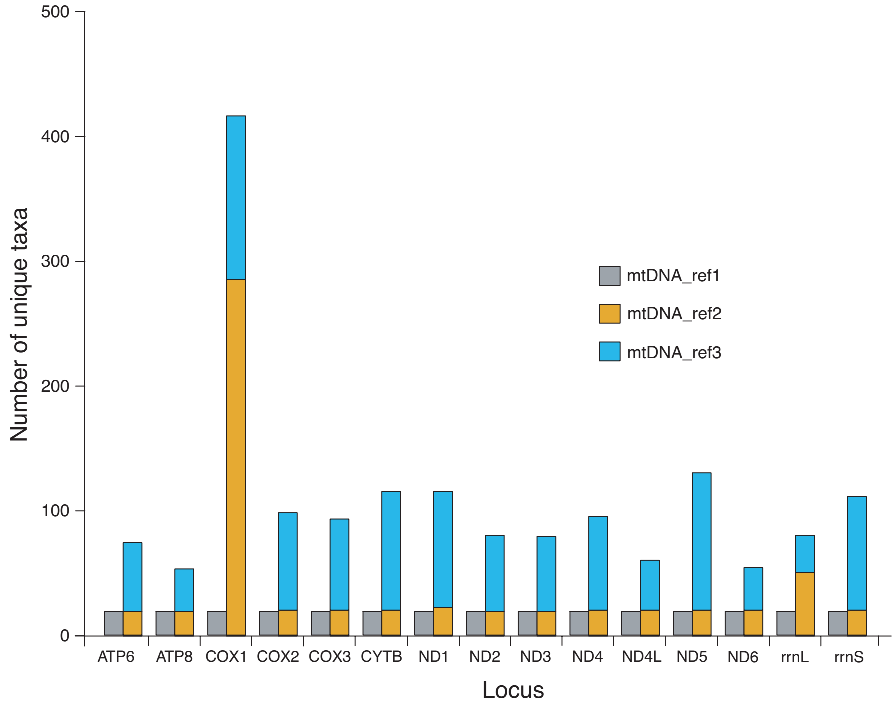

FIGURE 1 Unique reference taxa per mitochondrial legacy locus. mtDNA_ref1, 19 coreoid mitogenomic references; mtDNA_ref2, mtDNA_ref1 supplemented with individually accessioned coreoid sequences for each mtDNA locus from NCBI; mtDNA_ref3, mtDNA_ref2 references + verified mtDNA_ref2 results; see Table 1 for locus abbreviations

for a given locus (30 [rrnL]–131 [COX1] taxa) when using mtDNA_ref2 references (Table S9; Figure 2).

Increasing numbers of reference sequences had a mixed benefit. When compared to the mtDNA_ref1 results, the mtDNA_ref2 references increased locus recovery for 35 taxa (e.g. up to two more loci recovered), while the recovery was the same for 244 taxa (Tables S6 and S8). However, there were three taxa where recovery was worse in the mtDNA_ref2 search; after manual evaluation, one recovered sequence from each taxon was determined to be inaccurately assigned to a locus from the mtDNA_ref2 search (i.e. two ND6 sequences and one ATP8 sequence) and were subsequently removed (Tables S6 and S8).

Using mtDNA_ref3, we observed recovery for those taxa that had no recovery for a given locus from mtDNA_ref2 searches (Table 2, Tables S6 and S8); one additional locus was recovered on average, across 193 taxa. Most of these newly recovered sequences had lengths less than half of the known length of the targeted locus, with relatively few having obtained the full length of a locus (Table 2, Table S9). For newly recovered sequences, no taxa recovered all 15 loci, and two taxa recovered at least eight or more loci (Table 2, Tables S6 and S8). On average, 26 taxa had sequences recovered for a given locus for newly recovered sequences searching with mtDNA_ref3 (8 [COX3]–81 [rrnS] taxa; Table S9; Figure 2). Additionally, as mtDNA_ref3 builds off the newly extracted sequences recovered using mtDNA_ref2, when looking at locus recovery per taxon across all mtDNA reference files (i.e. mtDNA_ref1, mtDNA_ref2 and mtDNA_ref3), there were three cases where mtDNA_ref1 recovered a locus that was not recovered with mtDNA_ref2 and mtDNA_ref3 due to removal of problematic sequencing during screening (Table 2). Across all reference files, 13 taxa recovered complete mitogenomes, and 72 taxa recovered at least 50% of their mitogenomes (Table 2).

3.4 | Nuclear locus recovery varies across taxa

As two of our legacy loci (i.e. abd-A and scr) were targeted by baits, we only report results for the five loci not targeted. Our nucDNA reference data set obtained from NCBI (Table 1) included 3 (dfd, pb and ubx each) to 128 (28S rRNA) reference sequences across loci (mean = 25 reference sequences per locus) from 53 cumulative total taxa. Of the 192 reference sequences (Table 1), 78.1% were <50% complete, 16.7% were ≥50% to <100% complete, and 5.2% were within the complete locus length range (Table S2).

For searches using nucDNA_ref1, locus recovery varied across taxa, with about two loci recovered on average across 257 taxa and with 80 taxa on average represented for a given locus (2 [ubx]–246 [28S rRNA] taxa; Tables S5 and S7; Figure 3). As with the mtDNA results, most recovered sequences were less than 50% of the known locus length, with very few reaching between ≥50% and 1bp from the full length of the corresponding locus (Table 2 and Table S10). When looking at the cumulative results of nucDNA_ref1 and nucDNA_ref2, we found that nucDNA_ref2 had negligible impacts on our recovery measures (i.e. only one more locus recovered for three taxa), with patterns like those of nucDNA_ref1 alone (Table 2, Tables S5, S7 and S10; Figure 3).

3.5 | Phylogenetic result of a UCE +legacy data set

Our concatenated matrix—comprised of three legacy loci and Forthman et al.'s (2020) 1,000 UCEs—had a total of 329,771 sites. The proportion of parsimony-informative, uninformative and invariant sites (Table 3) were similar to those reported in Forthman et al. (2020). However, our UCEs exhibited a higher proportion of informative and variable sites than the three legacy loci combined (~37% more), with our legacy data largely comprised of invariant sites (~80%).

Our analysis resulted in a topology (Figure 4) generally congruent with Forthman et al.'s (2020) maximum-likelihood analyses. However, support was clearly impacted by the inclusion of legacy data as many nodes had weak to moderate support compared with most nodes in Forthman et al. (2020) that were highly supported (i.e. 90%–100%). Of our 37 NCBI reference taxa included in the analysis, nine were recovered as sister group to congeners that had UCE data (e.g. *Catorhintha*, *Merocoris* and *Chariesterus*), often with moderate to high support. Several other NCBI taxa were recovered in clades with members of their respective tribes (e.g. *Anoplocnemis curvipes* [Fabricius, 1781] recovered with other mictines, *Athaumastus haematicus* [Stål] recovered with other acanthocerines, and *Leptoscelis obscura* Dallas, 1852 recovered with other anisoscelines [Forthman et al.'s (2020) 'Anisoscelini Lineage 1']).

Three NCBI reference species of *Acanthocoris* were recovered within one of the polyphyletic lineages of Hypselonotini rather than with other acanthocorines. Of the *Acanthocoris* species that had UCE data, only *Acanthocoris* sp. CMF331 also had COX1 data; this sequence was substantially shorter than that of the NCBI reference taxa (189 bp vs. 600–1,534 bp, respective) and had little overlap in sequence with them. Another NCBI reference taxon, *Zicca annulata* (Burmeister, 1835), was recovered as closely related to *Hypselonotus* species rather

TABLE 2 Mitochondrial and nuclear-recovered locus and sequence information vary according to the reference file used

	mtDNA_ref1	mtDNA_ref2	mtDNA_ref3	mtDNA_ref1 + mtDNA_ref2 + mtDNA_ref3	nucDNA_ref1	nucDNA_ref2	nucDNA_ref1 + nucDNA_ref2
Average number of loci recovered	4	4	1	5	2	0	2
% Sequences <50% targeted length	56.8	57.8	83.6	64.7	84.4	66.7	84.3
% Sequences ≥50%–<100% targeted length	27.7	27.2	14.8	23.9	15.6	33.3	15.7
% Sequences ≥100% targeted length	15.5	15.0	1.6	11.4	0	0	0
Number of taxa recovering at least 1 locus	229	232	193	253	257	3	258
Number of taxa recovering complete mitogenomes (all 15 loci)	6	8	0	13	N/A	N/A	N/A
Number of taxa recovering ≥50% mitogenomes	44	45	2	72	N/A	N/A	N/A

Abbreviations: ≥50% mitogenomes, eight or more loci recovered; mtDNA_ref1, 19 coreoid mitogenomic references; mtDNA_ref2, mtDNA_ref1 supplemented with individually accessioned coreoid sequences for each mtDNA locus from NCBI; mtDNA_ref3, results here are only based on those taxa and loci that were not part of the mtDNA_ref2 reported results; nucDNA_ref2, results here are only based on those taxa and loci that were not part of the nucDNA_ref1 reported results.

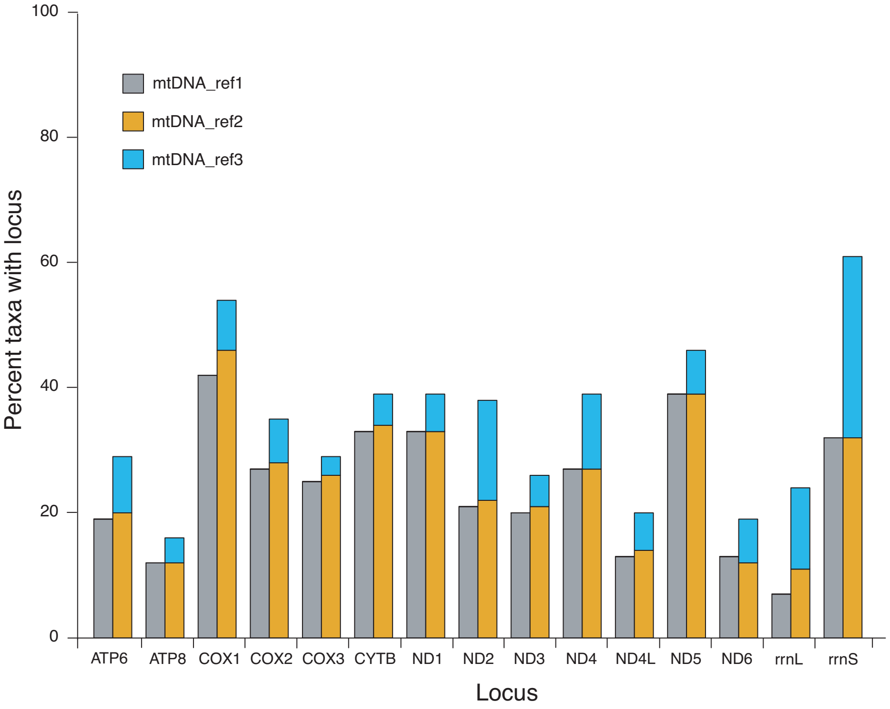

FIGURE 2 Mitochondrial legacy loci recovered. mtDNA_ref1, 19 coreoid mitogenomic references; mtDNA_ref2, mtDNA_ref1 supplemented with individually accessioned coreoid sequences for each mtDNA locus from NCBI; mtDNA_ref3, mtDNA_ref2 references + verified mtDNA_ref2 results. see Table 1 for locus abbreviations. Those sequences recovered reported as mtDNA_ref3 are new sequences not obtained with mtDNA_ref2

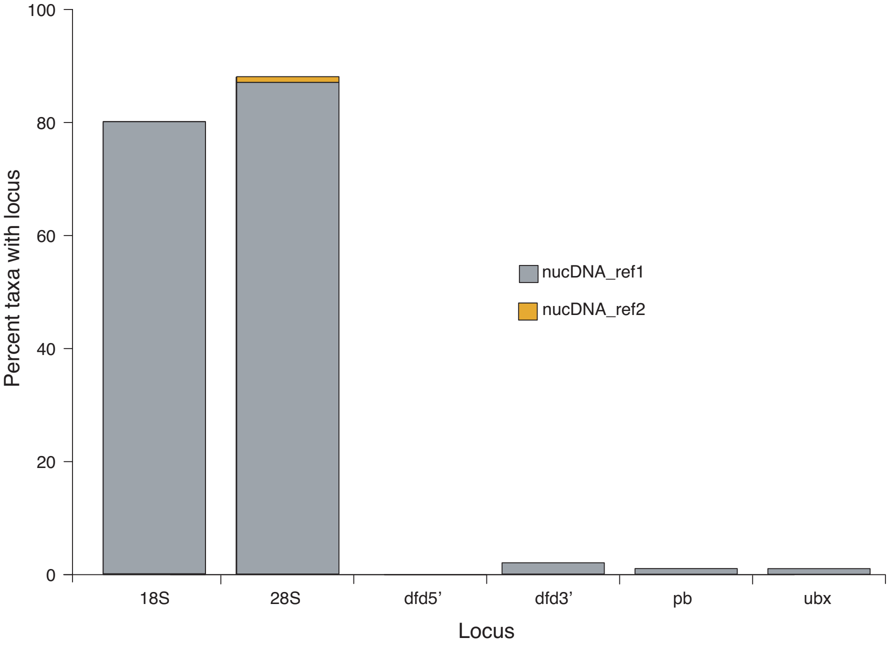

FIGURE 3 Nuclear legacy loci recovered. nucDNA_ref1, nuclear DNA reference file including sequences retrieved from the National Center for Biotechnology Information (NCBI); nucDNA_ref2, nucDNA_ref1 + verified legacy locus sequences extracted based on nucDNA_ref1; see Table 1 for locus abbreviations. Those legacy loci that were already targeted by baits (i.e. abd-A and scr) are not included here. Those sequences recovered with nucDNA_ref2 are new and not obtained with nucDNA_ref1

than to other species of *Zicca*. This species had 18S and 28S rRNA sequences, of which the former had no overlap with 18S sequences from the other *Zicca* species and the latter had relatively little overlap with the 28S sequence from *Zicca rubricator* (Fabricius, 1803).

4 | DISCUSSION

Sequence capture protocols, while effective, remain imperfect, with as little as 40% on-target recovery in some cases (e.g. Amaral et al., 2015; Asan et al., 2011; Guo et al., 2012, 2013; Samuels et al., 2013; Sulonen et al., 2011). The remaining portion of the sequence data, that is off-target sequences, is often ignored in many sequence capture studies, but several recent studies have exploited this by-catch to extract legacy loci used in past phylogenetic studies (e.g. Barrow et al., 2017; Branstetter et al., 2021; Caparroz et al., 2018; Gasc et al., 2016; Guo et al., 2012; Łukasik et al., 2019; Lyra et al., 2017; Matsuura et al., 2018; Percy et al., 2018; Simon et al., 2019; Taucce et al., 2018). Our study sought to recover legacy loci from

TABLE 3 Summary of site patterns in legacy loci (combined) and the UCE + legacy locus concatenated matrix compared with the UCE concatenated matrix in Forthman et al. (2020) (their '50p total evidence' data set)

Data set	No. of loci	Total sites	% Parsimony-informative sites	% Uninformative sites	% Invariant sites
Legacy (combined)	3	7,687	12.44	7.81	79.75
UCE + legacy	1,003	329,771	49.74	6.58	43.67
Forthman et al. (2020): '50p total evidence' UCE	1,000	322,084	50.64	6.55	42.81

Abbreviation: UCE, ultraconserved elements.

off-target sequences in a sequence capture data set using published legacy data, explore approaches to improve legacy locus recovery, and perform a phylogenetic analysis on a concatenated data set comprised of legacy loci and UCE data. Most legacy loci commonly used in past coreoid molecular phylogenetic studies were successfully recovered, and most were not already targeted by our sequence capture baits. However, there was a large amount of variation in the recovery of different types of loci (e.g. mtDNA vs. nucDNA) across taxa, as well as among loci within a locus type (e.g. COX1 vs. ATP8 for mtDNA). Our results suggest complementing an mtDNA reference file of complete mitogenomes with sequence data for individual loci available in genetic repositories does not offer much improvement for recovery of most legacy loci yet adds substantially to the time to assemble the set of reference sequences. However, adding extracted legacy sequences to mtDNA reference files can modestly improve recovery for some legacy loci (though this should be done with caution). There was little benefit to including extracted loci for nucDNA loci. We further found that legacy data can be integrated with sequence capture data to increase informative sites available for analysis and taxon representation throughout the phylogeny; however, this has some challenges and can impact topology and/or support. The approaches outlined here can be a potentially cost-effective alternative to designing legacy locus baits to combine with sequence capture bait kits.

The variation seen in the recovery of legacy loci among taxa may be due to a combination of the type of legacy locus extracted, the number of species represented by legacy sequences and taxon sampling of reference sequences. Some loci, such as mtDNA (Bogenhagen & Clayton, 1974; Samuels et al., 2013) or ribosomal (rRNA) (Eickbush & Eickbush, 2007) loci, have a high copy number within cells and may comprise a relatively large proportion of off-target sequences in sequence capture data (Branstetter et al., 2017, 2021; Simon et al., 2019). In contrast, nuclear protein-coding loci generally have comparably lower copy numbers. Thus, it may be expected that mtDNA and rRNA legacy loci are recovered for a greater proportion of taxa than nuclear protein-coding loci. We observed a pattern supporting this expectation, with far fewer taxa (<10%) having nuclear protein-coding legacy loci extracted from off-target data compared with mtDNA and rRNA loci (~10%–90% of taxa).

The number of reference legacy sequences may also explain some of the observed variations in locus recovery. The bioinformatic resources employed in this study to retrieve legacy data utilized publicly accessible DNA sequences as references. In general, the amount of mtDNA references available for coreoids far exceeds that of nucDNA references. Furthermore, certain loci have far more available reference sequences than other loci (e.g. mtDNA locus COX1 and rRNA loci 18S and 28S). The skewed distribution in the abundance of certain types of DNA loci available in repositories is common in other taxonomic groups due to Sanger-based sequencing of few well-known loci in past molecular studies (e.g. Brower & Desalle, 1994; Caterino et al., 2000; Kjer et al., 2016; McDonagh et al., 2016; Shao & Barker, 2007; Weirauch & Schuh, 2011; Zhang & Hewitt, 1997). Thus, this skew in available reference data, at least for coreoids, likely affects the recovery of legacy loci, as we had greater recovery occurring for loci with greater numbers of sequences in reference data sets.

A third contributing factor that may affect legacy locus recovery from off-target capture data is taxon sampling between reference and focal taxa, which can be related to the number of reference sequences available and the conservation of a sequence. For example, conserved loci (e.g. rrnS) are recovered more commonly than some of the other mitochondrial protein-coding regions. In genetic repositories such as NCBI, taxon sampling for a given locus may still be sparse across a broad taxonomic range, even if there are many sequences for that locus, which may account for the variation in legacy locus recovery seen in this study. In general, the number of reference sequences available for the nuclear protein-coding loci targeted in this study included far fewer taxa that may have been too distantly related from most taxa within the sequence capture data set. Increased taxon sampling can introduce more closely related species, which may improve recovery of legacy loci, specifically those taxa whose recovery was

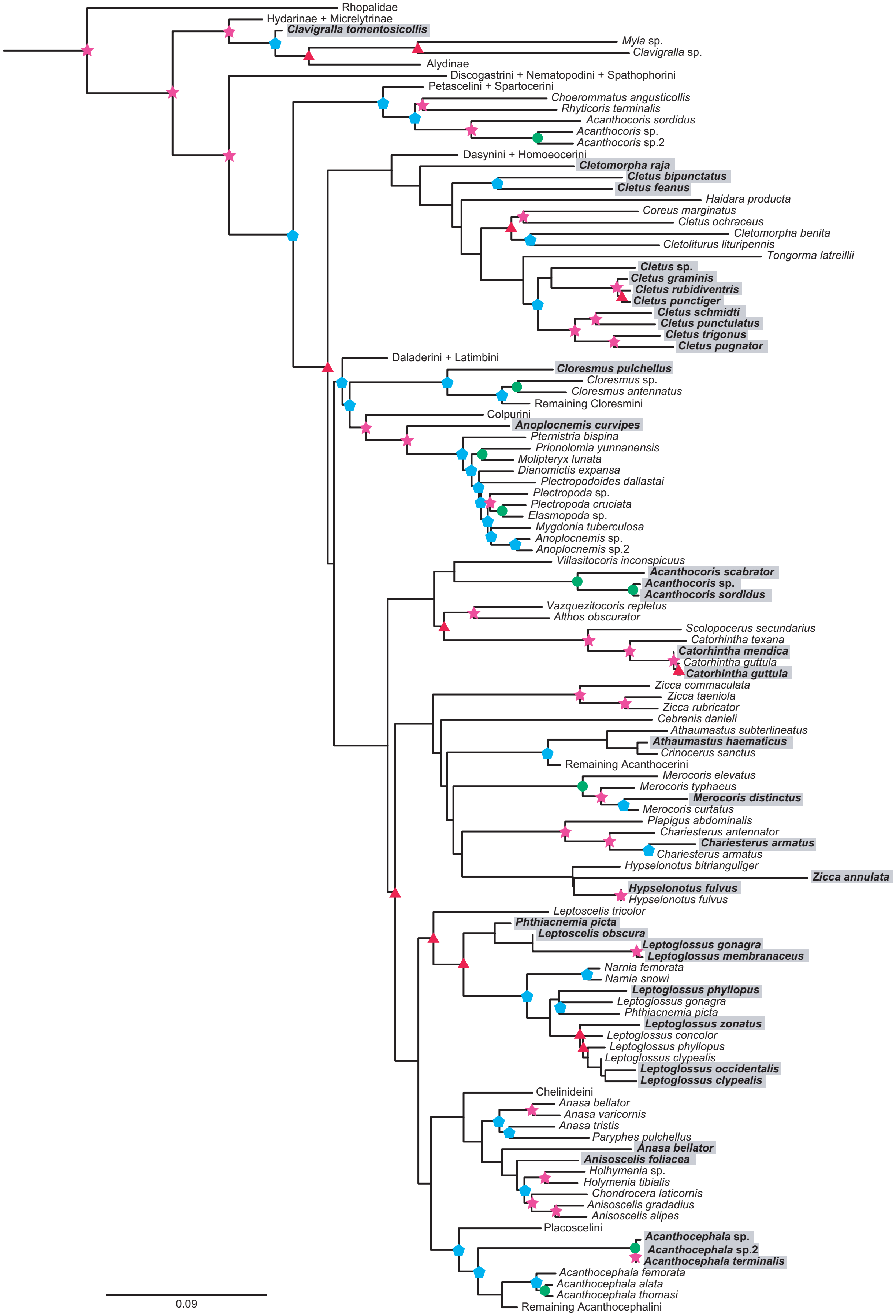
Rhopalidae
Hydarinae + Micrelytrinae
Clavigralla tomentosicollis
Myla sp.
Clavigralla sp.
Alydinae
Discogastrini + Nematopodini + Spathophorini
Petascelini + Spartocerini
Choerommatus angusticollis
Rhyticoris terminalis
Acanthocoris sordidus
Acanthocoris sp.
Acanthocoris sp.2
Dasynini + Homoeocerini
Cletomorpha raja
Cletus bipunctatus
Cletus feanus
Haidara producta
Coreus marginatus
Cletus ochraceus
Cletomorpha benita
Cletoliturus lituripennis
Tongorma latreillii
Cletus sp.
Cletus graminis
Cletus rubidiventris
Cletus punctiger
Cletus schmidti
Cletus punctulatus
Cletus trigonus
Cletus pugnator
Daladerini + Latimbini
Cloresmus pulchellus
Cloresmus sp.
Cloresmus antennatus
Remaining Cloresmini
Colpurini
Anoplocnemis curvipes
Pternistria bispina
Prionolomia yunnanensis
Molipteryx lunata
Dianomictis expansa
Plectropodoides dallastai
Plectropoda sp.
Plectropoda cruciata
Elasmopoda sp.
Mygdonia tuberculosa
Anoplocnemis sp.
Anoplocnemis sp.2
Villasitocoris inconspicuus
Acanthocoris scabrator
Acanthocoris sp.
Acanthocoris sordidus
Vazquezitocoris repletus
Althos obscurator
Scolopocerus secundarius
Catorhintha texana
Catorhintha mendica
Catorhintha guttula
Catorhintha guttula
Zicca commaculata
Zicca taeniola
Zicca rubricator
Cebrenis danieli
Athaumastus subterlineatus
Athaumastus haematicus
Crinocerus sanctus
Remaining Acanthocerini
Merocoris elevatus
Merocoris typhaeus
Merocoris distinctus
Merocoris curtatus
Plapigus abdominalis
Chariesterus antennator
Chariesterus armatus
Chariesterus armatus
Hypselonotus bitrianguliger
Zicca annulata
Hypselonotus fulvus
Hypselonotus fulvus
Leptoscelis tricolor
Phthiacnemia picta
Leptoscelis obscura
Leptoglossus gonagra
Leptoglossus membranaceus
Narnia femorata
Narnia snowi
Leptoglossus phyllopus
Leptoglossus gonagra
Phthiacnemia picta
Leptoglossus zonatus
Leptoglossus concolor
Leptoglossus phyllopus
Leptoglossus clypealis
Leptoglossus occidentalis
Leptoglossus clypealis
Chelinideini
Anasa bellator
Anasa varicornis
Anasa tristis
Paryphes pulchellus
Anasa bellator
Anisoscelis foliacea
Holhymenia sp.
Holymenia tibialis
Chondrocera laticornis
Anisoscelis gradadius
Anisoscelis alipes
Placoscelini
Acanthocephala sp.
Acanthocephala sp.2
Acanthocephala terminalis
Acanthocephala femorata
Acanthocephala alata
Acanthocephala thomasi
Remaining Acanthocephalini
0.09

FIGURE 4 Maximum-likelihood best tree generated from a concatenated matrix of three legacy loci (COX1 mtDNA, 18S rRNA and 28S rRNA) and 1,000 ultraconserved element loci (from Forthman et al. (2020)) (noncoreoid outgroups pruned and some clades without NCBI reference taxa collapsed to a single branch that is indicated by higher taxonomic name[s] for visualization; see Data S2 for unmodified tree). NCBI reference taxa indicated in bolded text on a grey background. Bootstrap support at nodes indicated by the following symbols: green circles = 100%, pink stars = 90%–99%, blue pentagon = 70%–89%, red triangle = 50%–69%, and no asterisk = <50%

worse for these loci. Future Sanger-based phylogenetic studies using legacy loci and mitogenome sequencing will likely continue to increase taxon sampling in genetic repositories. However, future sequence capture studies may alternatively improve legacy locus recovery from off-target sequences by sequencing, for example the complete mitogenome from one or a few representatives of clades with taxa exhibiting poor recovery.

Other factors, such as sample quality (e.g. fresh or degraded material) or differences in molecular benchwork protocols, could also explain the variation in legacy locus recovery observed in the present study. We further examined this possibility by comparing locus recovery with sample quality, tissues sampled, DNA extraction protocol, target enrichment protocol, etc. Our results did not exhibit any discernible patterns between these factors and legacy locus recovery (Table S11), indicating that differences in sample quality or molecular protocols do not explain the variation observed in this study. However, other modifications to these or the use of different molecular protocols, as well as the quality of tissues sampled, may impact legacy locus recovery in other studies and should be explored more.

In considering the factors above, we explored approaches to improve legacy locus recovery across taxa by expanding the number of reference sequences to search for loci in off-target capture data. Concerning mtDNA, it may be beneficial to utilize both mitogenomes and single mtDNA locus sequences from genetic repositories rather than mitogenomes alone (i.e. our mtDNA_ref2 vs. mtDNA_ref1, respectively); while legacy locus recovery did not change for many focal taxa, some taxa had slight improvements (up to two more loci) when mitogenomic sequences were complemented with single mtDNA locus sequences from NCBI. Our finding may be due to the increase in reference sequence data, which may allow for the inclusion of more closely related reference taxa relative to the focal taxa (Figure 1). These more closely related taxa can counter the effects of the elevated substitution rates of mtDNA that would otherwise make it difficult to identify loci among distantly related species (Bernt et al., 2013; Kumar, 1996; Meiklejohn et al., 2014; Rota-Stabelli et al., 2010). However, the lack of considerable improvement for the majority of taxa in our mtDNA legacy data set suggests that either the amount of reference mitogenomic sequences or taxonomic representation was enough to capture any sequenced legacy loci or that such loci may not have been identified due to a lack of closely related sequences.

Our findings from the initial searches for both mtDNA and nucDNA led us to explore whether legacy locus recovery could be further improved among taxa if our reference sequences were complemented with legacy sequences extracted from focal taxa. We expected that the inclusion of mtDNA and nucDNA sequences extracted from capture data would improve recovery for those taxa that did not have sequences matching our references. The addition of extracted mtDNA legacy locus sequences supported our expectations; however, after screening these results, it became clear that many of these newly extracted sequences were incorrectly identified as legacy loci (e.g. 37% across all mtDNA). However, we did not observe a similar pattern for nucDNA loci as only 9% of putative matches were removed. Thus, our results highlight that the addition of extracted sequences can improve legacy locus recovery, but that there can be a substantial amount of error that requires filtering before downstream analyses. Unless extensive postrecovery filtering is performed, it may be safest to exclude extracted sequences.

We found that our sequence capture baits already targeted the nuclear protein-coding legacy loci abd-A and scr. In sequence capture studies that extract legacy loci from off-target sequences, there is no indication whether similar steps were taken to verify that legacy loci were not already targeted by baits (e.g. Derkarabetian et al., 2019). Verifying whether legacy loci are targeted by capture baits is critical before extracting them from off-target data. Failure to do so may result in legacy loci that are recovered as both targeted loci and off-target loci (i.e. they are recovered twice) and weighted relative to other loci in phylogenomic analyses.

Integrating legacy data with our UCE sequence capture data generally supported previous hypotheses based on phylogenomic data. This was likely due to the much larger amount of data (i.e. 1,000 UCE loci) shared among the majority of taxa (124 taxa) included in the analysis. However, the greater amount of missing data due to the inclusion of NCBI reference taxa (i.e. taxa without any UCE data) was likely responsible for the widespread decrease in support across many nodes in our tree. Regardless, extracting legacy loci from our focal taxa allowed us to successfully place some NCBI reference taxa with their congeners or with other members of the same tribe or lineage according to Forthman et al.'s (2020) hypothesis. Only a few NCBI taxa were recovered in doubtful phylogenetic positions, possibly due to little overlap in legacy locus sequences with other congeners or members of their tribes. Given our approach extracts legacy data from off-target sequences for our focal taxa, variation in sequence length and distribution among these taxa is expected. As such, for some legacy loci, there may be little overlap with respect to taxonomic coverage and/or among the individual sequences due to variation in length among focal and reference taxa, which can impact phylogenetic inference if there are little data to group congeners or conspecifics together. Thus, utilizing legacy locus data to increase the number of informative sites and/or taxonomic sampling for analysis should only be effective if there is enough sequence overlap between reference and focal taxa, which may be expected when using commonly sampled legacy loci.

However, other potential issues could explain the questionable placement of some of the NCBI reference taxa. Perhaps one obvious issue is the potential for inaccurate species or sequence identities of NCBI reference taxa. While we screened our NCBI reference data to assess the validity of sequence identities and found several instances where inaccurate gene assignments were made (i.e. *L. occidentalis* CYTB sequences mistakenly listed as COX1 sequences), we could not verify taxonomic assignments in the absence of physical voucher specimens; thus, some of the reference taxa used in our analysis could have been misidentified by submitters. Another factor that could have affected the phylogenetic placement of the these NCBI reference taxa is the presence of mitochondrial pseudogenes in nuclear DNA (numts) that are similar in identity to COX1. Numts can be preferentially amplified with polymerase chain reaction if suboptimal primers are used (i.e. mtDNA sequence divergent enough from the primers used) or when numts occur in greater copy numbers than the mtDNA sequence (Bensasson et al., 2000, 2001; Collura & Stewart, 1995). Even if the reference mtDNA sequences are accurately identified, the placement of some reference taxa may still be the result of discordance between mitochondrial and nuclear loci, particularly if these taxa do not have any nuclear data (the predominant type of sequence data in our study) for analysis. One advantage of combining legacy data with our UCEs was the ability to identify some samples that may be problematic and require further exploration before use in other phylogenetic studies. Despite some of these challenges and potential issues, our results generally show that legacy data can be integrated with sequence capture data successfully with positive results after careful screening of legacy data (i.e. sequences and taxa) obtained from genetic depositories.

5 | CONCLUSION

This study assessed the ability to recover legacy data of interest from off-target sequences of a sequence capture data set in leaf-footed bugs. The results of this study showed the successful recovery of some mitochondrial and nuclear legacy loci, but this recovery varied greatly across taxa and is likely dependent on available genetic resources, the type of legacy loci targeted, taxon sampling and analytical approaches. Additionally, screening of legacy loci against sequences targeted by baits is critical for quality control. Our results demonstrate the benefit of integrating legacy locus data with sequence capture data, while also highlighting some of the potential issues (e.g. sample identity, sequence overlap, numts, mitonuclear discordance) that can arise from such an exercise. Overall, the approaches employed in this study emphasize how using existing public resources and data sets may reduce costs and circumvent limitations of other commonly used methods for targeting legacy loci, as well as benefit phylogenetic investigations.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their helpful feedback on our manuscript and the following individuals and institutions that contributed specimens used in this study: Christiane Weirauch, Dimitri Forero, Ummat Somjee, Zach Emberts, Takahisa Miyatake, Jason Cryan, Mark Deyrup, Wei Song Hwang, Li You, Oliver Keller, John Leavengood, Petr Kment, Joe Eger, Martin Steinbauser, Marcos Roca-Cusachs, Sam Noble Oklahoma Museum of Natural History, Field Museum of Natural History, Florida State Collection of Arthropods, California Academy of Sciences and California State Collection of Arthropods. Nathan Friedman assisted with molecular protocols. Gavin Naylor and Shannon Corrigan (University of Florida) allowed the use of a Covaris ultrasonicator. Collection of Florida (USA) specimens was permitted by the Florida Park Service and Division of Recreation and Parks (permits #05051610 and #09011720). Cebisile N. Magagula (University of Swaziland) and the Savannah Research Center facilitated the collection of specimens in the Kingdom of eSwatini. Collection of South African specimens was permitted by the Ezemvelo KZN Wildlife (permit #OP 172/2018) and iSimangaliso Wetland Park Authority (World Heritage Site). Specimens from Thailand were acquired under NSF DEB 0542864 (awarded to Michael Sharkey and Brian Brown). The collection of Singapore specimens (National Park permit #NP/RP17-012) was supported by the Lee Kong Chian Natural History Museum fellowship and the Nation Science Foundation grant OISE-1614015 (awarded to Zachary Emberts). The collection of Australian specimens (Regulation 17 permit # 01-000204-1) was supported by a University of Florida

Research Abroad for Doctoral Students (awarded to Zachary Emberts). The collection of additional samples was supported by funding awarded to Ummat Somjee: Society of the Study of Evolution Rosemary Grant Award; Smithsonian Tropical Research Institute Short Term Fellowship; and Systematics, Evolution, and Biodiversity Endowment Award (Entomological Society of America). We thank the University of Florida's University Scholars Program (awarded to C.D. Miller) for their support. This study was funded by NSF IOS-1553100 (awarded to C.W. Miller).

DATA AVAILABILITY STATEMENT

Legacy locus sequences extracted in this study are deposited in GenBank and FigShare (see Table S12 for accessions and additional information on sequences deposited in FigShare).

ORCID

Michael Forthman https://orcid.org/0000-0002-6987-8503

REFERENCES

Allard, M. W., Luo, Y., Strain, E., Li, C., Keys, C. E., Son, I., Stones, R., Musser, S. M., & Brown, E. W. (2012). High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. *BMC Genomics*, *13*(1), 32. https://doi.org/10.1186/1471-2164-13-32

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892–902. https://doi.org/10.1111/1755-0998.13160

Amaral, F. R. D., Neves, L. G., Resende, M. F. R., Mobili, F., Miyaki, C. Y., Pellegrino, K. C. M., & Biondo, C. (2015). Ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PLoS One*, *10*(9), e0138446. https://doi.org/10.1371/journal.pone.0138446

Asan, Xu, Y. U., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., Wang, J., Wu, M., Liu, X., Tian, G., Wang, J., Wang, J., Yang, H., & Zhang, X. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biology*, *12*(9), R95. https://doi.org/10.1186/gb-2011-12-9-r95

Banho, C. A. (2016). *Caracterização filogenética de percevejos terrestres das famílias Coreidae e Pentatomidae (Heteroptera: Pentatomomorpha) por meio de marcadores moleculares*. Master thesis, Universidade Estadual Paulista. Repositório Institucional UNESP. https://repositorio.unesp.br/handle/11449/136193

Barrow, L. N., Soto-Centeno, J. A., Warwick, A. R., Lemmon, A. R., & Lemmon, E. M. (2017). Evaluating hypotheses of expansion from refugia through comparative phylogeography of southeastern Coastal Plain amphibians. *Journal of Biogeography*, *44*(12), 2692–2705. https://doi.org/10.1111/jbi.13069

Bensasson, D., Zhang, D.-X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology and Evolution*, *16*(6), 314–321. https://doi.org/10.1016/S0169-5347(01)02151-6

Bensasson, D., Zhang, D.-X., & Hewitt, G. M. (2000). Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Molecular Biology and Evolution*, *17*(3), 406–415. https://doi.org/10.1093/oxfordjournals.molbev.a026320

Bernt, M., Braband, A., Schierwater, B., & Stadler, P. F. (2013). Genetic aspects of mitochondrial genome evolution. *Molecular Phylogenetics and Evolution*, *69*(2), 328–338. https://doi.org/10.1016/j.ympev.2012.10.020

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., & Moritz, C. (2013). Unlocking the vault: Next-generation museum population genomics. *Molecular Ecology*, *22*(24), 6018–6032. https://doi.org/10.1111/mec.12516

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 Evolutionary Biology*, *15*(1), 271. https://doi.org/10.1186/s12862-015-0552-5

Bogenhagen, D., & Clayton, D. A. (1974). The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. *Journal of Biological Chemistry*, *249*(24), 7991–7995. https://doi.org/10.1016/S0021-9258(19)42063-2

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768–776. https://doi.org/10.1111/2041-210x.12742

Branstetter, M. G., Müller, A., Griswold, T. L., Orr, M. C., & Zhu, C. (2021). Ultraconserved element phylogenomics and biogeography of the agriculturally important mason bee subgenus *Osima* (*Osima*). *Systematic Entomology*, *46*(2), 453–472. https://doi.org/10.1111/syen.12470

Brower, A. V., & Desalle, R. (1994). Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. *Annals of the Entomological Society of America*, *87*(6), 702–716. https://doi.org/10.1093/aesa/87.6.702

Cameron, S. L. (2014). Insect mitochondrial genomics: Implications for evolution and phylogeny. *Annual Review of Entomology*, *59*(1), 95–117. https://doi.org/10.1146/annurev-ento-011613-162007

Caparroz, R., Rocha, A. V., Cabanne, G. S., Tubaro, P., Aleixo, A., Lemmon, E. M., & Lemmon, A. R. (2018). Mitogenomes of two neotropical bird species and the multiple independent origin of mitochondrial gene orders in Passeriformes. *Molecular Biology Reports*, *45*(3), 279–285. https://doi.org/10.1007/s11033-018-4160-5

Caterino, M. S., Cho, S., & Sperling, F. A. (2000). The current state of insect molecular systematics: A thriving tower of babel. *Annual Review of Entomology*, *45*(1), 1–54. https://doi.org/10.1146/annurev.ento.45.1.1

Collura, R. V., & Stewart, C.-B. (1995). Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature*, *378*, 485–489. https://doi.org/10.1038/378485a0

CoreoideaSF Team (2021). *Coreoidea Species File Online*. Version, 5.0/5.0. http://coreoidea.speciesfile.org

de Souza, H. V. (2013). *Relacionamento filogenético de espécies pertencentes às famílias Coreidae e Pentatomidae (Heteroptera) a partir dos genes mitocondriais COI, 16S e nuclear 18S*. Doctoral dissertation, Universidade Estadual Paulista. Repositório Institucional UNESP. https://repositorio.unesp.br/handle/11449/102757

DeLeo, D. M., Pérez-Moreno, J. L., Vázquez-Miranda, H., & Bracken-Grissom, H. D. (2018). RNA profile diversity across Arthropoda: Guidelines, methodological artifacts, and expected outcomes. *Biology Methods and Protocols*, *3*(1), bpy012. https://doi.org/10.1093/biomethods/bpy012

Derkarabetian, S., Benavides, L. R., & Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Molecular Ecology Resources*, *19*(6), 1531–1544. https://doi.org/10.1111/1755-0998.13072

Eickbush, T. H., & Eickbush, D. G. (2007). Finely orchestrated movements: Evolution of the ribosomal RNA genes. *Genetics*, *175*(2), 477–485. https://doi.org/10.1534/genetics.107.071399

Emberts, Z., St. Mary, C. M., Howard, C. C., Forthman, M., Bateman, P. W., Somjee, U., Hwang, W. S., Li, D., Kimball, R. T., & Miller, C. W. (2020). The evolution of autotomy in leaf-footed bugs. *Evolution*, *74*(5), 897–910. https://doi.org/10.1111/evo.13948

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. https://doi.org/10.1093/bioinformatics/btv646

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. *Systematic Biology*, *61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

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*(6), e65923. https://doi.org/10.1371/journal.pone.0065923

Forthman, M., Miller, C. W., & Kimball, R. T. (2019). Phylogenomic analysis suggests Coreidae and Alydidae (Hemiptera: Heteroptera) are not monophyletic. *Zoologica Scripta*, *48*(4), 520–534. https://doi.org/10.1111/zsc.12353

Forthman, M., Miller, C. W., & Kimball, R. T. (2020). Phylogenomics of the leaf-footed bug subfamily Coreinae (Hemiptera: Coreidae): Applicability of ultraconserved elements at shallower depths. *Insect Systematics and Diversity*, *4*(4), 2. https://doi.org/10.1093/isd/ixaa009

Gasc, C., Peyretaillade, E., & Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, *44*(10), 4504–4518. https://doi.org/10.1093/nar/gkw309

Guo, Y., Li, J., Li, C., Shyr, Y., & Samuels, D. C. (2013). MitoSeek: Extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, *29*(9), 1210–1211. https://doi.org/10.1093/bioinformatics/btt118

Guo, Y., Long, J., He, J., Li, C.-I., Cai, Q., Shu, X.-O., Zheng, W., & Li, C. (2012). Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, *13*(1), 194. https://doi.org/10.1186/1471-2164-13-194

Hamilton, C. A., Lemmon, A. R., Lemmon, E. M., & Bond, J. E. (2016). Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evolutionary Biology*, *16*(1), 212. https://doi.org/10.1186/s12862-016-0769-y

Hughes, L. C., Ortí, G., Saad, H., Li, C., White, W. T., Baldwin, C. C., Crandall, K. A., Arcila, D., & Betancur-R, R. (2021). Exon probe sets and bioinformatics pipelines for all levels of fish phylogenomics. *Molecular Ecology Resources*, *21*(3), 816–833. https://doi.org/10.1111/1755-0998.13287

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Kieran, T. J., Gordon, E. R. L., Forthman, M., Hoey-Chamberlain, R., Kimball, R. T., Faircloth, B. C., Weirauch, C., & Glenn, T. C. (2019). Insight from an ultraconserved element bait set designed for hemipteran phylogenetics integrated with genomic resources. *Molecular Phylogenetics and Evolution*, *130*, 297–303. https://doi.org/10.1016/j.ympev.2018.10.026

Kjer, K. M., Simon, C., Yavorskaya, M., & Beutel, R. G. (2016). Progress, pitfalls and parallel universes: A history of insect phylogenetics. *Journal of the Royal Society Interface*, *13*(121), 20160363. https://doi.org/10.1098/rsif.2016.0363

Kumar, S. (1996). Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics*, *143*(1), 537–548. https://doi.org/10.1093/genetics/143.1.537

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*(3), 772–773. https://doi.org/10.1093/molbev/msw260

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*(5), 727–744. https://doi.org/10.1093/sysbio/sys049

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822

Lesieur, V., Lombaert, E., Guillemaud, T., Courtial, B., Strong, W., Roques, A., & Auger-Rozenberg, M.-A. (2019). The rapid spread of *Leptoglossus occidentalis* in Europe: A bridgehead invasion. *Journal of Pest Science*, *92*, 189–200. https://doi.org/10.1007/s10340-018-0993-x

Li, J.-M., Deng, R.-Q., Wang, J.-W., Chen, Z.-Y., Jia, F.-L., & Wang, X.-Z. (2005). A preliminary phylogeny of the Pentatomomorpha (Hemiptera: Heteroptera) based on nuclear 18S rDNA and mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, *37*, 313–326. https://doi.org/10.1016/j.ympev.2005.07.013

Longino, J. T., & Branstetter, M. G. (2020). Phylogenomic species delimitation, taxonomy, and 'bird guide' identification for the Neotropical ant genus *Rasopone* (Hymenoptera: Formicidae). *Insect Systematics and Diversity*, *4*(2), 1. https://doi.org/10.1093/isd/ixaa004

Łukasik, P., Chong, R. A., Nazario, K., Matsuura, Y. U., Bublitz, D. A. C., Campbell, M. A., Meyer, M. C., Van Leuven, J. T., Pessacq, P., Veloso, C., Simon, C., & McCutcheon, J. P. (2019). One hundred mitochondrial genomes of cicadas. *Journal of Heredity*, *110*(2), 247–256. https://doi.org/10.1093/jhered/esy068

Lyra, M. L., Joger, U., Schulte, U., Slimani, T., Mouden, E. H., Bouazza, A., Künzel, S., Lemmon, A. R., Lemmon, E. M., & Vences, M. (2017). The mitochondrial genomes of atlas geckos (*Quedenfeldtia*): Mitogenome assembly from transcriptomes and anchored hybrid enrichment datasets. *Mitochondrial DNA Part B*, *2*(1), 356–358. https://doi.org/10.1080/23802359.2017.1339212

Maddison, W. P., & Maddison, D. R. (2019). *Mesquite: A modular system for evolutionary analysis. Version 3.61*. http://www.mesquiteproject.org

Marçais, G., Yorke, J. A., & Zimin, A. (2015). QuorUM: An error corrector for Illumina reads. *PLoS One*, *10*(6), e0130821. https://doi.org/10.1371/journal.pone.0130821

Matsuura, Y. U., Moriyama, M., Łukasik, P., Vanderpool, D., Tanahashi, M., Meng, X.-Y., McCutcheon, J. P., & Fukatsu, T. (2018). Recurrent symbiont recruitment from fungal parasites in cicadas. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(26), E5970–E5979. https://doi.org/10.1073/pnas.1803245115

McDonagh, L. M., West, H., Harrison, J. W., & Stevens, J. R. (2016). Which mitochondrial gene (if any) is best for insect phylogenetics? *Insect Systematics & Evolution*, *47*(3), 245–266. https://doi.org/10.1163/1876312x-47032142

Meiklejohn, K. A., Danielson, M. J., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2014). Incongruence among different mitochondrial regions: A case study using complete mitogenomes. *Molecular Phylogenetics and Evolution*, *78*, 314–323. https://doi.org/10.1016/j.ympev.2014.06.003

Meza-Lázaro, R. N., Poteaux, C., Bayona-Vásquez, N. J., Branstetter, M. G., & Zaldívar-Riverón, A. (2018). Extensive mitochondrial heteroplasmy in the Neotropical ants of the *Ectatomma ruidum* complex (Formicidae: Ectatomminae). *Mitochondrial DNA Part A*, *29*(8), 1203–1214. https://doi.org/10.1080/24701394.2018.1431228

National Center for Biotechnology Information (NCBI) [Internet] (1988). National Library of Medicine (US), National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., McLean, J., Lasken, R., Clingenpeel, S. R., Woyke, T., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. In M. Deng, R. Jiang, F. Sun, & X. Zhang (Eds.), *Lecture notes in computer science research in computational molecular biology* (pp. 158–170). Springer. https://doi.org/10.1007/978-3-642-37195-0_13

Parks, M., Cronn, R., & Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology*, *7*(1), 84. https://doi.org/10.1186/1741-7007-7-84

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A., Bowie, R. C. K., & Moritz, C. (2014). Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources*, *14*(5), 1000–1010. https://doi.org/10.1111/1755-0998.12249

Percy, D. M., Crampton-Platt, A., Sveinsson, S., Lemmon, A. R., Lemmon, E. M., Ouvrard, D., & Burckhardt, D. (2018). Resolving the psyllid tree of life: Phylogenomic analyses of the superfamily Psylloidea (Hemiptera). *Systematic Entomology*, *43*(4), 762–776. https://doi.org/10.1111/syen.12302

Pierce, M. P., Branstetter, M. G., & Longino, J. T. (2017). Integrative taxonomy reveals multiple cryptic species within Central American *Hylomyrma* Forel, 1912 (Hymenoptera: Formicidae). *Myrmecological News*, *25*, 131–143. https://doi.org/10.25849/myrmecol.news_025:131

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J. L., Telford, M. J., Pisani, D., Blaxter, M., & Lavrov, D. V. (2010). Ecdysozoan mitogenomics: Evidence for a common origin of the legged invertebrats, the Panarthropoda. *Genome Biology and Evolution*, *2*, 425–440. https://doi.org/10.1093/gbe/evq030

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. *Trends in Genetics*, *29*(10), 593–599. https://doi.org/10.1016/j.tig.2013.07.006

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A. III, Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*, 687–695. https://doi.org/10.1038/265687a0

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Schuster, S. C. (2007). Next-generation sequencing transforms todays biology. *Nature Methods*, *5*(1), 16–18. https://doi.org/10.1038/nmeth1156

Shao, R., & Barker, S. C. (2007). Mitochondrial genomes of parasitic arthropods: Implications for studies of population genetics and evolution. *Parasitology*, *134*(2), 153–167. https://doi.org/10.1017/s0031182006001429

Simon, C., Gordon, E. R. L., Moulds, M. S., Cole, J. A., Haji, D., Lemmon, A. R., Lemmon, E. M., Kortyna, M., Nazario, K., Wade, E. J., Meister, R. C., Goemans, G., Chiswell, S. M., Pessacq, P., Veloso, C., McCutcheon, J. P., & Łukasik, P. (2019). Off-target capture data, endosymbiont loci and morphology reveal a relict lineage that is sister to all other singing cicadas. *Biological Journal of the Linnean Society*, *128*(4), 865–886. https://doi.org/10.1093/biolinnean/blz120

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Ströher, P. R., Zarza, E., Tsai, W. L., McCormack, J. E., Feitosa, R. M., & Pie, M. R. (2016). The mitochondrial genome of *Octostruma stenognatha* and its phylogenetic implications. *Insectes Sociaux*, *64*(1), 149–154. https://doi.org/10.1007/s00040-016-0525-8

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Sulonen, A.-M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A., & Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biology*, *12*(9), R94. https://doi.org/10.1186/gb-2011-12-9-r94

Tamashiro, R. A., White, N. D., Braun, M. J., Faircloth, B. C., Braun, E. L., & Kimball, R. T. (2019). What are the roles of taxon sampling and model fit in tests of cyto-nuclear discordance using avian mitogenomic data? *Molecular Phylogenetics and Evolution*, *130*, 132–142. https://doi.org/10.1016/j.ympev.2018.10.008

Taucce, P. P., Canedo, C., Haddad, C. F., Lemmon, A. R., Lemmon, E. M., Vences, M., & Lyra, M. (2018). The mitochondrial genomes of five frog species of the Neotropical genus *Ischnocnema* (Anura: Brachycephaloidea: Brachycephalidae). *Mitochondrial DNA Part B*, *3*(2), 915–917. https://doi.org/10.1080/23802359.2018.1501312

Tautz, D., Hancock, J. M., Webb, D. A., Tautz, C., & Dover, G. A. (1988). Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Molecular Biology and Evolution*, *5*(4), 366–376. https://doi.org/10.1093/oxfordjournals.molbev.a040500

Tian, X., Xie, Q., Li, M., Gao, C., Cui, Y., Xi, L., & Bu, W. (2011). Phylogeny of pentatomomorphan bugs (Hemiptera-Heteroptera: Pentatomomorpha) based on six Hox gene fragments. *Zootaxa*, *2888*(1), 57–68. https://doi.org/10.11646/zootaxa.2888.1.5

Wang, N., Hosner, P. A., Liang, B., Braun, E. L., & Kimball, R. T. (2017). Historical relationships of three enigmatic phasianid genera (Aves: Galliformes) inferred using phylogenomic and mitogenomic data. *Molecular Phylogenetics and Evolution*, *109*, 217–225. https://doi.org/10.1016/j.ympev.2017.01.006

Weirauch, C., & Schuh, R. T. (2011). Systematics and evolution of Heteroptera: 25 years of progress. *Annual Review of Entomology*, *56*(1), 487–510. https://doi.org/10.1146/annurev-ento-120709-144833

Xie, Q., Yu, S., Wang, Y., Rédei, D., & Bu, W. (2013). Secondary structure models of 18S and 28S rRNAs of the true bugs based on complete rDNA sequences of *Eurydema maracandica* Oshanin, 1871 (Heteroptera, Pentatomidae). *ZooKeys*, *319*, 363–377. https://doi.org/10.3897/zookeys.319.4178

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L., Heimes, P., Kaplan, M., & McCormack, J. E. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (Sarcohyla; Hylidae). *PeerJ*, *6*, e6045. https://doi.org/10.7717/peerj.6045

Zhang, D., & Hewitt, G. M. (1997). Assessment of the universality and utility of a set of conserved mitochondrial COI primers in insects. *Insect Molecular Biology*, *6*(2), 143–150. https://doi.org/10.1111/j.1365-2583.1997.tb00082.x

Zhang, Y., Deng, S., Liang, D., & Zhang, P. (2019). Sequence capture across large phylogenetic scales by using pooled PCR-generated baits: A case study of Lepidoptera. *Molecular Ecology Resources*, *19*(4), 1037–1051. https://doi.org/10.1111/1755-0998.13026

Zhang, Y. M., Williams, J. L., & Lucky, A. (2019). Understanding UCEs: A comprehensive primer on using ultraconserved elements for arthropod phylogenomics. *Insect Systematics and Diversity*, *3*(5), 3. https://doi.org/10.1093/isd/ixz016

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Miller, C. D., Forthman, M., Miller, C. W., & Kimball, R. T. (2021). Extracting 'legacy loci' from an invertebrate sequence capture data set. *Zoologica Scripta*, 00, 1–18. https://doi.org/10.1111/zsc.12513