Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

Understanding UCEs: A comprehensive primer on using Ultraconserved Elements for arthropod phylogenomics

Y. Miles Zhang*, Jason L. Williams, Andrea Lucky

University of Florida, Department of Entomology & Nematology, Gainesville, FL, 32608

*** Corresponding author email: Yuanmeng.zhang@gmail.com**

Abstract:

Targeted enrichment of ultraconserved elements (UCE) has emerged as a promising tool for inferring evolutionary history in many taxa, with utility ranging from phylogenetic and phylogeographic questions at deep time scales to population level studies at shallow time scales. However, the methodology can be daunting for beginners. Our goal is to introduce UCE phylogenomics to a wider audience by summarizing recent advances in arthropod research, and to familiarize readers with background theory and steps involved. We define terminology used in association with the UCE approach, evaluate current laboratory and bioinformatic methods and limitations, and, finally, provide a roadmap of steps in the UCE pipeline to assist phylogeneticists in making informed decisions as they employ this powerful tool. By facilitating increased adoption of UCE in phylogenomics studies that deepen our comprehension of the function of these markers across widely divergent taxa, we aim to ultimately improve understanding of the arthropod tree of life.

Keywords: Arachnida, Insecta, Phylogenomics Methods, Target Enrichment, Ultraconserved Elements

© 2019 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

22

23 Introduction & Background

24 The advent of massively-parallel sequencing technology and the subsequent emergence
25 of the field of phylogenomics has invigorated evolutionary biology in a relatively short time span
26 (reviewed in Philippe et al. 2011, Jones and Good 2016). This molecular revolution has offered
27 unprecedented opportunities to generate large-scale datasets, and with the concurrent explosion
28 of analytic and bioinformatics tools, has made it possible to address previously intractable
29 challenges due to limited genetic markers. However, the rapidity with which new technologies
30 have emerged has made it difficult for scientists to stay up to date about useful new tools;
31 understanding the steps involved in using new methods presents a challenge for researchers.

32 Genome-scale studies are rapidly supplanting the Sanger sequencing-based, multi-locus
33 molecular phylogenetic methods that dominated from the mid-1990's through the early 2000's;
34 today, genomic-scale studies dwarf previous approaches in the sheer scale of data they generate
35 (Bravo et al. 2019). While the cost and scale of whole-genome sequencing are prohibitive for
36 many researchers, recent advances in sequencing technology and laboratory protocols have made
37 it possible to generate high quality genomic datasets using a combination of next-generation
38 sequencing, genomic reduction, and sample multiplexing (Lemmon and Lemmon 2013,
39 McCormack et al. 2013a). These so-called 'genome reduction' or 'reduced representation'
40 approaches can rapidly generate datasets with thousands of loci, at relatively low cost, for model
41 and non-model taxa alike. Methods such as restriction enzyme-associated DNA sequencing
42 (RADseq; Miller et al. 2007, Baird et al. 2008, Peterson et al. 2012), transcriptomics (Bi et al.
43 2012), and target enrichment methods such as Anchored Hybrid Enrichment (AHE) (Lemmon et

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

al. 2012) or target capture of Ultraconserved Elements (UCE) (Faircloth 2017) are now widely used for generating genomic-scale data for phylogenomic studies. These phylogenomics methods are similar in some respects, but each has strengths and weaknesses which may not be easily discerned by researchers new to this field. Because of the proliferation of new approaches and tools in phylogenetics, selecting a method to use in the era of 'big data' can be daunting. Potential users need guidance in choosing methods appropriate to their research questions, and in navigating confusing terminologies, bioinformatics-heavy data processing, and computationally intensive analyses.

UCE-based phylogenomics continues to develop rapidly, and the lack of comprehensive review has been a significant challenge for potential users to overcome when exploring this option. This paper summarizes recent advances in UCE phylogenomics in arthropod research; we start by familiarizing readers with background theory and terminology, and describing the steps involved in generating and analyzing UCE data, and then provide quality-control tips to ensure that data collection and downstream analyses can be performed with confidence.

What are UCEs?

Ultraconserved Elements are highly-conserved regions within the genome that are shared among evolutionarily distant taxa (Bejerano et al. 2004). The DNA adjacent to each 'core' UCE region, known as flanking DNA, increases in variability with distance from the region (Faircloth et al. 2012). UCEs and flanking regions can be selectively captured, and used to reconstruct the evolutionary history of taxa at various time scales, from deep to shallow phylogenetic divergences (Faircloth et al. 2012, McCormack et al. 2012).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

The UCE approach belongs to the broad category of 'target enrichment' phylogenomic techniques, which involve selective capture of genomic regions from DNA prior to sequencing (Mamanova et al. 2010). Similar methods include AHE (Anchored Hybrid Enrichment), BaitFisher (Mayer et al. 2016), and Hyb-Seq (Weitemier et al. 2014). AHE has been the most widely used method for animal studies, to date, but all target enrichment methods have been successfully used across a variety of taxa. These techniques universally involve identifying loci of interest, then designing custom-made molecular probes (also known as baits) which are hybridized to loci of interest, and ultimately sequencing selected genomic region on a massively-parallel platform. The main difference between the AHE and UCE approaches is the nature of the loci targeted; AHE focuses on fewer loci (300-600) that are exclusively exonic, while UCEs target more loci (>1000) using fewer probes – these may include both exonic and intronic regions, depending on the organism (Crawford et al. 2012, McCormack et al. 2012, Faircloth et al. 2015). While AHE can cope with sequence variation at target loci by using a more diverse set of probes per locus, the details of the methodology are not available for scrutiny as they are, in part, proprietary (Lemmon et al. 2012). The UCE approach, in contrast, is fully open source, which has contributed to recent interest in using these markers for arthropod phylogenomics.

Advantages of UCE Phylogenomics

The UCE approach has become an increasingly popular target enrichment method for generating phylogenomic data, as it offers advantages over traditional Sanger sequencing methods in terms of quantity of data generated. UCEs have successfully been used in studies across a broad array of animal taxa including birds (McCormack et al. 2013b, Musher and Cracraft 2018), mammals (McCormack et al. 2012, Mclean et al. 2018), fish (Faircloth et al.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

2013, Alda et al. 2018), amphibians (Newman and Austin 2016, Zarza et al. 2018), reptiles (Crawford et al. 2012, Streicher and Wiens 2017, Myers et al. 2019), sponges (Ryu et al. 2012), cnidarians (Quattrini et al. 2018), echindoderms (Ryu et al. 2012), and arthropods (Faircloth et al. 2015, Baca et al. 2017b, Branstetter et al. 2017c, Hedin et al. 2018b, Kieran et al. 2019). These studies range widely in evolutionary scale, from phylogenetic and phylogeographic questions at deep time scales (Faircloth et al. 2013, Smith et al. 2014, Branstetter et al. 2017c) to population level studies at shallow time scales (Harvey et al. 2016, Manthey et al. 2016, Zarza et al. 2018, Branstetter and Longino 2019, Myers et al. 2019).

Benefits of using UCE include openly shared resources such as probe sets (https://www.ultraconserved.org/), lab protocols (https://baddna.uga.edu/protocols.html/), and bioinformatics tools (https://phyluce.readthedocs.io/en/latest/), making it an easy method to learn and use in comparison to more proprietary alternatives such as AHE. Complete library preparation for around 100 samples can be completed in approximately two weeks by one person, or a month if counting DNA extraction and possible troubleshooting. UCE datasets can be easily standardized, even from multiple studies, by using the same probe set. In this way, data from studies using the same probe set or with exon/transcriptome data (Bossert et al. 2019, Kieran et al. 2019) can be combined and can incorporate legacy methods if the probe set includes Sanger genes (Branstetter et al. 2017a). These are distinct advantages over restriction enzyme-based methods such as traditional RADseq, which lacks repeatability due to the random nature of the restriction enzyme digestion that generates random genomic fragments. An additional advantage of target enrichment methods is the high success rate with degraded or low-quantity samples; older, dried museum specimens may be unusable in traditional restriction enzyme-based and transcriptomics studies as they require large quantities of high-quality DNA or RNA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

from fresh or carefully-preserved tissues (Blaimer et al. 2016a, Lim and Braun 2016, Ruane and Austin 2017). It is worth noting that newer RAD-based methods such as RADcap (Hoffberg et al. 2016), Rapture (Ali et al. 2016), and hyRAD (Suchan et al. 2016) address these limitations by using a combination of restriction enzyme digestion and hybridization capture probes to overcome traditional RAD-based problems such as allele dropout, and can successfully capture degraded DNA from older museum samples.

UCE and Arthropods

The UCE approach was first demonstrated outside of vertebrates in the insect order Hymenoptera (See Table 1). To date, two published probe sets exist for Hymenoptera: hym-v2 (31,829 probes for 2,590 UCEs, Branstetter et al. 2017a) includes most of the original hym-v1 probe sets (2,749 probes for 1,510 UCEs, Faircloth et al. 2015), and excludes poorly performing loci. Other arthropods groups for which published UCE data exist include Arachnida, Coleoptera, and Hemiptera; as well as multiple upcoming studies for Diptera (E. Buenaventura, C. Cohen, K. Noble, *pers. comm.*). Psocodea and Lepidoptera probe sets have been developed but not yet tested *in vitro* (Table 1). The utility of UCEs extends beyond purely phylogenetic and taxonomic research. For example, UCE-based community phylogenomics has been used to reveal the importance of bee phylodiversity in agriculture (Grab et al. 2019); UCE-based phylogeny and geometric morphometrics have been used in combination to explore the evolution of parasitic wasp body shape (Santos et al. 2019); single nucleotide polymorphisms (SNPs) generated from UCE data have been used to demonstrate the success of unsupervised machine learning in species delimitation of harvestmen (Derkarabetian et al. 2019), and UCE phylogenies have been integrated with environmental niche modeling to examine phylogeographic patterns of ants across the Brazilian Atlantic Forest (Ströher et al. 2019).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

This study compiles all currently available UCE-based literature related to arthropods as of July 2019 (n = 32, Figure 1), but will undoubtedly increase exponentially in the future (Table 1). Our aim is to provide a step-by-step guide to make the UCE research pipeline more approachable to researchers working across different arthropod groups.

UCE Phylogenomics Pipeline

The steps in the UCE pipeline are 1) probe selection and design; 2) wet lab work and sequencing; 3) bioinformatics; and 4) phylogenomic analyses. Below, we visualize the process in a workflow diagram (Figure 2) and describe the choices a researcher must make at each stage. A glossary of technical terms is provided as a supplementary document (S1).

Probe Selection & Design

Probe sets are a collection of oligonucleotides that will bind to specific, conserved genome regions of interest, often called baits as they can 'fish' out the region of interest. These probes are sometimes interchangeably called 'baits' as they are used to fish out target loci from a 'pond' of randomly sheared, adaptor-ligated DNA (Gnirke et al. 2009). However, to avoid confusion we recommend reserving the term 'baits' for the intermediate stage in probe design; by contrast, 'probes' refer to the final products that is synthesized for commercial use (Gustafson et al. 2019). A probe set functions through a collection of biotinylated oligonucleotides that are designed to bind with specific genome regions of interest. Probes are combined with denatured and cooled DNA, allowing for 'in solution' hybridization to targets. Streptavidin-coated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

magnetic beads, which have high affinity for biotin, are added into the solution. The beads then bind to the probe-DNA hybrids through the biotin on the probe set. Any unwanted DNA fragments are then washed away, leaving only the desired regions attached to the beads (Gnirke et al. 2009).

Probes are designed based on UCE loci identified from published genomes for each taxonomic group. Currently probe sets for Arachnida, Coleoptera, Diptera, Hemiptera, and Hymenoptera are available for purchase through Arbor Biosciences (https://arborbiosci.com/products/uces/). Other taxonomic groups either have no probe sets available, or have not been tested *in vitro*. Designing new probe sets may prove challenging in the absence of published genomes for a group of interest. Nevertheless, low coverage genome sequencing (5x) may be an increasingly affordable appropriate first step (Zhang et al. 2019). The sequenced genomes selected as the basis for probe design should ideally reflect diversity within the group of interest. Ideally multiple genomes should be used for probe design, but minimally probe sets designed based on only two genomes (hym-v1) were shown to be successful in capturing UCEs across the diverse order Hymenoptera (Faircloth et al. 2015).

Whereas probe selection is straightforward (they are either available for the group of interest or they are not), probe design for new taxonomic groups is a time-consuming process for any target enrichment method, as the probe sets can differ in number and composition depending on the target taxa and evolutionary scale. Currently published probe sets for arthropods target 1,100 – 2,700 UCEs loci, and have been made publicly available under public domain license (CC-0), thus allowing for restriction-free commercial synthesis, testing, use and improvements by other research groups (http://ultraconserved.org/#protocols) (Branstetter et al. 2017a, Faircloth 2017, Gustafson et al. 2019). A generalized workflow for identifying conserved

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

sequences shared among divergent genomes and enrichment probes design is available (Faircloth 2017), and a new pipeline has been described using low-coverage genome sequencing that can also be used to design UCE probes (Zhang et al. 2019). In brief, the probe design sequence is 1) select base genome (s); 2) generate short reads as exemplars of the focal group's diversity and align to base genome(s); 3) merge approximate reads and find overlapping regions shared among exemplar taxa and base genome (conserved regions); 4) design temporary bait set from base genome against conserved regions and align to exemplar genome assemblies to remove duplicates; 5) design exemplar-specific probes for each locus where temporary baits match exemplar genome assemblies.

How to best optimize the probe design process is an area of active research. Both base genome choice and initial bait design stringency parameters can greatly affect the number of resultant probes and, subsequently, the number of loci detected and recovered in Adephagan beetles (Gustafson et al. 2019). The optimal base genome can be selected by conducting a base genome experiment by iteratively selecting each taxon as the base genome and finding candidate loci shared among exemplar taxa, or selected from taxon with the smallest average genetic distance to the other exemplar taxa through independently generated Sanger markers. Probe sets can also be modified to incorporate additional loci. The Hymenoptera probe set hym-v1 was improved by the publication of hym-v2, which included most of the original hym-v1 loci as well as new loci and probes targeting 16 commonly-sequenced nuclear genes to allow for 'back compatibility' with Sanger-era data (Branstetter et al. 2017a). The resulting capability of combining new genomic data with older sequences obtained from 'legacy' markers is vital to phylogenetic studies, as DNA quality tissue for many rare but vital taxa to phylogenetic studies may be difficult or impossible to obtain repeatedly. *In silico* tests of existing probe sets

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

demonstrate moderate success with sister outgroups, such as using the Hemiptera probe set to capture UCEs from thrips (Insecta, Order: Thysanoptera) (Faircloth 2017). Importantly, expense, time, and computational resources needed should be taken into consideration when designing new probe sets. The cost of development should be weighed against the potential future use of the probe set beyond the initial study; UCE probes for larger clades, for example, may be more likely to be adopted for multiple uses than those designed for species-poor groups.

Wet Lab Work & Sequencing

Specimen selection and DNA extraction. Selecting appropriate specimens for DNA extraction is vital to any phylogenetic endeavor. The first major requirement for molecular phylogenetics is to capture high-quality DNA. DNA capture success rates can be negatively affected by specimen age and preservation method (Short et al. 2018); Arthropod studies are often limited by DNA degradation, as most natural history collections have historically preserved specimens dry (pinned) or stored in 70% ethanol at room temperature which can lead rapidly deterioration of DNA (Short et al. 2018). Other complications include the number of freeze/thaw cycles (as few as possible), and the number/frequency of alcohol changes (regular enough to maintain 95% EtOH concentration and keep specimens submerged).

The degraded DNA of older specimens preserved by less-than-ideal methods can, fortunately, be captured by massive-parallel methods successfully incorporate shorter, more degraded DNA fragments than can be used for sanger sequencing. One illuminating study generated nearly 1000 UCEs loci from pinned bee specimens up to 121 years old (Blaimer et al. 2016a). This study demonstrated that pinned specimens less than 20 years old had significantly higher pre- and post-library concentrations, UCE contig lengths, and locus counts compared to older specimens. The small size (<5mm), and often corresponding low DNA yield, of many

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

arthropod specimens is another challenge to successful capture of genomic data, and a problem that may be exacerbated using non-destructive sampling to retain voucher specimens. Total yield of genetic material can be increased with the use of DNA amplification kits, albeit at a higher cost (Cruaud et al. 2018). UCE data has been successfully generated from minute, non-destructively sampled chalcidoid wasps (average DNA input = 25ng, Cruaud et al. 2019). This study used a modified protocol to maximize DNA yield from a commercially available extraction kit (Qiagen DNeasy Blood and Tissue Kit, Valencia, CA), by using LoBind tubes (Eppendorf) and heating the elution buffer for longer periods, while decreasing the number of purification steps. While no correlation between input DNA quantity and number of UCE loci captured in this dataset, a large amount of missing data ultimately resulted in a data matrix that was only 25% complete. Ultimately, in order to ensure high quality DNA generation, using fresh, well-preserved specimens preserved in 95% EtOH and stored in -80°C or -20°C, is recommended. Pinned specimens collected within the past 20 years are also suitable. Destructive sampling will likely generate higher DNA yield, but for rare specimens non-destructive soaking within lysis buffer will suffice.

Careful selection of tissue types can significantly lower the potential of contamination by non-target organisms. Precautions can be taken by decontaminating specimens using UV light, as well as separating areas used for DNA extraction from amplification areas (Yeates et al. 2016). Additional recommendations include removing appendages used by predators to capture prey (Bossert and Danforth 2018), targeting life stages, such as adults, that are less likely to host endoparasitoids. Contamination can also be reduced by using either strict bioinformatic processing parameters, or methods such as the *phyluce_assembly_match_contigs_to_barcodes*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

247 script in PHYLUCE which extracts the *COI* barcode region, which is used for validating the
248 presence of a single or multiple species (Bossert and Danforth 2018).

249 *Library Preparation.* Once DNA has been extracted from target organisms, wet lab
250 protocol for preparing the DNA libraries for sequencing varies little across taxa. Depending on
251 the quality of DNA or level of DNA degradation, the extracted and quantified genomic DNA
252 may need to be sheared using sonication or enzymatic digestion to reach the target size of 400–
253 600bp. The degree of DNA degradation will determine the duration of sonication needed; this
254 can be assessed using gel electrophoresis, or automated electrophoresis systems such as
255 TapeStation or Bioanalyzer.

256 At this stage, UCE sample preparation consists of seven main steps: 1) DNA
257 quantification; 2) adaptor ligation; 3) PCR amplification and initial pooling of specimens; 4)
258 hybrid enrichment; 5) amplification of enriched libraries; 6) Quantification and final pooling;
259 and 7) size selection and final quantification (detailed in (Branstetter et al. 2017a)).

260

261 **Bioinformatics**

262 Once sequencing is complete it is time to proceed to data analysis. Like other genomic
263 datasets, one of the advantages of UCEs is the volume of data returned; managing datasets at this
264 scale also presents a challenge to researchers new to genomics. Processing UCE data involves
265 three principal steps: 1) demultiplexing, filtering, and trimming the raw Illumina reads; 2) contig
266 assembly; and 3) UCE processing for phylogenomic analysis. Currently, the most widely-used
267 bioinformatics pipeline for UCE data processing is PHYLUCE (Faircloth 2015), which includes
268 a suite of Python wrapper scripts for these steps by calling other programs (detailed below) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

batch processing many samples at once. Additional bioinformatic programs not currently included within PHYLUCE can also be used to process data, as data can easily be imported back into the pipeline. Alternatively, the SECAPR (Andermann et al. 2018) pipeline also functions similarly to PHYLUCE and can be used for batch processing of UCE data, while MitoFinder (Allio et al. 2019) pipeline can be used to extract both UCE and mitogenomic data.

1*) Demultiplexing, filtering and trimming of raw Illumina reads*. Analyzing Illumina data always begins with batch trimming of adapters and low-quality bases of de-multiplexed data. In the PHYLUCE pipeline this is achieved using Illumiprocessor (Faircloth 2013), which is built around the Trimmomatic program (Bolger et al. 2014). Alternatively, external trimming programs such as Trim Galore! (https://github.com/FelixKrueger/TrimGalore) can be used instead of Illumiprocessor.

2) *Contig Assembly*. Currently PHYLUCE supports multiple programs such as velvet (Zerbino and Birney 2008), Trinity (Grabherr et al. 2011), ABySS (Simpson et al. 2009), and SPAdes (Bankevich et al. 2012) for genome assembly. While Trinity has been the most widely used of the assembly methods in published papers, updates to PHYLUCE are in the process of eliminating compatibility with Trinity due to technical issues. Both ABySS and velvet require an input for *k*-mer value, which is as part of the De Bruijn graph assembly algorithm. Smaller *k*-mers result in the assembly of shorter contigs with more connections, while large *k*-mers can result in longer but fewer contigs. However, it is difficult to determine the *k*-mer size for UCE data as the depth of coverage for each locus is variable due to capture efficiency. Therefore, testing multiple *k*-mer values is recommended, starting at the default of 35 and moving up to 55–65 to find the best trade-off in terms of contig size vs. *k*-mer number. Automatic estimation of *k*-mers is possible using SPAdes or the VelvetOptimiser wrapper script along with velvet.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

Mitogenomic assemblers, such as MetaSPAdes (Nurk et al. 2017), are a promising alternative to currently used genomic and transcriptomic *de novo* assemblers (Allio et al. 2019). These new tools are designed to account for variance in sequencing coverage, and are thus capable of generating larger and more complete supermatrices in a fraction of the time required by Trinity.

3) *UCE processing for phylogenomics*. Once assembled, contigs must be processed to determine which ones represent enriched UCEs loci. Orthologs are identified by aligning the assembled contigs to a FASTA file of target enrichment baits, and paralogs are subsequently removed (Faircloth 2015). The output is then screened to identify 1) assembled contigs match by probes targeting different loci, and 2) different contigs match by probes targeting the same loci. The latter must be removed from downstream analysis because they will be identified as potentially paralogous genes by PHYLUCE (Faircloth 2015), which can be problematic if the probes are not well-designed (see Current and Future Challenges below). The resulting FASTA files are then aligned using MUSCLE (Edgar 2004) or MAFFT (Katoh et al. 2002) within PHYLUCE, followed by trimming for data matrix completeness using GBlocks (Castresana 2000) or TrimAl (Capella-Gutiérrez et al. 2009). Finally, the completed data matrices can be exported in a variety of commonly used formats (e.g. phylip, nexus, etc.) for downstream phylogenomic analyses.

Allelic phasing

Allelic phasing is an additional, optional data processing step that extracts SNPs from UCE loci by separating (phasing) the heterozygous sites into two allele sequences; this approach can be used to increase resolution for shallow-level phylogenetic or species delimitation studies (Zarza et al. 2018, Andermann et al. 2019, Derkarabetian et al. 2019). Allelic phasing can be has been shown to provide more accurate estimation of tree topology and divergence times than using contig sequences, especially at shallow phylogenetic levels under multispecies coalescent

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

(MSC) models (Andermann et al. 2019), and can be performed in both PHYLUCE and SECAPR. This is in part due to common assembler programs not originally designed for heterozygous sequences or genomes, and as a result contig sequences generated by these programs will mask information by eliminating one of the two variants at a heterozygous site (Bodily et al. 2015). Another benefit of phasing the sequence doubles the sample size, as each diploid individual will have two strands of DNA sequences (Andermann et al. 2019). While this isn't always necessary for deep level phylogenomic studies, we recommend performing allelic phasing for UCE datasets intended for shallow-scale evolutionary studies, such as species delimitation or population genomics. However, sufficient sequence coverage is needed to ensure the quality of phased results, as contigs with lower coverage risk being phased inaccurately.

Phylogenomic Analyses

At this stage, data are nearly ready for use in phylogenetic reconstruction. Before beginning, however, it is advisable to perform inspection of sequence alignments for each gene, whether using programs such as GUIDANCE2 (Sela et al. 2015) or custom scripts , rather than labor-intensive inspection by eye. Preparation of raw data for tree building has become highly automated in response to the large volumes of data generated by high-throughput methods. Standardized sequence inspection helps reduce errors and inconsistencies, but can also be responsible for introducing errors in UCE datasets.

Data Filtering

Data filtering is a vital step in quality control of phylogenomic studies, as sequencing thousands of genes across many samples can lead to missing data in certain taxa. We advise

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

using different filtering criteria to generate multiple datasets and thereby find a balance between maintaining sequence quantity and quality. For example, GC bias has been demonstrated to be negatively correlated with topological support in bees (Bossert et al. 2017), and incongruences among analyses have been found to be exacerbated in studies of ants that used only "high signal" loci with highest average bootstrap (Borowiec 2019). While there is no current consensus on the best approach to filtering UCE data, multiple strategies exist. The program BaCoCa (Kuck and Struck 2014) can be used to filter out genes based on statistical properties such as saturation of nucleotides, compositional bias and heterogeneity, and proportion of shared missing data. The program Phylo-MCOA (de Vienne et al. 2012) can be used to detect outlier genes or species that cause topological incongruences; these can be subsequently filtered out for phylogenetic reconstruction. Another approach to data filtering is to retain only protein-coding genes rather than every UCE locus, some of which may include non-coding regions (see Current and Future Challenges below for more details). Also promising is analysis of protein-coding genes, which evolve under purifying selection and can be analyzed separately as amino acids; a custom script is now available for extracting putative protein-coding genes from UCE data (Borowiec 2019).

Data Partitioning

Approaches to partitioning UCE data can be divided into three strategies: 1) assign all UCE loci to a single partition; this assumes that every site in the alignment has evolved under a common evolutionary process; 2) assign each UCE locus to a separate partition; this allows for variation in rates and patterns of evolution between UCEs but assumes that all sites within each UCE locus have evolved under the same Markov process; or 3) *k*-means clustering of sites based on evolutionary rates (Frandsen et al. 2015), which subdivides data into partitions based on

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

evolutionary rates, thus avoiding *a priori* partitioning by the user. All three are used, however, recent studies have shown the *k*-means algorithm could be unreliable for UCE data, as it generates a partition comprised of all the invariant sites in the dataset, possibly misleading phylogenetic inference methods (Baca et al. 2017a). A promising new method for partitioning UCE data is the Sliding-Window Site Characteristics (SWSC, Tagliacollo and Lanfear 2018), which divides each UCE locus into three data blocks (right flank, core, and left flank) as the UCE core regions are conserved, while the two flanking regions become increasingly more variable (Faircloth et al. 2012). Different methods can be used by SWSC to evaluate sites, but the site entropies (EN), in particular, have been shown to most accurately account for within-UCE heterogeneity (Tagliacollo and Lanfear 2018). Using the SWSC-EN partitioning schemes account for within-UCE heterogeneity and leads to an increase in model fit (Tagliacollo and Lanfear 2018, Branstetter and Longino 2019).

Tree Building

Once datasets have been generated, downstream analyses on UCE data are similar to phylogenetic analyses performed on most other data types (e.g. Sanger sequencing, SNPs, etc.). A variety of tree-building methods (Figure 3) can be used for reconstructing phylogeny from UCE datasets, including Maximum Likelihood (ML), Bayesian Inference (BI), or Multispecies Coalescent/Species Tree (MSC). While the intricacies of phylogenetic analyses are beyond the scope of this paper, excellent and detailed overviews – both theoretical and practical – are available (Yang and Rannala 2012, Liu et al. 2015, Bromham et al. 2018, Bravo et al. 2019).

Maximum Likelihood

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

Maximum likelihood (ML) is a statistical methodology for estimating unknown parameters in a model. ML is widely used in phylogenetic studies due to its use of complex substitution models and its robustness to many violations to the assumptions of these models (Yang and Rannala 2012). The most widely used programs for phylogenetic reconstruction in the ML framework includes RAxML (Stamatakis 2006, Kozlov et al. 2018) and IQ-TREE (Nguyen et al. 2014). One advantage of these programs is their speed, with the former being the dominant method within UCE literature despite having very limited evolutionary model choices. IQ-TREE has gained momentum in recent years for its ability to produce accurate trees without sacrificing speed (Zhou et al. 2018). It includes functions such as ModelFinder for finding appropriate evolutionary models (Kalyaanamoorthy et al. 2017); approximation-based methods such as ultrafast bootstrap (UFBoot) and Shimodaira-Hasegawa like approximate likelihood ratio test (SH-aLRT), which greatly decreases computational time compared to traditional nonparametric bootstrap methods (Guindon et al. 2010, Hoang et al. 2017); and gene/site concordance factors as alternative support measures to illustrate disagreement among loci and sites (Minh et al. 2018).

Bayesian Inference and Divergence Dating

Like ML, Bayesian inference (BI) is also a general methodology of statistical inference that has been widely adopted for phylogenetic analyses. Bayesian inference differs from ML in that parameters in the models are considered to be random variables within statistical distributions rather than unknown fixed constants (Yang and Rannala 2012). Today BI using the Markov chain Monte Carlo (MCMC) sampling is a widely adopted method used for phylogenetic analysis, as the incorporation of prior knowledge into the analysis offers an appealing alternative to ML even at the cost of slower computational speed (Nascimento et al. 2017). The commonly used Bayesian programs for phylogenomics data include BEAST

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

(Drummond and Rambaut 2007), BEAST 2 (Bouckaert et al. 2014), and ExaBayes (Aberer et al. 2014). Bayesian analyses are extremely sensitive to prior probabilities set by users, and often default priors may not be appropriate for the data being analyzed as they can affect resulting topologies (Nascimento et al. 2017). Because setting priors can be daunting for beginners, we advise users to resist the all-too-common tendency to employ default settings and instead urge users to follow steps outlined in Bromham et al. (2018) to make informed choices when setting up Bayesian analyses. Running an 'empty' analysis without data to allow MCMC algorithm sampling from the prior is a good way of checking whether the data were informative enough to return posterior distributions different from the marginal priors, and to assess for good convergence and mixing of the MCMC chains (Nascimento et al. 2017, Blaimer et al. 2018b).

Divergence time estimation analyses can also be implemented for UCE phylogenies to generate dated chronograms, using carefully selected fossils as calibration points. The commonly used node-dating approach assigns the oldest fossil that can be confidently identified to the youngest internal node, imposing the age of the fossil a minimum age constraint (Arcila et al. 2015). An alternative method called total-evidence, or tip-dating methods can include all available paleontological information, ameliorating fossil-placement uncertainty while simultaneously incorporating fossil ages into the analysis (Ronquist et al. 2012a). Both of these methods can be implemented for divergence date estimation using Bayesian inference programs such as MrBayes (Ronquist et al. 2012b), BEAST/BEAST2, and the MCMCTree package in PAML (Yang 2007). While MCMCTree is faster computationally, the setup for prior distributions on fossil calibrations is less intuitive. BEAST/BEAST2, by comparison, is easier to understand and offers more analytical options such as the incorporation of fossils directly into the phylogeny with the newly developed node-dating method using the fossilized birth-death

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

model (Heath et al. 2014). The fossilized birth-death model offers an advantage over other methods by combining morphological and molecular data as well as stratigraphic range data from the fossil record, and can be implemented directly in RevBayes (Höhna et al. 2016), or in BEAST2 with add on package sampled-ancestor (Gavryushkina et al. 2014). In general, large data volumes associated with UCEs makes most Bayesian analyses too computationally intensive to be practical. To overcome this limitation, many studies reduce data size by removing taxa or loci in order to reduce the analysis time (Blaimer et al. 2018b, Borowiec 2019). It is also worth noting that tip-dating models have been shown to recover older ages than traditional node-dating models, and might produce inaccurate date estimations (Arcila et al. 2015). Regardless of the approach, the resulting dated chronogram can be then used as input for additional analyses such as ancestral state reconstruction, historical biogeographic analysis, or diversification rates estimation.

Multispecies Coalescent/Species Tree

One key advance in molecular phylogenetics has been the acknowledgement that high levels of incomplete lineage sorting (ILS) or other stochastic errors can yield misleading results for traditional methods concatenation methods (Liu et al. 2015, Bravo et al. 2019). Incorporation of discordance between gene trees and species trees as a result of high incomplete lineage sorting (ILS), under the MSC model (Heled and Drummond 2009) can alleviate this problem. Commonly used MSC tree summary-based methods such as ASTRAL (Mirarab et al. 2014, Mirarab and Warnow 2015, Zhang et al. 2018) and MP-EST (Liu et al. 2010) are performed in two steps, wherein gene trees are estimated first and separately, then used as input to generate a species tree based on various summaries of coalescent process (Bravo et al. 2019). Because the accuracy of the individual input gene trees directly affects the resulting species tree, these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

summary-based methods are especially suspectable to gene tree estimation errors (Molloy and Warnow 2018). Therefore, checking individual gene trees for incongruences is advised to ensure species tree accuracy in summary-based methods. Methods such as concordance factors (Ané et al. 2006, Minh et al. 2018) should be used to provide insight into the influence of ILS versus other factors such as introgression on the resulting topology. Alternatively, site-based coalescent method such as SVDquartets (Chifman and Kubatko 2014) and SVDquest (Vachaspati and Warnow 2018) bypass gene tree estimation, and is comparable or even more accurate than summary methods in cases of high ILS (Chou et al. 2015, Molloy and Warnow 2018). Finally, the newly developed StarBEAST2 (Ogilvie et al. 2017) package for BEAST2 is a promising implementation of the full MSC model which can jointly infer gene trees and species trees, but the current version is too computationally intensive to use on large UCE datasets.

Resources and Costs

Most steps of the wet lab protocol can be performed in standard molecular labs that have access to equipment such as a centrifuge and thermocycler. More specialized equipment such as a sonicator for shearing DNA, TapeStation/Bioanalyzer for quantifying DNA, and BluePippin/PippinHT for size selection can all be substituted with cheaper, albeit less accurate alternatives such as restriction enzymes, gel electrophoresis, and magnetic beads.

Illumina platforms (HiSeq, NextSeq, NovaSeq) are generally used for UCE studies due to their high throughput and low cost per base pair. The current estimated cost per specimen is approximately $30 – 40 USD, accounting for costs of all reagents in library preparation and paired-end Illumina run (See Supp Table 1 for sample cost breakdown). Some commercial

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

laboratories (e.g. RAPiD Genomics, Gainesville, FL, USA) also offer UCE enrichment services, handling all library preparation, enrichment, and sequencing; customers simply submit DNA extracts and then receive sequence data. Costs associated with such 'concierge service' are considerably higher (approximately ~$120 per specimen), but this may be an attractive option for researchers lacking the infrastructure or personnel to undertake wet lab protocols.

Having access to high performance computing (HPC) greatly expedites bioinformatic and phylogenomic analyses, especially when processing large batches of samples. While PHYLUCE and many associated data-processing programs can be run in local Linux/Unix environment, the parallelization using HPC will reduce execution time in computationally intensive steps such as demultiplexing and assembly. Similarly, many phylogenomics programs discussed above can also be expedited through this process.

Data Availability Recommendations

One hallmark feature of UCE data is its open source nature, probe sets, protocols, and previously published data are made publicly available, ensuring repeatability – the foundation of open scientific research. To this end, untrimmed raw Illumina reads should be uploaded to public database such as Sequence Read Archive (SRA) once studies are published, giving interested readers the full ability to download and process the data using different trimming settings. All analytical methods such as software and code used to process data should also be made publicly available on repositories such as Dryad or GitHub. UCE contigs can be uploaded to GenBank as targeted locus studies, making the data available for BLAST. The voucher specimens from which DNA was extracted should be deposited in recognized scientific collections and museums;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

associated information such as collection locality, identification, etc., should be included as metadata with all molecular sequence (Bravo et al. 2019).

Current and Future Challenges

UCE and similar methods offer the ability to generate massive amounts of data from many loci, and yet, despite the increase in data volume, the same concerns that have long plagued phylogenetic analyses remain as relevant as ever: taxon sampling, choice of alignment methods, and composition bias (Bossert et al. 2017, Mclean et al. 2018). Recent research also suggests phylogenomic results can be strongly affected by a tiny proportion of highly biased loci or sites (Shen et al. 2017), and reduction of phylogenetic noise resulting from compositional heterogeneity and saturation can increase congruence among different analytic methods (Borowiec 2019). With that in mind, we strongly encourage performing sensitivity analyses to test the robustness of results when interpreting phylogenomic data (Borowiec 2019, Camacho et al. 2019). As these large datasets are less prone to uncertainty, and instead may give strongly supported wrong results if model violations are not carefully evaluated (Borowiec et al. 2019).

The fact that the function of UCEs remains largely unknown is the basis of active research and a current challenge for identifying and modeling UCEs in a phylogenomic context (Bejerano et al. 2004). Vertebrate UCEs are characterized as predominantly non-coding sequences, non-randomly distributed across chromosomes and acting as regulators and/or enhancers of gene expression (Baira et al. 2008, Polychronopoulos et al. 2017). By contrast, studies of invertebrate UCEs reveal that most flanking regions captured include exons (Branstetter et al. 2017a), with the most widely shared loci being either exclusively conserved

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

exons or partially exonic regions in Hymenoptera and Arachnida (Bossert and Danforth 2018, Hedin et al. 2019). This is an exciting discovery, as the exonic flanking regions captured by the UCE process and transcriptome sequence data within these groups can be meaningfully combined, without the need to design specific probe sets to target them, as demonstrated in Apidae (Bossert et al. 2019). However, since the genomic landscapes of different animal taxa can differ substantially, the wider application of combining transcriptomic data with UCEs in other taxonomic groups still needs to be tested. Currently, it appears that the function of UCEs is highly variable, with flanking regions containing exons and introns; whether this variability will affect downstream analyses remains to be seen.

Continued refinement of existing probe sets is needed to increase capture success while minimizing duplicates and paralogous loci. It has been shown in the arachnid probe set, different UCE probes sometimes target regions of the same protein, or include non-orthologous sequences (Hedin et al. 2019). This is unsurprising given the wide phylogenetic depth of the probe set, which was designed to target all arachnids, but given that PHYLUCE cannot detect these non-orthologous sequences as the program only removes different contigs hit by probes targeting the same loci (Faircloth 2015), additional manual filtering is needed to ensure the exclusion of misleading paralogous sequences into the final data matrix (Hedin et al. 2019).

Conclusion

Ultraconserved elements-based phylogenomic studies have been rapidly adopted by researchers working on arthropod taxa since their introduction by Faircloth et al. (2012). This review described the versatility of UCE data at both deep and shallow evolutionary scale, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

provided a step-by-step guide to generating and analyzing UCEs; we then summarized current practices, challenges, and unresolved questions that surround this active field. Our hope is to make UCE-based phylogenomic studies more accessible to users with diverse taxonomic interests, and thereby deepen our collective understanding of the roles and functions of UCEs across widely divergent taxa. As our understanding of UCEs develops through studies of different organisms, identifying individual genes and incorporation of functional genomics will yield interesting comparative studies across deeply divergent taxonomic groups and provide new insights in the continued pursuit of building the tree of life.

Acknowledgements

We would like to thank Michael Branstetter and two anonymous reviewers for comments that greatly improved the manuscript, Eliana Buenaventura, Chris Cohen, Shahan Derkarabetian, Grey Gustafson, Katherine Noble, and Matthew van Dam for informative discussions on UCE research in their perspective groups. We thank Rachel Atchison and Suzy Rodriguez for graphic design assistance.

References

Aberer, A. J., K. Kobert, and A. Stamatakis. 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Molecular Biology and Evolution 31: 2553-2556.

Alda, F., V. A. Tagliacollo, M. J. Bernt, B. T. Waltz, W. B. Ludt, B. C. Faircloth, M. E. Alfaro, J. S. Albert, and P. Chakrabarty. 2018. Resolving deep nodes in an ancient radiation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

561 of neotropical fishes in the presence of conflicting signals from incomplete lineage sorting.
562 Systematic Biology 68: 573-593.
563 **Ali, O. A., S. M. O'Rourke, S. J. Amish, M. H. Meek, G. Luikart, C. Jeffres, and M. R.**
564 **Miller. 2016.** RAD Capture (Rapture): Flexible and efficient sequence-based genotyping.
565 Genetics 202: 389-400.
566 **Allio, R., A. Schomaker-Bastos, J. Romiguier, F. Prosdocimi, B. Nabholz, and F. Delsuc.**
567 **2019.** MitoFinder: efficient automated large-scale extraction of mitogenomic data in target
568 enrichment phylogenomics. bioRxiv 685412.
569 **Andermann, T., A. Cano, A. Zizka, C. Bacon, and A. Antonelli. 2018.** SECAPR-a
570 bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina
571 sequences, from raw reads to alignments. PeerJ 6: e5175.
572 **Andermann, T., A. M. Fernandes, U. Olsson, M. Topel, B. Pfeil, B. Oxelman, A. Aleixo, B.**
573 **C. Faircloth, and A. Antonelli. 2019.** Allele phasing greatly improves the phylogenetic utility
574 of ultraconserved elements. Systematic Biology 68: 32-46.
575 **Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2006.** Bayesian Estimation of
576 Concordance among Gene Trees. Molecular Biology and Evolution 24: 412-426.
577 **Arcila, D., R. Alexander Pyron, J. C. Tyler, G. Orti, and R. R. Betancur. 2015.** An
578 evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (Teleostei:
579 Percomorphaceae). Molecular Phylogenetics and Evolution 82 Pt A: 131-145.
580 **Baca, S. M., E. F. A. Toussaint, K. B. Miller, and A. E. Z. Short. 2017a.** Molecular
581 phylogeny of the aquatic beetle family Noteridae (Coleoptera: Adephaga) with an emphasis on
582 data partitioning strategies. Molecular Phylogenetics and Evolution 107: 282-292.
583 **Baca, S. M., A. Alexander, G. T. Gustafson, and A. E. Z. Short. 2017b.** Ultraconserved
584 elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly
585 of 'Hydradephaga'. Systematic Entomology 42: 786-795.
586 **Baira, E., J. Greshock, G. Coukos, and L. Zhang. 2008.** Ultraconserved elements: genomics,
587 function and disease. RNA Biology 5: 132-134.
588 **Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U.**
589 **Selker, W. A. Cresko, and E. A. Johnson. 2008.** Rapid SNP discovery and genetic mapping
590 using sequenced RAD markers. PLoS One 3: e3376.
591 **Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M.**
592 **Lesin, S. I. Nikolenko, S. Pham, and A. D. Prjibelski. 2012.** SPAdes: a new genome assembly
593 algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:
594 455-477.
595 **Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D.**
596 **Haussler. 2004.** Ultraconserved elements in the human genome. Science 304: 1321-1325.
597 **Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012.**
598 Transcriptome-based exon capture enables highly cost-effective comparative genomic data
599 collection at moderate evolutionary scales. BMC Genomics 13: 403.
600 **Blaimer, B. B., J. R. Mawdsley, and S. G. Brady. 2018a.** Multiple origins of sexual
601 dichromatism and aposematism within large carpenter bees. Evolution 72: 1874-1889.
602 **Blaimer, B. B., M. W. Lloyd, W. X. Guillory, and S. G. Brady. 2016a.** Sequence capture and
603 phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens.
604 PLoS One 11: e0161531.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

Blaimer, B. B., J. S. LaPolla, M. G. Branstetter, M. W. Lloyd, and S. G. Brady. 2016b. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus *Acropyga*. Molecular Phylogenetics and Evolution 102: 20-29.

Blaimer, B. B., P. S. Ward, T. R. Schultz, B. L. Fisher, and S. G. Brady. 2018b. Paleotropical diversification dominates the evolution of the hyperdiverse ant tribe Crematogastrini (Hymenoptera: Formicidae). Insect Systematics and Diversity 2: 3.

Blaimer, B. B., S. G. Brady, T. R. Schultz, M. W. Lloyd, B. L. Fisher, and P. S. Ward. 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. BMC Evolutionary Biology 15: 271.

Bodily, P. M., M. S. Fujimoto, C. Ortega, N. Okuda, J. C. Price, M. J. Clement, and Q. Snell. 2015. Heterozygous genome assembly via binary classification of homologous sequence. BMC Bioinformatics 16: S5.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Borowiec, M. L. 2019. Convergent evolution of the army ant syndrome and congruence in big-data phylogenetics. Systematic Biology 68: 642–656.

Borowiec, M. L., C. Rabeling, S. G. Brady, B. L. Fisher, T. R. Schultz, and P. S. Ward. 2019. Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. Molecular Phylogenetics and Evolution 134: 111-121.

Bossert, S., and B. N. Danforth. 2018. On the universality of target-enrichment baits for phylogenomic research. Methods in Ecology and Evolution 9: 1453-1460.

Bossert, S., E. A. Murray, B. B. Blaimer, and B. N. Danforth. 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Molecular Phylogenetics and Evolution 111: 149-157.

Bossert, S., E. A. Murray, E. A. B. Almeida, S. G. Brady, B. B. Blaimer, and B. N. Danforth. 2019. Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. Molecular Phylogenetics and Evolution 130: 121-131.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10: e1003537.

Branstetter, M. G., and J. T. Longino. 2019. Ultra-conserved element phylogenomics of New World *Ponera* (Hymenoptera: Formicidae) illuminates the origin and phylogeographic history of the Endemic Exotic Ant *Ponera exotica*. Insect Systematics and Diversity 3: 1.

Branstetter, M. G., J. T. Longino, P. S. Ward, B. C. Faircloth, and S. Price. 2017a. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. Methods in Ecology and Evolution 8: 768-776.

Branstetter, M. G., A. Jesovnik, J. Sosa-Calvo, M. W. Lloyd, B. C. Faircloth, S. G. Brady, and T. R. Schultz. 2017b. Dry habitats were crucibles of domestication in the evolution of agriculture in ants. Proceedings of the Royal Society of London B: Biological Sciences 284.

Branstetter, M. G., B. N. Danforth, J. P. Pitts, B. C. Faircloth, P. S. Ward, M. L. Buffington, M. W. Gates, R. R. Kula, and S. G. Brady. 2017c. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. Current Biology 27: 1019-1025.

Bravo, G. A., A. Antonelli, C. D. Bacon, K. Bartoszek, M. P. K. Blom, S. Huynh, G. Jones, L. L. Knowles, S. Lamichhaney, T. Marcussen, H. Morlon, L. K. Nakhleh, B. Oxelman, B. Pfeil, A. Schliep, N. Wahlberg, F. P. Werneck, J. Wiedenhoeft, S. Willows-Munro, and S.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

650 **V. Edwards. 2019.** Embracing heterogeneity: coalescing the Tree of Life and the future of
651 phylogenomics. PeerJ 7: e6399.
652 **Bromham, L., S. Duchene, X. Hua, A. M. Ritchie, D. A. Duchene, and S. Y. W. Ho. 2018.**
653 Bayesian molecular dating: opening up the black box. Biological Reviews 93: 1165-1191.
654 **Camacho, G. P., M. R. Pie, R. M. Feitosa, and M. S. Barbeitos. 2019.** Exploring gene tree
655 incongruence at the origin of ants and bees (Hymenoptera). Zoologica Scripta 48: 215-225.
656 **Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009.** trimAl: a tool for
657 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-
658 1973.
659 **Castresana, J. 2000.** Selection of conserved blocks from multiple alignments for their use in
660 phylogenetic analysis. Molecular Biology and Evolution 17: 540-552.
661 **Chifman, J., and L. Kubatko. 2014.** Quartet inference from SNP data under the coalescent
662 model. Bioinformatics 30: 3317-3324.
663 **Chou, J., A. Gupta, S. Yaduvanshi, R. Davidson, M. Nute, S. Mirarab, and T. Warnow.**
664 **2015.** A comparative study of SVDquartets and other coalescent-based species tree estimation
665 methods. BMC Genomics 16: S2.
666 **Cooke, C. 2018.** Forest Micro-Hymenoptera, including those attacking trees (Cynipidae: Oak
667 Gall Wasps) and those potentially defending them (Parasitic Pteromalidae). PhD, University of
668 Maryland, College Park, MD, USA.
669 **Crawford, N. G., B. C. Faircloth, J. E. McCormack, R. T. Brumfield, K. Winker, and T. C.**
670 **Glenn. 2012.** More than 1000 ultraconserved elements provide evidence that turtles are the sister
671 group of archosaurs. Biology Letters rsbl20120331.
672 **Cruaud, A., G. Groussier, G. Genson, L. Saune, A. Polaszek, and J. Y. Rasplus. 2018.**
673 Pushing the limits of whole genome amplification: successful sequencing of RADseq library
674 from a single microhymenopteran (Chalcidoidea, *Trichogramma*). PeerJ 6: e5640.
675 **Cruaud, A., S. Nidelet, P. Arnal, A. Weber, L. Fusu, A. Gumovsky, J. Huber, A. Polaszek,**
676 **and J. Y. Rasplus. 2019.** Optimized DNA extraction and library preparation for minute
677 arthropods: Application to target enrichment in chalcid wasps used for biocontrol. Molecular
678 Ecology Resources.
679 **de Vienne, D. M., S. Ollier, and G. Aguileta. 2012.** Phylo-MCOA: a fast and efficient method
680 to detect outlier genes and species in phylogenomics using multiple co-inertia analysis.
681 Molecular Biology and Evolution 29: 1587-1598.
682 **Derkarabetian, S., S. Castillo, P. K. Koo, S. Ovchinnikov, and M. Hedin. 2019.** A
683 demonstration of unsupervised machine learning in species delimitation. Molecular
684 Phylogenetics and Evolution 139.
685 **Derkarabetian, S., J. Starrett, N. Tsurusaki, D. Ubick, S. Castillo, and M. Hedin. 2018.** A
686 stable phylogenomic classification of Travunioidea (Arachnida, Opiliones, Laniatores) based on
687 sequence capture of ultraconserved elements. ZooKeys 1-36.
688 **Drummond, A. J., and A. Rambaut. 2007.** BEAST: Bayesian evolutionary analysis by
689 sampling trees. BMC Evolutionary Biology 7: 214.
690 **Edgar, R. C. 2004.** MUSCLE: multiple sequence alignment with high accuracy and high
691 throughput. Nucleic Acids Research 32: 1792-1797.
692 **Faircloth, B. 2013.** Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality
693 trimming.
694 **Faircloth, B. C. 2015.** PHYLUCE is a software package for the analysis of conserved genomic
695 loci. Bioinformatics 32: 786-788.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

Faircloth, B. C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods in Ecology and Evolution 8: 1103-1112.

Faircloth, B. C., L. Sorenson, F. Santini, and M. E. Alfaro. 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., M. G. Branstetter, N. D. White, and S. G. Brady. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Molecular Ecology Resources 15: 489-501.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology 61: 717-726.

Forthman, M., C. W. Miller, and R. T. Kimball. 2019. Phylogenomic analysis suggests Coreidae and Alydidae (Hemiptera: Heteroptera) are not monophyletic. Zoologica Scripta 48: 520-534.

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

Gavryushkina, A., D. Welch, T. Stadler, and A. J. Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Computational Biology 10: e1003919.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, and C. Russ. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182.

Grab, H., M. G. Branstetter, N. Amon, K. R. Urban-Mead, M. G. Park, J. Gibbs, E. J. Blitzer, K. Poveda, G. Loeb, and B. N. Danforth. 2019. Agriculturally dominated landscapes reduce bee phylogenetic diversity and pollination services. Science 363: 282-284.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, and Q. Zeng. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59: 307-321.

Gustafson, G. T., A. Alexander, J. S. Sproul, J. M. Pflug, D. R. Maddison, and A. E. Z. Short. 2019. Ultraconserved element (UCE) probe set design: Base genome and initial design parameters critical for optimization. Ecology and Evolution 9: 6933-6948.

Harvey, M. G., B. T. Smith, T. C. Glenn, B. C. Faircloth, and R. T. Brumfield. 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Systematic Biology 65: 910-924.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences of the United States of America 111: E2957-2966.

Hedin, M., S. Derkarabetian, J. Blair, and P. Paquin. 2018a. Sequence capture phylogenomics of eyeless *Cicurina* spiders from Texas caves, with emphasis on US federally-endangered species from Bexar County (Araneae, Hahniidae). ZooKeys 49-76.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

Hedin, M., S. Derkarabetian, M. J. Ramírez, C. Vink, and J. E. Bond. 2018b. Phylogenomic reclassification of the world's most venomous spiders (Mygalomorphae, Atracinae), with implications for venom evolution. Scientific Reports 8: 1636.

Hedin, M., S. Derkarabetian, A. Alfaro, M. J. Ramírez, and J. E. Bond. 2019. Phylogenomic analysis and revised classification of atypoid mygalomorph spiders (Araneae, Mygalomorphae), with notes on arachnid ultraconserved element loci. PeerJ 7: e6864.

Heled, J., and A. J. Drummond. 2009. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27: 570-580.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and S. V. Le. 2017. UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution msx281.

Hoffberg, S. L., T. J. Kieran, J. M. Catchen, A. Devault, B. C. Faircloth, R. Mauricio, and T. C. Glenn. 2016. RAD cap: sequence capture of dual-digest RAD seq libraries with identifiable duplicates and reduced missing data. Molecular Ecology Resources 16: 1264-1278.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65: 726-736.

Ješovnik, A. N. A., J. Sosa-Calvo, M. W. Lloyd, M. G. Branstetter, F. FernÁNdez, and T. R. Schultz. 2017. Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming ant genus Sericomyrmex Mayr (Hymenoptera: Formicidae): ultraconserved elements (UCEs) resolve a recent radiation. Systematic Entomology 42: 523-542.

Jones, M. R., and J. M. Good. 2016. Targeted capture in evolutionary and ecological genomics. Molecular Ecology 25: 185-202.

Kalyaanamoorthy, S., B. Q. Minh, T. K. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14: 587.

Katoh, K., K. Misawa, K. i. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059-3066.

Kieran, T. J., E. R. Gordon, M. Forthman, R. Hoey-Chamberlain, R. T. Kimball, B. C. Faircloth, C. Weirauch, and T. C. Glenn. 2019. Insight from an ultraconserved element bait set designed for hemipteran phylogenetics integrated with genomic resources. Molecular Phylogenetics and Evolution 130: 297-303.

Kozlov, A., D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. 2018. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. bioRxiv 447110.

Kuck, P., and T. H. Struck. 2014. BaCoCa--a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Molecular Phylogenetics and Evolution 70: 94-98.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Systematic Biology 61: 727-744.

Lemmon, E. M., and A. R. Lemmon. 2013. High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics 44: 99-121.

Lim, H. C., and M. J. Braun. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

Liu, L., L. Yu, and S. V. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 10: 302.

Liu, L., Z. Xi, S. Wu, C. C. Davis, and S. V. Edwards. 2015. Estimating phylogenetic trees from genome-scale data. Annals of the New York Academy of Sciences 1360: 36-53.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner. 2010. Target-enrichment strategies for next-generation sequencing. Nature Methods 7: 111.

Manthey, J. D., L. C. Campillo, K. J. Burns, and R. G. Moyle. 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., M. Sann, A. Donath, M. Meixner, L. Podsiadlowski, R. S. Peters, M. Petersen, K. Meusemann, K. Liere, and J.-W. Wägele. 2016. BaitFisher: a software package for multispecies target DNA enrichment probe design. Molecular Biology and Evolution 33: 1875-1886.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013a. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66: 526-538.

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

Mclean, B. S., K. C. Bell, J. M. Allen, K. M. Helgen, and J. A. Cook. 2018. Impacts of inference method and data set filtering on phylogenomic resolution in a rapid radiation of ground squirrels (Xerinae: Marmotini). Systematic biology 68: 298-316.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17: 240-248.

Minh, B. Q., M. Hahn, and R. Lanfear. 2018. New methods to calculate concordance factors for phylogenomic datasets. bioRxiv 487801.

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31: i44-i52.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541-i548.

Molloy, E. K., and T. Warnow. 2018. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. Systematic Biology 67: 285-303.

Musher, L. J., and J. Cracraft. 2018. Phylogenomics and species delimitation of a complex radiation of Neotropical suboscine birds (Pachyramphus). Molecular Phylogenetics and Evolution 118: 204-221.

Myers, E. A., R. W. Bryson Jr, R. W. Hansen, M. L. Aardema, D. Lazcano, and F. T. Burbrink. 2019. Exploring Chihuahuan Desert diversification in the gray-banded kingsnake, *Lampropeltis alterna* (Serpentes: Colubridae). Molecular Phylogenetics and Evolution 131: 211-218.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

Nascimento, F. F., M. dos Reis, and Z. Yang. 2017. A biologist's guide to Bayesian phylogenetic analysis. Nature Ecology & Evolution 1: 1446-1454.

Newman, C. E., and C. C. Austin. 2016. Sequence capture and next-generation sequencing of ultraconserved elements in a large-genome salamander. Molecular Ecology 25: 6162-6174.

Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32: 268-274.

Nurk, S., D. Meleshko, A. Korobeynikov, and P. A. Pevzner. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27: 824-834.

Ogilvie, H. A., R. R. Bouckaert, and A. J. Drummond. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular Biology and Evolution 34: 2101-2114.

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

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. Littlewood, M. Manuel, G. Worheide, and D. Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biology 9: e1000602.

Pierce, M. P., M. G. Branstetter, and J. T. Longino. 2017. Integrative taxonomy reveals multiple cryptic species within Central American *Hylomyrma* Forel, 1912 (Hymenoptera: Formicidae). Myrmecological News 25: 131-143.

Polychronopoulos, D., J. W. King, A. J. Nash, G. Tan, and B. Lenhard. 2017. Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Research 45: 12611-12624.

Prebus, M. 2017. Insights into the evolution, biogeography and natural history of the acorn ants, genus *Temnothorax* Mayr (hymenoptera: Formicidae). BMC Evolutionary Biology 17: 250.

Quattrini, A. M., B. C. Faircloth, L. F. Dueñas, T. C. Bridge, M. R. Brugler, I. F. Calixto-Botía, D. M. DeLeo, S. Foret, S. Herrera, and S. M. Lee. 2018. Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: New approaches to long-standing problems. Molecular Ecology Resources 18: 281-295.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology 61: 973-999.

Ronquist, F., M. Teslenko, P. Van Der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61: 539-542.

Ruane, S., and C. C. Austin. 2017. Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. Molecular Ecology Resources 17: 1003-1008.

Ryu, T., L. Seridi, and T. Ravasi. 2012. The evolution of ultraconserved elements with different phylogenetic origins. BMC Evolutionary Biology 12: 236.

Santos, B. F., A. Perrard, and S. G. Brady. 2019. Running in circles in phylomorphospace: host environment constrains morphological diversification in parasitic wasps. Proceedings of the Royal Society of London B: Biological Sciences 286.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

874 **Sela, I., H. Ashkenazy, K. Katoh, and T. Pupko. 2015.** GUIDANCE2: accurate detection of
875 unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic
876 Acids Research 43: W7-W14.
877 **Shen, X. X., C. T. Hittinger, and A. Rokas. 2017.** Contentious relationships in phylogenomic
878 studies can be driven by a handful of genes. Nature Ecology & Evolution 1: 126.
879 **Short, A. E. Z., T. Dikow, and C. S. Moreau. 2018.** Entomological Collections in the Age of
880 Big Data. Annual Review of Entomology 63: 513-530.
881 **Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. 2009.**
882 ABySS: a parallel assembler for short read sequence data. Genome Research 19: 1117-1123.
883 **Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2014.** Target
884 capture and massively parallel sequencing of ultraconserved elements for comparative studies at
885 shallow evolutionary time scales. Systematic Biology 63: 83-95.
886 **Stamatakis, A. 2006.** RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
887 thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.
888 **Starrett, J., S. Derkarabetian, M. Hedin, R. W. Bryson Jr, J. E. McCormack, and B. C.**
889 **Faircloth. 2017.** High phylogenetic utility of an ultraconserved element probe set designed for
890 Arachnida. Molecular Ecology Resources 17: 812-823.
891 **Streicher, J. W., and J. J. Wiens. 2017.** Phylogenomic analyses of more than 4000 nuclear loci
892 resolve the origin of snakes among lizard families. Biology Letters 13: 20170393.
893 **Ströher, P. R., A. L. S. Meyer, E. Zarza, W. L. E. Tsai, J. E. McCormack, and M. R. Pie.**
894 **2019.** Phylogeography of ants from the Brazilian Atlantic Forest. Organisms Diversity &
895 Evolution.
896 **Suchan, T., C. Pitteloud, N. S. Gerasimova, A. Kostikova, S. Schmid, N. Arrigo, M.**
897 **Pajkovic, M. Ronikier, and N. Alvarez. 2016.** Hybridization capture using RAD probes
898 (hyRAD), a new tool for performing genomic analyses on collection specimens. PloS One 11:
899 e0151651.
900 **Tagliacollo, V. A., and R. Lanfear. 2018.** Estimating improved partitioning schemes for
901 ultraconserved elements. Molecular Biology and Evolution 35: 1798-1811.
902 **Vachaspati, P., and T. Warnow. 2018.** SVDquest: Improving SVDquartets species tree
903 estimation using exact optimization within a constrained search space. Molecular Phylogenetics
904 and Evolution 124: 122-136.
905 **Van Dam, M. H., M. Trautwein, G. S. Spicer, and L. Esposito. 2019.** Advancing mite
906 phylogenomics: Designing ultraconserved elements for Acari phylogeny. Molecular Ecology
907 Resources 19: 465-475.
908 **Van Dam, M. H., A. W. Lam, K. Sagata, B. Gewa, R. Laufa, M. Balke, B. C. Faircloth, and**
909 **A. Riedel. 2017.** Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-
910 weevils. PLoS One 12: e0188044.
911 **Ward, P. S., and M. G. Branstetter. 2017.** The acacia ants revisited: convergent evolution and
912 biogeographic context in an iconic ant/plant mutualism. Proceedings of the Royal Society of
913 London B: Biological Sciences 284.
914 **Weitemier, K., S. C. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A.**
915 **Liston. 2014.** Hyb-Seq: Combining target enrichment and genome skimming for plant
916 phylogenomics. Applications in Plant Sciences 2: 1400042.
917 **Wood, H. M., V. L. Gonzalez, M. Lloyd, J. Coddington, and N. Scharff. 2018.** Next-
918 generation museum genomics: Phylogenetic relationships among palpimanoid spiders using

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

sequence capture techniques (Araneae: Palpimanoidea). Molecular Phylogenetics and Evolution 127: 907-918.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586-1591.

Yang, Z., and B. Rannala. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13: 303-314.

Yeates, D. K., A. Zwick, and A. S. Mikheyev. 2016. Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections. Current opinion in insect science 18: 83-88.

Zarza, E., E. M. Connors, J. M. Maley, W. L. E. Tsai, P. Heimes, M. Kaplan, and J. E. McCormack. 2018. Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (*Sarcohyla*; Hylidae). PeerJ 6.

Zerbino, D., and E. Birney. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research 18: 821-829.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19: 153.

Zhang, F., Y. Ding, C. Zhu, X. Zhou, M. C. Orr, S. Scheu, and Y.-X. Luan. 2019. Phylogenomics from Low-coverage Whole-genome Sequencing. Methods in Ecology and Evolution 10: 507-517.

Zhou, X., X. X. Shen, C. T. Hittinger, and A. Rokas. 2018. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. Molecular Biology and Evolution 35: 486-503.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

943

944 **Table 1**. Published studies using arthropod UCE datasets as of July 2019.

Taxonomic Group	Probe Sets	References
Hymenoptera	Hymenoptera 1.5Kv1 Hymenoptera 2.5Kv2 Full Hymenoptera 2.5Kv2 Ant-Specific Hymenoptera 2.5Kv2 Bee-Ant-Specific	(Blaimer et al. 2015, Faircloth et al. 2015, Blaimer et al. 2016b, Blaimer et al. 2016a, Branstetter et al. 2017c, Branstetter et al. 2017b, Branstetter et al. 2017a, Ješovnik et al. 2017, Pierce et al. 2017, Prebus 2017, Ward and Branstetter 2017, Blaimer et al. 2018a, Blaimer et al. 2018b, Cooke 2018, Borowiec 2019, Bossert et al. 2019, Branstetter and Longino 2019, Cruaud et al. 2019, Grab et al. 2019, Santos et al. 2019, Ströher et al. 2019)
Arachnida	Arachnida 1.1Kv1 Mite-v2	(Faircloth 2017, Starrett et al. 2017, Derkarabetian et al. 2018, Hedin et al. 2018a, Hedin et al. 2018b, Wood et al. 2018, Derkarabetian et al. 2019, Hedin et al. 2019, Van Dam et al. 2019)
Coleoptera	Coleoptera 1.1Kv1 Adephaga_2.9Kv1	(Baca et al. 2017b, Faircloth 2017, Van Dam et al. 2017, Gustafson et al. 2019)
Hemiptera	Hemiptera 2.7Kv1	(Faircloth 2017, Forthman et al. 2019, Kieran et al. 2019)
Diptera	Diptera 2.7Kv1	(Faircloth 2017)
Lepidoptera	(*in silico* only)	(Faircloth 2017)
Psocodea	Phthiraptera-2.8Kv1 (*in silico* only)	(Zhang et al. 2019)

945

946

947

948

949

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

950 **Supplemental Table 1.** Sample cost breakdown of DNA extraction, library preparation, UCE enrichment, and Illumina sequencing

951 estimated at around $40 USD per specimen for 96 samples. (as of July 2019).

Reagent / Kit	Unit of Total Kit	Total Kit Cost	Cost Per Prep	Number Preps	Total Cost
DNeasy Blood and Tissue Kit	250 rxns	$644.00	$2.58	96	$247.30
Library Prep Kits	96 reactions	$2,496.00	$6.50	96	$624.00
HotStart ReadyMix	500 x 25 µL reactions	$525.00	$1.05	108	$113.40
Magnetic SpeedBeads	15 mL	$385.50	$25.70	1	$25.70
			$35.83		**$1,010.40**
Dynabeads MyOne Streptavidin T1	10000 µL	$1,624.00	$8.12	12	$97.44
MYbaits-1	12 reactions - 5.5µL per reaction	$2,400.00	$32.73	12	$392.73
Enrichment Reagents	Misc. Chemicals / Reagents	NA	$15.00	12	$180.00
			$55.85		**$670.17**
qPCR Library Quant	500 x 20 µL reactions	$604.00	$0.60	90	$54.36
Gel Cassettes, BluePippin	Cassette for 5 samples	$450.00	$50.00	1	$50.00
High Sensitivity D1000	Tape for 16 samples	$513.51	$50.00	1	$50.00
			$114.50		**$154.36**
HiSeq 125 cycle paired-end	1 lane	$2,140.00	$2,140.00	1	**$2,140.00**
					$3,974.92

952

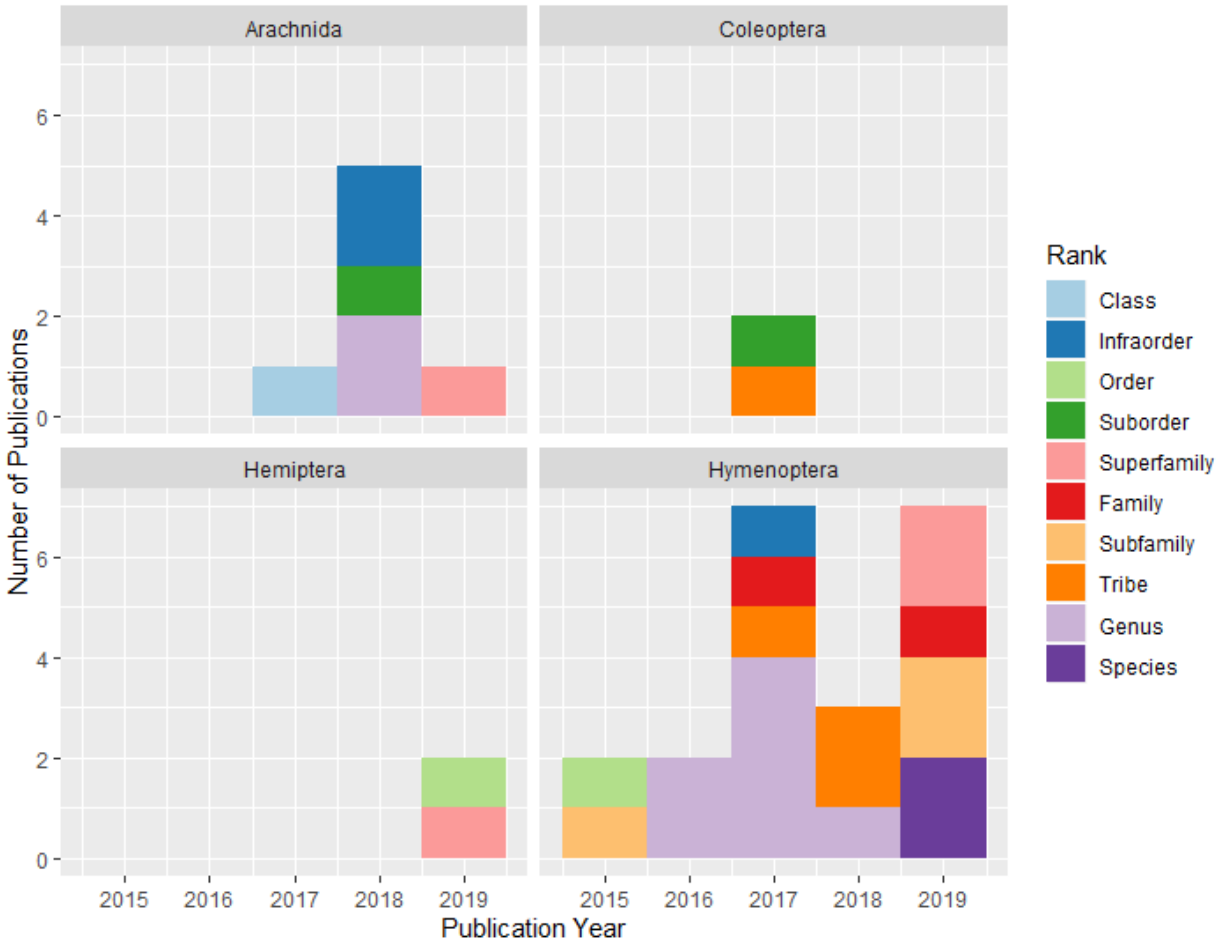
953 **Supplementary Table 2.** Breakdown of methods used by used in 32 UCE-based arthropod studies (as of July 2019).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

954 **Figure 1.** Breakdown of the number of arthropod UCEs-based publications per year (as of July

955 2019) by taxonomic group and taxonomic hierarchy.



956

957

958

959

960

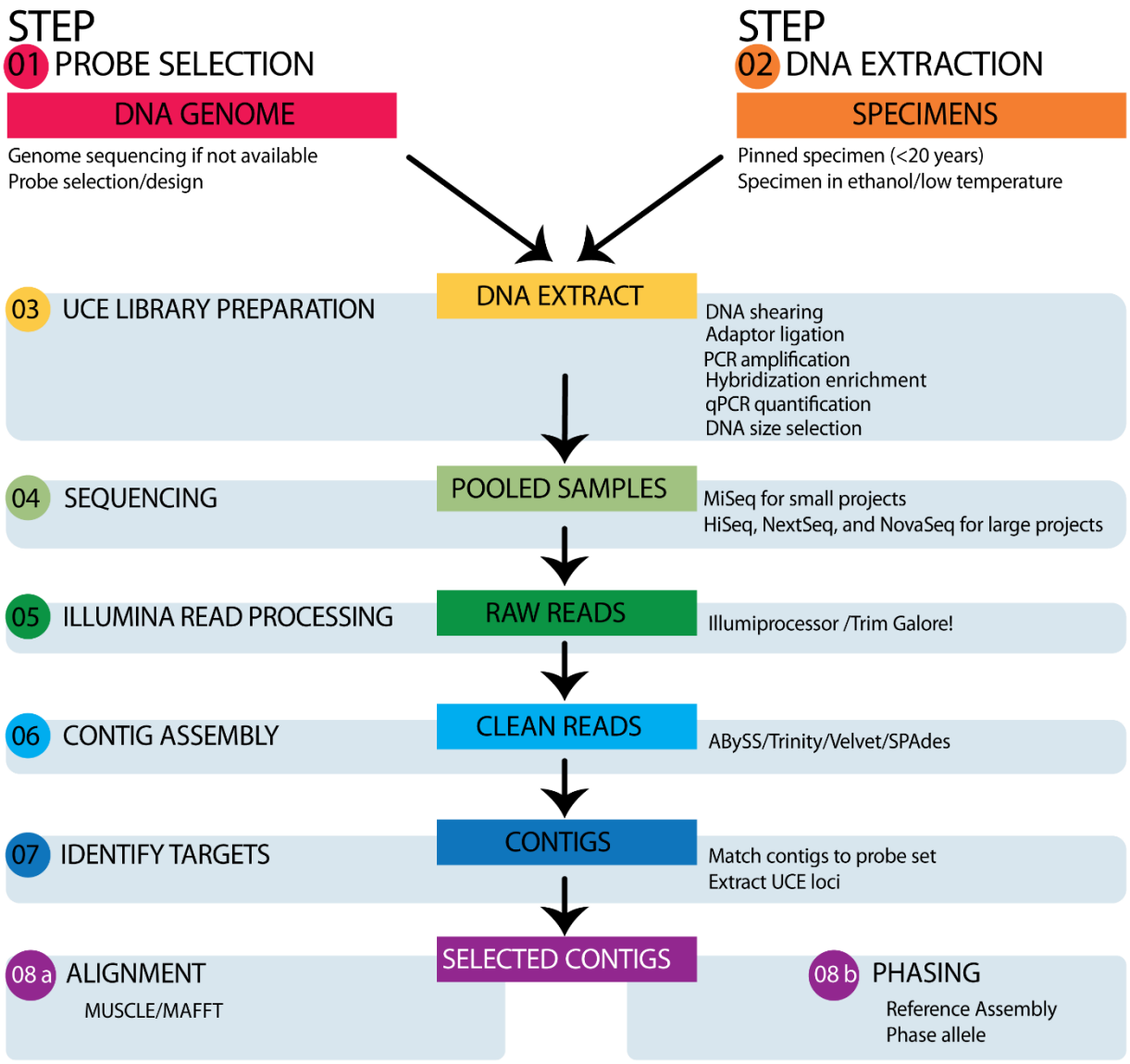
961

962

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

963 **Figure 2**. Generalized workflow of the UCE pipeline.



964

965

966

967

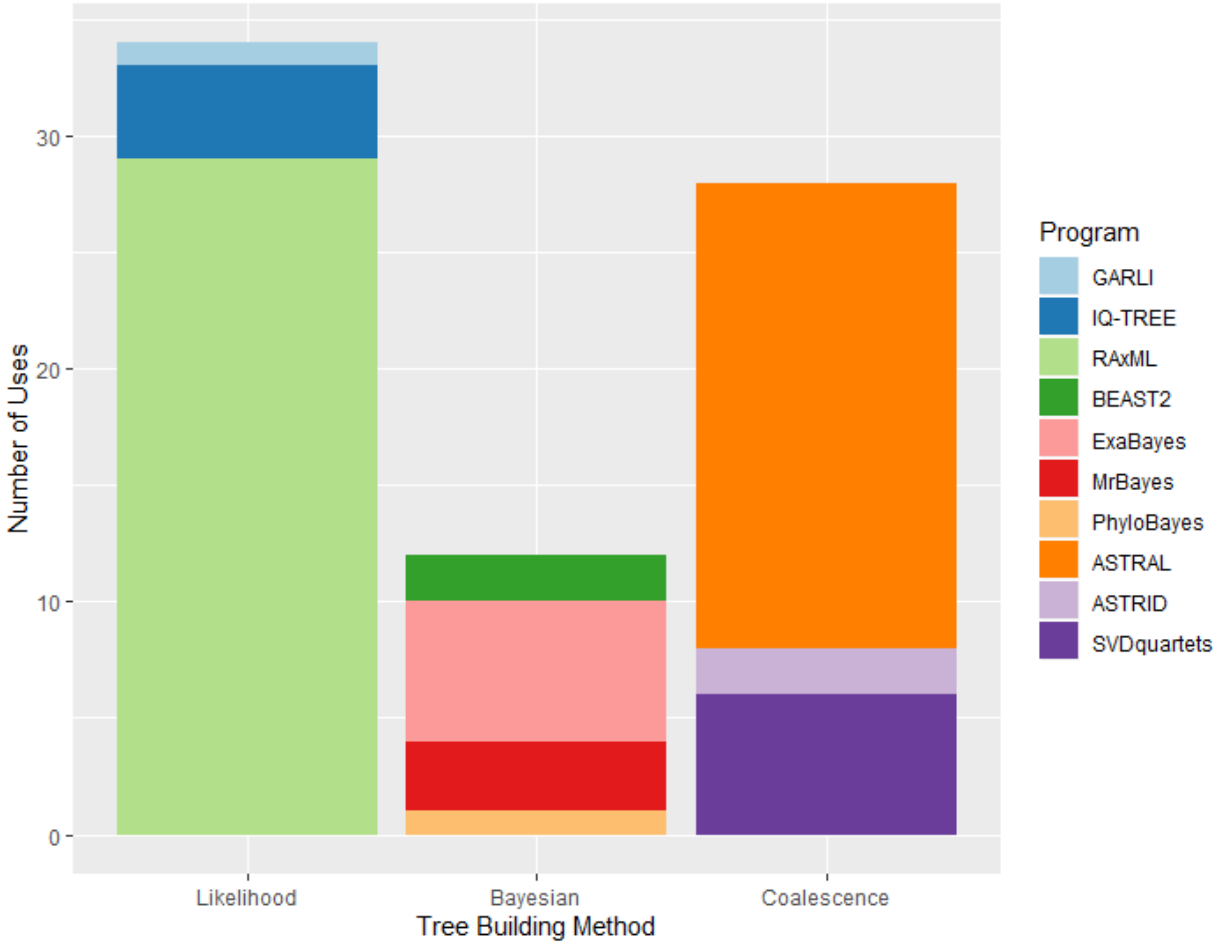
968

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at Insect Systematics and Diversity 2019, 3, 3; doi:10.1093/isd/ixz016

969 **Figure 3.** Breakdown of the phylogenetic programs used by arthropod UCE-based publications

970 (as of July 2019).



971

972

973 **Supplementary Data. Glossary of commonly encountered terminology in UCE generation**
974 **and analyses.**

975

976 **AHE:** Anchored Hybrid Enrichment = A target enrichment method that uses in solution
977 hybridization to capture exonic genes for phylogenomic studies.

978

979 **Adapter:** A short piece of known DNA attached to the genomic DNA of interest to identify the
980 sample once mixed with other samples.

981

982 **Assembly:** Assembly of fragment sequences into higher order structures based on their overlap
983 and reference sequence, where appropriate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

Biotinylation: The process of attaching biotin to proteins and other macromolecules, in this case to bind the DNA regions of interest to the streptavidin magnet beads during the in-solution hybridization process. Streptavidin has a high affinity for biotin, being one of the strongest non-covalent interactions known in nature.

Contig: A contiguous stretch of DNA sequence that is the result of assembly of multiple overlapping sequence reads into a single consensus sequence.

de Bruijn Graph: A graph theory method for assembling a long sequence from overlapping fragments. The de Bruijn graph is a set of unique substrings (words) of a fixed length (a *k*-mer) that contain all possible words in the data set exactly once. The sequence reads are split into all possible *k*-mers, and overlapping *k*-mers are linked by edges in the graph. Reads are then mapped onto the graph of overlapping *k*-mers in a single pass, greatly reducing the computational complexity of genome assembly.

***de novo* Assembly:** Assembly of contigs without a reference genome.

Exon: A portion of a gene that is transcribed and spliced to form the final messenger RNA (mRNA). Exons contain protein-coding sequence and untranslated upstream and downstream regions (3′ UTR and 5′ UTR). Exons are separated by introns, which are sequences that are transcribed by RNA polymerase, but spliced out after transcription and not included in the mature mRNA.

Flanking regions: The areas immediately to the left and right of the UCE core, which are variable and, therefore, the target of UCE capture.

***K*-mer:** A motif (or a small word) of length *k* observed more than once in a genomic or sequenced sequence. E.g., a dinucleotide is a *k*-mer where *k*=2.

In Solution Hybridization: Binding of biotinylated probes with denatured genomic regions of interest in the process of several hours in liquid.

Library: A set of nucleic acid fragments which has undergone all processing steps and is ready for actual sequencing.

Multiplex: A library containing various samples labelled with adapters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2019 doi:10.20944/preprints201905.0384.v2

Peer-reviewed version available at *Insect Systematics and Diversity* **2019**, *3*, 3; doi:10.1093/isd/ixz016

Paired-End Read: A technology that obtains sequence reads from both ends of a DNA fragment template. The use of paired-end sequencing can greatly improve de novo sequencing applications by allowing contigs to be joined when they contain read pairs from a single template fragment, even if no reads overlap.

Probe/Bait: A collection of oligonucleotides that will bind to specific, conserved genome regions of interest, often called baits as they can 'fish' out the region of interest. In our review we refer to baits as the temporary oligonucleotides used during the probe design process, whereas probes are the final product synthesized to capture UCE loci during library prep.

Read: Data output from the analysis of a single fragment (sequence).

SNP: Single-Nucleotide Polymorphism = sequence divergence in the range of a single base.

Target Enrichment: Capturing genomic regions of interest by hybridization to target-specific biotinylated probes, which are then isolated by magnetic beads.

UCE: Ultraconserved Elements = highly-conserved regions within the genome that are shared among evolutionarily distant taxa.