Molecular evolution of venom proteins in Ctenidae (Order: Araneae) spiders

by

T. Jeffrey Cole

May, 2021

Director of Dissertation: Michael Scott Brewer

Major Department: Biology

Spiders comprise the largest group of venomous animals and are a pivotal component of the global ecosystem with approximately 50,000 species spread across nearly every habitat on Earth. The family Ctenidae Keyserling, 1877 comprises small to large nocturnal wandering spiders. There are drastic differences in the venom potency amongst wandering spiders. The bite of the highly aggressive Brazilian wandering spider (Phoneutria nigriventer) causes pain, cramps, priapism, and arrhythmia, whereas the bite of ctenids dwelling in the temperate forests of North America have no recorded adverse symptoms. Inhibitor Cystine Knot toxins (ICKs) make up the majority of the venom composition across the spider tree of life and exist as complex multi-copy gene families in which one species may express up to 100 homologs.

To develop an understanding of venom evolution in wandering spiders, we reconstructed a species-level phylogeny using transcriptomic sequences and conducted feeding experiments to evaluate the variation in venom biology of ctenid spiders in the U.S. We found that U.S. ctenids do not represent a single lineage, and that venom utilization strategies differed within the family.

To characterize the patterns of molecular evolution of ICK toxins in wandering spiders. To do this, we used venom gland transcriptomics and proteomics to identify and characterize

the molecular evolution of 626 unique coding sequences of ICK peptides. No amino acid sites were found to have evidence of pervasive positive selection, though 12 out of 80 sites contained a portion of branches within the gene family phylogeny with evidence of episodic positive selection.

The final objective was to compare the genomic architecture and patterns of duplication in spiders with publicly available genome sequences. We demonstrated that the current state of spider genome assemblies presents limitations in the analyses that can be performed We were not able to identify many ICKs in the assemblies of all species but were able to construct a phylogeny of ICKs present as duplicates in three species. Through this investigation, we revealed that ICKs have undergone duplication events before and after speciation events, indicating these events have occurred throughout the evolutionary history of spiders.

Molecular evolution of venom proteins in Ctenidae (Order: Araneae) spiders

A Dissertation

Presented to the Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy in Interdisciplinary Biological Sciences

by

Timothy Jeffrey Cole

May, 2021

Molecular evolution of venom proteins in Ctenidae (Order: Araneae) spiders

by

Timothy Jeffrey Cole

APPROVED BY:

DIRECTOR OF
DISSERTATION

Michael Brewer, PhD

COMMITTEE MEMBER:

Christopher Balakrishnan, PhD

COMMITTEE MEMBER:

Alfred Lamb, PhD

COMMITTEE MEMBER:

Jessica Garb, PhD

CHAIR OF THE DEPARTMENT
OF
Biology

Dave Chalcraft, PhD

DEAN OF THE
GRADUATE SCHOOL:

Paul J. Gemperline, PhD

Dedication

I dedicate this dissertation to my father, who passed away two days before I graduated high school. He gave me three things that I carry with me proudly each and every day. First, he gave me his name. He started a stucco and stone construction company when I was in elementary school by the name of “TIJECO”, which was an abbreviation of our name “TImothy JEffrey COle”. When he passed away the business was shut down, but I carried the name tijeco on as my GitHub username. Second, he instilled in me a love for nature. Many weekends we walked in the woods and he would show me the different species of trees on our land. We would go to the widest tree on our property, which was a loblolly pine so wide that our arms could not wrap all the way around it. I often thought of him during my travels to some incredibly beautiful parts of the southeastern United States, where I walked in the woods to collect spiders for my dissertation. Third, he taught me the value of problem solving. He did not have a high school diploma, but he was among the smartest people I have ever met. When my parents first moved out to the land I grew up on, there was an easement dispute that prevented them from getting municipal electricity for nearly a year. During that year, he used the alternator from a lawnmower his brother had given him as a gag-gift to charge an array of car batteries he picked up at a junkyard, which he used to wire a series of taillight bulbs throughout their home so that they could flip on a switch to have light in each room. When presented with a problem, he was never afraid to think outside the box and be resourceful to find a solution. Throughout my dissertation, I carried that lesson with me to be as creative as possible to solve problems no one has ever solved before.

Acknowledgements

This research was made possible by the National Science Foundation Graduate Research Fellowship, the American Museum for Natural History Theodore Roosevelt travel grant, as well as East Carolina University's Department of Biology startup funds for the Brewer Lab. I would also like to thank my advisor Michael Brewer, and committee members Christopher Balakrishnan, Trip Lamb, and Jessica Garb for their support and advisement throughout the entirety of my dissertation. I would also like to thank collaborators Drew Hataway, Brad Bennet, and Antonio Brescovit for providing tissue samples from Brazil. I'm grateful to my lab-mate Chris Cohen as well as fellow graduate students Xinjun Wu and Tim McDaniel for fieldwork assistance. I would also like to acknowledge the amazing undergraduate assistants that worked with behavioural experiments: Aayushi Patel, Chase Magsig, Radha Patel, and Khaleb Enoch. I am eternally grateful to my mother for providing me a place to stay during fieldwork, and my beautiful girlfriend Nora for providing edits on the dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Phylogenomics and venom biology of wandering spiders of temperate zone North America (Araneae: Ctenidae)

1.1 Introduction

The family Ctenidae Keyserling, 1877 comprises small to large nocturnal wandering spiders that are fast-moving ground hunting predators that make little use of silk for prey capture[1,2]. They typically inhabit forest floors and understory vegetation[3,4], while a small number are arboreal[5]. Of the 117 currently described spider families, Ctenidae is the 20th most diverse family in Araneae with 533 species distributed in 48 genera[6]. Much of this diversity lies in the neotropics, approximately 48% of the species in this family[6]. Many of these species are unique to forest habitats, so, unfortunately, deforestation and habitat fragmentation have had a significant impact on their densities[7,8].

The most notorious members of this family belong to the genus *Phoneutria* Perty, 1833, often referred to as “armed spiders” or “banana spiders”. They are known to be aggressive and are considered to be among the most medically important spiders in the world due to their highly noxious venom with neurotoxic action[9]. In 2006, bites from members of this genus were responsible for 2,687 hospitalizations in Brazil[10]. Despite this, the degree to which noxious venom is expressed in other members of this family is poorly understood.

Part of the lack of understanding of the venom biology of other members of this family is related to its taxonomic history. Historically, investigations of ctenid taxa were hindered by inadequate original descriptions and little to no knowledge of evolutionary relationships; however, in the past few decades advances in taxonomic revision and phylogenetic analysis have greatly improved knowledge regarding the evolutionary histories of the family[3,11–18]. Much work is still needed, however, as most of the species (˜40%) in the family are assigned to the genus *Ctenus* Walckenaer, 1805, which historically has been a repository for ill-defined species[13] and is considered to be polyphyletic[3,14,17,19].

Unlike their close relatives within the Lycosidae Sundevall, 1833, which are widespread throughout New World tropical and temperate forests[20], the presence of ctenids in the North American temperate zone is sparse and limited (Figure 1.1). A particularly interesting assemblage of ctenids that has colonized the North American temperate zone comprises four species that are currently assigned to the genus *Ctenus* Keyserling, 1877. All species in this assemblage have a fairly localized and allopatric geographic distribution (Figure 1.1). Additionally, all species inhabit caves or karst terrain, but as wanderers, they can also be found in other habitats[20]. *Ctenus hibernalis* Hentz, 1844 is typically restricted to the southern foothills of the Appalachian mountains in Alabama[21]. It frequently inhabits deciduous forest floors, under logs and detritus, and has been found in numerous caves. *Ctenus captiosus* Gertsch, 1935 is only found in central Florida and inhabits mesic hammock forests under pine bark and detritus and some small cavities[22]. Though it is similar to *C. hibernalis* in size and color, it can be readily distinguished by its genitalia. *Ctenus exlineae* Peck, 1981 is only found in northwestern Arkansas in the Ouachita and Ozark forests in large populations in locally restricted habitats[20]. *C. exlineae* is easily distinguished from *C. hibernalis* and *C. captiosus* by its reddish-brown coloration (Figure 1.2). Though it has not been recorded in caves specifically, *C. exlineae* has been known to inhabit southern facing talus slopes in karst terrain with numerous small cavities, but rarely under logs. The final temperate North American member of this genus, *C. valverdiensis* Peck, 1981, is somewhat of an anomaly compared to the other three members of this assemblage[20]. There are only two recorded localities for this species in south-central Texas; one is a small cave along the side of a highway, and the other is an abandoned mine shaft just thirty kilometers north of that. This species is over 1,500 kilometers from the nearest range of the other three species in this assemblage, which themselves are only around 600 kilometers apart[23]. Due to their restricted locality, *C. valverdiensis* will be henceforth excluded from all mentions of U.S. ctenids.

Apart from the aforementioned *Ctenus* assemblage, there are only two other species of Ctenidae in the North American temperate zone, in the genera *Leptoctenus* Koch, 1878 and *Anahita* Karsch, 1879. *Leptoctenus byrrhus* Simon, 1888 is only found as far north as the hill country of central Texas and as far south as the coastal plains and mountains of northeast Mexico[24]. They have been collected around boulders and detritus in cedar stands along granite and limestone outcroppings[20]. *Anahita punctulata* is the smallest of the temperate North American species and has the largest distribution[21]. It has been found as far north as the hills of southern Indiana but is primarily confined to the mesic forests of the southeastern states in the U.S., overlapping with *C. hibernalis*[20].

Venom is a foraging strategy for spiders, so it has the benefit of a high energy reward for successful prey capture[25]. In spiders, venoms generally have evolved to immobilize their prey, so it is not always the case that the venom is capable of outright killing their prey[26]. This is likely due to the energetic costs of using venom to kill prey rather than merely immobilizing them. For free-hunting spiders, such as ctenids, there are seven steps for prey capture: recognition, orientation, approach, grasp, envenomation, prey paralysis, and digestion[27]. Wandering spiders do not build webs so a speedy grasp is vital to ensuring a successful prey capture. Beyond a speedy grasp, envenomation location is also critical for successful

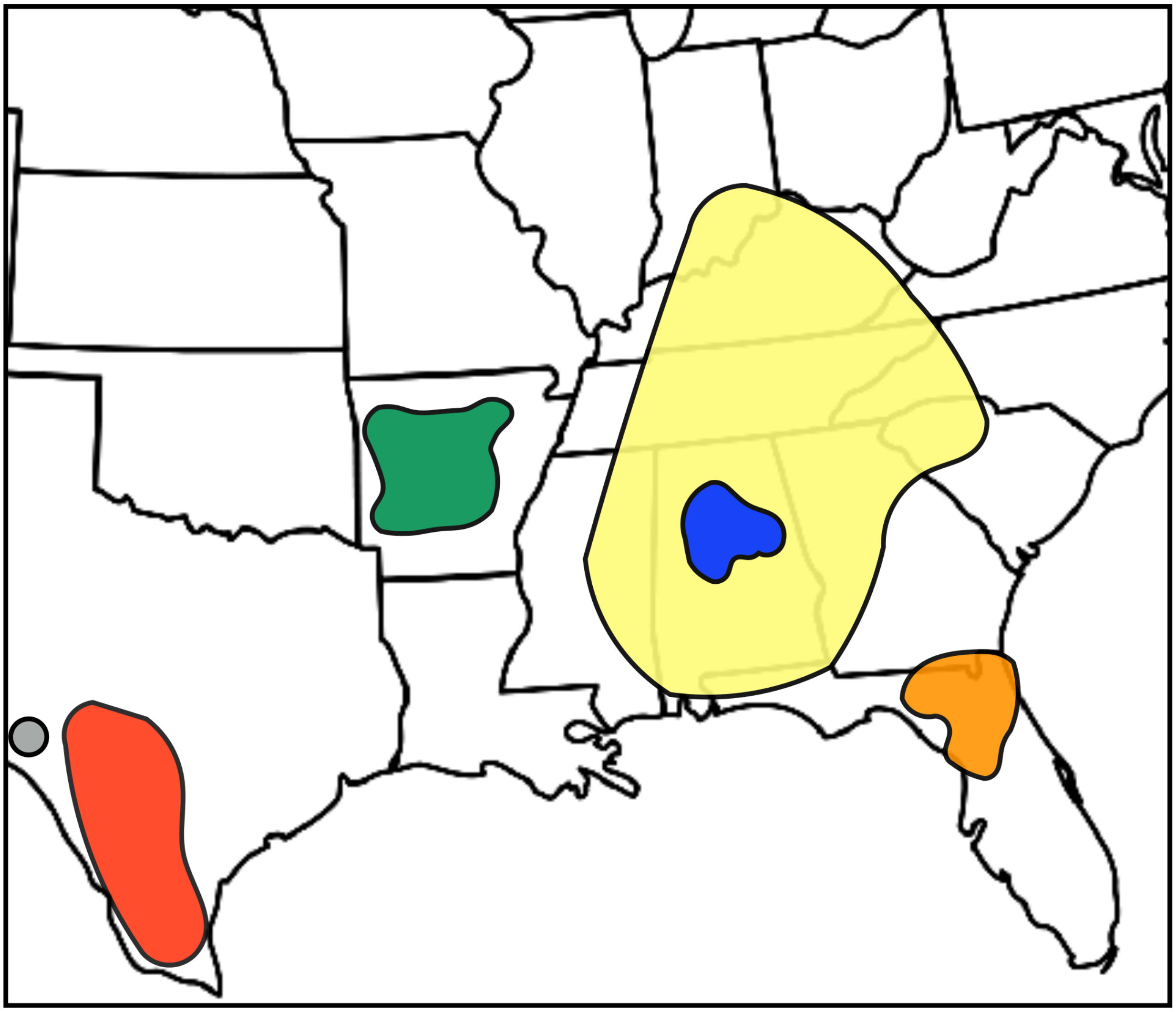

Figure 1.1: Distribution of U.S. ctenids. *Leptoctenus byrrhus* (red), *Ctenus exlineae* (green), *Anahita punctulata* (yellow), *Ctenus hibernalis* (blue), *Ctenus captiosus* (orange), *Ctenus valverdiensis* (gray) .

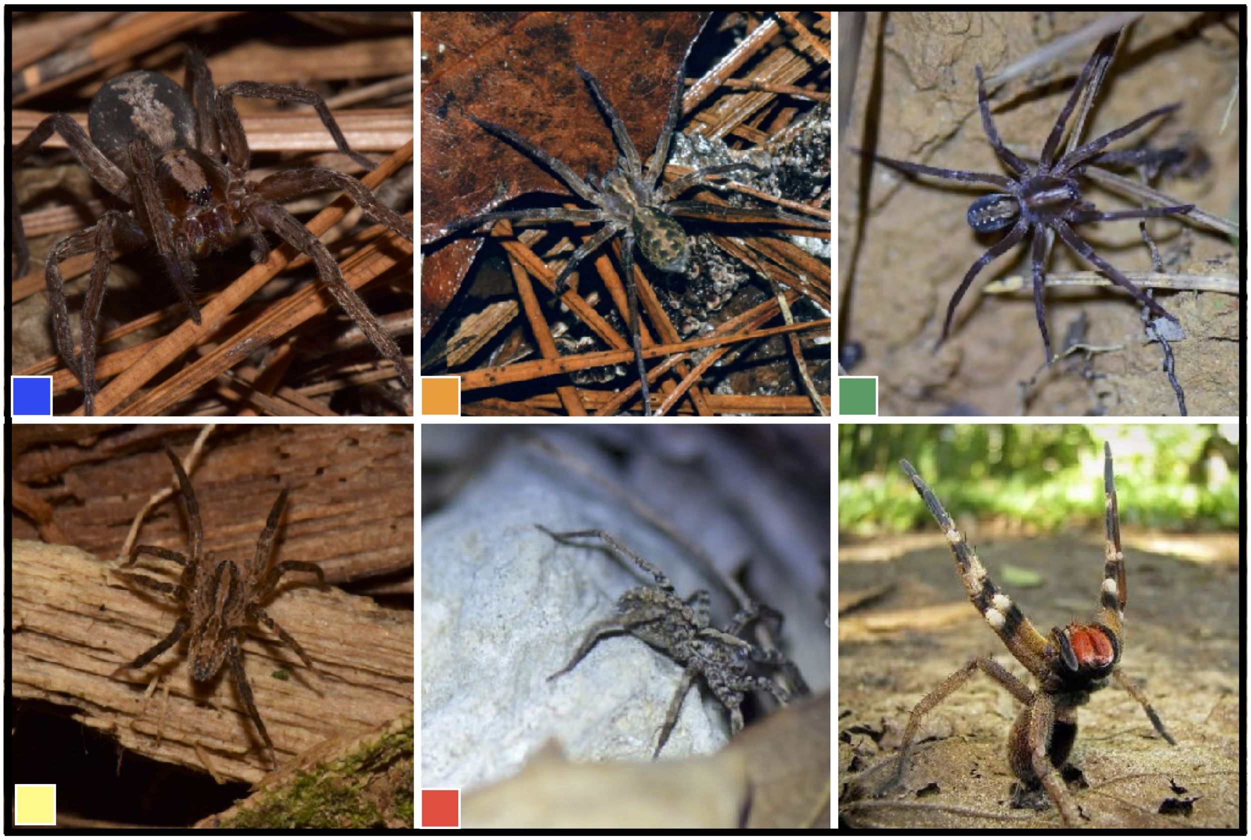

Figure 1.2: Photographs of adult female U.S. ctenids (colored box at the bottom left corresponds to species code from Figure 1, the photos were taken by T. Cole), and *Phoneutria nigriventer* (bottom right, the photo was taken by João Burini).

paralysis, since the venom acts as a neurotoxin it serves its purpose most effectively when injected near nervous tissues. Prey paralysis is thus an especially critical step in this process. If the venom is not sufficiently potent or not enough is delivered, then the prey can escape. Understanding the variation in attack speeds and prey paralysis efficiency that exists between spider species can provide insights into the underlying drivers of venom evolution. To do this in ctenid spiders, a well supported phylogeny is first needed to provide an evolutionary framework. In this study, we reconstructed a species-level phylogeny using transcriptomic sequences and conducted feeding experiments to evaluate the variation in venom biology of ctenid spiders in the U.S.

1.2 Methods

Much of the phylogenetic work in Ctenidae has focused on using morphological characters. Unfortunately, none of these phylogenetic treatments of the family have included all of the species found in North America, so there is currently no phylogenetic hypothesis regarding those species and their neotropical counterparts[16–18]. Inferring accurate phylogenies solely with morphological characters is problematic, however, because of the complex processes in which morphology evolves and the high amount of convergent evolution, parallel evolution, and trait reversals that occur[28]. While phylogenies reconstructed using a small handful of genetic loci provide great improvements in accuracy over morphological data, a genome-wide multi-locus approach is ideal in terms of maximizing resolving power and phylogenetic signal[29].

Taxon sampling: Ctenidae belongs to the superfamily Lycosoidea, which is a member of a clade of spiders that possess a retrolateral tibial apophysis (RTA), a backward-facing projection on the tibia of the male pedipalp. The RTA-clade is a highly diverse group of spiders with higher diversification rates than any other group of spiders[30]. Although relationships of superfamilies deep within the spider tree of life have improved greatly through phylogenomic efforts[30–32], lycosoid relationships remain elusive. Five recent studies have addressed lycosoid relationships using traditional molecular markers[31,33–36], and each reconstructed different relationships. More recently, however, Cheng & Piel [37] utilized transcriptomic data to reconstruct a highly supported lycosoid phylogeny. Thus, the same outgroups used by Cheng & Piel [37] were used in this analysis. A list of all species previously mentioned, and their localities are detailed in (Table 1.1).

Three adult males and females were collected from all members of Ctenidae in the United States, excluding *C. valverdiensis*. *C. hibernalis*, *C. exlineae*, *C. captiosis*, *L. byrrhus*, and *A. punctula* were collected from the following respective localities: North-Central Alabama, Northwest Arkansas, Central Florida, South-Central Texas, and Northwest Georgia. Four male and female *Phoneutria nigriventer* (Keyserling, 1891) along with four female *Isoctenus sp.* were collected from Brazil.

Table 1.1: Transcriptomes used in this phylogenetic analysis. Samples contributed by this study are designated by "novel contribution" in the SRA Run Accession column, otherwise are abbreviated as C, (Cheng & Piel [37]); B, (Bond *et al.* [38]); Fe, (Fernandez *et al.* [32]); G, (Garrison *et al.* [30]); M, (Meng *et al.* [39]).

Species	Family	SRA Run Accession	Tissue	Specimen Origin
Anahita punctulata	Ctenidae	SRR3144072, G	Whole body	Auburn, AL, USA
Anahita punctulata	Ctenidae	Novel contribution	Venom gland	Cave Spring, GA, USA
Ctenus captiosus	Ctenidae	Novel contribution	Venom gland	Ocala, FL, USA

Species	Family	SRA Run Accession	Tissue	Specimen Origin
Ctenus corniger	Ctenidae	SRR6360557, C	Whole body	Singapore
Ctenus exlineae	Ctenidae	Novel contribution	Venom gland	Langley, AR, USA
Ctenus hibernalis	Ctenidae	Novel contribution	Venom gland	Homewood, AL, USA
Isoctenus sp.	Ctenidae	Novel contribution	Venom gland	Brazil
Leptoctenus byrrhus	Ctenidae	Novel contribution	Venom gland	Vanderpool, TX, USA
*Phoneutria nigrivente*r	Ctenidae	Novel contribution	Venom gland	Brazil
Dolomedes triton	Pisauridae	SRR3144094, G	Whole body	Opelika, AL, USA
Fecenia protensa	Psechridae	SRR6360558, C	Whole body	Singapore
Habronattus signatus	Salticidae	SRR1514888, B	Whole body	Ocotillo, CA, USA
Hibana sp.	Anyphaenidae	SRR3144074, G	Whole body	Auburn, AL, USA
Hippasa holmerae sundaica	Lycosidae	SRR6360559, C	Whole body	Singapore
Homalonychus theologus	Homalonychidae	SRR3144075, G	Whole body	Imperial Co, CA, USA

Species	Family	SRA Run Accession	Tissue	Specimen Origin
Misumenoides formosipes	Thomisidae	SRR3144080, G	Whole body	Opelika, AL, USA
Nilus albocinctus	Pisauridae	SRR6360560, C	Whole body	Singapore
Odo patricius	Xenoctenidae	SRR6360553, C	Whole body	Iquique, Chile
Oxyopes sp.	Oxyopidae	SRR6360554, C	Whole body	Singapore
*Pardosa pseudoannulat*a	Lycosidae	SRR1833279, M	Whole body	Nanjing, China
Peucetia longipalpis	Oxyopidae	SRR1514898, B	Whole body	Opelika, AL, USA
Pisaurina mira	Pisauridae	SRR1365651, Fe	Whole body	University Park, MD
Psechrus singaporensis	Psechridae	SRR6360555, C	Whole body	Singapore
*Schizocosa rovner*i	Lycosidae	SRR1514894, B	Whole body	Oxford, MS, USA
*Sergiolus capulatu*s	Gnaphosidae	SRR1514903, B	Whole body	Opelika, AL, USA
Sosippus placidus	Lycosidae	SRR6360556, C	Whole body	Placid Lakes, FL, USA
Sphedanus quadrimaculatus	Pisauridae	SRR6360561, C	Whole body	Singapore

		SRA Run		
Species	Family	Accession	Tissue	Specimen Origin
Thomisus spectabilis	Thomisidae	SRR6360562, C	Whole body	Singapore

Locus sampling: Whole RNA was isolated from the venom glands from the five U.S species and the two Brazilian species using TRIzol® (Life Technologies, Carlsbad, CA) and the Qiagen RNeasy kit (Qiagen, Valencia, CA). RNA concentration and integrity was evaluated using Quant-iT™ PicoGreen and Bioanalyzer. The RNA extraction was sent to the Genomic Services Lab at HudsonAlpha (Huntsville, AL) for library preparation with poly(A) selection and sequencing on a 100 bp paired-end run on an Illumina HiSeq 2500, comprising 25 million reads forward and reverse (50 million total reads) per sample. For all samples sequenced after 2018, due to sequence facility updates, the same setup was used but with Illumina NovaSeq. Additionally, RNAseq reads from all outgroup species were retrieved from NCBI Short Read Archive.

Prior to assembly, removal of adapters, correcting sequence errors, and trimming low-quality base calls ensured maximal accuracy of transcript recovery[40]. All read processing was executed with FASTP v 0.19.6[41]. De novo assemblies typically recover an unexpectedly large number of transcripts, sometimes well over 100,000[42]. This happens for three main reasons. First, an increased depth of sequencing combined with improved transcript recovery algorithms increases the recovery of transcripts that are expressed at levels lower than would otherwise be considered biologically relevant. Second, common contaminants, such as bacteria and fungi, unavoidably make their way into samples, thus inflating the mRNA transcript pool diversity. Third, in eukaryotic systems, alternative splicing yields a significant increase in recoverable transcripts per gene locus as isoforms. Further, when several RNA-seq experiments from different species are sequenced together, cross contamination inevitably occurs[43–49]. The aforementioned issues can have tremendous effects on downstream phylogenetic inferences made from problematic transcripts[50]. To alleviate these issues, reads were first mapped to a transcript database of common contaminants (bacteria, fungi, human, and nematodes) using salmon v1.3.0 with default mapping parameters, all unmapped reads were retained as the finalized library of processed reads[51]. The processed reads were de novo assembled into transcripts using TRINITY v2.8.4[52]. TRINITY utilizes the following three-step approach to assemble transcripts de novo: First, it assembles unique portions of alternatively spliced transcripts. Second, it clusters those contiguous sequences and constructs de Bruijn graphs. Finally, it processes each graph to assemble full-length transcripts for alternatively spliced isoforms.

Coding sequences within transcripts were inferred using TRANSDECODER v3.0.1[53]. TRANSDECODER uses the following criteria to identify the single best coding sequence in a given transcript: available open reading frame (ORF) of a minimum length of 30 codons, log-likelihood score of the coding sequence, predictions of start and stop codons as refined by a Position-Specific Scoring Matrix. The completeness of the assemblies was evaluated using BUSCO v3.0.2 (Benchmarking Universal Single-Copy Orthologs)[54] with the arthropod database. In addition to providing a metric of assembly completeness, complete BUSCO transcripts serve as a robust set of loci for phylogenetic analysis.

Phylogenetic reconstruction: A well supported phylogeny provides a necessary evolutionary framework for comparative analysis of venom evolution. To reconstruct the North American ctenid species relationships, additional RNAseq reads from 20 outgroup species were retrieved from NCBI Short Read Archive. Loci sampling for phylogenetic analysis involved the following procedure. Only complete coding sequences inferred from TRANSDECODER that were the longest isoform of a given TRINITY gene assignment were used for this analysis. Coding sequences that contained a single complete match to a BUSCO term were retrieved from the assemblies. Multiple protein alignments were generated with MAFFT v7.221[55] for BUSCO matches and retained if 30 out of 48 samples were present and <50% of the sequences were identical. This resulted in 245 BUSCO term alignments that were then trimmed with TRIMAL v1.4.1[56] to remove uninformative sites. Model selection was performed for the trimmed concatenated matrix to elucidate the best-fit model using the "TESTONLY" option of IQ-TREE v1.6.10[57]. IQ-TREE then generated 1,000 random starting trees, and maintained a pool of 100 candidate trees during subsequent analysis. IQ-TREE then randomly selected one of those trees and applied a random Nearest Neighbor Interchange (NNI), then it used the resultant tree to initiate an NNI-based hill-climbing tree search. Modifications that improved the tree likelihood resulted in the prior tree being replaced in the tree pool. Otherwise, the modification was discarded and the analysis terminated after 1,000 unsuccessful iterations resulting in the final tree. Subsequently, 1,000 ultrafast bootstrap replicates were performed to calculate node support. For the species tree analysis, the phylogenies of each gene tree was reconstructed using IQTREE with the same settings as previously mentioned[57]. The gene trees were then provided as input for ASTRAL v2.0[58] to reconstruct the species phylogeny.

Feeding behavior observations: Adult females from the following ctenid species were used for these experiments: *L. byrrhus*, *C. hibernalis*, *C. exlineae*, and *A. punctulata*. Additionally, lycosid species were included as outgroup comparisons: *Hogna carolinensis* (Walckenaer, 1805) and *Rabidosa rabida* (Walckenaer, 1837). Spiders were housed in a 500 cm^3 plastic container and were watered and cleaned weekly. The temperature was maintained between 72-77° C, which is near the average temperature that these animals experience in their natural habitats. Also, these animals live in relatively humid environments, so the humidity levels stayed between 50-70%. The day-night cycle simulated that of what these animals experience during the summer, with 16 hours of light and 8 hours of darkness.

Each week for three weeks the spiders were weighed and placed in a 500 cm^3 acrylic box, then a juvenile house cricket (*Aecheta domesticus* (Linnaeus, 1758)) was also weighed and placed in the container with the spider. The interaction between the spider and the cricket was videotaped and then analyzed to determine if the spider successfully subdued the cricket as well as the initial attack location on the body of the cricket. Biases in attack location were evaluated with a Chi-squared goodness of fit test with R v 3.6.3. Further, the following three time points were collected in seconds: (1) the time the cricket was placed in the container, (2) the time the spider envenomated the cricket, and (3) the time the cricket became paralyzed as determined by the ceasing of movement of its antennae. The difference between the first two time points was used to determine the approximate time to attack, and the difference between the 3rd and 2nd time points was used to reflect time to paralysis. Both seconds to attack and seconds to paralysis were transformed using hyberbolic arcsin function, (to have a similar result as logarithmic transformation but allowing for values of zero) before using an ANOVA with R to determine if either varied between species.

Three mixed effects models were constructed to evaluate the effect that species has on the relationship between the mass of the cricket relative to the spider and the time it took for the cricket to become paralyzed using the lmerTest package in R[59]. For the simplest model 1.1, we let $Y_i TimeToParalysis$ be the time in seconds to paralysis for every i_{th} individual spider as the dependent variable, and $X_i CricketMass$ be the weight of every i_{th} cricket relative to the weight of every i_{th} individual spider as the independent variable. Further, due to repeated measures, we let I_i be the individual-specific random effect for every i_{th} individual spider. The other two models included an additional explanatory variable $\beta_1 Species_i$ of every i_{th} individual spider, respectively. Models 1.2 and 1.3 included $\beta_1 Species_i$, while model 1.3 allowed for species to have non-equal slopes. The R library lmerTest was used to construct all these mixed effects models, and then the ANOVA function from lmerTest was used to compare each of the models to the simplest model to keep the model that is significantly better and has a lower AIC score. The formulation for all four models can be found below.

$$Y_i TimeToParalysis = X_i CricketMass + I_i \tag{1.1}$$

$$Y_i TimeToParalysis = X_i CricketMass + \beta_1 Species_i + I_i \tag{1.2}$$

$$Y_i TimeToParalysis = X_i CricketMass \times \beta_1 Species_i + I_i \tag{1.3}$$

1.3 Results

Transcriptome assembly statistics: The 48 assemblies in this analysis included an average transcript recovery of 106,410 (s.d = 43,708), representing an average of 84,474 (s.d = 32,908) genes as designated by Trinity. The 21,936 genes with alternative transcripts designated by Trinity had an average of 3.04 isoforms (s.d = 1.99).

The average percentage of complete BUSCO hits within the assemblies was 84.96% (s.d = 15.48%), and an average of 60.33% (s.d = 8.90%) were single copy. 245 BUSCO loci met the threshold of 60% of species represented with at least 50% of the sequences being nonidentical. The untrimmed alignments had an average matrix width of 348.8 nucleotides (s.d = 179.0); trimming the alignments reduced the average to 314.7 nucleotides (s.d = 162.8). The total size of the concatenated matrix was 78,594 nucleotides.

Phylogenetic results: There were no topological differences between the concatenated matrix phylogeny and the ASTRAL species tree. The only topological difference that occurred between this study and Cheng et al. 2018 was that we recovered Oxyopidae as sister to a clade comprising Ctenidae + Psechridae + Lycosidae + Pisauridae, instead of being sister to the Thomisidae. Temperate zone North American ctenids do not form a single lineage, as *Anahita punctulata* is sister to the North American *Ctenus-Leptoctenus* clade and the Neotropical *Phoneutria-Isoctenus* clade. Within the Ctenidae, the genus *Ctenus* is polyphyletic in regard to *Ctenus corniger*, the only Old World representative of the family, being sister to all New World member ctenids included here.

Feeding observations: A total of 180 feeding observations among 83 individual spiders from 6 species were measured, resulting in an average of 2.17 observations per species (s.d = 1.24). The two species with the most individuals with feeding observations were *Ctenus*

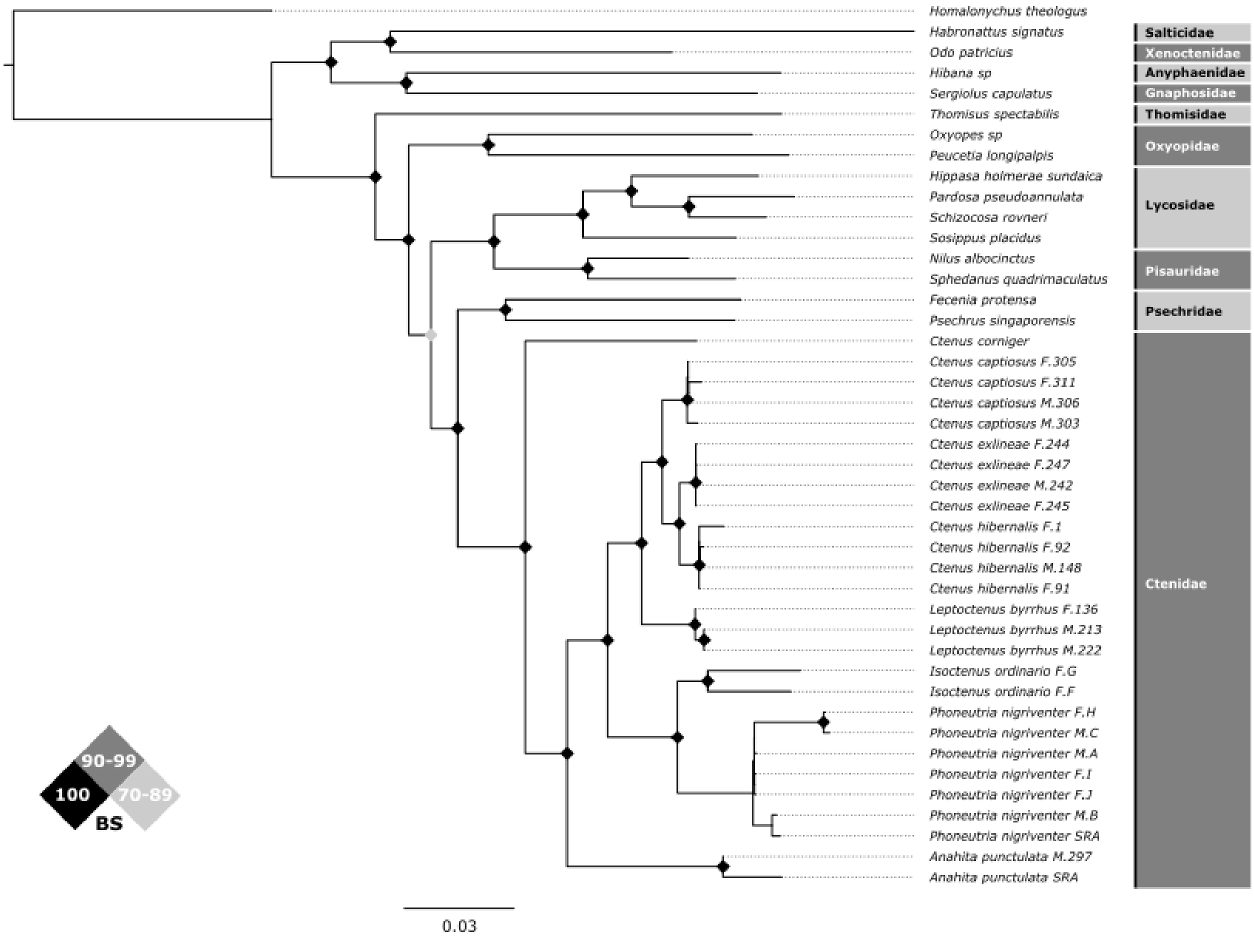

Figure 1.3: Reconstructed species-level phylogeny from concatenated matrix using IQTREE. Bootstrap support values are indicated by the color of the diamond placed on the inner nodes.

hibernalis and *Hogna carolinensis* with 64 and 57 individuals respectively. Next, was *Ctenus exlineae* and *Leptoctenus byrrhus* each comprised 29 and 17 observations respectively. The species with the fewest individuals observed were *Anahita punctulata* and *Rabidosa rabida* with 7 and 6 individuals respectively. Males and juveniles made up less than 10% of individuals with observations, so all downstream analyses were restricted to female observations.

The three ctenid species with the most observations (*C. hibernalis*, *C. exlineae*, and *L. byrrhus*) and the lycosid *H. carolinensis* had at least 5 observations per attack location (abdomen, head, leg, and thorax) necessary to carry out an attack location bias estimation[60]. All four species showed significant biases towards attacking the thoracic region (Figure 1.4). The most significant were from *H. carolinensis* ($\chi^2_3 = 108.4, p = 2.2 \times 10^{-26}$) and *C. hibernalis* ($\chi^2_3 = 117.88, p = 2.2 \times 10^{-16}$). *L. byrrhus* also showed significant attack biases towards the thorax ($\chi^2_3 = 30.3, p = 1.2 \times 10^{-6}$), whereas the bias exhibited by *C. exlineae* were split between throacic and abdominal region, though still more towards the thorax ($\chi^2_3 = 26.3, p = 8.2 \times 10^{-6}$).

Time to attack in seconds (transformed using hyperbolic arcsin) varied between species ($F_{5,174} = 11.16$,$p = 2.45 \times 10^-9$), as can be seen in Figure 1.5. According to Tukey HSD *post hoc* analysis, *H. carolinensis* was significantly faster than all 4 ctenid spiders with p-values less than 0.005, except when compared to *C. hibernalis* (p = 0.0068). *C. exlineae* was significantly slower to attack than *C. hibernalis* and *R. rabida* (p < 0.05). Finally, *R. rabida* was significantly faster to attack than *L. byrrhus* (p = 0.025).

Time to paralysis in seconds (transformed using hyperbolic arcsin) only slightly varied between species, though was just 0.0025 above the $\alpha = 0.05$ significance threshold ($F_{5,174} = 2.24$,$p = 0.0525$). Further inspection using Tukey HSD *post hoc* analysis revealed no pairwise significant differences in time to paralysis between species.

Only one individual was paired with a cricket that was larger than it, a *C. hibernalis* so it was removed from the the three mixed effects models used to to evaluate the effect that species has on the relationship between the mass of the cricket relative to the spider and

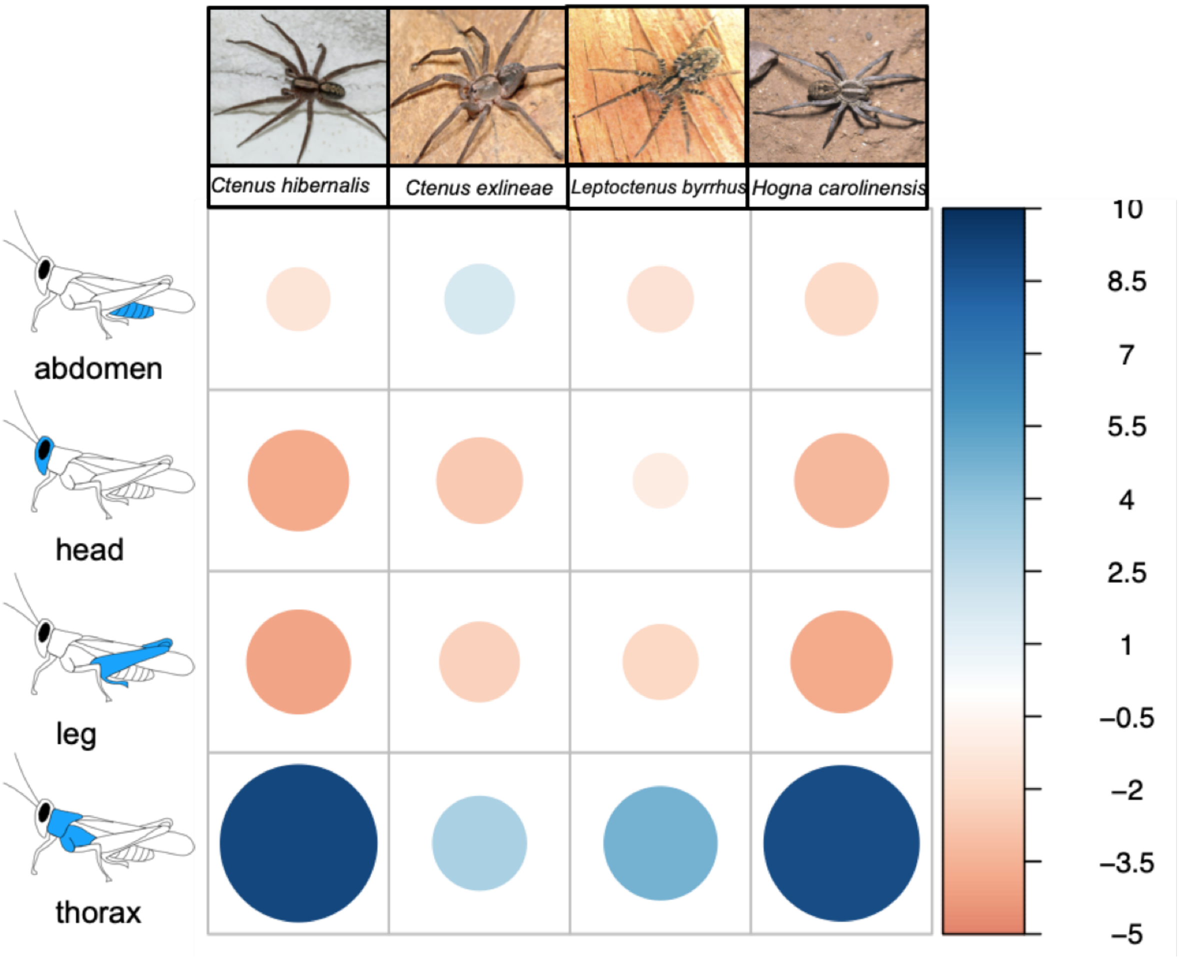

Figure 1.4: Residuals of Chi-squared goodness of fit analysis of biases in attack location.

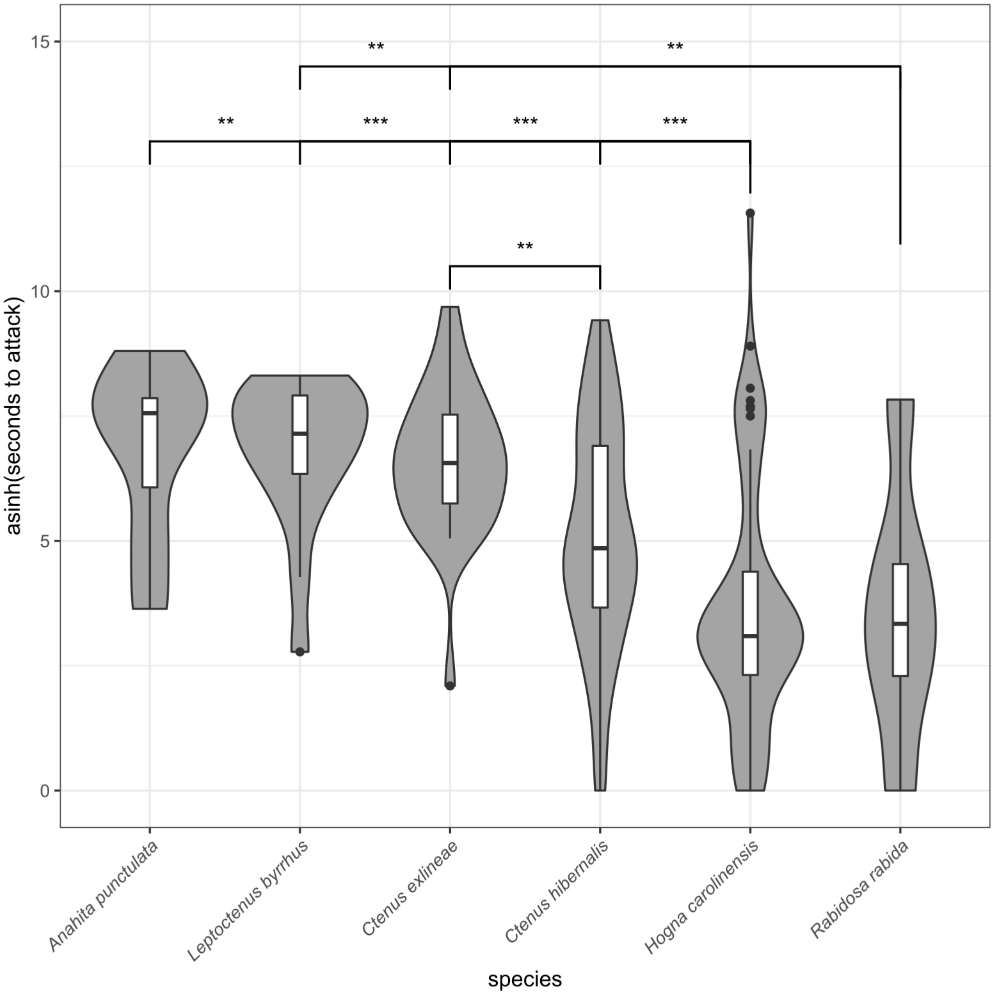

Figure 1.5: A boxplot of the seconds passed before attack was initated between species (hyperbolic arcsin transformed). Bars indicate pairwise significant differences.

the time it took for the cricket to become paralyzed. Following an ANOVA of models 1.2 and 1.3 to the simplest model 1.1, model 1.3 which allowed for species to have non-equal slopes, was a significantly better fit than the other two models ($\chi^2_3 = 349.12, p = 0.005$). As visualized in Figure 1.6, *C. exlineae* and *C. hibernalis* both had negative slopes (-0.21 and -0.38 respectively), while *H. carolinensis* and *L. byrrhus* had positive slopes (0.57 and 0.46 respectively).

Table 1.2: Summary of ANOVA of the three models constructed with relative mass of the cricket (hyperbolic arcsin transformed) as the independent variable and seconds to paralysis (hyperbolic arcsin transformed) as the dependent variable. Model 1.2 included species as a fixed effect, and model 1.3 allowed species to have non-equal slopes

model	parameters	AIC	BIC	log-likelihood	deviance	Chi-squared	Df	p-value
1.1	4	373.73	385.08	-182.87	365.73			
1.2	7	375.76	395.62	-180.88	361.76	3.97	3	0.27
1.3	10	369.12	397.48	-174.56	349.12	12.64	3	0.005

1.4 Discussion

Phylogenetic results: Prior to this investigation, the origins of the handful of species of ctenids in the temperate zone of North America was unclear. Interestingly, *Ctenus hibernalis* and *Ctenus exlineae* are more closely related to each other than *Ctenus captiosis*, despite *C. hibernalis* being nearly morphologically indistinguishable from *C. captiosis*. The family has a Gondwanan distribution, with most of the diversity found in the neotropics, suggesting that

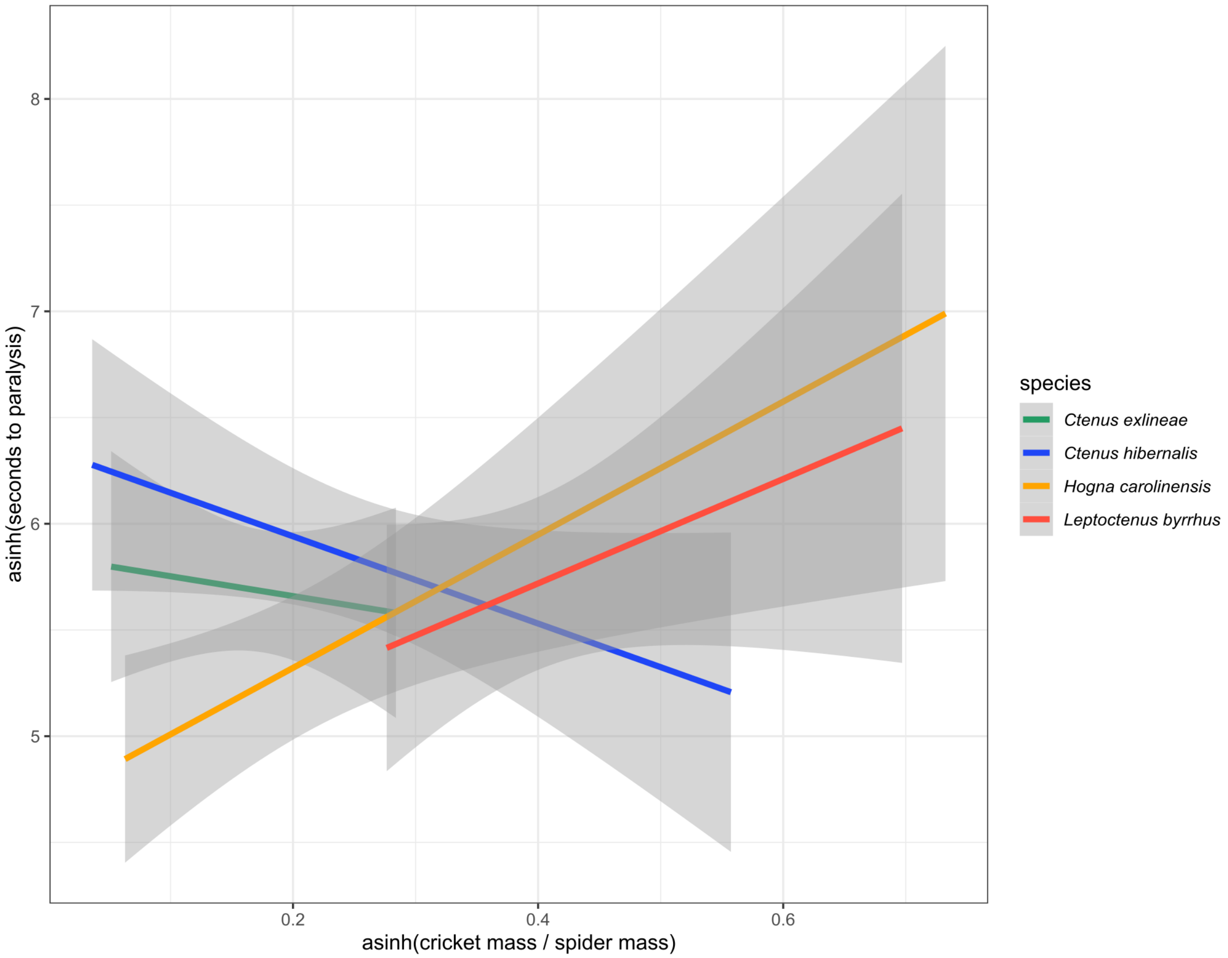

Figure 1.6: Relationship between seconds (hyperbolic arcsin transformed) passed until the cricket became paralyzed and the mass of the cricket relative to the mass of the spider

the few species in the U.S migrated from the tropics[3]. Because all U.S ctenids do not form a single clade, ctenids appear to have colonized North America independently more than once. *Leptoctenus* and *"Ctenus"* represent a single colonization event, though the biogeographical history of *"Ctenus"* remains unclear.

Comparative venom biology: Prior to this investigation, to what extent venom utilization differed between the U.S ctenid species was uncertain.Two critical components of venom utilization that precede envenomation are the time it takes to initiate the envenomation and the anatomical location of envenomation on the prey item. When comparing the two representative lycosid species to the three ctenid species, lycosids initiated the envenomation faster than ctenids. Though, the sample sizes of representative species are too small to infer that lycosids are generally faster than ctenids at attacking. Within North American *"Ctenus"*, *C. hibernalis* was faster to initiate envenomation than *C. exlineae*, despite being recovered as sister groups in our phylogenetic reconstruction, suggesting behavioral divergences have occurred since the two species diverged. With increased sampling spanning more species, the phylogeny inferred in this study will be an invaluable tool in testing evolutionary hypothesis in regards to the venom biology of ctenids, at the current state however, such phylogenetic comparative analysis are not yet feasible.

All species exhibited similar anatomical biases in where they initiated envenomation of cricket prey. For the most part, both lycosids and ctenids favored biting the thoracic region of the cricket. This behavior has been demonstrated in other spider species[61], though it is remarkable to have an additional line of evidence to suggest that spiders are somehow able to target the thoracic region, which is much smaller than the cricket's abdomen but contains a higher density of nervous tissue that is more sensitive to envenomation[62].

The variation in venom potency is still unclear in these species. Anecdotally, the U.S ctenids are expected to have less noxious venom, at least to vertebrate physiologies, than their ctenid relavtives in the genus *Phoneutria*, but the degree to which venom toxicity varies is still uncertain. There were no notable differences in the time to cricket prey paralysis between lycosids and the ctenids used in this investigation. If there were drastic differences in venom toxicity within these species, the species with the most toxic venom would have been able to disrupt cricket physiology faster. Beyond comparing the time to paralysis, comparing the relationship between the time to paralysis and the relative weight of the cricket provided more insights. The best fitting model included species as a fixed effect and allowed for unequal slopes. Overall, for both *L. byrrhus* and *H. carolinensis* exhibited a positive correlation between time to paralysis and relative cricket size, whereas *C. hibernalis* and *C. exlineae* displayed a negative correlation. As cricket mass increases, more venom is required to paralyze it, so delivering a constant amount of venom would lead to an increase in time to paralysis as the relative size of the cricket increases. Changes to this relationship indicate a potential alteration in the amount of venom delivered. Thus, for *C. hibernalis* and *C. exlineae* this could be indicative venom metering, in which there is an adjustment of the amount of venom delivered as relative cricket size increases. Venom metering has been documented in other spider species[63,64], but increased taxon sampling is needed to understand how prevalent the phenomena is throughout the spider tree of life.

Crude venom toxicity assays still need to be performed. Comparing the lethal median dosage of venom between species is a non-trivial task because what is highly toxic for one animal may have no effect at all on even close relatives. This is due to the fact that venomous predators and their prey often participate in a cyclical evolutionary arms races where natural selection gives rise to prey that resist the effects of venom and escape predators, which then in return give rise to predators with more toxic venoms[65–68]. This results in venom that is moderately toxic to prey that it co-evolved with but highly toxic to closely related species that the predator did not co-evolve with[69–72]. Thus, the venom activity of a particular species is highly influenced by the predator-prey dynamics in which it has evolved. For example, within ctenids, members of the genus *Phoneutria* occasionally prey upon tetrapods and have venom that is highly noxious to humans[73,74], whereas several other groups of spiders such as fishing spiders within Pisauridae Simon, 1890, prey upon vertebrates but do not have venom that is particularly noxious to humans[75]. Thus, it is unclear how highly noxious ctenid venoms evolved and the degree to which noxious venom persists within the family.

1.5 Conclusion

Overall, this study introduced the first comprehensive evolutionary framework for wandering spiders in the U.S, with additional descriptions of their basic venom biology. We demonstrated species-specific variation in venom biology related behaviors and have introduced a statistical approach to determine venom utilization differences from behavioral data. With increased sampling, coupled with toxicity assays, this could become the framework for a more holistic understanding of the evolution of noxious venom in wandering spiders.

Chapter 2

Molecular evolution of Inhibitor Cystine Knot toxins in wandering spiders (Araneae: Ctenidae)

2.1 Introduction

Animal venoms are a biomolecular cocktail with high concentrations of protein and peptide toxins[76]. There is a tremendous diversity of proteins expressed across venomous taxa, totaling upwards of a million unique protein toxins[77]. Unlike defensive venom systems (e.g. venom utilized by echinoderms, helodermatid lizards, lepidopteran larvae, most venomous fishes, and other insects) that are streamlined and highly conserved for immediate and extremely localized pain, predatory venom systems (e.g. venom utilized by snakes, spiders, scorpions, centipedes, cephalopods, gastropods, cnidarians, some lizards, and some insects) are highly variable in composition and molecular functionality[78]. Among predatory venomous taxa, venoms serve as foraging adaptations, so natural selection on venom composition is a likely consequence. The molecular diversity found in venom proteins arises in most cases through the duplication and subsequent neofunctionalization of non-toxic physiological proteins that are actively selected on[79].

Recent advances in RNA-Seq technologies and Mass Spectrometry (MS) proteomics have played a pivotal role in expanding knowledge of molecular evolution through the generation of an abundance of protein coding sequence data across all levels of biodiversity[80]. In venom systems, these tools have allowed for high-throughput investigations of venom gene expression. Venom gland transcriptomics provides insights into venom composition and expression levels but likely overestimates the number of venom genes that are actually translated to a final protein product[42]. While MS proteomics provides the aforementioned lacking information about venom gene translation, it still relies on a sequence database for searching the peptide fragment mass spectra against. Thus, a proper investigation of venom composition requires a venom gland transcriptome and a venom proteome, which are used in concert to form a "venome".

Venomic investigations have largely focused on species whose bites result in life-threatening symptoms in humans[81,82]. This is especially true in spiders, which are the largest group of venomous animals, yet only the small handful of spiders that have medically relevant venoms have been the subject of comprehensive venomic characterization. Despite their toxicity, little to no transcriptomic or proteomic investigations have occurred in wandering spiders outside of the medically significant genus *Phoneutria* Perty, 1833 (Arachnida: Araneae: Ctenidae). Within *Phoneutria* there have been numerous proteomic characterizations and targeted approaches to delineate the active noxious components[83,84]. One of these components is toxin Tx1, which exerts inhibitory effects on neuronal sodium channels in a highly selective manner and has a lethal median dosage (LD_{50}) of 47 μg/kg in *Mus musculus*[85]. This is nearly four times more toxic than the lethal nerve agent Sarin, which is classified as a Schedule 1 substance by the Chemical Weapons Convention of 1993 and has an LD_{50} of 172 g/kg in *M. musculus* when injected subcutaneously. Another noxious component is toxin Tx2-6 which is the responsible agent for priapism, occasionally an envenomation symptom of *Phoneutria* in human males[86]. Expression of these components greatly contributes to the danger of the bite of *Phoneutria*, yet many species exist within the same family that are confirmed to share homologs to the noxious components of *Phoneutria* but do not share the dangerous bite[87].

A recent investigation of the venom gland transcriptome and proteome of *P. nigriventer* (Keyserling, 1891) revealed that its noxious venom components are among a diverse class of toxins known as inhibitor cystine knot (ICK) toxins[88]. The core ICK cysteine framework consists of three pairs of cysteines forming disulfide bridges between C_1-C_4, C_2-C_5, C_3-C_6 to take on an unusually stable conformation. That investigation recovered 98 cysteine rich toxins that represented nine additional cysteine frameworks, six of which were verified to be ICKs. The number of cysteine residues per group ranged from six to fourteen. They also spanned a broad range of predicted functionality, from ion channel modulators of varying specificity (Ca^{+2},K^{+}, and Na^{+}), to protease inhibitors and NMDA receptor modulators. The cysteine-rich peptide toxins represented 93.24% of the relative abundance of peptides expressed in the venom when accounting for expression levels. The observation that ICKs make up the majority of the venom composition appears to be common across the spider tree of life[89].

Functional diversity is an important asset of venom in a predatory system, because it allows for the predator to behave as a generalist that can incapacitate a broad array of prey items[90]. In arachnids, venom has evolved as a predation strategy independently in spiders and scorpions, though the evolution of functionally diverse venoms followed different trajectories in those lineages. In both spiders and scorpions, ICKs were recruited into venom tissues via the process of gene duplication followed by subsequent neofunctionalization[25]. As is evident in *P. nigriventer*, ICKs in spiders exist as multi-copy gene families spanning numerous cysteine frameworks and molecular functionalities, which in part gives rise to functional diversity. In scorpions, however, ICKs are single copy per species with one cysteine framework, and broader functional diversity was achieved through the recruitment of additional toxin components. Thus, ICKs in spiders represent a unique opportunity to serve as a case study for the molecular evolution of functional diversity within a gene family. The heavily characterized ICKs of *Phoneutria* make ctenids a great model for ICK evolution in spiders. The North American representatives of this family are convenient to collect, but also consists of multiple unrelated lineages from three genera which allows for phylogenetic independence. For the first time, in this study we investigate the evolutionary processes that have given rise to the diverse class of ICK toxins in spiders with a focus on wandering spiders and their close relatives.

2.2 Methods

Taxon sampling: Three adult males and females were collected from all members of Ctenidae in the United States, excluding the narrow Texas cave endemic *C. valveriensis*. *C. hibernalis*, *C. exlineae*, *C. captiosus*, *L. byrrhus*, and *A. punctula* were collected from the following respective localities: North-Central Alabama, Northwest Arkansas, Central Florida, South-Central Texas, and Northwest Georgia. Four male and female *Phoneutria nigriventer* along with four female *Isoctenus sp.* were collected from Brazil. Ctenidae belongs to the superfamily Lycosoidea, which is a member of a clade of spiders that possess a retrolateral tibial apophysis (RTA), a backward-facing projection on the tibia of the male pedipalp. To allow for an investigation of the broader evolutionary context of ICK toxins in wandering spiders, whole body transcriptomes from outgroups within the RTA clade were retrieved from NCBI's Short Read Archive.

Venom and RNA isolation: Spiders were housed in a 500 cm^3 plastic container and were watered and cleaned and fed crickets (*Acheta domesticus*) weekly. The temperature was maintained between 72-77° C, which is near the average temperature that these animals experience in their natural habitats. Prior to venom collection, individuals were anesthetized with CO_2 using a modified procedure as described by Barrio & Brazil [91]. Venom was collected using electrostimulation with 7 V of AC current, similar to previous studies[92–94].

Anesthetized individuals were placed on clamped forceps attached to an electrode. One prong of the forceps was wrapped in non-conductive insulating tape to create a point of contact for the spider that retards current, while the other prong of the forceps was wrapped with a cotton thread and soaked in saline to create a point of contact with the spider to promote electrical conductivity. A capillary tube was then placed over the fang in order to collect the venom. Finally, the second electrode was placed on a syringe connected to a vacuum pump which was touched to the base of the chelicerae in order to complete the circuit and allow the muscles around the venom gland to contract and eject venom into the capillary tube while simultaneously allowing regurgitate to be vacuum pumped through the syringe to prevent contamination. The collected venom was then stored at -80° C. Two days after venom milking, the venom glands of each ctenid were dissected out, whole RNA was isolated from the venom glands from the five U.S species and the two Brazilian species using TRIzol® (Life Technologies, Carlsbad, CA) and the Qiagen RNeasy kit (Qiagen, Valencia, CA). RNA concentration and integrity was evaluated using Quant-iT™ PicoGreen and Bioanalyzer.

Sequencing and processing: The RNA extraction was sent to the Genomic Services Lab at HudsonAlpha (Huntsville, AL) for library preparation with poly(A) selection and sequencing on a 100 bp paired-end run on an Illumina HiSeq 2500, comprising 25 million reads forward and reverse (50 million total reads) per sample. For all samples sequenced after 2018, due to sequence facility updates, the same setup was used but with Illumina NovaSeq. Additionally, RNAseq reads from all outgroup species were retrieved from NCBI Short Read Archive.

Prior to assembly, FASTP v 0.19.6[41] was used to remove of adapters, correct sequencing errors, and trim low-quality base calls, ensuring maximal accuracy of transcript recovery[40]. All read processing was executed with. *De novo* assemblies typically recover an unexpectedly large number of transcripts, sometimes well over 100,000[42]. This happens for three main reasons. First, an increased depth of sequencing combined with improved transcript recovery algorithms increases the recovery of transcripts that are expressed at levels lower than would otherwise be considered biologically relevant. Second, common contaminants, such as bacteria and fungi, unavoidably make their way into samples, thus inflating the mRNA transcript pool diversity. Third, in eukaryotic systems, alternative splicing yields a significant increase in recoverable transcripts per gene locus as isoforms. Further, when several RNA-seq experiments from different species are sequenced together, cross contamination inevitably occurs[43–49]. The aforementioned issues can have tremendous effects on downstream phylogenetic inferences made from problematic transcripts[50]. To alleviate these issues, reads were first mapped to a transcript database of common contaminants (bacteria, fungi, human, and nematodes) using salmon v1.3.0 with default mapping parameters, all unmapped reads were retained as the finalized library of processed reads[51].

Transcript reconstruction and expression quantification: The processed reads were *de novo* assembled into transcripts using TRINITY v2.8.4[52]. TRINITY utilizes the following three-step approach to assemble transcripts *de novo*: First, it assembles unique portions of alternatively spliced transcripts. Second, it clusters those contiguous sequences and constructs de Bruijn graphs. Finally, it processes each graph to assemble full-length transcripts for alternatively spliced isoforms.

An important aspect of elucidating the functional relevance of a given protein is the quantification of its expression level. This can be achieved at the transcript level through quantifying read coverage by mapping the reads from each sample back to their assembled transcripts. Salmon uses a quasi-mapping approach and is one of the fastest, most efficient, and most accurate methods for quantifying expression in RNAseq experiments[51]. Since TRINITY also assembles splice variants and alleles of the same gene, these were consolidated into SuperTranscripts[95] prior to mapping so that the inferred expression values are at the gene level and not the isoform level. SuperTranscripts are formed by collapsing common and unique regions of sequences among splicing isoforms into a singular consolidated linear sequence. Then, the SuperTranscripts were provided as input along with the processed reads to salmon to quantify expression levels as Transcripts per million (TPM).

Venom proteomics: To characterize the venom profiles, venom proteins were isolated using HPLC followed by MALDI-TOF Mass Spectrometry. MALDI-TOF is the preferred method for mass analysis in proteins due to only singly charging analytes, compared to commonly used ESI techniques which apply multiple charges to analytes and complicate downstream analysis. The finalized dataset of candidate venom encoding transcripts were elucidated by cross-referencing proteomic MS data to the transcriptome using the CRUX pipeline[96].

Inhibitor cystine knot annotation: Coding sequences within transcripts were inferred using TRANSDECODER v3.0.1[53]. TRANSDECODER uses the following criteria to identify the single best coding sequence in a given transcript: available open reading frame (ORF) of a minimum length of 30 codons, log-likelihood score of the coding sequence, and predictions of start and stop codons as refined by a Position-Specific Scoring Matrix. To ensure phylogenetic independence, only the longest inferred coding sequence per isoform designated by Trinity was retained.

To identify inhibitor cystine knot toxins in the transcriptome assemblies, a database of verified ICKs from spiders was retrieved from the KNOTTIN database[97]. Only ICKs with complete coding sequences and verified disulfide connectivity were retained in the final verified ICK database. The database was provided as input to BLASTp to search against the inferred protein sequences from TRANSDECODER[98]. Additionally, a multiple sequence alignment was generated from the verified ICK database to create a Hidden Markov Model (HMM) that could be searched against the genomic protein sequences using HMMER v3.3.1[99]. For both BLASTp and HMMER, only matches with at least six cysteines and up to 200 amino acids in length were kept for downstream analysis. Additionally, putative matches were only kept if they contained a signal peptide as indicated by signalP v5.0, which predicts the presence of signal peptide cleavage sites[100]. A homology network of the finalized peptides was generated using an all against all BLASTp search, and then provided as input to SiLiX v1.2.11 to group the peptides into putative gene families[101]. Cysteine frameworks were designated using the following approach. Cysteines that were directly adjacent to each other were designated as CC. Cysteines separated by one residue were denoted as CXC. Finally, cysteines separated by more than one residue were designated as $C-C$. Each framework was given a numeric code to represent the number of cysteines they contain along with a unique identifier. To ensure that only ICKs are included in the analysis, only cysteine frameworks representing the top 80% of peptides in the largest family were included for downstream disulfide connectivity predictions and phylogenetic analysis. A non-redundant dataset was then created to only include unique coding sequences.

Disulfide connectivity predictions: Disulfide connectivity is of great importance in understanding structural homology in ICK toxins when sequence similarity is greatly reduced. Determining disulfide connectivity normally requires empirical structural validation but can be reasonably predicted using computational predictions. The number of possible disulfide bonds explodes in a combinatorial fashion, to the point where exhaustively comparing disulfide connection possibilities in peptides with more than 6 cysteines is not computationally tractable[102]. To alleviate this, a number of heuristic approaches have been developed. For this dataset the following four approaches were used to generate disulfide connection predictions for a random representative mature peptide from the top cysteine frameworks.

1. DiANNA v1.1 uses PSIPRED[103] to produce secondary structure predictions that are provided to PSIBLAST[98] that are searched against SwissProt which form a multiple sequence alignment and make a profile for each pair of cysteines used for disulfide connectivity prediction with a diresidue neural network[104]. This was used to generate a single prediction per ICK representative per unique cysteine framework.
2. DISULFIND collectively decides the bonding state assignment of the entire chain using a Support Vector Machine binary classifier followed by a refinement stage[105]. DISULFIND v1.1 was used to generate a total of three alternative disulfide connection predictions.
3. CYSCON uses a hierarchical order reduction protocol to identify the most confident

disulfide bonds and then evaluate what remains using Support Vector Regression[106]. CYSCON v2015.09.27 was used to generate a single disulfide prediction per ICK representative per unique cysteine framework.

4. CRISP v1.0 not only predicts disulfide bonds, but also the entire structure of a cysteine rich peptide by searching a customized template database with cysteine-specific sequence alignment with three separate machine learning models to filter templates, rank models, and estimate model quality[107]. CRISP was used to generate five structural models for each ICK representative per unique cysteine framework.

A chord diagram was constructed for each cysteine framework to demonstrate the variability in disulfide connectivities for every prediction attempt of each of the approaches using the D3 JavaScript library[108]. A consensus disulfide connection prediction of all the approaches in conjunction with previously published disulfide connections, as found through Arachnoserver[109], were used to generate the finalized disulfide connectivity predictions for the three disulfide bridges homologous to all cysteine frameworks.

Phylogenetic tests for selection: Aligning ICKs, or any cysteine rich peptide, is difficult due to nonhomologous cysteines mistakenly being aligned. Thus, the finalized consensus disulfide connectivities were used to inform the alignment of ICKs using a similar approach to Pineda *et al.* [89]. Rather than align everything at once and then manually adjusting misaligned cysteines followed by realignment of regions between the two adjusted cysteines, only amino acids between cysteines participating in disulfide bonds common to all ICKs in the dataset were used for the alignment. Additionally, the regions between homologous cysteines were aligned separately while using the barcode “WWYHWYYHMM” to replace flanking cysteines to prevent inner cysteines from misaligning with flanking cysteines similar to the approach by Shafee *et al.* [110]. The alignment was then provided as input to IQTREE v1.5.5[57] to test the amino acid composition using a Chi-squared test. Any sequences that failed the test were removed from the alignment, and the sub-regions were realigned following the previously described procedure. The resulting alignment was reverse translated to form a coding sequence alignment using PRANK[111].

The same outgroup as used by Pineda *et al.* [89] (disulfide-directed β-hairpin from the whip scorpion *Mastigoproctus giganteus*) was added to the protein alignment using MAFFT v7.455[55]. The phylogenetic relationships of the ICKs in this alignment were reconstructed using IQTREE and the default settings.

Adaptive molecular evolution is typically inferred in coding sequences by comparing ratios of the rates of nonsynonymous substitution and synonymous substitution (d_N/d_S or ω), where d_N exceeding d_S indicates positive selection, d_S exceeding d_N indicates negative selection, and d_N/d_S approaching unity indicates neutral evolution. One of the broadest questions to ask in regards to how a gene has evolved is "Has a particular gene evolved under positive selection?" To do this, the HYPHY[112] implementation of Branch-Site Unrestricted Statistical Test (BUSTED) for Episodic Diversification assesses whether a gene has experienced positive selection at at least one site on at least one branch, by fitting a codon model with three rate classes constrained as $\omega_1 \leq \omega_2 \leq 1 \leq \omega_3$[113]. This unconstrained model is compared to a null model disallowing positive selection where $\omega_3 = 1$. To determine if ICKs have experienced positive selection, the codon multiple sequence alignment and phylogeny were provided as input to BUSTED using default parameters with the entire phylogeny set as the background.

In ICKs, specific amino acid sites may play an important role in the structure-function (i.e., binding specificity) and adaptive evolution. To identify specific amino acid sites that have undergone pervasive positive selection, the HYPHY implementation of a Fast, Unconstrained Bayesian AppRoximation (FUBAR) employs a pre-specified discrete grid of dN and dS values to be applied across sites[114]. The codon multiple sequence alignment and phylogeny were provided as input to FUBAR with default parameters.

There may only be specific episodes where certain amino acids receive strong bouts of positive selection. To determine if a certain number of branches have amino acid sites undergoing positive selection, the HYPHY implementation of a Mixed Effects Model of

Evolution (MEME)[115] by comparing a null model of rate parameters to an alternative model of rate parameters at each branch. Both models include a single d_S or α value, and two separate d_N or β values (β^+ and β^-), where α is shared between both β values per site. In the null model, both β values are restricted to be less than or equal to α, whereas β^+ is unrestricted in the alternative model. Positive selection is inferred when $\beta^+ > \alpha$. To determine if certain amino acid sites have undergone episodic positive selection, the codon multiple sequence alignment and phylogeny were provided as input to MEME with default parameters and the entire phylogeny set as the background.

To evaluate specific instances on a phylogeny where positive selection has occurred, branch-site models are typically implemented. Much like how MEME is unable to statistically specify the exact branches within a site undergoing episodic positive selection, branch-site models are only able to identify specific branches where a certain portion of sites have undergone positive selection. This can be accomplished using the HYPHY implementation of adaptive Branch-Site Random Effects Likelihood (aBSREL)[116], by modeling both site-level and branch-level ω heterogeneity. Since some branches may feature more or less complex evolutionary patterns, aBSREL infers the optimal number of ω rate classes for each branch using the sample Akaike information criterion (AIC_c). A likelihood ratio test is then performed to compare the full adaptive model to a null model where branches are not allowed to have rate classes where ω exceeds unity.

Aside from evaluating signatures of positive selection through calculations of codon substitution rates, it is also useful to detect the presence of the co-occurrence between amino acid positions in ICKs, which may provide useful inferences into the evolution of their structure/function. This can be achieved using the HYPHY implementation of the Bayesian Graphical Model (BGM)[117], which maps amino acid substitutions to a phylogeny and reconstructs ancestral states for a given model of codon substitution rates that is then followed up by a series of 2x2 contingency table analyses.

2.3 Results

Venom gland transcriptome and proteome: The 48 assemblies in this analysis included an average transcript recovery of 106,410 (s.d = 43,708), representing an average of 84,474 (s.d = 32,908) genes as designated by Trinity (Table 2.1). The 21,936 genes with alternative transcripts designated by Trinity had an average of 3.04 isoforms (s.d = 1.99). From the longest isoforms, we recovered on average 12,173 complete coding sequences per species (s.d = 4,993), with an average amino acid length of 264 (s.d = 9).

Table 2.1: Transcriptome assembly statistics for all ctenid samples contributed to this study, including number of SuperTranscripts and coding sequences.

species	sex	sample	transcripts	superTranscripts	CDS
Anahita punctulata	male	297	120,048	88,950	55,238
Ctenus captiosus	female	305	124,919	99,712	59,434
Ctenus captiosus	female	311	140,647	110,550	64,437
Ctenus captiosus	male	303	157,109	123,061	70,951
Ctenus captiosus	male	306	158,795	122,878	71,795
Ctenus exlineae	female	244	105,140	85,561	49,630
Ctenus exlineae	female	245	111,981	89,516	52,365
Ctenus exlineae	female	247	112,224	90,112	52,434
Ctenus exlineae	male	242	95,088	78,974	44,942
Ctenus exlineae	male	246	54,776	46,375	24,166
Ctenus hibernalis	female	91	194,576	148,947	83,512
Ctenus hibernalis	female	92	161,519	124,108	71,340
Ctenus hibernalis	male	148	202,764	157,422	81,381
Leptoctenus byrrhus	female	136	99,257	81,488	47,189

species	sex	sample	transcripts	superTranscripts	CDS
Leptoctenus byrrhus	male	213	108,687	88,723	50,313
Leptoctenus byrrhus	male	222	101,706	83,634	47,527

Inhibitor cystine knot annotation: A total of 1,259 cysteine rich peptides were recovered that met the following criteria: signal peptide present, less than 200 amino acids, mature peptide had at least 6 cysteines, and was a match with the KNOTTIN database from either a BLAST search or HMMER with an e-value cutoff of 1×10^{-3} . On average each sample contained 26.6 cysteine rich peptides (s.d = 8.3). Cysteine rich peptides made up less than 10% of the total peptides in the proteomes of *C. exlineae* and *C. hibernalis*, though in at least two samples they made up over 50% of the relative abundance when accounting for expression level (Table 2.2).

Table 2.2: Cysteine rich peptide composition in the proteomes of *C. exlineae* and *C. hibernalis*, with comparison to the expression levels from the venom gland transcriptomes of each sample per species.

species	sex	sample	peptides	ICKs	%ICK	sum TPM	ICK TPM	%TPM
C. exlineae	male	242	113	5	4.42%	18,214.3	9,670.5	53.1%
C. exlineae	female	245	127	7	5.51%	16,755.8	7,002.0	41.8%
C. exlineae	male	246	200	13	6.50%	131,454	51,011.1	38.8%
C. exlineae	female	244	339	5	1.47%	42,546.4	12,227.6	28.7%
C. exlineae	female	247	194	5	2.58%	27,378.7	5,436.3	19.9%
C. hibernalis	female	4926	196	8	4.08%	109,103	57,283.2	52.5%
C. hibernalis	female	91	525	12	2.29%	111,759	14,703.9	13.2%
C. hibernalis	male	148	170	7	4.12%	31,121.2	2,581.5	8.3%

species	sex	sample	peptides	ICKs	%ICK	sum TPM	ICK TPM	%TPM
C. hibernalis	female	92	176	5	2.84%	37,128.6	2,366.6	6.4%

SiLiX grouped the cysteine rich peptides into 53 putative gene families. The largest family comprised 1,148 peptides, and the top seven frameworks corresponding to the ICKs described by Diniz *et al.* [88] represented 960 putative ICKs. The largest cysteine framework recovered was C8.0 with 538 peptides, whereas C6.0, the next largest, represented 123 peptides (Table 2.3).

Table 2.3: Summary of the number of peptides recovered per cysteine framework as well as the corresponding numeral indication designated by Diniz *et al.* [88].

	Dinez		
identifier	Numeral	motif	total
6.0	I	C_1-C_2-C_3C_4-C_5-$C6$	123
8.0	II	C_1-C_2-C_3C_4-C_5XC6-C_7XC_8	538
10.0	V	C_1-C_2-$C_3XC_4C_5$-$C6XC_7$-C_8XC_9-$C10$	117
10.1		C_1-C_2-$C_3C_4C_5$-$C6XC_7$-C_8XC_9-$C10$	23
12.0	VI	C_1-C_2-$C_3XC_4C_5$-$C6XC_7$-C_8XC_9-C_{10}-C_{11}-C_{12}	100
12.1	VII	C_1-C_2-$C_3XC_4C_5XC6$-C_7XC_8-C_9XC_{10}-C_{11}-C_{12}	27
14.0	VIII	C_1-C_2-$C_3XC_4C_5$-$C6XC_7$-C_8XC_9-C_{10}-C_{11}-C_{12}-C_{13}-C_{14}	33

The largest cysteine framework (C8.0), was the most abundant framework for all species except for *Homalonychus theologus*. One framework (C12.1) was only recovered in Psechridae and Ctenidae. A novel cysteine framework (C10.1) not reported by Diniz *et al.* [88], was not recovered in *P. nigriventer*, though it was recovered in five other species of Ctenidae (Table 2.4).

Table 2.4: Number of ICK peptides recovered for each cysteine framework per species.

Family	species	6.0	8.0	10.0	10.1	12.0	12.1	14.0
Homalonychidae	*Homalonychus theologus*	1	4	2	0	9	0	0
Salticidae	*Habronattus signatus*	6	13	2	1	1	0	0
Xenoctenidae	*Odo patricius*	2	21	2	0	3	0	0
Anyphaenidae	*Hibana sp*	3	10	0	1	1	0	0
Gnaphosidae	*Sergiolus capulatus*	1	11	0	1	2	0	0
Thomisidae	*Thomisus spectabilis*	2	13	3	0	2	0	1
Thomisidae	*Misumenoides formosipes*	1	4	0	1	0	0	0
Oxyopidae	*Oxyopes sp*	0	15	9	0	2	0	0
Oxyopidae	*Peucetia longipalpis*	1	11	4	1	0	0	0
Lycosidae	*Hippasa holmerae*	1	11	5	1	3	0	0
Lycosidae	*Pardosa pseudoannulata*	0	7	1	1	0	0	0
Lycosidae	*Schizocosa rovneri*	0	10	0	0	1	0	0
Lycosidae	*Sosippus placidus*	5	26	1	1	3	0	2
Pisauridae	*Nilus albocinctus*	1	8	2	1	3	0	0
Pisauridae	*Sphedanus quadrimaculatus*	1	6	2	1	4	0	0
Pisauridae	*Pisaurina mira*	0	1	0	1	1	0	0
Pisauridae	*Dolomedes triton*	1	11	0	0	10	0	1
Psechridae	*Fecenia protensa*	1	17	4	0	2	1	1
Psechridae	*Psechrus singaporensis*	0	13	3	1	2	1	0
Ctenidae	*Ctenus corniger*	11	19	5	1	3	1	1
Ctenidae	*Anahita punctulata*	2	15	1	0	2	1	1
Ctenidae	*Ctenus captiosus*	4	10	2	1	3	1	2
Ctenidae	*Ctenus exlineae*	2	9	2	1	2	1	1

Family	species	6.0	8.0	10.0	10.1	12.0	12.1	14.0
Ctenidae	*Ctenus hibernalis*	2	10	3	1	3	1	1
Ctenidae	*Isoctenus sp*	3	11	4	0	0	1	2
Ctenidae	*Leptoctenus byrrhus*	1	12	2	1	1	1	1
Ctenidae	*Phoneutria nigriventer*	5	12	3	0	1	1	2

Disulfide connectivity predictions Disulfide connectivity predictions varied greatly between the different prediction approaches; predictions for peptides with fewer cysteines were more consistent between approaches (Figure 2.1). The three disulfide bridges homologous to all cysteine frameworks were cross referenced to peptides with the same cysteine framework in Arachnoserver and used to guide the subsequent multiple sequence alignment (Figure 2.2).

Phylogenetic tests for selection: The first four inner cysteine loops shared by all ICKs were aligned following the schema defined by Table 2.5. This resulted in a multiple sequence alignment with a width of 80 amino acids for 626 peptides with no redundant coding sequences, and no sequences that failed the Chi-squared sequence composition test.

Table 2.5: Multiple sequence alignment schema for ICKs using the four pairs of structurally homologous cysteine residues.

class	loop 1	loop 2	loop 3	loop 4
C6.0	C_1-C_2	C_2——C_3	C_4——C_5	C_5————C_6
C8.0	C_1-C_2	C_2——C_3	C_4——C_5	C_5XC_6-C_7XC_8
C10.0	C_1-C_2	C_2-C_3XC_4	C_5——C_6	C_6XC_7-C_8XC_9
C10.1	C_1-C_2	C_2——C_3	C_4-C_5-C_6	C_6XC_7-C_8XC_9
C12.0	C_1-C_2	C_2-C_3XC_4	C_5——C_6	C_6XC_7-C_8XC_9
C12.1	C_1-C_2	C_2-C_3XC_4	C_5XC_6-C_7	C_7XC_8-C_9XC_{10}

class	loop 1	loop 2	loop 3	loop 4
C14.0	C_1-C_2	C_2-C_3XC_4	C_5——C_6	C_6XC_7-C_8XC_9

Based on the reconstructed phylogeny of (Figure 2.2) all other cysteine frameworks appear to have originated from framework C8.0. Framework C6.0 appears to have evolved via a loss of a pair of cysteines (C_6 and C_7) from the original C8.0 framework. The largest monophyletic grouping of C6.0 was entirely unique to ctenids. The framework C10.1 represents an entirely separate lineage from C10.0 and is monophyletic. The remaining frameworks follow a general trend in terms of their evolutionary history, though they are not all completely monophyletic as some peptides from spurious cysteine frameworks were placed within what would otherwise be clades.

BUSTED, with synonymous rate variation found evidence (LRT, p-value ≤ 0.05) of gene-wide episodic diversifying selection in the entire phylogeny (Table 2.6). Therefore, there is evidence that at least one site on at least one branch has experienced diversifying selection. The site by site variation in test statistics is visualized in Figure 2.3, though BUSTED does not possess the statistical power to infer which specific sites or branches display evidence of episodic diversifying selection.

Table 2.6: A statistical summary of the models fit to the ICK alignment. "Unconstrained model" refers to the BUSTED alternative model for selection, and "Constrained model" refers to the BUSTED null model for selection.

Model	log(likelihood)	parameters	AICc	\$ω1|ω2|ω\$3		
Unconstrained	-37648.7	1169	77691.4	0.06	0.09	3.35
Constrained	-37733.9	1168	77859.6	0.03	0.03	1.00

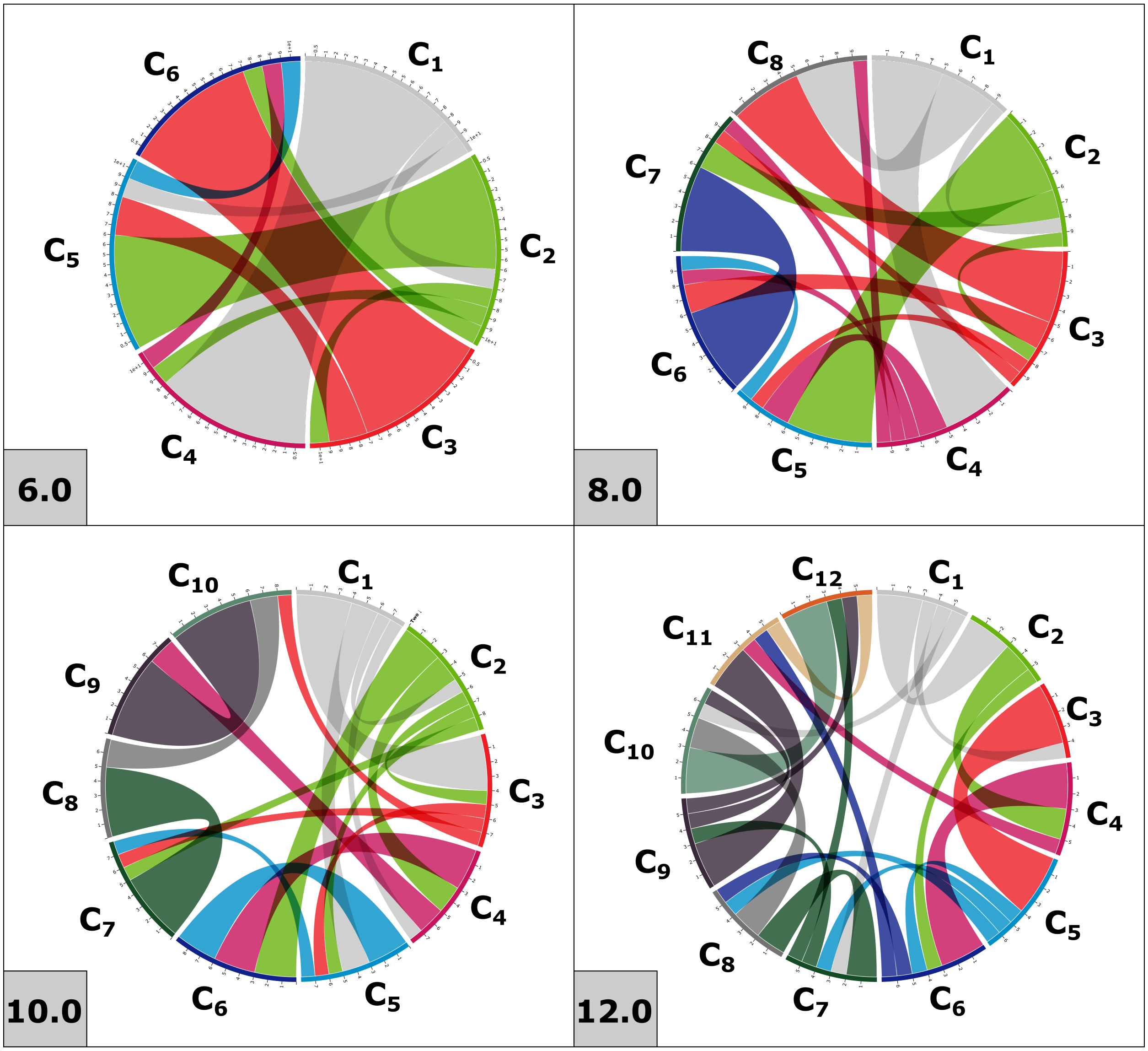

Figure 2.1: Pairwise disulfide connectivity predictions for cysteine motifs 6.0, 8.0, 10.0 and 12.0. Predictions come from a combination of four different prediction methods.

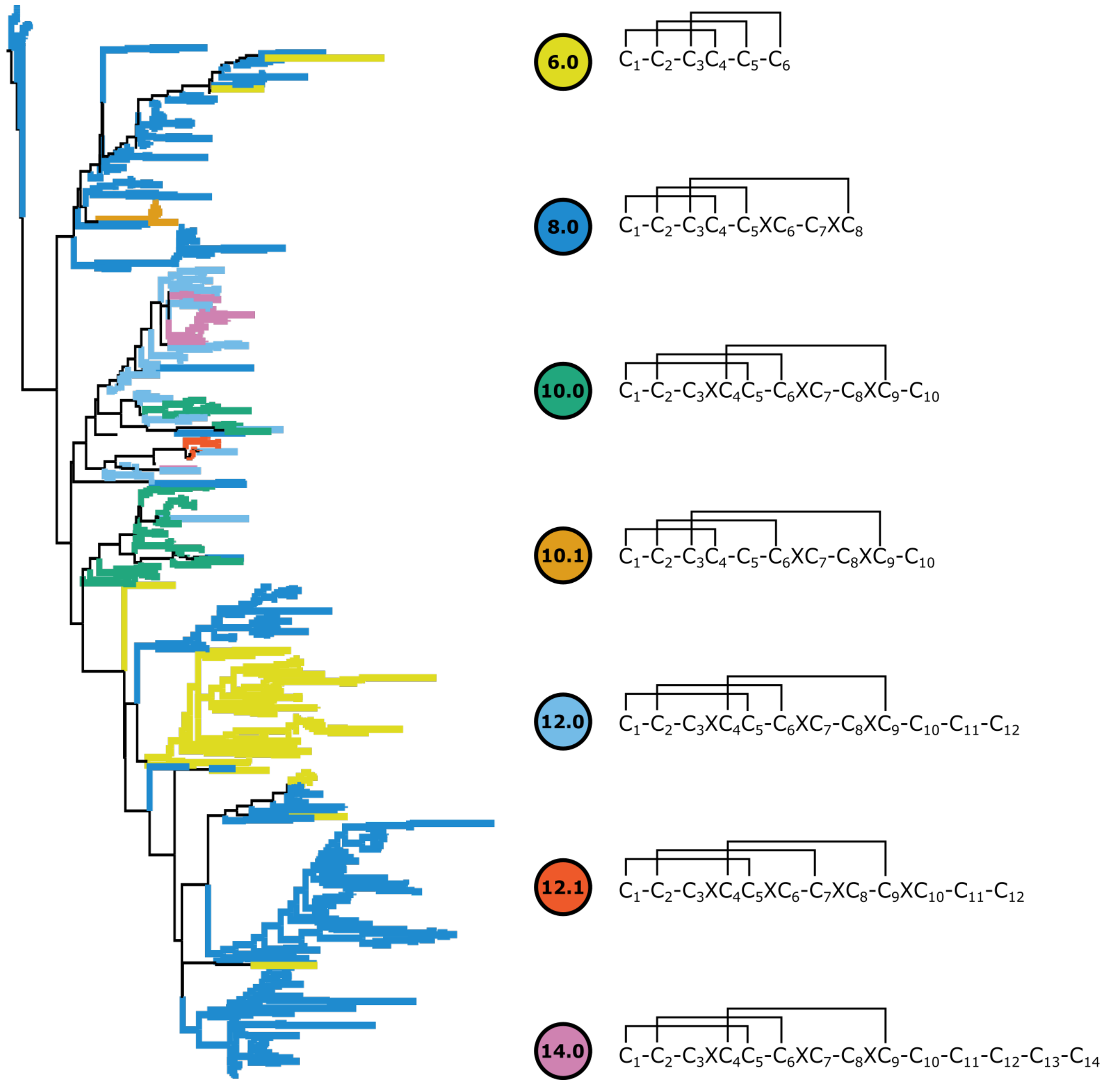
6.0
C_1-C_2-C_3C_4-C_5-C_6
8.0
C_1-C_2-C_3C_4-C_5XC_6-C_7XC_8
10.0
C_1-C_2-$C_3XC_4C_5$-C_6XC_7-C_8XC_9-C_{10}
10.1
C_1-C_2-C_3C_4-C_5-C_6XC_7-C_8XC_9-C_{10}
12.0
C_1-C_2-$C_3XC_4C_5$-C_6XC_7-C_8XC_9-C_{10}-C_{11}-C_{12}
12.1
C_1-C_2-$C_3XC_4C_5XC_6$-C_7XC_8-C_9XC_{10}-C_{11}-C_{12}
14.0
C_1-C_2-$C_3XC_4C_5$-C_6XC_7-C_8XC_9-C_{10}-C_{11}-C_{12}-C_{13}-C_{14}

Figure 2.2: Reconstructed phylogeny of the 626 ICKs recovered from ctenids and lycosoid outgroups. Terminals are colored by their respective cysteine framework. Predicted disulfide connectivities representing all three homologous disulfide bridges shared among all ICK classes are shown to the right.

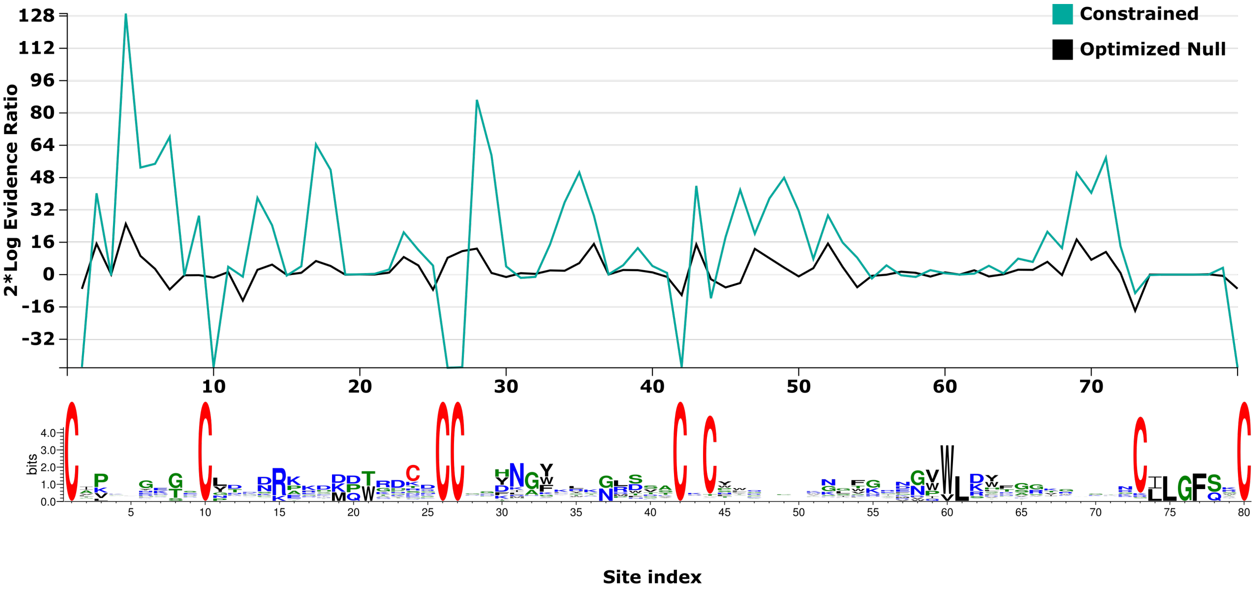

Figure 2.3: Model test statistics per site using 2*Log evidence ratio for BUSTED constrained and optimized null. Sequence logos for the alignment are presented beneath.

FUBAR did not find evidence of pervasive positive/diversifying selection at any sites, but evidence of negative/purifying selection was detected at 57 sites with a posterior probability of 0.9. The line of best-fit from a linear regression of d_S as the independent variable and d_N as the dependent variable for each of the 80 sites had a slope of 0.49 ($F_{1,78} = 44.78, R^2 = 0.36, p = 3.02 \times 10^{-9}$). Only four sites had d_N estimates that exceeded d_S, though the highest posterior probability of those sites was 0.54 (Figure 2.4).

MEME found evidence of positive/diversifying selection under a portion of branches at 12 sites with p-value threshold of 0.05, after correcting for multiple testing (Figure 2.5). Four were within the first loop between the first and second cysteine residues. None were within the second loop between the second and third cysteine residues. Two of those sites were directly upstream of the adjacent pair of cysteines (sites 26 and 27 of the alignment).

aBSREL found evidence of episodic diversifying selection on two out of 1,158 branches in the phylogeny. Significance was assessed using the Likelihood Ratio Test at a threshold of p $\leq$ 0.05, after correcting for multiple testing. One branch was a clade of four peptides with the C8.0 cysteine framework expressed by one ctenid (*C. corniger*), two oxyopids (*Peucetia*

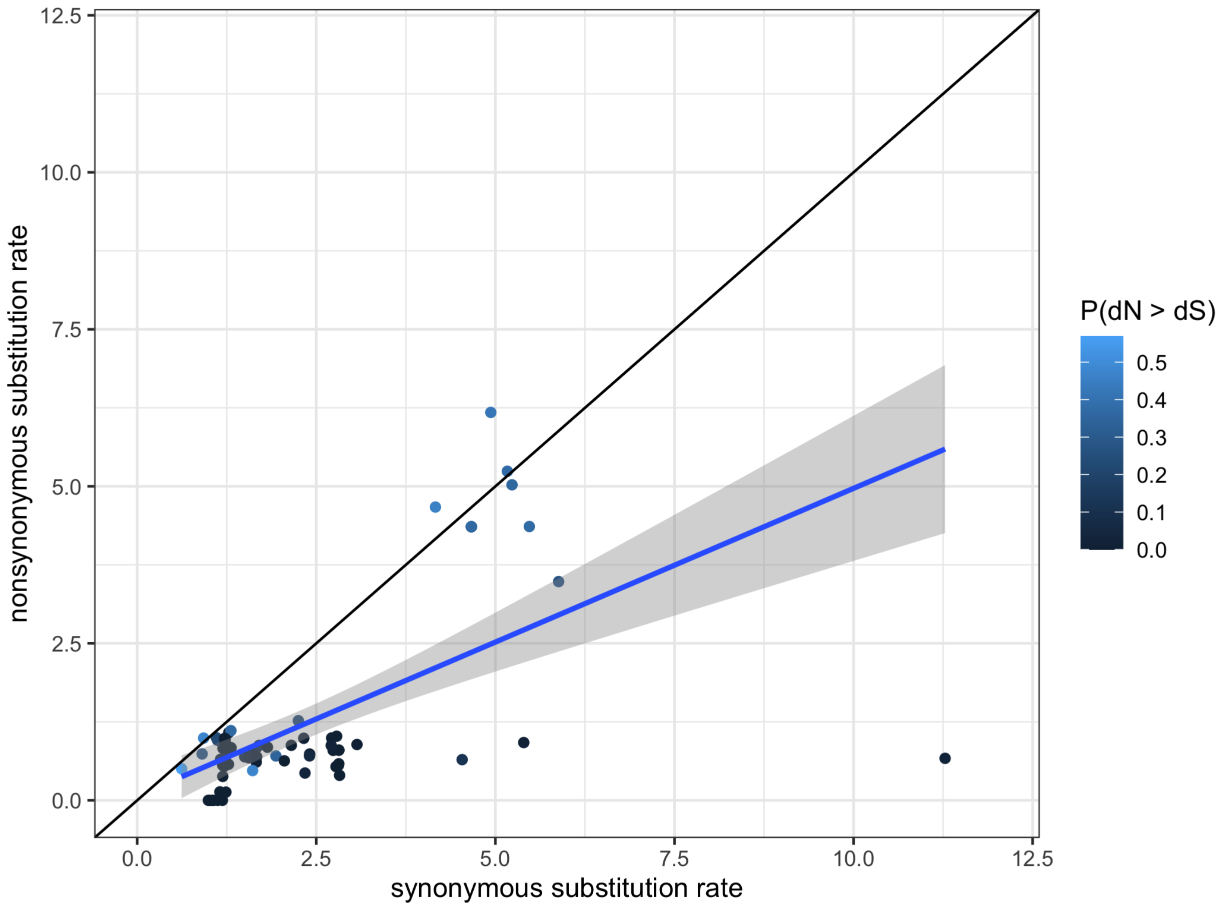

Figure 2.4: Scatter plot of synonymous substitution rate versus nonsynonymous substitution rate for each of the 80 sites of the ICK alignment. The black diagonal line indicates neural evolution where the two rates are equal. Points are colored to indicate the posterior probability that a given site had evidence of pervasive positive selection. Line of best-fit is in blue, with the 95% confidence interval shaded in gray.

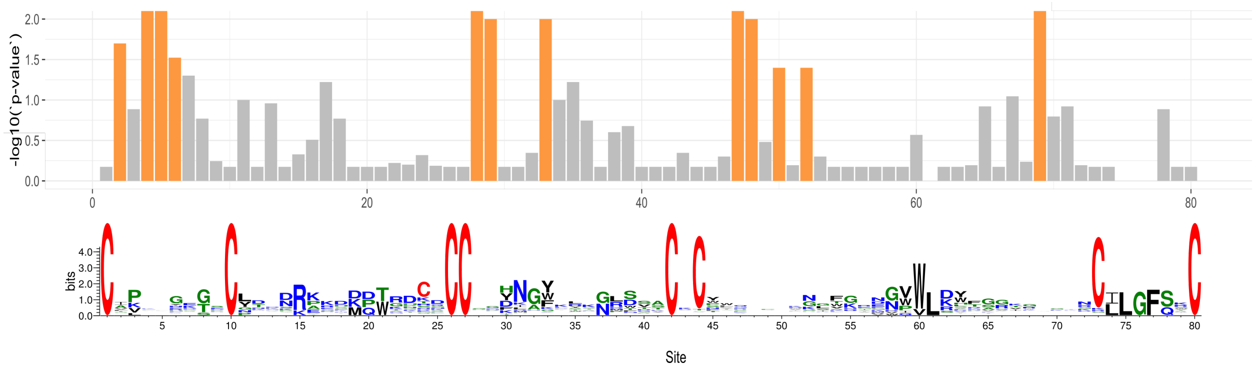

Figure 2.5: Bar plot of negative-log transformed p-values that a portion of branches for each site have evidence of episodic/diversifying selection. Sites with p-values < 0.05 are highlighted in orange. Sequence logo for the alignment presented beneath.

longipalpis and *Oxyopes sp*) and one psechrid (*Fecenia protensa*). The other branch was a clade of 19 peptides with the C14.0 cysteine framework which comprises only ctenids (excluding *C. corniger*).

BGM found 30 pairs of coevolving sites (posterior probability ≥ 0.95), of which 10 had a posterior probability ≥ 0.99. Of the 30 pairs of coevolving sites, 11 were at least three residues apart, whereas the furthest distance between two coevolving sites was 66 amino acid residues. Of particular interest was the fifth amino acid residue, which was found to be coevolving with three other residues, more than any of the others. Site 5, was found to be coevolving with sites 9, 23 , and 35. Site 5 and 9 are both found within the first loop between cysteines one and two, whereas site 23 is right in the middle of the second loop, and site 35 is centrally located in the third loop (Table 2.7).

Table 2.7: Summary of amino acid coevolution analysis. Each of the 11 pairs of amino acid residues found to be coevolving by BGM that had a posterior probability greater than 0.95, and were separated by at least 3 amino acids. Substitutions as well as substitution probabilities are listed for each amino acid pair.

					aa_i	aa_j	Shared
aa_i	aa_j	P($aa_i \rightarrow aa_j$)	P($aa_j \rightarrow aa_i$)	P($aa_i \leftrightarrow aa_j$)	subs	subs	subs
2	9	0	0.99	0.99	155	161	54
4	70	0.01	0.95	0.96	265	163	80
5	9	0.98	0.017	1	193	161	67
5	23	0.95	0	0.95	193	76	37
5	35	0.57	0.42	0.99	193	117	55
7	71	0.01	0.98	0.99	201	165	65
13	28	0.22	0.77	0.99	154	187	66

					aa_i	aa_j	Shared
aa_i	aa_j	P($aa_i \rightarrow aa_j$)	P($aa_j \rightarrow aa_i$)	P($aa_i \leftrightarrow aa_j$)	subs	subs	subs
28	43	0.97	0.01	0.98	187	193	73
38	41	0.29	0.68	0.97	17	23	6
44	73	0.009	0.98	0.99	22	21	8

2.4 Discussion

In this study, we identified and characterized the molecular evolution of 626 unique coding sequences for ICK peptides in wandering spiders and their free-hunting lycosoid relatives. The molecular functionality of these toxins is still unknown for the most part. Increased efforts in neurophysiology assays and molecular modeling will allow broader insights into the evolution of the molecular targets of these toxins. The best disulfide connectivity servers currently available were incredibly imprecise at predicting disulfide connections in spider ICKs. This is especially true for ICK structural elaborations with more than four cysteine pairs, illustrating the need for a spider-specific approach to elucidating structural predictions. Unfortunately, a large bottleneck in making that a possibility is empirical investigations in determining the structure of ICK elaborations in spiders.

It appears that the original ICK toxin in spiders may not be what is typically referred to as the “core” ICK cysteine framework with three pairs of cysteines. Instead, the most abundant cysteine framework with four pairs appears to be the original ICK toxin, and the 6-cysteine framework evolved from the 8-cysteine framework via a loss of a pair of cysteines. Though this study only focused on lycosoid spiders, 8-cysteine toxins appear to be the most abundant throughout the spider tree of life, so a more comprehensive analysis of ICKs across all spiders may yield the same results.

It is postulated that antagonistic co-evolution through predator-prey interactions has shaped venom function via reciprocal selective pressures in an evolutionary “arms race"[65,66,68,118]. Though in rattlesnakes, the predator was observed to be the locally adapted antagonist by being evolutionarily ahead of their prey[119]. It is unclear if spiders are also evolutionarily ahead of their prey, though the fact that they have colonized nearly every terrestrial habitat on earth suggests this is likely the case[120], at least at broad scales. At the protein level, selective pressures on venom have been observed through patterns of rapid evolution of amino acid sequences[121]. More specifically, according to the Rapid Accumulation of Variations in Exposed Residues (RAVER) model of venom evolution, structurally important residues receive strong negative selection while there is a rapid accumulation of variation in the molecular surface of the toxin under a coevolutionary "arms race" scenario[122]. The coevolution of venom resistance in prey and increasingly potent venom in the predator are theorized to exert reciprocal selection pressures.

Broadly speaking, there was evidence of gene-wide episodic diversifying selection in ICK toxins of lycosoid spiders. There was no evidence of pervasive positive selection in any of the codon sites of the ICK alignment. This is consistent with what has been reported in previous tests of pervasive positive selection in spider ICKs[123]. ICKs in spiders date back ˜300 MY, so it is not unusual that ˜70% of the amino acid sites demonstrated evidence of negative selection, because selection "erases its traces" of early bouts of positive selection with persistent negative selection to preserve the potency of the toxin[123,124]. What is particularly striking, though, is that evidence of episodic positive selection was detected in a portion of branches for 12 amino acid sites. This is consistent with the two-speed model of venom evolution proposed by Sunagar & Moran [123], in which positive selection pervades early in venom evolution (such as what is observed in the toxins of contemporary snakes and cone snails) followed by bouts of negative selection, then subsequent bouts of positive selection. These later episodic bouts of positive selection may be indicative of ecological specialization, such as dietary shifts and range expansions, resulting in a rapid diversification of venom arsenal.

Structurally, none of the residues between the first and second cysteine showed evidence of episodic diversification. This could indicate that those residues are necessary to maintain structural integrity and sustain venom potency. One of the two branches on the ICK phylogeny that had evidence of positive selection was a clade of 19 14-cysteine ICKs entirely unique to ctenids. It is possible that this ICK elaboration has played an important role in the range and diet expansion of ctenids. There was also strong evidence of amino acid co-evolution between one residue within the first loop and another amino acid four residues upstream in the same loop and two additional separate residues found midway through the second and third loop. This may indicate that these residues play an important role in the structural integrity or potency of the venom as they had a much higher than expected rate of co-occurrence.

2.5 Conclusion

In this study, we provided evolutionary insights into the ICK toxins of spiders. These insights may prove useful in the field of bioprospecting and peptide design, in which the ICK scaffold is useful for agricultural and pharmacological applications. What remains unresolved are the evolutionary mechanisms giving rise to molecular functions of these toxins, which will become a possibility as more structural and functional assays in spider ICKs are performed. None of the species included in our analysis have publicly available genome sequences, so our analyses relied on incomplete data. However, we demonstrated that these toxins exist as multi-copy gene families across different species. What has yet to be determined are the specific mechanisms that have given rise to these large gene families. Sequencing the genomes of these spiders would provide valuable insights into the evolution of ICK toxins in spiders and finally allow investigations regarding the diversification and formation of cysteine framework elaborations of ICKs in spiders.

Chapter 3

A comparison of the genomic architecture and localization of Inhibitor Cystine Knot toxins in spiders

3.1 Introduction

Genomes hold the key to investigating fundamental biological processes. The advent of *de novo* genome assembly from short-read sequences has allowed for the proliferation of genomic information from numerous species. Such technology is limited, however, to recovering good quality, highly-accurate, and contiguous genomes in only a small percentage of biodiversity, such as haploids and vertebrate animals (e.g., birds and mammals) with genomes of relatively low complexity. For example, venomous animals represent ~15% of animal diversity[125], but because they typically have relatively large and highly repetitive genomes with low GC content, they have received a disproportionately low amount of genome sequencing effort[126].

Spiders comprise the largest group of venomous animals with 48,963recognized species (as of 30 November 2020) spread across nearly every habitat on Earth[6]. Spiders have received recent genome sequencing efforts through the i5k project, which was an initiative to sequence the genomes of 5,000 medically and agriculturally relevant arthropods. Through this initiative, in 2013, the genomes of three species of spiders were published[127]. This included two of the medically relevant species that live in the U.S, the brown recluse (*Loxosceles reclusa* (Family: Sicariidae) Gertsch & Mulaik, 1940) and the western black widow (*Latrodectus hesperus* (Family: Theridiidae) Chamberlin & Ivie, 1935). Additionally, this initiative provided genomic information for the common house spider, *Parasteatoda tepidariorum* (Family: Theridiidae) (C. L. Koch, 1841). Since then, the genomes of seven additional species have been sequenced. In 2014, the genomes of the social velvet spider, *Stegodyphus mimosarum* (Family: Eresidae) Pavesi, 1883 and the Brazilian whiteknee tarantula *Acanthoscurria geniculata* (Family: Theraphosidae) (C.L. Koch, 1841) were seqeunced[128]. In 2017, the genome for the golden orb weaver, *Trichonephila clavipes* (Family: Araneidae) (Linnaeus, 1767) was sequenced[129]. Finally, in 2019 the genomes of *Dysdera silvatica* (Family: Dysderidae) Schmidt, 1981, *Anelosimus studiosus* (Family: Theridiidae) (Hentz, 1850), *Pardosa pseudoannulata* (Family: Lycosidae) (Bösenberg & Strand, 1906), *Stegodyphus dumicola* (Family: Eresidae) Pocock, 1898 were sequenced and made available to the public[130–133].

These recent advances in spider genomics provide an unprecedented opportunity to answer long standing evolutionary questions. For instance, the evolutionary origins of the inhibitor cystine knot toxins expressed by spiders can now be investigated using genomic information. Inhibitory cystine knot (ICK) neurotoxins comprise a large fraction of the noxious components found in venomous invertebrates (e.g., spiders, scorpions, cone snails, and robber flies). These cysteine rich toxins take on an unusually stable conformation via the formation of disulfide bonds and often target voltage gated ion channels[97]. It is largely accepted that ICKs were independently recruited to the venom arsenal of scorpions and spiders hundreds of millions of years ago. What remains puzzling, however, is that the evolutionary trajectories of scorpion ICKs followed a very different path than spiders ICKs. Scorpion ICKs are, for the most part, single copy genes with the same 3-loop core ICK topological motif[134]. Spider ICKs exist as complex multi-copy gene families in which one species may express up to 100 homologs (as is the case with *Phoneutria nigriventer*)[88]. There are at least eight structural elaborations that have evolved with up to seven additional pairs of cysteines that form disulfide bridges, and a double ICK knot has also evolved at least once[89]. These structural elaborations are entirely unique to spiders, a contributing factor to the broad functional diversity of spider venom[135].

Due to the multi-copy nature of ICK toxins in spiders, it is expected that gene duplication events have played a role in their diversification. It is also likely that alternative splicing of the same genetic locus has contributed to the observed diversity of ICKs, though this has only been verified in one toxin from *Parasteatoda tepidariorum* that expressed two structural variations as alternative isoforms from the same genetic locus[136]. Thus, it is expected that gene duplications play a larger role than alternative splicing. Furthermore, the specific mechanisms of duplication that underly ICK diversity in spiders remains unclear. Whole genome duplications have been documented in spiders, though at least seven whole genome duplication events would be required to account for the number of ICKs expressed per species[137]. What seems likely, is that there was a combination of whole genome duplications that amplified localized tandem duplication events. Tandem duplication has been identified in the ICK toxins expressed by *Pardosa pseudoannulata* and *Stegodyphus mimosarum*, though this has not been fully investigated in the genome sequences of other species. The evolutionary mechanisms driving patterns of tandem duplication in ICKs has thus far remained uninvestigated. In this study, we implement an approach to (1) identify ICK peptides in all publicly available spider genome assemblies to highlight the current state of spider genome assemblies and their annotations and (2) to provide preliminary insights into the genomic architecture and patterns of molecular evolution of ICK toxins in spiders.

3.2 Methods

Data Retrieval and Sequence Analysis: All available spider genome assemblies, along with annotations and protein sequences, were retrieved from NCBI's public online database. To provide comparative benchmarks for genome assembly completeness, the assemblies were provided as input to the Benchmarking sets of Universal Single-Copy Orthologs (BUSCO v3.0.2) using the arachnid dataset[54]. Other genome assembly statistics were recorded using seqkit v0.13.2[138], which included the distribution of scaffold length and GC content.

Inhibitor cystine knot toxin identification: To identify inhibitor cystine knot toxins in the genome assemblies, a database of verified ICKs from spiders was retrieved from the KNOTTIN database[97]. Only ICKs with complete coding sequences and verified disulfide connectivity were retained in the final verified ICK database. The database was provided as input to BLASTp to search against the protein sequences from the genome annotations[98]. Additionally, a multiple sequence alignment was generated from the verified ICK database to create a Hidden Markov Model (HMM) that could be searched against the genomic protein sequences using HMMER v3.3.1[99]. For both BLASTp and HMMER, only matches with at least six cysteines and up to 200 amino acids in length were kept for downstream analysis. Additionally, putative matches were only kept if they contained a signal peptide as indicated by signalP v5.0, which predicts the presence of signal peptide cleavage sites[100].

To determine if the total number of recovered ICKs was directly proportional to the percentage of complete BUSCOs recovered, a linear regression was performed using percent complete BUSCO as the independent variable and number of ICKs as the dependent variable for each species using the *lm* function in R v3.6.3. To correct for the incompleteness of the genome assembly, the expected relative number of ICKs (rICKs) was calculated using the following expression $rICK = \frac{\%BUSCO}{totalICK}$. To evaluate if there was a disproportionate number of rICKs between species, a Chi-Squared goodness of fit test was implemented.

To explore the evolutionary history of duplicated ICK toxins, only toxins with the core ICK motif was retained that had a duplicate present on the same scaffold. Any ICK elaborations with additional cysteines were present on the same scaffold were retained as well. The same outgroup from Pineda *et al.* [89] (a disulfide-directed -hairpin expressed by the whip scorpion *Mastigoproctus giganteus*) was included, and then the multiple sequence alignment was inferred using MAFFT v7.221 with the automatic setting[55,89]. A phylogenetic tree was reconstructed using IQTREE v1.6.10 with the default settings[57]. Nodes whose descendants belonged to the same scaffold were denoted as tandem duplication nodes; nodes that represented a speciation event were labeled as speciation nodes; and all others were labeled as other duplication event.

To provide further insights into ICKs belonging to a tandem duplicate cassette versus those with only one copy per scaffold, a Welch two-sample one-sided t-test was implemented to test if the length of scaffolds with duplicated toxins with the core ICK motif were longer than those with only one copy. Further, to evaluate motifs in the noncoding DNA at the exon-intron junction, a sequence logo was constructed using the eight flanking nucleotides of the intron donor and acceptor sites using WebLogo v3.7.4[139].

3.3 Results

Genome size, scaffold length and GC content: All genome assemblies were larger than one Gb in total length, with an average of 2.84 Gb (s.d = 1.79). The assembly for *Acanthoscurria geniculata* was over double the average genome size, at 7.17 Gb. Assembly methods and sequencing technology between species varied greatly; likewise the distribution of assembled scaffold length between species was highly variable (Table 3.1). The average N50 was 568.9 kb (s.d = 1.25 Mb).

Table 3.1: Summary statistics for the genome assemblies for each species. Species with asterisks indicate assemblies with no publicly available annotations. N50 is the median scaffold length.

Species (Family)	Assembly	size (Gb)	N50 (kb)	%GC	BUSCO	ICKs
T. clavipes	GCA_002102615.1	2.44	63.0	22.4	76.7	2
D. silvatica	GCA_006491805.1	1.36	38.0	35.1	75.1	3
S. dumicola	GCA_010614865.1	2.55	254.1	33.5	91.4	17
S. mimosarum	GCA_000611955.2	2.74	480.6	34.7	96.3	24
L. reclusa	GCA_001188405.1	3.26	63.2	19.6	39.2	1
L. hesperus	GCA_000697925.1	1.14	13.9	18.7	49.6	5
P. tepidariorum	GCF_000365465.2	1.45	4,055.4	20.9	97.3	17
*P. pseudoannulata**	GCA_008065355.1	4.21	711.4	31.9	89.0	
*A. studiosus**	GCA_008297655.1	2.03	4.8	20.9	36.5	
*A. geniculata**	GCA_000661875.1	7.18	20.3	38.6	32.8	

The mean GC content for all species was 27.6% (s.d = 7.74%). The lowest was *Latrodectus_hesperus*, with a mean GC content of 18.7% (s.d = 8.33%). The species with the most variable GC content was *Trichonephila clavipes* (relative standard deviation of 53.3%). The assemblies with the least variation in GC content were *Dysdera silvatica* and *Stegodyphus dumicola*, (relative standard deviations of 7.0% and 6.3%, respectively). Interestingly, GC content skew had a noticeable effect on the assembled scaffold length. For all species, scaffolds with GC-skew of nearly zero were higher than those with GC-skew greater than or less than zero, forming a bell-shaped relationship (Figure 3.1).

The average percentage of single copy complete BUSCO genes from the assemblies of each species was 66.5% (s.d = 23.8%). Three species had assemblies with BUSCO scores

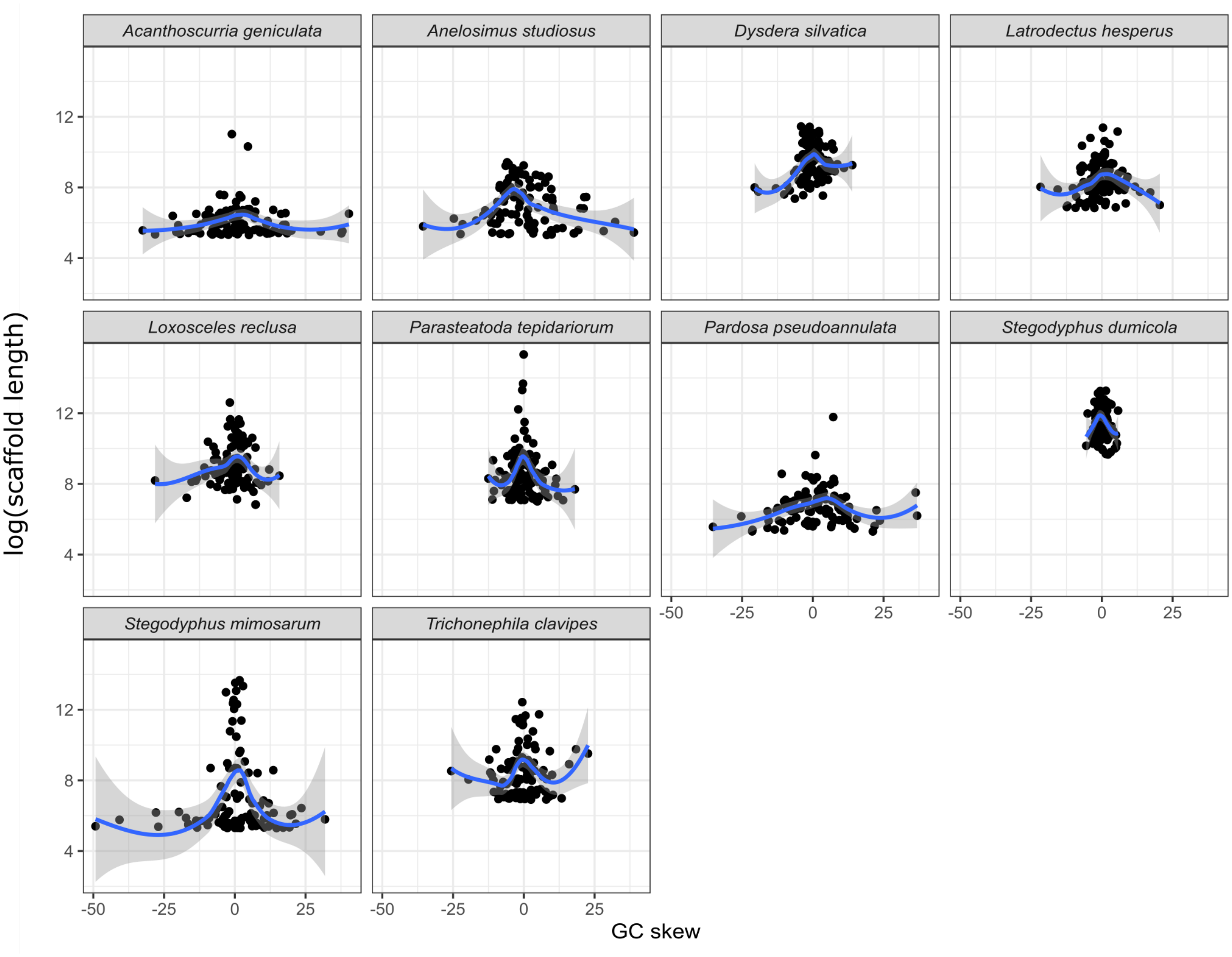

Figure 3.1: Scatter plot of scaffold length in relationship to GC-skew, with smoothing line created using loess function and shaded region representing a 95% confidence interval range.

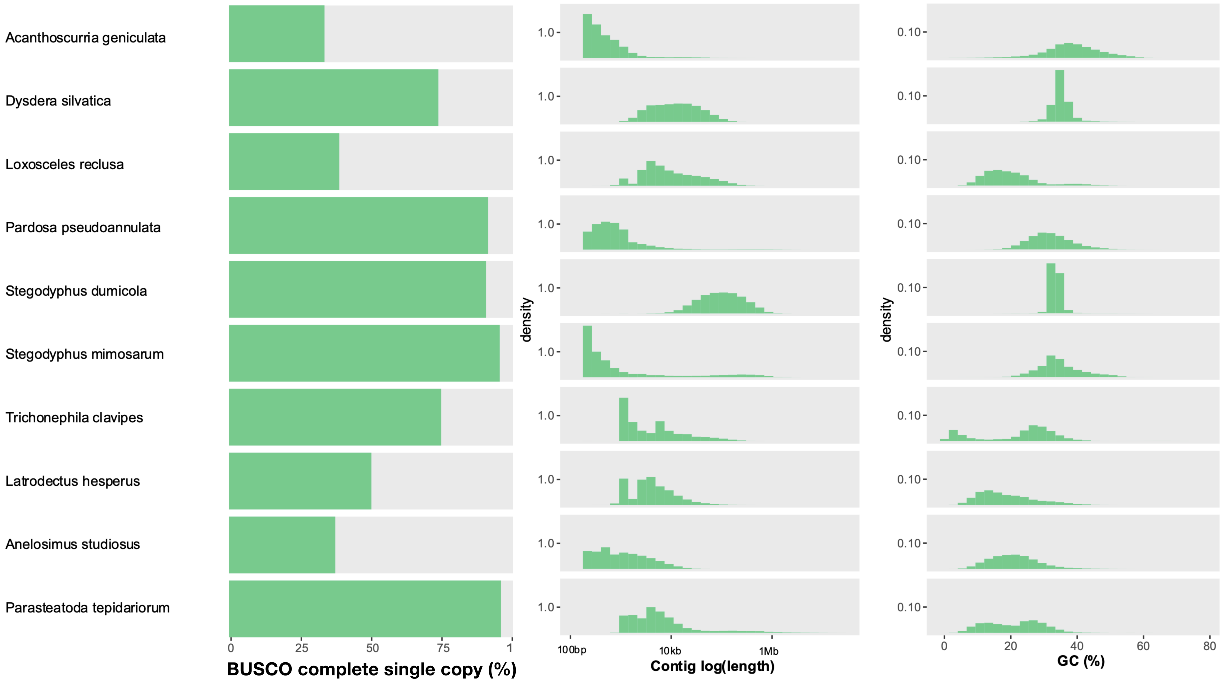

Figure 3.2: Distribution of Single copy BUSCO hits, scaffold length, and GC content of scaffolds for all genome assemblies across the ten spider species

less than 40% (*Acanthoscurria geniculata*, *Anelosimus studiosus*, and *Loxosceles reclusa*), whereas two species had BUSCO scores greater than 90% (*Pardosa_pseudoannulata*, and *Stegodyphus mimosarum*), as is visualized in Figure 3.2.

Inhibitor cystine knot toxin identification: A total of 69 putative ICK toxins were recovered from the seven species whose genomes contained sufficient annotations for analysis. Of those, 30 were found as duplicates on the same scaffold, and six were splice variants. The number of ICKs recovered was directly proportional to the percentage of complete BUSCOs recovered in the genome ($F_{1,5} = 9.15, R^2 = 0.65, p = 0.03$), as is visualized in Figure 3.3. A comparison of ICKs recovered for each species was derived by calculating the number of ICKs divided by the percentage of complete BUSCOs, henceforth referred to as relative ICK (rICK). The average rICK was 11.9 (s.d = 9.3). The highest rICK was found in *Stegodyphus mimosarum* with 25.8, wheras the lowest was from *Loxosceles reclusa* with 2.6. There was a significant bias in the number of rICKs identified between species ($\chi^2_6 = 42.2, p = 1.68\times10^{-7}$).

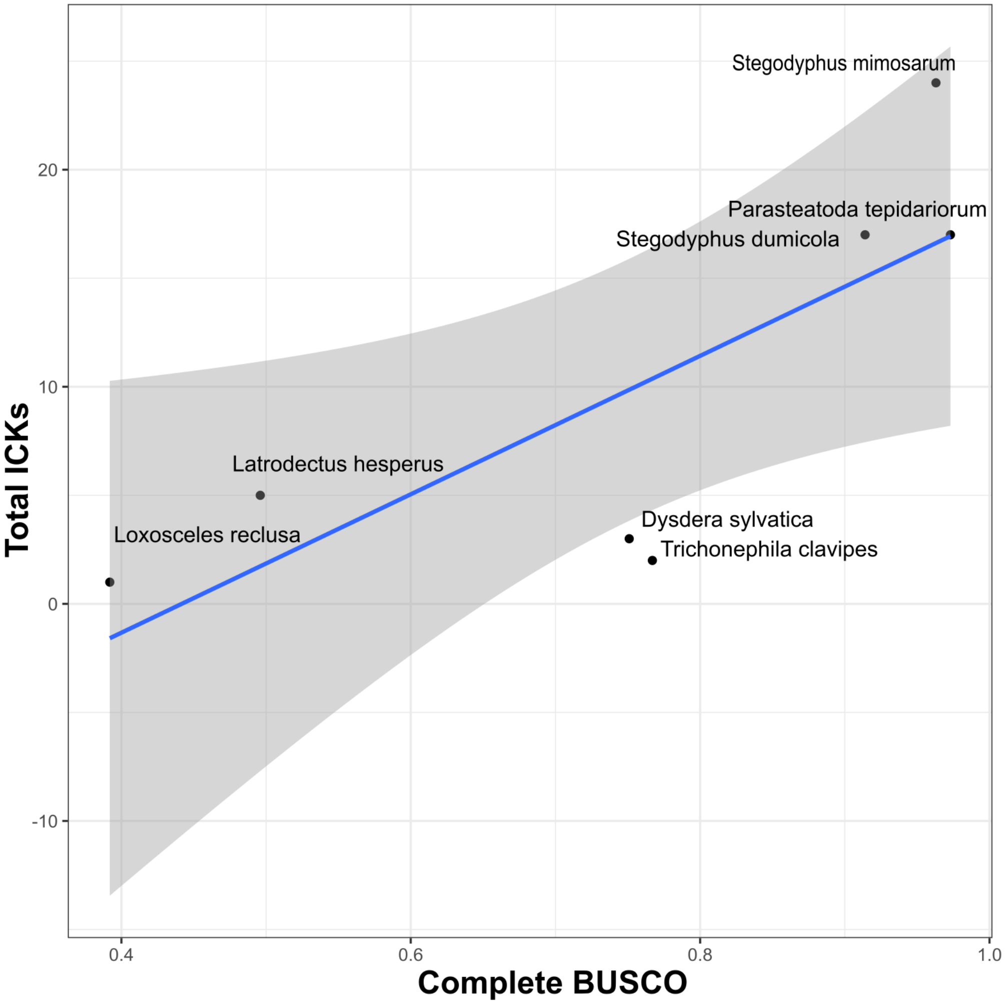

Figure 3.3: Scatter plot of number of ICKs versus percent complete BUSCO genes recovered from the genome assemblies of all species. Shaded region represents the 95% confidence interval range.

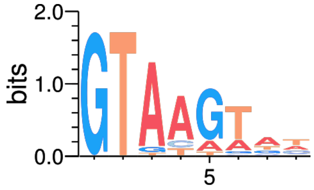

Figure 3.4: Sequence logo of the six nucleotides downstream of the GT donor site of the intronic sequences for toxins containing the core ICK motif.

The 69 putative ICKs identified comprised a total of 16 unique cysteine motifs, 11 of which were singleton cysteine motifs. The most abundant cysteine motif was the 8-cysteine core ICK motif, which had 31 members. Four of these had a 9th cysteine but still maintained the same core motif. The core ICK motif was not recovered in *T. clavipes*, *L. reclusa*, or *D. sylvatica*. Only two peptides with the core ICK motif were recovered from *L. reclusa*, whereas 12, nine and eight were recovered from *P. tepidariorum*, *S. mimosarum*, and *S. dumicola*, respectively. Of the 31 peptides with the core ICK motif, 16 belonged to a duplicated gene cassette. All core ICK motif containing genes contained at least two introns, with one having five. All ICK gene introns contained the canonical GT donor and AG acceptor sites. Nucleotide diversity increased incrementally downstream from the GT donor site (Figure 3.4), whereas there was an increased abundance of thymine upstream of the AG acceptor site (Figure 3.5). A total of six ICK gene cassettes were recovered from *P. tepidariorum*, *S. mimosarum*, and *S. dumicola*, though none of them included additional ICK peptides belonging to additional ICK motif elaborations. The average length of scaffolds containing duplicate ICKs was 2.4 times greater than scaffolds containing a single ICK and the median length was 4.9 times greater (Figure 3.6), though the difference was not statistically significant ($t = 1.01, df = 8.80, p = 0.17$).

The reconstructed phylogeny of the tandem duplicated toxins containing the core ICK motif from *P. tepidariorum*, *S. mimosarum*, and *S. dumicola* indicated that tandem duplication events occurred before and after speciation events (Figure 3.7). A total of eight tandem duplication events were identified that would have given rise to the six tandem duplicate

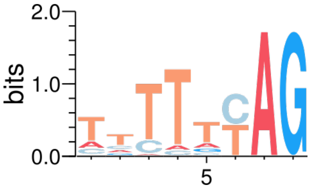

Figure 3.5: Sequence logo of the six nucleotides upstream of the AG acceptor site of the intronic sequences for toxins containing the core ICK motif.

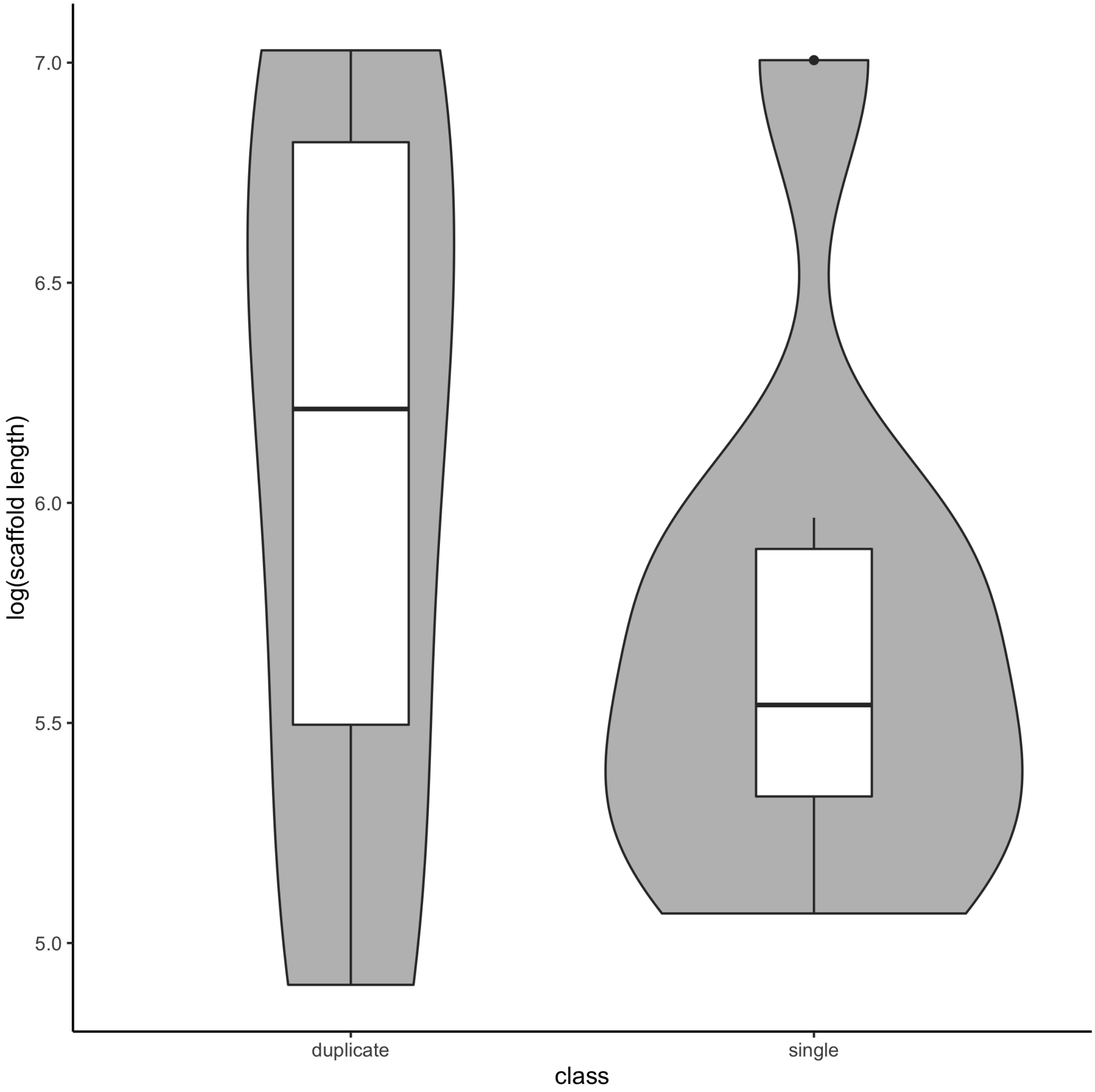

Figure 3.6: Box plot of the log transformed length of scaffolds containing either duplicated ICKs or single copy ICKs.

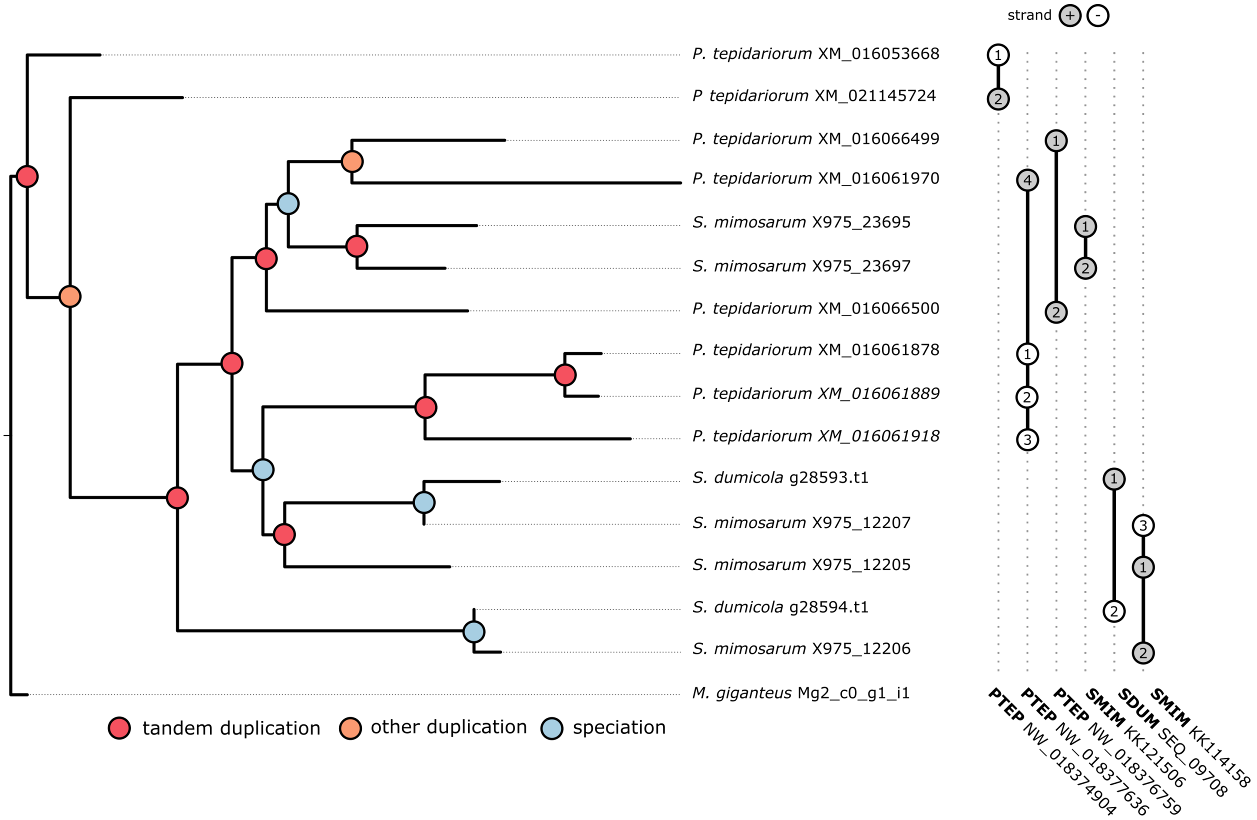

Figure 3.7: Reconstructed phylogeny of duplicated toxins containing the core ICK motif. Nodes are colored to indicate an inferred tandem duplication event, other duplication event, or speciation event. Dots connected by bars to the right indicate toxins that are on the same scaffold, numbers indicate synteny order from 5'-3', gray indicates positive strand and white indicates negative strand. Scaffold names are indicated at the bottom right.

cassetes. Four of those six duplicated cassettes included duplicates present on both strands.

3.4 Discussion

Spider genomes are large, AT rich, and full of repetitive DNA that has contributed to their assemblies being highly variable in terms completeness and fragmentation[126]. It was previously observed that GC content is significantly positively correlated with the median scaffold length in the genome assemblies of arthropods[140]. We explored this further at the individual scaffold level and determined a sharp trend in the effect that GC skew has on the length of an assembled scaffold. This indicates that GC content acts as a limiting factor that must be overcome through technological advancement (e.g., long-read versus short-read sequencing platforms and advances in bioinformatic tools) in order to assemble highly contiguous genome assemblies in spiders. The BUSCO scores for certain assemblies were strikingly low and had notable impacts on our ability to investigate ICK toxins. At present, re-sequencing efforts in species with BUSCO scores lower than 50% is the only way to alleviate this, as no computational methods currently available would be adequate to overcome the aforementioned issues.

There is a large discrepancy in the ICK toxins recovered in the genome sequences and those reported from transcriptomic and proteomic investigations. We were only able to identify two putative ICK toxins from the genome annotations of the brown recluse (*L. reclusa*), though there are at least four verified ICKs from proteomic investigations that were curated and deposited into the KNOTTIN database[97]. Proteomic and transcriptomic investigations in the Brazilian wandering spider *Phoneutria nigriventer* reported nearly 100 ICK toxins expressed among ten ICK functional classes[88].). We were unable to recover even two-thirds of that number from all seven spider genome assemblies combined. No genome assembly exists for *P. nigriventer*, but it may be possible that ICK copy number is highly variable between species, resulting in a larger than expected number of ICK variants in *P. nigriventer*. We detected a significant bias in the relative number of ICKs recovered when correcting for genome assembly completeness, though a proper inference regarding the variability in ICK copy numbers between spiders will require improved genome assemblies. A combination of deeper sequencing efforts with longer reads and improved annotation methods could greatly reduce this discrepancy. Gene models may need to be custom tailored to identifying ICKs, specifically in genome assemblies, much in the way that the BUSCO pipeline identifies BUSCO genes.

Despite the limitations in the completeness and fragmentation of spider genome assemblies, we were able to develop a preliminary understanding of the evolutionary history of ICKs in spider genomes. The presence of numerous duplicated ICKs on the same scaffold are indicative of large-scale duplication, but because scaffolds do not represent chromosomes, the true frequency of tandem duplication may be underestimated. In fact, of the 31 core ICK peptides recovered, scaffolds with a single ICK were 2.3 Mb shorter on average than scaffolds containing duplicated ICKs, though the difference was not statistically significant, likely due to reduced statistical power and a small sample size. It seems likely that improved assembly techniques to generate larger scaffolds would potentially recover more tandemly duplicated ICKs and certainly provide better insight into the chromosomal localization of these toxins.

Based on the analysis of ICKs from the genome assemblies of *S. mimosarum*, *S. dumicoa*, and *P. tepidariorum*- it is already evident that tandem duplication events have played an important role in the diversification of ICKs in spiders. What is particularly striking is that numerous potential tandem duplication events have been identified before and after speciation events amongst these three species. If that trend continues in a broader taxonomic sampling, then it would indicate that tandem duplication events have occurred all throughout the evolution of spiders. Other duplication events have likely played an important role in concert with tandem duplications as well. An early whole genome duplication, such as the one that occurred early in the evolutionary history of spiders, may have amplified pre-existing tandem duplication cassettes[137].

3.5 Conclusion

With increased sequencing efforts, ICKs in spiders will be uniquely positioned as an invaluable model for exploring the interplay of numerous aspects of genome and molecular evolution. Duplication events have significant effects on genomic architecture, as well as the tempo of chromosome evolution[141]. Additionally, duplication events act as a source of functional novelty. The two most common classifications of gene duplication (whole genome duplication and tandem duplication) have played a critical role in the evolution of ICKs in spiders allowing for an expanded exploration of peptide space. Spider ICKs can also provide a test case scenario for the ortholog conjecture, to determine the extent to which duplication events lead to greater evolutionary change than speciation events through the process of neofunctionalization and subfunctionalization. Additionally, with improved bioinformatic method development catered to ICKs, it may be possible to identify pseudogenes, allowing for a more comprehensive exploration into the fate of a duplicated gene. With nearly 50,000 species to choose from, spider genomes provide a valuable opportunity to explore the evolution of neurotoxins. This will undoubtedly translate to improved peptide space exploration in the field of bioprospecting, which has already found numerous agricultural and biomedical applications for ICK peptides.

References

1. Höfer, H. & Brescovit, A. D. Species and guild structure of a Neotropical spider assemblage (Araneae) from Reserva Ducke, Amazonas, Brazil. *Andrias* **15,** 99–119 (2001).

2. Griswold, C. E., Ramirez, M. J., Coddington, J. A. & Platnick, N. I. Atlas of phylogenetic data for entelegyne spiders (Araneae: Araneomorphae: Entelegynae) with comments on their phylogeny. *Proceedings of the California Academy of Sciences* (2005).

3. Davila, D. S. Higher-level relationships of the spider family Ctenidae (Araneae: Ctenoidea). *Bulletin of the American Museum of natural History* **2003,** 1–86 (2003).

4. Jocqué, R., Samu, F. & Bird, T. Density of spiders (Araneae: Ctenidae) in Ivory Coast rainforests. *Journal of Zoology* **266,** 105–110 (2005).

5. Polotow, D. & Brescovit, A. D. Revision of the neotropical spider genus Gephyroctenus (Araneae: Ctenidae: Calocteninae). *Revista Brasileira de Zoologia* **25,** 705–715 (2008).

6. World Spider Catalog, W. *World spider catalog, version 19.5.* 2018.

7. Rego, F. N., Venticinque, E. M. & Brescovit, A. D. Densidades de aranhas errantes (Ctenidae e Sparassidae, Araneae) em uma floresta fragmentada. *Biota Neotropica* **5,** 45–52 (2005).

8. Rego, F. N., Venticinque, E. M. & Brescovit, A. D. Effects of forest fragmentation on four Ctenus spider populations (Araneae: Ctenidae) in central Amazonia, Brazil. *Studies on Neotropical Fauna and Environment* **42,** 137–144 (2007).

9. Brazil, V. & Vellard, J. A. *Contribuição ao estudo do veneno das aranhas* (Insituto Butantan, 1925).

10. Bucaretchi, F., Bertani, R., De Capitani, E. & Hyslop, S. Envenomation by wandering spiders (Genus Phoneutria). *Clin. Tox* **63,** 1–49 (2016).

11. Lachmuth, U., Grasshoff, M. & Barth, F. G. Taxonomische revision der gattung Cupiennius Simon 1891 (Arachnida: Araneae: Ctenidae). Revisión taxonómica del género Cupiennius Simon 1891 (Arachnida: Araneae: Ctenidae). *Senckenbergiana Biologica.* **65,** 329–372 (1985).

12. Simó, M. & von Eickstedt, V. Revisión de la sistematica del género Asthenoctenus Simon, 1897 (Araneae, Ctenidae). *Arachnologia* **22,** 1–12 (1994).

13. Brescovit, A. D. & Simó, M. On the Brazilian Atlantic Forest species of the spider genus Ctenus Walckenaer, with the description of a neotype for C. dubius Walckenaer (Araneae, Ctenidae, Cteninae). *Arachnology* **14,** 1–17 (2007).

14. Simó, M. & Brescovit, A. D. Revision and cladistic analysis of the Neotropical spider genus Phoneutria Perty, 1833 (Araneae, Ctenidae), with notes on related Cteninae. *BULLETIN-BRITISH ARACHNOLOGICAL SOCIETY* **12,** 67–82 (2001).

15. Martins, R. & Bertani, R. The non-Amazonian species of the Brazilian wandering spiders of the genus Phoneutria Perty, 1833 (Araneae: Ctenidae), with the description of a new species. *Zootaxa* **1526,** 1–36 (2007).

16. Polotow, D. & Brescovit, A. D. Phylogenetic relationships of the Neotropical spider genus Itatiaya (Araneae). *Zoologica Scripta* **40,** 187–193 (2011).

17. Polotow, D. & Brescovit, A. D. Phylogenetic analysis of the tropical wolf spider subfamily Cteninae (Arachnida, Araneae, Ctenidae). *Zoological Journal of the Linnean Society* **170,** 333–361 (2014).

18. Polotow, D., Carmichael, A. & Griswold, C. E. Total evidence analysis of the phylogenetic relationships of Lycosoidea spiders (Araneae, Entelegynae). *Invertebrate Systematics* **29,** 124–163 (2015).

19. Polotow, D., Brescovit, A. D., *et al.* Revision of the new wandering spider genus Ohvida and taxonomic remarks on Celaetycheus Simon, 1897 (Araneae: Ctenidae). *Zootaxa* **2115,** 1–20 (2009).

20. Peck, W. & WB, P. The Ctenidae of temperate zone North America (1981).

21. Hentz, N. M. Descriptions and figures of the araneides of the United States. *Boston J. nat. Hist.* **5,** 443–478 (1847).

22. Gertsch, W. J. New American spiders with notes on other species. American Museum novitates; no. 805 (1935).

23. Sissom, W., Peck, W. & Cokendolpher, J. New records of wandering spiders from Texas, with a description of the male of Ctenus valveriensis (Areneae: Ctenidae). *Entomological news* (1999).

24. Simon, E. *Etudes arachnologiques. 21e Mémoire. XXIX. Descriptions d'espèces et de genres nouveaux de l'Amérique centrale et des Antilles* in *Annales de la Société entomologique de France* **6** (1888), 203–216.

25. Casewell, N. R., Wüster, W., Vonk, F. J., Harrison, R. A. & Fry, B. G. Complex cocktails: the evolutionary novelty of venoms. *Trends in ecology & evolution* **28,** 219–229 (2013).

26. Herzig, V. & Hodgson, W. C. Neurotoxic and insecticidal properties of venom from the Australian theraphosid spider Selenotholus foelschei. *Neurotoxicology* **29,** 471–475 (2008).

27. Eggs, B., Wolff, J. O., Kuhn-Nentwig, L., Gorb, S. N. & Nentwig, W. Hunting without a web: how lycosoid spiders subdue their prey. *Ethology* **121,** 1166–1177 (2015).

28. Scotland, R. W., Olmstead, R. G. & Bennett, J. R. Phylogeny reconstruction: the role of morphology. *Systematic Biology* **52,** 539–548 (2003).

29. Lee, M. S. & Palci, A. Morphological phylogenetics in the genomic age. *Current Biology* **25,** R922–R929 (2015).

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

31. Wheeler, W. C. *et al.* The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. *Cladistics* **33,** 574–616 (2017).

32. Fernandez, R. *et al.* Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Current Biology* **28,** 1489–1497 (2018).

33. Agnarsson, I., Gregorič, M., Blackledge, T. A. & Kuntner, M. The phylogenetic placement of P sechridae within E ntelegynae and the convergent origin of orb-like spider webs. *Journal of Zoological Systematics and Evolutionary Research* **51,** 100–106 (2013).

34. Bayer, S. & Schönhofer, A. L. Phylogenetic relationships of the spider family Psechridae inferred from molecular data, with comments on the Lycosoidea (Arachnida: Araneae). *Invertebrate Systematics* **27,** 53–80 (2013).

35. Blackledge, T. A., Kuntner, M., Marhabaie, M., Leeper, T. C. & Agnarsson, I. Biomaterial evolution parallels behavioral innovation in the origin of orb-like spider webs. *Scientific Reports* **2,** 833 (2012).

36. Moradmand, M., Schönhofer, A. L. & Jäger, P. Molecular phylogeny of the spider family Sparassidae with focus on the genus Eusparassus and notes on the RTA-clade and 'Laterigradae'. *Molecular phylogenetics and evolution* **74,** 48–65 (2014).

37. Cheng, D.-Q. & Piel, W. H. The origins of the Psechridae: Web-building lycosoid spiders. *Molecular phylogenetics and evolution* **125,** 213–219 (2018).

38. Bond, J. E. *et al.* Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Current Biology* **24,** 1765–1771 (2014).

39. Meng, X., Zhang, Y., Bao, H. & Liu, Z. Sequence analysis of insecticide action and detoxification-related genes in the insect pest natural enemy Pardosa pseudoannulata. *PloS one* **10,** e0125242 (2015).

40. MacManes, M. D. The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ* **6,** e5428 (2018).

41. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34,** i884–i890 (2018).

42. Hart, T., Komori, H. K., LaMere, S., Podshivalova, K. & Salomon, D. R. Finding the active genes in deep RNA-seq gene expression studies. *BMC genomics* **14,** 778 (2013).

43. Longo, M. S., O'Neill, M. J. & O'Neill, R. J. Abundant human DNA contamination identified in non-primate genome databases. *PLoS One* **6,** e16410 (2011).

44. Lusk, R. W. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PloS one* **9,** e110808 (2014).

45. Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* **2,** e675 (2014).

46. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics* **32,** 3196–3198 (2016).

47. Edgar, R. C. UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. *bioRxiv,* 088666 (2016).

48. Borner, J. & Burmester, T. Parasite infection of public databases: a data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies. *BMC genomics* **18,** 100 (2017).

49. Lafond-Lapalme, J., Duceppe, M.-O., Wang, S., Moffett, P. & Mimee, B. A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm. *Bioinformatics* **33,** 1293–1300 (2017).

50. Ballenghien, M., Faivre, N. & Galtier, N. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC biology* **15,** 1–16 (2017).

51. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14,** 417 (2017).

52. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8,** 1494 (2013).

53. Haas, B., Papanicolaou, A., *et al.* TransDecoder (find coding regions within transcripts). *Github, nd https://github. com/TransDecoder/TransDecoder (accessed May 17, 2018)* (2015).

54. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).

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

56. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).

57. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32,** 268–274 (2015).

58. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30,** i541–i548 (2014).

59. Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., *et al.* lmerTest package: tests in linear mixed effects models. *Journal of statistical software* **82,** 1–26 (2017).

60. McDonald, J. H. *Handbook of biological statistics* (sparky house publishing Baltimore, MD, 2009).

61. Marques, E. S., Vasconcelos-Netto, J. & de Mello, M. B. Life history and social behavior of Anelosimus jabaquara and Anelosimus dubiosus (Araneae, Theridiidae). *Journal of Arachnology,* 227–237 (1998).

62. Altman, J. S. & Kien, J. Suboesophageal neurons involved in head movements and feeding in locusts. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **205,** 209–227 (1979).

63. Wullschleger, B. & Nentwig, W. Influence of venom availability on a spider's prey-choice behaviour. *Functional ecology* **16,** 802–807 (2002).

64. Nelsen, D. R., Kelln, W. & Hayes, W. K. Poke but don't pinch: risk assessment and venom metering in the western black widow spider, Latrodectus hesperus. *Animal Behaviour* **89,** 107–114 (2014).

65. Van, V. & Van Valen, L. A new evolutionary law. (1973).

66. Dawkins, R. & Krebs, J. R. Arms races between and within species. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **205,** 489–511 (1979).

67. Endler, J. A. *Natural selection in the wild* (Princeton University Press, 1986).

68. Daltry, J. C., Wüster, W. & Thorpe, R. S. Diet and snake venom evolution. *Nature* **379,** 537–540 (1996).

69. Heatwole, H. & Poran, N. S. Resistances of sympatric and allopatric eels to sea snake venoms. *Copeia,* 136–147 (1995).

70. Biardi, J. E. & Coss, R. G. Rock squirrel (Spermophilus variegatus) blood sera affects proteolytic and hemolytic activities of rattlesnake venoms. *Toxicon* **57,** 323–331 (2011).

71. Jansa, S. A. & Voss, R. S. Adaptive evolution of the venom-targeted vWF protein in opossums that eat pitvipers. *PLoS One* **6,** e20997 (2011).

72. Barlow, A., Pook, C. E., Harrison, R. A. & Wüster, W. Coevolution of diet and prey-specific venom activity supports the role of selection in snake venom evolution. *Proceedings of the Royal Society B: Biological Sciences* **276,** 2443–2449 (2009).

73. Pacheco, E. O., Ferreira, V. G., Pedro, F. M. S. R. & Santana, D. J. Predation on Scinax crospedospilus (Anura: Hylidae) by Phoneutria nigriventer (Aranae: Ctenidae) in a Atlantic Forest fragment in the South-east of Brazil. *Herpetology Notes* **9,** 315–316 (2016).

74. Foerster, N. E., Carvalho, B. H. G. & Conte, C. E. Predation on Hypsiboas bischoffi (Anura: Hylidae) by Phoneutria nigriventer (Araneae: Ctenidae) in southern Brazil. *Herpetology Notes* **10,** 403–404 (2017).

75. Wang, H. *et al.* The venom of the fishing spider Dolomedes sulfurous contains various neurotoxins acting on voltage-activated ion channels in rat dorsal root ganglion neurons. *Toxicon* **65,** 68–75 (2013).

76. Calvete, J. J. Venomics: integrative venom proteomics and beyond. *Biochemical Journal* **474,** 611–634 (2017).

77. Prashanth, J. R., Hasaballah, N. & Vetter, I. Pharmacological screening technologies for venom peptide discovery. *Neuropharmacology* **127,** 4–19 (2017).

78. Harris, R. J. & Jenner, R. A. Evolutionary ecology of fish venom: adaptations and consequences of evolving a venom system. *Toxins* **11,** 60 (2019).

79. Kordiš, D. & Gubenšek, F. Adaptive evolution of animal toxin multigene families. *Gene* **261,** 43–52 (2000).

80. Todd, E. V., Black, M. A. & Gemmell, N. J. The power and promise of RNA-seq in ecology and evolution. *Molecular ecology* **25,** 1224–1241 (2016).

81. Fry, B. G. *et al.* Functional and structural diversification of the Anguimorpha lizard venom system. *Molecular & Cellular Proteomics* **9,** 2369–2390 (2010).

82. Von Reumont, B. M. Studying smaller and neglected organisms in modern evolutionary venomics implementing RNASeq (transcriptomics)—A critical guide. *Toxins* **10,** 292 (2018).

83. Arújo, D. A., Cordeiro, M. N., Diniz, C. R. & Beirão, P. S. Effects of a toxic fraction, PhTx 2, from the spider Phoneutria nikriventer on the sodium current. *Naunyn-Schmiedeberg's archives of pharmacology* **347,** 205–208 (1993).

84. Gomez, M. V., Kalapothakis, E., Guatimosim, C. & Prado, M. A. Phoneutria nigriventer venom: a cocktail of toxins that affect ion channels. *Cellular and molecular neurobiology* **22,** 579–588 (2002).

85. Richardson, M. *et al.* Comparison of the partial proteomes of the venoms of Brazilian spiders of the genus Phoneutria. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **142,** 173–187 (2006).

86. Nunes, K. P. *et al.* Tx2-6 toxin of the Phoneutria nigriventer spider potentiates rat erectile function. *Toxicon* **51,** 1197–1206 (2008).

87. Cole, T. J., Buszka, P. A., Mobley, J. A. & Hataway, R. A. Characterization of the proteome for the wandering spider, Ctenus hibernalis (Aranea: Ctenidae). *Toxicon* **100,** 373–374 (2016).

88. Diniz, M. R. *et al.* An overview of Phoneutria nigriventer spider venom using combined transcriptomic and proteomic approaches. *PloS one* **13** (2018).

89. Pineda, S. S. *et al.* Structural venomics reveals evolution of a complex venom by duplication and diversification of an ancient peptide-encoding gene. *Proceedings of the National Academy of Sciences* **117,** 11399–11408 (2020).

90. Fry, B. G. *et al.* The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annual review of genomics and human genetics* **10,** 483–511 (2009).

91. Barrio, A. & Brazil, O. V. Ein neues verfahren der Giftentnahme bei spinnen. *Experientia* **6,** 112–113 (1950).

92. Munekiyo, S. M. & Mackessy, S. P. Effects of temperature and storage conditions on the electrophoretic, toxic and enzymatic stability of venom components. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **119,** 119–127 (1998).

93. Binford, G. J. & Wells, M. A. The phylogenetic distribution of sphingomyelinase D activity in venoms of Haplogyne spiders. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **135,** 25–33 (2003).

94. Clarke, T. H. *et al.* Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. *BMC genomics* **15,** 365 (2014).

95. Davidson, N. M., Hawkins, A. D. & Oshlack, A. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome biology* **18,** 148 (2017).

96. McIlwain, S. *et al.* Crux: rapid open source protein tandem mass spectrometry analysis. *Journal of proteome research* **13,** 4488–4491 (2014).

97. Gelly, J.-C. *et al.* The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucleic acids research* **32,** D156–D159 (2004).

98. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25,** 3389–3402 (1997).

99. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39,** W29–W37 (2011).

100. Armenteros, J. J. A. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* **37,** 420–423 (2019).

101. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC bioinformatics* **12,** 116 (2011).

102. Rubinstein, R. & Fiser, A. Predicting disulfide bond connectivity in proteins by correlated mutations analysis. *Bioinformatics* **24,** 498–504 (2008).

103. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16,** 404–405 (2000).

104. Ferrè, F. & Clote, P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic acids research* **33,** W230–W232 (2005).

105. Ceroni, A., Passerini, A., Vullo, A. & Frasconi, P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic acids research* **34,** W177–W181 (2006).

106. Yang, J., He, B.-J., Jang, R., Zhang, Y. & Shen, H.-B. Accurate disulfide-bonding network predictions improve ab initio structure prediction of cysteine-rich proteins. *Bioinformatics* **31,** 3773–3781 (2015).

107. Liu, Z.-L., Hu, J.-H., Jiang, F. & Wu, Y.-D. CRiSP: accurate structure prediction of disulfide-rich peptides with cystine-specific sequence alignment and machine learning. *Bioinformatics* **36,** 3385–3392 (2020).

108. Bostock, M., Ogievetsky, V. & Heer, J. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* **17,** 2301–2309 (2011).

109. Herzig, V. *et al.* ArachnoServer 2.0, an updated online resource for spider toxin sequences and structures. *Nucleic acids research* **39,** D653–D657 (2010).

110. Shafee, T. M., Robinson, A. J., van der Weerden, N. & Anderson, M. A. Structural homology guided alignment of cysteine rich proteins. *Springerplus* **5,** 1–7 (2016).

111. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).

112. Pond, S. L. K. & Muse, S. V. in *Statistical methods in molecular evolution* 125–181 (Springer, 2005).

113. Murrell, B. *et al.* Gene-wide identification of episodic selection. *Molecular biology and evolution* **32,** 1365–1371 (2015).

114. Murrell, B. *et al.* FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution* **30,** 1196–1205 (2013).

115. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8,** e1002764 (2012).

116. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular biology and evolution* **32,** 1342–1353 (2015).

117. Poon, A. F., Lewis, F. I., Frost, S. D. & Kosakovsky Pond, S. L. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* **24,** 1949–1950 (2008).

118. Endler, J. Defence against predators. *Predator-prey relationships* (1986).

119. Holding, M. L., Biardi, J. E. & Gibbs, H. L. Coevolution of venom function and venom resistance in a rattlesnake predator and its squirrel prey. *Proceedings of the Royal Society B: Biological Sciences* **283,** 20152841 (2016).

120. Turnbull, A. Ecology of the true spiders (Araneomorphae) (1973).

121. Juárez, P., Comas, I., González-Candelas, F. & Calvete, J. J. Evolution of snake venom disintegrins by positive Darwinian selection. *Molecular biology and evolution* **25,** 2391–2407 (2008).

122. Sunagar, K. *et al.* Three-fingered RAVERs: Rapid Accumulation of Variations in Exposed Residues of snake venom toxins. *Toxins* **5,** 2172–2208 (2013).

123. Sunagar, K. & Moran, Y. The rise and fall of an evolutionary innovation: contrasting strategies of venom evolution in ancient and young animals. *PLoS Genet* **11,** e1005596 (2015).

124. Haller, B. C. & Hendry, A. P. Solving the paradox of stasis: squashed stabilizing selection and the limits of detection. *Evolution* **68,** 483–500 (2014).

125. Holford, M., Daly, M., King, G. F. & Norton, R. S. Venoms to the rescue. *Science* **361,** 842–844 (2018).

126. Garb, J. E., Sharma, P. P. & Ayoub, N. A. Recent progress and prospects for advancing arachnid genomics. *Current opinion in insect science* **25,** 51–57 (2018).

127. I5K Consortium. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity* **104,** 595–600 (2013).

128. Sanggaard, K. W. *et al.* Spider genomes provide insight into composition and evolution of venom and silk. *Nature communications* **5,** 1–12 (2014).

129. Babb, P. L. *et al.* The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. *Nature Genetics* **49,** 895–903 (2017).

130. Sánchez-Herrero, J. F. *et al.* The draft genome sequence of the spider Dysdera silvatica (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. *GigaScience* **8,** giz099 (2019).

131. Purcell, J. & Pruitt, J. N. Are personalities genetically determined? Inferences from subsocial spiders. *BMC genomics* **20,** 1–10 (2019).

132. Yu, N. *et al.* Genome sequencing and neurotoxin diversity of a wandering spider Pardosa pseudoannulata (pond wolf spider). *bioRxiv,* 747147 (2019).

133. Liu, S., Aagaard, A., Bechsgaard, J. & Bilde, T. DNA methylation patterns in the social spider, Stegodyphus dumicola. *Genes* **10,** 137 (2019).

134. Santibáñez-López, C. E. *et al.* Integration of phylogenomics and molecular modeling reveals lineage-specific diversification of toxins in scorpions. *PeerJ* **6,** e5902 (2018).

135. Saez, N. J. *et al.* Spider-venom peptides as therapeutics. *Toxins* **2,** 2851–2871 (2010).

136. Haney, R. A., Matte, T., Forsyth, F. S. & Garb, J. E. Alternative transcription at venom genes and its role as a complementary mechanism for the generation of venom complexity in the common house spider. *Frontiers in ecology and evolution* **7,** 85 (2019).

137. Schwager, E. E. *et al.* The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC biology* **15,** 1–27 (2017).

138. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one* **11,** e0163962 (2016).

139. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14,** 1188–1190 (2004).

140. Brewer, M. S., Cotoras, D. D., Croucher, P. J. & Gillespie, R. G. New sequencing technologies, the development of genomics tools, and their applications in evolutionary arachnology. *The Journal of Arachnology* **42,** 1–15 (2014).

141. Durand, D. Vertebrate evolution: doubling and shuffling with a full deck. *TRENDS in Genetics* **19,** 2–5 (2003).