Title: Improved target capture with lower hybridization temperatures for invertebrate loci with different baiting strategies: a case study of the leaf-footed bugs and allies (Hemiptera: Coreoidea)

Running title: Improved target capture in leaf-footed bugs

Michael Forthman[1,2*], Eric R. L. Gordon[3], Rebecca T. Kimball[4]

[1] California State Collection of Arthropods, Plant Pest Diagnostics Branch, California Department of Food and Agriculture, 3294 Meadowview Road, Sacramento, CA 95832, USA

[2] Entomology & Nematology Department, University of Florida, 1881 Natural Area Drive, Gainesville, FL 32611, USA

[3] Department of Ecology & Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA

[4] Department of Biology, University of Florida, 876 Newell Drive, Gainesville, FL 32611, USA

*Corresponding author: California State Collection of Arthropods, Plant Pest Diagnostics Branch, California Department of Food and Agriculture, 3294 Meadowview Road, Sacramento, CA 95832, USA; e-mail: michael.forthman@cdfa.ca.gov

Abstract

Target capture approaches are widely used in phylogenomic studies, yet few experimental comparisons of critical parameters, e.g. hybridization temperature, have been published. Even fewer studies have focused on invertebrates where bait-target divergences may be high. Most capture studies use a fixed hybridization temperature of 65°C to maximize the proportion of on-target data, but lower temperatures, which might improve locus recovery, are not commonly employed. We used fresh and degraded specimens of leaf-footed bugs and relatives (Hemiptera: Coreoidea) to investigate the effect of hybridization temperature on capture success of previously published ultraconserved elements (UCE) targeted by baits derived from divergent hemipteran genomes and other loci targeted by newly designed baits derived from less divergent coreoid transcriptomes. We found touchdown capture approaches with lower hybridization temperatures generally resulted in lower proportions of on-target reads and lower coverage but were associated with more assembled contigs and improved recovery of targeted UCE loci. Low hybridization temperatures were also associated with increased numbers of putative paralogs of targeted UCE loci and recovery of well-known loci (from off-target reads) with historical uses in Sanger-based molecular phylogenetic studies. Hybridization temperatures did not generally affect recovery of newly targeted loci, which we attributed to lower bait-target divergences (compared to higher divergences between UCE baits and targets) and greater bait tiling density. Thus, optimizing *in vitro* target capture conditions to accommodate low hybridization temperatures can provide a cost-effective, widely applicable solution to improve recovery of protein-coding loci in invertebrates, while retrieving other potentially useful data for downstream comparative analyses.

Keywords: Annealing temperature, Coreoidea, Pentatomomorpha, target capture, touchdown hybridization

## 1. Introduction

Many biological disciplines have witnessed a dramatic increase in the amount of genomic data sampled due to recent advances and declining cost of next-generation sequencing technologies. While whole genome sequencing may be cost-effective for some research questions, genome reduction approaches that enrich for genomic regions of interest prior to high-throughput sequencing are often more feasible (reviewed in Lemmon and Lemmon, 2013). One of the most frequently employed genome reduction techniques in phylogenomic studies are target capture approaches, which include exon capture, anchored hybrid enrichment (AHE), and capture of ultraconserved elements (UCEs) (e.g., Bi et al., 2012; Faircloth et al., 2012; Lemmon et al., 2012; Li et al., 2013). In general, target capture leverages existing genomic resources to synthesize short (60–120 bp) nucleotide bait sequences complementary to genomic regions of interest. Baits are then hybridized to DNA libraries, and unbound DNA (i.e., non- or off-targets) is removed via a series of washing steps prior to sequencing.

Several recent studies have investigated experimental conditions that may affect the success of *in vitro* target capture approaches, including, e.g., GC content and tiling of baits, starting amount of DNA or baits, bait-target divergence, and washing stringency (e.g., Ávila-Arcos et al., 2011; Li et al., 2013; Cruz-Dávalos et al., 2017). Another condition that can affect capture success is hybridization temperature during bait-target annealing. Many capture studies have employed low hybridization temperatures (e.g., 50°C) to improve capture success, particularly for divergent DNA sequences relative to capture baits (e.g., Mason et al., 2011;

Peñalba et al., 2014; Li et al., 2015). However, to our knowledge, only four experimental studies have explicitly tested the effect of hybridization temperature on capture success (Li et al., 2013; Paijmans et al., 2016; Cruz-Dávalos et al., 2017; Mohandesan et al., 2017) (Table 1). Each of these studies has compared a commonly used high hybridization temperature (fixed at 65°C) in capture experiments with lower hybridization temperatures (e.g., 50°C, either at a fixed setting or through incremental decreases [i.e., touchdown]). Results from these experiments have provided conflicting conclusions regarding the benefits of lower hybridization temperatures (fixed or touchdown) in target capture studies based on various metrics, e.g., the proportion of on-target reads, mismatch tolerance between bait and targets, and overall number of loci recovered. However, these conflicting results may be attributable to differing experimental properties among studies, such as the type of capture employed, sample quality, or type of loci targeted (Table 1).

What has not been well explored are the effects of hybridization temperatures on other aspects of capture success, such as the number of putative paralogs between different hybridization temperature conditions. Additionally, none have explored intermediate temperatures to evaluate if there is no further improvement to locus recovery or if costs outweigh the benefits at a particular temperature. Furthermore, these studies have not investigated the effects of hybridization temperatures on invertebrate capture success, where divergences may be greater in some taxa than seen in some of the vertebrate taxa investigated. Thus, given a desire to maximize recovery of loci using these approaches, further effort is warranted to understand general impacts of lower hybridization temperatures on capture success.

Invertebrate target capture studies often recover a low to moderate proportion of the targeted loci (particularly for UCE studies) (e.g., Faircloth et al., 2015; Hamilton et al., 2016; Baca et al., 2017; Van Dam et al., 2017; Dietrich et al., 2017; Kieran et al., 2019). Optimizing

existing invertebrate target capture bait sets to be more tailored to focal taxa has been shown to improve recovery (e.g., Branstetter et al., 2017; Gustafson et al., 2020), suggesting that baits may often be too divergent from some taxa to allow effective recovery. However, genomic resources that permit such optimization for many other groups are still lacking. As such, optimizing one or more *in vitro* target capture conditions may provide a more cost-effective solution to improve locus recovery. Studies in invertebrates typically employ a fixed hybridization temperature at 65°C as suggested by standard protocols (although temperatures are often not reported in AHE studies), with few studies having used lower temperatures during bait-target hybridization (Zhang et al., 2019; Braby et al., 2020; Emberts et al., 2020; Forthman et al., 2020; Miller et al.; 2022). However, no studies have experimentally investigated the effect of altering hybridization temperatures on capture success of targeted loci in invertebrates.

Lower hybridization temperatures, whether fixed or achieved through touchdown, may improve on-target and locus recovery due to relaxed specificity between baits and targets. This may be particularly advantageous if some baits are more divergent from their targets (as is commonly the case in invertebrates) or have lower optimal annealing temperatures than other baits. However, relaxing specificity to allow for partial matching between baits and targets should also increase the risk of baits to potentially hybridize with paralogous sequences exhibiting some degree of divergence from the corresponding target sequence and/or may increase the number of off-target sequences (e.g., Cruz-Dávalos et al., 2017) and thus reduce read numbers for targeted regions.

However, off-targets reads may contain sequences from loci traditionally used in phylogenetic studies (herein referred to as “legacy loci”) (e.g., Amaral et al., 2015; Wang et al., 2017; Simon et al., 2019; Miller et al., 2022). Integrating legacy loci with target capture datasets

has benefits, such as increasing the resolution power for phylogenetic inference and the inclusion of rare species with existing legacy data that are difficult to sample repeatedly (Branstetter et al., 2017; Derkarabetian et al., 2019; Zhang et al., 2019). While legacy locus data can be integrated with target capture data by designing baits from legacy loci (Branstetter et al., 2017; Simon et al., 2019; Hughes et al., 2021), this may increase the cost of custom probe kits because more baits may be required across more species due to higher substitution rates of some loci (e.g., mtDNA) and/or these baits may be included in a separate kit to prevent high copy number loci, like those on the mtDNA genome or the rRNA operon, from dominating capture data (Ströher et al., 2016; Pierce et al., 2017; Allio et al., 2020; Branstetter et al., 2021; Miller et al., 2022). Extracting legacy loci from off-target reads can circumvent some of these issues (Miller et al., 2022), and they have been successfully integrated with capture data despite their often-fewer numbers (compared to, e.g., 1000+ UCE loci) and/or introduction of large amounts of missing data (e.g., Simon et al., 2019; Miller et al., 2022). Thus, having off-target reads may not always be detrimental in capture studies when designing legacy locus baits is less desirable, as long as targeted regions are also recovered.

Here, we evaluated the impact of four different protocols that varied hybridization temperatures on invertebrate target capture success in leaf-footed bugs and allies (Hemiptera: Coreoidea) and closely related taxa using samples from various sources (fresh versus older, dried specimens) and library qualities to reflect conditions typical of empirical studies. One protocol used a fixed, standard hybridization temperature (65°C) while the remaining three protocols employed touchdown approaches with different final temperatures. Specifically, we addressed the following questions: 1) do touchdown target capture approaches with lower hybridization temperatures result in a greater total of on-target reads, total assembled contigs, total targeted

loci, and longer targeted assemblies compared to the commonly used standard hybridization temperature?; 2) does the touchdown target capture protocol with the lowest final temperature produce the most data for the variables listed in question 1?; and 3) does a lower hybridization temperature from touchdown capture protocols generate a greater proportion of off-target reads (including useful legacy loci) and/or paralogous sequences?

Our study utilized a subsampled version of an existing Hemiptera-derived UCE bait set (Faircloth, 2017; see Forthman et al., 2019), but we also introduced newly designed baits (with slightly greater tiling) derived from coreoid transcriptomes and evaluated the ability of these new baits to enrich samples *in vitro*. Given the different bait designs, we also examined whether the effects of hybridization temperature exhibited different patterns across bait design strategies; specifically, we investigated if 1) the proportion of on-target reads and coverage exhibit similar trends across hybridization temperature conditions regardless of bait design strategy, 2) the capture of loci with greater divergences from baits shows the greatest improvement at lower hybridization temperatures than loci with less divergence from their baits, 3) the number of putative paralogous loci increase as hybridization temperatures decrease regardless of bait design strategy, and 4) an increase in bait tiling improves coverage of captured loci.

## 2. Material and methods

### *2.1. Sample material*

Our target capture experiment was performed on 39 taxa (36 species of Coreoidea, 3 outgroup taxa), of which 30 were ethanol, frozen, or silica bead (“fresh”) preserved samples (collected 2008–2017) and nine were degraded samples from pinned museum material of varying ages (1935–2017) (Table S1). All taxa had previously been subjected to target capture protocols

and sequenced prior to the start of this study (i.e., freshly preserved samples or dried preserved samples were subjected to the standard or TD-60 protocols shown in Fig. 1, respectively); 27 taxa have already been published following protocols described in Forthman et al. (2019, 2020) and Emberts et al. (2020) (see Table S1 and references therein). These 39 taxa were selected for this experiment based on the availability of extra DNA libraries for additional target captures and to include a diversity of preservation methods and specimen ages (recent/fresh vs. historical/dried), as well as library qualities (best, moderate, and marginal quality based on initial sequencing outcomes relative to other samples).

*2.2. Target capture baits*

For a list of terms and their definitions used in this study, see Table 2. A summary of bait properties from our different bait design strategies are given in Fig. 2.

We used our previously published custom myBaits kit (Forthman et al., 2019), which subsampled a Hemiptera-wide derived UCE bait set (at ~1.33x tiling) designed by Faircloth (2017) to only include two pentatomomorphan taxa that are more closely-related to but not included in our ingroup taxa (herein, referred to as “Pentatomomorpha-derived baits”; Table 2; Fig. 2). This kit also included an independently designed set of baits that were derived from coreoid transcriptomes, but these have not yet been introduced in the literature prior to this study. Thus, we introduce our coreoid bait design procedures here (herein, collectively referred to as “Coreoidea-derived baits”) and assess the effectiveness of these baits. Below, we describe two bait design strategies for our Coreoidea-derived baits, wherein baits were designed from individual exons while others were designed across entire transcripts.

For baits designed from individual coreoid exon sequences (herein, “exon-derived baits”, which is a subset of the Coreoidea-derived baits; Table 2; Fig. 2), we first retrieved an annotated draft genome of *Oncopeltus fasciatus* (Dallas, 1852) (Lygaeidae) from the Baylor College of Medicine – Human Genome Sequencing Center (https://www.hgsc.bcm.edu/arthropods/milkweed-bug-genome-project). We extracted exon sequences from the *O. fasciatus* genome using BEDTools v2.29.0 (Quinlan and Hall, 2010). Exon sequences were then filtered to exclude those that were <200 bp in length or that had a GC-content <30% or >70%.

We then obtained sequence reads for five published coreoid transcriptomes that have not been annotated: *Alydus pilosulus* Herrich-Sch□ffer, 1847 (Johnson et al., 2018); NCBI Bioproject PRJNA272214), *Anasa tristis* (De Geer, 1773) (Johnson et al., 2018; NCBI Bioproject PRJNA272215), *Anoplocnemis curvipes* (Fabricius, 1781) (Agunbiade et al., 2013); NCBI Bioproject PRJNA192258), *Boisea trivittata* (Say, 1825) (Johnson et al., 2018; NCBI Bioproject PRJNA272221), and *Clavigralla tomentosicollis* Stål, 1855 (Agunbiade et al., 2013; NCBI Bioproject PRJNA192261). Sequence reads were processed using PRINSEQ-lite v0.20.4 (Schmieder and Edwards, 2011) and QuorUM v1.1.0 (Marçais et al., 2015), as well as *de novo* assembled in Trinity (Grabherr et al., 2011), following Forthman et al. (2019). For each coreoid transcriptome, a localized reciprocal blastn search using *O. fasciatus* individual exon sequences was performed using a custom python script (e-value threshold set to 1e-20 and percent identity to 60%) (reciprocal_blast.py).

The best reciprocal hit was extracted from each transcriptome using a custom perl script (extract_rbh_to_exons.pl), with sequences from multiple transcriptomes corresponding to the same exon grouped together in a single fasta file. These exons were then searched against all

coreoid transcriptome sequences using blastn to confirm orthology and to identify additional sequences that may not have been found in reciprocal blast hits with the more distant *O. fasciatus* genome. We then used the RepeatMasker web server (https://repeatmasker.org) (options: rmblast search engine and *Drosophila melanogaster* DNA source [all other options at default]) to exclude exons with low complexity sequences and simple repeats. Sequences were then aligned with MAFFT v7.305b (Katoh et al., 2002; Katoh and Standley, 2013) using the G-INS-i algorithm, and alignments were visually inspected in Geneious v9 to further confirm orthology. Sequences for a total of 456 individual exons were then selected for part of the Coreoidea-derived bait set.

We also used a pipeline from Portik et al. (2016) to design baits across transcript sequences (i.e., RNAseq-derived sequences that may include more than one exon, with baits potentially spanning across one or more introns; herein, “transcript-derived baits”; Table 2; Fig. 2). We first used Portik et al.’s (2016) 4-Annotation.pl script (with some modifications to process our data) and amino acid sequences of the *O. fasciatus* genome to annotate the assembled coreoid transcriptomes by transcript ID (e-value threshold set to 1e-20 and percent identity to 60%). We then used their 6-MarkerSelectionTRANS.pl script with default settings to find orthologous transcript sequences across our transcriptomes. This script requires an orthologous sequence to be present across all transcriptomes to be selected for bait design. Few transcript sequences were selected when all five of our transcriptomes were used because of this requirement; in attempting this, we found that the inclusion of the *B. trivittata* transcriptome was associated with the low number of transcript sequences selected. Thus, we excluded the *B. trivittata* transcriptome from subsequent searches. Furthermore, to maximize the number of transcript sequences selected for bait design while “allowing” for missing taxa, we performed two separate searches that excluded the *Ano. curvipes* or *Ana. tristis* transcriptomes, respectively. Transcript sequences were aligned

and visually inspected as described above for exon-derived baits. Based on the annotations, 141 transcripts (out of 172 initially selected) did not correspond to any of the exon-derived baits. Of these, we selected 81 transcripts for part of the Coreoidea-derived baits that were found across most of our coreoid transcriptomes and that ranged between 256 bp and 1000 bp to reduce the number of probes potentially targeting multiple exons interspersed by long introns.

The final Coreoidea-derived bait sequences (i.e., exon- and transcript-derived baits) were submitted to Arbor Biosciences (Ann Arbor, MI) to produce 120 bp baits with ~2x tiling density. We also preliminarily compared our Coreoidea-derived baits against the Pentatomomorpha-derived baits using blastn (e-value threshold set to 1e-20) to determine whether any were potentially associated with a locus already targeted by the latter set of baits. Those Coreoidea-derived bait sequences matching to Pentatomomorpha-derived baits were not removed from the final selection of baits, as their inclusion could allow for some targeted loci to be captured by more baits from more closely-related species (herein, " Pentatomomorpha-Coreoidea [PC] dual baits"; Table 2; Fig. 2).

*2.3. DNA extraction and library preparation*

See references in Table S1 for details on library preparation and target capture for samples previously published. For new samples, genomic DNA was extracted from any part of the body or the entire body from ethanol-preserved, silica-bead preserved, frozen, or dried specimens to sample similar amounts of tissue across taxa, where possible (Table S1). Freshly preserved specimens were extracted with either the Gentra Puregene Tissue or Qiagen DNeasy Blood and Tissue kit (hereafter DNeasy) (Table S1). For the Puregene kit, we followed the manufacturer's protocol for 5–10 mg tissue and optional recommendations, but we made the following

modifications: 10 µL of proteinase K was added to samples; samples were incubated for 24–48 hr; 600 µL of 100% ethanol was used for the first wash, and the sample was then centrifuged for 10 mins; and 50–100 µL of molecular grade water or Puregene DNA Hydration Solution was used to resuspend isolated DNA. For the DNeasy kit, we also followed the manufacturer's protocol but with fewer modifications: tissue was incubated in 180–190 µL Buffer ATL and 10–20 µL proteinase K for 24–48 hr, and depending on the source of the tissue, DNA was eluted once or twice with 50 µL Buffer AE.

For degraded museum specimens, DNA was extracted using a modified version of the DNeasy protocol, following Knyshov et al. (2019) (i.e., Qiagen DNeasy Blood and Tissue kit coupled with Qiagen QIAquick PCR purification kit; hereafter DNQIA) (Table S1). The protocol is designed to extract DNA >100 bp in length. The DNQIA protocol follows the DNeasy protocol to the first centrifugation step, but a QIAquick spin column is used. The samples are then subjected to the manufacturer's Qiagen QIAquick PCR purification protocol by replacing AW1 and AW2 washes with PE buffer. Samples were then eluted in 30 µL EB buffer.

We assessed DNA quality and quantity with 1% agarose gel electrophoresis and a Qubit 2.0 fluorometer, respectively. Samples were normalized to 10–20 ng/µL. High molecular weight samples were then fragmented into 200–1000 bp using a Bioruptor UCD-300 sonication device (4–10 cycles of 30 s on/30 s) or a Covaris M220 Focused-ultrasonicator (20–60 s) (Table S1).

Libraries were constructed with a modified KAPA Hyper Prep Kit protocol following Forthman et al. (2019). Briefly, we used half volume reactions for all steps. iTru universal adapter stubs and 8 bp dual indexes were used (Glenn et al., 2016). Library amplification conditions involved initial denaturation at 98°C for 3 min; 14 cycles of 98°C for 30 s, 60°C for 30 s, and 72°C for 30 s; and a final extension at 72°C for 5 min. Amplified libraries quality and quantity

were assess with gel electrophoresis and Qubit, respectively. Libraries were then combined into 1000 ng pools using equimolar amounts, dried at 60°C, and resuspended in 14 µL IDTE.

*2.4. Target capture experimental design*

To evaluate the effects of different hybridization conditions on target capture, we compared four experimental protocols using ½ or ¼ volume baits for fresh and dried samples, which differed only in hybridization temperatures used over a 24- (common in standard target capture protocols) or 36-hour period (see Fig. 1). Due to the limited availability of extra DNA libraries, previously sequenced samples could only be assigned to one of three different target capture protocols (see Table S2). In assigning samples to treatments, we attempted to distribute sample preservation methods, sample age, and library quality (based on results of our initial sequencing that had used a constant 65°C hybridization temperature).

All post-capture protocols followed Forthman et al. (2019), with the exception that captures were washed at temperatures corresponding to the final hybridization temperature used in each capture protocol (i.e., 65°C for standard, 60°C for TD-60, etc.). All enriched library pools were combined in equimolar amounts and sequenced on an Illumina HiSeq3000 lane (2x100) at the University of Florida's Interdisciplinary Center for Biotechnology Research.

*2.5. Sequence data processing and analysis*

Unless otherwise stated, all data processing steps and analyses mentioned below used default settings. Sequence reads were demultiplexed by the sequencing facility. Adapters were trimmed with illumiprocessor (Faircloth, 2013; Bolger et al., 2014). Duplicate reads were filtered with PRINSEQ-lite. Reads were then error corrected using QuorUM and subsequently assembled

*de novo* using SPAdes v3.13.0 with the single-cell and auto coverage cutoff options invoked (Bankevich et al., 2012; Nurk et al., 2013; Prjibelski et al., 2020). PHYLUCE v1.5.0 (Faircloth, 2016) was then used to extract targeted loci from assembled contigs.

Because our preliminary comparison of loci targeted by Coreoidea-derived baits against the Pentatomomorpha-derived baits prior to bait design indicated that some loci were targeted by both sets of baits, a more thorough confirmation was performed after *in vitro* target capture. Using captured loci targeted by our Coreoidea-derived baits, we performed a tblastx (e-value threshold = 1e-10) search against those loci captured by the Pentatomomorpha-derived baits and extracted matches with ALiBaSeq (Knyshov et al., 2021). Of the loci captured by the Pentatomomorpha-derived baits, we found that 103 of these were targeted by both Pentatomomorpha- and Coreoidea-derived baits. It is worth noting that during this process, we also determined that some targeted transcript loci were also targeted by multiple, adjacent UCE loci by Pentatomomorpha-derived baits; in such cases, we treated these loci as a single locus. Thus, we had 376 loci targeted by exon-derived baits, 58 targeted by transcript-derived baits, 2566 targeted by Pentatomomorpha-derived baits, and 103 targeted by both Pentatomomorpha- and Coreoidea-derived baits (i.e., PC dual baits), resulting in a total of 3103 targeted loci.

We calculated the number and lengths of assembled contigs and captured loci using PHYLUCE. Because our Pentatomomorpha- and Coreoidea-derived baits were designed from different sets of genomes and transcriptomes of varying divergences, we also calculated the average minimum distances between our baits and captured loci for exon-derived, transcript-derived, Pentatomomorpha-derived, and PC dual baits. We then calculated coverage using our filtered reads and the total number of on-target filtered reads using BBMap v38.44 (Bushnell,

2014). We determined the number of captured loci with putative paralogs by invoking the keep-duplicates option in the PHYLUCE phyluce_assembly_match_contigs_to_probes.py script.

We extracted mitochondrial and nuclear legacy loci from off-target contigs in our target capture dataset. Briefly, we retrieved sequence data for 15 mitochondrial (13 protein-coding and two ribosomal regions) and two nuclear ribosomal loci (18S and 28S) from the National Center for Biotechnology Information's database. We used MitoFinder v1.1 (Allio et al., 2020) to extract mitochondrial sequences. To identify nuclear legacy loci, we created a local nucleotide database using BLAST for each locus of interest and queried our capture data against them using blastn (e-value set to $1 \times 10^{-50}$). We then calculated the number of legacy loci recovered.

As one part of the experiment, we also wanted to quantify the effect of different tiling strategies (~1.33x vs. ~2x tiling density) on locus recovery. As most loci with baits tiled differently also exhibited major differences in bait-target divergences (see Section 3.2.), we were not able to directly measure the effect of tiling strategy for most loci. However, some loci (i.e., 40 out of 2566) captured with Pentatomomorpha-derived baits had low average minimum bait-target divergences as seen in loci captured with Coreoidea-derived bait (see Sections 3.2. and 3.4.3.). Thus, we had a limited opportunity to explore the effect of tiling strategy on coverage while controlling for bait-target divergences. For this, we selected loci captured with Coreoidea- or Pentatomomorpha-derived baits that had divergences ranging from 0.05–0.10 and calculated coverage.

Sequencing depth between different sequence lanes/runs may affect, e.g., how many loci are recovered or the proportion of on-target reads. Furthermore, the effective sample size across different sequencing efforts can vary as some samples may fail or have poor sequencing outcomes, which can have an impact on sequencing depth across samples. Three different

sequencing efforts were performed across our samples, with each producing different effective sample sizes: 1) 99 samples combined for previously sequenced samples subjected to the standard capture conditions, 2) 96 samples combined for previously sequenced samples subjected to the TD-60 capture conditions and 3) 88 samples combined for sequenced samples subjected to the different target capture conditions conducted in this study. Thus, to investigate the influence of different sequencing depths on our metrics of capture success, we equalized sequencing depth by subsampling 2,000,000 raw reads generated under the different capture protocols for 28 taxa using Seqtk v1.3 (https://https://github.com/lh3/seqtk) (random seed [-s option] = 100); 11 taxa were not included in this analysis because at least one of the target capture protocols for each of these taxa were associated with fewer than 3,000,000 raw reads total. The subsampled reads were then processed and evaluated as described above to determine if any patterns observed in the subsampled dataset differed from what was observed in the original data.

## 3. Results

### *3.1. Raw read and assembled contig yield across target capture conditions*

Overall, lower hybridization temperatures tended to be associated with more raw sequence reads. Of the 39 samples in our dataset, 26 had more raw reads sequenced when these samples were subjected to lower hybridization temperatures in pairwise comparisons (median increase = 272%) (Table S3). On average, TD-50 generated the most raw reads (Fig. 3A), followed by TD-55, compared to the standard and TD-60 protocols. Three degraded samples that failed or nearly failed to produce any raw reads under the standard protocol had over 33,000 reads sequenced at a modestly lower hybridization temperature (TD-60). Of those samples in which their respective touchdown protocol produced fewer total raw reads than the standard, all were

fresh quality samples, most were subjected to the TD-60 protocol, and the decrease was <72% of the standard protocol (median = 48%).

In general, touchdown protocols were also associated with a greater total of assembled contigs and nucleotides, with few exceptions (Table S3). Overall, the TD-50 protocol assembled the most contigs (Fig. 3B). Contigs had similar ranges of median lengths between capture protocols, although the TD-50 and TD-55 protocols generally had longer median contig lengths than the other protocols (Fig. S1A), and data generated under protocols with lower hybridization temperatures yielded some of the longest contigs recovered (Table S3). Thus, while lower hybridization temperatures varied in their success, the lowest hybridization temperature resulted in the most raw reads and contigs, on average.

*3.2. Bait-target distances, reads on- and off-target, coverage, and locus length*

We found that loci targeted by Coreoidea exon- and transcript-derived bait sequences followed similar trends in our experiment, as well as similar average minimum bait-target divergences (Fig. S1B). Due to this and the relatively few number of loci compared to many more loci targeted by Pentatomomorpha-derived UCE baits, we combined results from the exon- and transcript-targeted loci together, which we refer to as Coreoidea-derived baits. For separate exon- and transcript targeted locus results, see Tables S4–S6 and Figs. S1 and S2.

Low average minimum bait-target divergences were observed for loci captured by Coreoidea-derived baits and PC dual baits (with less variation) compared to those captured by Pentatomomorpha-derived baits (Fig. 3C), as we expected.

In pairwise comparisons of protocols for a given taxon, lower hybridization temperatures were generally associated with increases in the total number of on-target filtered reads (Table S4).

When partitioning captured loci by the baits used (i.e., Coreoidea-derived, Pentatomomorpha-derived, and PC dual baits), this was most apparent with loci captured by Pentatomomorpha-derived baits (Table S4). However, when comparing among capture conditions, hybridization temperature did not appear to affect the number of on-target reads for loci targeted by Coreoidea- or Pentatomomorpha-derived baits (Fig. S1C). We found the highest median locus length in either the TD-50 or TD-55 protocols regardless of bait design strategy, while the TD-60 protocol was associated with the shortest median locus length for all target sequences (Fig. S1F).

Despite the general increases in the number of filtered on-target reads in pairwise comparisons as hybridization temperature was lowered, there was often a decrease in the proportion of on-target reads from the TD-55 and TD-50 capture protocols (Table S4; Figs. 3D, E). For any given target capture protocol, in general, lower hybridization temperatures noticeably reduced the percentage of on-target reads (Fig. 3D). When comparing across protocols, the TD-50 and TD-55 protocols were associated with the highest percentages of off-target reads (Fig. 3E).

In pairwise comparisons, there was frequently a decrease in coverage at lower temperatures regardless of bait design strategy (Tables 4, S4). Coverage was relatively similar between loci captured with Coreoidea-derived baits and PC dual baits, which were higher than those captured using Pentatomomorpha-derived baits and did not appear to be generally affected by hybridization temperatures (Table S4, Fig. 3F).

*3.3. Targeted locus recovery*

Pairwise comparisons within each of our 39 samples showed that lower hybridization temperatures were generally associated with a greater percentage of loci and longer median

sequence lengths captured compared to the standard temperature and longer median sequence lengths (Tables S5–S8). Comparison of capture protocols across all samples generally showed little to no improvements in locus recovery or median locus length across hybridization temperatures when looking at Coreoidea-derived and PC dual bait design strategies (Figs. 4A, 4B, S1F). However, lower hybridization temperatures generally resulted in improved recovery and slightly longer median locus lengths for Pentatomomorpha-baited loci (Figs. 4C, S1F).

In pairwise comparisons and comparison target capture conditions across all samples, we did not observe any patterns with respect to hybridization temperatures and the number of putative paralogs detected for Coreoidea-derived and PC dual baited strategies (Tables S5–S7; Figs. 4D). In contrast, we observed slight to moderate increases in putative paralogs at lower temperatures for the Pentatomomorpha-derived bait strategy (Table S8; Fig. 4D).

### *3.4. Additional measures and parameters investigated*

#### *3.4.1. Tissue quality*

We observed similar trends across hybridization temperatures for the number of raw reads generated, total number of assembled contigs, proportion of on-target reads, number of putative paralogs of loci captured by Coreoidea-derived and PC dual baits, and number of legacy loci recovered even when partitioning the data by tissue quality (i.e., preserved fresh or dried) (see Tables S2–S9; Figs. S3–S6). Freshly preserved samples also showed similar trends for most other measured variables (see Tables S2–S9, Figs. S3–S6) as seen when the data is not partitioned by tissue quality (i.e., Figs. 3–5).

We found that all samples preserved dry and subjected to the standard protocol produced very little, if any, data for analysis. Sequencing these degraded samples after target capture at low

hybridization temperatures substantially rectified this issue, even at very modest temperature reductions (i.e., ≤60°C) (Tables S2–S9); more raw reads and assembled contigs were obtained when these degraded samples were subjected to low hybridization temperatures. However, unlike freshly preserved samples, we observed no apparent trends for preserved, dried samples with respect to overall coverage, locus recovery, locus length, or number of putative paralogous loci (see Tables S2–S9, Figs. S3–S6).

*3.4.2. Library quality*

When partitioning the data by library quality (i.e., best, moderate, or marginal) (see Tables S2–S9, Figs. S7–S10), we observed trends similar to those seen for all the variables measured when the data is not partitioned (Figs. 3–5).

*3.4.3. Tiling strategies*

We quantified the effect of different tiling strategies on one aspect of locus recovery: coverage. Average coverage per locus appeared to differ when considering tiling strategy. For loci captured with Coreoidea- or Pentatomomorpha-derived baits (~2x and ~1.33x tiling depth, respectively) and that exhibited similar average minimum bait-target divergences (i.e., 0.05–0.10 divergence), we found that loci captured by Coreoidea-derived baits at ~2x tiling had greater coverage than Pentatomomorpha-derived baited loci at 1.33x tiling (Fig. S1E).

*3.4.4. Legacy loci*

When comparing overall legacy locus recovery across target capture protocols, as well as in pairwise comparisons of protocols for each taxon, lower hybridization temperatures improved

(or had no negative effect on) legacy locus recovery from off-target data (Tables 8, S9; Fig. 5). This was particularly true for mitochondrial legacy loci, in which we observed a drastic increase in the number of loci recovered for most samples subjected to the TD-50 protocol instead of the standard one. While nuclear legacy loci appeared to be recovered more often at lower hybridization temperatures, this was not as consistent among samples compared to mitochondrial legacy loci.

*3.4.5. Effects of sequencing depth*

To investigate the influence of sequencing depth on our metrics of capture success, we evaluated whether the capture success metrics for 28 taxa were still congruent when reads were subsampled evenly. In general, patterns across hybridization temperatures remained congruent for the following metrics compared to when reads were not subsampled: proportion of on-target reads, coverage, total number of assembled contigs, number of loci captured by Coreoidea-derived baits, and number of legacy loci recovered (Figs. S11–S13). However, the median length of assembled contigs exhibited similar ranges of variation across hybridization temperatures, but the median quartile was much lower in the TD-50 protocol compared to the others (Fig. S11B). Similarly, the median length of capture loci was often shorter at lower hybridization temperatures (Fig. S12E). Lastly, while we found the number of loci captured by our Pentatomomorpha-derived baits to increase at lower hybridization temperatures compared to the standard protocol (congruent with our non-subsampled data), we observed that the TD-50 protocol did not capture as many loci as TD-60 and TD-55 protocols (Fig. S12D).

## 4. Discussion

Invertebrate target capture studies employing a fixed, high hybridization temperatures often recover relatively low proportions of targeted loci. For many taxonomic groups, like the leaf-footed bugs studied here, genomic resources are lacking or too limited to adequately optimize target capture baits to improve locus recovery. Furthermore, optimizing or re-designing target capture baits to improve locus recovery may result in more bait sets that are too narrowly targeted from a taxonomic perspective (e.g., a family) when there can be advantages of making bait sets more broadly applicable (e.g., a superfamily, infraorder, or higher ranks). Thus, modifying target capture conditions may be the best approach to improve the success of capture experiments across a broader taxon sampling.

Our evaluation of four different in-solution capture protocols found hybridization temperatures lower than the standard 65°C generally led to more assembled contigs and improved recovery of targeted loci, which held true even after normalizing sequencing depth. This improvement was particularly apparent for loci targeted by the more divergent Pentatomomorpha-derived baits despite touchdown capture ending at 50°C generally having a negative impact on the proportion of on-target reads and coverage, as well as increased numbers of paralogous loci. Lower hybridization temperatures also led to a greater recovery of legacy loci from off-target reads. Given our results, we find that optimizing *in vitro* target capture conditions to accommodate low hybridization temperatures can provide a cost-effective solution to improve the successful recovery of targeted loci in invertebrates (albeit with some costs), while also providing opportunities to gain additional data.

Our touchdown capture protocols were performed at longer durations than the standard protocol (36 hrs vs. 24 hrs, respectively), and one of these touchdown protocols had four incremental temperature decreases every 9 hrs instead of three incremental decreases every 12 hrs

(Fig. 1). Thus, the total capture duration and duration of the incremental temperature decreases might be considered a confounding factor in our experiment. However, two of our target capture protocols (TD-60 and TD-55) used the same timing, and it still appeared that a lower hybridization temperature generally improved target capture success. Thus, the duration of capture or the duration of bait-target hybridization at specific temperatures may not be a major confounding factor in our study.

*4.1. Considerations for using Coreoidea-derived baits*

We employed a bait design procedure in which coreoid transcriptomes were used to identify individual exons that baits would target, as well as entire transcripts that could contain multiple exons with baits potentially dissected by introns (i.e., a single bait that is derived from two exons). Based on our metrics irrespective of target capture conditions (particularly the number of targeted loci recovered), we found that our exon-derived baits were highly successful (78% recovery, on average) compared to the transcript-derived baits (33% recovery, on average). This finding indicates that most of the 58 loci targeted by transcript-derived baits cannot be captured. The poor *in vitro* performance of our transcript-derived baits may be due to transcript sequences comprised of many short exons leading to many baits dissected by introns. This could result in multiple assembled contigs if the entire intron(s) was not sequenced, which would result in the associated transcript baits matching multiple “different” contigs and in PHYLUCE excluding these sequences due to putative paralogy. Thus, our transcript-derived baits do not appear very useful in Coreoidea target capture studies, and our transcript bait design strategy may not be suitable for other groups in general unless exon-intron boundaries are known or different bioinformatic strategies are used.

*4.2. Considerations for using low hybridization temperature*

Low hybridization temperatures were associated with improved recovery of loci only targeted by the Pentatomomorpha-derived UCE baits, but not for loci targeted by Coreoidea-derived baits or PC dual baits, suggesting that other factors, such as bait-target divergence and bait tiling strategy, also impacted our capture experiments. Previous studies suggest that the efficacy of target capture markedly decreases when bait-target divergences exceed 5–10% (Vallender, 2011; Bi et al., 2012; Paijmans et al., 2016), although successful capture has been reported when divergences are much higher (e.g., 40% in Li et al. [2013]). The Pentatomomorpha-derived baits used here were derived from two taxa in the same infraorder as the leaf-footed bugs (Faircloth, 2017), but these taxa diverged from coreoids for ~160–230 my (Li et al., 2017; Liu et al., 2019; Wang et al., 2019). The Pentatomomorpha-derived baits generally exhibited relatively high levels of divergence from the targeted loci in coreoids, sometimes as high as that seen in Li et al. (2013). By reducing target capture stringency via lower hybridization temperatures, a greater mismatch tolerance between divergent baits and targets can lead to improved locus recovery, which our study supports. Given that our Coreoidea-derived baits or PC dual baits were designed from coreoid transcriptomes, we were able to target loci with much lower divergences from their respective baits (<10%).

Bait tiling strategy also appears to have played a role, although we could not thoroughly explore its effects on many of our metrics. Despite some loci targeted by the Pentatomomorpha-derived baits exhibiting similar levels of divergences as those targeted by Coreoidea-derived baits, we found that a modest increase in bait tiling density (in this case, the latter set of baits) was associated with higher coverage of the targeted loci. Lower bait-target divergences and/or

greater tiling of baits across targeted loci likely explain the high proportion of loci recovered regardless of hybridization temperature for these Coreoidea-derived baits.

If locus recovery is the main goal of invertebrate capture studies (especially if bait-target divergences exhibit a broad range of variation as in our study), using less stringent hybridization temperatures can be particularly beneficial when baits and targets are expected to exhibit high divergences and/or when baits are tiled at a low density. The lowest hybridization temperature tested (TD-50) resulted in the best target locus recovery, especially for the more divergent target loci. This result is likely due in part to the higher amount of data (i.e., reads and contigs) obtained for this treatment. However, when we subsampled raw reads across all capture conditions to normalize sequencing depth, we found that the TD-60 protocol provided the best balance with respect to the percentage of on-target reads and lower hybridization stringency allowing for maximal target locus recovery. Thus, to maximize target data for many samples with relatively lower sequencing depth, the TD-60 protocol may be preferable, but if sequencing depth is not a critically limiting factor, the TD-50 protocol may be a more desirable option.

It is expected that lowering hybridization temperature to reduce target capture stringency should result in a greater proportion of off-target than on-target sequences in capture data due to greater mismatch tolerance. Previous studies have observed this pattern (Paijmans et al., 2016 [ancient and archival specimens]; Cruz-Dávalos et al., 2017), which we also support here. More off-target data may be considered undesirable for many studies, and for phylogenomic studies, this could include potential sequences deemed paralogous to target data. When putative paralogs are detected in some sequence processing pipelines like PHYLUCE, both the target and putative paralog are excluded from further analysis, so even targeted data may be excluded from final analyses. We found more paralogs recovered for loci targeted by Pentatomomorpha-derived baits,

but not other targeted loci, which may be related to the degree of bait-target divergences or tiling strategy. Thus, for capture studies using baits expected to be quite divergent from ingroup taxa, lower hybridization temperatures can result in more targets recovered, but possibly at the cost of recovering more paralogous loci and a loss of some corresponding true targets. However, further evaluation of putative paralogs and potential targets can often prevent loss of data, especially since bioinformatic pipelines like PHYLUCE can report information on multiple contigs matching to a bait. This then allows analyzing sequences more carefully, such as with BLAST, inspecting alignments containing reference sequences, or refining ortholog binning procedures in bioinformatic pipelines to also use untargeted paralogs in analyses. Accurate paralog detection steps are very important regardless of hybridization temperatures (and how many paralogs are expected to be recovered) as any undetected paralogs have the potential to mislead phylogenetic reconstruction.

While off-target data may be considered undesirable, it can provide a greater chance to extract more legacy loci than what might be acquired under standard capture conditions. This can be advantageous for phylogenomic studies that seek to generate more comprehensive datasets (taxon and character sampling) by integrating capture data with well-known markers of historical use in molecular phylogenetic studies. Several target capture studies, especially in vertebrates, have extracted legacy loci from off-target sequences in capture data (e.g., Meiklejohn et al., 2014; Amaral et al., 2015; Wang et al., 2017; Derkarabetian et al., 2019; Simon et al., 2019; Miller et al., 2022). Legacy locus data extracted for focal taxa can then be combined with loci retrieved from genetic depositories for other taxa of interest prior to phylogenetic analysis, as well as to assess possible errors and/or contamination. While the relatively low number of legacy loci available for taxa in depositories does not always provide enough information for perfect

phylogenetic resolution, sometimes these loci are the only available data for rare taxa. If such taxa are closely related to others in a target capture experiment, they can be accurately placed in phylogenetic analyses despite the large amounts of missing data (e.g., Kieran et al., 2021; Miller et al., 2022). Given the sequencing depth of a standard Illumina HiSeq or NextSeq lane often used for sequencing target capture data and typical pooling of 60–96 samples for sequencing, the presence of some usable off-target data may be desirable compared to the limited range of on-target loci sequenced at very high, sometimes excessive, depth.

Tissue and library quality generally appear to have little influence on the trends we observed with respect to hybridization temperatures and many of our metrics of capture success, as we still observed the same general patterns when we partitioned our data based on these characteristics. One notable exception was degraded samples (i.e., those preserved dry). Our three touchdown hybridization temperatures improved locus recovery for degraded samples when compared to the fixed standard 65°C, but recovery did not differ when comparing across the touchdown protocols. It may be that for samples with highly fragmented DNA, the hybridization temperature of baits to bind to short target DNA (shorter than DNA usually targeted from fresh samples) only requires a modest reduction to obtain the maximum amount of targeted data possible. Furthermore, slightly lowering hybridization temperature may have the benefit of recovering enough loci to ensure a taxon can be represented in a final analysis where missing data is minimized. Thus, based on our assessment, employing capture approaches with hybridization temperatures less than 65°C generally have positive outcomes especially with samples of low quality.

## 5. Conclusion

Our study primarily focused on the effects of a single target capture parameter, i.e., hybridization temperature, on several metrics of capture success in an invertebrate protein-coding dataset. However, given that our baits were derived from different genomic resources, we were also able to explore the role of bait-target divergence and bait tiling strategy in our experiment, as well as tissue quality. While hybridization temperature, bait-target divergence, and bait tiling strategy had effects on some of our metrics, we recognize other parameters likely affect capture efficacy and remain to be thoroughly investigated in target capture experiments similar to ours (i.e., in-solution capture, sample quality, etc.). Such parameters may also include base genomes used in probe design (Gustafson et al., 2019), GC content of probes (Tewhey et al., 2009; Ávila-Arcos et al., 2011), amount of probes used (Cruz-Dávalos et al., 2017), and post-hybridization washing stringency (Li et al., 2013), among many others. Regardless, our study suggests that lowering the hybridization temperature during capture can be beneficial to similar studies (especially those with high bait-target divergences) seeking to improve target recovery in fresh and degraded invertebrate material without major negative impacts overall, while also retrieving other potentially useful data for comparative analyses.

Acknowledgements

We thank Joe Eger, Christiane Weirauch, and the Field Museum of Natural History for providing new samples for our study. We also thank Kevin Johnson for kindly providing transcriptome data prior to their publication. Caroline Miller and Min Zhao assisted with molecular benchwork. Gavin Naylor and Shannon Corrigan kindly provided use of their Covaris ultrasonication. Edward L. Braun helped with some perl and python scripts used in this study, and Ke Bi modified the Portik et al. (2016) scripts in order to accommodate our data. We also thank

Christine W. Miller and members of the Miller lab for their support and feedback. This study was funded by the National Science Foundation (grant number IOS-1553100 [awarded to Christine W. Miller]).

References

Agunbiade, T. A., Sun, W., Coates, B. S., Djouaka, R., Tamò, M., Ba, M. N., Binso-Dabire, C., Baoua, I., Olds, B. P., Pittendrigh, B. R. 2013. Development of reference transcriptomes for the major field insect pests of cowpea: a toolbox for insect pest management approaches in West Africa. *PLoS ONE*, *8*(11), e79929. doi: 10.1371/journal.pone.0079929

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. *Mol. Ecol. Resour.*, *20*(4), 892–905. doi: https://doi.org/10.1111/1755-0998.13160

Amaral, F. R. do, Neves, L. G., Resende, M. F. R. R., Jr., 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. doi: 10.1371/journal.pone.0138446

Ávila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V., Rasmussen, M., Fordyce, S. L., Montiel, R., Vielle-Calzada, J.-P., Willerslev, E., Gilbert, M. T. P. 2011. Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.*, *1*(1), 74. doi: 10.1038/srep00074

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*(4), 786–795. doi: 10.1111/syen.12244

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, *19*(5), 455–477. doi: 10.1089/cmb.2012.0021

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., Good, J. M. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. doi: 10.1186/1471-2164-13-403

Bolger, A. M., Lohse, M., Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 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 Ecol. Evol.*, *8*(6), 768–776. doi: 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*). *Syst. Entomol.*, *46*, 453–472. doi: 10.1111/syen.12470

Bushnell, B. 2014. *BBMap: A Fast, Accurate, Splice-Aware Aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). Retrieved from https://www.osti.gov/biblio/1241166

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Scheu, A., Beneke, N., Ludwig, A., Cooper, A., Willerslev, E., Orlando, L. 2017. Experimental conditions improving in-solution target enrichment for ancient DNA. *Mol. Ecol. Resour.*, *17*(3), 508–522. doi: https://doi.org/10.1111/1755-0998.12595

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

Dietrich, C. H., Allen, J. M., Lemmon, A. R., Lemmon, E. M., Takiya, D. M., Evangelista, O., Walde, K. K. O., Grady, P. G. S., Johnson, K. P. 2017. Anchored hybrid enrichment-based phylogenomics of leafhoppers and treehoppers (Hemiptera: Cicadomorpha: Membracoidea). *Insect Syst. Divers.*, *1*(1), 57–72. doi: 10.1093/isd/ixx003

Emberts, Z., Mary, C. M. S., 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. doi: 10.1111/evo.13948

Faircloth, B. C. 2013. Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. Retrieved from https://github.com/faircloth-lab/illumiprocessor

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

Faircloth, B. C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods Ecol. Evol.*, *8*(9), 1103–1112. doi: 10.1111/2041-210X.12754

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*(3), 489–501. doi: 10.1111/1755-0998.12328

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*(5), 717–726. doi: 10.1093/sysbio/sys004

Forthman, M., Miller, C. W., Kimball, R. T. 2019. Phylogenomic analysis suggests Coreidae and Alydidae (Hemiptera: Heteroptera) are not monophyletic. *Zool. Scr.*, *48*(4), 520–534. doi: 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). *Insect Syst. Divers.*, *4*(2). doi: 10.1093/isd/ixaa009

Glenn, T. C., Nilsen, R. A., Kieran, T. J., Finger, J. W., Pierson, T. W., Bentley, K. E., Hoffberg, S. L., Louha, S., León, F. J. G.-D., del Rio Portilla, M. A., Reed, K. D., Anderson, J. L., Meece, J. K., Aggery, S. E., Rekaya, R., Alabady, M., Bélanger, M., Winker, K., Faircloth, B. C. 2016. Adapterama I: universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). *bioRxiv*, 049114.

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*(7), 644–652. doi: 10.1038/nbt.1883

Gustafson, G. T., Alexander, A., Sproul, J. S., Pflug, J. M., Maddison, D. R., Short, A. E. Z. 2019. Ultraconserved element (UCE) probe set design: base genome and initial design parameters critical for optimization. *Ecol. Evol.*, *9*(12), 6933–6948. doi: 10.1002/ece3.5260

Gustafson, G. T., Baca, S. M., Alexander, A. M., Short, A. E. Z. 2020. Phylogenomic analysis of the beetle suborder Adephaga with comparison of tailored and generalized ultraconserved element probe performance. *Syst. Entomol.*, *45*(3), 552–570. doi: https://doi.org/10.1111/syen.12413

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 Evol. Biol.*, *16*(1), 212. doi: 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. *Mol. Ecol. Resourc.*, *21*, 816–833. doi: 10.1111/1755-0998.13287

Johnson, K. P., Dietrich, C. H., Friedrich, F., Beutel, R. G., Wipfler, B., Peters, R. S., Allen, J. M., Petersen, M., Donath, A., Walden, K. K. O., Kozlov, A. M., Podsiadlowski, L., Mayer, C., Meusemann, K., Vasilikopoulos, A., Waterhouse, R. M., Cameron, S. L., Weirauch, C., Swanson, D. R., Percy, D. M., Hardy, N. B., Terry, I., Liu, S., Zhou, X., Misof, B., Robertson, H. M., Yoshizawa, K. 2018. Phylogenomics and the evolution of hemipteroid insects. *P. Natl. Acad. Sci.*, *115*(50), 12775–12780. doi: 10.1073/pnas.1815820115

Katoh, K., Misawa, K., Kuma, K., Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, *30*(14), 3059–3066. doi: 10.1093/nar/gkf436

Katoh, K., Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, *30*(4), 772–780. doi: 10.1093/molbev/mst010

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. *Mol. Phylogenet. Evol.*, *130*, 297–303. doi: 10.1016/j.ympev.2018.10.026

Kieran, T. J., Gordon, E. R. L., Zaldívar-Riverón, A., Ibarra-Cerdeña, C. N., Glenn, T.C., Weirauch, C. 2021. Ultraconserved elements reconstruct the evolution of Chagas disease-vectoring kissing bugs (Reduviidae: Triatominae). *Syst. Entomol.*, *46*, 725–740, doi: 10.1111/syen.12485.

Knyshov, A., Gordon, E. R. L., Weirauch, C. 2019. Cost-efficient high throughput capture of museum arthropod specimen DNA using PCR-generated baits. *Methods Ecol. Evol.*, *10*, 841–852.

Knyshov, A., Gordon, E. R. L., Weirauch, C. 2021. New alignment-based sequence extraction software (ALiBaSeq) and its utility for deep level phylogenetics. *PeerJ*, *9*, e11019. doi: 10.7717/peerj.11019

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

Lemmon, E. M., Lemmon, A. R. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, *44*(1), 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Li, C., Hofreiter, M., Straube, N., Corrigan, S., Naylor, G. J. P. 2013. Capturing protein-coding genes across highly divergent species. *Biotechniques*, *54*(6), 321–326. doi: 10.2144/000114039

Li, F. W., Melkonian, M., Rothfels, C. J., Villarreal, J. C., Stevenson, D. W., Graham, S. W., Wong, G. K.-S., Pryer, K. M., Mathews, S. 2015. Phytochrome diversity in green plants and the origin of canonical plant phytochromes. *Nat. Commun.*, *6*, 1–12. doi: 10.1038/ncomms8852

Li, H., Leavengood, J. M., Chapman, E. G., Burkhardt, D., Song, F., Jiang, P., Liu, J., Zhou, Z., Cai, W. 2017. Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs. *Proc. R. Soc. B, Biol. Sci.*, *284*(1862), 20171223. doi: 10.1098/rspb.2017.1223

Liu, Y., Li, H., Song, F., Zhao, Y., Wilson, J.-J., Cai, W. 2019. Higher-level phylogeny and evolutionary history of Pentatomomorpha (Hemiptera: Heteroptera) inferred from mitochondrial genome sequences. *Syst. Entomol.*, *44*(4), 810–819. doi: https://doi.org/10.1111/syen.12357

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

Mason, V. C., Li, G., Helgen, K. M., Murphy, W. J. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from

noninvasively sampled museum specimens. *Genome Res.*, *21*, 1695–1704. doi: 10.1101/gr.120196.111

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. *Mol. Phylogenet. Evol.*, *78*, 314–323. doi: 10.1016/j.ympev.2014.06.003

Miller, C. D., Forthman, M., Miller, C. W., Kimball, R. T. 2022. Extracting "legacy loci" from an invertebrate sequence capture dataset. *Zool. Scr.*, *51*(1), 14–31. doi: 10.1111/zsc.12513

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.), *Research in Computational Molecular Biology* (pp. 158–170). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-37195-0_13

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., Förster, D. W. 2016. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol. Ecol. Res.*, *16*(1), 42–55. doi: 10.1111/1755-0998.12420

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. *Mol Ecol. Res.*, *14*, 1000–1010. doi: 10.1111/1755-0998.12249

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

Portik, D. M., Smith, L. L., Bi, K. 2016. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol. Ecol. Res.*, *16*(5), 1069–1083. doi: https://doi.org/10.1111/1755-0998.12541

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A. 2020. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, *70*(1), e102. doi: https://doi.org/10.1002/cpbi.102

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

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

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 genes and morphology reveal a relict lineage that is sister to all other singing cicadas. *Biol. J. Linn. Soc.*, *128*(4), 865–886. doi: 10.1093/biolinnean/blz120

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. *Insect. Soc.*, *64*, 149–154. doi: 10.1007/s00040-016-0525-8

Tewhey, R., Nakano, M., Wang, X., Pabón-Peña, C., Novak, B., Giuffre, A., Lin, E., Happe, S., Roberts, D. N., LeProust, E. M., Topol, E. J., Harismendy, O., Frazer, K. A. 2009. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, *10*(10), R116. doi: 10.1186/gb-2009-10-10-r116

Vallender, E. J. 2011. Expanding whole exome resequencing into non-human primates. *Genome Biol.*, *12*(9), R87. doi: 10.1186/gb-2011-12-9-r87

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*(11), e0188044. doi: 10.1371/journal.pone.0188044

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. *Mol. Phylogenet. Evol.*, *109*, 217–225. doi: 10.1016/j.ympev.2017.01.006

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. 2019. When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera–Heteroptera. *Cladistics*, *35*(1), 42–66. doi: 10.1111/cla.12232

Zhang, Y. M., Williams, J. L., Lucky, A. 2019. Understanding UCEs: a comprehensive primer on using ultraconserved elements for arthropod phylogenomics. *Insect Syst. Divers.*, *3*, 3. doi: 10.1093/isd/ixz016

Data accessibility

Newly generated sequence read files are available on NCBI's Sequence Read Archive under Bioproject PRJNA763985. For sequence read data previously published, see NCBI

Bioprojects PRJNA531965 (Forthman et al., 2019), PRJNA546248 (Forthman et al., 2020), and PRJNA609116 (Emberts et al., 2020).

Tables

Table 1. Experimental design of four studies that investigated the effects of target capture hybridization temperatures on various metrics used to assess capture performance.

| **Characteristics of study** | **Li et al. (2013)** | **Paijmans et al. (2016)** | **Cruz-Dávalos et al. (2017)** | **Mohandesan et al. (2017)** |
|---|---|---|---|---|
| Taxon | Gnathostomata | Felidae | Equidae | *Camelus* (Camelidae) |
| Target data type | CDS | Mitogenome | SNP | Mitogenome |
| No. of targets | 1449 | 1 (entire mitogenome) | ~5000 (SNPs) | 1 (entire mitogenome) |
| Bait type | Biotinylated RNA | “Oligonucleotides” | Biotinylated RNA | Biotinylated RNA |
| Bait length (bp) | 120 | 60 | 60 | 80 |
| Tiling density | 2x | 30x | 3x | 4x |
| Sample quality | Fresh | Fresh & degraded | Degraded | Degraded |
| Target capture method | In-solution | Solid-state | In-solution | In-solution |
| Capture Protocol (CP) 1 | 65°C hybridization (duration not reported) | 65°C hybridization; captured targets used for additional capture under same conditions (duration not reported) | 65°C hybridization (24 hrs) | 65°C hybridization (36 hrs) |
| Capture Protocol (CP) 2 | 65°C to 50°C hybridization (5°C decrease every 11 hrs) | 50°C hybridization; captured targets used for additional capture under same conditions (duration not reported) | 55°C hybridization (40 hrs) | 65°C to 50°C hybridization (5°C decrease every 12 hrs) |

| | | | | |
|---|---|---|---|---|
| Capture Protocol (CP) 3 | CP3 used captured targets & same capture conditions from CP 2 | 65°C to 50°C hybridization (5°C decrease every 16.25 hrs); captured targets used for additional capture under same conditions | | |
| Metrics evaluated | Proportion of target loci recovered<br>Bait-target divergence<br>GC content<br>Target seq. length recovered<br>Chromosomal position | Proportion of on-target reads<br>Bait-target divergence (fresh tissues only)<br>Target coverage | Fold enrichment<br>Proportion of on-target reads<br>Target coverage<br>Read length | Proportion of on-target reads<br>Percentage of endogenous DNA<br>Percentage of PCR duplicate reads |
| Recommended CP based on metrics | CP 3 | CP 3 (fresh)<br>CP 1 (archival & ancient) | CP 2 (samples with low to medium endogenous DNA)<br>CP 1 & CP 2 not significantly different (samples with high endogenous DNA) | CP 1 & CP 2 performance not significantly different |
| Relevant conclusions | Lower hybridization temperatures improve proportion of target loci recovered, particularly when bait-target divergence is high | Sample type & quality likely determines which hybridization temperature results in best capture based on proportion of on-target | Lower hybridization temperatures result in greater enrichment folds when samples have greater amounts of contamination but no | Lower hybridization temperatures perform similarly with standard capture conditions for recovery of uniquely mapped reads or |

| | | | |
|---|---|---|---|
| | reads; high temperatures may be better for degraded samples with greater amounts of contamination | impact when samples have higher amounts of endogenous DNA | capture efficiency |

Table 2. Terminology related to bait design and targeted loci.

| Term | Definition |
|---|---|
| Ultraconserved element (UCE) baits | 120 bp baits designed to target highly conserved genomic regions at ~1.33x tiling. A Hemiptera-wide UCE bait set was designed by Faircloth (2017), which targets the conserved core of protein-coding loci (Kieran et al., 2019); however, contigs captured by UCE baits may also include introns and/or untranslated regions (UTRs). |
| Pentatomomorpha-derived baits | The Faircloth (2017) Hemiptera-wide UCE bait set was subsampled to include only those baits designed from the genomes of two species within the infraorder Pentatomomorpha, which are the closest relatives to the ingroup of this study. |
| Exon-derived baits | 120 bp baits designed at ~2x tiling from individual exon sequences (i.e., baits do not span across introns) exhibiting ≥60% conservation across transcriptomes from five species of Coreoidea. While the baits do not target UTRs, given that they are derived from transcriptome sequences, contigs captured by these baits may include introns and/or UTRs. |
| Transcript-derived baits | 120 bp baits designed at ~2x tiling from sequences that may include one or more exons and that exhibit ≥60% conservation across transcriptomes from five species of Coreoidea. While the baits do not target UTRs, given that they are derived from transcriptome sequences, baits may span across one or more introns and, thus, contigs captured by these baits may include introns and/or UTRs. |
| Coreoidea-derived baits | Baits designed primarily from transcriptomes of five species of Coreoidea (i.e., exon- and transcript-derived baits combined). |
| Pentatomomorpha-Coreoidea (PC) dual baits | Loci targeted by Pentatomomorpha-derived UCE baits as well as by exon-derived or transcript-derived baits (i.e., Coreoidea-derived baits). |

Figures

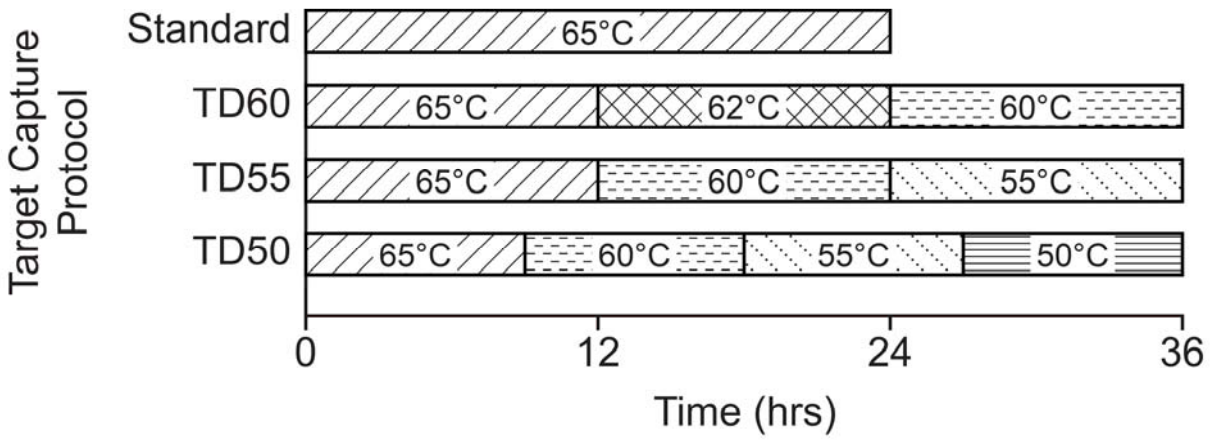
Figure 1. Experimental target capture set-up. Hybridization temperatures across time (in hours) are reported. Abbreviations: TD-60, touchdown hybridization approach starting at 65°C for 12 hrs followed by 62°C for 12 hrs and ending at 60°C for 12 hrs; TD-55, touchdown hybridization approach starting at 65°C for 12 hrs followed by 60°C for 12 hrs and ending at 55°C for 12 hrs; TD-50, touchdown hybridization approach starting at 65°C for 9 hrs followed by 60°C for 9 hrs, 55°C for 9 hrs, and ending at 50°C for 9 hrs.

Figure 2. Design and properties of bait used in this study. Abbreviations: bp, base pair; UCE, ultraconserved element; UTR, untranslated region.

Figure 3. (A) Effects of target capture protocols on the total number of raw reads and (B) assembled contigs. (C) Average minimum bait-target distances and (D) percentage of filtered reads on-target by bait design strategy. (E) Effects of target capture protocols on the percentage of filtered reads off-target. (F) Effects of target capture protocols on coverage by bait design strategy. Numbers in parentheses above x-axis denote sample size. Abbreviations: PC dual baits, loci targeted by both Pentatomomorpha- and Coreoidea-derived baits; see Fig. 1 for additional abbreviations.

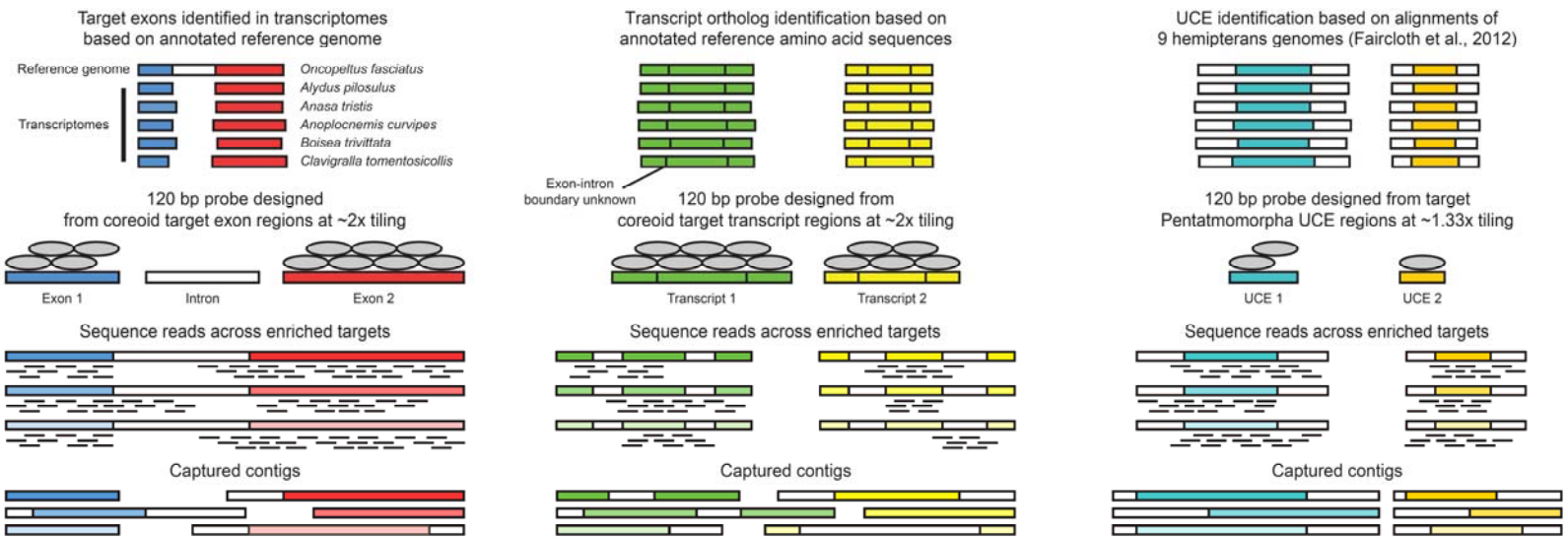
Figure 4. Effects of target capture protocols on the number of loci captured by (A) Coreoidea-derived baits, (B) PC dual baits, and (C) Pentatomomorpha-derived baits. (D) Effects of target capture protocols on the number of putative paralogs of loci captured by bait design strategy. Numbers in parentheses above x-axis denote sample size. Abbreviations: see Figs. 1 and 3.

Figure 5. Effects of target capture protocols on the total number of legacy loci recovered. Numbers in parentheses above x-axis denote sample size. Abbreviations: see Fig. 1.


Target Capture Protocol
Standard
TD60
TD55
TD50
65°C
62°C
60°C
55°C
50°C
0
12
24
36
Time (hrs)

892

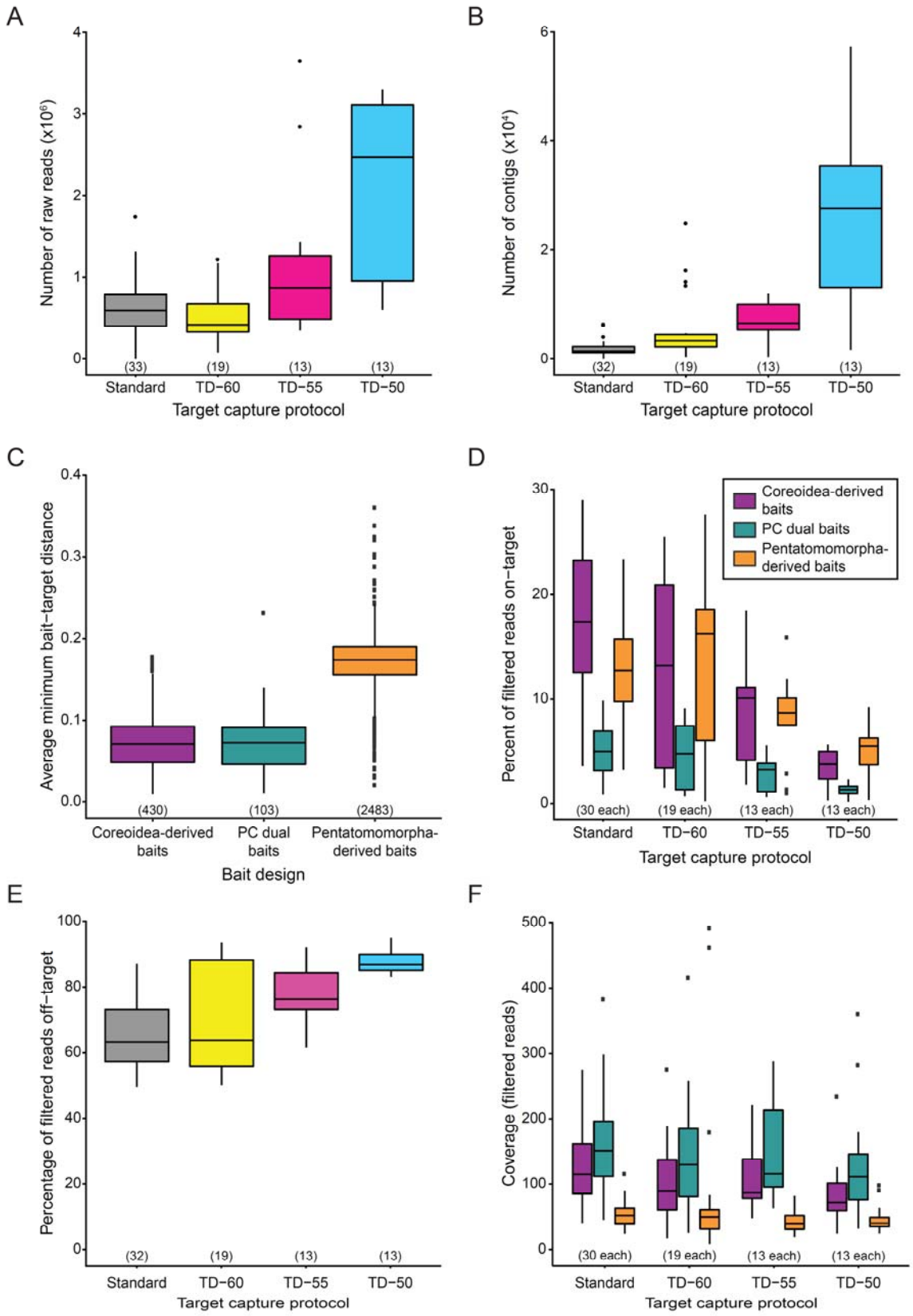
893



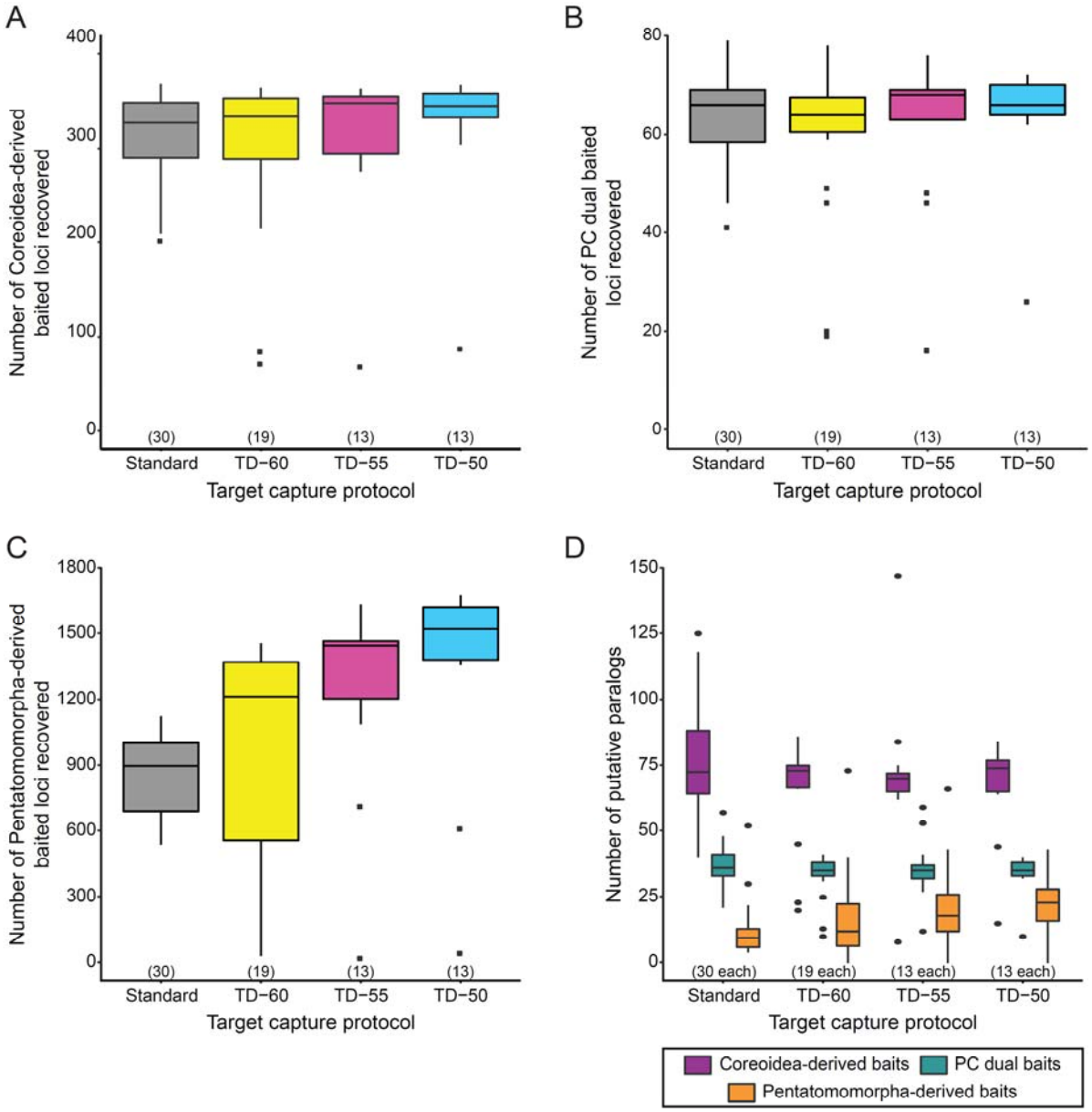

| Coreoidea-derived baits | | Pentatomomorpha-derived baits |
|---|---|---|
| Exon-derived baits | Transcript-derived baits | UCE-derived baits |
| Protein-coding target<br>May include flanking UTRs | Protein-coding target<br>May include flanking UTRs | Protein-coding conserved core<br>May include flanking UTRs |
| 5 coreoid base transcriptomes | 5 coreoid base transcriptomes | 2 pentatomomorphan base genomes |
| 456 exons initally targeted | 81 transcripts initally targeted | 2669 UCEs initally targeted |
| 376 final exons targeted | 58 final transcripts targeted | 2566 final UCEs targeted |
| 0.61–17.37%<br>average minimum<br>bait-target divergence | 2.82–17.50%<br>average minimum<br>bait-target divergence | 1.75–35.83%<br>average minimum<br>bait-target divergence |
| ~2x tiling density | ~2x tiling density | ~1.33x tiling density |

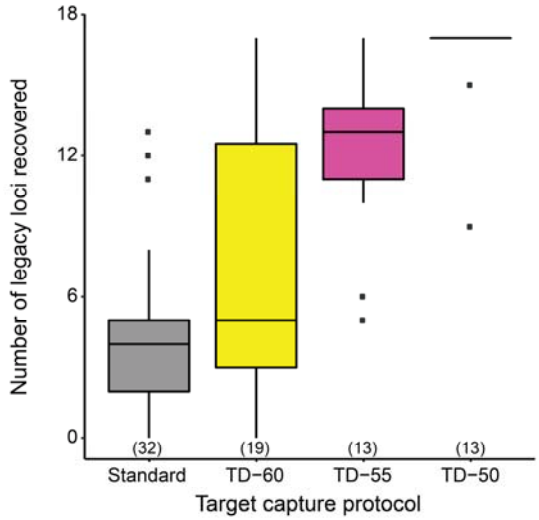
| Pentatomomorpha-Coreoidea dual baits (PC-dual baits) |
|---|
| 0 loci initally targeted |
| 103 final exon/transcript + UCE loci targeted |
| 0.69–22.92% average minimum bait-target divergence |
| ~2x tiling density |

894


A
Number of raw reads (x10^6)
(33)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol
B
Number of contigs (x10^4)
(32)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol
C
Average minimum bait−target distance
(430)
(103)
(2483)
Coreoidea-derived baits
PC dual baits
Pentatomomorpha-derived baits
Bait design
D
Percent of filtered reads on−target
Coreoidea-derived baits
PC dual baits
Pentatomomorpha-derived baits
(30 each)
(19 each)
(13 each)
(13 each)
Standard
TD−60
TD−55
TD−50
Target capture protocol
E
Percentage of filtered reads off−target
(32)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol
F
Coverage (filtered reads)
(30 each)
(19 each)
(13 each)
(13 each)
Standard
TD−60
TD−55
TD−50
Target capture protocol

895


A
Number of Coreoidea-derived baited loci recovered
400
300
200
100
0
(30)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol
B
Number of PC dual baited loci recovered
80
60
40
20
0
(30)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol
C
Number of Pentatomomorpha-derived baited loci recovered
1800
1500
1200
900
600
300
0
(30)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol
D
Number of putative paralogs
150
125
100
75
50
25
0
(30 each)
(19 each)
(13 each)
(13 each)
Standard
TD−60
TD−55
TD−50
Target capture protocol
Coreoidea-derived baits
PC dual baits
Pentatomomorpha-derived baits


18
12
6
0
Number of legacy loci recovered
(32)
(19)
(13)
(13)
Standard
TD−60
TD−55
TD−50
Target capture protocol

898

Supplementary material

Table S1. Information regarding sample age, preservation method, and DNA extraction and library preparation protocols used. Abbreviations: DNeasy, Qiagen, DNeasy Blood and Tissue Kit; DNQIA, DNeasy with Qiagen QIAquick PCR Purification Kit.

Table S2. Target capture experimental design. Freshly preserved samples or samples preserved dried were subjected to the standard and TD-60 protocols, respectively, prior to the start of this study. Abbreviations: TD-60, touchdown hybridization approach starting at 65°C for 12 hrs followed by 62°C for 12 hrs and ending at 60°C for 12 hrs; TD-55, touchdown hybridization approach starting at 65°C for 12 hrs followed by 60°C for 12 hrs and ending at 55°C for 12 hrs; TD-50, touchdown hybridization approach starting at 65°C for 9 hrs followed by 60°C for 9 hrs, 55°C for 9 hrs, and ending at 50°C for 9 hrs.

Table S3. Summary data for raw and filtered sequence reads, reads on-target and coverage across all targeted loci, and contigs. Abbreviations: FRO, filtered reads on-target; see Table S2 for additional abbreviations.

Table S4. Summary data for on-target reads and coverage of targeted loci, partitioned based on type of baits used. Abbreviations: Cov., coverage; exon, loci targeted by exon-derived baits; C-baits, loci targeted by exon- or transcript-derived baits (i.e., Coreoidea-derived baits); transcript, loci targeted by transcript-derived baits; PC dual, loci targeted by both Pentatomomorpha ultraconserved element (UCE) baits and Coreoidea-derived baits (i.e., Pentatomomorpha-Coreoidea dual baits); P-baits, loci targeted only by Pentatomomorpha-derived UCE baits; see Tables S2 and S3 and Table 1 of main text for explanation of terms used.

Table S5. Summary data for captured loci targeted by exon-derived baits (376 loci targeted). Abbreviations: see Tables S2–S4.

Table S6. Summary data for captured loci targeted by transcript-derived baits (58 loci targeted). Abbreviations: see Tables S2–S4.

Table S7. Summary data for captured loci targeted by Pentatomomorpha-Coreoidea dual baits (103 loci targeted). Abbreviations: see Tables S2–S4.

Table S8. Summary data for captured loci targeted by Pentatomomorpha-derived baits (2566 loci targeted). Abbreviations: see Tables S2–S4.

Table S9. Summary data for legacy loci (15 mitochondrial [mtDNA] and two nuclear ribosomal [rRNA] loci targeted). Abbreviations: see Table S2.

Table S10. Summary data for raw and filtered sequence reads, reads on-target and coverage across all targeted loci, and contigs from 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Tables S2 and S3.

Table S11. Summary data for on-target reads and coverage of targeted loci, partitioned based on type of baits used for 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Tables S2–S4 and Table 1 of main text.

Table S12. Summary data for captured loci targeted by exon-derived baits (376 loci targeted) from 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Tables S2–S4.

Table S13. Summary data for captured loci targeted by transcript-derived baits (58 loci targeted) from 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Tables S2–S4.

Table S14. Summary data for captured loci targeted by Pentatomomorpha-Coreoidea dual baits (103 loci targeted) from 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Tables S2–S4.

Table S15. Summary data for captured loci targeted by Pentatomomorpha-derived baits (2566 loci targeted) from 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Tables S2–S4.

Table S16. Summary data for legacy loci (15 mitochondrial [mtDNA] and two nuclear ribosomal [rRNA] loci targeted) from 28 taxa that had 2,000,000 million raw reads subsampled to equalize sequencing depth across capture conditions. Abbreviations: see Table S2.

Figure S1. (A) Effects of target capture protocols on mean contig. (B) Average minimum bait-target distances by bait design strategy. (C) Effects of target capture protocols on the number of filtered reads on-target by bait design strategy. (D) Effects of target capture protocols on coverage by bait design strategy. (E) Effects of tiling strategy (Coreoidea-derived ~2x tiling density; Pentatomomorpha-derived ~1.33x tiling density) on average coverage per locus (capture loci exhibiting 0.05–0.10 average minimum bait-target divergences for each bait design strategy). (F) Effects of target capture protocols on median locus lengths by bait design strategy. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S2. Effects of target capture protocols on the number of loci captured by (A) exon-derived baits and (B) transcript-derived baits. (C) Effects of target capture protocols on the number of putative paralogs of loci captured by bait design strategy. Numbers in parentheses above x-axis denote sample size. See Table S2 for abbreviations.

Figure S3. Effects of target capture protocols on (A) number of raw reads, (B) number of contigs, (C) percentage of filtered reads on-target, and (D) coverage by sample preservation method. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S4. Effects of target capture protocols on the number of loci captured with (A) Coreoidea-derived baits, (B) PC dual baits, and (C) Pentatomomorpha-derived baits by sample preservation method. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S5. Effects of target capture protocols on the median length of loci captured with (A) Coreoidea-derived baits, (B) PC dual baits, and (C) Pentatomomorpha-derived baits by sample preservation method. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S6. Effects of target capture protocols on (A) the number of putative paralogs of loci capture with Coreoidea-derived baits, (B) the number of putative paralogs of loci capture with PC dual baits, (C) the number of putative paralogs of loci capture with Pentatomomorpha-derived baits, and (D) the number of legacy loci recovered by sample preservation method. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S7. Effects of target capture protocols on (A) number of raw reads, (B) number of contigs, (C) percentage of filtered reads on-target, and (D) coverage by library quality. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S8. Effects of target capture protocols on the number of loci captured with (A) Coreoidea-derived baits, (B) PC dual baits, and (C) Pentatomomorpha-derived baits by library quality. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S9. Effects of target capture protocols on the median length of loci captured with (A) Coreoidea-derived baits, (B) PC dual baits, and (C) Pentatomomorpha-derived baits by library quality. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S10. Effects of target capture protocols on (A) the number of putative paralogs of loci capture with Coreoidea-derived baits, (B) the number of putative paralogs of loci capture with PC dual baits, (C) the number of putative paralogs of loci capture with Pentatomomorpha-derived baits, and (D) the number of legacy loci recovered by library quality. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S11. When controlling for sequencing depth, effects of target capture protocols on (A) the number of assembled contigs, (B) median contig length, (C) percentage of filtered reads on-target by bait design strategy, and (D) coverage by bait design strategy. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S12. When controlling for sequencing depth, effects of target capture protocols on the number of loci captured by (A) exon-derived baits, (B) transcript-derived baits, (C) PC dual baits, and (D) Pentatomomorpha-derived baits when sequencing depth is equalized across taxa and capture conditions, as well as (E) effects on the median locus length and (F) number of putative

paralogs of loci captured by bait design strategy. Numbers in parentheses above x-axis denote sample size. See Tables S2–S4 for abbreviations.

Figure S13. When controlling for sequencing depth, effect of target capture protocols on the total number of legacy loci recovered. Numbers in parentheses above x-axis denote sample size. See Table S2 for abbreviations.