

## Molecular Phylogenetics, Phylogenomics, and Phylogeography

# To design, or not to design? Comparison of beetle ultraconserved element probe set utility based on phylogenetic distance, breadth, and method of probe design

Grey T. Gustafson<sup>1,\*</sup>, Rachel D. Glynn<sup>1</sup>, Andrew E.Z. Short<sup>2</sup>, Sergei Tarasov<sup>3</sup>, Nicole L. Gunter<sup>4</sup>

<sup>1</sup>Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA, <sup>2</sup>Department of Ecology and Evolutionary Biology and Division of Entomology, Biodiversity Institute, University of Kansas, Lawrence, KS, USA, <sup>3</sup>Finnish Museum of Natural History, Pohjoinen Rautatiekatu 13, FI-00014 Helsinki, Finland, <sup>4</sup>The Cleveland Museum of Natural History, Cleveland, OH, USA  
\*Corresponding author, mail: [grey.gustafson@nau.edu](mailto:grey.gustafson@nau.edu)

Subject Editor: Adriana Marvaldi

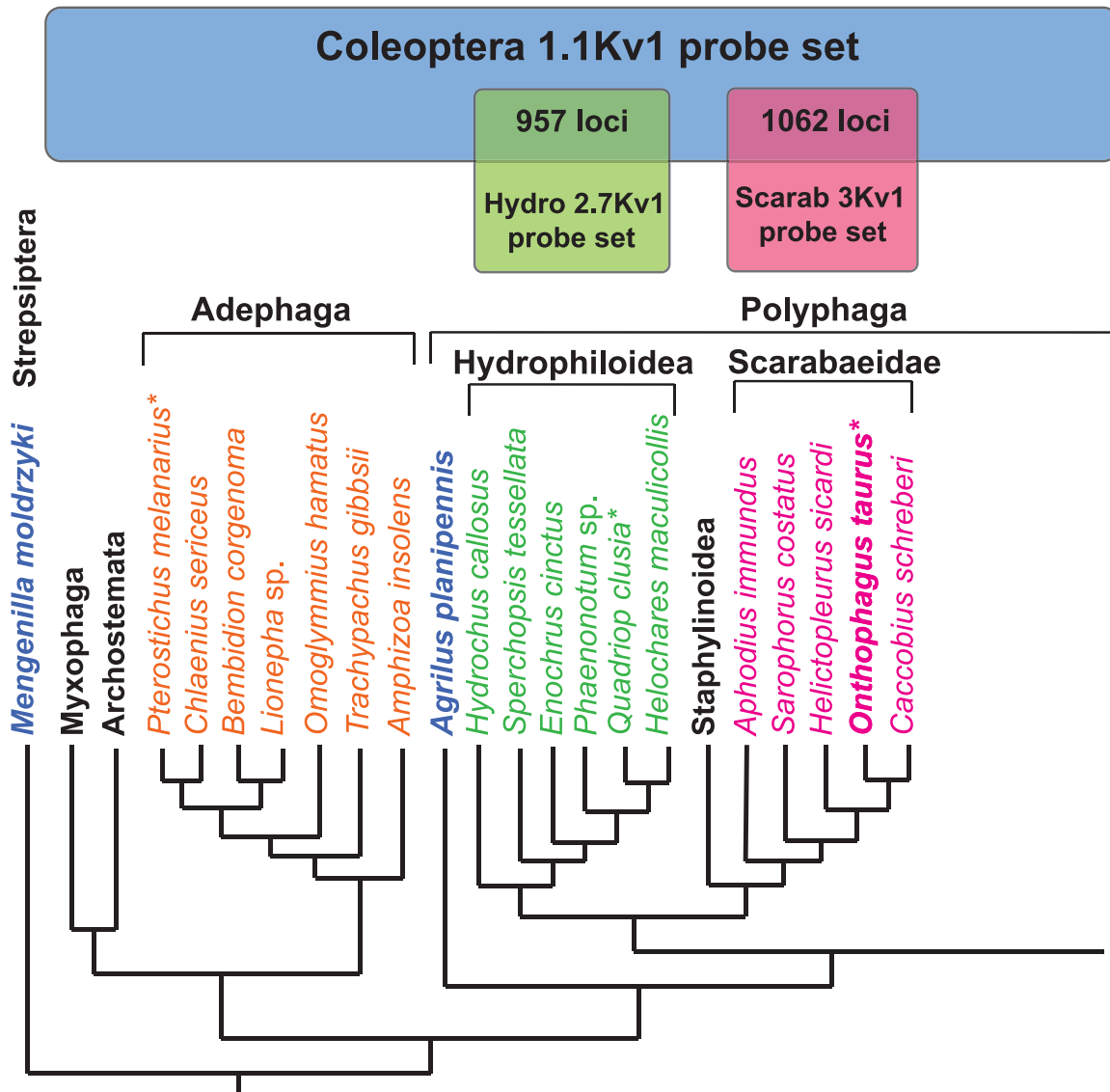
Received on 9 January 2023; revised on 16 May 2023; accepted on 22 June 2023

Tailoring ultraconserved element (UCE) probe set design to focal taxa has been demonstrated to improve locus recovery and phylogenomic inference. However, beyond conducting expensive in vitro testing, it remains unclear how best to determine whether an existing UCE probe set is likely to suffice for phylogenomic inference or whether tailored probe design will be desirable. Here we investigate the utility of 8 different UCE probe sets for the in silico phylogenomic inference of scarabaeoid beetles. Probe sets tested differed in terms of (i) how phylogenetically distant from Scarabaeoidea taxa those used during probe design are, (ii) breadth of phylogenetic inference probe set was designed for, and (iii) method of probe design. As part of this study, 2 new UCE probe sets are produced for the beetle family Scarabaeidae and superfamily Hydrophiloidea. We confirm that probe set utility decreases with increasing phylogenetic distance from target taxa. In addition, narrowing the phylogenetic breadth of probe design decreases the phylogenetic capture range. We also confirm previous findings regarding ways to optimize UCE probe design. Finally, we make suggestions regarding assessment of need for de novo probe design.

Key words: Hydrophiloidea, phylogenomics, probe design, Scarabaeidae, ultraconserved elements

## Graphical Abstract

- Two new UCE probe sets developed: Hydro 2.7Kv1 & Scarab 3Kv1
- UCE probe set design recommendations
- Insights for picking among existing UCE probe sets



## Introduction

Ultraconserved elements (UCEs) sensu (Faircloth et al. 2012) have become widely used for phylogenomic analysis of different insect groups including Coleoptera (Baca et al. 2017, 2021, Van Dam et al. 2017, Gustafson et al. 2020, Kobayashi et al. 2021, Bradford et al. 2022, Sota et al. 2022), Diptera (Buenaventura et al. 2021), Hemiptera (Forthman et al. 2019, 2020, Kieran et al. 2019), Hymenoptera (Blaimer et al. 2015, Faircloth et al. 2015, Branstetter et al. 2017b, Bossert et al. 2019, Cruaud et al. 2019, Zhang et al. 2020, Longino and Branstetter 2021, Pisanty et al. 2022), Isoptera

(Hellemans et al. 2022), and Psocodea (Zhang et al. 2019a, Manchola et al. 2022). Their unique features consisting of a highly conserved core with variable flanking regions allow their use in reconstructing both deep- and shallow-level evolutionary relationships (Crawford et al. 2012, Faircloth et al. 2012, Smith et al. 2014, Manthey et al. 2016, Longino and Branstetter 2020). The shorter length of UCE loci relative to other targeted capture approaches (Karin et al. 2020), also facilitates the incorporation of museum specimens into phylogenomic analyses further increasing their utility (Blaimer et al. 2016, McCormack et al. 2016, Ruane and Austin 2017, Van Dam et

al. 2017). It is also becoming increasingly evident that UCE loci are largely exonic in nature in invertebrates, as opposed to vertebrates where intronic loci are predominant (Hedin et al. 2019, Van Dam et al. 2021). Thus, UCEs have further practicality in their ability to incorporate transcriptomic data, allowing taxa sampled as part of the major 1Kite initiative (<https://1kite.org/>) to be included in UCE-based phylogenomic studies (Bossert et al. 2019, Baca et al. 2021).

Given their broad phylogenetic viability and increasing use within insect taxa, it is little surprise there has been a proliferation in the number of available UCE probe sets (reviewed in Zhang et al. 2019b, see also Buenaventura et al. 2021, Hellemans et al. 2022, Liu et al. 2022, Van Dam et al. 2022b). Faircloth (2017) produced several publicly available probe sets for Coleoptera, Diptera, Hemiptera, and Lepidoptera to join the original Hymenoptera probe set (Faircloth et al. 2015), all of which were designed to capture loci across entire orders. However, unlike the situation in vertebrates where the original Tetrapod 2.5kv1 and 5kv1 probe sets (Faircloth et al. 2012) provide abundant locus recovery for analysis (i.e., analyzed matrices contained between ~1,500–4,000 UCE loci) across birds (McCormack et al. 2013), mammals (Esselstyn et al. 2017), squamates (Streicher and Wiens 2017), and clades within Anura (Streicher et al. 2018); the insect probe sets do not appear to have the same universal applicability. The original Hymenoptera probe set, hym-v1, developed using 2 genomes: *Nasonia vitripennis* and *Apis mellifera*, showed decreased locus recovery within the order as phylogenetic distance away from these taxa increased (Faircloth et al. 2015). This led Branstetter et al. (2017a) to refine the existing hym-v1 probes using additional genomic resources and develop novel probes for newly identified loci across a more diverse swath of hymenopteran genomes, producing the hym-v2 probe set which consists of old and new probes. As part of this new probe design, the inclusion of 2 ant genomes resulted in probes tailored for use in ants. Implementing the ant-tailored probes resulted in significantly improved locus recovery in taxa for phylogenomic analysis (Branstetter et al. 2017a). Baca et al. (2017) used the Coleoptera 1kv1 probe set for phylogenomic inference of the beetle suborder Adephaga but found locus recovery strongly diminished. Despite its design for use across the entire order Coleoptera, the genomes used in the Coleoptera 1.1kv1 probe design came entirely from the suborder Polyphaga (Fig. 1), given the genomic resources available at the time (Faircloth 2017). Because of this, Gustafson et al. (2019) designed a novel UCE probe set using adephagan genomes, tailored specifically for use in Adephaga (Fig. 1). Like the hym-v2, the Adephaga 2.9kv1 probe set included probes from the original Coleoptera 1.1kv1 design alongside novel tailored probes, and similar to Branstetter et al. (2017a), in vitro testing of the tailored probe set revealed increased locus recovery over the original probe set (Gustafson et al. 2020).

While reduced locus recovery in beetles belonging to an entirely different suborder than those used in probe design may be expected, the same is not necessarily true regarding a family of polyphagan beetles that had representation during design. Van Dam et al. (2017) encountered reduced in vitro locus recovery in the entimine weevil tribe Eupholini, despite a member of Curculionidae (a scolytine: *Dendroctonus ponderosae*) having been included in the Coleoptera 1.1kv1 probe design (Fig. 1). The estimated divergence time between Entiminae and Scolytinae is ~90 Ma (Shin et al. 2018), which is comparable to the split between palaeognath birds (ostriches etc.) and the remainder of Aves (Kimball et al. 2019). Given the original tetrapod probe set worked across all of birds and for even deeper divergences within Tetrapoda, this suggests a comparable single universal probe for all Coleoptera (and likely other insect orders as well) is intractable. The reason for this may have to do with the massive diversity of Coleoptera, including at the genomic level, where

varied life histories and diets have produced considerable differences in genome size and complexity through the incorporation and proliferation of horizontally transferred genes (Kirsch et al. 2014, Van Dam et al. 2017, McKenna 2018, McKenna et al. 2019, Lata et al. 2022). Thus, to obtain thousands of loci for analysis, it will likely be necessary to design probe sets tailored to focal clades of beetles. However, a relatively uniform set of homologous loci for use across related taxa within a lineage is desirable for affording cross compatibility between phylogenetic studies, as opposed to the proliferation of isolated probe sets tailored only to the goals of individual studies. Additionally, many research groups may not have the means to generate the genomic resources necessary for producing a novel tailored probe set. Thus, assessing attributes associated with a UCE probe set's utility for providing robust data for study will aid both in determining the need for a novel probe design based on a focal clade and existing probe design, while optimizing odds of in vitro success a priori when a pre-existing probe set must be selected.

In this study, we investigate the utility of different coleopteran probe sets (Fig. 1) for recovering phylogenetically informative UCE data based on several features including (i) phylogenetic distance from taxa used in probe design, (ii) breadth of phylogenetic coverage probe set was designed for, and (iii) use of optimized probe design features like relaxed initial locus identification, and inclusion of both tailored and generalized probes. Our goal is to use our findings to provide recommendations for selecting an appropriate existing probe set or identifying if probe design is likely needed. We also set forth to assess previous findings regarding optimization of probe design, to provide recommendations for best practices to produce a tailored probe set. The results of this work have led to the production of 2 UCE probe sets for the beetle family Scarabaeidae and superfamily Hydrophiloidea.

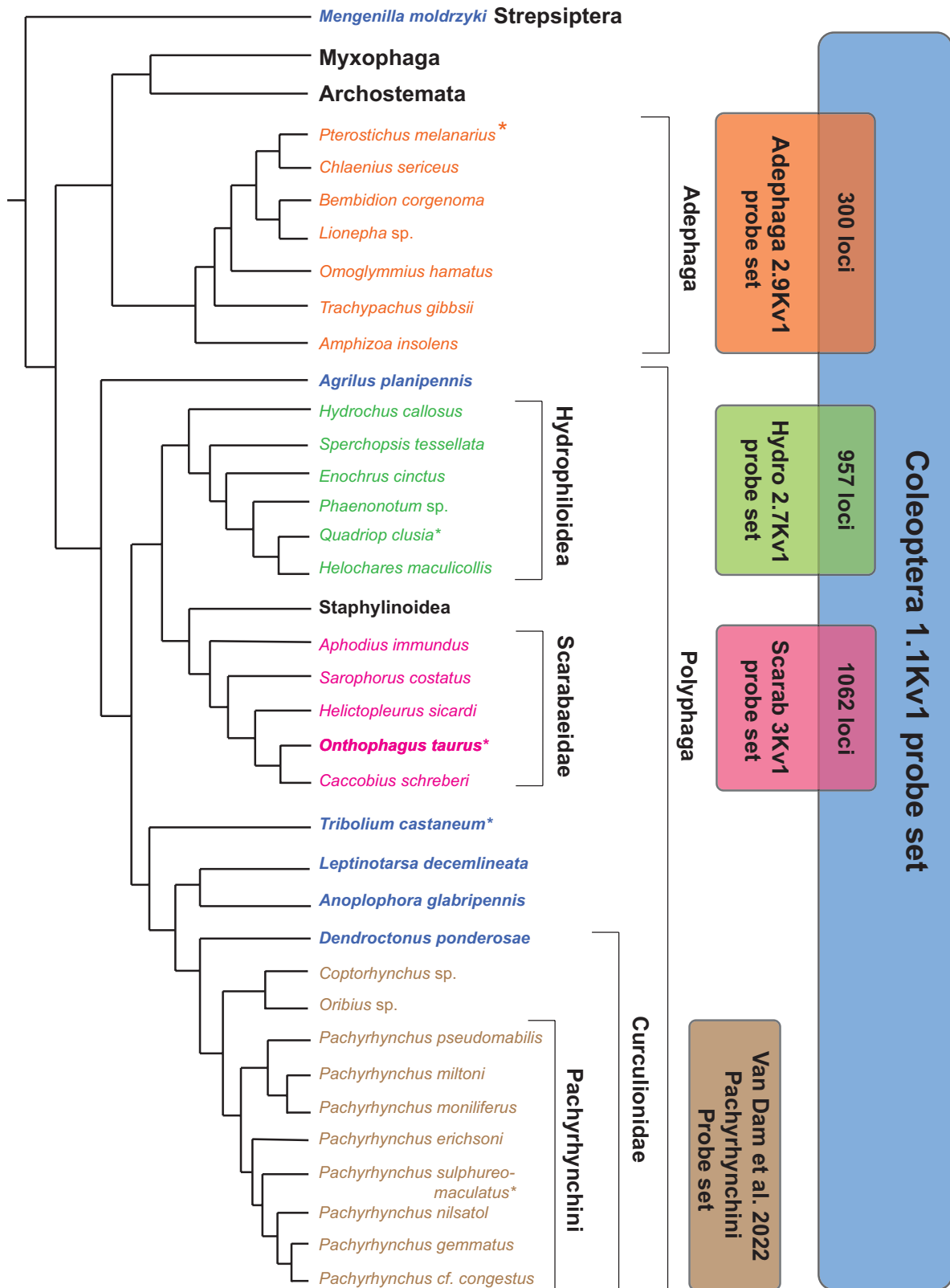
## Materials and Methods

### Study Overview

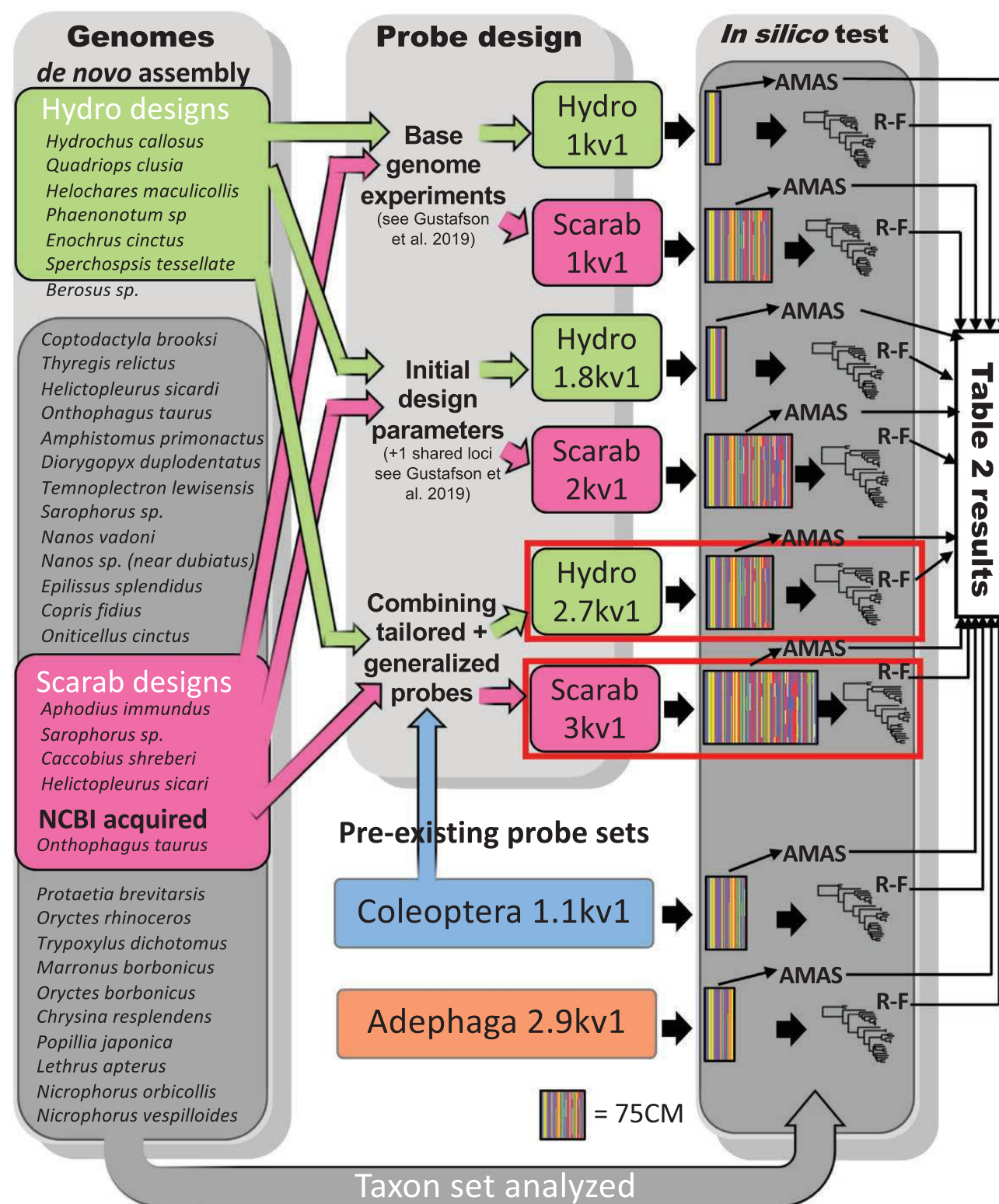
Our focal taxa are members of Hydrophiloidea and Scarabaeoidea within the beetle suborder Polyphaga (Fig. 1). These 2 lineages have consistently been placed in a clade together along with Staphylinoidea, with recent support for Scarabaeoidea sister to Staphylinoidea (McKenna et al. 2015, 2019, Zhang et al. 2018). Figure 2 provides a visualization of our study's workflow to accompany this text. To begin, we generated de novo genomic assemblies for Hydrophiloidea and Scarabaeidae to be used in UCE probe design (Fig. 2). Next during probe design, we experimented with the use of different base genomes (i.e., the base genome experiments of Gustafson et al. 2019), the use of different initial design parameters (see Gustafson et al. 2019), and finally combining tailored probes with pre-existing probes to assess the importance of these 3 factors in probe set design. This resulted in 6 different novel probe sets (Fig. 2): 1 optimized for base genome, another for initial design parameters, and a final probe set for both Hydrophiloidea and Scarabaeidae that combines fully optimized, tailored probes and generalized probes taken from the Coleoptera 1.1kv1. Finally, to test the utility of these different probe sets and the publicly available Coleoptera 1.1kv1 and Adephaga 2.9kv1 probe set (Figs. 1–2), we conducted in silico tests using scarabaeoid genomes available through the NCBI database or generated by this study (see Table 1, Fig. 2, and Supplementary Tables S1 and S22). Below we provide detailed methods for each step.

### Novel Genomic Resource Generation

A diverse sampling of taxa from across the Hydrophilidae, representing 5 of 6 currently recognized subfamilies and 1 member from



**Fig. 1.** Phylogenetic relationships based on McKenna et al. (2019) among select Coleoptera taxa relevant to or included in UCE probe design. Color (online), brackets, and fonts of species' names indicates inclusion in a particular probe design: orange/Adephaga bracket, Adephaga 2.9kv1 (Gustafson et al. 2020); green/Hydrophiloidea bracket, Hydro 2.7kv (current study); magenta/Scarabaeidae bracket, Scarab 3kv1 (current study); brown/Pachyrhynchini bracket plus *Coptorhynchus* sp. and *Oribius* sp., Pachyrhynchini probe set (Van Dam et al. 2022b); remaining species and those in blue and/or bold font, Coleoptera 1.1kv1 probe design (Faircloth 2017). Asterisk indicates base genome used in final probe design. Boxes associated with the probe set denote the taxonomic breadth probe set was designed to work across. Boxes' overlap and associated number indicate existence and quantity of loci in common.



**Fig. 2.** Schematic overview of the workflow for the present study. Workflow proceeds from left to right and top to bottom. The tablets correspond to the three broader segments of the study: genomic resource generation, probe design, and in silico testing; within tablet boundaries can be found associated taxon sets, data, experimentation, and results. Arrows indicate the flow of data, associated results, and the location results can ultimately be found. Color corresponds to membership within a taxon set or probe set. The red boxes around Hydro 2.7kv1 and Scarab 3kv1 denote final optimized, tailored probe set design tailored for Hydrophiloidae and Scarabaeidae based on the results of this study. Length of 75CM box corresponds to alignment length. Abbreviations used: NCBI, National Center for Biotechnology Information; 75CM, 75% complete matrix; AMAS, alignment manipulation and summary statistics (Borowiec 2016); R-F, Robinson-Foulds distance (Robinson and Foulds 1981).

**Table 1.** Novel genomic resources generated, their corresponding assembly metrics, and involvement in probe design

| Family        | Subfamily      | Tribe                | Taxon                                    | BioSample    | genome size | N50     | N90    | L50    | L90     | BUSCO %S | BUSCO%F | BUSCO%M | Probe de-sign |
|---------------|----------------|----------------------|------------------------------------------|--------------|-------------|---------|--------|--------|---------|----------|---------|---------|---------------|
| Hydrochidae   |                |                      | <i>Hydrochus callosus</i>                | SAMN35529347 | 228.812 Mb  | 18,680  | 5,503  | 2,041  | 7,686   | 88.4     | 7.4     | 2.8     | yes           |
| Hydrophilidae | Acidocer-inae  |                      | <i>Quadriops clusia</i>                  | SAMN35529348 | 271.788 Mb  | 7,580   | 3,549  | 10,497 | 33,032  | 64       | 10.0    | 1.9     | yes           |
| Hydrophilidae | Acidocer-inae  |                      | <i>Helochares maculicollis</i>           | SAMN35529349 | 531.846 Mb  | 12,248  | 4,601  | 8,085  | 26,073  | 80.8     | 13.4    | 4.0     | yes           |
| Hydrophilidae | Sphaeridi-inae | Coelostomat-ini      | <i>Phaeno-notum</i> sp.                  | SAMN35529350 | 876.463 Mb  | 4,726   | 3,274  | 18,463 | 43,521  | 46.8     | 40.3    | 12.1    | yes           |
| Hydrophilidae | Enochr-inae    |                      | <i>Enochrus cinctus</i>                  | SAMN35529351 | 611.295 Mb  | 5,714   | 3,394  | 12,790 | 33,766  | 70.1     | 17.2    | 11.1    | yes           |
| Hydrophilidae | Hydrophil-inae | Berosini             | <i>Berosus</i> sp.                       | SAMN35529352 | –           | 4,008   | 3,160  | 15,682 | 34,107  | 24.5     | 42.1    | 33.0    | no            |
| Hydrophilidae | Hydrophil-inae | Hydrobius-ini        | <i>Sperchopsis tessellata</i>            | SAMN35529353 | 1.279 Gb    | 7,997   | 3,834  | 36,831 | 104,105 | 60.9     | 26.3    | 11.6    | yes           |
| Scarabaeidae  | Aphodi-inae    |                      | <i>Aphodius immundus</i>                 | SAMN35529361 | –           | 3,997   | 3,143  | 455    | 1,022   | 33.9     | 37.1    | 28.7    | yes           |
| Scarabaeidae  | Scarabae-inae  | basal scarabaeinae   | <i>Sarophorus costatus</i>               | SAMN35529365 | –           | 4,073   | 3,154  | 1,861  | 4,167   | 43.9     | 31.7    | 24.4    | yes           |
| Scarabaeidae  | Scarabae-inae  | Australasian lineage | <i>Coptodact-yla brooksi</i>             | SAMN35529354 | –           | 7,149   | 3,662  | 15,912 | 44,778  | 60.6     | 23.2    | 14.4    | no            |
| Scarabaeidae  | Scarabae-inae  | Australasian lineage | <i>Thyregis relictus</i>                 | SAMN35529355 | –           | 3,850   | 3,124  | 3,763  | 8,078   | 9.6      | 34.1    | 56.3    | no            |
| Scarabaeidae  | Scarabae-inae  | Oniticellini         | <i>Helictopleu-rus sicardi</i>           | SAMN35529369 | –           | 9,856   | 3,513  | 168    | 813     | 6.3      | 159.0   | 77.4    | yes           |
| Scarabaeidae  | Scarabae-inae  | Onthophag-ini        | <i>Caccobius schreberi</i>               | SAMN35529359 | –           | 3,542   | 3,082  | 479    | 983     | 11.8     | 26.4    | 61.8    | yes           |
| Scarabaeidae  | Scarabae-inae  | Onthophag-ini        | <i>Onthopha-gus taurus de novo</i>       | SAMN35529360 | –           | 3,490   | 3,089  | 130    | 265     | 10.7     | 25.4    | 63.9    | no            |
| Scarabaeidae  | Scarabae-inae  | Onthophag-ini        | <i>Onthopha-gus taurus</i> NCBI          | SAMN02628949 | –           | 346,038 | 43,217 | 154    | 1,001   | 95.4     | 0.5     | 1.3     | yes           |
| Scarabaeidae  | Scarabae-inae  | Australasian lineage | <i>Amphisto-mus primonactus</i>          | SAMN35529356 | –           | 3,816   | 3,124  | 3,497  | 7,489   | 5.8      | 17.8    | 76.4    | no            |
| Scarabaeidae  | Scarabae-inae  | Australasian lineage | <i>Diorygopyx duploden-tatus</i>         | SAMN35529357 | 582.669 Mb  | 7,253   | 3,688  | 12,692 | 36,947  | 66.9     | 20.2    | 11      | no            |
| Scarabaeidae  | Scarabae-inae  | Australasian lineage | <i>Temnoplect-ron lewisensis</i>         | SAMN35529358 | –           | 4,476   | 3,170  | 131    | 333     | 2.4      | 4.8     | 92.8    | no            |
| Scarabaeidae  | Scarabae-inae  | basal scarabaeinae   | <i>Sarophorus costatus</i>               | SAMN35529364 | –           | 4,081   | 3,160  | 1,997  | 4,457   | 43.9     | 33.0    | 22.8    | no            |
| Scarabaeidae  | Scarabae-inae  | Madagascan lineage 1 | <i>Nanos</i> sp. (near <i>dubiatu</i> s) | SAMN35529367 | –           | 3,296   | 3,023  | 3      | 6       | 0.7      | 1.5     | 97.8    | no            |
| Scarabaeidae  | Scarabae-inae  | Madagascan lineage 1 | <i>Nanos vandoni</i>                     | SAMN35529368 | –           | 3,666   | 3,004  | 1      | 2       | 2.2      | 9.5     | 88.3    | no            |
| Scarabaeidae  | Scarabae-inae  | Madagascan lineage 2 | <i>Epilissus splendidus</i>              | SAMN35529366 | –           | 4,358   | 3,171  | 110    | 275     | 1.9      | 3.6     | 94.5    | no            |
| Scarabaeidae  | Scarabae-inae  | Coprini              | <i>Copris fidiu</i> s                    | SAMN35529363 | –           | 3,277   | 3,066  | 16     | 31      | 6.7      | 24.6    | 68.6    | no            |
| Scarabaeidae  | Scarabae-inae  | Oniticellini         | <i>Euoniticellu</i> s <i>fulvus</i>      | SAMN35529362 | –           | 3,311   | 3,049  | 60     | 119     | 7.5      | 24.2    | 68.3    | no            |
| Scarabaeidae  | Scarabae-inae  | Oniticellini         | <i>Helictopleu-rus sicardi</i>           | SAMN35529370 | –           | 6,536   | 3,381  | 378    | 1,133   | 5.7      | 5.8     | 87.8    | no            |



the closely related hydrophiloid family Hydrochidae (Short and Fikáček 2013), were chosen for genomic sequencing and probe design (Table 1). For scarabs, sampling targeted the subfamily Scarabaeinae (13 genera from 7 lineages that are equivalent to tribes) and 1 member of the closely related subfamily Aphodiinae (Table 1). All specimens were preserved in  $\geq 95\%$  ethanol and kept frozen at  $-20^{\circ}\text{C}$  prior to extraction. Total genomic DNA was extracted using a DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's protocols. Extractions were quantified via a Qubit (Life Technologies, Carlsbad, CA, USA) system to estimate input DNA for library preparation.

For hydrophiloid taxa, library preparation was done by technicians at RAPiD Genomics LLC (Gainesville, FL, USA) and the Genomic Sequencing Core at the University of Kansas (KU). There, DNA was mechanically sheared to an average size of 400 base pairs (bp), with libraries constructed by repairing the ends of the sheared fragments followed by the ligation of an adenine residue to the 3'-end of the blunt-end fragments. Then barcoded adapters suited for the Illumina Sequencing platform were ligated to the libraries, with ligated fragments being PCR-amplified using standard cycling protocols (e.g., Mamanova et al., 2010). Genomic sequencing for *Quadriops clusia* Girón & Short (Coleoptera: Hydrophilidae), *Berosus* sp. Leach, and *Hydrochus callosus* LeConte (Coleoptera: Hydrochidae) was done on an Illumina HiSeq 3000 with paired-end 150 bp reads. Genomic sequencing for *Helochares maculicollis* Mulsant (Coleoptera: Hydrophilidae), *Phaenotum* sp. Sharp, *Sperchopsis tessallata* (Ziegler), and *Enochrus cinctus* (Say) was done on an Illumina NextSeq 550 with paired-end 150 bp reads. Demultiplexing was also done by techs at RAPiD Genomics LLC and the KU Genomic Sequencing Core. For the scarab taxa used in probe design (non-Australasian lineages, Table 1), a Nextera Flex library (Illumina) was sequenced using an Illumina NextSeq 500 sequencer. For Australasian dung beetles, genomic DNA was sent to SNPsaurus (Eugene, OR) for shotgun sequencing. Following their standard workflow, a Nextera tagmentation kit was used for library generation, samples were subsequently sequenced on an Illumina HiSeq 4000 sequencer with paired-end 150 bp reads, and adapters trimmed with BBDuk (BBMap version 38.41).

Raw demultiplexed Illumina reads were first cleaned using the TRIMMOMATIC (Bolger et al. 2014) with final adapter contamination trimmed using CUTADAPT 1.18 (Martin 2011) or FASTP (Chen et al. 2018). Raw, cleaned and trimmed reads were quality inspected using FASTQC (Andrews 2010). For hydrophiloid taxa the best  $k$ -mer length for de novo genome assembly was inferred using KMERGENIE (Chikhi and Medvedev 2014). This program was also used to estimate genome size using  $k$ -mers from assemblies that were found to be  $\geq 50\%$  complete (Table 1). Unless otherwise indicated, genome assembly was done using SPADes 3.7.1 (Bankevich et al. 2012) under the default  $k$ -mer settings for 150 bp paired-end reads including the  $-m$  1,000 memory option. For *Quadriops clusia*, *Hydrochus callosus*, and *Helochares maculicollis*, which were identified as having the best  $k$ -mer assembly size above 77, the  $-k$  option was used with the  $k$ -mer values of 21, 33, 55, 77, 99, 127. For scarab taxa used in probe design (non-Australasian lineages, Table 1) SPARSEASSEMBLER (Ye et al. 2012) was used to assemble low-coverage genomes using forward and reverse reads with the command "SparseAssembler g 10 k 31 LD 0 GS 1000000000 NodeCovTh 1 EdgeCovTh 0 1" as suggested earlier (Kypke 2018, Brunke et al. 2021).

Genomic assemblies were first evaluated using QUAST 5.0.2 and 5.2.0 (Mikheenko et al. 2018) to get the following metrics:

total contig number, total assembly length, coverage, GC content, N50, N90, L50, L90 (Table 1, Supplementary Table S1). Next BUSCO 3.0.2 (Simão et al. 2015) was used to measure assembly completeness. BUSCO was run using *Tribolium castaneum* as the AUGUSTUS (Stanke et al. 2004) input species and the 2,442 Endopterygota single copy orthologous protein coding reference genes from OrthoDB v9 (Zdobnov et al. 2017). This provided each genome assembly with BUSCO completeness scores (Table 1). Full genome assembly stats are provided in Supplementary Table S1 in Supplemental Materials.

### Probe Design: Hydrophiloidea

The hydrophiloid probe design follows the methods described in Gustafson et al. (2019) for the identification of a base genome and optimization of the probe set. Optimal base genome identification can be done either by conducting base genome trials or by selecting the taxon with the shortest overall genetic distance from all other taxa (see Gustafson et al. (2019) for more details). A major goal of this design, other than to produce a hydrophiloid probe set, was to test the findings of Gustafson et al. (2019) regarding probe set design optimization.

First, we gathered sequence data for six molecular markers commonly used in phylogenetic studies of the Hydrophilidae (Supplementary Table S2) (Short and Fikáček 2013) for each genus used in probe design or a sister taxon based on the phylogeny from Short and Fikáček (2013). These loci were generated via Sanger sequencing and are independent of the genomic sequencing conducted here. Additionally, we randomly selected 10 of the genes successfully extracted directly from each genomic assembly that were used to estimate BUSCO completeness scores to generate additional genetic distance measures for comparative purposes.

For these 2 sets of markers genetic distances among taxa were estimated using 4 different models: Jukes Cantor (Jukes and Cantor 1969), Tajima Nei (Tajima and Nei 1984), Tamura Nei (Tamura and Nei 1993), and the maximum composite likelihood model (Tamura et al. 2004), as implemented in MEGA ver. 6.06 (Tamura et al. 2013). From the resulting genetic distance measures taxa were ranked from smallest to largest for each model. These rank scores were then averaged across models for each marker. From the average scores for each marker the overall genetic distances ranks were established for each taxon within the 2 sets of markers (i.e., the six Sanger loci and the 10 BUSCO genes).

Both the Sanger and BUSCO genetic markers identified *Quadriops* as having the shortest average overall genetic distance from all other taxa used in probe design. Therefore, *Quadriops* was selected as the initial base genome for UCE probe design. Design was conducted in PHYLUCe ver. 1.7.1 following the methods outlined in detail in Faircloth (2017) and the generalized workflow discussed in Gustafson et al. (2019). Specific probe design commands (PDC) and settings are available in the Supplemental materials. Our temporary baits were initially designed to target putative loci shared among the base taxon and all 5 other taxa (PDC1). Following alignment of the temporary baits back to the genomic assemblies (PDC2), the final probe set was designed to target only loci present in all six taxa (PDC3).

After the initial probe design described above was completed, monolithic fasta files were produced using the same methods with each hydrophilid taxon serving as the base genome during probe design. This was done (i) as a base genome test to confirm previous findings that base genome choice significantly alters locus recovery and further test the ability of genetic distance measures alone for selecting an optimal base genome; (ii) apply a more stringent paralogy filter to our final probe set using a pipeline consisting of BLAST

(Altschul et al. 1990) and R scripts (R Core Team 2022) for comparing monolithic fasta files (i.e., comparing\_monolithic\_UCE\_fasta v0.2 and whittle\_uce\_probes.R (Alexander 2018 and Gustafson et al. 2019); and (iii) receive further information about locus length and recovery in other taxa given a particular base genome using these R scripts (i.e., select\_base\_genome.R) (Alexander 2018). Because the *Berosus* genome was largely incomplete (Table 1), it was not included in probe design as another hydrophiline, *Sperchopsis*, was available. This first probe set design represents optimization for base genome.

To test the effect of initial design parameters on final probe design, 1 additional de novo probe design was conducted using the optimal base genome. For this design, putative locus identification for the development of temporary baits was done using loci shared between the base genome and just 1 other taxon (i.e., +1), rather than all other taxa (PDC4). Probe design then continued identically to that described above (PDC5–6), with the exception of swapping base genomes to produce multiple probe sets. Note the paralogy filter was still applied using the same monolithic fastas from the prior probe design but with the new optimal base genome +1 monolithic fasta file replacing the corresponding +6 fasta. This second probe set design represents optimization for initial design parameters.

In addition to the novel UCE loci identified, we wanted to include probes from the Coleoptera 1.1kv1 probe set designed by Faircloth (2017) to produce the final probe set. This will afford cross-compatibility and it has been demonstrated that combining tailored probes alongside the generalized probes of the Coleoptera 1.1kv1 can actually improve phylogenetic analysis (Gustafson et al. 2020). When combining probe sets designed in PHYLUCE there are a couple of considerations: (i) there could be homologous loci identified in common between them; (ii) the probe naming scheme can lead to homonymous probes that target non-homologous loci (i.e., both probe sets may target a UCE locus 300, but “locus 300” may not be homologous between the 2). To combine the probe sets, given these considerations, we generated novel R scripts (available from <https://github.com/sergeitarasov/Rename-and-Merge-UCES>) that first BLASTs between monolithic fasta files produced by the 2 probe sets on focal design taxa to identify homologous loci. Next, probe naming conventions are compared such that new probes targeting homologous loci in common with the original probe set (in this case, Coleoptera 1.1kv1) are renamed to match the original probe set (so probes in both sets target the same locus 300). Finally, the 2 sets of probes are merged such that duplicate probes from the original probe set are removed to prevent overlap; and homonymous probe names between the 2, now target correspondingly homologous UCE loci.

### Probe Design: Scarabaeidae

For the Scarabaeidae probe design, we followed all the methodology above with the exception of performing genetic distance measures, as we encountered difficulties finding 10 BUSCO genes in common across all scarab genomes used in probe design, likely due to their relatively low completeness (Table 1). Additionally, we ultimately wanted the scarab probe design to utilize the available NCBI genome assembly of *Onthophagus taurus* (Schreber) (Coleoptera: Scarabaeidae). There were several reasons for this: (i) the assembly is highly complete (Table 1); (ii) its annotation will allow all loci targeted by the resultant probe design to be identifiable, thus preventing inclusion of “off-target” probes (see Van Dam et al. 2022a); and (iii) it performed relatively well as the base genome (Supplementary Table S19).

### UCE Locus Identification

We used the scripts provided by Van Dam et al. (2021) to identify loci targeted by our probe sets. These scripts require an annotated genome to function. No annotated hydrophiloid genome is currently available, thus we only utilized the *Onthophagus taurus* annotation. To identify hydrophiloid UCE loci using this annotation, we produced *Onthophagus*-based probes by aligning the final Hydrophiloid probe set against the *Onthophagus taurus* genome as “temporary baits” (PDC7), then based on which probes hit (PDC8–9), we produced finalized *Onthophagus*-based hydrophiloid probes using the same methods described above (PDC10).

### In Silico Testing

For assessing optimality of de novo probe design, i.e., base genome experimentation and relaxed initial design parameters, via PHYLUCE we compared the number of loci extracted by the finalized probe set when applied only to genomes used in the corresponding probe design.

To test the broader phylogenetic utility of the optimized-, final-, and existing UCE probe sets, we downloaded scarabaeoid genomes available through the NCBI website (Fig. 2 and Supplementary Table S22). Two Staphylinoid taxa were included as an outgroup. Scarabaeoid taxa were selected because (i) an abundance of genomic resources relative and (ii) they afford a range of phylogenetic distances from the Coleoptera UCE probe sets (Fig. 1). Specifically, Scarabaeidae is nested within Scarabaeoidea, which in turn together with Staphylinoida is sister to Hydrophiloidea, and the aforementioned clade is nested inside the suborder Polyphaga, which is sister to the suborder Adephaga. In other words, these relationships range in varying degrees of “off-targetedness” from low (the Scarab probe designs) to high (Adephaga probe design) simulating decision making in the absence of a fully tailored probe set (i.e., one whose design includes Scarabaeoidea + Staphylinoida taxa – the ingroup and outgroup respectively, Fig. 1). Through the PHYLUCE 1.7.1 pipeline, first probes were aligned to the genomic data and the surrounding sequence information extracted using the command `phyluce_probe_slice_sequence_from_genomes` with flank value set to 500. Next UCE loci were pulled from these contigs using each UCE probe set with the command `phyluce_assembly_match_contigs_to_probes` and the minimum identity set at 80 and the minimum coverage at 50. Extracted UCE loci were aligned using MAFFT (Katoh and Standley 2013) under the default PHYLUCE wrapper settings, but with the additional command “no-trim” to provide internal trimming. Internal trimming is recommended for analysis of divergences over 50 million years ago (MYA) (Faircloth 2016), and most splits among our taxa meet this requirement (i.e., *Onthophagus* + *Aphodius* Hellwig split was estimated ca. 100 MYA in McKenna et al. 2019). Next, alignments were trimmed with the GBLOCKS (Castresana 2000, Talavera and Castresana 2007) PHYLUCE wrapper and the commands `b1 0.5`, `b2 0.85`, `b3 8`, `b4 10`. Finally, an unpartitioned 75% complete concatenated data matrix was produced and the program AMAS (Borowiec 2016) was used to generate summary statistics associated with each data set.

Maximum likelihood analyses were conducted using IQ-TREE ver. 1.6.12 (Nguyen et al. 2015) and the MFP option (Kalyaanamoorthy et al. 2017) to select the optimal evolutionary model using corrected Akaike information criterion scores, and apply it for phylogenetic inference. Nodal support was assessed with 1000 ultrafast bootstraps (Hoang et al. 2018). As relationships among Scarabaeoidea remain unclear, we generated Robinson–Foulds (R–F) distances (Robinson and Foulds 1981) in RAXML ver. 8.2.12 (Stamatakis 2014) to quantitatively assess differences among the resultant phylogenetic trees.



## Results

### Base Genome Identification and Base Genome Optimized Probe Sets

The overall average genetic distance measures using both common Sanger gene fragments for phylogenetics (Fig. 3A, Supplementary Tables S2–S8) and 10 randomly selected BUSCO genes (Fig. 3B, Supplementary Tables S2, S9–S18) both identified *Quadriops clusia* as having the shortest distance from all other taxa used in probe design, implying its identification as the optimal base genome. The Hydrophiloid base genome tests unambiguously confirmed *Quadriops clusia* as being the optimal base genome for probe design in that it resulted in: the highest number of total UCE loci recovered (Fig. 3A and B), the highest number of loci recovered in individual taxa (Supplementary Table S19), the highest number of paralogy-free loci across all taxa when comparing monolithic fasta files, and the longest locus lengths. Relative to the second best performing base genome, *Helochares*, use of *Quadriops* as the base genome resulted in a probe set that recovers several thousand more loci, representing a 66% increase in total locus recovery (Supplementary Table S1). We found base genome performance, as assessed by total number of UCE loci recovered during in silico testing, to be negatively correlated with average genetic distance rank (Fig. 3A and B). Genome assembly metrics (Table 1) like N50 value (Fig. 3C) and BUSCOs (Fig. 3D), showed little-to-no correlation with base genome

performance. Given the above findings, *Quadriops clusia* was selected as the base genome to produce the hydrophiloid base genome optimized probe set, which after paralogy filtering targets 1,153 loci using 13,636 probes. This probe set will hereafter be referred to as the Hydro 1kv1 (Fig. 2).

The Scarabaeidae base genome trials similarly revealed different base genomes resulted in considerable differences in the number of loci recovered (Supplementary Table S19). The highest total loci recovered was 12% higher than the second highest total using a different genome. The *Onthophagus taurus* genome (Table 1) produced the third highest total recovered loci out of the 5 base genomes. For the reasons discussed above, the NCBI *Onthophagus taurus* genome was selected to serve as the base genome to produce the first Scarabaeid probe set, which after paralogy filtering targets 1,521 loci using 15,139 probes. This probe set will hereafter be referred to as the Scarab 1kv1 (Fig. 2).

### Optimization of Initial Design Parameters

Changing the number of taxa the base genome must share a locus with for identification of putative loci from all to just 1, dramatically increased the number of recovered loci during in silico testing. For the Hydrophiloid probe design, optimizing initial design parameters in this way significantly increased locus recovery ( $P = 4.794e^{-5}$  paired, two-tailed  $t$ -test on number of loci recovered in

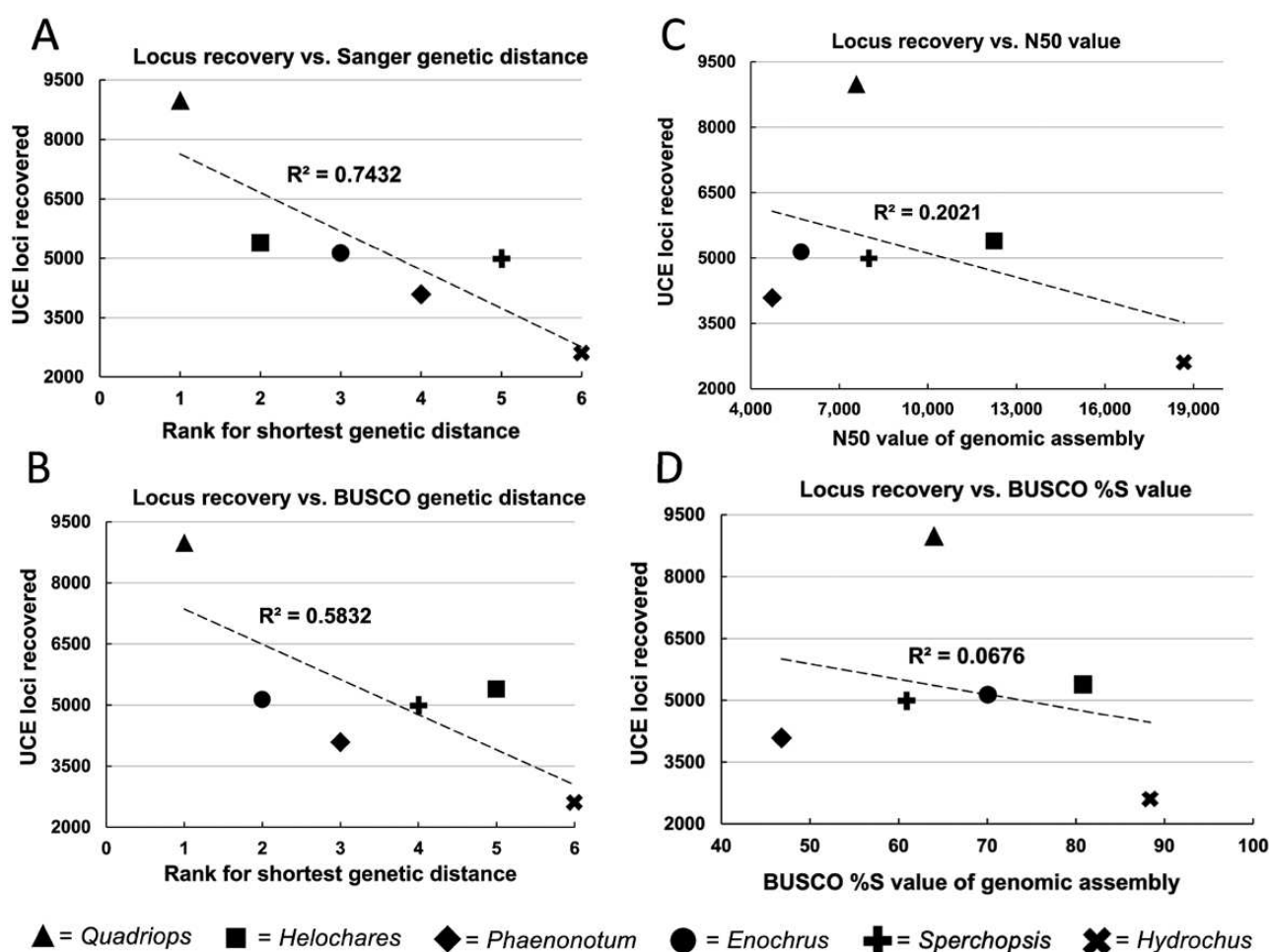


Fig. 3. UCE loci recovery during in silico testing plotted against different metrics: A) average genetic distance estimated based on common gene fragments used in phylogenetics; B) average genetic distance estimated using BUSCO genes; C) N50 assembly metrics; and D) BUSCO S values.

each taxon) and resulted in 85% more total loci recovered. Using both the optimal base genome of *Quadriops clusia* and initial design parameters of +1 for putative locus identification, resulted in a probe set which after paralogy filtering targets 1,886 loci using 22,341 probes. This probe set will hereafter be referred to as the Hydro 1.8kv1 (Fig. 2).

The Scarabaeidae probe design showed similar results where optimizing initial design parameters also significantly increased locus recovery ( $P = 2.764 \times 10^{-6}$  paired, two-tailed  $t$ -test on number of loci recovered in each taxon), but with a far more substantial increase in the total number of loci recovered at 205% (tripling the total). With *Onthophagus taurus* as the base genome and initial design parameters set to +1 for putative locus identification, after paralogy filtering this probe set targeted 2,343 loci using 23,278 probes. However, 3 loci could not be identified within the *Onthophagus taurus* genome and were removed, leaving 2,340 loci and 23,260 probes. This probe set is dubbed the Scarab 2kv1 (Fig. 2).

### Combining Probe Sets

When the entire Coleoptera 1.1kv1 probe set was applied to our hydrophiloid taxa in silico, 957 loci could be recovered. We thus whittled the Coleoptera 1.1kv1 probes down to only these loci using the R script `whittle_uce_probes.R` (Alexander 2018), prior to merger with the fully optimized Hydro 1.8kv1 probe set. Of the 1,886 loci identified through the de novo Hydro 1.8kv1 probe design, only 82 were found in common with the 957 loci from the Coleoptera 1.1kv1 during merger. The final Hydrophiloid probe set, Hydro

2.7kv1, targets 2,761 loci (1,804 tailored, 957 generalized) using 32,657 probes (Fig. 2).

*Onthophagus taurus* was included in the Coleoptera 1.1kv1 probe set design (Fig. 1), and as such has probes tailored to it. This design also utilized a strepsipteran, *Mengenilla moldrzyki*, as an outgroup for highly conserved locus identification (Fig. 1) (Faircloth 2017). Therefore, for the final Scarabaeidae probe set, using PHYLUC we limited the Coleoptera 1.1kv1 probes to only those designed with *Onthophagus* and *Mengenilla* prior to merger. Of the 2,340 loci identified through the de novo Scarab 2kv1 probe design, 228 loci were in common with the 1,062 loci of the Coleoptera 1.1kv1 design. The final Scarabaeid probe set, Scarab 3kv1, targets 3,174 loci (2,112 tailored, 1,062 generalized) using 25,786 probes (Fig. 2).

### Identification of Loci

Use of *Onthophagus* as the base genome for the Scarab probe sets allowed all loci to be identified (Fig. 4, Supplementary Table S21). We were also successful in identifying slightly over half of the loci associated with the final hydrophiloid probe set (Fig. 4, Supplementary Table S21). Comparison of the proportions of different types of loci shows the majority are exonic, followed by intronic, then intergenic, with few being both intronic and exonic (Fig. 4). Of the identifiable loci in the Hydro 2.7kv1 probe set, the same trends are evident (Fig. 4). Optimization of initial design parameters increased the proportion of exonic loci, as did merging the probe set tailored probe set with the existing generalized probe set (Fig. 4).

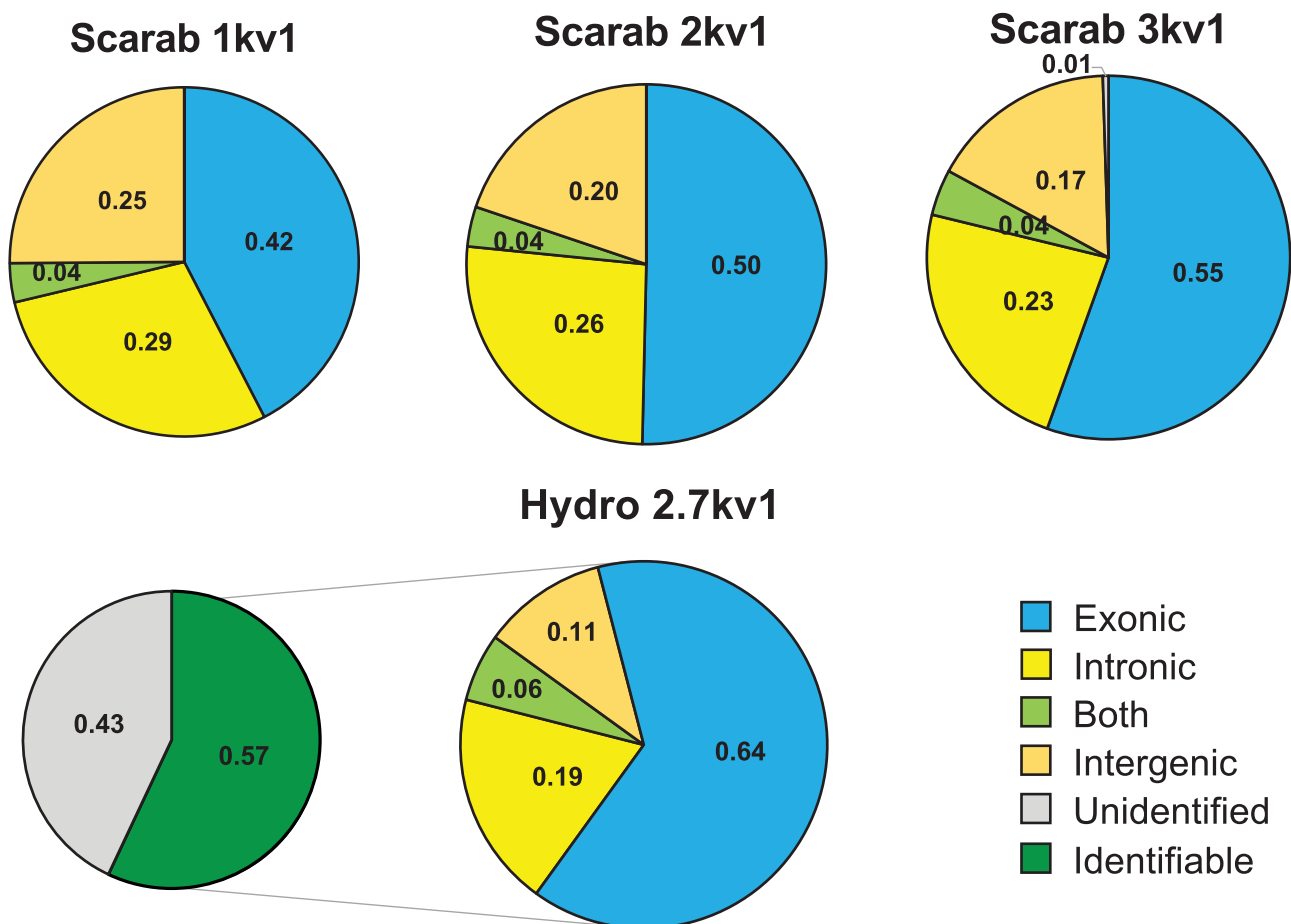


Fig. 4. The proportion of different types of loci targeted by scarab and hydrophiloid probe sets.

## Comparison of Phylogenetic Utility of the Probe Sets

The probe set used had major impacts on aspects of the 75% complete data matrix (hereafter 75CM) assembled (Table 2, Fig. 2). Use of any of the Scarab probe sets, which are phylogenetically closest to the scarabaeoid taxa (Fig. 1), resulted in the largest alignments with the highest proportions of both variable- and parsimony informative sites. Phylogenetic analysis of the 75CM produced with the Scarab 3kv1 probe set resulted in a fully resolved tree whose topology was almost completely maximally supported (Fig. 5A, Supplementary Fig. S1). The other scarab probe sets produced trees most similar to that shown in Fig. 5A, Supplementary Figs. S2 and S3.

Among Hydrophiloidea probe sets (Table 2), representing the next closest clade phylogenetically to the Scarabaeoidea (Fig. 1), the Hydro 2.7kv1 produced a 75CM comparable to that of the Scarab 1kv1 and an identical phylogenetic tree (Supplementary Figs. S3 and S4). On the contrary, the Hydro 1- and 1.8kv1 probe sets consisting solely of probes tailored for Hydrophiloidea, produced the shortest alignments with relatively low proportions of parsimony informative sites. The Hydro 1.8kv1 (Supplementary Fig. S5) produced trees closer to the scarab probe sets than others, but Hydro 1kv1 produced a tree (Supplementary Fig. S6) 10 R-F distance points away from both the Hydro 2.7 and Scarab 1kv1 (Table 2).

The Coleoptera 1.1kv1 was developed as a generalized or “universal” probe set for Coleoptera, however, genomes used in locus identification for probe design consisted entirely of members of Polyphaga, the suborder of beetles Scarabaeoidea belongs to (Fig. 1). Therefore, this probe set represents the penultimate in terms of phylogenetic distance. The 75CM produced (Table 2) was larger in alignment length than both tailored hydrophilid probe sets but had the overall lowest proportions of variable- and parsimony informative sites. The phylogenetic tree produced (Supplementary Fig. S7) was equidistant with the Hydro 1kv1 from the Scarab 3kv1.

The probe set most phylogenetically distant from Scarabaeoidea is the Adephaga 2.9kv1, which was tailored for use in an entirely different suborder of beetles (Fig. 1). This probe set produced the most different tree (Fig. 5B, Supplementary Fig. S7) as indicated by the largest R-F distance from all other trees (Table 2). The 75CM produced an alignment shorter than the generalized Coleoptera 1.1kv1, but longer than both tailored Hydro probe sets and with higher proportions of parsimony informative sites than all 3 (Table 2).

## Comparison of Resulting Phylogenetic Trees

In general, the 2 contrasting phylogenetic trees for Scarabaeidae generated from different probe sets (Fig. 5A and B), are in agreement with previous studies. A clade in which Dynastinae is nested within

Rutelinae emerges as sister to Cetoniinae, while Scarabaeinae is recovered as sister to Aphodiinae (Ahrens et al. 2014, McKenna et al. 2019). The scarabaeine genus *Sarophorus* Erichson is found to be the sister lineage to the rest of Scarabaeinae; Oniticeellini+Onthophagini and Australasian scarabaeines are recovered to be monophyletic; 2 Madagascan lineages, represented by the genera *Nanos* and *Epilissus*, come out as non-monophyletic (Gunter et al. 2016, Tarasov and Dimitrov 2016).

However, some differences were observed between the trees generated using different probe sets. Specifically, within Scarabaeidae, the position of *Lethrus apterus* (Geotrupidae) alternated between being at the base of all Scarabaeidae (Fig. 5B), or at the base of the Pleurostict scarabs (Fig. 5A). Recent studies have supported the former relationship, which appears more plausible (Ahrens et al. 2014, McKenna et al. 2019). Within Scarabaeinae, the position of *Nanos* in relation to *Copris* and *Epilissus*, and *Amphistomus* and *Diorygopyx* within the Australasian Endemic clade varied. Considering that the present analysis was limited to taxa with genomes, it should be noted that differences were generally to be expected. Consequently, we anticipate that these discrepancies will be resolved once the taxon sampling is expanded.

## Discussion

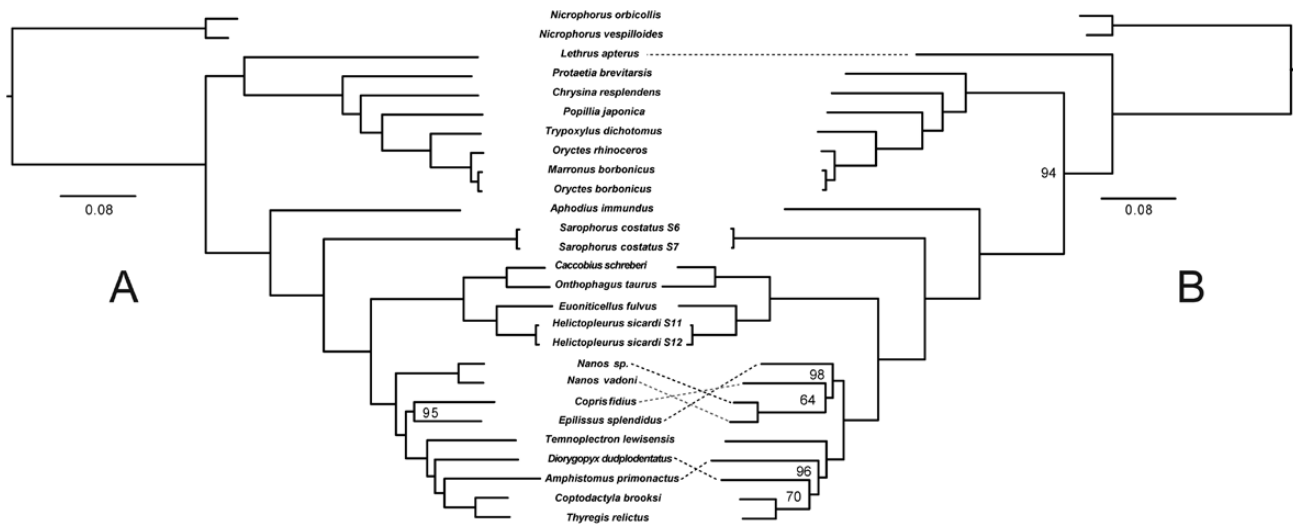
### Identification and Importance of Selecting an Optimal Base Genome

Our study confirms the previous findings of Gustafson et al. (2019) that base genome choice significantly alters probe set performance, based on in silico testing. Having an optimal base genome will result in the most loci recovered (Fig. 3A and B; Supplementary Table S19), and based on comparisons of monolithic fasta files, loci that are longer than those developed using a less optimal base genome. Both factors are likely important for maximizing phylogenetic utility, as one will assemble matrices with larger alignment sizes through the increased locus recovery, and the longer size of these loci likely afford more variable sites for analysis.

In regards to how researchers interested in developing their own UCE probe sets can identify an optimal base genome efficiently, our results support Gustafson et al.'s (2019) finding that base genome performance negatively correlates with overall genetic distance away from taxa used in probe design (Fig. 3A and B). If highly complete genome assemblies are available, genes extracted as part of BUSCO assessment can be used to estimate genetic distances, or gathering traditional Sanger loci used in phylogenetic studies provides similar results (Fig. 3A and B). Our results also corroborate that genome

**Table 2.** Results of the in silico test of different probe sets in terms of the 75% complete unpartitioned concatenated matrix and Robinson-Foulds distance of the resultant phylogenetic tree

| Probe design      | Alignment length (bp) | Proport. variable sites | Parsimony informative sites (bp) | Proport. parsimony informative sites | Sum of plain pairwise R-F distances |
|-------------------|-----------------------|-------------------------|----------------------------------|--------------------------------------|-------------------------------------|
| Adephaga 2.9kv1   | 54,189                | 0.421                   | 19,549                           | 0.361                                | 52                                  |
| Coleoptera 1.1kv1 | 67,806                | 0.384                   | 21,562                           | 0.318                                | 36                                  |
| Hydro 1kv1        | 25,798                | 0.417                   | 8,533                            | 0.331                                | 36                                  |
| Hydro 1.8kv1      | 39,584                | 0.431                   | 13,592                           | 0.343                                | 28                                  |
| Hydro 2.7kv1      | 157,675               | 0.422                   | 55,745                           | 0.354                                | 26                                  |
| Scarab 1kv1       | 148,870               | 0.455                   | 54,944                           | 0.369                                | 26                                  |
| Scarab 2kv1       | 214,531               | 0.465                   | 81,820                           | 0.381                                | 24                                  |
| Scarab 3kv1       | 267,822               | 0.450                   | 98,823                           | 0.369                                | 24                                  |



**Fig. 5.** Phylogenetic trees of Scarabaeidae generated using UCEs; node values indicate bootstrap support. A) The tree produced with the Scarab 3kv1 probe set. B) The topology produced with the Adephaga 2.9kv1 probe set.

assembly metrics do not correlate with base genome performance (Fig. 3C and D). Thus, genome assembly completeness alone, is not a good justification for selecting a taxon to serve as the base genome, as this does not guarantee optimal probe design.

### Relaxed Initial Locus Identification Provides More Exonic Loci

Decreasing the number of taxa candidate loci must be identified within to only 1 during initial probe design dramatically increased the total number of loci UCE loci recovered (Supplementary Table S20), another result consistent with Gustafson et al. (2019). Furthermore, after filtering paralogous loci, the additional loci targeted appear to be skewed towards exonic regions, based on the evident increase in the ratio of exonic loci between the Scarab 1kv1 and Scarab 2kv1 (Fig. 4). Investigators developing their own UCE probe sets wanting to increase the number of loci targeted, and those interested in maximizing exonic loci for analysis, should consider relaxed initial locus identification with candidate loci required to be shared between the base genome and just one other taxon. Under this scenario, stringent paralogy filtering is recommended, which can be accomplished using the scripts available from Alexander (2018).

### Combining Generalized and Tailored Probe Sets

Combining the generalized probes from the Coleoptera 1.1kv1 (Fig. 1) with tailored hydrophiloid and scarab probes improved phylogenetic utility of final probe sets via increased alignment length and higher proportion of parsimony informative sites in the 75CM (Table 2, Fig. 2). The addition of these probes also had the effect of decreasing the proportion of variable sites. The Coleoptera 1kv1 probe set was developed to target loci across highly divergent lineages of beetles (Fig. 1), and as a result, the loci and its probes pull contain decreased variability (Table 2). While this aspect may seem like a potential drawback, a mixture of highly conserved loci and more variable tailored loci has been shown to improve phylogenetic analysis, especially in regards to placement of phylogenetically distant taxa (Gustafson et al. 2020). Consistent with this, the Hydro 2.7kv1, which contains generalized probes, outperformed both tailored-only hydrophiloid probe sets when applied to the distantly related scarabaeoid taxa (Table 2, Fig. 2).

An additional important aspect of combining probe sets is affording cross compatibility between probe sets and phylogenetic studies utilizing them (Fig. 1). Thus, UCE enriched read data generated using the 2 novel probe sets developed here, will share targeted loci and be combinable with those from studies utilizing the Coleoptera 1.1kv1 like Kobayashi et al. (2021), Bradford et al. (2022), Sota et al. (2022), etc.; and to a lesser extent those using the Adephaga 2.9kv1 probe set (Fig. 1).

### Selecting an Existing Probe Set for UCE Phylogenomics

Predictably, as evolutionary distance of taxa used in probe design increased away from Scarabaeoidea (Fig. 1), phylogenetic utility of probe sets decreased, as evidenced in smaller alignment lengths, decreased proportions of variable- and parsimony informative sites, and lower numbers of parsimony informative sites of the 75CM (Table 2). Furthermore, trees produced using probe sets further away from Scarabaeoidea received lower support (Supplementary Figs. S1–S8) and had more strongly differing topologies (Table 2, Supplementary Figs. S1–S8). Notably, taxonomic depth of probe design also had a strong impact. The overall worst performing probe set (Table 2, Fig. 2) was not the Adephaga 2.9kv1, the furthest away phylogenetically (Fig. 1), but the Hydro 1kv1 probe set tailored for use in the hydrophiloid clade (Fig. 1). The reason probe sets containing only tailored hydrophiloid probes performed in the most limited fashion, likely has to do with locus identification being limited only to the focal clade Hydrophiloidea (Fig. 1). When identifying conserved regions for probe design, PHYLUCS requires sequence divergences between the base genome and remaining taxa be no greater than 5% (Branstetter et al. 2017a). Thus, loci identified during probe design with only Hydrophiloid taxa are more likely to be unique to this clade relative to a more broadly sampled probe design taxa where this limited sequence divergence must span greater phylogenetic diversity (Fig. 1). Therefore, these hydrophiloid clade-specific loci are unlikely to be present, or at least conserved enough, at deeper levels in the Coleoptera tree, including within Scarabaeoidea, for UCE probes to recover them. This also explains both why few of the Coleoptera 1.1kv1 loci were found in common between the tailored loci for both Hydrophiloidea and Scarabaeidae (see Combining



probe sets above): as limiting probe design to a focal clade identifies conserved loci unique to that lineage; and functionally how tailoring an *ultraconserved* element probe set can substantially improve locus recovery, relative to both more generalized probe sets and those tailored for other taxa (Branstetter et al. 2017a, Van Dam et al. 2019, Gustafson et al. 2020, Kulkarni et al. 2020).

Optimizing initial locus identification improved phylogenetic performance in the tailored-only Hydro 1.8kv1 (Table 2), especially regarding the tree produced, but the 75CM was still limited relative to the general Coleoptera 1.1kv1. This latter probe set was meant to serve as a generalized Coleoptera probe set (Fig. 1), and as such it provides conserved loci with the lowest proportions of variable sites, but ultimately functions better than probe sets tailored for specific clades outside of our focal Scarabaeoidea (Table 2). Notably, combining the generalized Coleoptera probes with the optimized hydrophiloid tailored probes in the Hydro 2.7kv1 expanded the probe set's phylogenetic utility to be comparable to an incompletely optimized scarab probe set like the Scarab 1kv1 (Table 2, Fig. 2).

Given the above, when selecting an existing probe set for a focal clade (i.e., in the absence of generating a tailored probe set), the primary concern is phylogenetic distance (Fig. 1). All scarab probe sets performed the best on scarabaeoid taxa during in silico testing (Table 2). However, care should be taken to ensure the probe set includes some probes developed with taxonomic breadth beyond a highly derived focal clade, as tailored-only probes for the sister clade outside of scarabaeoids (i.e., the Hydro 1kv1 and 1.8kv1, Fig. 1) performed comparable to- or worse than a generalized only probe set like the Coleoptera 1.1kv1 (Table 2). For example, using Fig. 1, researchers interested in selecting an existing probe set for use within Staphylinidea, the Scarab 3kv1 probe set is likely to perform the best given Scarabaeidae is sister to Staphylinidea and includes generalized probes. Because the Hydro 2.7kv1 also includes generalized probes, it would likely function better than picking the existing Coleoptera 1.1kv1 alone, but our results suggest not as well as the Scarab 3kv1. For researchers interested in selecting an existing UCE probe set for broader use within Curculionidae (Fig. 1), unless the study is focused on Entiminae, the choice is less clear. While the Pachyrhynchini probe set of Van Dam et al. (2022b) was developed specifically for a tribe of entimine weevils and thus is the phylogenetically closest probe set available, suggesting potential optimal performance, there is the possibility that its design for a highly derived clade could limit its utility in the absence of generalized probes. A possible solution would be to include all *Dendroctonus* probes from the Coleoptera 1.1kv1 and the entire Pachyrhynchini set, but the size of this probe set: 225,632 probes targeting 12,522 (Van Dam et al. 2022b) may prohibit its combined use, as third-party sequencing companies often have price increases associated with higher number of probes synthesized.

## Conclusion

Our study confirms previous recommendations for developing an optimized UCE probe design. Selecting a taxon to serve as the base genome that has the shortest overall genetic distance from all other taxa involved in probe design, and setting the initial locus identification parameters to loci shared between this base genome and just 1 other taxon, will result in an optimized probe set that recovers the most loci during in silico testing, and which have the longest per-locus alignments. We also found that genome assembly metrics, such as completeness measures, do not correlate well with base genome performance, and alone are not a good justification for selecting a taxon to serve as the base genome. However, use of an annotated

genome as the base will allow identification of UCE loci targeted by the probe design, which can in turn be used to merge cogenic loci for analysis, and prevent development of off-target probes, both of which are important aspects to consider (Van Dam et al. 2021, 2022a). Finally, we found further evidence that combining generalized probes with tailored probes improves phylogenetic performance of the probe set. Applying these aspects to probe design we developed 2 novel beetle probe sets: the Hydro 2.7kv1 targeting 2,761 loci (1,804 tailored, 957 generalized) using 32,657 probes; and the Scarab 3kv1 targeting 3,174 loci (2,112 tailored, 1,062 generalized) using 25,786 probes. Both are available under the public domain license (CC-0) and can be accessed at <https://zenodo.org/record/7926017#.ZF0tQC9BxBw>, DOI: 10.5281/zenodo.7926017.

While developing tailored UCE probes for focal taxa appears to be the best method to ensure optimal in vitro UCE locus recovery, selecting an existing probe set for use may be necessary. Our results suggest that when selecting among UCE probe sets, those phylogenetically closest to the ingroup are likely to perform the best. Additionally, probe sets containing probes developed across a broader taxonomic breadth, in addition to tailored probes for the focal clade, are likely to perform better than selecting a “universal” probe set consisting solely of generalized probes. We also found that probe sets consisting solely of probes tailored to a more derived clade, phylogenetically adjacent to focal taxa, may perform worse than generalized probes alone. Ultimately, all in silico testing, including that conducted here, should be considered exploratory, requiring in vitro validation, as in vitro results differ for a variety of reasons including physical probe synthesis, probe efficacy, and stochastic error. While our study has focused on beetles, we expect our findings to be broadly applicable to arthropod UCE probe design and application, given UCEs in both groups are primarily exonic (Hedin et al. 2019, Van Dam et al. 2021) and its members possess broad genomic- and taxonomic diversity.

## Supplementary Material

Supplementary material is available at *Insect Systematics and Diversity* online.

## Specimen Collection Statement

The authors attest that all legal and regulatory requirements, including export and import collection permits, have been followed for the collection of specimens from source populations at any international, national, regional, or other geographic level for all relevant field specimens collected as part of this study.

## Acknowledgments

We are thankful to Giulio Montanaro and Michele Rossini for species identifications, and to Vasily Grebennikov and Geoff Monteith for providing specimens used in this study. Two anonymous reviewers are also thanked for their comments improving this manuscript. Computational analyses and data processing reported in this study were conducted on Northern Arizona University's Monsoon computing cluster, funded by Arizona's Technology and Research Initiative Fund and the University of Kansas Center for Research computer cluster. This material is based upon work supported by the US National Science Foundation under grants DEB-1942193 to NLG and DEB-1453452 to AEZS. The work of ST was supported by the Academy of Finland grant: 331631, and 3-yr grant from the University of Helsinki.

## Author Contributions

Grey T. Gustafson (Conceptualization-Equal, Formal analysis-Equal, Funding acquisition-Equal, Project administration-Lead, Writing – original draft-Lead,



Writing – review & editing-Lead), Rachel Glynn (Data curation-Equal, Formal analysis-Equal, Writing – original draft-Equal, Writing – review & editing-Equal), Andrew Short (Funding acquisition-Equal, Writing – original draft-Equal, Writing – review & editing-Equal), Sergei Tarasov (Conceptualization-Equal, Funding acquisition-Equal, Software-Equal, Writing – original draft-Equal, Writing – review & editing-Equal), Nicole L. Gunter (Conceptualization-Equal, Funding acquisition-Equal, Writing – original draft-Equal, Writing – review & editing-Equal)

## Data Availability

The Supplementary Materials, including the designed UCE probe sets, are available at Zenodo (<https://zenodo.org/record/7926017#.ZF0tQC9BxBw>), doi: 10.5281/zenodo.7926017).

## References

- Ahrens D, Schwarzer J, Vogler AP. The evolution of scarab beetles tracks the sequential rise of angiosperms and mammals. *Proc R Soc Lond B Biol Sci*. 2014;281(1791):20141470.
- Alexander A. Comparing\_monolithic\_UCE\_fastas v0.2 computer program. 2018. [https://doi.org/https://github.com/laninsky/comparing\\_monolithic\\_UCE\\_fastas](https://doi.org/https://github.com/laninsky/comparing_monolithic_UCE_fastas).
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Andrews, S. FastQC: A quality control tool for high throughput sequence data [Online]. 2010. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Baca SM, Alexander A, Gustafson GT, Short AEZ. Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of “Hydradephaga”. *Syst Entomol*. 2017;42(4):786–795. <https://doi.org/10.1111/syen.12244>
- Baca SM, Gustafson GT, Alexander AM, Gough HM, Toussaint EFA. Integrative phylogenomics reveals a Permian origin of Adephaga beetles. *Syst Entomol*. 2021;46(4):968–990. <https://doi.org/10.1111/syen.12506>
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Comput Biol*. 2012;19:455–477.
- Blaimer BB, Lloyd MW, Guillory WX, Brady SG. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS One*. 2016;11(8):e0161531. <https://doi.org/10.1371/journal.pone.0161531>
- Blaimer BB, Brady SG, Schultz TR, Lloyd MW, Fisher BL, Ward PS. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evol Biol*. 2015;15:271. <https://doi.org/10.1186/s12862-015-0552-5>
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics*. 2014;30:2114–2120.
- Borowiec ML. AMAS: a fast tool for alignment and computing of summary statistics. *PeerJ*. 2016;4:e1660. <https://doi.org/10.7717/peerj.1660>
- Bossert S, Murray EA, Almeida EAB, Brady SG, Blaimer BB, Danforth BN. Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. *Mol Phylogenet Evol*. 2019;130:121–131. <https://doi.org/10.1016/j.ympev.2018.10.012>
- Bradford TM, Ruta R, Cooper SJB, Libonatti ML, Watts CHS. Evolutionary history of the Australasian Scirtinae (Scirtidae; Coleoptera) inferred from ultraconserved elements. *Invertebr Syst*. 2022;36:291–305.
- Branstetter MG, Longino JT, Ward PS, Faircloth BC. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods Ecol Evol*. 2017a;8(6):768–776. <https://doi.org/10.1111/2041-210X.12742>
- Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. Phylogenomic insights into the evolution of stinging wasps and the origin of ants and bees. *Curr Biol*. 2017b;27(7):1019–1025. <https://doi.org/10.1016/j.cub.2017.03.027>
- Brunke AJ, Hansen AK, Salnitska M, Kypke JL, Predeus AV, Escalona H, Chapados JT, Eyres J, Richter R, Smetana A, et al. The limits of Quediini at last (Staphylinidae: Staphylininae): a rove beetle mega-radiation resolved by comprehensive sampling and anchored phylogenomics. *Syst Entomol*. 2021;46(2):396–421. <https://doi.org/10.1111/syen.12468>
- Buenaventura E, Lloyd MW, Perillalópez M, González V, Thomas-Cabianca A, Dikow T. Protein-encoding ultraconserved elements provide a new phylogenomic perspective of Oestroidea flies (Diptera, Calyptratae). *Syst Entomol*. 2021;46:5–27.
- Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17(4):540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ pre-processor. *Bioinformatics*. 2018;34(17):i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 2014;30(1):31–37. <https://doi.org/10.1093/bioinformatics/btt310>
- Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Lett*. 2012;8:783–786.
- Cruaud A, Nidelet S, Arnal P, Weber A, Fusu L, Gumovsky A, Huber J, Polaszek A, Rasplus JY. Optimized DNA extraction and library preparation for minute arthropods: application to target enrichment in chalcid wasps used for biocontrol. *Mol Ecol Resour*. 2019;19(3):702–710. <https://doi.org/10.1111/1755-0998.13006>
- Esselstyn JA, Oliveros CH, Swanson MT, Faircloth BC. Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved loci. *Genome Biol Evol*. 2017;9(9):2308–2321. <https://doi.org/10.1093/gbe/evx168>
- Faircloth BC. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*. 2016;32(5):786–788. <https://doi.org/10.1093/bioinformatics/btv646>
- Faircloth BC. Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods Ecol Evol*. 2017;8(9):1103–1112. <https://doi.org/10.1111/2041-210X.12754>
- Faircloth BC, Branstetter MG, White ND, Brady SG. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol Ecol Resour*. 2015;15(3):489–501. <https://doi.org/10.1111/1755-0998.12328>
- Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*. 2012;61(5):717–726. <https://doi.org/10.1093/sysbio/sys004>
- Forthman M, Miller CW, Kimball RT. Phylogenomic analysis suggests Coreidae and Alydidae (Hemiptera: Heteroptera) are not monophyletic. *Zool Scr*. 2019;48(4):520–534. <https://doi.org/10.1111/zsc.12353>
- Forthman M, Miller CW, Kimball RT. Phylogenomics of the leaf-footed bug subfamily Coreinae (Hemiptera: Coreidae). *Insect Syst Divers*. 2020;4(4):2.
- Gunter NL, Weir TA, Slipinksi A, Bocak L, Cameron SL. If dung beetles (Scarabaeidae: Scarabaeinae) arose in association with dinosaurs, did they also suffer a mass co-extinction at the K-Pg boundary?. *PLoS One*. 2016;11(5):e0153570. <https://doi.org/10.1371/journal.pone.0153570>
- Gustafson GT, Baca SM, Alexander A, Short AEZ. Phylogenomic analysis of the beetle suborder Adephaga with comparison of tailored and generalized ultraconserved element probe performance. *Syst Entomol*. 2020;45:552–570.
- Gustafson GT, Alexander A, Sproul JS, Pflug JM, Maddison DR, Short AEZ. Ultraconserved element (UCE) probe set design: base genome and initial design parameters critical for optimization. *Ecol Evol*. 2019;9(12):6933–6948. <https://doi.org/10.1002/ece3.5260>
- Hedin M, Derkarabetian S, Alfaro A, Ramirez MJ, Bond JE. Phylogenomic analysis and revised classification of atypoid mygalomorph spiders (Araneae, Mygalomorphae), with notes on arachnid ultraconserved element loci. *PeerJ*. 2019;7:e6864.
- Hellemans S, Wang M, Hasegawa N, Šobotník J, Scheffrahn RH, Bourguignon T. Using ultraconserved elements to reconstruct the termite tree of life.

- Mol Phylogenet Evol. 2022;173:107520. <https://doi.org/10.1016/j.ympev.2022.107520>
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35(2):518–522. <https://doi.org/10.1093/molbev/msx281>
- Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969. p. 21–132.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–589. <https://doi.org/10.1038/nmeth.4285>
- Karin BR, Gamble T, Jackman TR. Optimizing phylogenomics with rapidly evolving long exons: comparison with anchored hybrid enrichment and ultraconserved elements. Mol Biol Evol. 2020;37(3):904–922. <https://doi.org/10.1093/molbev/msz263>
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780. <https://doi.org/10.1093/molbev/mst010>
- Kieran TJ, Gordan ERL, Forthman M, Hoey-Chamberlain H, Kimball RT, Faircloth BC, Weirauch C, Glenn TC. Insight from an ultraconserved element bait set designed for hemipteran phylogenetics integrated with genomic resources. Mol Phylogenet Evol. 2019;130:297–303.
- Kimball RT, Oliveros CH, Wang N, White ND, Barker FK, Field DJ, Ksepka DT, Chesser RT, Moyle RG, Braun MJ, et al. A phylogenomic supertree of birds. Diversity. 2019;11(7):109. <https://doi.org/10.3390/d11070109>
- Kirsch R, Gramzow L, Therißen G, Siegfried DB, Ffrench-Constant RH, Heckel DG, Pauchet Y. Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: Key events in the evolution of herbivory in beetles. Insect Biochem Mol Biol. 2014;52:33–50.
- Kobayashi T, Hayashi M, Kamite Y, Sota T. Molecular phylogeny of Elmidae (Coleoptera: Byrrhoidea) with a focus on Japanese species: implications for intrafamilial classification. Syst Entomol. 2021;46(4):870–886. <https://doi.org/10.1111/syen.12499>
- Kulkarni S, Wood H, Lloyd M, Hormiga G. Spider-specific probe set for ultraconserved elements offers new perspectives on the evolutionary history of spiders (Arachnida, Araneae). Mol Ecol Resour. 2020;20(1):185–203. <https://doi.org/10.1111/1755-0998.13099>
- Kypke JL. Phylogenetics of the world's largest beetle family (Coleoptera: Staphylinidae): a methodological exploration. Copenhagen, Denmark: Natural History Museum of Denmark, Faculty of Science, University of Copenhagen; 2018. [https://doi.org/https://soeg.kb.dk/permalink/45KBDK\\_KGL/1pioq0f/alma9912277823905763](https://doi.org/https://soeg.kb.dk/permalink/45KBDK_KGL/1pioq0f/alma9912277823905763)
- Lata D, Coates BS, Walden KKO, Robertson HM, Miller NJ. Genome size evolution in the beetle genus *Diabrotica*. G3 Genes Genomes Genet. 2022;12:jkac052.
- Liu D, Niu M, Lu Y, Wei J, Zhang H. Taxon-specific ultraconserved element probe design for phylogenetic analyses of scale insects (Hemiptera: Sternorrhyncha: Coccoidea). Front Ecol Evol. 2022;10:984396.
- Longino JT, Branstetter MG. Phylogenomic species delimitation, taxonomy, and “Bird Guide” identification for the Neotropical ant genus *Rasopone* (Hymenoptera: Formicidae). Insect Syst Divers. 2020;4(2):1–33.
- Longino JT, Branstetter MG. Integrating UCE phylogenomics with traditional taxonomy reveals a trove of New World *Syscia* species (Formicidae: Dorylinae). Insect Syst Divers. 2021;5(2):1–51.
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111–118.
- Manchola OFS, Sáenz ES, Herrera SV, D'Alessio LM, Aldrete ANG, Johnson KP. Mining ultraconserved elements from transcriptome and genome data to explore the phylogenomics of the free-living lice suborder Psocomorpha (Insecta: Psocodea). Insect Syst Divers. 2022;6:1–11.
- Manthey JC, Campillo LC, Burns KJ, Moyle RG. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid Tanagers (Aves, Genus: *Piranga*). Syst Biol. 2016;65:640–650.
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–12. <https://doi.org/10.14806/ej.17.1.200>
- McCormack JE, Tsai WL, Faircloth BC. Sequence capture of ultraconserved elements from bird museum specimens. Mol Ecol Resour. 2016;16(5):1189–1203. <https://doi.org/10.1111/1755-0998.12466>
- McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One. 2013;8(1):e54848. <https://doi.org/10.1371/journal.pone.0054848>
- McKenna DD. Beetle genomes in the 21st century: prospects, progress and priorities. Curr Opin Insect Sci. 2018;25:76–82. <https://doi.org/10.1016/j.cois.2017.12.002>
- McKenna DD, Wild AL, Kanda K, Bellamy CL, Beutel RG, Caterino MS, Farnum CW, Hawks DC, Ivie MA, Jameson ML, et al. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. Syst Entomol. 2015;40:835–880.
- McKenna DD, Shin S, Ahrens D, Balke M, Beza-Beza C, Clarke DJ, Donath A, Escalona HE, Friedrich F, Letsch H, et al. The evolution and genomic basis of beetle diversity. Proc Natl Acad Sci U S A. 2019;116:24729–24737.
- Mikheenko A, Pribelski AD, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142–i150. <https://doi.org/10.1093/bioinformatics/bty266>
- Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–274. <https://doi.org/10.1093/molbev/msu300>
- Pisanty G, Richter R, Martin T, Dettman J, Cardinal S. Molecular phylogeny, historical biogeography and revised classification of andrenine bees (Hymenoptera: Andrenidae). Mol Phylogenet Evol. 2022;170:107151. <https://doi.org/10.1016/j.ympev.2021.107151>
- R Core Team. R: a language and environment for statistical computing computer program, version by R Core Team. Vienna (Austria); 2022
- Robinson DR, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53(1–2):131–147. [https://doi.org/10.1016/0025-5564\(81\)90043-2](https://doi.org/10.1016/0025-5564(81)90043-2)
- Ruane S, Austin CC. Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. Mol Ecol Resour. 2017;17(5):1003–1008. <https://doi.org/10.1111/1755-0998.12655>
- Shin S, Clarke DJ, Lemmon AR, Lemmon EM, Aitken AL, Haddad S, Farrell BD, Marvaldi AE, Oberprieler RG, McKenna DD. Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. Mol Biol Evol. 2018;35:823–836.
- Short AEZ, Fikáček M. Molecular phylogeny, evolution and classification of the Hydrophilidae (Coleoptera). Syst Entomol. 2013;38(4):723–752. <https://doi.org/10.1111/syen.12024>
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst Biol. 2014;63(1):83–95. <https://doi.org/10.1093/sysbio/syt061>
- Sota T, Takami Y, Ikeda H, Liang H, Karagyan G, Scholtz C, Hori M. Global dispersal and diversification in ground beetles of the subfamily Carabinae. Mol Phylogenet Evol. 2022;167:107355. <https://doi.org/10.1016/j.ympev.2021.107355>
- Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32(Web Server issue):W309–W312. <https://doi.org/10.1093/nar/gkh379>
- Streicher JW, Wiens JJ. Phylogenomic analyses of more than 4000 nuclear loci resolve the origin of snakes among lizard families. Biol Lett. 2017;13(9):20170393. <https://doi.org/10.1098/rsbl.2017.0393>
- Streicher JW, Miller EC, Guerrero PC, Correa C, Ortiz JC, Crawford AJ, Pie MC, Wiens JJ. Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2214 loci. Mol Phylogenet Evol. 2018;119:128–143.

- Tajima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol.* 1984;1(3):269–285. <https://doi.org/10.1093/oxfordjournals.molbev.a040317>
- Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 2007;56(4):564–577. <https://doi.org/10.1080/10635150701472164>
- Tamura K, Nei M. Estimation of number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 1993;10(3):512–526. <https://doi.org/10.1093/oxfordjournals.molbev.a040023>
- Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A.* 2004;101(30):11030–11035. <https://doi.org/10.1073/pnas.0404206101>
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. Mega6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol.* 2013;30(12):2725–2729. <https://doi.org/10.1093/molbev/mst197>
- Tarasov S, Dimitrov D. Multigene phylogenetic analysis redefines dung beetle relationships and classification (Coleoptera: Scarabaeidae: Scarabaeinae). *BMC Evol Biol.* 2016;16(1):1–19.
- Van Dam AR, Covas Orizondo JO, Lam AM, McKenna DD, Van Dam MH. Metagenomic clustering reveals microbial contamination as an essential consideration in ultraconserved element design for phylogenomics with insect museum specimens. *Ecol Evol.* 2022a;12:e8625.
- Van Dam MH, Lam AW, Sagata K, Gewa BLaufa R, Balke M, Faircloth BC, Riedel A. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-weevils. *PLoS One.* 2017;12:e0188044.
- Van Dam MH, Trautwein M, Spicer GS, Esposito L. Advancing mite phylogenomics: designing ultraconserved elements for Acari phylogeny. *Mol Ecol Resour.* 2019;19(2):465–475. <https://doi.org/10.1111/1755-0998.12962>
- Van Dam MH, Henderson JB, Esposito L, Trautwein M. Genomic characterization and curation of UCEs improves species tree reconstruction. *Syst Biol.* 2021;70(2):307–321. <https://doi.org/10.1093/sysbio/syaa063>
- Van Dam MH, Cabras AA, Lam AW. How the Easter Egg Weevils got their spots: phylogenomics reveals Müllerian mimicry in *Pachyrhynchus* (Coleoptera, Curculionidae). *Syst Biol.* 2022b;72(3):syac064.
- Ye C, Ma Z, Cannon C, Pop M, Yu DW. Exploiting sparseness in de novo genome assembly. *BMC Bioinform.* 2012;13:1–8.
- Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 2017;45(D1):D744–D749. <https://doi.org/10.1093/nar/gkw1119>
- Zhang F, Ding Y, Zhu CD, Zhou X, Orr MC, Scheu S, Luan YX. Phylogenomics from low-coverage whole-genome sequencing. *Methods Ecol Evol.* 2019a;10(4):507–517. <https://doi.org/10.1111/2041-210x.13145>
- Zhang SQ, Che LH, Li Y, Liang D, Pang H, Ślipiński SA, Zhang P. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nat Commun.* 2018;9:205.
- Zhang YM, Williams JL, Lucky A. Understanding UCEs: a comprehensive primer on using ultraconserved elements for arthropod phylogenies. *Insect Syst Divers.* 2019b;3(5):31–12.
- Zhang YM, Buffington ML, Looney C, László Z, Shorthouse JD, Ide T, Lucky A. UCE data reveal multiple origins of rose gallers in North America: global phylogeny of *Diplolepis* Geoffroy (Hymenoptera: Cynipidae). *Mol Phylogenet Evol.* 2020;153:106949. <https://doi.org/10.1016/j.ympev.2020.106949>