Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017 doi: 10.1111/2041-210X.12754

Identifying conserved genomic elements and designing universal bait sets to enrich them

Brant C. Faircloth*

Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA

Summary

1. Targeted enrichment of conserved genomic regions is a popular method for collecting large amounts of sequence data from non-model taxa for phylogenetic, phylogeographic and population genetic studies. For example, two available bait sets each allow enrichment of thousands of orthologous loci from >20 000 species (Faircloth *et al*. Systematic Biology, 61, 717–726, 2012; Molecular Ecology Resources, 15, 489–501, 2015). Unfortunately, few open-source workflows are available to identify conserved genomic elements shared among divergent taxa and to design enrichment baits targeting these regions. Those that do exist require extensive bioinformatics expertise and significant amounts of time to use. These shortcomings limit the application of targeted enrichment methods to additional organismal groups.

2. Here, I describe a universal workflow for identifying conserved genomic regions in available genomic data and for designing targeted enrichment baits to collect data from these conserved regions. These methods require less expertise, less time and better use commonly available information to identify conserved loci and design baits to capture them.

3. I apply this computational approach to the understudied arthropod groups Arachnida, Coleoptera, Diptera, Hemiptera or Lepidoptera to identify thousands of conserved loci in each group and design target enrichment baits to capture these loci. I then use *in silico* analyses to demonstrate that targeted enrichment of the conserved loci can be used to reconstruct the accepted relationships among genome sequences from the focal arthropod orders.

4. The software workflow I created allowed me to identify thousands of conserved loci in five diverse arthropod groups and design sequence capture baits to target them. This suite of capture bait designs should enable collection of phylogenomic data from >900 000 arthropod species. Although the examples in this manuscript focus on understudied arthropod groups, the approach I describe is applicable to all organismal groups having some form of pre-existing genomic information (e.g. other invertebrates, plants, fungi and microbes). Finally, the documentation, design steps, software code and bait sets developed here are available under an open-source license for restriction-free testing, use, and additional modification by any research group.

Key-words: bait design, conserved elements, genomics, phylogenetics, phylogenomics, phylogeography, ultraconserved elements

Introduction

Collecting sequence data from non-model taxa has undergone a revolution during the previous 10 years, driven by advancements in sequencing technologies (Bentley *et al*. 2008) and molecular methods (Hardenbol *et al*. 2003; Baird *et al*. 2008; Gnirke *et al*. 2009). Ecologists and evolutionary biologists have typically focused on a narrower subset of these approaches, collectively known as 'reduced representation' methods, which include varieties of restriction enzyme-based (Baird *et al*. 2008; Elshire *et al*. 2011; Peterson *et al*. 2012), transcriptomic (Dunn *et al*. 2008; Smith *et al*. 2011; Misof *et al*. 2014) and targeted enrichment (Bi *et al*. 2012; Faircloth *et al*. 2012; Peñalba *et al*. 2014; Ali *et al*. 2015; Hoffberg *et al*. 2016; Hugall *et al*. 2016; Suchan *et al*. 2016) approaches. These methods allow the collection of large numbers of loci from large numbers of organisms and are less expensive and potentially less complicated than whole-genome sequencing or genome-resequencing approaches, particularly when collecting data from tens or hundreds of individuals.

One popular reduced representation approach is the targeted enrichment (Gnirke *et al*. 2009) of conserved or ultraconserved genomic elements (*sensu* Faircloth *et al*. 2012). In this approach, researchers identify genomic regions of high conservation shared among divergent lineages, design synthetic oligonucleotide 'baits' that are complementary to these regions, hybridize genomic libraries to these oligonucleotide baits, 'fish' out the hybridized bait + library structure, remove the bait sequence and sequence the remaining pool of enriched, targeted DNA. Although the baits target and enrich conserved regions of the genome, the library preparation, enrichment, sequencing and assembly procedures ensure that the approach

*Correspondence author. E-mail: brant@faircloth-lab.org

also captures variable flanking sequence that sits to each side of each conserved region (Faircloth *et al*. 2012; Smith *et al*. 2014).

The power of this approach is that a single tube of synthetic oligonucleotide baits can be used by multiple studies to collect data across very broad taxonomic scales – for example, amniotes (Crawford *et al*. 2012; McCormack *et al*. 2013; Hosner *et al*. 2015; Streicher & Wiens 2016) or fishes (Faircloth *et al*. 2013; McGee *et al*. 2016) or bees, ants, and wasps (Blaimer *et al*. 2015; Faircloth *et al*. 2015). Additionally, the sequence reads from these enriched, conserved loci can be analysed in different ways to address questions at different scales, from deep-time phylogenetic studies (Faircloth *et al*. 2013) to shallower level phylogeographic studies (Smith *et al*. 2014) to population-level studies (Harvey *et al*. 2016; Manthey *et al*. 2016). Finally, because the targeted enrichment approach is DNA based, it can be applied to degraded and low-quantity samples, such as those in many specimen or tissue collections (Bi *et al*. 2013; McCormack, Tsai & Faircloth 2015; Blaimer *et al*. 2016; Lim & Braun 2016; Ruane & Austin 2017).

Although targeted enrichment of conserved elements offers many benefits (Harvey *et al*. 2016), there are few well-described, easy-to-use workflows for identifying conserved loci shared among organismal genomes or for designing sequence capture baits targeting these regions (cf. Johnson *et al*. 2016; Mayer *et al*. 2016). Here, I describe a workflow I have recently developed to accomplish these tasks, and I demonstrate its utility by: (i) identifying large suites of conserved elements shared within five diverse and understudied arthropod groups (Arachnida, Coleoptera, Diptera, Hemiptera, Lepidoptera), and (ii) designing five sets of capture baits targeting conserved regions shared among members of each taxonomic group. This updated workflow differs from previous approaches (Faircloth *et al*. 2012; McCormack *et al*. 2012) by aligning small, random pieces of DNA from several genomes to a focal reference genome using a permissive read aligner and then using overlapping coordinates shared among multiple taxa to identify regions of shared conservation. This technique greatly increases the number of conserved regions detected relative to synteny based, genome-genome alignment procedures (e.g. Harris 2007) used in earlier manuscripts (Faircloth *et al*. 2012; McCormack *et al*. 2012). I then use *in silico* target enrichment experiments to show that these bait sets collect conserved loci that can be used to reconstruct the known phylogenetic relationships within the respective class/orders. We empirically test one of the bait sets described below by enriching conserved loci from a diverse group of Arachnids in a separate manuscript (Starrett *et al*. in press), and we describe a second empirical test of the workflow described here to improve available resources for hymenopteran phylogenetics in Branstetter *et al*. (in press).

I make all documentation and computer code for this workflow available under an open-source license, allowing researchers to generalize the approach to other organisms having some genomic data. I also make all of the bait sets for Arachnida, Coleoptera, Diptera, Hemiptera and Lepidoptera available under a public domain license (CC-0), facilitating restriction-free commercial synthesis, testing, use and improvement of these bait sets by other research groups interested in phylogenetic, phylogeographic and population-level analyses of arthropods.

Materials and methods

STUDY GROUP

The arthropod groups Arachnida, Coleoptera, Diptera, Hemiptera and Lepidoptera are among the most diverse invertebrate classes/orders, encompassing more than 900 000 species (Harvey 2002; Zhang 2011). Yet, our understanding of the evolutionary factors responsible for generating the extreme diversity within each of these groups is poor. Large projects, like the i5k (i5K Consortium 2013) are transforming our knowledge of major relationships among arthropod lineages (Misof *et al*. 2014). However, the extreme diversity of many arthropod groups makes the large-scale collection of transcriptome data across clades difficult, suggesting that less expensive, genome reduction techniques that work with DNA (vs. RNA) could be useful for understanding finer grained evolutionary relationships among hundreds or thousands of arthropod species within major taxonomic groups. Targeted enrichment of conserved DNA regions shared among these species offers one approach for beginning to fill these gaps, particularly because the technique is useful with older, degraded DNA, similar to that collected from arthropod (Faircloth *et al*. 2015; Blaimer *et al*. 2016) and other museum specimens (Bi *et al*. 2013; McCormack, Tsai & Faircloth 2015).

GENERAL WORKFLOW

Although some implementation details differ for the groups described below in terms of the specific data used, the general workflow (Fig. 1) for identifying conserved loci and designing capture baits to target them begins with the selection of an appropriate 'base' genome that is within or related to the focus group and against which data from other exemplar taxa sampled within the focus group will be aligned. The base genome sequence can be an ingroup or outgroup taxon, and it is reasonable to select the best-assembled and annotated genome that is closely related to or nested within the focus group rather than focusing intently on ingroup or outgroup status. This choice facilitates downstream analysis or selection of conserved loci based on desirable properties derived from annotation or positional information (exon, intron, intergenic, unlinked, etc.), although how, exactly, to identify the best-assembled genome is a matter of debate (Earl *et al*. 2011; Bradnam *et al*. 2013). I generally focus on selecting assemblies as the 'base' when they contain relatively large and complete scaffolds and have reasonable annotation (ideally evidence-based, although gene predictions are also useful).

Following the selection of a base genome, the workflow proceeds by generating short reads from organisms that serve as exemplars of the focus group's diversity using either: (i) low-coverage (4–6×), massively parallel sequencing reads and/or (ii) short reads simulated from other genome sequences that exist for the focus group. The next step in the workflow is to align all sets of exemplar reads to the base genome using a permissive raw-read aligner such as *stampy* (Lunter & Goodson 2011), produce a BAM (Li *et al*. 2009) file, and use *samtools* (Li *et al*. 2009) to reduce the size of the BAM file by selecting only the reads from each BAM file that align to the base genome.

The workflow proceeds by converting successfully aligned reads to interval (BED) format using *bedtools* (Quinlan & Hall 2010), which allows fast and easy manipulation of alignment data. Using the resulting BED files, the next step in the workflow is to sort the alignment

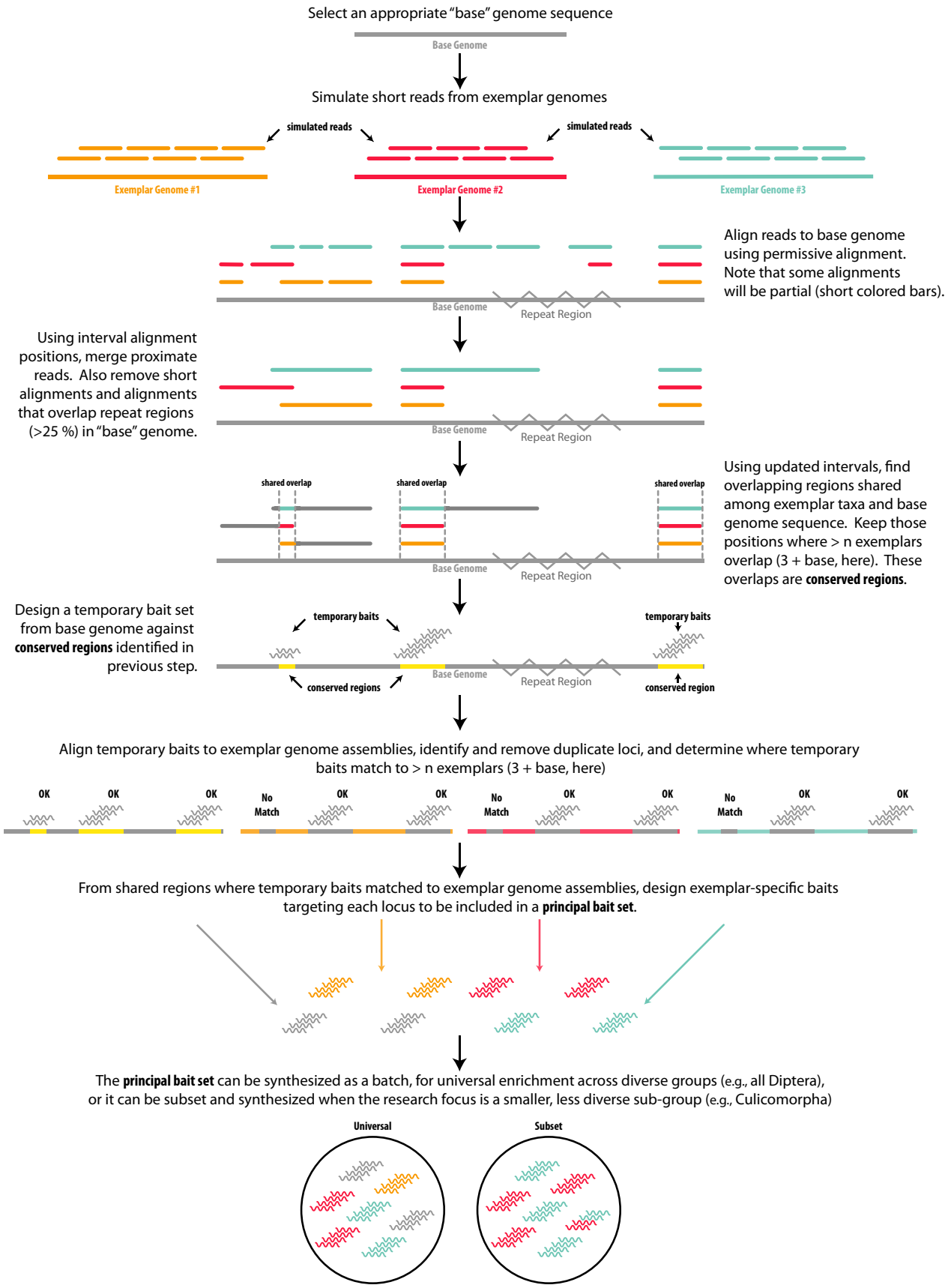

Fig. 1. Illustration of the steps involved in the conserved element identification and bait design workflow.

coordinates; merge together alignment positions in each file that are close (<100 bp) to one another; and remove alignments that are short (<80 bp), overlap masked loci (>25% of length) and/or contain ambiguous (N or X) bases. The workflow then proceeds by processing the filtered BEDs to create a relational database of overlapping alignment positions shared between the base genome sequence and each of the exemplar taxa. Because reads from each exemplar taxon were aligned to the same base genome and because masked loci have been removed, overlapping alignment positions shared among taxa represent loci that are putatively conserved between genomes. Users can query this database to generate a BED file of the genomic locations of each conserved locus in the base genome sequence that are also shared by a subset of some or all of the exemplar taxa.

The final stages of the workflow focus on designing oligonucleotide baits to target the conserved loci identified in the steps described above. The first step of this process is to extract the conserved loci from the base genome as FASTA-formatted records, and design temporary oligonucleotide bait sequences targeting these loci. The workflow then uses LASTZ (Harris 2007) to align these temporary baits designed from the base genome to genomic data from a set of exemplar taxa (which can be the same, a subset, or a superset of the organisms used for locus identification) and builds a relational database of loci detected in each of the exemplar taxa. From this relational database, users can determine which base genome bait sequences 'hit' in which exemplar taxa, and they can select to output those loci consistently detected in a majority of exemplar taxa. Users input this list of loci to a program that designs bait sequences from all exemplar taxa where each conserved locus was consistently detected. The final stage of the workflow is to screen and remove bait sequences that appear to target duplicate loci within and between all exemplar taxa by aligning all baits to themselves, identifying baits designed from one locus that 'hit' other loci (≥50% sequence identity over ≥50% of sequence length), and reciprocally removing all loci where any bait matched any portion of another locus. The final output of the workflow is a file containing bait sequences for each conserved locus that were selected from each exemplar genome, such that Locus 1 may have baits designed from Taxon A, Taxon B and Taxon C. This design approach increases the likelihood that baits will capture the targeted locus when combined with DNA libraries prepared from organisms having genome sequences divergent from the exemplar taxa. I called this final bait design file the 'principal' FASTA file of bait sequences or the 'principal bait set'.

ARACHNIDA

Because few arachnids have genomic data available, I collected sequence data from a diverse group of arachnids using low-coverage genome sequencing. I extracted DNA from legs or legs + cephalothorax of samples using Qiagen DNeasy kits, adding 2 μL RNase (1 mg mL^{-1}) to each extraction. I visualized DNA extracts on 1·5% (w/v) agarose, and I sheared the resulting DNA to 400–500 bp using a Bioruptor (Diagenode, Inc., Denville, NJ, USA). After shearing, I prepared sequencing libraries from 100 to 500 ng sheared DNA using a commercial library preparation kit (Kapa Biosystems, Inc., Wilmington, MA, USA) with a set of custom sequence tags to identify each library (Faircloth & Glenn 2012). I amplified 15 μL of each library in a reaction mix of 25 μL Kapa HiFi HS Master Mix (Kapa Biosystems, Inc.), 5 μL Illumina primer mix (5 μM each), 15 μL of adapter-ligated DNA and 5 μL of ddH_2O using a thermal profile of 98 °C for 45 s followed by 14 cycles of 98 °C for 15 s, 60 °C for 30 s, 72 °C for 60 s followed by a 72 °C extension for 5 min. After amplification, I cleaned libraries 1:1 with a SPRI-substitute (Rohland & Reich 2012), and I checked the quality of resulting libraries by visualizing 1 μL of each (5 ng μL^{-1}) on a BioAnalyzer (Agilent, Inc., Santa Clara, CA, USA). Because I observed adapter-dimer peaks following initial quality control, I cleaned libraries one to two additional times using 1:1 SPRI-substitute. After validating the removal of dimer peaks from the sequencing libraries, I qPCR quantified 5 ng μL^{-1} aliquots of each library using a commercial kit (Kapa Biosystems, Inc.), and I combined libraries at equimolar ratios to make a 10 μM pool that I sequenced using PE150 reads on an Illumina HiSeq 2000 (UCSC Genome Technology Center). Once I received sequence data from the sequencing centre, I trimmed sequencing reads for adapter contamination and low-quality bases using an automated wrapper around trimmomatic (Faircloth 2013; Bolger, Lohse & Usadel 2014), and I merged read pairs into a single file. To add a lineage representing ticks to the sequence data generated from other arachnids, I used *art* (Huang *et al*. 2012) to simulate error-free, paired-end reads of 100 bp from the *Ixodes scapularis* genome assembly (GCA_000208615.1).

Following the steps outlined in Data S1, Supporting Information, I aligned reads for all organisms in Table S1 to the *Limulus polyphemus* genome assembly (GCA_000517525.1; hereafter limPol1) using *stampy* (Lunter & Goodson 2011) with a substitution rate of 0·10, and I streamed the resulting SAM data to a BAM file using *samtools* (Li *et al*. 2009). I used a higher substitution rate for arachnids, relative to the other arthropod groups described below, to: (i) account for the older estimated divergence times of crown arachnid lineages (~400–500 MYA; Sanders and Lee 2010) relative to the other arthropod groups (~150–250 MYA; Misof *et al*. 2014), and (ii) account for sequencing error (Ross *et al*. 2013) in the arachnid Illumina data relative to the simulated reads generated for other arthropod taxa (which derive from a putatively more accurate consensus genome assembly). After alignment, I reduced the BAM file to contain only those reads mapping to the limPol1 genome, I converted each BAM file to a BED file and I screened the resulting interval data to remove those intervals in each BED file that overlapped masked, short (<80 bp), or ambiguous segments of the limPol1 genome. The intervals that were not filtered represent conserved sequence regions shared between the base genome (limPol1) and each of the exemplar taxa. I used the phyluce_probes_get_multi_merge_table program to determine which of these conserved intervals were shared among some/all of the exemplar taxa, and I output a list of those intervals shared by limPol1 and six arachnid exemplars. I buffered these intervals shared by limPol1 and arachnids to 160 bp, and I extracted FASTA sequence from the limPol1 genome corresponding to the buffered intervals (phyluce_probes_get_genome_sequences_from_bed). Then, I designed a temporary set of sequence capture baits by tiling two bait sequences over each interval (phyluce_probe_get_tiled_probes) where baits overlapped by 40 bp. This produced a set of temporary enrichment baits designed from limPol1, and I screened this set of temporary baits to remove baits sequences that were ≥50% identical over >50% of their length.

To design a more diverse bait set that included baits from a larger selection of arachnids, I downloaded several arachnid genome assemblies (Table S1) and also included a new genome assembly from a tick (NCBI PRJNA374336). Then, I aligned baits from the temporary bait set to each genome using a wrapper (phyluce_probe_run_multiple_lastzs_sqlite) around *lastz* (Harris 2007) with liberal alignment parameters (≥50% sequence identity required to map). Using the alignment data, I removed loci that were hit by baits targeting different conserved regions or multiple loci that were hit by the same bait (phyluce_slice_sequence_from_genomes), and I buffered remaining, non-duplicate loci to 180 bp. I used a separate program (phyluce_probes_get_multi_fasta_table) to determine which loci I detected across the arachnid genome assemblies, and I created a list of those loci detected in 6 of the 10 arachnid genome assemblies. I then designed a

bait set targeting these loci by tiling baits across each locus in each of the 10 arachnid genomes where I detected the locus, and I screened the resulting bait set to remove putative duplicates. I called this the principal arachnid bait set.

To check the sanity of the data returned from the principal arachnid bait set, I performed an *in silico* targeted enrichment experiment. First, I aligned the baits to 10 arachnid genomes (Table S1) using a program (phyluce_probe_slice_sequence_from_genomes) from the PHYLUCE package. After identifying conserved loci that aligned to baits in the principal bait set, I buffered the match locations by ±500 base pairs, and I extracted FASTA data from the buffered intervals. Then, I input the FASTA-formatted contigs from the previous step to the standard PHYLUCE workflow for phylogenomic analyses (Faircloth 2015). Briefly, I performed additional orthology and duplicate screening steps (phyluce_assembly_match_contigs_to_probes; –min_coverage 80, –min_identity 80), exported non-duplicate conserved loci to FASTA format, aligned the FASTA data using *mafft* (Katoh & Standley 2013) and trimmed the resulting alignments using *gblocks* (Castresana 2000; Talavera & Castresana 2007). I created a dataset in which all alignments contained at least 7 of the 10 taxa (70% complete matrix), and I concatenated the resulting alignment data into a supermatrix. I used *RAxML* v8.0.19 (Stamatakis 2014) to: (i) perform a maximum likelihood (ML) search for the tree best-fitting the data using the GTRGAMMA site rate substitution model, (ii) perform nonparametric bootstrapping of the data, and (iii) reconcile the 'best' ML tree with the bootstrap support values.

To further assess the performance of the principal arachnid bait set, we performed extensive *in vitro* enrichments of the identified loci as part of a separate manuscript (Starrett *et al.* in press).

COLEOPTERA

To design baits targeting conserved loci in Coleoptera (specific steps are outlined in Data S2), I downloaded available genomes for several coleopteran lineages (Table S2), and I used *art* (Huang *et al.* 2012) to simulate error-free, paired-end reads of 100 bp at 2× coverage from each genome sequence. I merged paired reads for each taxon into a single file, and I aligned the merged, simulated reads to the genome sequence of *Tribolium castaneum* (GCA_000002335.2; triCas1 hereafter) using *stampy* (Lunter & Goodson 2011) with a substitution rate of 0·05 and streaming the resulting SAM alignment data to BAM format using *samtools* (Li *et al.* 2009). Subsequent processing steps were similar to the workflow for Arachnida. In brief, I remove unaligned reads from the BAM file, converted the BAM file to BED format, and screened the resulting interval data to remove intervals in each BED file that overlapped masked, short (<80 bp) or ambiguous segments of the triCas1 genome. I subsequently created a table of conserved regions shared between the base genome (triCas1) and each of the exemplar taxa, and I queried this table to output a list of intervals shared by all of the exemplar taxa and the base taxon. I selected this stricter threshold (relative to those used for other arthropod groups) because of the extreme diversity of the beetle clade and because conserved loci shared among all of the exemplar beetle taxa were more likely to be present in all beetle lineages. I output the list of these loci, designed a temporary bait set using FASTA data from the triCas1 genome, and re-aligned the temporary baits to the genomes of each exemplar taxon (Table S2), as well as one species representing a strepsipteran outgroup to beetles (*Mengenilla moldrzyki*, GCA_000281935.1). I included this species to add additional diversity to the bait set and better represent earlier diverging clades in the beetle tree relative to the clades represented by the other beetle genomes I used for bait design. I extracted sequence in FASTA format for each conserved locus from each exemplar taxon assembly, and I designed a hybrid set of bait sequences targeting each of these loci from the genomes of the exemplar taxa. I filtered putative duplicate baits/loci from this dataset, and I called the resulting file the principal coleopteran bait set.

I performed an *in silico* sanity check of the bait set using an approach identical to that described above. I aligned the principal bait set to the genomes of the taxa that I used to design the principal bait set, sliced FASTA sequences from each genome that flanked the conserved locus location by ±400 bp, performed additional orthology and duplicate screening steps (–min_coverage 67, –min_identity 80), used *mafft* to align FASTA slices for each locus across all taxa, trimmed resulting alignments using *gblocks*, created a dataset in which all alignments contained at least five of the seven taxa (70% complete matrix), and concatenated these into a PHYLIP supermatrix which I analysed using the best ML (GTRGAMMA) and bootstrap searches in *RAxML* v8.0.19 (Stamatakis 2014). I reconciled the best ML tree with the bootstrap replicates using *RAxML* v8.0.19.

DIPTERA

The bait design process for dipterans (Data S3) followed the same workflow I used to design the principal coleopteran bait set. I downloaded available genomes for several dipteran lineages (Table S3) and simulated paired-end reads at 2× coverage. After merging read pairs, I aligned the simulated reads to the genome sequence of *Aedes aegypti* (aedAeg1 hereafter) with a substitution rate of 0·05, converting the output to BAM format. After removing unaligned reads, I followed the workflow for coleopterans by creating a table of conserved regions shared between the base genome (aedAeg1) and exemplar taxa, outputting the list of intervals shared by all of the exemplar taxa and the base taxon, designing a temporary bait set from the aedAeg1 genome, and re-aligning the temporary baits to the genomes of exemplar taxa representing a diverse group of dipteran species (Table S3). As above, I extracted sequence in FASTA format for each conserved locus from each exemplar taxon assembly, designed a set of hybrid bait sequences targeting each of these loci from the genomes of each exemplar taxa, and filtered putative duplicate baits/loci from this set. I called the resulting file the principal dipteran bait set.

I performed an *in silico* check of the bait set by reconstructing the relationships between members of two dipteran clades (Culicomorpha and Drosophilidae), where relationships have previously been resolved with reasonable support (Drosophila 12 Genomes Consortium *et al.* 2007; van der Linde *et al.* 2010; Neafsey *et al.* 2015). To do this, I aligned the principal dipteran bait set to the genomes of additional dipteran lineages and an outgroup assembly from *Limnephilus lunatus* (GCA_000648945.1; Table S3). Then, I sliced FASTA sequences from each genome that flanked the conserved locus location by ±400 bp and performed additional orthology and duplicate screening steps (–min_coverage 67, –min_identity 80). I then created one dataset containing members of Culicomorpha with *L. lunatus* as an outgroup taxon, and I created a second dataset containing members of the Drosophilidae with *Musca domestica* and *Lucilia cuprina* as outgroup taxa. For each dataset, I followed the same alignment, alignment trimming, filtering (70% complete matrix) and analysis procedures described for Coleoptera.

HEMIPTERA

The bait design process for hemipterans (Data S4) was similar those previously described: I downloaded available genomes for hemipteran lineages (Table S4) and simulated reads from each genome at 2×

coverage. I merged read pairs and aligned the merged reads to the genome sequence of *Diaphorina citri* (GCA_000475195.1; diaPsy1 hereafter) with a substitution rate of 0·05. After removing unaligned reads, I followed the workflow for dipterans. I queried the resulting table of alignment intervals and output a list of intervals shared by the base taxon and three of the five exemplar taxa. I designed a temporary bait set targeting these loci from the diaPsy1 genome, re-aligned the temporary bait set to the available genomes of exemplar taxa representing hemipteran diversity (Table S4), designed a set of hybrid bait sequences targeting each conserved locus from the genomes of each exemplar taxon and filtered duplicate baits/loci from this set. I called the resulting file the principal hemipteran bait set.

I followed the same procedures described above to perform an *in silico* check of the bait set. The only differences were that I aligned the baits to the genomes of taxa I used to design the principal hemipteran bait set, as well as two additional genome-enabled hemipteran lineages and an outgroup thysanopteran genome, *Frankliniella occidentalis* (Table S4). I also performed the additional orthology and duplicate screening steps with slightly stricter parameters (−min_coverage 80, −min_identity 80). I followed the same alignment, alignment trimming, filtering and analysis procedures described for Coleoptera.

LEPIDOPTERA

Similar to the groups above, I downloaded available genomes for five lepidopteran lineages (Table S5, Data S5) and the genome assembly for *L. lunatus*, a caddisfly (Order Trichoptera). I simulated paired-end reads, and merged read pairs for alignment to the *Bombyx mori* genome (GCA_000151625.1; bomMor1 hereafter) with a substitution rate of 0·05. After removing unaligned reads, I followed the workflow for dipterans, although I did not merge aligned reads that were <100 bp from one another because subsequent filtering steps for removing duplicate loci also removed these overlapping regions. I queried the table of alignment intervals and output a list of intervals shared by the base taxon and all five of the exemplar taxa. I designed a temporary bait set from bomMor1 to target these loci, and I aligned the temporary bait set to the available genomes of exemplar taxa representing lepidopteran diversity (Table S5). From these matches, I designed a set of hybrid bait sequences targeting each conserved locus from the genomes of each exemplar taxon, and I filtered duplicate baits/loci from this set. I called the resulting file the principal lepidopteran bait set.

I performed an *in silico* check of the principal lepidopteran bait set following the same procedures described above. In addition to re-aligning baits to the genomes of taxa I used for the bait set design, I included genome assemblies from 15 lepidopterans as well as the outgroup assembly from *L. lunatus* (Table S5). I performed the orthology and duplicate screening steps with parameters identical to those used for dipterans, and I followed the same alignment, alignment trimming, filtering and analysis procedures described for Coleoptera, except that each alignment in the concatenated matrix contained data for at least 12 of the 16 taxa (75% complete matrix).

Results

ARACHNIDA

I collected an average of 39 M (95 CI: 7·3 M) sequencing reads from each low-coverage arachnid library (Table S1). An average of 1·45% (95 CI: 0·8%) of reads aligned to the limPol1 base genome sequence. After converting the alignments to BED format, merging overlapping alignment regions and filtering BEDs of short loci or loci that aligned to large repeat regions in the limPol1 genome assembly, I selected 5975 loci from the relational database that were shared by limPol1 and all six arachnid exemplars used for conserved locus identification. I designed a temporary bait set targeting 5733 of these loci identified in the limPol1 genome assembly, and I re-aligned the temporary baits to the genomes of nine arachnids and limPol1. I selected a set of 1168 conserved loci that were shared by limPol1 and at least five of the nine exemplar arachnid taxa, and I designed a hybrid bait set targeting these loci using the genomes of all nine arachnids and limPol1. After bait design and duplicate filtering, the principal arachnid bait set contained 14 799 baits targeting 1120 loci.

During *in silico* testing, I detected an average of 1029 conserved loci among arachnid genome assemblies and the outgroup (limPol1) genome assembly, while the average number of non-duplicate, conserved loci was 692·8 (95 CI: 59·1). The 70% complete matrix contained 550 trimmed alignments that were 399 bp in length (95 CI: 16·73), totalled 219 372 characters and contained 99 882 informative sites (mean per locus ± 95 CI: 182 ± 8). The resulting ML phylogeny (Fig. S1) reconstructed the established orders as monophyletic while recovering recognized relationships within spiders (Garrison *et al*. 2016) with high support at all nodes. Additional details regarding *in vitro* tests of this bait set can be found in (Starrett *et al*. in press).

COLEOPTERA

I simulated an average of 7·0 M (95 CI: 3·8 M) sequencing reads from each coleopteran genome assembly (Table S2), and approximately 1·2% (95CI: 0·2%) of these reads aligned to the triCas1 genome. After converting the alignments to BED format, merging overlapping regions and filtering BEDs of short loci or loci that overlapped repetitive regions in the triCas1 genome, I selected 1822 loci from the relational database that were shared by triCas1 and five exemplar taxa. I designed a temporary bait set from the triCas1 genome targeting 1805 conserved loci, and I aligned the temporary baits to the genomes of six coleopterans and the strepsipteran outgroup. I selected a set of 1209 conserved loci that were shared by triCas1 and at least four of the coleopteran and strepsipteran exemplar taxa, and I designed a hybrid bait set targeting these loci using the genomes of all seven coleopteran lineages and one strepsipteran lineage. The principal coleopteran bait set contained 13 674 baits targeting 1172 conserved loci.

During *in silico* testing, I detected an average of 994 conserved loci among coleopteran genome assemblies and the strepsipteran outgroup assembly, while the average number of non-duplicate, conserved loci detected in each taxon was 837·7 (95 CI: 105·9). After alignment and alignment trimming, the 70% complete matrix contained 865 loci that were 626·9 (95 CI: 9·8) bp in length, totalled 542 324 characters and contained 163 681 informative sites (mean per locus ± 95 CI: 189·2 ± 3·6). The resulting ML phylogeny reconstructed recognized relationships among coleopteran superfamilies (Mckenna, Wild & Kanda 2015) with high support at all nodes (Fig. S2).

DIPTERA

I simulated 2·9 M (95 CI: 0·3 M) sequencing reads from each of two dipteran genome assemblies (Table S3). An average of 2·1% (95 CI: 1·8%) of these reads aligned to the aedAeg1 genome. After converting the alignments to BED format, merging overlapping reads and filtering BEDs of short loci or loci that overlapped repetitive regions in the aedAeg1 genome, I selected 4904 conserved loci from the relational database that were shared by aedAeg1 and the two exemplar taxa. I designed a temporary bait set targeting these loci using the aedAeg1 genome assembly, and I aligned the temporary baits to the genomes of seven dipterans. I selected a set of 2834 conserved loci that were shared by aedAeg1 and at least four other dipteran genome assemblies, and I designed a hybrid bait set targeting these loci using the genomes of seven dipteran lineages. The principal dipteran bait set contained 31 328 baits targeting 2711 conserved loci.

During *in silico* testing, I detected an average of 2413 conserved loci among dipteran genome assemblies and the trichopteran outgroup assembly, while the average number of non-duplicate, conserved loci detected in each taxon was 1774·0 (95 CI: 213·6). I constructed two phylogenetic data matrices and inferred phylogenies from each. The 75% complete matrix for Culicomorpha contained 1202 loci that were 676·4 (95 CI: 11·2) bp in length, totalled 813 084 characters and contained 266 806 informative sites (mean per locus ± 95 CI: 222·0 ± 4·0). The resulting ML phylogeny (Fig. S3a) reconstructed the relationships among major mosquito/black fly lineages with high support at all nodes (Wiegmann *et al*. 2011), and the best ML topology was identical to a tree inferred from whole-genome sequence data (Neafsey *et al*. 2015). The 75% complete matrix for Drosophilidae contained 1658 loci that were 721·2 (95 CI: 8·3) bp in length, totalled 1 195 791 characters, and contained 471 185 informative sites (mean per locus ± 95 CI: 284·2 ± 3·5). The resulting ML phylogeny (Fig. S3b) reconstructed the relationships among and within drosophilid lineages with high support at all nodes, and the best ML topology was similar to those of other studies (van der Linde *et al*. 2010; Wiegmann *et al*. 2011; Neafsey *et al*. 2015). The primary difference in topology between the conserved element tree and topologies inferred by other studies was the placement of *D. willistoni* sister to the *virilis* + *repleta* + *grimshawi* groups + subgenus Sophophora, a difference that could be explained by rooting the conserved element tree on *M. domestica* + *L. cuprina*.

HEMIPTERA

I simulated 14·2 M (95 CI: 4·7 M) sequencing reads from each of the hemipteran genome assemblies (Table S4). An average of 0·7% (95 CI: 0·3%) of these reads aligned to the diaPsy1 genome assembly. After converting the alignments to BED format, merging overlapping reads, and removing short and repetitive loci, I selected 6210 loci from the relational database that were shared by diaPsy1 and three exemplar taxa. I designed a temporary bait set targeting these loci from the diaPsy1 genome assembly, and I aligned the temporary baits to the genomes of eight hemipterans. I selected a set of 2878 conserved loci shared by diaPsy1 and at least five of the hemipteran genome assemblies, and I designed a hybrid bait set targeting these loci using the genome assemblies of nine hemipteran lineages. The principal hemipteran bait set contained 40 207 baits targeting 2731 conserved loci.

During *in silico* testing, I detected an average of 2381 conserved loci among hemipteran genome assemblies and the thysanopteran outgroup assembly, while the average number of non-duplicate, conserved loci detected in each taxon was 1673·8 (95 CI: 223·1). The 75% complete matrix contained 1444 loci that were 386·4 (95 CI: 7·1) bp in length, and the concatenated data matrix contained 557 988 characters and 260 127 informative sites (mean per locus ± 95 CI: 180·1 ± 3·1). The resulting ML phylogeny (Fig. S4) reconstructed recognized relationships among hemipteran lineages (Cryan & Urban 2012), particularly those within Heteroptera (Wang *et al*. 2016), with high support.

LEPIDOPTERA

I simulated an average of 6·9M (95 CI: 1·3 M) sequencing reads from each of the lepidopteran genome assemblies (Table S5). An average of 4% (95 CI: 1·2%) of these reads aligned to the bomMor1 base genome sequence. After converting the alignments to BED format, merging overlapping alignment regions and filtering BEDs of short loci or loci that aligned to large repeat regions in the bomMor1 genome, I selected 2162 conserved loci from the relational database that were shared among bomMor1 and the five exemplar taxa used for conserved region identification. I designed a temporary bait set containing 4181 baits targeting 2120 loci in bomMor1, and aligned that to the genome sequence of each exemplar taxon. I selected a set of 1417 conserved loci that were shared by bomMor1 and at least three of the five exemplar taxa, and I designed a hybrid bait set targeting these loci using the genome assemblies of six lepidopteran lineages. After designing the baits and filtering duplicates, the principal lepidopteran bait set contained 14 363 baits targeting 1381 conserved loci.

During *in silico* testing, I detected an average of 1141 conserved loci among lepidopteran genome assemblies and the trichopteran outgroup assembly, while the average number of non-duplicate, conserved loci detected in each taxon was 920·6 (95 CI: 39·9). The 75% complete matrix contained 876 conserved loci that were 463·3 (95 CI: 17·2) bp in length, and the concatenated data matrix contained 405 849 characters and 158 187 informative sites (mean per locus ± 95 CI: 180·5 ± 7·1). The resulting ML phylogeny (Fig. S5) reconstructed lepidopteran relationships that largely agree with recent phylogenomic studies (Kawahara & Breinholt 2014; Cong *et al*. 2015). Relationships within Papilionoidea do not differ from other studies. However, the placement of Pyraloidea sister to Papilionoidea in the conserved element phylogeny conflicts with previous studies that suggest Pyraloidea is sister to Macroheterocera + Mimallonidae (Bazinet *et al*. 2013; Kawahara & Breinholt 2014). Bootstrap support for this

relationship is low, and a post hoc DensiTree (Bouckaert 2010) analysis of the bootstrap replicates (Fig. S6) suggests that one cause of the low support for this relationship, as well as a source of the low support for the node uniting the Macroheterocera, is instability regarding to the placement of the Pyraloidea in the concatenated phylogenetic analysis.

Discussion

I created a generalized workflow for (i) identifying conserved sequences shared among divergent genomes and (ii) designing enrichment baits to collect these conserved regions from DNA libraries for downstream phylogenetic and phylogeographic analyses. Application of this workflow to several diverse groups of arthropods suggests that the method identifies thousands of conserved loci shared among divergent taxa using a handful of relatively simple steps. *In silico* testing suggests that these enrichment baits can be used to collect data from hundreds of loci across entire organismal groups, and *in silico* results also suggest that each bait set can be extended to divergent outgroups with moderate success. *In vitro* testing of the bait set designed for arachnids (Starrett *et al*. in press) suggests that *in silico* tests provide a reasonably accurate measure of success when baits are used to collect sequence data from real DNA libraries. A separate effort using this workflow to update a hymenopteran bait set (Branstetter *et al*. in press) shows that the approach described here improves capture success. However, as with all newly designed target enrichment bait sets, readers should be cautioned that every bait set is 'experimental' until validated *in vitro*. The number of bait sets I designed using the workflow described above puts *in vitro* testing of each beyond the scope of this manuscript. However, each bait set is available, restriction-free, under a public domain license (CC-0), and individual research groups interested in testing remaining bait sets are free to use, modify and extend any of the arthropod bait sets I have developed.

The workflow presented here differs from related efforts by combining target locus identification with enrichment bait design (Mayer *et al*. 2016) and because the process of conserved locus identification that I used is agnostic to the class of loci being interrogated (Johnson *et al*. 2016). This means that the conserved loci identified by the workflow I describe can be exons, introns or intergenic regions. The set of conserved loci can be further subdivided into different classes using annotation information available from genomes to which the bait set is aligned or other data, such as transcript sequences. Furthermore, different algorithms for bait sequence selection and bait design (Mayer *et al*. 2016) can be applied to the conserved regions identified by the workflow I created to find improved or optimal bait designs.

Because the workflow I described is generalized, its application is not limited to specific vertebrate or invertebrate classes – any organismal group having some genomic resources can be used for locus identification and subsequent bait design. And, the locus identification process can be tailored by users to be more or less strict than the moderate approach I used for each arthropod group, a strategy that allows researchers to identify variable numbers of conserved loci shared among focal taxa that scales with the risk each research group is willing to accept. For example, targeting those few hundred loci found in six out of six divergent taxa representing a given organismal group is less risky than targeting those few thousand loci that are putatively shared by only three of six divergent taxa.

It is important to keep in mind that the workflow I designed attempts to produce target enrichment baits from orthologous loci, and the steps of the workflow try to ensure orthology by culling repetitive regions and using several rounds of sequence similarity searches and duplicate sequence removal during locus identification and bait design. While these steps help to ensure homology and reduce paralogy, they do not guarantee orthology. Unfortunately, orthology can be hard to validate, particularly when reasonably robust phylogenetic hypotheses of relationships within a given organismal group do not exist. For organismal groups where sufficient data exist, additional orthology assessment steps could and likely should be implemented by laboratories using this workflow to identify conserved genomic elements and design enrichment baits to target them. These steps include filtering BAM alignments for only unique matches (*samtools* -q), quantifying sequence divergence between taxa at each conserved locus and removing 'outlier loci' before bait design, or using gene trees derived from simulated or empirical studies to test whether the conserved loci identified using this pipeline show topological patterns consistent with duplication events (i.e. they are paralogous). For organismal groups where there are insufficient data to conduct these tests, this workflow and/or the baits I designed may provide a good 'first-pass' mechanism for collecting empirical data that can then be subjected to these additional tests.

By making all of the design steps, documentation, software code and bait sets developed here available under an open-source license, I hope that the workflow I described will facilitate the collection of genome scale data from a diversity of organismal groups and provide additional insight into common and different patterns of diversification we see across the Tree of Life.

Acknowledgements

I thank two anonymous reviewers and M. Gilbert, whose comments improved this manuscript. I also thank the genome sequencing centres and individual research groups who make their genome assemblies publicly available – the work I described above would not be possible without these phenomenally useful resources. Specific thanks are due to Richard K. Wilson and The Genome Institute, Washington University School of Medicine for providing access to the *L. polyphemus* genome assembly through NCBI Genbank. I also thank the Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu), who made all of the data generated as part of the i5k Initiative (i5K Consortium 2013) available to others. G. Dasch provided early access to the *Amblyomma americanum* genome assembly (NCBI PRJNA374336). A. Chase assisted with DNA library preparation and M. Branstetter, R. Bryson, S. Derkarabetian, T. Glenn, M. Hedin, J. McCormack, N. Pourmand and J. Starrett contributed, in various ways, to the development process described above. C. Carlton, M. Forthman, C. Mitter, K. Noble and C. Weirauch provided comments on phylogenetic trees inferred during *in silico* tests for different organismal groups, although any errors in describing those trees are my own. This work was supported by startup funds from Louisiana State University, with additional computational support from NSF DEB-1242260. Parts of this work were also encouraged by DEB-1352978 to David O'Brochta, whose invitation to speak at

the IGTRCN symposium in 2014 spurred me to design baits targeting conserved loci in different insect groups. This study was also supported, in part, by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology. Other portions of this research were conducted with high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu).

Conflicts of interest

None declared.

Data accessibility

I have integrated the workflow described above as a major revision to the open-source PHYLUCE package (v1.6+; https://www.github.com/faircloth-lab/phyluce/). A generalized tutorial implementing the methods described as part of this manuscript is available from http://phyluce.readthedocs.io/en/latest/tutorial-four.html.

Raw reads used to identify conserved loci in arachnids are available from NCBI PRJNA324685 or directly from the SRA (Table S1). Supplemental data, including tables, figures and files describing the bait design process, the final bait sets designed for this manuscript, and data from *in silico* testing are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.v0k4h; Faircloth 2017). All principal bait sets are also available from FigShare (https://doi.org/10.6084/m9.figshare.c.3472383), where I will maintain updated/improved versions.

References

Ali, O.A., O'Rourke, S.M., Amish, S.J., Meek, M.H., Luikart, G., Jeffres, C. & Miller, M.R. (2015) RAD capture (rapture): flexible and efficient sequence-based genotyping. *Genetics*, **202**, 389–400.

Baird, N., Etter, P., Atwood, T., Currey, M., Shiver, A., Lewis, Z., Selker, E., Cresko, W. & Johnson, E. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bazinet, A.L., Cummings, M.P., Mitter, K.T. & Mitter, C.W. (2013) Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study *PLoS ONE*, **8**, e82615.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

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**, 403.

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**, 6018–6032.

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**, 271.

Blaimer, B.B., Lloyd, M.W., Guillory, W.X. & Brady, S.G. (2016) Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS ONE*, **11**, e0161531.

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bouckaert, R.R. (2010) DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, **26**, 1372–1373.

Bradnam, K.R., Fass, J.N., Alexandrov, A. *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, **2**, 10.

Branstetter, M.G., Longino, J.T., Ward, P.S. & Faircloth, B.C. (in press) Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, 13943696. doi: 10.1111/2041-210X.12742.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N.V. (2015) Skipper genome sheds light on unique phenotypic traits and phylogeny. *BMC Genomics*, **16**, 639.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K. & Glenn, T.C. (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, **8**, 783–786.

Cryan, J.R. & Urban, J.M. (2012) Higher-level phylogeny of the insect order Hemiptera: is Auchenorrhyncha really paraphyletic? *Systematic Entomology*, **37**, 7–21.

Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.

Dunn, C.W., Hejnol, A., Matus, D.Q. *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.

Earl, D., Bradnam, K., St John, J. *et al.* (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research*, **21**, 2224–2241.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. & Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Faircloth, B.C. (2013) illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. Available at: https://github.com/faircloth-lab/illumiprocessor (accessed 19 September 2015).

Faircloth, B.C. (2015) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, **32**, 786–788.

Faircloth, B.C. (2017) Data from: Identifying conserved genomic elements and designing universal probe sets to enrich them. *Dryad Digital Repository*, http://dx.doi.org/10.5061/dryad.v0k4h.

Faircloth, B.C. & Glenn, T.C. (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS ONE*, **7**, e42543.

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**, 717–726.

Faircloth, B.C., Sorenson, L., Santini, F. & Alfaro, M.E. (2013) A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS ONE*, **8**, e65923.

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. *Molecular Ecology Resources*, **15**, 489–501.

Garrison, N.L., Rodriguez, J., Agnarsson, I. *et al.* (2016) Spider phylogenomics: untangling the Spider Tree of Life. *PeerJ*, **4**, e1719.

Gnirke, A., Melnikov, A., Maguire, J. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Hardenbol, P., Banér, J., Jain, M. *et al.* (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology*, **21**, 673–678.

Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. PhD thesis, The Pennsylvania State University.

Harvey, M.S. (2002) The neglected cousins: what do we know about the smaller arachnid orders? *The Journal of Arachnology*, **30**, 357–372.

Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C. & Brumfield, R.T. (2016) Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, **65**, 910–924.

Hoffberg, S., Kieran, T.J., Catchen, J.M., Devault, A., Faircloth, B.C., Mauricio, R. & Glenn, T.C. (2016) RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. bioRxiv, **16**, 1264–1278.

Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L. & Kimball, R.T. (2015) Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, **33**, 1110–1125.

Huang, W., Li, L., Myers, J.R. & Marth, G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Hugall, A.F., O'Hara, T.D., Hunjan, S., Nilsen, R. & Moussalli, A. (2016) An exon-capture system for the entire class Ophiuroidea. *Molecular Biology and Evolution*, **33**, 281–294.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J.C. & Wickett, N.J. (2016) HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, **4**, 1600016.

i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *The Journal of Heredity*, **104**, 595–600.

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

Kawahara, A.Y. & Breinholt, J.W. (2014) Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proceedings. Biological Sciences/The Royal Society*, **281**, 20140970.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. & Proc, 1000 Genome Project Data (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lim, H.C. & Braun, M.J. (2016) High-throughput SNP genotyping of historical and modern samples of five bird species via sequence capture of ultraconserved elements. *Molecular Ecology Resources*, **16**, 1204–1223.

van der Linde, K., Houle, D., Spicer, G.S. & Steppan, S.J. (2010) A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genetics Research*, **92**, 25–38.

Lunter, G. & Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Manthey, J.D., Campillo, L.C., Burns, K.J. & Moyle, R.G. (2016) Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: Piranga). *Systematic Biology*, **65**, 640–650.

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

McCormack, J.E., Tsai, W.L.E. & Faircloth, B.C. (2015) Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, **16**, 1189–1203.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T. & Glenn, T.C. (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome Research*, **22**, 746–754.

McCormack, J.E., Harvey, M.G., Faircloth, B.C., Crawford, N.G., Glenn, T.C. & Brumfield, R.T. (2013) A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS ONE*, **8**, e54848.

McGee, M.D., Faircloth, B.C., Borstein, S.R., Zheng, J., Darrin Hulsey, C., Wainwright, P.C. & Alfaro, M.E. (2016) Replicated divergence in cichlid radiations mirrors a major vertebrate innovation. *Proceedings of the Royal Society B*, **283**, 20151413.

Mckenna, D.D., Wild, A.L. & Kanda, K. (2015) The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Systematic Entomology*, **40**, 835–880.

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

Neafsey, D.E., Waterhouse, R.M., Abai, M.R. *et al.* (2015) Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, **347**, 1258522.

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**, 1000–1010.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. & Hoekstra, H.E. (2012) Double digest RADseq: an inexpensive method for de novo snp discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

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

Rohland, N. & Reich, D. (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C. & Jaffe, D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biology*, **14**, R51.

Ruane, S. & Austin, C.C. (2017) Phylogenomics using formalin-fixed and 100+ year old intractable natural history specimens. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12655

Sanders, K.L. & Lee, M.S.Y. (2010) Arthropod molecular divergence times and the Cambrian origin of pentastomids. *Systematics and Biodiversity*, **8**, 63–74.

Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G. & Dunn, C.W. (2011) Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, **480**, 364–367.

Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C. & Brumfield, R.T. (2014) Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales. *Systematic Biology*, **63**, 83–95.

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

Starrett, J., Derkarabetian, S., Hedin, M., Bryson Jr., R.W., McCormack, J.E. & Faircloth, B.C. (in press) High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12621

Streicher, J.W. & Wiens, J.J. (2016) Phylogenomic analyses reveal novel relationships among snake families. *Molecular Phylogenetics and Evolution*, **100**, 160–169.

Suchan, T., Pitteloud, C., Gerasimova, N.S., Kostikova, A., Schmid, S., Arrigo, N., Pajkovic, M., Ronikier, M. & Alvarez, N. (2016) Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE*, **11**, e0151651.

Talavera, G. & Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

Wang, Y.-H., Cui, Y., Rédei, D. *et al.* (2016) Phylogenetic divergences of the true bugs (Insecta: Hemiptera: Heteroptera), with emphasis on the aquatic lineages: the last piece of the aquatic insect jigsaw originated in the Late Permian/Early Triassic. *Cladistics: The International Journal of the Willi Hennig Society*, **32**, 390–405.

Wiegmann, B.M., Trautwein, M.D., Winkler, I.S. *et al.* (2011) Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 5690–5695.

Zhang, Z.-Q. (2011) Phylum Arthropoda von Siebold, 1848. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: an outline of higher-level classification and survey of taxonomic richness. *Zootaxa*, **3148**, 99–103.

Received 16 December 2016; accepted 30 January 2017
Handling Editor: M. Gilbert

Supporting Information

Details of electronic Supporting Information are provided below.

Fig. S1. Maximum likelihood phylogeny inferred from *in silico* testing of baits targeting conserved loci in Arachnida.

Fig. S2. Maximum likelihood phylogeny inferred from *in silico* testing of baits targeting conserved loci in Coleoptera.

Fig. S3. Maximum likelihood phylogeny inferred from *in silico* testing of baits targeting conserved loci in Diptera.

Fig. S4. Maximum likelihood phylogeny inferred from *in silico* testing of baits targeting conserved loci in Hemiptera.

Fig. S5. Maximum likelihood phylogeny inferred from *in silico* testing of baits targeting conserved loci in Lepidoptera.

Fig. S6. DensiTree plot of bootstrap replicates demonstrating instability regarding placement of the Pyraloidea in the concatenated phylogenetic analysis.

Table S1. Arachnid species used for conserved locus identification, bait design and *in silico* testing of the resulting bait design.

Table S2. Coleopteran species used for conserved locus identification, bait design and *in silico* testing of the resulting bait design.

Table S3. Dipteran species used for conserved locus identification, bait design and *in silico* testing of the resulting bait design.

Table S4. Hemipteran species used for conserved locus identification, bait design and *in silico* testing of the resulting bait design.

Table S5. Lepidopteran species used for conserved locus identification, bait design and *in silico* testing of the resulting bait design.