diversity

MDPI

Article

In Silico Assessment of Probe-Capturing Strategies and Effectiveness in the Spider Sub-Lineage Araneoidea (Order: Araneae)

Yi-Yen Li [1,2], Jer-Min Tsai [3], Cheng-Yu Wu [1], Yi-Fan Chiu [1], Han-Yun Li [1], Natapot Warrit [4], Yu-Cen Wan [1], Yen-Po Lin [2,5], Ren-Chung Cheng [2,*] and Yong-Chao Su [1,*]

1 Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung 80708, Taiwan; yiyenl30@gmail.com (Y.-Y.L.); u109551003@kmu.edu.tw (C.-Y.W.); ivan06513i@gmail.com (Y.-F.C.); hanyun1012@gmail.com (H.-Y.L.); u109551007@gap.kmu.edu.tw (Y.-C.W.)
2 Department of Life Sciences, National Chung Hsing University, Taichung 40227, Taiwan; yplin@tesri.gov.tw
3 Department of Information and Communication, Kun Shan University, Tainan 710, Taiwan; tjm@fhl.net
4 Center of Excellence in Entomology and Department of Biology, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand; natapot.w@chula.ac.th
5 Taiwan Endemic Species Research Institute, No.1, Minsheng East Rd., Jiji Township, Nantou 55244, Taiwan
* Correspondence: bolasargiope@email.nchu.edu.tw (R.-C.C.); ycsu527@kmu.edu.tw (Y.-C.S.); Tel.: +886-4-228-40416 (ext. 707) (R.-C.C.); +886-7-312-1101 (ext. 6983) (Y.-C.S.)

Abstract: Reduced-representation sequencing (RRS) has made it possible to identify hundreds to thousands of genetic markers for phylogenomic analysis for the testing of phylogenetic hypotheses in non-model taxa. The use of customized probes to capture genetic markers (i.e., ultraconserved element (UCE) approach) has further boosted the efficiency of collecting genetic markers. Three UCE probe sets pertaining to spiders (Araneae) have been published, including one for the suborder Mesothelae (an early diverged spider group), one for Araneae, and one for Arachnida. In the current study, we developed a probe set specifically for the superfamily Araneoidea in spiders. We then combined the three probe sets for Araneoidea, Araneae, and Arachnid into a fourth probe set. In testing the effectiveness of the 4 probe sets, we used the captured loci of the 15 spider genomes in silico (6 from Araneoidea). The combined probe set outperformed all other probe sets in terms of the number of captured loci. The Araneoidea probe set outperformed the Araneae and Arachnid probe sets in most of the included Araneoidea species. The reconstruction of phylogenomic trees using the loci captured from the four probe sets and the data matrices generated from 50% and 75% occupancies indicated that the node linked to the *Stegodyphus* + RTA (retrolateral tibial apophysis) clade has unstable nodal supports in the bootstrap values, gCFs, and sCFs. Our results strongly indicate that developing ad hoc probe sets for sub-lineages is important in the cases where the origins of a lineage are ancient (e.g., spiders ~380 MYA).

Keywords: target sequencing; reduced representation sequencing (RRS); spider phylogenomics; deep phylogeny

Citation: Li, Y.-Y.; Tsai, J.-M.; Wu, C.-Y.; Chiu, Y.-F.; Li, H.-Y.; Warrit, N.; Wan, Y.-C.; Lin, Y.-P.; Cheng, R.-C.; Su, Y.-C. In Silico Assessment of Probe-Capturing Strategies and Effectiveness in the Spider Sub-Lineage Araneoidea (Order: Araneae). *Diversity* **2022**, *14*, 184. https://doi.org/10.3390/d14030184

Academic Editor: Luc Legal

Received: 18 December 2021
Accepted: 28 February 2022
Published: 3 March 2022

1. Introduction

High-throughput sequencing is widely used for the generation of genomic data in phylogenomic research [1–4]. Reduced-representation sequencing (RRS) methods [5] have made it possible to collect hundreds to thousands of genetic markers at a fraction of the cost of whole-genome sequencing [6]. The ultraconserved elements approach (UCE approach), a form of target DNA sequencing, is becoming particularly prevalent [7–9]. The UCE approach using customized probes makes it possible for researchers to capture thousands of genetic markers from non-model taxa, thereby making it possible to test hypotheses about phylogeny from shallow (e.g., <5 MYA) to deep (e.g., >200 MYA) divergence times [10]. Despite the importance of the UCE approach in phylogenomics, the design of ad hoc probe

sets remains a technical gap such that many researchers are forced to use probe sets designed for similar taxa or for different taxonomic levels. In the current study, we compared the effectiveness of an ad hoc probe set for spiders in the superfamily Araneoidea to the existing probe sets that are known to be applicable to higher taxonomic levels in arachnids [9,11].

UCEs are non-variable genomic fragments that occur across species in a given taxonomic group [12]. These genomic fragments, which are often in >200 bps conserved regions [13], have been detected in a variety of taxa [14]. The functions of these UCEs are unknown [15], and the types of UCEs vary among taxonomic groups [16]. Hedin et al. [17] showed that the spider UCEs mostly correspond to exons. In a pioneering work, Faircloth et al. [18] captured 854 UCE loci to reconstruct the phylogenomic tree of birds. Subsequent research assessed the utility of the UCE approach in applying phylogenomic hypotheses to taxa dated from 5 MYA to 200 MYA [7,19,20]. The UCE approach has also been extended to the reconstruction of species trees and coalescent methods [21,22]. Recent advances in the UCE approach have strengthened phylogenetic hypothesis testing and phylogenetic tree reconstruction, including for arthropods [9,10,23].

The UCE approach was first applied to arachnids by Starrett et al. [23], and to Araneae by Kulkarni et al. [11]. Note that the order Araneae includes 49,877 species [24] in three subclades, suborder Mesothelae, infraorder Mygalomorphae, and infraorder Araneomorphae, with evolutionary time extending back to more than 300 MYA [25,26]. The application of the probe sets designed for the higher levels (for order and class) could be problematic because the probes may not be fully targeted, thus reducing the number of captured loci when testing the hypotheses of the phylogeny within suborders and lower taxonomic levels. Xu et al. [27] tested a customized probe set in the suborder Mesothelae. Hedin et al. [23] reconstructed the UCE phylogenomic tree in the Mygalomorphae species. In the current study, we developed an ad hoc UCE probe set for the superfamily Araneoidea, which contains 17 families, 25% spider diversity, and a variety of web architectures [28].

We developed the Araneoidea probe set in accordance with the pipeline outlined by Faircloth [9]. We then compared the effectiveness of our Araneoidea probe set with those for arachnid and Araneae. Finally, we combined these three probe sets as the fourth probe set. Note that we did not assess the Mesothelae probe set because it is clearly applicable at that suborder level [27]. We evaluated the effectiveness of the four probe sets in two schemes. (1) We performed in silico testing on the number of captured UCE loci in 15 genomes of Araneoidea and other spider species. (2) We compared the phylogenomic trees reconstructed using the concatenation and gene–tree–species–tree approaches with various data matrices to compare the tree topologies and node supports.

2. Materials and Methods

2.1. Data Sources of UCE Loci

As data sources for our in silico testing, we employed two published probe sets for ultraconserved elements [9,11], including 14 published genomes (Table 1) and 1 de novo assembled genome (*Argyrodes miniaceus*).

2.2. Genome Assembly

We assembled the genomes de novo using the procedure below. We used TRIMMOMATIC [29] for raw read trimming and adaptor removal. KMERGENIE [30] was then used to estimate the optimal k-mer length for genomic assembly. Finally, ABYSS 2.0 [31] was used to assemble the genome for *Argyrodes miniaceus* using the following settings: k = 55, B = 30 G. ABYSS–FAC was used to evaluate the quality of the genome assembly.

Table 1. Genomes fetched from GenBank. Information for all 14 genomes used in this research, which were fetched from GenBank. This table displays species name, assembly accession, assembly level, assembly submission date, N50 of contigs, coverage rate, and references for each genome. Adding *Argyrodes miniaceus*, 15 genomes are included in this study.

Organism Name	Assembly Accession	Total Sequence Length	Assembly Level	Submission Date	Contig N50	Coverage	Reference
Acanthoscurria geniculata	GCA_000661875.1	7,178,402,394	Contig	2014-04-29	541	21.5×	[32]
Anelosimus studiosus	GCA_008297655.1	2,033,432,615	Scaffold	2019-09-05	1132	79.0×	[33]
Araneus ventricosus	GCA_013235015.1	3,656,621,265	Scaffold	2019-08-02	22,999	70×	[34]
Argiope bruennichi	GCA_015342795.1	1,670,285,661	Chromosome	2020-11-16	284,772	70×	[35]
Dolomedes plantarius	GCA_907164885.1	2,381,335,874	Chromosome	2021-05-16	292,830	19.4×	[36]
Dysdera silvatica	GCA_006491805.2	1,365,686,336	Scaffold	2021-07-07	21,954	96.9×	[37]
Latrodectus hesperus	GCA_000697925.2	1,233,806,489	Scaffold	2018-02-05	15,961	80.0×	[38]
Loxosceles reclusa	GCA_001188405.1	3,262,478,678	Contig	2015-04-27	1834	55×	[38]
Oedothorax gibbosus	GCA_019343175.1	821,427,276	Chromosome	2021-08-05	979,336	14.0×	[39]
Parasteatoda tepidariorum	GCA_000365465.3	1,228,972,128	Scaffold	2019-06-14	66,479	48.0×	[40]
Pardosa pseudoannulata	GCA_008065355.1	4,207,954,893	Scaffold	2019-08-22	23,226	423.95×	[41]
Stegodyphus dumicola	GCF_010614865.1	2,551,871,595	Scaffold	2020-02-14	254,130	49.0×	[42]
Stegodyphus mimosarum	GCA_000611955.2	2,738,704,917	Scaffold	2014-08-01	40,146	86.0×	[32]
Trichonephila clavipes	GCA_002102615.1	2,439,301,466	Scaffold	2017-04-20	7993	140.0×	[43]

2.3. Design of UCE Probe Set for Araneidae

The design of UCE probes was based on the PHYLUCE pipeline [9,44]. The genomes of *Argyrodes miniaceus*, *Latrodectus hesperus*, *Loxosceles reclusa*, *Trichonephila clavipes*, *Parasteatoda tepidariorum*, and *Stegodyphus mimosarum* were used in UCE probe design as follows: (1) ART v2016.06.05 [45] was used to simulate genomic fragments of *Argyrodes miniaceus*, *Latrodectus hesperus*, *Loxosceles reclusa*, *Trichonephila clavipes*, and *Parasteatoda tepidariorum* into 100-bps reads. (2) Simulated short reads were aligned to *Stegodyphus mimosarum* (i.e., the base genome [32]) using STAMPY (substitution rate = 0.05 and insert size = 400) [46]. Misaligned fragments were removed using SAMTOOLS [47], and the aligned fragments were combined in the browser-extensible data (BED) format using BEDTOOLS [48]. (3) Duplicated genomic fragments were removed using PHYLUCE script (phyluce_probe_strip_masked_loci_from_set) to detect and remove fragments that were mapped but designated too short (<80 bps) or within masked regions of the *Stegodyphus mimosarum* genome (more than 25%). (4) SQLITE v 3.34.0 [49] was used to construct a database of candidate UCE sites for *Argyrodes miniaceus*, *Latrodectus hesperus*, *Loxosceles reclusa*, *Trichonephila clavipes*, and *Parasteatoda tepidariorum* to determine the shared conserved regions. PHYLUCE was then used to remove duplicated candidate probes, and LASTZ [50] was used to align the candidate probes with a given genome to enable the extraction of UCE sites for a given species. (5) Finally, we relaxed the similarity to 50% and reconstructed the database in SQLITE to create the final UCE probes, whereupon we repeated the duplicate-probe removal process in PHYLUCE. The final probe length was 120 bps with tiling, with 60 bps overlapping (thus covering 180 bps) per target locus.

2.4. In Silico Simulation of Probe Sets Aimed at Capturing Affinity

Simulations were conducted using four probe sets for Arachnids (Arachnid probe set [9]), Araneae (Araneae probe set [11]), Araneoidea (Araneoidea probe set), and a combination of these three (combined probe set). In generating the combined probe set, we compiled the probe sets for Arachnid, Araneae, and Araneoidea and removed potential duplicated probes using LASTZ Python script (phyluce_probe_remove_duplicate_hits_from_probes_using_lastz [44]). Each probe set was tested on 15 genomes using standardized testing procedures. (1) The probes were aligned with the targeted genome using LASTZ [50]. (2) The probes were then aligned and mapped to the targeted genome using PHYLUCE [44] to extract the 500-bps regions on both sides of the UCE probe sites. Note that our objective was to simulate the fragment length when conducting in-solution cap-

ture. (3) The probes were aligned with each extracted sequence using LASTZ by running phyluce_assembly_match_contigs_to_probes to identify which loci sequences belonged. The duplicates were again sorted out. Following the capture and filtering of fragments from each probe set and each targeted genome, the number of captured loci and proportion of identified loci (defined as the capture rate) in all simulated contigs were calculated per genome per probe set. The captured loci per probe set were then used to reconstruct the phylogenomic tree per data matrix from each probe set.

2.5. Reconstruction of Phylogenomic Tree Using the Captured Data Matrix for Each Probe Set

We generated the data matrices for different occupancy (the smallest percentage of data per locus in a matrix). We counted the number of in silico captured loci in each genome under occupancies from 10% to 100%, with 10% as the increment. We then decided the occupancies to use in the final tree-reconstruction analyses. The script phyluce_align_get_only_loci_with_min_taxa was used to output the locus matrices of the probe set using occupancies 50% and 75%, respectively. This allowed the omission of up to 50% and up to 25% missing taxa per locus, which resulted in two matrices per probe set. In total, we used eight locus matrices in reconstructing the phylogenomic trees of available spider species.

The MAFFT [51] script, phyluce_align_seqcap_align, was used to align each UCE locus, whereupon the ends of the aligned fragments were trimmed using GBLOCKS (default arguments of PHYLUCE: −b1 = 0.5, −b2 = 0.85, −b3 = 8, and −b4 = 10) [52] via the script phyluce_align_get_gblocks_trimmed_alignments_from_untrimmed. We then concatenated the aligned loci using phyluce_align_concatenate_alignments to produce a matrix for each probe-set per occupancy and then output the matrices in PHYLIP format (partition scheme) and in NEXUS format. After detecting the models with the best fit in each locus using MODELFINDER [53], IQTREE-2.0.3 [54] was used to reconstruct the phylogenomic trees via concatenation involving 1000 bootstrap operations. We also used the gene–tree–species–tree approach in IQTREE-2.0.3 to infer the gene trees and calculate the gene concordance factors (gCFs) and site concordance factors (sCFs) for the nodes associated with species tree [55]. In accordance with the methods outlined by Wheeler et al. [56], *Acanthoscurria geniculata* (Theraphosidae) was used as an outgroup. Finally, FIGTREE [57] was used to visualize phylogenomic trees.

3. Results

3.1. De Novo Genome Assembly

Using Illumina Hi-seq short-read sequences, we assembled a genome, *Argyrodes miniaceus*. From 7,051,281 contigs in *Argyrodes miniaceus*, we obtained a total assembled length of 35.51×10^6 bps with N50 = 618 bps and a maximum assembled contig of 5834 bps (for genome assembly statistics, see Table S1). This de novo assembled draft genome was then intended to be used to detect UCE probes.

3.2. Probe Detection

Using *Stegodyphus mimosarum* as the base genome in accordance with the methods outlined by Faircloth et al. [9], we detected 12,679 probes related to 1374 UCE loci using *Argyrodes miniaceus*, *Latrodectus hesperus*, *Loxosceles reclusa*, *Trichonephila clavipes*, and *Parasteatoda tepidariorum*.

3.3. In Silico Testing of Capture Efficiency

We used four probe sets for the in silico capture of targeted loci from 15 genomes. We detected a total of 7357 loci using the newly designed Araneoidea probe set (Figure 1 and Tables S2–S5). From the Arachnid probe set, we detected a total of 4579 loci. From the Araneae probe set, we detected a total of 9103 loci. Accordingly, even though we mostly used the genomes in Araneoidea in this study, we collected fewer loci compared with

the Araneae probe set, which mostly included the well-assembled genomes fetched from GenBank (Table 2). From the combined probe set, we detected 14,271 loci.

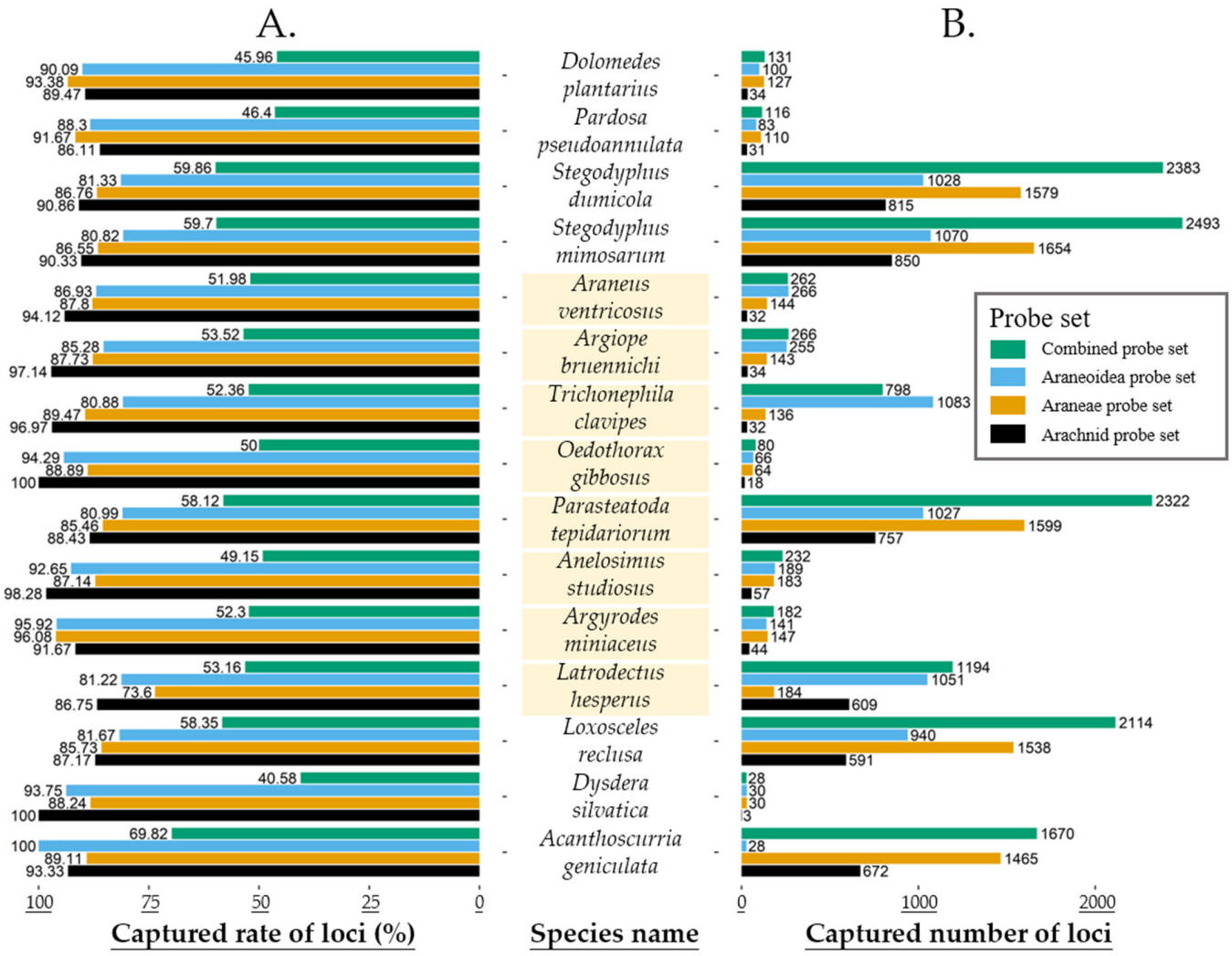

Figure 1. Results of in silico tests. Capture rate of loci (**A**) and captured loci number (**B**) in each probe set. The Araneoidea species are shaded in orange.

The performance of each probe set in capturing the loci in each genome varied as a function of the taxonomic group. In eight of the genomes, the Araneae probe set outperformed the Araneoidea probe set in the capture of loci (Figure 1). The Arachnid probe set outperformed the Araneoidea probe set in only one species, *Acanthoscurria geniculata* (Mygalomorphae). Nonetheless, the Araneoidea probe set outperformed two published probe sets, mostly in the Araneoidea species (e.g., *Trichonephila clavipes*, *Latrodectus hesperus*, *Argiope bruennichi*, *Anelosimus studiosus*, *Oedothorax gibbosus*, and *Araneus ventricosus*). The combined probe set outperformed all probe sets in most species except *Araneus ventricosus*, *Argiope bruennichi*, and *Dysdera silvatica*, which presented very few loci (3–30 loci) in each probe set (Figure 1).

The recovery rates of the probe sets were assessed using the combined probe set as targeted contigs to determine the number of captures and capture rates. The re-capture number and re-capture rates were lowest in the Arachnid probe set, followed by the Araneoidea probe set and the Araneae probe set (Figure 2).

Table 2. List of the genomes for probe-set design and the numbers of captured loci. All probe sets used in this research are listed below, ordered by published year. This table displays the targeted taxa of probe-set design, species names of the genomes used to identify UCE loci, species names of the genomes used to design probes, number of UCE loci, number of probes, and the published year of the probe set.

Target Taxon	Genomes Used to Identify UCEs	Genomes Used to Design Probes	Number of UCE Loci	Number of Probes	Publication Year	Reference
Arachnida	*Trithyreus pentapaltis*, *Atypoides riversi*, *Phrynus marginemaculatus*, *Cryptocellus goodnighti*, *Mitopus morio*, *Bothriurus keyserlingi*, *Pseudouroctonus apacheanus*, *Hadogenes troglodytes*, *Vaejovis deboerae*, *Ixodes scapularis*, *Limulus polyphemus*	*Ixodes scapularis*, *Limulus polyphemus*, *Acanthoscurria geniculata*, *Centruroides exilicauda*, *Latrodectus hesperus*, *Mesobuthus martensii*, *Parasteatoda tepidariorum*, *Stegodyphus mimosarum*, *Amblyomma americanum*	1120	14,799	2017	[9]
Araneae	*Parasteatoda tepidariorum*, *Acanthoscurria geniculata*, *Stegodyphus mimosarum*	*Parasteatoda tepidariorum*, *Acanthoscurria geniculata*, *Stegodyphus mimosarum*	2021	15,051	2020	[11]
Araneoidea	*Argyrodes miniaceus*, *Latrodectus hesperus*, *Loxosceles reclusa*, *Trichonephila clavipes*, *Parasteatoda tepidariorum*, *Stegodyphus mimosarum*	*Argyrodes miniaceus*, *Latrodectus hesperus*, *Loxosceles reclusa*, *Trichonephila clavipes*, *Parasteatoda tepidariorum*, *Stegodyphus mimosarum*	1374	12,679	2021	This article
-	-	-	3344	30,379	2021	This article

3.4. Capture Rates and Number of Loci in Various Occupancies

We present the number of captured loci in each genome under occupancies from 10% to 100%, in increments of 10%. The numbers of captured loci were higher in the combined probe set and Araneae probe set under occupancies of 10% to 30%. The numbers of captured loci did not vary considerably under occupancies of >50%. We observed similar trends in the retention ratio, with the highest retention in the Araneoidea probe set, and a merging of results at occupancies of >50% (Figure 3). Thus, in accordance with the UCE-phylogenomic results published earlier, we used occupancies of 50% and 75% in reconstructing the phylogenomic trees [11,25]. Note, however, that this strategy reduced the in silico capture number to less than 350 loci in each genome (see Figure 4; for other trees, see Figures S1–S7).

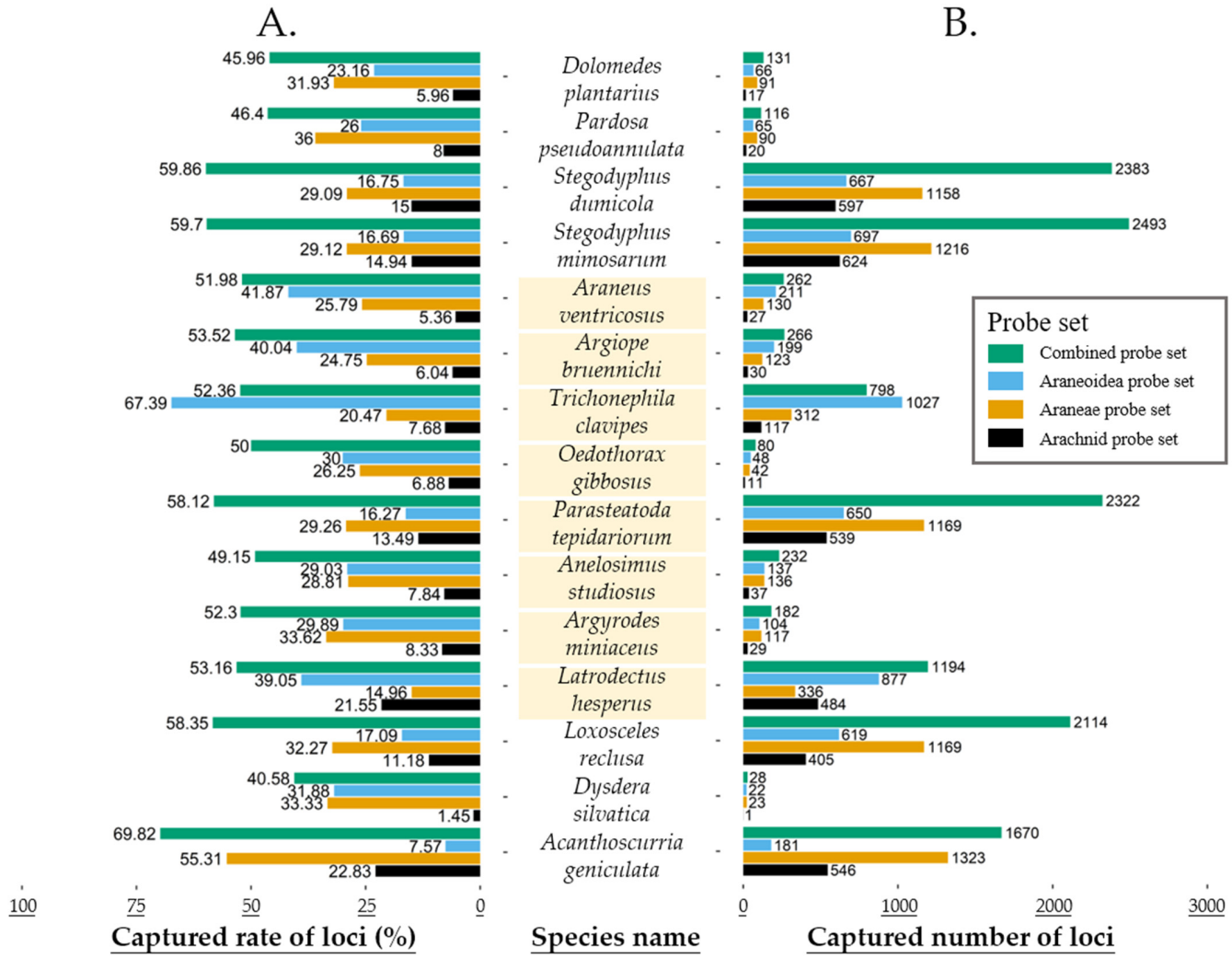

Figure 2. Comparison of all three probe sets with the combined probe set. Capture rate of loci (**A**) and captured loci number (**B**) when each probe set aligned with simulated contigs using the sequences of the combined probe set.

3.5. Tree Reconstruction Using Simulated Captured Loci

We reconstructed the phylogenomic trees using two occupancies (50% and 75%) for the loci captured from the four probe sets, thereby resulting in eight data matrices. The resulting topologies were similar to previous findings (e.g., Kulkarni et al. [58]), and the supports (i.e., bootstrap, gCF, and sCF) of each node were similar between the results of these two datasets (shown in Figure 4B,C).

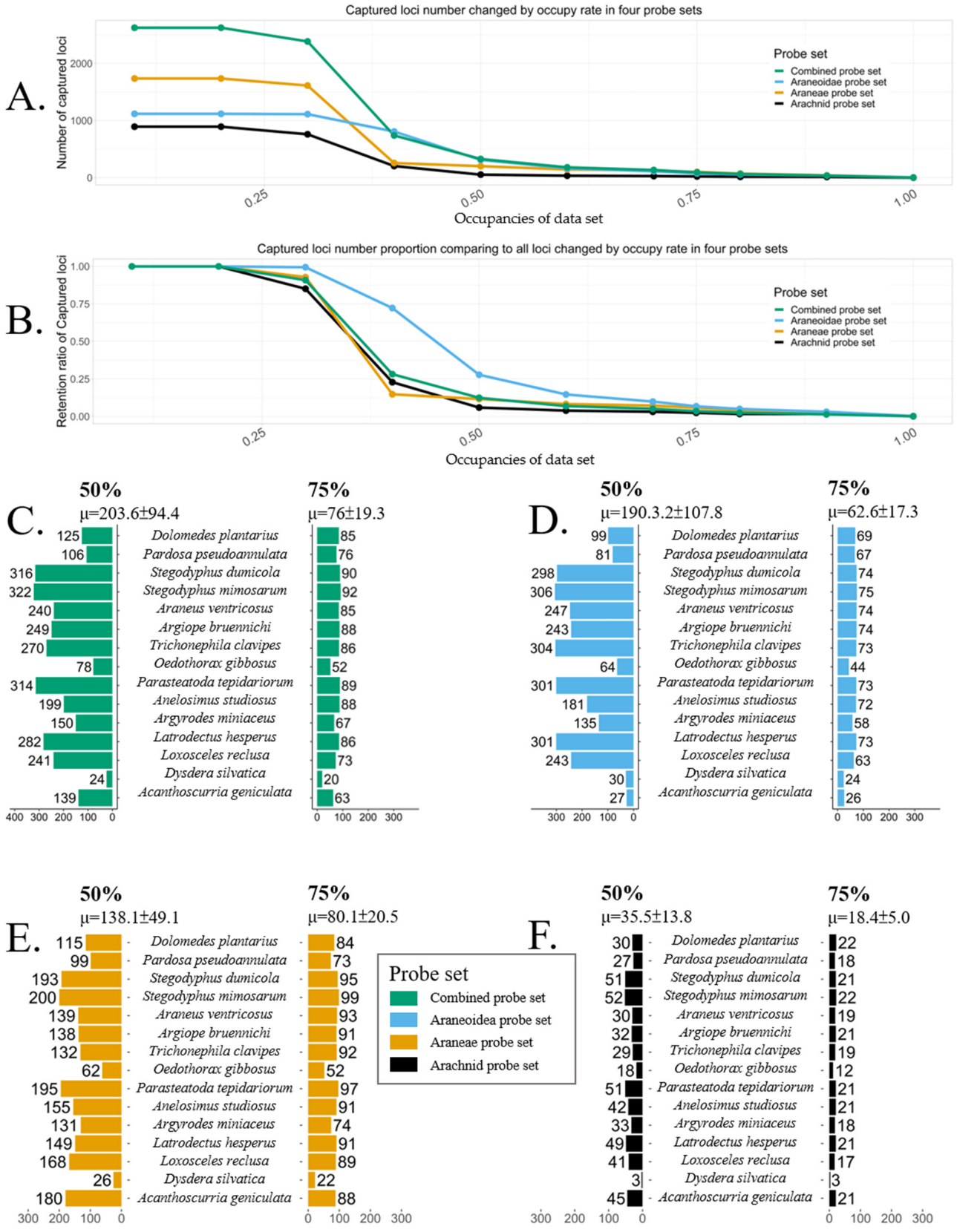

Figure 3. Loci numbers of data sets. (**A**,**B**) Number of loci changed by filtered increment of 10% (from 10% to 100%) occupancies (for data, see Table S6). (**C**–**F**) Number of loci used in tree reconstruction of each data set (Table S7).

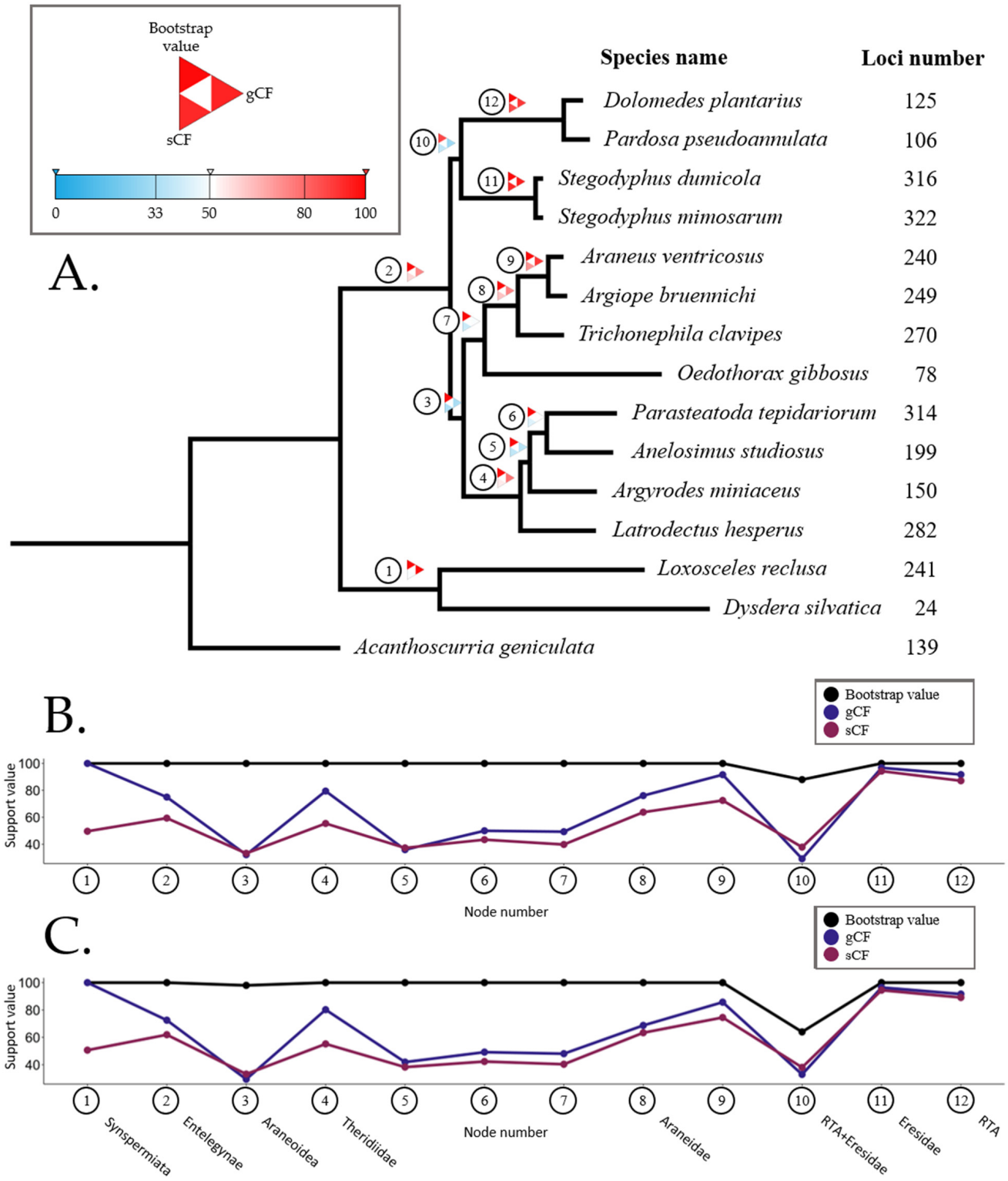

Figure 4. Phylogenetic tree of the data set that had the most captured loci (combined probe set filtered with 50% occupancy). (**A**) Tree topology: color of triangles represents three types of support values—bootstrap, gCF, and sCF—and dashed lines indicate the loci numbers used in tree reconstruction. (**B**,**C**) Support values for different nodes on reconstructed trees using combined probe set of 50% (**B**) and 75% (**C**) occupancies. Other trees, see Figures S1–S7.

4. Discussion

This study suggests that even when dealing with a monophyletic group (e.g., Araneae), an ancient evolutionary origin (e.g., ~380 MYA), the use of a specific probe set to test phylogenetic hypotheses within a sub-lineage could benefit via more lineage-specific loci, and potentially, more captured loci. A specific probe set is meant to enable the capture of a larger number of specific loci to facilitate phylogenomic analysis when combined with probes designed for higher taxonomic levels. The number of loci revealed by the Araneoidea probe set (7357 loci) was lower than that of the Araneae probe set (9103 loci). However, the loci captured in the Araneoidea species outperformed the probe sets designed for higher taxonomic levels (the Arachnid and Araneae probe sets, Figures 1 and 2). Incremental testing of occupancy from 10% to 100% revealed that the probe set designed specifically for Araneoidea presented a more gradual loss of retention than the other probe sets, including the combined probe set (Figure 3A,B). The higher retention rate made possible by the specific probe sets produced a larger number of orthologous loci that only occurred in the targeted clade. In tree reconstruction, the tree topologies were consistent across the eight data matrices in the basal nodes and the node related to Araneoidea (Figure 4). Note, however, that the nodal supports (node 10, Figure 4) in the *Stegodyphus* + RTA (retrolateral tibial apophysis) clade were unstable, thereby supporting our claim that a specially designed probe set is necessary for a sub-lineage (e.g., the RTA clade).

Our in silico results showed that the numbers of captured loci using the combined probe set generally outperformed other probe sets. Among the specifically designed taxon probe sets, the Araneoidea probe set captured a larger number of loci in five of the six genomes used to develop the probe set. However, both probe sets performed poorly in the RTA clade (Figures 1 and 2). We detected 490.4 $\pm$ 464.7 loci (range = 28 to 1083) in the Araneoidea probe set, and 951.4 $\pm$ 974.4 loci (range = 28 to 2493) in the combined probe set. Note that the newly assembled draft genome for *Argyrodes miniaceus* returned a relatively low number of loci (182 and 142 in the combined and Araneoideae probe sets, respectively). The other Araneoidea genomes included in this study, which assembled in lower qualities (Table 1), generated <270 loci, thereby demonstrating that the number of captured loci was biased toward the well-assembled genomes used in the design of the probes. The Araneae and Arachnid probe sets generated low numbers of loci in Araneoidea genomes (lower than the Araneoidea probe set, except *Parasteatoda tepidariorum* and *Argyrodes miniaceus*). These showed a taxon-specific trend that the probe sets designed for higher taxonomic levels tended to capture fewer, and nearly insufficient, loci for phylogenomic analyses. Together with the results obtained using the four probe sets, we found that the quality and completeness of the genomes could have a deterministic effect on the number of captured loci. Moreover, the taxonomic group played a role in the number of captured loci, i.e., if there were no representative genomes in a clade, a low number of captured loci would be observed (see the RTA clade in Figure 1).

We did not observe large variations in the tree topologies reconstructed using different data matrices (i.e., with loci captured from different probe sets). However, the nodal supports dropped in both traditional bootstrap statistics and in the concordance factors (gCF and sCF) when there were no representative genomes used for probe design (i.e., *Stegodyphus* + RTA clade, in our case) (Figure 4). Within Araneoidea, nodes 3 and 5 did not perform well in gCF and sCF; however, the results still met an acceptable level of >33, thereby indicating a possible downside of using existing probe sets to resolve these nodes. Bootstrap values tended to generate optimistically high support, as observed in other phylogenomic studies [55]. In the current study, we used 50% and 75% occupancies to generate data matrices for phylogenomic analysis, with the mean number of loci varying from 35.5 to 203.6 per genome in 50% occupancy matrices, and a mean of 18.4 to 80.1 per genome in 75% occupancy matrices. The number of captured loci in silico was significantly lower than would be expected in real-world, in-solution captured data. The mean number of captured loci was 589.3 in the Arachnid probe set [23] and 553.71 in the Araneae probe set [11]. Our in silico results, in rough estimation, only captured up to 1/3 of the loci compared with

the in-solution captured results. We inferred that for in silico testing, we constrained the sequence identity to 80% for the capture of loci. From a practical perspective, in-solution capture could likely have required a lower degree of similarity to capture DNA fragments. As we aimed to relatively compare the numbers of the captured loci from different probe sets under the same in silico condition, we therefore expected to capture a larger number of loci when using our probe set for in-solution capture under laboratory conditions in the Araneoidea species.

5. Conclusions

This study designed specific probe sets using six genomes to facilitate the testing of phylogenetic hypotheses pertaining to Araneoidea. When using in silico capture, the data matrices generated using the combined and Araneoidea probes resolved most of the nodes in the sub-clades in Araneoidea, resulting in several hundred loci (relatively more loci compared with other non-targeted taxa). We expected that when conducting in-solution capture in a wet lab, it should be possible, using the estimated 1/3 in silico/in-solution ratio, to capture more than one thousand loci per genome. In our preliminary test using Argyrodinae as a targeted taxon, we captured 897.5 $\pm$ 62.9 loci per genome, which is about 4$\times$ the number captured in our in silico *Argyrodes miniaceus* results (182 or 142 loci). However, there are disadvantages to using this newly designed probe set, e.g., (1) fewer applicability to other taxa such as the RTA clade, and (2) potentially higher costs when synthesizing this customized probe and the combined probe sets. Moreover, our combined approach showed that it broadens the application of the probe sets given there are representative genomes collected from a sub-lineage. However, these approaches should be tested in a wet lab to validate the applicability of the Araneoidea and combined probe sets.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/d14030184/s1: all generated phylogenetic trees in this article, assembled genome statistics, and the raw data in Figures 1–4; Figure S1: Phylogenetic tree of the data set of the combined probe set filtered by 75% occupancy; Figure S2: Phylogenetic tree of the data set of the Arachnid probe set filtered by 50% occupancy; Figure S3: Phylogenetic tree of the data set of the Arachnid probe set filtered by 75% occupancy; Figure S4: Phylogenetic tree of the data set of the Araneae probe set filtered by 50% occupancy; Figure S5: Phylogenetic tree of the data set of the Araneae probe set filtered by 75% occupancy; Figure S6: Phylogenetic tree of the data set of the Araneoidea probe set filtered by 50% occupancy; Figure S7: Phylogenetic tree of the data set of the Araneoidea probe set filtered by 75% occupancy; Table S1: Statistics of genome and raw reads of Argyrodes miniaceus; Table S2: Results of probe set designed for Arachnid probe set in silico test on each genome; Table S3: Results of probe set designed for Araneae probe set in silico test on each genome; Table S4: Results of probe set designed for Araneoidea probe set in silico test on each genome; Table S5: Results of probe set designed for combined probe set in silico test on each genome; Table S6: Retention number of loci in data sets after filtering by different standards; Table S7: Loci number of each genome after filtering by occupancy.

Author Contributions: Conceptualization, Y.-C.S. and R.-C.C.; methodology, Y.-Y.L. and Y.-C.S.; software and hardware maintenance, J.-M.T. and Y.-C.W.; validation, C.-Y.W., Y.-F.C. and H.-Y.L.; formal analysis and data curation, Y.-Y.L., Y.-C.S. and C.-Y.W.; writing—original draft preparation, Y.-Y.L., Y.-F.C. and H.-Y.L.; writing—review and editing, Y.-C.S., R.-C.C. and N.W.; visualization, C.-Y.W. and Y.-Y.L.; funding acquisition, Y.-P.L.; supervision, project administration, and funding acquisition, Y.-C.S., R.-C.C. and N.W. All authors have read and agreed to the published version of the manuscript.

Funding: The funding sources are the Ministry of Science and Technology, Taiwan (110-2621-B-037 -001-MY3 and 107-2621-B-037-001-MY2 to Y.-C.S.; 108-2621-B-005-001 and 109-2621-B-005-003-MY2 to R.-C.C.), and the National Science and Technology Development Agency, Thailand (NSTDA: JRA-CO-2563-11148-TH to N.W.). Part of this research is from Y.-Y.L.'s thesis.

Institutional Review Board Statement: Our in silico study does not require Institutional Review Board approval as it does not involve human or vertebrate materials.

Data Availability Statement: The de novo assembled genome and the probe sets are available at GitHub: https://github.com/yiyenl/designing-probe-set-for-Araneoidea/blob/main/combine_probe.pl. Other visualized results are in the Supplementary Materials.

Acknowledgments: We thank all Ecology and Evolutionary Lab members at Kaohsiung Medical University, Taiwan.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barrett, C.F.; Bacon, C.D.; Antonelli, A.; Cano, Á.; Hofmann, T. An Introduction to Plant Phylogenomics with a Focus on Palms. *Bot. J. Linn. Soc.* **2016**, *182*, 234–255. [CrossRef]
2. Pettengill, J.B.; Luo, Y.; Davis, S.; Chen, Y.; Gonzalez-Escalona, N.; Ottesen, A.; Rand, H.; Allard, M.W.; Strain, E. An Evaluation of Alternative Methods for Constructing Phylogenies from Whole Genome Sequence Data: A Case Study with *Salmonella*. *PeerJ* **2014**, *2*, e620. [CrossRef]
3. Brewer, M.S.; Cotoras, D.D.; Croucher, P.J.P.; Gillespie, R.G. New Sequencing Technologies, the Development of Genomics Tools, and Their Applications in Evolutionary Arachnology. *J. Arachnol.* **2014**, *42*, 1–15. [CrossRef]
4. Giribet, G. New Animal Phylogeny: Future Challenges for Animal Phylogeny in the Age of Phylogenomics. *Org. Divers. Evol.* **2016**, *16*, 419–426. [CrossRef]
5. Hirsch, C.D.; Evans, J.; Buell, C.R.; Hirsch, C.N. Reduced Representation Approaches to Interrogate Genome Diversity in Large Repetitive Plant Genomes. *Brief. Funct. Genom.* **2014**, *13*, 257–267. [CrossRef]
6. Ekblom, R.; Galindo, J. Applications of next Generation Sequencing in Molecular Ecology of Non-Model Organisms. *Heredity* **2011**, *107*, 1–15. [CrossRef]
7. McCormack, J.E.; Faircloth, B.C.; Crawford, N.G.; Gowaty, P.A.; Brumfield, R.T.; Glenn, T.C. Ultraconserved Elements Are Novel Phylogenomic Markers That Resolve Placental Mammal Phylogeny When Combined with Species-Tree Analysis. *Genome Res.* **2012**, *22*, 746–754. [CrossRef]
8. Mamanova, L.; Coffey, A.J.; Scott, C.E.; Kozarewa, I.; Turner, E.H.; Kumar, A.; Howard, E.; Shendure, J.; Turner, D.J. Target-Enrichment Strategies for next-Generation Sequencing. *Nat. Methods* **2010**, *7*, 111–118. [CrossRef]
9. Faircloth, B.C. Identifying Conserved Genomic Elements and Designing Universal Bait Sets to Enrich Them. *Methods Ecol. Evol.* **2017**, *8*, 1103–1112. [CrossRef]
10. Zhang, Y.M.; Williams, J.L.; Lucky, A. Understanding UCEs: A Comprehensive Primer on Using Ultraconserved Elements for Arthropod Phylogenomics. *Insect Syst. Divers.* **2019**, *3*, 3. [CrossRef]
11. 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*, 185–203. [CrossRef]
12. Ryu, T.; Seridi, L.; Ravasi, T. The Evolution of Ultraconserved Elements with Different Phylogenetic Origins. *BMC Evol. Biol.* **2012**, *12*, 236. [CrossRef]
13. Bejerano, G.; Pheasant, M.; Makunin, I.; Stephen, S.; Kent, W.J.; Mattick, J.S.; Haussler, D. Ultraconserved Elements in the Human Genome. *Science* **2004**, *304*, 1321–1325. [CrossRef]
14. Siepel, A.; Bejerano, G.; Pedersen, J.S.; Hinrichs, A.S.; Hou, M.; Rosenbloom, K.; Clawson, H.; Spieth, J.; Hillier, L.W.; Richards, S.; et al. Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes. *Genome Res.* **2005**, *15*, 1034–1050. [CrossRef]
15. Habic, A.; Mattick, J.S.; Calin, G.A.; Krese, R.; Konc, J.; Kunej, T. Genetic Variations of Ultraconserved Elements in the Human Genome. *OMICS A J. Integr. Biol.* **2019**, *23*, 549–559. [CrossRef]
16. Van Dam, M.H.; Henderson, J.B.; Esposito, L.; Trautwein, M. Genomic Characterization and Curation of UCEs Improves Species Tree Reconstruction. *Syst. Biol.* **2021**, *70*, 307–321. [CrossRef]
17. Hedin, M.; Derkarabetian, S.; Alfaro, A.; Ramírez, M.J.; Bond, J.E. Phylogenomic Analysis and Revised Classification of Atypoid Mygalomorph Spiders (*Araneae*, *Mygalomorphae*), with Notes on Arachnid Ultraconserved Element Loci. *PeerJ* **2019**, *7*, e6864. [CrossRef]
18. Faircloth, B.C.; McCormack, J.E.; Crawford, N.G.; Harvey, M.G.; Brumfield, R.T.; Glenn, T.C. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Syst. Biol.* **2012**, *61*, 717–726. [CrossRef]
19. Thom, G.; Amaral, F.R.D.; Hickerson, M.J.; Aleixo, A.; Araujo-Silva, L.E.; Ribas, C.C.; Choueri, E.; Miyaki, C.Y. Phenotypic and Genetic Structure Support Gene Flow Generating Gene Tree Discordances in an Amazonian Floodplain Endemic Species. *Syst. Biol.* **2018**, *67*, 700–718. [CrossRef]
20. Winker, K.; Glenn, T.C.; Faircloth, B.C. Ultraconserved Elements (UCEs) Illuminate the Population Genomics of a Recent, High-Latitude Avian Speciation Event. *PeerJ* **2018**, *6*, e5735. [CrossRef]

21. Meiklejohn, K.A.; Faircloth, B.C.; Glenn, T.C.; Kimball, R.T.; Braun, E.L. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. *Syst. Biol.* **2016**, *65*, 612–627. [CrossRef] [PubMed]
22. Bossert, S.; Murray, E.A.; Pauly, A.; Chernyshov, K.; Brady, S.G.; Danforth, B.N. Gene Tree Estimation Error with Ultraconserved Elements: An Empirical Study on *Pseudapis* Bees. *Syst. Biol.* **2021**, *70*, 803–821. [CrossRef] [PubMed]
23. Starrett, J.; Derkarabetian, S.; Hedin, M.; Bryson, R.W.; McCormack, J.E.; Faircloth, B.C. High Phylogenetic Utility of an Ultraconserved Element Probe Set Designed for Arachnida. *Mol. Ecol. Resour.* **2017**, *17*, 812–823. [CrossRef] [PubMed]
24. Gloor, D.; Nentwig, W.; Blick, T.; Kropf, C. World Spider Catalog. *Nat. Hist. Mus. Bern.* **2022**. [CrossRef]
25. Selden, P.A.; Shear, W.A.; Sutton, M.D. Fossil Evidence for the Origin of Spider Spinnerets, and a Proposed Arachnid Order. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20781–20785. [CrossRef]
26. Garrison, N.L.; Rodriguez, J.; Agnarsson, I.; Coddington, J.A.; Griswold, C.E.; Hamilton, C.A.; Hedin, M.; Kocot, K.M.; Ledford, J.M.; Bond, J.E. Spider Phylogenomics: Untangling the Spider Tree of Life. *PeerJ* **2016**, *4*, e1719. [CrossRef]
27. Xu, X.; Su, Y.-C.; Ho, S.Y.W.; Kuntner, M.; Ono, H.; Liu, F.; Chang, C.-C.; Warrit, N.; Sivayyapram, V.; Aung, K.P.P.; et al. Phylogenomic Analysis of Ultraconserved Elements Resolves the Evolutionary and Biogeographic History of Segmented Trapdoor Spiders. *Syst. Biol.* **2021**, *70*, 1110–1122. [CrossRef]
28. Dimitrov, D.; Benavides, L.R.; Arnedo, M.A.; Giribet, G.; Griswold, C.E.; Scharff, N.; Hormiga, G. Rounding up the Usual Suspects: A Standard Target-Gene Approach for Resolving the Interfamilial Phylogenetic Relationships of Ecribellate Orb-Weaving Spiders with a New Family-Rank Classification (*Araneae*, *Araneoidea*). *Cladistics* **2017**, *33*, 221–250. [CrossRef]
29. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
30. Chikhi, R.; Medvedev, P. Informed and Automated K-Mer Size Selection for Genome Assembly. *Bioinformatics* **2014**, *30*, 31–37. [CrossRef]
31. Jackman, S.D.; Vandervalk, B.P.; Mohamadi, H.; Chu, J.; Yeo, S.; Hammond, S.A.; Jahesh, G.; Khan, H.; Coombe, L.; Warren, R.L.; et al. ABySS 2.0: Resource-Efficient Assembly of Large Genomes Using a Bloom Filter. *Genome Res.* **2017**, *27*, 768–777. [CrossRef] [PubMed]
32. Sanggaard, K.W.; Bechsgaard, J.S.; Fang, X.; Duan, J.; Dyrlund, T.F.; Gupta, V.; Jiang, X.; Cheng, L.; Fan, D.; Feng, Y.; et al. Spider Genomes Provide Insight into Composition and Evolution of Venom and Silk. *Nat. Commun.* **2014**, *5*, 3765. [CrossRef] [PubMed]
33. Purcell, J.; Pruitt, J.N. Are Personalities Genetically Determined? Inferences from Subsocial Spiders. *BMC Genom.* **2019**, *20*, 867. [CrossRef] [PubMed]
34. Kono, N.; Nakamura, H.; Ohtoshi, R.; Moran, D.A.P.; Shinohara, A.; Yoshida, Y.; Fujiwara, M.; Mori, M.; Tomita, M.; Arakawa, K. Orb-Weaving Spider Araneus Ventricosus Genome Elucidates the Spidroin Gene Catalogue. *Sci. Rep.* **2019**, *9*, 8380. [CrossRef] [PubMed]
35. Sheffer, M.M.; Hoppe, A.; Krehenwinkel, H.; Uhl, G.; Kuss, A.W.; Jensen, L.; Jensen, C.; Gillespie, R.G.; Hoff, K.J.; Prost, S. Chromosome-Level Reference Genome of the European Wasp Spider *Argiope Bruennichi*: A Resource for Studies on Range Expansion and Evolutionary Adaptation. *GigaScience* **2021**, *10*, giaa148. [CrossRef] [PubMed]
36. Wellcome Sanger Institute. 25 Genomes for 25 Years. Available online: https://www.sanger.ac.uk/collaboration/25-genomes-for-25-years/ (accessed on 16 December 2021).
37. Sánchez-Herrero, J.F.; Frías-López, C.; Escuer, P.; Hinojosa-Alvarez, S.; Arnedo, M.A.; Sánchez-Gracia, A.; Rozas, J. The Draft Genome Sequence of the Spider *Dysdera Silvatica* (*Araneae*, *Dysderidae*): A Valuable Resource for Functional and Evolutionary Genomic Studies in Chelicerates. *GigaScience* **2019**, *8*, giz099. [CrossRef]
38. i5K Consortium. The I5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *J. Hered.* **2013**, *104*, 595–600. [CrossRef]
39. Hendrickx, F.; De Corte, Z.; Sonet, G.; Van Belleghem, S.M.; Köstlbacher, S.; Vangestel, C. A Masculinizing Supergene Underlies an Exaggerated Male Reproductive Morph in a Spider. *Nat. Ecol. Evol.* **2022**, *6*, 195–206. [CrossRef]
40. Schwager, E.E.; Sharma, P.P.; Clarke, T.; Leite, D.J.; Wierschin, T.; Pechmann, M.; Akiyama-Oda, Y.; Esposito, L.; Bechsgaard, J.; Bilde, T.; et al. The House Spider Genome Reveals an Ancient Whole-Genome Duplication during Arachnid Evolution. *BMC Biol.* **2017**, *15*, 62. [CrossRef]
41. Yu, N.; Li, J.; Liu, M.; Huang, L.; Bao, H.; Yang, Z.; Zhang, Y.; Gao, H.; Wang, Z.; Yang, Y.; et al. Genome Sequencing and Neurotoxin Diversity of a Wandering Spider Pardosa Pseudoannulata (Pond Wolf Spider). *BioRxiv.* **2019**. [CrossRef]
42. Liu, S.; Aagaard, A.; Bechsgaard, J.; Bilde, T. DNA Methylation Patterns in the Social Spider, *Stegodyphus Dumicola*. *Genes* **2019**, *10*, 137. [CrossRef] [PubMed]
43. Babb, P.L.; Lahens, N.F.; Correa-Garhwal, S.M.; Nicholson, D.N.; Kim, E.J.; Hogenesch, J.B.; Kuntner, M.; Higgins, L.; Hayashi, C.Y.; Agnarsson, I.; et al. The Nephila Clavipes Genome Highlights the Diversity of Spider Silk Genes and Their Complex Expression. *Nat. Genet.* **2017**, *49*, 895–903. [CrossRef] [PubMed]
44. Faircloth, B.C. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* **2016**, *32*, 786–788. [CrossRef] [PubMed]
45. Huang, W.; Li, L.; Myers, J.R.; Marth, G.T. ART: A next-Generation Sequencing Read Simulator. *Bioinformatics* **2012**, *28*, 593–594. [CrossRef]

46. Lunter, G.; Goodson, M. Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads. *Genome Res.* **2011**, *21*, 936–939. [CrossRef]
47. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
48. Quinlan, A.R.; Hall, I.M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
49. Hipp, R.D. SQLite Home Page. Available online: https://www.sqlite.org/index.html (accessed on 25 December 2020).
50. Harris, R.S. *Improved Pairwise Alignment of Genomic DNA*; The Pennsylvania State University: State College, PA, USA, 2007.
51. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
52. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552. [CrossRef]
53. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef]
54. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [CrossRef] [PubMed]
55. Minh, B.Q.; Hahn, M.W.; Lanfear, R. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Mol. Biol. Evol.* **2020**, *37*, 2727–2733. [CrossRef] [PubMed]
56. Wheeler, W.C.; Coddington, J.A.; Crowley, L.M.; Dimitrov, D.; Goloboff, P.A.; Griswold, C.E.; Hormiga, G.; Prendini, L.; Ramírez, M.J.; Sierwald, P.; et al. The Spider Tree of Life: Phylogeny of Araneae Based on Target-Gene Analyses from an Extensive Taxon Sampling. *Cladistics* **2017**, *33*, 574–616. [CrossRef]
57. Rambaut, A. FigTree v1.4.4. Available online: https://github.com/rambaut/figtree/releases (accessed on 25 October 2021).
58. Kallal, R.J.; Kulkarni, S.S.; Dimitrov, D.; Benavides, L.R.; Arnedo, M.A.; Giribet, G.; Hormiga, G. Converging on the Orb: Denser Taxon Sampling Elucidates Spider Phylogeny and New Analytical Methods Support Repeated Evolution of the Orb Web. *Cladistics* **2021**, *37*, 298–316. [CrossRef] [PubMed]