Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

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

Y. Miles Zhang[1*], Jason L. Williams[1], Andrea Lucky[1]

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

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

Abstract:

Targeted enrichment of ultraconserved elements (UCEs) 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, UCEs are underutilized in arthropod phylogenomics, and the methodology can be daunting for beginners. Our goal is to introduce UCEs phylogenomics to a wider audience by summarizing recent advances in UCE phylogenomics in arthropod research to familiarize readers with background theory and steps involved in UCEs phylogenomics. 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 UCEs pipeline to assist phylogeneticists in making informed decisions as they employ this powerful tool. The UCEs pipeline can be divided into the following steps: 1) probe selection and design 2) wet lab work and sequencing, 3) bioinformatics, and 4) phylogenomic analyses. we provide quality-control tips to ensure that best results in data collection and downstream analyses. Our hope is to

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

encourage increased adoption of UCEs in phylogenomics studies, deepen our understanding of the function of UCEs themselves across widely divergent taxa, and toward increased understanding of the tree of life.

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

Introduction & Background

The advent of massively-parallel sequencing technology and the subsequent emergence of the field of phylogenomics has invigorated multiple research areas in biology in a relatively short time span (reviewed in Philippe et al. 2011, Jones and Good 2016). This molecular revolution has opened up unprecedented opportunities to generate large-scale datasets and with the concurrent explosion of analytic and bioinformatics tools has made it possible to address previously intractable systematic challenges. However, the rapidity with which new technologies have emerged has made it difficult for scientists to stay up to date about useful new tools; understanding the steps involved in using new methods presents an additional challenge.

Genome-scale studies are rapidly supplanting the Sanger sequencing-based, multi-locus molecular phylogenetic methods that dominated from the mid-1990's through the early 2000's; today, genomic-scale studies dwarf previous approaches in the sheer scale of data they generate. While the cost and scale of whole-genome sequencing is prohibitive for many researchers, recent advances in sequencing technology and laboratory protocols have made it possible to generate high quality genome-scale datasets using a combination of next-generation sequencing, genomic reduction, and sample multiplexing (Lemmon and Lemmon 2013, McCormack et al. 2013a).

These so-called 'genome reduction' or 'reduced representation' approaches can rapidly generate datasets with thousands of loci, at relatively low cost, for model and non-model taxa alike. Methods such as restriction enzyme-associated DNA sequencing (RADseq; Miller et al. 2007, Baird et al. 2008, Peterson et al. 2012), transcriptomics (Bi et al. 2012), and target enrichment methods such as Anchored Hybrid Enrichment (AHE) (Lemmon et al. 2012) or Ultraconserved Elements (UCEs) (Faircloth 2017) are now widely used for generating genomic-scale data for phylogenomic studies. While these phylogenomics methods are similar in some respects, each has particular 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, the problem of methods choice overload in the era of 'big data' can be daunting for beginners. Potential users need guidance to help them choose methods appropriate to the scale of their research questions, and to navigate the challenges of understanding confusing terminologies, bioinformatics-heavy data processing, and computationally-intensive analyses.

This paper summarizes recent advances in UCEs phylogenomics in arthropod research to familiarize readers with background theory and steps involved in UCEs phylogenomics, and provides quality-control tips to ensure that data collection and downstream analyses are performed with confidence. UCE-based phylogenomics continues to develop rapidly, and the lack of comprehensive review is a significant challenge for potential users to overcome when exploring this option. We define terminology used in association with the UCEs approach, evaluate current laboratory and bioinformatic methods and limitations, and provide a roadmap of steps in the UCEs pipeline to assist phylogeneticists in making informed decisions as they employ this powerful tool.

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). Using existing sequencing technology and laboratory protocols, 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).

The UCEs approach belongs to the broad category of 'target enrichment' phylogenomics techniques, which involve selective capture of genomic regions from a DNA sample before sequencing (Mamanova et al. 2010). Other methods, similar to UCEs, 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, followed by designing custom-made molecular probes (also known as baits) which are hybridized to loci of interest, and sequenced on a massively-parallel platform. The group-specific hybridization of 'baits' used to 'fish' target loci (Faircloth et al. 2012, Lemmon et al. 2012, Faircloth 2017) imposes few limitations on genetic material that can be used, in terms of quantity and quality. The main difference between AHE and UCEs is the nature of the loci targeted, with AHE focusing on fewer loci (300-600) that are exclusively exonic, while the UCEs use fewer probes to target more loci (>1000) and 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

a more diverse set of probes per locus, the details of the methodology are largely proprietary (Lemmon et al. 2012). The focus of this review is on the use of UCEs for phylogenomics of Arthropoda, in response to recent interest in its usage.

Advantages of UCEs Phylogenomics

The UCEs approach has become an increasingly popular target enrichment method for generating phylogenomic data, as it offers advantages over traditional Sanger sequencing methods, both in terms of quantity and quality of data generated. UCEs have successfully been used in studies across a broad array of taxa including birds (McCormack et al. 2013b, Musher and Cracraft 2018), mammals (McCormack et al. 2012), fish (Faircloth et al. 2013, Chakrabarty et al. 2017), 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. 2017a, 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) to population level studies at shallow time scales (Harvey et al. 2016, Manthey et al. 2016, Branstetter and Longino 2019, Myers et al. 2019).

Benefits of using UCEs 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 compared to more proprietary alternatives such as AHE. UCE datasets can be easily standardized using the same probe set and re-applied over multiple studies, allowing for

combinable data across studies using the same probe set, exon/transcriptome (Bossert et al. 2019), 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 enzyme digestion process. Additionally, because target enrichment is DNA-based, it works well with degraded or low-quantity samples such as older, dried museum specimens (Blaimer et al. 2016a, Lim and Braun 2016, Ruane and Austin 2017), which may be unusable in both traditional restriction enzyme-based or transcriptomics studies as they require large quantities of high-quality DNA or RNA from fresh or carefully-preserved tissues. 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) use 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.

UCEs and Arthropods

A limited number of arthropod-focused research groups have adopted UCEs for use in phylogenetic studies, despite the availability of probe sets for use across diverse arthropod taxa (Faircloth 2017), and in contrast to the wide adoption of UCEs in vertebrate phylogeny. This slow adoption may be due to the use of other target enrichment methods, but also to lack of knowledge about UCE utility and versatility within the arthropod phylogenomics community. The UCEs approach was first demonstrated outside of vertebrates in the insect order Hymenoptera (See Table 1). To date two published probe sets exist for Hymenoptera, with the hym-v2 (31,829 probes for 2,590 UCEs) by Branstetter et al. (2017a) including most of the original hym-v1 probe sets (2,749 probes for 1,510 UCEs) developed by Faircloth et al. (2015),

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

while excluding poorly performed loci that were found to be <1kb apart. Other arthropods groups for which published UCE data exist include Arachnida, Coleoptera, and Hemiptera (Table 1). The utility of UCEs extends beyond purely phylogenetic research, as demonstrated by the Grab et al. (2019) use of UCE-based community phylogenomics to reveal the importance of bee phylodiversity in agriculture. Similarly, Santos et al. (2019) used a combination of UCE-based phylogeny and geometric morphometrics to explore the evolution of parasitic wasp body shape, while Derkarabetian et al. (2018a) used single nucleotide polymorphisms (SNPs) generated from UCE data to demonstrate the success of unsupervised machine learning in species delimitation. The number and taxonomic breadth of UCE based phylogenies will undoubtedly increase in the taxa for which UCE data are available, and hope that this step-by-step guide will make the UCEs research pipeline more approachable to researchers working on additional groups.

UCEs Phylogenomics Pipeline

Below we outline the steps in the UCE pipeline: 1) probe selection and design 2) wet lab work and sequencing, 3) bioinformatics, and 4) phylogenomic analyses. We describe choices at each stage, and visualize the process in a workflow diagram (Figure 1). A glossary of technical terms are provided in the supplementary glossary.

Probe Selection & Design

Probe sets are designed based on UCE loci identified from published genomes for each taxonomic group. This could prove challenging if there are no published genomes for the group of interest. Nevertheless, low coverage genome sequencing may be an appropriate first step as it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

becomes increasingly affordable. The choice of sequenced genomes should ideally reflect diversity within the group of interest; even 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).

A probe set is a collection of biotinylated oligonucleotides that are designed to bind with specific genome regions of interest. These probes are often 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). The denatured and cooled DNA and the probes are combined together, allowing for 'in solution' hybridization to targets. Then streptavidin-coated magnetic beads are added into the solution, which bind to the probe-DNA hybrids through the biotin on the probe set. Any unwanted DNA fragments are then washed away thus leaving only the desired regions (Gnirke et al. 2009).

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 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). Identifying conserved sequences shared among divergent genomes consists of two principal steps: 1) identify genomic regions of high conservation among divergent lineages; and 2) design synthetic oligonucleotide 'baits' to collect these conserved regions from DNA libraries. A generalized workflow for identifying conserved sequences shared among divergent genomes and enrichment probes design can be found in Faircloth (2017). These probe sets can also be updated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

to incorporate additional loci, for example the Hymenoptera probe set hym-v1 was improved by the publication of hym-v2, which included most of the original hym-v1 loci, additional, new loci, and probes targeting 16 nuclear genes that have been commonly sequenced in insects to allow for 'back compatibility' with Sanger-era data (Branstetter et al. 2017a). This ability to combine 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 are difficult or impossible to obtain multiple times. *In silico* testing of the existing probe sets have shown moderate success in sister outgroups, such as using the Hemiptera probe set to capture Thysanoptera (Thrips) UCEs (Faircloth 2017). The recommended best practice, however, is designing new probe sets for a target taxon to ensure greatest sequence capture success.

Wet Lab Work & Sequencing

Selection of appropriate specimens for DNA extraction is vital to any phylogenetic work. The first major limitation for molecular phylogenetics is capture of high-quality DNA, which can be affected by age and specimen preservation methods (Short et al. 2018). This is especially problematic for arthropods, as most natural history collections historically preserved specimens pinned dry or stored in 70% ethanol at room temperature; both methods result in DNA degradation over time, with especially rapid decline of the latter (Short et al. 2018). Other complications include the percentage of ethanol used for preservation (95% is recommended), the number of freeze/thaw cycles (as few as possible), and the number/frequency of alcohol changes (regular enough to maintain high EtOH concentration and submerged specimens).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

Fortunately, massive-parallel methods can capitalize on shorter, more degraded DNA fragments from these specimens, as shown by Blaimer et al. (2016a) in a study that generated nearly 1000 UCEs loci from pinned bee specimens up to 121 years old. Based on their study, pinned specimens less than 20 years old had significantly higher pre- and post-library concentrations, UCEs contig lengths, and locus counts compared to older specimens (Blaimer et al. 2016a). Another challenge for many arthropod specimens is their small size (<5mm) and therefore low DNA yield, especially if non-destructive sampling is required in order to retain specimens. The total yield can be increased with the use of DNA amplification kits; albeit at a higher cost (Cruaud et al. 2018). Alternatively, newly developed library prep kits such as the Kapa Hyper and PacBio SMRT are capable of extracting genomic data even with very low (<5ng) DNA input (Kingan et al. 2019). Thus, in order to ensure high quality DNA generation, fresh, well-preserved material (in 95% EtOH stored in -80°C or -20°C) is recommended, with pinned specimens collected within the past 20 years being suitable.

Due to the wide range of taxa targeted by available UCE probe sets, careful selection of tissue types can significantly lower the potential of contamination by non-target organisms. Additional 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* script in PHYLUCE which extracts the *COI* barcode region be used for validating the presence of a single or multiple species (Bossert and Danforth 2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

Once DNA has been extracted from the organism of choice, the wet lab protocol is similar across taxa. Depending on the stage of degradation, the extracted genomic DNA may need to be sheared to a target size of 400–600bp using sonication or enzymatic digestion. The degree of DNA degradation will determine the duration of sonication needed; this can be assessed using gel electrophoresis, or automated electrophoresis systems such as TapeStation or Bioanalyzer. UCE sample preparation consists of four main steps: 1) adapter ligation to bind to Illumina flow cell; 2) hybridization of RNA probes to bind to the region of interest; 3) pooling enrichment of DNA libraries using PCR; and 4) DNA size selection for sequencing (details in Faircloth et al. (2015)). Multiple clean up steps using magnetic beads, and pooling of samples are also interspersed between each step.

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. Some commercial laboratories also offer UCE enrichment services, handling all library preparation, enrichment, and sequencing; customers simply submit DNA extracts and then receive sequence data. This ‘concierge service’ increases the cost considerably

(approximately ~$120 per specimen), but is an attractive option for researchers lacking the infrastructure or personnel to undertake wet lab protocols.

Access to high performance computing (HPC) will greatly expedite bioinformatic and phylogenomic analyses, especially when processing large batches of samples. PHYLUCE and many associated data-processing programs can be run in parallel, greatly reducing execution time in computationally intensive steps such as demultiplexing and assembly. Similarly, many phylogenomics programs (discussed below) can also be expedited through this process.

Bioinformatics

Processing UCE data involves three principal steps: 1) Demultiplexing, filtering and trimming of raw Illumina reads; 2) Contig assembly; and 3) UCEs processing for phylogenomic analysis. Currently, the most widely-used bioinformatics pipeline for UCEs data processing is PHYLUCE (Faircloth 2015), which includes a suite of Python wrapper scripts for these steps by calling other programs (detailed below) and batch processing many samples at once. Alternatively, the SECAPR pipeline also functions similarly to PHYLUCE and can be used for batch processing of UCE data (Andermann et al. 2018).

1*) 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, programs such as Trim Galore! (https://github.com/FelixKrueger/TrimGalore) can be used.

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, we recommend testing multiple k-mer values starting at the default of 35 and moving up to 55–65 to find the best trade off in terms of contig size vs. number. Alternatively, using SPAdes and the VelvetOptimiser wrapper script for velvet can automatically estimate k-mers.

3) *UCEs processing for phylogenomics*. Once assembled, contigs are then processed to determine which ones represent enriched UCEs loci, followed by removal of potential paralogs from the data set. 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 matrices can be exported in a variety of commonly-used formats (e.g. phylip, nexus, etc.) for downstream phylogenomics analyses.

Allelic phasing

Allelic phasing is an additional, optional data processing step, that extracts SNPs from UCEs loci by separating (phasing) the heterozygous sites into two allele sequences (Andermann et al. 2019). This can be performed in both PHYLUCE and SECAPR, and has been shown to

provide more accurate estimation of tree topology and divergence times than using contig sequences, especially at shallow phylogenetic levels under MSC models (Andermann et al. 2019). Common assembler programs are not optimized 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). Additionally, phasing the sequence double the effective sample size, as two sequences are compiled for diploid individual (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 quality phased results, as contigs with lower coverage will not be phased accurately.

Phylogenomic Analyses

Preparation of raw data for tree building has become highly automated in response to the large volumes of data generated by high-throughput methods. This standardization helps reduce errors and inconsistencies, but can also be responsible for introducing errors in UCEs datasets as it is impossible to check thousands of DNA alignments by eye.

Data Filtering

Data filtering is a vital step in quality control of phylogenomic studies. Sequencing thousands of genes across many samples can often to missing data in certain taxa. We advise using different filtering criteria to generate multiple datasets and thereby find a balance between maintaining sequence quantity and quality. Although more stringent filtering will ensure

inclusion of only high-quality loci shared by most of the individuals, it can also result in exclusion of many potentially useful sequences. Additionally, filtering data-based gene information content (e.g. rate of evolution, GC bias) may benefit UCE data sets, as both the UCEs core and flanking regions tend to show lower variability than other classes of more rapidly evolving loci (Mclean et al. 2018). What information content should be filtered and by how much is largely dependent on the study organism and evolutionary scenario. For example, Bossert et al. (2017) demonstrated that GC bias is negatively correlated with topological support in bees while Mclean et al. (2018) were able to improve gene tree precision in a ground squirrel study by filtering based on the number of variable sites.

Data Partition

Partitioning of UCEs data can be divided into two 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, or 2) assign each UCEs 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 (Tagliacollo and Lanfear 2018). In recent years, selecting best-fit partitioning schemes often employed the k-means clustering of sites based on evolutionary rates (Frandsen et al. 2015), which subdivides data into partitions based on evolutionary rates, thus avoiding *a priori* partitioning by the user. However, recent studies have shown the k-means algorithm could be unreliable for UCEs data, as it generates a partition comprised of all the invariant sites in the dataset, possibly misleading phylogenetic inference methods (Baca et al. 2017b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

Currently, the most promising method for partitioning UCE data seems to be the Sliding-Window Site Characteristics (SWSC) by 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 often produces better log-likelihood than alternative partitioning schemes such as partitioning-by-locus (Tagliacollo and Lanfear 2018, Branstetter and Longino 2019).

Tree Building

Once datasets have been generated, downstream analyses on UCEs data are similar to phylogenetic analyses performed on most other data types (e.g. Sanger sequencing). A variety of tree-building programs can be used for reconstructing phylogeny from UCE datasets, including Maximum Likelihood (ML), Bayesian Inference (BI), or Multispecies Coalescent/Species Tree (MSC) methods.

Maximum Likelihood

The most popular programs for phylogenetic reconstruction are those that employ ML methods such as RAxML (Stamatakis 2006) and IQ-TREE (Nguyen et al. 2014). IQ-TREE in particular have 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 to illustrate disagreement among loci and sites (Minh et al. 2018). These functionalities may soon be common across platforms; for example, the developers of RAxML are releasing RAxML-NG, which will include features comparable to IQ-TREE for genomic-scale data (Kozlov et al. 2018).

Bayesian Inference

Commonly used Bayesian programs for phylogenomics data include BEAST (Drummond and Rambaut 2007), BEAST 2 (Bouckaert et al. 2014), and ExaBayes (Aberer et al. 2014). While the intricacies of Bayesian phylogenetic analyses are beyond the scope of this paper, excellent and detailed overviews – both theoretical and practical – are available (reviewed in dos Reis et al. 2016, Nascimento et al. 2017, Bromham et al. 2018). Bayesian analyses are extremely sensitive to priors set by users, as they can affect resulting topologies. 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 divergence dating analyses.

Divergence Dating

Both ML and BI approaches can be used for divergence date estimation using fossils as calibration points, and for generating dated chronograms. Divergence date estimation can be performed using Bayesian inference programs such as BEAST/BEAST2 mentioned above, and the mcmctree package in PAML (Yang 2007). While mcmctree is faster computationally, it is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

not intuitive on how to set prior distributions on fossil calibrations. BEAST/BEAST2 by comparison is easier to understand, and offers more analytical options such as the incorporation of fossils directly into the phylogeny using the fossilized birth-death model (Heath et al. 2014). The fossilized birth-death model can be implemented in RevBayes (Höhna et al. 2016), or in BEAST2 using the 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). 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 incorporation of gene tree stochasticity into the inference of species phylogenies, under the MSC model (Heled and Drummond 2009). Thus, MSC tree methods such as ASTRAL (Mirarab et al. 2014) and SVDQuartets (Chifman and Kubatko 2014) are often presented in conjunction with traditional concatenation methods for genomic datasets. The theory and advantages of species trees are comprehensively reviewed by Liu et al. (2015), who highlight the main advantage of coalescent-based summary methods as circumventing the discordance between gene trees and species trees as a result of high incomplete lineage sorting (Chou et al. 2015).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

Data Availability

One hallmark feature of UCE data is its open source nature, probe sets, protocols, and previously published data are made publicly available, ensuring repeatability – foundation of 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; 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

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 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 by Bossert et al. (2019). However, as 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 seems 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.

New probe sets can be designed for specific groups of interest, which should limit contamination from distantly related taxa (Bossert and Danforth 2018). Probe design can be tailored to desired evolutionary scale, as fewer longer loci might be better for shallow scales than many, shorter loci. Additionally, Hedin et al. (2019) that, in the arachnid probe set, different UCE probes sometimes target regions of the same protein, or include non-orthologous sequences. This is unsurprising given the wide phylogenetic depth of the probe set, which was designed to target all arachnids, and suggests the likelihood this being true in probe sets designed for other arthropod groups. Therefore, continued refinement of existing probe sets is needed to increase capture success while minimizing duplicates and paralogous loci is needed. As our understanding of UCEs develops through studies on different organisms, identifying and incorporation of functional genomics will yield interesting comparative studies across deeply-divergent taxonomic groups (Bossert and Danforth 2018).

UCEs 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, degree of missing

data, choice of alignment methods, and composition bias (Bossert et al. 2017, Camacho et al. 2019). Recent research also suggests phylogenomic results can be strongly affected by a tiny proportion of highly biased loci or sites (Shen et al. 2017). 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).

Conclusion

Ultraconserved elements-based phylogenomic studies have been rapidly adopted by researchers working on a subset of arthropod taxa since their introduction by Faircloth et al. (2012). This review aimed to demonstrate the versatility of UCEs data at both deep and shallow evolutionary scale, by providing a step-by-step guide to UCEs and a summary of 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. We expect this exciting new phylogenomic tool will provide new insights in the continued pursuit of building the tree of life.

Acknowledgements

We would like to thank Michael Branstetter for comments that greatly improved the manuscript, Eliana Buenaventura, Shahan Derkarabetian, Katherine Noble, and Matthew van

Dam for informative discussions on UCEs research in their perspective groups. We thank 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.
Ali, O. A., S. M. O'Rourke, S. J. Amish, M. H. Meek, G. Luikart, C. Jeffres, and M. R. Miller. 2016. RAD Capture (Rapture): Flexible and efficient sequence-based genotyping. Genetics 202: 389-400.
Andermann, T., A. Cano, A. Zizka, C. Bacon, and A. Antonelli. 2018. SECAPR-a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. PeerJ 6: e5175.
Andermann, T., A. M. Fernandes, U. Olsson, M. Topel, B. Pfeil, B. Oxelman, A. Aleixo, B. C. Faircloth, and A. Antonelli. 2019. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. Systematic Biology 68: 32-46.
Baca, S. M., A. Alexander, G. T. Gustafson, and A. E. Z. Short. 2017a. Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of 'Hydradephaga'. Systematic Entomology 42: 786-795.
Baca, S. M., E. F. A. Toussaint, K. B. Miller, and A. E. Z. Short. 2017b. Molecular phylogeny of the aquatic beetle family Noteridae (Coleoptera: Adephaga) with an emphasis on data partitioning strategies. Molecular Phylogenetics and Evolution 107: 282-292.
Baira, E., J. Greshock, G. Coukos, and L. Zhang. 2008. Ultraconserved elements: genomics, function and disease. RNA Biology 5: 132-134.
Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376.
Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, and A. D. Prjibelski. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19: 455-477.
Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. Science 304: 1321-1325.
Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13: 403.
Blaimer, B. B., J. R. Mawdsley, and S. G. Brady. 2018a. Multiple origins of sexual dichromatism and aposematism within large carpenter bees. Evolution 72: 1874-1889.
Blaimer, B. B., M. W. Lloyd, W. X. Guillory, and S. G. Brady. 2016a. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS One 11: e0161531.

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 syy088.
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. V. Edwards. 2019. Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ 7: e6399.
Bromham, L., S. Duchene, X. Hua, A. M. Ritchie, D. A. Duchene, and S. Y. W. Ho. 2018. Bayesian molecular dating: opening up the black box. Biological Reviews 93: 1165-1191.

Camacho, G. P., M. R. Pie, R. M. Feitosa, and M. S. Barbeitos. 2019. Exploring gene tree incongruence at the origin of ants and bees (Hymenoptera). Zoologica Scripta 48: 215-225.
Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-1973.
Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17: 540-552.
Chakrabarty, P., B. C. Faircloth, F. Alda, W. B. Ludt, C. D. Mcmahan, T. J. Near, A. Dornburg, J. S. Albert, J. Arroyave, and M. L. Stiassny. 2017. Phylogenomic systematics of ostariophysan fishes: Ultraconserved elements support the surprising non-monophyly of characiformes. Systematic Biology 66: 881-895.
Chifman, J., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30: 3317-3324.
Chou, J., A. Gupta, S. Yaduvanshi, R. Davidson, M. Nute, S. Mirarab, and T. Warnow. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16: S2.
Cooke, C. 2018. Forest Micro-Hymenoptera, including those attacking trees (Cynipidae: Oak Gall Wasps) and those potentially defending them (Parasitic Pteromalidae). PhD, University of Maryland, College Park, MD, USA.
Crawford, N. G., B. C. Faircloth, J. E. McCormack, R. T. Brumfield, K. Winker, and T. C. Glenn. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters rsbl20120331.
Cruaud, A., G. Groussier, G. Genson, L. Saune, A. Polaszek, and J. Y. Rasplus. 2018. Pushing the limits of whole genome amplification: successful sequencing of RADseq library from a single microhymenopteran (Chalcidoidea, *Trichogramma*). PeerJ 6: e5640.
Derkarabetian, S., S. Castillo, P. K. Koo, S. Ovchinnikov, and M. Hedin. 2018a. An Empirical Demonstration of Unsupervised Machine Learning in Species Delimitation. bioRxiv 429662.
Derkarabetian, S., J. Starrett, N. Tsurusaki, D. Ubick, S. Castillo, and M. Hedin. 2018b. A stable phylogenomic classification of Travunioidea (Arachnida, Opiliones, Laniatores) based on sequence capture of ultraconserved elements. ZooKeys 1-36.
dos Reis, M., P. C. Donoghue, and Z. Yang. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nature Reviews Genetics 17: 71-80.
Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7: 214.
Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.
Faircloth, B. 2013. Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming.
Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32: 786-788.
Faircloth, B. C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods 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.
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.
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.
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.
Kingan, S. B., H. Heaton, J. Cudini, C. C. Lambert, P. Baybayan, B. D. Galvin, R. Durbin, J. Korlach, and M. K. Lawniczak. 2019. A high-quality *de novo* genome assembly from a single mosquito using PacBio sequencing. Genes 10: 62.
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.
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.
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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

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.
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 B: Biological Sciences 286.
Shen, X. X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology & Evolution 1: 126.
Short, A. E. Z., T. Dikow, and C. S. Moreau. 2018. Entomological Collections in the Age of Big Data. Annual Review of Entomology 63: 513-530.
Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19: 1117-1123.
Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Systematic Biology 63: 83-95.
Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.
Starrett, J., S. Derkarabetian, M. Hedin, R. W. Bryson Jr, J. E. McCormack, and B. C. Faircloth. 2017. High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. Molecular Ecology Resources 17: 812-823.
Streicher, J. W., and J. J. Wiens. 2017. Phylogenomic analyses of more than 4000 nuclear loci resolve the origin of snakes among lizard families. Biology Letters 13: 20170393.
Suchan, T., C. Pitteloud, N. S. Gerasimova, A. Kostikova, S. Schmid, N. Arrigo, M. Pajkovic, M. Ronikier, and N. Alvarez. 2016. Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. PloS One 11: e0151651.
Tagliacollo, V. A., and R. Lanfear. 2018. Estimating improved partitioning schemes for ultraconserved elements. Molecular Biology and Evolution 35: 1798-1811.
Van Dam, M. H., M. Trautwein, G. S. Spicer, and L. Esposito. 2019. Advancing mite phylogenomics: Designing ultraconserved elements for Acari phylogeny. Molecular Ecology Resources 19: 465-475.
Van Dam, M. H., A. W. Lam, K. Sagata, B. Gewa, R. Laufa, M. Balke, B. C. Faircloth, and A. Riedel. 2017. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-weevils. PLoS One 12: e0188044.
Ward, P. S., and M. G. Branstetter. 2017. The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. Proceedings of the Royal Society B: Biological Sciences 284.

 doi:10.20944/preprints201905.0384.v1

Weitemier, K., S. C. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Applications in Plant Sciences 2: 1400042.
Wood, H. M., V. L. Gonzalez, M. Lloyd, J. Coddington, and N. Scharff. 2018. Next-generation museum genomics: Phylogenetic relationships among palpimanoid spiders using 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.
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.
Zerbino, D., and E. Birney. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research 18: 821-829.
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.

Table 1. Published studies using arthropod UCEs datasets as of May 2019.

Taxonomic Group	Probe Sets	Publications	References
Hymenoptera	Hymenoptera 1.5Kv1 Hymenoptera 2.5Kv2 Full Hymenoptera 2.5Kv2 Ant-Specific	20	(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, Grab et al. 2019, Santos et al. 2019)
Arachnida	Arachnida 1.1Kv1 Mite-specific (*in silico* only)	10	(Faircloth 2017, Starrett et al. 2017, Derkarabetian et al. 2018a, Derkarabetian et al. 2018b, Hedin et al. 2018a, Hedin et al. 2018b, Wood et al. 2018, Hedin et al. 2019, Van Dam et al. 2019)
Coleoptera	Coleoptera 1.1Kv1	3	(Baca et al. 2017a, Faircloth 2017, Van Dam et al. 2017)
Hemiptera	Hemiptera 2.7Kv1	3	(Faircloth 2017, Forthman et al. 2019, Kieran et al. 2019)
Diptera	Diptera 2.7Kv1	1	(Faircloth 2017)
Lepidoptera	(*in silico* only)	1	(Faircloth 2017)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201905.0384.v1

Figure 1. Generalized workflow of the UCEs pipeline.