1 **Venom-gland transcriptomics and venom proteomics of the black-**
2 **back scorpion (*Hadrurus spadix*) reveal detectability challenges**
3 **and an unexplored realm of animal toxin diversity**

4 Darin R. Rokyta^{a,*} and Micaiah J. Ward^a

5 ^a Department of Biological Science, Florida State University, Tallahassee, Florida 32306

6 * Corresponding author: drokyta@bio.fsu.edu

7 **Running title:** The venom of *Hadrurus spadix*

8 **Keywords:** Scorpion, venom, transcriptome, proteome

9 **Corresponding author:**

10 Darin R. Rokyta

11 Florida State University

12 Department of Biological Science

13 319 Stadium Dr.

14 Tallahassee, FL USA 32306-4295

15 email: drokyta@bio.fsu.edu

16 phone: (850) 645-8812

Abstract

The order Scorpiones is one of the most ancient and diverse lineages of venomous animals, having originated approximately 430 million years ago and diversified into 14 extant families. Although partial venom characterizations have been described for numerous scorpion species, we provided the first quantitative transcriptome/proteome comparison for a scorpion species using single-animal approaches. We sequenced the venom-gland transcriptomes of a male and female black-back scorpion (*Hadrurus spadix*) from the family Caraboctonidae using the Illumina sequencing platform and conducted independent quantitative mass-spectrometry analyses of their venoms. We identified 79 proteomically confirmed venom proteins, an additional 69 transcripts with homology to toxins from other species, and 596 nontoxin proteins expressed at high levels in the venom glands. The venom of *H. spadix* was rich in antimicrobial peptides, K⁺-channel toxins, and several classes of peptidases. However, the most diverse and one of the most abundant classes of putative toxins could not be assigned even a tentative functional role on the basis of homology, indicating that this venom contained a wealth of previously unexplored animal toxin diversity. We found good agreement between both transcriptomic and proteomic abundances across individuals, but transcriptomic and proteomic abundances differed substantially within each individual. Small peptide toxins such as K⁺-channel toxins and antimicrobial peptides proved challenging to detect proteomically, at least in part due to the significant proteolytic processing involved in their maturation. In addition, we found a significant tendency for our proteomic approach to overestimate the abundances of large putative toxins and underestimate the abundances of smaller toxins.

Introduction

Modern approaches for high-throughput screening for biomedically useful or promising molecules is significantly increasing the importance of characterizations of animal venom components, particularly those from the many taxa that have received little or no study. The purification and characterization of individual components from whole venoms has traditionally been a rate-limiting step in the process of realizing the biomedical potential of venom components, and the additional constraint of low venom yields has limited the use of many taxa comprising smaller species. Zhang et al. (2016) recently described a high-throughput screening procedure that involves construction of venom libraries starting only from a toxin-sequence database. Their approach was used, as a test case, to successfully identify toxins that would bind with high affinity to Kv1.3 channels. If DNA/RNA sequences of toxins can be mined directly for desirable functional activities through this or similar approaches, venom-gland transcriptomic studies of previously uncharacterized taxa, even those posing no serious threat to humans, have significantly improved practical utility in the biomedical sciences.

Our generally accepted understanding of the long-term coevolutionary dynamics between venomous animals and their prey (e.g., Duda and Palumbi, 1999; Gibbs and Rossiter, 2008) has recently been challenged, necessitating further study of venom composition and evolution in ancient venomous lineages. Sunagar and Moran (2015) argued that two opposing modes of venom evolution were apparent in the sequences of animal venoms. The venom components of snakes, cone snails, and other venom neophytes have long been known to evolve at high rates within and between species (e.g., Duda and Palumbi, 1999; Duda and Remigio, 2008; Lynch, 2007; Gibbs and Rossiter, 2008; Margres et al., 2015). These high evolutionary rates are generally attributed to strong, continual positive selection related to incessant coevolution with prey (or predators). Sunagar and Moran (2015) suggested that older lineages tend instead toward slower evolutionary rates dictated by purifying selection. Unfortunately, the available data are strongly biased toward the younger venomous lineages (e.g., snakes and cone snails). The older venomous lineages (e.g., centipedes, scorpions, and spiders) have received far less study relative to the younger lineages. In contrast to work in venomous snakes (e.g., Sanz et al., 2006; Rokyta et al., 2013, 2015b), few data are available on evolutionary rates of venoms between closely related taxa, preventing a systematic comparison of evolutionary rates over comparable time scales in both old and young venomous lineages. Power to detect positive selection in DNA sequences decreases with high levels of sequence divergence (Anisimova et al., 2001) because multiple substitutions at sites ultimately become undetectable. We need comparable sampling of old and new venomous lineages to properly test the hypothesis of Sunagar and Moran (2015).

The scorpions are among the oldest terrestrial and venomous lineages, and, although the venoms of certain taxa have been well characterized, the vast majority of scorpion toxin biodiversity remains unexplored. The order Scorpiones originated approximately 430 million years ago and diversified into 14 extant recognized families (Soleglad and

Fet, 2003). Sanger-sequencing-based, partial venom-gland transcriptomes have been described for members of the families Buthidae (Ruiming et al., 2010; Ma et al., 2012; Alvarenga et al., 2012; Valdez-Velázquez et al., 2013; Diego-García et al., 2014; Mille et al., 2014), Caraboctonidae (Schwartz et al., 2007), Chaerilidae (He et al., 2013), Euscorpiidae (Ma et al., 2009, 2012), Scorpionidae (Ma et al., 2010; Diego-García et al., 2012), and Vaejovidae (Quintero-Hernández et al., 2015). Little venom-gland transcriptomic work has been conducted using high-throughput sequencing (e.g., 454 or Illumina), and this work has been limited to the families Buthidae (Rendón-Anaya et al., 2012; de Oliveira et al., 2015) and Urodacidae (Luna-Ramírez et al., 2015). A complete characterization of a venom requires both a high-coverage venom-gland transcriptome and a proteomic analysis of the corresponding venom, particularly for poorly studied groups for which homology-dependent toxin identification is likely to fail for at least some venom components. Such a full analysis of venom has not, to our knowledge, been completed for a scorpion species.

We provided the first high-throughput venom-gland transcriptomic and venom proteomic characterization of a scorpion in the family Caraboctonidae. The genus *Hadrurus* within the family Caraboctonidae, commonly known as the giant hairy scorpions, is native to Mexico and the United States of America (Stahnke, 1945) and is comprised of some of the largest scorpion species in North America and the New World, reaching sizes of 12–13 cm and 20–25 g (Francke and Prendini, 2008). These scorpions are primarily burrowers (Stahnke, 1945), and, despite their impressive sizes, are relatively harmless to humans. The venoms of species of *Hadrurus* are nonlethal to rats, and human envenomations typically result only in local swelling, pain, and discoloration (Stahnke, 1945). A low-coverage venom-gland transcriptome for *Hadrurus gertschi* (Schwartz et al., 2007) was previously sequenced by means of Sanger sequencing, but no proteomic verification of any of the toxins was provided. The black-back scorpion (*Hadrurus spadix*) is the northernmost species of this genus and is characterized by its unique, solid-black prosoma. We independently sequenced venom-gland transcriptomes from a male and female of *H. spadix* and conducted a parallel LC-MS/MS analysis of their venoms to provide one of the most complete characterizations of a scorpion venom to date.

Materials and Methods

Scorpions, venoms, and venom glands

Specimens were collected in Owyhee County in southern Idaho near the northern edge of the range of *H. spadix* (Soleglad et al., 2011). Venom proteomic and venom-gland transcriptomic analyses were performed independently on two specimens labeled C0195 and C0196. Sex was determined by means of pectine teeth counts and by examination of the proportional sizes of the most proximal parts of the middle lamellae (Stahnke, 1945). On the basis of these criteria, C0195 was determined to be female (pectine counts: 30 right and 29 left, and wide middle lamellae relative to pectine length), and C0196 was

determined to be male (pectine counts: 33 right and 32 left, and narrow middle lamellae relative to pectine length).

Venom was extracted by means of electrical stimulation at the base of the telson, following anesthesia through exposure to CO₂ for five minutes. Venom was lyophilized and stored at -80°C . Venom glands were removed four days after venom extraction using a Fisher Scientific 420 series zoom stereoscopic microscope and micro-surgical dissection instruments. To remove venom glands, scorpions were first anesthetized for approximately 15 minutes by means of exposure to CO₂. Once fully anesthetized, the entire metasoma and attached telson (tail and stinger) were removed using surgical scissors. The metasoma was taped to a sterile surface so the telson would be visible under the microscope lens. The telson was cut using a micro-surgical blade from the base of the telson to the tip, slightly curved out at each end so the final cut resembled an "I". Tweezers were used to peel back each side of the telson to access the venom glands. The venom glands were removed using curved surgical tweezers by scraping out the inside of each telson side. Gland tissue was immediately transferred to 100 μl of RNAlater. The glands in RNAlater were held at 4°C overnight and transferred to -80°C until further use. The body and metasoma (with removed venom glands) of each scorpion were preserved in 95% ethanol and stored at -80°C .

Transcriptome sequencing

Scorpion venom-gland RNA extraction was performed as previously described (Rokyta et al., 2011). Briefly, 500 μl of Trizol (Invitrogen) was added to the 100 μl RNAlater containing the venom-gland tissue. Tissue was homogenized while submerged in Trizol using a sterile syringe with a 20 gauge needle. We then added 500 μl of Trizol, followed by 20% chloroform and centrifugation in phase lock heavy gel tubes (5Prime) to separate the RNA from DNA and other cellular debris. Isolated RNA was pelleted with isopropyl alcohol and washed with 75% ethanol. Purified RNA was then washed with 70% ethanol, and its quality was verified using a Bioanalyzer with an RNA 6000 Pico Kit (Agilent Technologies) following the manufacturer's instructions. The total RNA yields for C0195 and C0196 based on Bioanalyzer results were 6,377.5 ng and 1,849.8 ng, respectively.

Approximately 100 ng of total RNA was used to isolate mRNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs). A fragmentation time of 15.5 minutes was used to achieve fragment sizes of approximately 370 nucleotides (adapter-ligated), which is appropriate for 150-nucleotide paired-end sequencing. The purified mRNA was immediately used for cDNA library preparation using the NEBNext Ultra RNA Library Prep Kit and Multiplex Oligos for Illumina (New England Biolabs). PCR was performed using the NEBNext High-Fidelity 2X Hot Start PCR Master Mix and 14 cycles of PCR to achieve the desired DNA concentration for sequencing. DNA was purified using Agencourt AMPure XP PCR Purification Beads throughout and at the end of the protocol. The library quality was assessed using a Bioanalyzer with a High Sensitivity DNA Kit (Agilent Technologies) following the manufacturer's instructions.

The total cDNA yield for C0195 was 31.7 ng with an average fragment size of 384 bp, and the total yield for C0196 was 216.2 ng with an average fragment size of 410 bp. To determine the amplifiable concentration, KAPA PCR was performed on individual samples by the Florida State University Molecular Cloning Facility. The amplifiable concentrations for C0195 and C0196 were determined to be 5.23 nM and 23.18 nM, respectively. These samples were pooled with cDNA libraries from other samples to achieve the desired final concentration for sequencing (~ 5 nM) such that each individual library was equally represented. The quality of the pooled DNA samples was assessed using a Bioanalyzer with a High Sensitivity DNA Kit (Agilent Technologies), and an additional round of KAPA PCR was performed to confirm amplifiable concentration of the pooled sample prior to sequencing. Sequencing was performed on an Illumina HiSeq 2000 by the Florida State University College of Medicine Translational Science Laboratory.

Proteomics

Protein concentrations were quantified using the Qubit Protein Assay kit with a Qubit 1.0 Fluorometer (Thermo Fisher Scientific). Approximately 5 μ g of whole venom was digested using the Calbiochem ProteoExtract All-in-One Trypsin Digestion Kit (Merck, Darmstadt, Germany) according to the manufacturer's instructions and using LC/MS grade solvents. This digestion resulted in an overall yield of approximately 4.3 μ g of digested venom protein. Digested samples were frozen and then dried using a SpeedVac for 4.5 hours.

Digested tryptic peptides were resuspended in 0.1% formic acid at a final concentration of 250 ng/ μ l. Three digested *Escherichia coli* proteins were used as internal standards. We used 2,500 fmol of P31697 (Chaperone protein FimC), 250 fmol of P31658 (Protein deglycase 1), and 25 fmol of P00811 (Beta-lactamase ampC) per injection. For the three internal standards, highly purified recombinant proteins were purchased from Abcam at known concentrations and mixed in the specified proportions (1000 $\times$) prior to digestion. The digested peptide mix was infused into samples to yield the desired final concentrations. A 2 μ l aliquot was run on LC-MS/MS as follows. An externally calibrated Thermo Q Exactive HF (high-resolution electrospray tandem mass spectrometer) was used in conjunction with Dionex UltiMate3000 RSLCnano System. A 2 μ l sample was aspirated into a 50 μ l loop and loaded onto the trap column (Thermo μ -Precolumn 5 mm, with nanoViper tubing 30 μ m i.d. $\times$ 10 cm). The flow rate was set to 300 nl/min for separation on the analytical column (Acclaim pepmap RSLC 75 μ Mx 15 cm nanoviper). Mobile phase A was composed of 99.9% H₂O (EMD Omni Solvent) and 0.1% formic acid, and mobile phase B was composed of 99.9% ACN and 0.1% formic acid. A 60 minute linear gradient from 3% to 45% B was performed. The LC eluent was directly nanosprayed into Q Exactive HF mass spectrometer (Thermo Scientific). During the chromatographic separation, the Q Exactive HF was operated in a data-dependent mode and under direct control of the Thermo Excalibur 3.1.66 (Thermo Scientific). MS data were acquired

using a data-dependent top-20 method for the Q Exactive HF, dynamically choosing the most abundant not-yet-sequenced precursor ions from the survey scans (350–1700). Sequencing was performed via higher energy collisional dissociation fragmentation with a target value of 10^5 ions determined with predictive automatic gain control. Full scans (350–1700 m/z) were performed at 60,000 resolution in profile mode. MS2 were acquired in centroid mode at 15,000 resolution. Ions with single charge or charges more than seven as well as unassigned charge were excluded. A 15-second dynamic exclusion window was used. All measurements were performed at room temperature. To facilitate label-free quantification, all measurements were done with three technical replicates to account for machine-related variability. Resultant raw files were searched with Proteome Discoverer 1.4 using SequestHT as the search engine with custom-generated FASTA databases and percolator as peptide validator. SequestHT search parameters used were as follows: enzyme name = Trypsin, maximum missed cleavage = 2, minimum peptide length = 6, maximum peptide length = 144, maximum delta Cn = 0.05, precursor mass tolerance = 10 ppm, fragment mass tolerance = 0.2 Da, dynamic modifications, carbamidomethyl +57.021 Da(C) and oxidation +15.995 Da(M). Protein and peptide identities were validated using Scaffold (version 4.3.4, Proteome Software Inc., Portland, OR, USA) software. Peptide identities were accepted based on a 1.0% false discovery rate (FDR) using the Scaffold Local FDR algorithm. Protein identities were also accepted with an FDR of 1.0% and a minimum of one recognized peptide.

To estimate proteomic abundances, we first calculated separate conversion factors for each of three replicates for the venom of each individual (C0195 and C0196) on the basis of the known concentrations of the control proteins and their observed Quantitative Values (i.e., normalized spectral counts) as calculated by Scaffold. We excluded the lowest-abundance control from the calculation of conversion factors because it was not detected in all replicates. To calculate the conversion factors, we found the slope of the best fit line relating the known concentrations of the controls to the observed normalized spectral counts, requiring an intercept at the origin. These conversion factors were then used to convert the normalized spectral counts for each protein in each replicate to concentrations. Final concentrations for each sample were averaged across the corresponding three replicates.

Transcriptome assembly and analysis

On the basis of our target insert size of approximately 180 nucleotides and our paired-end read lengths of 150 nucleotides, we expected most read pairs to show significant 3' overlap. We therefore merged reads using PEAR version 0.9.6 (Zhang et al., 2014) and used only merged reads for our subsequent analyses. Our primary transcriptome assembly was generated with DNASTar NGen version 12.3.1. We used 10 million merged reads and the default transcriptome assembly settings, except that we set the minimum number of sequences to retain a contig to 200. Because we did not necessarily expect known homologs of the toxins for this species to be present in public databases, we implemented

four search strategies for identifying and annotating proteins in the transcriptome. The first two strategies used the mass-spectrometry results from the venom (described above) and protein databases generated by applying TransDecoder version 2.0.1 to our assembled transcriptomes. We first used the protein sequences predicted by TransDecoder with a minimum protein length of 50 as the database against which to search our mass-spectrometry results. Results were filtered in Scaffold Viewer version 4.6.0 with protein and peptide false-discovery rates set to 1.0% and the minimum number of peptides set to one to accommodate any potential short peptides in the venom. The presence of signal peptides was verified with SignalP version 4.1 (Petersen et al., 2011). The second strategy used all possible protein or peptide sequences of at least 50 amino acids from all six reading frames generated by TransDecoder, not just the most likely. We also used less stringent filtering in Scaffold by setting the protein threshold to 99.9%, the peptide threshold to 95.0%, and the minimum number of peptides to one. We included this second analysis to ensure that small peptides were not being missed in the prediction step of the TransDecoder analyses. For this analysis, contigs already annotated based on the TransDecoder predictions were excluded from consideration. To identify proteins from the transcriptome with homology to known toxins, we conducted a blastx (version 2.2.30+) search of our transcripts generated by NGen against the UniProt animal toxins database (downloaded on November 16, 2015). We attempted to annotate full-length putative toxins that showed a match against at least 80% of the length of a known toxin. To generate a general database of proteins expressed in venom glands, including toxins and nontoxins, we conducted a blastx analysis of the transcripts generated by NGen against the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein database (downloaded on November 13, 2015). To ensure we only annotated highly expressed, high-quality proteins, we only considered transcripts comprised of at least 1,000 reads and showed a match against at least 95% of the length of a known protein. To help ensure that we did not miss any high-abundance transcripts in our NGen assembly, we also assembled the transcriptome using Extender (Rokyta et al., 2012). We assembled from 1,000 random reads. We only used reads with phred qualities of ≥ 30 at all positions and required an exact match of 120 nucleotides for extension. The resulting contigs were then searched against the UniProt animal toxins database with blastx. For each of the two individuals sequenced, we generated a consensus transcriptome by combining the results from all five analyses described above. We used cd-hit-est version 4.6 (Li and Godzik, 2006) on the coding sequences to remove any exact duplicates. We then screened for bad or chimeric sequences by aligning the merged reads with bowtie2 version 2.2.7 (Langmead and Salzberg, 2012) and screening for coverage anomalies. To generate the final representative transcriptome for the species, we merged the results of the two individuals. We clustered transcripts by analyzing the coding sequences with cd-hit-est, using a global sequence identity of 0.98. We estimated transcript abundances using RSEM (Li et al., 2011) version 1.2.28 based on bowtie (Langmead et al., 2009) version 1.1.2 alignments. Alignments were based on all merged reads for each individual. As discussed in detail elsewhere (Rokyta et al., 2015a), we used the centered logratio

transform (Aitchison, 1986) on all of our transcriptome and proteome abundances. This transform does not affect rank-based analyses and is equivalent to a log transform for linear analyses.

Because our two focal RNA-seq libraries were sequenced in a HiSeq lane with RNA-seq libraries from other species, we conducted a final quality-control step on our assembled transcripts to ensure that none of these transcripts resulted from assembly of reads that were misassigned during demultiplexing. The reads from each library were merged with PEAR version 0.9.6 (Zhang et al., 2014), and we aligned all of the merged reads resulting from each library sequenced on the same lane against the coding sequences of the *H. spadix* transcripts with bowtie2 version 2.2.7 (Langmead and Salzberg, 2012). Transcripts with coding sequences with full-length coverage, no fixed sequence differences, and $>100\times$ higher coverage for another library relative to one of the two *H. spadix* libraries was considered a contaminant and removed from the transcript set.

Data availability

The raw transcriptome reads were submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject PRJNA340270, BioSamples SAMN05711363 (C0195) and SAMN05711364 (C0196), and SRA accessions SRR4069277 (C0195) and SRR4069278 (C0196). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaíno et al., 2016) partner repository with the dataset identifier PXD005009 and 10.6019/PXD005009. The assembled transcripts were submitted to the NCBI Transcriptome Shotgun Assembly database. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GFAH00000000. The version described in this paper is the first version, GFAH01000000.

Results and Discussion

Venom-gland transcriptomes

For individual C0195 (female), we generated 12,205,269 read pairs passing the Illumina quality filter with 150 paired-end sequencing. Of these raw reads, 10,622,836 (87.0%) pairs were merged on the basis of their 3' overlaps, and ten million of these merged reads were assembled with NGen into 3,679 contigs. We annotated 135 coding sequences based on blastx hits to the UniProt toxins database, 331 coding sequences on the basis of blastx hits to the NCBI nr database, 65 coding sequences based on our MS-directed analysis using the TransDecoder predictions, and an additional 24 coding sequences on the basis of our MS-directed analysis using all ORFs. The extender assembly produced 45 annotated coding sequences after a blastx search against the UniProt toxins database. After combining all of these annotated sequences, screening for chimeras, and eliminating duplicates, we identified 420 coding sequences.

For individual C0196 (male), we generated 24,243,211 read pairs passing the Illumina quality filter. Of these raw reads, 19,857,433 (81.9%) pairs were merged on the basis of their 3' overlaps. NGen assembly using 10 million merged reads produced 5,568 contigs. We annotated 133 coding sequences based on blastx hits to the UniProt toxins database, 531 coding sequences on the basis of blastx hits to the NCBI nr database, 43 coding sequences based on our MS-directed analysis using the TransDecoder predictions, and an additional 34 coding sequences on the basis of our MS-directed analysis using all ORFs. The extender assembly produced 39 annotated coding sequences after a blastx search against the UniProt toxins database. After combining all of these annotated sequences, screening for chimeras, and eliminating duplicates, we identified 638 coding sequences.

We combined the annotated transcripts from individuals C0195 (420 total) and C0196 (638 total) and eliminated duplicates to produce a final combined transcriptome for *H. spadix* consisting of 744 transcripts with unique and complete protein-coding sequences. This set of sequences was used in all subsequent analyses of transcript abundances and LC-MS/MS results for both individuals. We divided these transcripts into three classes. The proteomically confirmed toxins (79 transcripts) encoded proteins that were detected proteomically in the venom of one or both individuals. We consider these transcripts to have a high likelihood of encoding toxic components of the *H. spadix* venom. Some of these, particularly those detected at low levels, may represent nontoxic proteins leaked into the venom, perhaps resulting from cell rupture in the gland tissue during venom extraction. The second class was the homology-based toxins (69 sequences). These transcripts encode proteins with homology to known animal toxins that were not detected proteomically in the venoms of either individual. Many of these encode small peptides that were expressed at high levels in the transcriptomes (Table 1) and were homologous to well-established scorpion toxins or other peptides in the proteomically detected classes (e.g., K⁺-channel toxins or antimicrobial peptides). The small sizes of these peptides and post-translational processing or modifications may have rendered them proteomically undetectable through our approach. In addition, tryptic peptides differ in their propensities for MS detection, and detectability is influenced by the sequences of the peptides and the regions flanking the peptides in their parent proteins (Kuster et al., 2005; Tang et al., 2006; Fusaro et al., 2009). Proteins were excluded from the homology-based toxins if they were large (>200 amino-acid residue precursors) and had no toxin matches in the UniProt toxins database from class Arachnida. We considered this class of proteins to have medium-to-high likelihoods of encoding toxic proteins or peptides. The third class, the nontoxins, consisted of the 596 transcripts not assigned to the two toxin classes and therefore were considered to have low likelihoods of encoding toxic proteins or peptides. These transcripts most likely encoded proteins important for basic cell function and the production of toxins. Of the mapped reads, the proteomically confirmed toxins accounted for 475,120.36 and 479,448.72 transcripts per million (TPM), the homology-based toxins accounted for 382,185.64 and 323,155.28 TPM, and the nontoxins accounted for 142,694.06 and 197,396.10 TPM for C0195 and

C0196, respectively.

Antimicrobial peptides

We identified 15 transcripts encoding proteins homologous to scorpion antimicrobial peptides (AMPs) in the venom-gland transcriptome of *H. spadix*. These peptides can be divided into those with and without disulfide bridges. Cysteine-containing AMPs usually have 3 or 4 disulfide bridges and interact with Na^+ , K^+ , Ca^{2+} , and Cl^- channels (Harrison et al., 2014). The first member of this protein group described from scorpion venoms was Scorpine from *Pandinus imperator* and showed activity against Gram-positive and Gram-negative bacteria (Conde et al., 2000). We identified four members of this group for *H. spadix*: AMP-1, AMP-4, AMP-7, and AMP-9. All four have the six conserved cysteine residues characteristic of Scorpine and a 19 amino-acid signal peptide. Three of the four were detected proteomically, but we failed to detect AMP-7, the least abundant of the four, in the venom proteome of either individual.

Nondisulfide bridge peptides show bradykinin-potentiating, antimicrobial, hemolytic, and immune-modulating activities (Zeng et al., 2005) and have mature peptides with 13–56 amino-acid residues (Almaaytah and Albalas, 2014). Some scorpion species spray venom onto their bodies, presumably to clean them (Torres-Larios et al., 2000), which might, in part, explain the presence of such components in their venoms. Zeng et al. (2005) defined the classification system for nondisulfide bridged proteins (NDBPs) on the basis of pharmacological activity, peptide length, and structural similarity and defined six subfamilies of NDBPs. We identified members of four of these six subfamilies within the venom-gland transcriptome of *H. spadix*. The NDBP-5 subfamily was the most diverse with five members: AMP-11, AMP-12, AMP-13, AMP-14, and AMP-15. All five members had highly conserved, 23 amino-acid signal peptides and 35 amino-acid C-terminal propeptides (Figure 1). The mature peptides were either 12 or 13 amino-acid residues in length and were, in contrast to their signal peptides and propeptides, highly diverged. None of these peptides were detected in the venom proteome, most likely as a result of their small size and extreme post-translational processing. We identified three members of the NDBP-4 subfamily: AMP-5, AMP-6, and AMP-8. Of these, only AMP-6 was detected in the venom proteome. Subfamily NDBP-2 was represented by two transcripts: AMP-2 and AMP-3. Only AMP-2 was detected proteomically. We identified only one member of subfamily NDBP-3, AMP-19, which was not detected proteomically.

The AMPs were one of the most highly expressed and diverse toxin types in the venom of *H. spadix* (Table 1 and Figure 2). AMP-3 was the most abundant transcript for C0195 (101,140.16 TPM) and the second most abundant for C0196 (61,030.53 TPM), but was not detected proteomically in the C0195 venom. AMP-3 was, however, detected in the venom of C0196. Altogether, the AMPs were the most abundant toxin type in individual C0195, accounting for a total of 333,853.45 TPM (i.e., approximately one third of the assigned transcriptional output), and the third most abundant toxin type in C0196 (190,218.94 TPM, behind the αKTx and the generic venom protein classes).

Ion-channel toxins

Scorpion venoms are typically characterized by the presence of a high diversity and abundance of toxins with effects accomplished by means of interfering with the action of ion channels (Quintero-Hernández et al., 2013), and the venom of *H. spadix* was consistent with this pattern (Figure 2). The vast majority of the identified putative ion-channel toxins were K⁺-channel toxins (KTxs), and we also identified toxins with homology to known Ca²⁺-channel toxins (CaTxs). We found no transcripts encoding proteins with clear homology to Na⁺-channel toxins. We identified 31 α KTxs, one β KTx, and one CaTx in the venom-gland transcriptome of *H. spadix* (Table 1).

The α KTx family is comprised of peptide toxins that interact with and block the action of K⁺ channels. The α KTx family of K⁺-channel toxins is the largest of four defined families of toxins affecting K⁺ channels from scorpions (Tytgat et al., 1999; de la Vega and Possani, 2004; Quintero-Hernández et al., 2013). Their mature peptides are 23–42 residues in length, with three or four disulfide bridges (Quintero-Hernández et al., 2013). These toxins were the most diverse group of toxins for *H. spadix* and the second most abundant class of putative toxin transcripts for both individuals with 247,583.81 TPM for C0195 and 204,542.79 TPM for C0196 (Table 1 and Figure 2). Tytgat et al. (1999) identified 12 subfamilies of α KTx on the basis of 49 peptides, and de la Vega and Possani (2004) updated the number of subfamilies to 18 on the basis of 120 peptides. Our sequences were generally too divergent from known sequences to classify within previously described classification systems. Two (α KTx-3 and α KTx-18) had six cysteine residues, and the remaining 29 had eight cysteine residues. Of the 31 unique α KTxs, we were only able to proteomically detect nine. Among those undetected was α KTx-29, which was the most highly expressed transcript in this toxin class for both individuals. These putative toxins have even smaller precursors than the AMPs, a class which also showed detectability challenges.

The β KTx family of KTxs are long-chain toxins (approximately 50–75 amino-acid residues in their mature peptides), and we identified one putative member of this family in the venom-gland transcriptome of *H. spadix* (Table 1). This putative toxin, β KTx-1, was also detected proteomically and only accounted for a small amount of the venom-gland expression (3,514.12 for C0195 and 4,553.94 TPM for C0196).

We identified a single putative Ca²⁺-channel toxin, CaTx-1, in the venom-gland transcriptome of *H. spadix* (Table 1). This toxin was expressed at a high level in C0195 (13,222.11 TPM) but at a much lower level in C0196 (285.73 TPM). We did not detect this putative venom component proteomically in the venom of either individual, but its predicted protein sequence differed at only two amino-acid sites relative to the well-characterized toxin Hadrucalcin from *Hadrurus gertschi* (Schwartz et al., 2009), suggesting that this transcript does encode a legitimate toxin. Hadrucalcin, a member of the calcin protein family, activates ryanodine receptors (i.e., Ca²⁺ channels of the sarcoplasmic reticulum), thereby inducing the release of Ca²⁺ (Schwartz et al., 2009).

Proteases and peptidases

We identified one aminopeptidase (Aminopeptidase-1) in both the venom-gland transcriptome and venom proteome. This protein had a signal peptide and an aminopeptidase N (PepN) domain. Although we found no close scorpion-toxin match in the nr database, this protein is clearly a component of the venom of *H. spadix* (Table 1). This transcript was approximately two times more abundant for individual C0195 than C0196 in their transcriptomes (1,228.46 versus 533.76 TPM, respectively). No blast matches were found for this protein in the UniProt toxins database, indicating that this transcript represents a new class of animal venom component. We detected five different serine proteases (SPs) transcriptomically, and the four most transcriptomically abundant of these were detected proteomically (Table 1). All five SPs had trypsin-like serine protease domains, signal peptides, and database matches to putative scorpion toxins. This toxin class accounted for 8,553.26 and 14,599.66 TPM for C0195 and C0196, respectively. We identified one cysteine peptidase (CP-1) in both the venom transcriptome and proteome (Table 1). This protein had three conserved domains: a cystatin-like domain, a cathepsin propeptide inhibitor domain, and papain-like cysteine-peptidase domain. A similar protein was detected in the venom-gland transcriptome of the buthid scorpion *Tityus bahiensis* (de Oliveira et al., 2015). This cysteine peptidase was at a moderate transcriptional abundance in both individuals, but approximately two times more abundant in C0196 than C0195 (1,340.18 and 531.77 TPM, respectively). We identified one M13 peptidase (PeptidaseM13-1) both proteomically and transcriptomically. We did not detect a signal peptide for this protein with SignalP, and it was at relatively low abundance in both transcriptomes (208.11 and 375.63 TPM for C0195 and C0196, respectively). The encoded protein had a single Peptide family M13 conserved domain. We identified four M2 peptidases (Table 1). All four were detected proteomically and had a signal peptide and a Peptidase family M2 Angiotensin converting enzyme (ACE) domain. This moderately abundant class accounted for 5,668.22 and 4,809.76 TPM for C0195 and C0196, respectively. Enzymes in this family are typically membrane-bound zinc-dependent peptidases that convert angiotensin I to angiotensin II, a potent vasopressor. Overall, peptidase and proteases accounted for 16,189.93 and 21,659.35 TPM for C0195 and C0196, respectively (Figure 2). The abundance of proteases may be partially related to the substantial post-translational processing many of the known scorpion toxins must undergo prior achieving a functional state (e.g., Figure 1).

Putative toxins with functionally characterized homologs

Three different phospholipases A₂ (PLA2s) were detected transcriptomically, and all three of these were detected proteomically (Table 1 and Figure 2). All three had signal peptides and PLA2-like domains. These PLA2s (PLA2-1, PLA2-2, and PLA2-3) were all more than 25% divergent from one another at the amino-acid level and were related to Phaiodactylipin (Valdez-Cruz et al., 2004), a heterodimeric PLA2 from the scorpion *Anuroctonus phaiodactylus* that is encoded as a single protein precursor. Maturation

of phaiodactylipin requires removal of a hexapeptide linker and formation of disulfide bridges to between the two resulting subunits. Phaiodactylipin is lethal to arthropods and causes muscular inflammation in mice (Valdez-Cruz et al., 2004).

The remaining putative toxins in this group with functionally characterized homologs accounted for only a minute portion of the transcriptome and proteome (Figure 2). We identified two Kunitz-type protease inhibitors (KUNs) that differed by 4.6% at the amino-acid level in the *H. spadix* venom-gland transcriptome. Both had signal peptides, but neither was detected proteomically. Both had a Kunitz/Bovine pancreatic trypsin inhibitor domain. The closest database match in nr was Hg1 (Table 3) from the *H. gertschi* venom-gland transcriptome (Schwartz et al., 2007), which was also not proteomically confirmed to be in the venom of *H. gertschi*. The expression levels of these were low (<550 TPM), which could explain the failure to detect them proteomically and could also indicate that these proteins serve a nontoxic role. We identified one transcript encoding a putative chitinase (Chitinase-1) that was proteomically confirmed in the venom. This protein had a signal peptide and a Chitin binding Peritrophin-A (CBM_14) domain. We identified two cysteine-rich secretory proteins (CRISP-1 and CRISP-2) in both the transcriptome and proteome. Both had signal peptides and cysteine-rich secretory protein family (CAP) domains. Despite a shared domain structure, these two genes showed no obvious homology to one another in alignments. We identified two cystatins (Cystatin-1 and Cystatin-2) in both the transcriptome and proteome. Both had signal peptides and Cystatin-like domains, which are associated with cysteine-protease inhibitors. These two proteins were 14.2% divergent at the amino-acid level. In addition, we identified a hyaluronidase (HYAL-1), a histidine phosphatase (HistP), and a nucleotidase (NUC-1) in both the venom-gland transcriptome and venom proteome. Finally, six low-abundance proteins were proteomically detected only in C0195: a C-reactive protein (CReactive), a glyceraldehyde 3-phosphate dehydrogenase (G3PDH2), a low density lipoprotein receptor (LDLR-1), superoxide dismutase (SupDismutase), a transferrin, and a transketolase. All six of these have their closest database matches to nontoxic homologs and probably represent proteins incidentally released into the venom.

Putative toxins without functionally characterized homologs

We identified three different La1-like proteins (La1-1, La1-2, and La1-3) in the venom-gland transcriptome (Table 1). The two most transcriptomically abundant of these (La1-1 and La1-2) were also detected proteomically. All three had signal peptides and a single von Willebrand factor type C (SVWC) domain. This class of proteins was one of the more abundant classes (Figure 2), despite having only three members, representing 40,704.15 and 120,079.00 TPM for C0195 and C0196 respectively. Expression levels of all three La1-like proteins were higher in the male (C0196) than the female (C0195). Related proteins have been identified in the venom-gland transcriptomes of other scorpion species, but their functions are unknown (Luna-Ramírez et al., 2013).

We identified 66 proteins and peptides that we were unable to classify and therefore

521 labeled generically as “venom proteins” (VP-1–66). Nearly half of our putative toxin-
 522 encoding transcripts (66 of 148) could not be classified, including 35 of 79 proteomi-
 523 cally confirmed toxins (Table 1). These unclassified toxins accounted for 176,385.26 and
 524 225,673.33 TPM for C0195 and C0196, respectively. All but four of these putative toxins
 525 had signal peptides detectable by SignalP, and their precursor lengths ranged from 56
 526 to 353 amino-acid residues. Among these 66 putative toxins, we identified seven groups
 527 with at least two homologs involving a total of 26 of 66 sequences (Table 1). The re-
 528 maining 40 putative VP toxins had no homology with functionally characterized toxins
 529 in other species or with other sequences in the *H. spadix* venom-gland transcriptome.
 530 Group I included VP-1 and VP-6, which differed at 0.9% of amino-acid sites as well
 531 as a 12 amino-acid insertion in VP-6 relative to VP-1. They had identical 20 amino-
 532 acid predicted signal peptides and no cysteines in the predicted mature peptide. As
 533 a group, they accounted for 2,313.73 and 3,995.14 TPM for C0195 and C0196 respec-
 534 tively, and only VP-6 was detected proteomically. Group II included VP-15, VP-32,
 535 and VP-49. These three putative toxins were highly diverged from one another, with
 536 at least 39.5% amino-acid divergence. All three had detectable signal peptides and 10
 537 conserved cysteine residues outside their signal peptides. Expression levels were relative
 538 low with 2,882.05 and 640.79 TPM for C0195 and C0196, respectively. None of these
 539 putative toxins were detected proteomically. Group III included VP-17, VP-18, VP-22,
 540 VP-23, VP-27, VP-57, and VP-61 (Table 1). All seven had a 23 amino-acid signal pep-
 541 tide, 10 conserved cysteine residues, and amino-acid divergences ranging from 6.4% to
 542 89.4%. All but VP-18 were detected proteomically, and altogether this group accounted
 543 for 16,419.30 and 41,126.10 TPM for C0195 and C0196, respectively. Group IV was the
 544 most abundant and diverse group of the VPs, including VP-21, VP-29, VP-31, VP-34,
 545 VP-35, VP-36, VP-44, and VP-45. All eight of these were detected proteomically, had
 546 18–19 amino-acid signal peptides, and shared eight conserved cysteine-residues. Pairwise
 547 amino-acid divergences ranged from 4.4% to 82.2%. This group of VPs accounted for
 548 48,800.56 and 54,941.61 TPM for C0195 and C0196, respectively. Group V included two
 549 members, VP-37 and VP-39, that showed 20.0% amino-acid divergence. Both putative
 550 toxins had signal peptides and 10 conserved cysteine residues, and neither was detected
 551 proteomically in the venom of either individual. This group was at a low abundance in
 552 the transcriptome, accounting for only 485.95 and 430.88 TPM in C0195 and C0196, re-
 553 spectively. Group VI included two members, VP-46 and VP-66, with 45.6% amino-acid
 554 divergence between them. Both had signal peptides and 12 conserved cysteine residues,
 555 and neither was detected proteomically. This group was also at a low abundance in the
 556 transcriptome with 347.63 and 75.4 TPM for C0195 and C0196, respectively. Group
 557 VII also included only two members, VP-48 and VP-53, that showed 62.6% amino-acid
 558 divergence. Both members of this group had 21 amino-acid signal peptides and 12 con-
 559 served cysteine residues, and neither was detected proteomically. This group was low
 560 abundance in the transcriptomes, accounting for 294.35 and 80.67 TPM for C0195 and
 561 C0196, respectively. The remaining 40 ungrouped VPs accounted for the majority of VP
 562 expression with 104,642.26 and 124,422.72 TPM for C0195 and C0196, respectively.

Comparison to known *Hadrurus* toxins

Few components of the venoms of scorpions in the genus *Hadrurus* have been previously characterized. Schwartz et al. (2006) characterized a K⁺-channel toxin (α KTx) from *Hadrurus gertschi* which they named HgeTx1 (UniProtKB/Swiss-Prot: P84864.1). Torres-Larios et al. (2000) characterized the antimicrobial peptide Hadrurin from the venom of *Hadrurus aztecus*. Only amino-acid sequences are available for these two toxins. The sequenced fragment of HgeTx1 was 36 amino-acid residues in length, and the closest match to it in the *H. spadix* transcriptome was α KTx-31, which differed from HgeTx1 at five amino-acid positions. This high percentage (13.9%) of amino-acid differences between proteins from closely related species could be indicative of positive selection, but α KTx-31 was not detected in the venom proteomes of either *H. spadix* individual, suggesting that relaxed selection in *H. spadix* is an equally valid possibility. The closest match in the *H. spadix* venom-gland transcriptome to the 41 amino-acid sequence of Hadrurin was AMP-3, which was detected in the venom proteome. The amino-acid sequence of AMP-3 differed from Hadrurin at four of the 41 residues (9.8%), again suggesting rapid sequence evolution. The α KTx and AMP toxin families were large and diverse in *H. spadix* and are presumably similarly diverse in *H. gertschi* and *H. aztecus*. We therefore have little reason to expect that these two comparisons actually involve orthologous sequences, and comparisons between different paralogs may have artificially inflated the amount of sequence divergence.

Schwartz et al. (2009) characterized Hadrucalcin (UniProtKB/Swiss-Prot: B8QG00.1; GenBank: EU496812), a toxin affecting Ca²⁺ channels, from the venom of *H. gertschi*. The closest match to Hadrucalcin in the *H. spadix* venom-gland transcriptome was CaTx-1. Both sequences were 225 nucleotides in length and differed at seven nucleotide positions (3.1%). The predicted precursor proteins differed at only two positions (2.7%). We estimated a pairwise dN/dS ratio of 0.16 using codeml version 4.6 (Yang, 1997, 2007), which is consistent with the action of purifying selection.

Schwartz et al. (2007) sequenced 160 ESTs from the venom-gland of a single individual of *H. gertschi*, which they clustered into 68 unique sequences. Because they did not conduct a proteomic analysis of the venom, they relied on homology to known toxins to identify toxin-encoding transcripts. They identified 10 transcripts with homology to known toxins. We clustered the coding sequences from *H. spadix* with the 10 putative toxins identified for *H. gertschi* by Schwartz et al. (2007), using cd-hit-est (Li and Godzik, 2006) with a global minimum match percentage of 80. We identified matches for seven of the ten *H. gertschi* sequences (Table 3). The best matches for four of the *H. gertschi* sequences were detected only in the venom-gland transcriptome and not in the venom proteome of *H. spadix* (Table 3). The remaining three had matches in the transcriptome that were also detected proteomically. The average nucleotide divergence between the *H. gertschi* and *H. spadix* coding sequences was 7.1%. Sunagar and Moran (2015) proposed that ancient venomous lineages such as scorpions should tend to show toxins evolving under negative or purifying selection whereas newer venomous lineages such as snakes should show more evidence of positive selection. We estimated pairwise dN/dS ratios for

the seven pairs of sequences using codeml (Yang, 1997, 2007) and found values ranging from 0.17 through 0.64 (Table 3), consistent with this hypothesis. Our sample size of proteomically confirmed pairs, however, was small, and estimates of dN/dS averaged over entire sequences have low power to detect positive selection compared to approaches that allow site-classes to have different dN/dS ratios (e.g., Yang et al., 2000). Unfortunately, the latter methods require more than two sequences to be applied. In addition, the components of peptide rich venoms such as those of scorpions often undergo extensive proteolytic processing prior to achieving their active states. As Figure 1 demonstrates, the active, functional regions of precursors can be a minority of the total amino-acid sites, and we would only expect strong positive selection on the regions that contribute to the active peptide. Overall, the precursors in Figure 1 showed evidence for purifying selection because the signal peptides and propeptides were highly conserved. The small region of the mature peptide, however, diverged substantially among paralogs. Every site in the mature toxins in Figure 1 showed nonsynonymous variation, but these sites represent less than 20% of the total sites in the precursor. Attempts at testing the hypothesis of Sunagar and Moran (2015) resulting in a failure to reject must rule out the effects of differential patterns of selection among sites due to such proteolytic processing to be conclusive. Sunagar and Moran (2015) argued that a higher proportion of sites showed positive selection in young lineages compared to older lineages, but this could simply reflect more extensive post-translational processing of some toxins in older lineages.

Transcript and protein abundances were similar across individuals

The mRNA expression levels of nontoxin-encoding proteins for the two individuals showed a high correlation (Spearman’s rank correlation $\rho = 0.88$, Pearson’s correlation coefficient $R = 0.87$, and $R^2 = 0.75$), indicating that our venom-gland transcriptomes were comparable (Figure 3). Expression levels were less consistent for the proteomically confirmed toxins (Spearman’s rank correlation $\rho = 0.66$, Pearson’s correlation coefficient $R = 0.49$, and $R^2 = 0.24$) and the homology-only toxins (Spearman’s rank correlation $\rho = 0.59$, Pearson’s correlation coefficient $R = 0.61$, and $R^2 = 0.37$). Whether this divergence in expression levels represents individual-level variation, sex-related differences (Miller et al., 2016), or noise in the data was not clear. The nontoxin results suggest that the observed differences for the two putative toxin classes were biological rather than technical. Most of the divergence was attributable to a handful of outlier transcripts that were highly expressed in one of the transcriptomes and lowly expressed in the other (Figure 3). Outliers include members of the major classes of venom components, including AMPs, α KTxs, and VPs.

The LC-MS/MS show patterns similar to those of the transcriptional levels. Considering only proteins detected in the venoms of both individuals (52 proteins), the protein abundances showed good agreement (Spearman’s rank correlation $\rho = 0.70$, Pearson’s correlation coefficient $R = 0.72$, and $R^2 = 0.52$), suggesting that our proteomic abundance measurements are reliable and consistent (Figure 4). The major proteomic outlier

was PeptidaseM2-4 (Figure 4), which was abundant in individual C0195, but barely detected in C0196. The venom proteome of C0195, however, had 22 proteins not detected for C0196, and the venom proteome of C0196 had five proteins not detected for C0195 (Table 2). Many of these presence/absence disagreements involved proteins detected only at very low levels in the proteome of one or the other individual. The most abundant protein in the venom of individual C0195 that was not detected in the venom of C0196 was LDLR-1, which was the 22nd most abundant protein of 74 in the C0195 venom proteome (Tables 2 and 1). This protein has a LDLa domain (low-density lipoprotein receptor domain class A) and is homologous to Prohormone 4 precursors from other invertebrate species. The most abundant protein in the venom of C0196 that was not detected in C0195 was VP-31, which was the 13th most abundant protein of 57 detected for C0196. This protein has a signal peptide and a precursor length of 117 amino-acids. VP-31 has no hits in the NCBI nonredundant (nr) database. Of the 27 proteins showing presence/absence differences (Table 2), nine of these had transcripts that were detected as outliers in the transcript-abundance analysis (Figure 3). We identified a total of 11 mRNA outliers, indicating significant expression differences between C0195 and C0196, and nine of these were manifested as presence/absence differences in the venom proteomes. The agreement across data types suggests that these differences are biological rather than technical. Many of the proteome presence/absence disagreements certainly resulted from detection thresholds, but those involving more abundant proteins such as LDLR-1 and VP-31 suggest significant toxin expression differences between C0195 and C0196. These differences could be sex related or simply intraspecific expression variation. Most of the expression-level (Figure 3), protein-level (Figure 4), and presence/absence (Table 2) differences were in the same direction between individuals; many of the venom components for the female (C0195) were downregulated in the male (C0196).

Transcript and protein abundance estimates differ

Rokyta et al. (2015a) showed that venom-component transcript and protein abundances are highly correlated for venomous snakes. They showed that for 11 transcriptome/proteome comparisons from snakes representing 10 species, six genera, and three families, estimates of abundances had Pearson’s correlation coefficients of $0.58 < R < 0.92$, Spearman’s rank correlation coefficients of $0.47 < \rho < 0.89$, and coefficients of determination of $0.34 < R^2 < 0.85$. Most comparisons were closer to the upper ends of these ranges. Our comparisons of transcriptomic (Figure 3) and proteomic (Figure 4) abundances across individuals showed high positive correlations, suggesting that our measurements were reliable and repeatable. We found a weaker relationship when comparing transcript and protein abundance estimates for each individual of *H. spadix* (Figure 5). For C0195, we found $R = 0.40$, $\rho = 0.40$, and $R^2 = 0.16$. For C0196, we found $R = 0.42$, $\rho = 0.49$, and $R^2 = 0.18$. These values were all far below any of the values measured by Rokyta et al. (2015a) for snakes except that ρ for C0196 was slightly higher than the lowest value detected by Rokyta et al. (2015a). These discrepancies were

even more apparent when considering class-level abundance estimates (Figure 2).

Casewell et al. (2014) claimed to identify major discrepancies between venom-gland transcriptomes and venom proteomes for snakes and attributed these to the interposition of post-transcriptional regulatory mechanisms (e.g., microRNAs) in the expression of venoms (although Rokyta et al. (2015a) raised concerns about their claims). As Rokyta et al. (2015a) pointed out, a failure to find agreement between different sets of measurements must be assumed to have originated in methodological issues or biases until those sources of disagreement can be absolutely eliminated as potential causes. For our results, class-level estimates clearly showed (Figure 2) that the major discrepancies involved the transcriptomically abundant small toxins such as the α KTxs and AMPs. We failed to proteomically detect many of these highly expressed short proteins (Table 1), suggesting a bias with our proteomic measurements.

To determine whether the transcriptome/proteome discrepancy resulted from biases related to the sizes of the proteins, we compared the difference between the protein and transcript abundance estimates to the length of their corresponding coding sequences (Figure 6). For both individuals, we found a significant positive correlation (C0195: $p = 2.0 \times 10^{-8}$ and C0196: $p = 7.9 \times 10^{-3}$). This positive correlation implies that protein abundance estimates were generally lower than expected based on transcriptomic estimates for shorter proteins and/or larger than expected for larger proteins (Figure 6). Although we found a significant trend in the relationship between protein size and divergence between transcript and protein abundance estimates, this relationship was far from universal, implying that a simple correction for length would not obviate all of the transcriptome/proteome discrepancies.

Post-translational proteolytic processing (e.g., Figure 1) could further exacerbate this issue by rendering peptide predictions used in the analysis of mass-spectrometry data inaccurate. This inaccuracy could be especially debilitating for short toxins for which we already expect few possible peptides. Tryptic peptides can also differ in their propensities for mass-spectrometry detection, and detectability can be influenced by the sequences of the peptides and the regions flanking the peptides in their parent proteins (Kuster et al., 2005; Tang et al., 2006; Fusaro et al., 2009). Undetectable peptides in a large protein would have less effect on protein detectability and abundance estimation than such peptides in smaller proteins. These types of issues may be somewhat scorpion-specific because of the prevalence of many small, highly processed peptides in their venoms; such issues are not likely to affect most snake venoms, particularly those of viperids, because these venoms are comprised of larger proteins. We therefore consider the transcriptomic estimates of abundances to be more accurate for the venom of *H. spadix*, with the whole-venom, shotgun proteomic analyses providing a confirmation of the presence of particular proteins and peptides in the venom.

Conclusions

We demonstrated data-collection and analysis protocols suitable to comprehensive, single-animal quantitative venom characterizations of invertebrate (or other small-animal) venoms, and our approach gave repeatable and consistent quantitative transcriptomic and proteomic estimates across individuals. We found the venom of the black-back scorpion (*H. spadix*) to be complex, with 79 proteomically confirmed venom components and an additional 69 identified on the basis of homology with known animal toxins. Transcriptomically, the venom of *H. spadix* showed abundant and diverse K^+ -channel toxins, antimicrobial peptides, and peptidases. A major portion of the putative toxins, however, could not be assigned a function on the basis of homology, indicating a wealth of unexplored toxin sequence space within the venom of just this single species. The whole-venom proteome of *H. spadix* proved to be challenging because of the diversity of small peptide toxins and, at least in some cases, unaccounted-for postranslational proteolytic processing. These putative toxins were often not detected proteomically despite clear homology to known toxins and high expression levels in the venom gland, and those that were detected tended to show lower abundances than predicted by their transcriptional levels. We identified several putative toxins with substantially different expression levels between our male and female individuals. The vast majority of these showed downregulation in the male relative to the female, implying that the venom of the female was more diverse and complex. With sample sizes of one for each sex, however, we could not yet attribute this to a sex-based venom difference.

Acknowledgments

Funding for this work was provided by the National Science Foundation (NSF DEB-1145978) and the Florida State University Council on Research and Creativity. We thank Rakesh Singh of the Florida State University College of Medicine Translational Science Laboratory for advice and assistance with proteomic analyses and Devin Rokyta for assistance in acquiring specimens.

References

- Aitchison, J., 1986. The statistical analysis of compositional data. Chapman and Hall, London.
- Almaaytah, A. and Q. Albalas, 2014. Scorpion venom peptides with no disulfide bridges. Peptides 51:35–45.
- Alvarenga, É. R., T. M. Mendes, B. F. Magalhães, F. F. Siqueira, A. E. Dantas, T. M. Barroca, C. C. Horta, and E. Kalapothakis, 2012. Transcriptome analysis of the *Tityus serrulatus* scorpion venom gland. Open Journal of Genetics 2:210–220.

- Anisimova, M., J. P. Bielawski, and Z. Yang, 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution* 18:1585–1592.
- Casewell, N. R., S. C. Wagstaff, W. Wüster, D. A. N. Cook, F. M. S. Bolton, S. I. King, D. Pla, J. J. Calvete, and R. A. Harrison, 2014. Medically important differences in snake venom composition are dictated by distinct postgenomic mechanisms. *Proceedings of the National Academy of Sciences of the USA* 111:9205–9210.
- Conde, R., F. Z. Zamudio, M. H. Rodríguez, and L. Possani, 2000. Scorpine, an anti-malaria and anti-bacterial agent purified from scorpion venom. *FEBS Letters* 471:165–168.
- Diego-García, E., F. Caliskan, and J. Tytgat, 2014. The Mediterranean scorpion *Mesobuthus gibbosus* (Scorpiones, Buthidae): transcriptome analysis and organization of the genome encoding chlorotoxin-like peptides. *BMC Genomics* 15:295.
- Diego-García, E., S. Peigneur, E. Clynen, T. Marien, L. Czech, L. Schoofs, and J. Tytgat, 2012. Molecular diversity of the telson and venom components from *Pandinus cavimanus* (Scorpionidae Latreille 1802): transcriptome, venomomics and function. *Proteomics* 12:313–328.
- Duda, T. F., Jr. and S. R. Palumbi, 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proceedings of the National Academy of Sciences of the USA* 96:6820–6823.
- Duda, T. F., Jr. and E. A. Remigio, 2008. Variation and evolution of toxin gene expression patterns of six closely related venomous marine snails. *Molecular Ecology* 17:3018–3032.
- Francke, O. F. and L. Prendini, 2008. Phylogeny and classification of the giant hairy scorpions, *Hadrurus* Thorell (Iuridae Thorell): a reappraisal. *Systematics and Biodiversity* 6:205–223.
- Fusaro, V. A., D. R. Mani, J. P. Mesirov, and S. A. Carr, 2009. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology* 27:190–198.
- Gibbs, H. L. and W. Rossiter, 2008. Rapid evolution by positive selection and gene gain and loss: PLA₂ venom genes in closely related *Sistrurus* rattlesnakes with divergent diets. *Journal of Molecular Evolution* 66:151–166.
- Harrison, P. L., M. A. Abdel-Rahman, K. Miller, and P. N. Strong, 2014. Antimicrobial peptides from scorpion venoms. *Toxicon* 88:115–137.

- He, Y., R. Zhao, Z. Di, Z. Li, X. Xu, W. Hong, Y. Wu, H. Zhao, W. Li, and Z. Cao, 2013. Molecular diversity of the Chaerilidae venom peptides reveals the dynamic evolution of scorpion venom components from Buthidae to non-Buthidae. *Journal of Proteomics* 89:1–14.
- Kuster, B., M. Schirle, P. Mallick, and R. Aebersold, 2005. Scoring proteomes with proteotypic peptide probes. *Nature Reviews Molecular Cell Biology* 6:577–583.
- Langmead, B. and S. L. Salzberg, 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359.
- Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology* 10:R25.
- Li, M., I. X. Wang, Y. Li, A. Bruzel, A. L. Richards, J. M. Toung, and V. G. Cheung, 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 333:53–58.
- Li, W. and A. Godzik, 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
- Luna-Ramírez, K., V. Quintero-Hernández, V. R. Juárez-González, and L. D. Possani, 2015. Whole transcriptome of the venom gland from *Urodacus yaschenkoi* scorpion. *PLOS One* 10:e0127883.
- Luna-Ramírez, K., V. Quintero-Hernández, L. Vargas-Jaimes, C. V. F. Batista, K. D. Winkel, and L. D. Possani, 2013. Characterization of the venom from the Australian scorpion *Urodacus yaschenkoi*: molecular mass analysis of components, cDNA sequences and peptides with antimicrobial activity. *Toxicon* 63:44–54.
- Lynch, V. J., 2007. Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A₂ genes. *BMC Evolutionary Biology* 7:2.
- Ma, Y., Y. He, R. Zhao, Y. Wu, W. Li, and Z. Cao, 2012. Extreme diversity of scorpion venom peptides and proteins revealed by transcriptomic analysis: implication for proteome evolution of scorpion venom arsenal. *Journal of Proteomics* 75:1563–1576.
- Ma, Y., R. Zhao, Y. He, S. Li, J. Liu, Y. Wu, Z. Cao, and W. Li, 2009. Transcriptome analysis of the venom gland of the scorpion *Scorpiops jendeki*: implication for the evolution of the scorpion venom arsenal. *BMC Genomics* 10:290.
- Ma, Y., Y. Zhao, R. Zhao, W. Zhang, Y. He, Y. Wu, Z. Cao, L. Guo, and W. Li, 2010. Molecular diversity of toxic components from the scorpion *Heterometrus petersii* venom revealed by proteomic and transcriptomic analysis. *Proteomics* 10:2471–2485.

- Margres, M. J., J. J. McGivern, , M. S. K. P. Wray, J. Facente, and D. R. Rokyta, 2015. Contrasting modes and tempos of venom expression evolution in two snake species. *Genetics* 199:265–176.
- Mille, B. G., S. Peigneur, and J. T. Elia Diego-García and, Reinhard Prede and, 2014. Partial transcriptomic profiling of toxins from the venom gland of the scorpion *Parabuthus stridulus*. *Toxicon* 83:75–83.
- Miller, D. W., A. D. Jones, J. S. Goldston, M. P. Rowe, and A. H. Rowe, 2016. Sex differences in defensive behavior and venom of the striped bark scorpion *Centruroides vitattus* (Scorpiones: Buthidae). *Integrative and Comparative Biology* 56:1022–1031.
- de Oliveira, U. C., D. M. Candido, V. A. C. Dorce, and I. de Loiola Meirelles Junqueira-de Azevedo, 2015. The transcriptome recipe for the venom cocktail of *Tityus bahiensis* scorpion. *Toxicon* 95:52–61.
- Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen, 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8:785–786.
- Quintero-Hernández, V., J. M. Jiménez-Vargas, G. B. Gurrola, H. H. Valdivia, and L. D. Possani, 2013. Scorpion venom components that affect ion-channels function. *Toxicon* 76:328–342.
- Quintero-Hernández, V., S. Ramírez-Carretero, M. T. Romero-Gutiérrez, L. L. Valdez-Velázquez, B. Becerril, L. D. Possani, and E. Ortiz, 2015. Transcriptome analysis of scorpion species belonging to the *Vaejovis* genus. *PLOS One* 10:e0117188.
- Rendón-Anaya, M., L. Delaye, L. D. Possani, and A. Herrera-Estrella, 2012. Global transcriptome analysis of the scorpion *Centruroides noxius*: new toxin families and evolutionary insights from an ancestral scorpion species. *PLOS One* 7:e43331.
- Rokyta, D. R., A. R. Lemmon, M. J. Margres, and K. Aronow, 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics* 13:312.
- Rokyta, D. R., M. J. Margres, and K. Calvin, 2015a. Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. *Genes|Genomes|Genetics* 5:2375–2382.
- Rokyta, D. R., K. P. Wray, A. R. Lemmon, E. M. Lemmon, and S. B. Caudle, 2011. A high-throughput venom-gland transcriptome for the eastern diamondback rattlesnake (*Crotalus adamanteus*) and evidence for pervasive positive selection across toxin classes. *Toxicon* 57:657–671.
- Rokyta, D. R., K. P. Wray, and M. J. Margres, 2013. The genesis of an exceptionally lethal venom in the timber rattlesnake (*Crotalus horridus*) revealed through comparative venom-gland transcriptomics. *BMC Genomics* 14:394.

- Rokyta, D. R., K. P. Wray, J. J. McGivern, and M. J. Margres, 2015b. The transcriptomic and proteomic basis for the evolution of a novel venom phenotype within the Timber Rattlesnake (*Crotalus horridus*). *Toxicon* 98:34–48.
- Ruiming, Z., M. Yibao, H. Yawen, D. Zhiyong, W. Yingliang, C. Zhijian, and L. Wenxin, 2010. Comparative venom gland transcriptome analysis of the scorpion *Lychas mucronatus* reveals intraspecific toxic gene diversity and new venomous components. *BMC Genomics* 11:452.
- Sanz, L., H. L. Gibbs, S. P. Mackessy, and J. J. Calvete, 2006. Venom proteomes of closely related *Sistrurus* rattlesnakes with divergent diets. *Journal of Proteome Research* 5:2098–2112.
- Schwartz, E. F., E. M. Capes, E. Diego-García, F. Z. Zamudio, O. Fuentes, L. D. Possani, and H. H. Valdivia, 2009. Characterization of hadrucalcin, a peptide from *Hadrurus gertschi* scorpion venom with pharmacological activity on ryanodine receptors. *British Journal of Pharmacology* 157:392–403.
- Schwartz, E. F., E. Diego-Garcia, R. C. R. de la Vega, and L. D. Possani, 2007. Transcriptome analysis of the venom gland of the Mexican scorpion *Hadrurus gertschi* (Arachnida: Scorpiones). *BMC Genomics* 8:119.
- Schwartz, E. F., C. A. Schwartz, F. Z. Z. Froylan Gómez-Laguna and, and L. D. Possani, 2006. HgeTx1, the first K⁺-channel specific toxin characterized from the venom of the scorpion *Hadrurus gertschi* Soleglad. *Toxicon* 48:1046–1053.
- Soleglad, M. E. and V. Fet, 2003. High-level systematics and phylogeny of the extant scorpions (Scorpiones: Orthosterni). *Euscorpius* 2003:1–56.
- Soleglad, M. E., V. Fet, and G. Lowe, 2011. Contributions to scorpion systematics. IV. Observations on the *Hadrurus* “*spadix*” subgroup with a description of a new species (Scorpiones: Caraboctonidae). *Euscorpius* 2011:1–36.
- Stahnke, H. L., 1945. Scorpions of the genus *Hadrurus* Thorell. *American Museum Novitates* 1298.
- Sunagar, K. and Y. Moran, 2015. The rise and fall of an evolutionary innovation: contrasting strategies of venom evolution in ancient and young animals. *PLOS Genetics* 11:e1005596.
- Tang, H., R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly, and P. Radivojac, 2006. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* 22:e481–e488.
- Torres-Larios, A., G. B. Gurrola, F. Z. Zamudio, and L. D. Possani, 2000. Hadrurin, a new antimicrobial peptide from the venom of the scorpion *Hadrurus aztecus*. *European Journal of Biochemistry* 267:5023–5031.

898 Tytgat, J., K. G. Chandy, M. L. Garcia, G. A. Gutman, M.-F. Martin-Eauclaire, J. J.
899 van der Walt, and L. D. Possani, 1999. A unified nomenclature for short-chain peptides
900 isolated from scorpion venoms: α -KTx molecular subfamilies. Trends in Pharmacological Sciences 20:444–447.
901

902 Valdez-Cruz, N. A., C. V. F. Batista, and L. D. Possani, 2004. Phaiodactylipin, a glyco-
903 sylated heterodimeric phospholipase a_2 from the venom of the scorpion *Anuroctonus*
904 *phaiodactylus*. European Journal of Biochemistry 271:1453–1464.

905 Valdez-Velázquez, L. L., V. Quintero-Hernández, M. T. Romero-Gutiérrez, F. I. V. Coronas,
906 and L. D. Possani, 2013. Mass fingerprinting of the venom and transcriptome of
907 venom gland of scorpion *Centruroides tecomanus*. PLOS One 8:e66486.

908 de la Vega, R. C. R. and L. D. Possani, 2004. Current views on scorpion toxins specific
909 for K⁺-channels. Toxicon 43:865–875.

910 Vizcaíno, J. A., A. Csordas, N. del Toro N, J. A. Dienes, J. Griss, I. Lavidas, G. Mayer,
911 Y. Perez-Riverol, F. Reisinger, T. Ternent, Q. W. Xu, R. Wang, and H. Hermjakob,
912 2016. 2016 update of the PRIDE database and related tools. Nucleic Acids Research
913 44:D447–D456.

914 Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum like-
915 lihood. Computer Applications in the Biosciences 13:555–556.

916 ———, 2007. PAML 4: a program package for phylogenetic analysis by maximum
917 likelihood. Molecular Biology and Evolution 24:1586–1591.

918 Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen, 2000. Codon-substitution
919 models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

920 Zeng, X.-C., G. Corzo, and R. Hahin, 2005. Scorpion venom peptides without disulfide
921 bridges. IUBMB Life 57:13–21.

922 Zhang, H., M. Du, J. Xie, X. Liu, J. Sun, W. Wang, X. Xin, L. D. Possani, K. Yea, and
923 R. A. Lerner, 2016. Autocrine-based selection of drugs that target ion channels from
924 combinatorial venom peptide libraries. Angewandte Chemie 128.

925 Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis, 2014. PEAR: a fast and accurate
926 Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.

Figure Legends

Figure 1. Five antimicrobial peptides (AMPs) of subfamily NDBP-5 were detected in the venom-gland transcriptome of *Hadrurus spadix*. All five had a highly conserved 23 amino-acid signal peptide and 35 amino-acid propeptide. The putative mature peptides were 12–13 amino acids in length and showed high levels of sequence variation in contrast to the rest of the precursor. This putative venom component clearly illustrates the challenges of proteomic detection and the analysis of patterns of selection. Because of the proteolytic processing, the actual peptides present in a tryptic digest differ substantially from computational predictions. The overall precursor appeared to have evolved mostly under purifying selection, but the mature peptide clearly did not.

Figure 2. Class-level abundance comparisons were similar across individuals for both venom-gland transcriptomes and venom proteomes, but transcriptomes and proteomes did not generally agree within an individual. Rather than posit hypothetical post-transcriptional regulatory mechanisms to account for the transcriptome/proteome discrepancies, we instead argue that our proteomic abundance estimates were rendered inaccurate due in part to unaccounted for post-translational processing of smaller peptide toxins and the general difficulty of detecting and quantifying small peptide toxins. In particular, the NDBP AMPs and ion-channel toxins appeared to be vastly underdetected in the proteome, and many of these are known to undergo extensive proteolytic processing prior to inclusion in the venom. Transcriptome abundances were based on transcripts per million (TPM) estimates, and proteome abundances were expressed as molar percentages. Transcriptome percentages refer only to reads mapped to putative toxins. Abbreviations: AMP—antimicrobial peptide, NDBP—nondisulfide bridged protein, PLA2—phospholipase A₂, SP—serine protease, VP—venom protein (unknown function).

Figure 3. A venom-gland transcript abundance comparison between a female (C0195) and male (C0196) *H. spadix* generally showed strong agreement. Transcript levels were highly correlated between the venom-gland transcriptomes of the two individuals for nontoxins, but less so for the proteomically confirmed venom components and putative toxins identified on the basis of homology. For the two toxin plots, the dashed lines represent the 99th percentile of differences between the two nontoxin measures. Points outside the dashed lines therefore represent toxins with unusually different expression differences relative to the nontoxins. Abbreviations: clr—centered logratio transform, n —number of transcripts, ρ —Spearman’s rank correlation coefficient, R —Pearson’s correlation coefficient, R^2 —coefficient of determination.

Figure 4. A venom proteomic comparison between a female (C0195) and male (C0196) *H. spadix* showed good agreement across individuals for proteins detected in both venom proteomes. Table 2 shows the proteomic presence/absence differences between the two individuals. Abbreviations: clr—centered logratio transform, n —number of transcripts, ρ —Spearman’s rank correlation coefficient, R —Pearson’s correlation coefficient, R^2 —coefficient of determination.

Figure 5. Transcript and protein abundances were positively correlated, but the agreement across data sets was weaker than for transcript or protein abundances across individuals for both the (A) female and (B) male. Abbreviations: clr—centered logratio transform, n —number of transcripts, ρ —Spearman’s rank correlation coefficient, R —Pearson’s correlation coefficient, R^2 —coefficient of determination.

Figure 6. We found a significant positive correlation between the discrepancy between protein and transcript abundance estimates and the sizes of the proteins (C0195: $p = 2.0 \times 10^{-8}$ and C0196: $p = 7.9 \times 10^{-3}$). Relative to transcript abundance estimates, small proteins tended to have lower-than-expected protein abundances, and larger proteins tended to have higher-than-expected protein abundances for both the (A) female and (B) male. Abbreviations: CDS—coding sequence, n —number of transcripts, nt—nucleotides, ρ —Spearman’s rank correlation coefficient, R —Pearson’s correlation coefficient, R^2 —coefficient of determination.

Tables

Table 1. Putative toxins identified in the venom-gland transcriptome of *Hadrurus spadix*.

	Putative toxin	Signal peptide	Precursor (aa)	C0195 TPM	C0196 TPM	C0195 fmol	C0196 fmol	Notes
	α KTx-1	Yes	65	13,037.99	29,537.81	—	—	
	α KTx-2	Yes	62	8,493.29	10,196.38	—	—	
	α KTx-3	Yes	69	439.39	296.56	—	—	
	α KTx-4	Yes	62	4,226.57	3,405.53	—	—	
	α KTx-5	Yes	63	3,623.94	2.36	83.90	105.43	
	α KTx-6	Yes	62	5,189.36	1,836.78	85.67	—	
	α KTx-7	Yes	61	4,351.83	1,997.06	—	—	
	α KTx-8	Yes	59	2,174.07	4,123.89	—	—	
	α KTx-9	Yes	59	2,499.25	5,909.05	—	—	
	α KTx-10	Yes	63	140.64	1,040.95	—	—	
	α KTx-11	Yes	62	2,351.78	1,152.83	—	—	
	α KTx-12	Yes	66	8,853.85	1,239.94	268.58	132.87	
	α KTx-13	Yes	63	31,488.32	13,847.59	—	25.92	
	α KTx-14	Yes	63	4,011.87	5,965.62	—	—	
	α KTx-15	Yes	66	17,255.29	2,429.14	120.38	108.49	
	α KTx-16	Yes	62	115.17	730.53	—	—	
	α KTx-17	Yes	68	25,340.46	17,970.44	723.73	623.21	
	α KTx-18	Yes	64	853.53	1,052.47	—	—	
	α KTx-19	Yes	62	1,787.60	24.48	—	—	
	α KTx-20	Yes	66	385.31	23.03	—	—	
	α KTx-21	Yes	62	1,161.63	262.57	—	—	
	α KTx-22	Yes	66	6,252.68	18,340.01	—	—	
	α KTx-23	Yes	69	4,338.69	363.88	67.50	—	
	α KTx-24	Yes	59	163.78	402.40	—	—	
	α KTx-25	Yes	63	524.83	141.09	—	—	
	α KTx-26	Yes	62	34,517.24	19,555.50	87.96	—	
	α KTx-27	Yes	63	898.76	1,627.25	158.53	505.82	
	α KTx-28	Yes	63	16,214.62	7,598.29	—	—	
	α KTx-29	Yes	63	46,653.41	53,186.62	—	—	
	α KTx-30	Yes	60	119.22	81.97	—	—	
	α KTx-31	Yes	61	119.44	200.77	—	—	
Aminopeptidase-1	AMP-1	Yes	95	35,832.35	28,587.21	1,089.75	1,293.95	No scorpion-toxin homologs
	AMP-2	Yes	78	4,292.15	175.25	—	—	Scorpine-like
	AMP-3	Yes	78	101,140.16	61,030.53	—	66.52	Subfamily NDBP-2
	AMP-4	Yes	102	12,361.31	4,612.40	27.52	—	Subfamily NDBP-2
	AMP-5	Yes	73	7,137.38	11,914.46	—	—	Scorpine-like
	AMP-6	Yes	73	9,745.12	32.58	—	—	Subfamily NDBP-4
	AMP-7	Yes	103	2,284.49	950.21	—	—	Subfamily NDBP-4
	AMP-8	Yes	74	40,693.00	16,277.25	—	—	Scorpine-like
	AMP-9	Yes	94	6,457.86	9,993.36	367.91	542.82	Subfamily NDBP-4
	AMP-10	Yes	80	17,748.31	14,094.27	—	—	Scorpine-like
	AMP-11	Yes	70	20,497.03	19,950.61	—	—	Subfamily NDBP-3
	AMP-12	Yes	70	24,929.66	7,534.74	—	—	Subfamily NDBP-5
	AMP-13	Yes	71	13,427.79	14,981.20	—	—	Subfamily NDBP-5
	AMP-14	Yes	71	17,771.20	36.05	—	—	Subfamily NDBP-5
	AMP-15	Yes	71	19,535.64	48.82	—	—	Subfamily NDBP-5
	β KTx-1	Yes	79	3,514.12	4,553.94	40.00	—	
	CaTx-1	Yes	74	13,222.11	285.73	—	—	Hadrucaucin-like
Chitinase-1	CP-1	Yes	96	3,631.86	8,976.95	131.60	72.11	
	CR	Yes	454	531.77	1,340.18	1,000.79	977.86	
	CR	Yes	240	474.70	1,438.43	176.28	—	
	CRISP-1	Yes	404	2,193.90	3,637.07	267.68	371.05	
	CRISP-2	Yes	392	1,869.95	680.69	503.96	268.74	
	Cystatin-1	Yes	136	1,178.17	3,188.46	136.50	133.40	

Table 1. (continued)

Putative toxin	Signal peptide	Precursor (aa)	C0195 TPM	C0196 TPM	C0195 fmol	C0196 fmol	Notes
Cystatin-2	Yes	131	446.60	1,017.77	85.39	140.86	
G3PDH2	No	338	649.31	618.07	13.14	—	
HistP	Yes	381	219.58	264.12	301.68	107.23	
HYAL-1	Yes	394	730.43	721.75	219.34	228.52	
KUN-1	Yes	88	359.87	159.01	—	—	
KUN-2	Yes	88	542.51	0.70	—	—	
La1-1	Yes	103	19,501.93	35,018.51	435.10	581.09	
La1-2	Yes	100	15,383.44	82,092.80	197.17	170.17	
La1-3	Yes	95	5,818.78	11,967.69	—	—	
LDLR-1	No	204	295.73	18.17	280.21	—	
NUC-1	Yes	579	442.71	1,146.75	1,479.78	1,047.15	
PeptidaseM13-1	No	740	208.22	375.99	63.88	13.74	No scorpion-toxin homologs
PeptidaseM2-1	Yes	590	1,916.69	3,370.37	3,091.11	5,310.24	
PeptidaseM2-2	Yes	629	646.35	858.88	81.86	14.03	
PeptidaseM2-3	Yes	522	423.20	526.40	80.13	24.68	
PeptidaseM2-4	Yes	599	2,681.98	54.11	1,049.55	3.36	
PLA2-1	Yes	243	5,819.25	1,336.30	190.63	26.31	Phaiodactylipin-like
PLA2-2	Yes	242	4,173.53	863.69	408.27	83.95	Phaiodactylipin-like
PLA2-3	Yes	236	1,462.75	1,045.43	187.44	91.75	Phaiodactylipin-like
SP-1	Yes	247	30.92	182.77	—	—	
SP-2	Yes	298	3,493.29	2,319.90	501.93	231.14	
SP-3	Yes	282	4,446.15	11,317.95	647.66	774.83	
SP-4	Yes	264	386.93	448.31	171.76	91.92	
SP-5	Yes	303	195.97	330.73	14.63	7.40	
SupDismutase	No	152	760.60	979.92	29.23	—	
Transferrin	Yes	710	363.84	229.45	17.88	—	
Transketolase	No	624	264.88	268.19	21.74	—	
VP-1	Yes	111	283.09	2,036.76	—	—	Group I
VP-2	Yes	77	4,018.77	1,429.64	—	—	
VP-3	Yes	280	3,733.18	6,779.21	868.77	434.42	
VP-4	Yes	66	9,083.85	7,629.79	496.07	646.40	
VP-5	Yes	58	2,458.81	2,478.78	—	—	
VP-6	Yes	123	2,030.64	1,958.38	5.17	3.36	Group I
VP-7	Yes	109	5,157.84	7,129.88	—	—	
VP-8	Yes	228	360.23	993.88	—	—	
VP-9	Yes	164	3,849.20	1,904.11	171.42	31.71	
VP-10	No	353	2,413.61	1,439.90	255.12	6.98	
VP-11	Yes	131	1,458.31	1,134.66	88.96	—	
VP-12	No	173	501.20	2,096.82	—	—	
VP-13	Yes	100	81.99	7,096.12	—	114.39	
VP-14	Yes	106	1,774.46	973.88	159.93	8.38	
VP-15	Yes	95	385.11	142.72	—	—	Group II
VP-16	Yes	94	2,955.34	1,720.25	74.59	47.65	
VP-17	Yes	97	3,711.22	34.35	26.61	—	Group III
VP-18	Yes	97	2,901.55	28.80	—	—	Group III
VP-19	Yes	136	1,418.78	417.04	38.07	—	
VP-20	Yes	95	35.97	64.31	—	—	
VP-21	Yes	115	7,552.69	6,340.78	986.73	659.67	Group IV
VP-22	Yes	97	1,342.95	7.12	16.89	—	Group III
VP-23	Yes	97	1,166.10	3,926.19	149.40	97.93	Group III
VP-24	Yes	242	737.61	651.11	—	—	
VP-25	Yes	166	412.88	626.95	21.51	5.03	
VP-26	Yes	65	3,846.89	2,223.05	90.93	15.36	
VP-27	Yes	97	3,797.18	34,443.71	176.08	301.36	Group III
VP-28	Yes	69	3,278.48	9,017.41	29.23	20.60	
VP-29	Yes	115	3,506.55	3,145.11	599.42	316.11	Group IV
VP-30	Yes	56	4,023.39	3,542.87	65.92	26.53	
VP-31	Yes	117	6.62	5,864.74	—	486.99	Group IV
VP-32	Yes	108	573.00	130.63	—	—	Group II
VP-33	Yes	113	687.40	4.46	6.87	—	

Table 1. (continued)

Putative toxin	Signal peptide	Precursor (aa)	C0195 TPM	C0196 TPM	C0195 fmol	C0196 fmol	Notes
VP-34	Yes	116	13,483.43	4,684.46	996.12	108.92	Group IV
VP-35	Yes	115	567.48	873.29	47.34	22.34	Group IV
VP-36	Yes	116	5,549.60	17,533.75	712.52	120.32	Group IV
VP-37	Yes	106	346.11	255.20	—	—	Group V
VP-38	No	165	262.13	592.24	—	—	
VP-39	Yes	102	139.84	175.68	—	—	Group V
VP-40	Yes	130	351.98	161.24	—	—	
VP-41	Yes	88	247.08	144.30	—	—	
VP-42	Yes	100	51.05	207.19	—	5.03	
VP-43	Yes	129	419.73	309.89	16.40	—	
VP-44	Yes	117	15,867.65	13,779.53	995.34	616.68	Group IV
VP-45	Yes	114	2,266.54	2,719.95	248.25	106.94	Group IV
VP-46	Yes	107	240.22	31.47	—	—	Group VI
VP-47	Yes	118	21,917.43	32,909.67	—	—	
VP-48	Yes	90	186.94	36.74	—	—	Group VII
VP-49	Yes	90	1,923.94	367.44	—	—	Group II
VP-50	Yes	91	476.79	149.40	—	—	
VP-51	Yes	76	20,583.57	22,014.79	—	—	
VP-52	Yes	131	168.68	48.82	18.71	—	
VP-53	Yes	95	279.84	3.95	—	—	Group VII
VP-54	Yes	134	769.97	1,501.64	—	—	
VP-55	No	135	161.57	692.88	—	—	
VP-56	Yes	95	103.51	39.49	—	—	
VP-57	Yes	87	2,919.66	2,685.31	68.40	—	Group III
VP-58	Yes	148	1,918.06	2,278.26	—	—	
VP-59	Yes	104	45.22	32.72	—	—	
VP-60	No	143	531.83	20.40	138.42	—	
VP-61	Yes	92	580.64	0.62	13.77	—	Group III
VP-62	Yes	78	968.28	2,702.56	81.86	—	
VP-63	Yes	134	419.12	9.85	—	—	
VP-64	Yes	141	177.87	269.81	—	—	
VP-65	Yes	109	2,780.20	983.45	250.02	59.57	
VP-66	Yes	104	107.41	43.93	—	—	Group VI

Abbreviations: AMP—antimicrobial peptide, CaTx—Ca²⁺-channel toxin, CP—cysteine-peptidase, Creactive—C-reactive protein, CRISP—cysteine-rich secretory protein, G3PDH2—glyceraldehyde 3-phosphate dehydrogenase, HistP—histidine phosphatase, HYAL—hyaluronidase, KTx—K⁺-channel toxin, KUN—Kunitz-type protease inhibitor, LDLR—low density lipoprotein receptor, NUC—nucleotidase, PLA2—phospholipase A₂, SP—serine proteinase, SupDismutase—superoxide dismutase, TPM—transcripts per million, VP—venom protein.

Table 2. Presence/absence differences in the two venom proteomes.

Protein	C0195 rep 1	C0195 rep 2	C0195 rep 3	C0196 rep 1	C0196 rep 2	C0196 rep 3	C0195 avg	C0196 avg
α KTx-6	79.46	100.11	77.42	—	—	—	85.67	—
α KTx-13	—	—	—	—	—	77.77	—	25.92
α KTx-23	73.80	66.78	61.93	—	—	—	67.50	—
α KTx-26	107.86	88.98	67.05	—	—	—	87.96	—
AMP-3	—	—	—	60.32	83.71	55.53	—	66.52
AMP-4	—	—	82.55	—	—	—	27.52	—
β KTx-1	39.73	38.96	41.30	—	—	—	40.00	—
CReactive	164.65	183.59	180.59	—	—	—	176.28	—
G3PDH2	22.73	16.70	—	—	—	—	13.14	—
LDLR-1	266.79	300.40	273.45	—	—	—	280.21	—
SupDismutase	34.06	27.83	25.81	—	—	—	29.23	—
Transferrin	—	27.83	25.81	—	—	—	17.88	—
Transketolase	22.73	16.70	25.81	—	—	—	21.74	—
VP-11	79.46	94.55	92.86	—	—	—	88.96	—
VP-13	—	—	—	105.56	109.86	127.76	—	114.39
VP-17	—	33.39	46.43	—	—	—	26.61	—
VP-19	28.40	44.52	41.30	—	—	—	38.07	—
VP-22	28.40	22.26	—	—	—	—	16.89	—
VP-31	—	—	—	462.47	470.81	527.67	—	486.99
VP-33	—	—	20.62	—	—	—	6.87	—
VP-42	—	—	—	15.08	—	—	—	5.03
VP-43	17.00	16.70	15.50	—	—	—	16.40	—
VP-52	22.73	33.39	—	—	—	—	18.71	—
VP-57	107.86	61.22	36.12	—	—	—	68.40	—
VP-60	147.59	133.50	134.16	—	—	—	138.42	—
VP-61	—	—	41.30	—	—	—	13.77	—
VP-62	85.13	77.91	82.55	—	—	—	81.86	—

Quantities are given in fmol. Abbreviations: AMP—antimicrobial peptide, avg—average, Creactive—C-reactive protein, G3PDH2—glyceraldehyde 3-phosphate dehydrogenase, KTx—K⁺-channel toxin, LDLR—low density lipoprotein receptor, rep—replicate, SupDismutase—superoxide dismutase, VP—venom protein.

Table 3. Comparison to *H. gertschi* transcriptome (Schwartz et al., 2007)

<i>H. gertschi</i> transcript	GenBank accession	Description	<i>H. spadix</i> match	Proteome	% nt	dN/dS
HGE024 Contig2	EL698898	α -KTx 6 subfamily	None	—	—	—
HGE025 Contig5	EL698899	α -KTx	None	—	—	—
HGE026 Hgscplike2	EL698900	Scorpine-like group	AMP-7	No	3.4	0.17
HGE027 NDPB_5.5	EL698901	NDBP group 5	AMP-12	No	14.0	0.35
HGE028 NDPB_5.6	EL698902	NDBP group 5	None	—	—	—
HGE029 NDPB_3.7	EL698903	NDBP group 3	AMP-10	No	6.5	0.50
HGE030 Hg1	EL698904	SP inhibitor	KUN-1	No	5.8	0.52
HGE031 PLA2	EL698905	Group III PLA2	PLA2-1	Yes	7.0	0.58
HGE034 Hgscplike1	EL698908	Scorpine-like group	AMP-1	Yes	6.6	0.60
HGE035 HgbetaKTx1	EL698909	β -KTx	β KTx-1	Yes	6.1	0.64

Abbreviations: AMP—antimicrobial peptide, KTx—K⁺-channel toxin, KUN—Kunitz-type protease inhibitor, NDBP—non-disulfide bridged protein, nt—nucleotide, PLA2—phospholipase A₂, SP—serine proteinase.