Toxicon 142 (2018) 14–29

Contents lists available at ScienceDirect

Toxicon

journal homepage: www.elsevier.com/locate/toxicon

Venom-gland transcriptomics and venom proteomics of the Hentz striped scorpion (*Centruroides hentzi*; Buthidae) reveal high toxin diversity in a harmless member of a lethal family

Micaiah J. Ward, Schyler A. Ellsworth, Darin R. Rokyta*

Department of Biological Science, Florida State University, Tallahassee, FL 32306, USA

ARTICLE INFO

Article history:
Received 9 November 2017
Received in revised form
7 December 2017
Accepted 11 December 2017
Available online 14 December 2017

Keywords:
Scorpion
Venom
Transcriptome
Proteome

ABSTRACT

Of the 14 extant scorpion families, Buthidae has the most thoroughly characterized venoms. Most of this characterization, however, has been limited to species with medically significant stings, including members of the *Centruroides* genus, which have caused human deaths (e.g., *Centruroides sculpturatus*). To understand the origin and evolution of highly toxic venoms, we should also characterize the more harmless venoms of close relatives. We used Illumina sequencing to separately characterize the venom-gland transcriptomes of a male and female Hentz striped scorpion (*Centruroides hentzi*) and performed independent quantitative mass-spectrometry analysis of the venom from each individual, providing the first full venom characterization of a *Centruroides* species that poses no serious threat to humans. We identified 59 venom proteins that were proteomically confirmed, 63 additional transcripts that were identified on the basis of homology to known toxins, and 355 nontoxins expressed in the venom-glands. The most abundant toxins belonged to the Na^+ and K^+-channel toxin classes. Antimicrobial peptides and peptidases were also identified, along with a large group of venom proteins that could not be classified based on homology, suggesting *C. hentzi* is a source of previously untapped toxin diversity.

1. Introduction

Scorpions are one of the most ancient extant venomous lineages, originating approximately 430 million years ago, and have since diversified into 14 recognized families and over 1700 described species (Soleglad and Fet, 2003; Stockmann and Ythier, 2010). Only approximately 30 scorpion species are known to be harmful to humans, and all but one of these (*Hemiscorpius lepturus* of the family Scorpionidae) are members of the family Buthidae (Pipelzadeh et al., 2007; Chippaux and Goyffon, 2008), which are known for their rich diversity and high abundance of ion-channel toxins (Fet et al., 2003). The impact on human populations has resulted in buthids being the most well characterized scorpion family in terms of venom (Ruiming et al., 2010; Ma et al., 2012; Alvarenga et al., 2012; Rendón-Anaya et al., 2012; Valdez-Velazquéz et al., 2013; Diego-García et al., 2014; Mille et al., 2014; de Oliveira et al., 2015), but has also led to a biased representation of scorpions in the literature by focusing only on scorpions that cause severe physiological symptoms in humans, while neglecting those that are much less harmful. To better understand why some species are harmful to humans, we should also investigate why some of their close relatives are not.

In addition to buthids, venom characterizations (either full or partial) have been completed for seven other scorpion families: Caraboctonidae (Schwartz et al., 2007; Rokyta and Ward, 2017), Chaerilidae (He et al., 2013), Euscorpiidae (Ma et al., 2009, 2012; Santibáñez-López et al., 2017), Scorpionidae (Ma et al., 2010; Diego-García et al., 2012), Superstitioniidae (Santibáñez-López et al., 2016), Urodacidae (Luna-Ramírez et al., 2015), and Vaejovidae (Quintero-Hernández et al., 2015). Only a handful of these characterizations have used high-throughput transcriptomic methods (Rendón-Anaya et al., 2012; Luna-Ramírez et al., 2015; de Oliveira et al., 2015; Santibáñez-López et al., 2016). The first combined high-coverage venom-gland transcriptomic and venom proteomic analysis for a scorpion species was recently completed for *Hadrurus spadix* (Rokyta and Ward, 2017), a member of the family Caraboctonidae. These venom characterizations revealed many previously unknown and medically significant venom

* Corresponding author. Florida State University, Department of Biological Science, 319 Stadium Dr., Tallahassee, FL 32306-4295, USA.
E-mail address: drokyta@bio.fsu.edu (D.R. Rokyta).

https://doi.org/10.1016/j.toxicon.2017.12.042
0041-0101/

components, such as antimicrobial peptides and anticancer peptides (Zeng et al., 2005; Harrison et al., 2014; Ortiz et al., 2015), as well as a plethora of ion-channel toxins with mammal and insect specificity (de la Vega et al., 2010; Quintero-Hernández et al., 2013). In *H. spadix*, 66 of 148 putative toxins identified could not be assigned a tentative function on the basis of homology to toxins with known functions (Rokyta and Ward, 2017). Previous work on non-buthids has demonstrated the biodiscovery potential that could result from shifting the emphasis of scorpion venom research from the well-characterized harmful species to the generally ignored harmless species.

Centruroides are members of the family Buthidae that are widely distributed throughout North, South, and Central America and are often referred to as bark scorpions. A partial venom-gland transcriptome and venom proteome characterization was described for *Centruroides tecomanus* (Valdez-Velazquéz et al., 2013, 2016), and high-throughput 454 transcriptome sequencing was completed for *Centruroides noxius* (Rendón-Anaya et al., 2012). A small number of individual toxins, primarily Na^+-channel toxins, have also been purified and characterized from *Centruroides sculpturatus* (Wang and Strichartz, 1983), *Centruroides noxius* (Possani et al., 1985), *Centruroides suffusus* (Martin et al., 1987), and *Centruroides margaritatus* (Garcia-Calvo et al., 1993). All of these species are considered medically significant within their ranges, and all are potentially deadly to humans (Chippaux and Goyffon, 2008). Venom from the Arizona bark scorpion (*C. sculpturatus*), for example, can give rise to symptoms of methamphetamine overdose and even lead to death (Skolnik and Ewald, 2013; Strommen and Shirazi, 2015). Some *Centruroides* species, however, are not considered harmful to humans. *Centruroides vittatus*, which is native primarily to the south central United States (Shelley and Sissom, 1995), has been reported to have mild stings (More et al., 2004; Kang and Brooks, 2017). *Centruroides hentzi*, which is native to the southeastern United States (Shelley and Sissom, 1995; Stevenson et al., 2012), has no known records of envenomations requiring hospitalization, or even treatment, in medical literature (Kang and Brooks, 2017), and stings that have been reported were said to be painful, but short-lived and with no medical consequences (Stevenson et al., 2012). To begin to address differences between the deadly *Centruroides* species and their significantly less-harmful relatives, we completed the first high-throughput venom-gland transcriptomic and proteomic characterization for a member of this genus, the Hentz striped scorpion (*C. hentzi*). We independently sequenced venom-gland transcriptomes from a male and female *C. hentzi* and conducted a parallel LC-MS/MS analysis of their venoms, providing a complete venom characterization for a member of the most well-studied, medically important scorpion families.

2. Materials and methods

2.1. Scorpions, venoms, and venom-glands

Venom-gland transcriptomic and venom proteomic analyses were performed independently on two individuals of *C. hentzi* labeled C0136 and C0148. These specimens were collected in the northern Florida counties of Madison (C0136) and Wakulla (C0148) in the north-central region of the range of *C. hentzi* (Shelley and Sissom, 1995; Stevenson et al., 2012). *Centruroides* species are sexually dimorphic; females have larger bodies and shorter metasomal (tail) segments, and males have smaller, more slender bodies and longer metasomal segments (Fig. 1; Polis and Sissom, 1990; Stahnke and Calos, 1977). On the basis of these distinguishing features, C0136 was determined to be male, and C0148 was determined to be female.

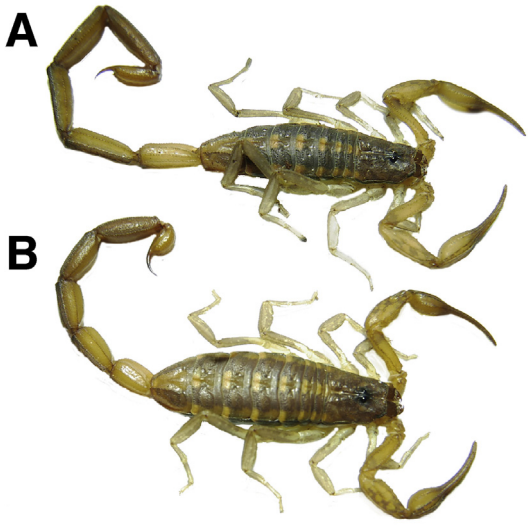

Fig. 1. Dorsal view of a male (A) and female (B) *C. hentzi* clearly illustrates sexual dimorphism in this species. Males have smaller, more slender bodies and longer metasomal segments in comparison to the females, which have larger bodies and shorter metasomal segments. Individual male and female *C. hentzi* shown here are representative only and do no correspond with individuals C0136 and C0148 discussed throughout the manuscript.

Venom and venom-glands were extracted using methods previously described in detail (Rokyta and Ward, 2017). After a 5 min anesthetization with CO_2, electrical stimulation was applied to the base of the telson (stinger) to induce a muscle contraction and expel the venom. Venom was then lyophilized and stored at −80°C until later use. Four days after venom extraction, venom-glands were removed under a stereoscopic microscope with micro-surgical dissection instruments. Prior to gland removal, scorpions were fully anesthetized with CO_2 for approximately 15 min. The metasoma and intact telson were then cut from the scorpion body and taped to a sterile surface such that the telson was visible under the microscope. Using a micro-surgical blade, the telson was gently cut open and tweezers were used to peel back each side of the telson to access the venom-glands. Venom-glands were removed by scraping out the inside of the telson using curved surgical tweezers, and this tissue was immediately transferred to 100 μL of RNAlater. The gland tissue in RNAlater was stored at 4 °C overnight and transferred to −80°C until further use. The scorpion body and remaining metasoma of each specimen were preserved in 95% ethanol and stored at −80°C.

2.2. Transcriptome sequencing

Scorpion venom-gland RNA extraction was performed as previously described (Rokyta and Ward, 2017). Briefly, 500 μL of Trizol (Invitrogen) was mixed with the 100 μL RNAlater containing the venom-gland tissue, and this mixture was homogenized using a sterile syringe with a 20 gauge needle. An additional 500 μL of Trizol was then added, along with 20% chloroform, and the full tissue mixture was transferred to phase lock heavy gel tubes (5Prime) and centrifuged to separate the RNA from DNA and other cellular debris. Isopropyl alcohol was used to pellet the isolated RNA. The isolated RNA was then washed with 75% ethanol, and the final, purified RNA was washed with 70% ethanol. After Qubit RNA quantification (Thermo Fisher Scientific), the quality of purified RNA was checked using the RNA 6000 Pico Bioanalyzer Kit (Agilent Technologies) according to manufacturer's instructions. The estimated total RNA yields for C0136 and C0148 were 82.7 ng and

91.6 ng, respectively.

After performing quantification and quality analyses, each total RNA sample was used to isolate mRNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs), resulting in inputs of 79.3 ng and 87.9 ng of total RNA for C0136 and C0148, respectively. Following methods described by Rokyta and Ward (2017), fragment sizes of approximately 370 nucleotides (adapter-ligated) were generated using a 15.5 min fragmentation step, and the purified mRNA was immediately used for cDNA library preparation using the NEBNext Ultra RNA Library Prep Kit with the High-Fidelity 2X Hot Start PCR Master Mix and Multiplex Oligos for Illumina (New England Biolabs) per manufacturer's instructions. Agencourt AMPure XP PCR Purification Beads were used to purify DNA both throughout and at the end of the protocol. A High Sensitivity DNA Bioanalyzer Kit (Agilent Technologies) was used to assess library quality, following the manufacturer's instructions. The total cDNA yield for C0136 was 50.6 ng with an average fragment size of 392 bp, and the total yield for C0148 was 222.8 ng with an average fragment size of 377 bp. The amplifiable concentrations for each sample were determined using KAPA PCR performed by the Florida State University Molecular Cloning Facility. The amplifiable concentration for C0136 was 5.53 nM, and the concentration for C0148 was 33.6 nM. These samples were diluted to ~5 nM and pooled with other 5 nM cDNA libraries to be run on the same sequencing lane. The quality of the pooled DNA library sample was then checked using a High Sensitivity DNA Bioanalyzer Kit (Agilent Technologies), and amplifiable concentration of the pooled sample was confirmed with an additional round of KAPA PCR. Sequencing was performed by the Florida State University College of Medicine Translational Laboratory using an Illumina HiSeq 2500.

2.3. Proteomics

Proteomic analyses were completed following methods previously described (Rokyta and Ward, 2017). Venom protein samples were quantified using the Qubit Protein Assay Kit with a Qubit 1.0 Fluorometer (Thermo Fisher Scientific), and approximately 5 μg of whole venom was digested using LC/MS grade solvents with the Calbiochem ProteoExtract All-in-One Trypsin Digestion Kit (Merck, Darmstadt, Germany) following the manufacturer's instructions. After digestion, approximately 4.3 μg of digested venom proteins remained, and these were then frozen and dried using a SpeedVac for 4.5 h.

LC-MS/MS was performed by the Florida State University College of Medicine Translational lab following methods described by Rokyta and Ward (2017). The digested venom protein samples were resuspended in 0.1% formic acid to reach a final concentration of 250 ng/μL. Three highly-purified recombinant *Escherichia coli* proteins were purchased from Abcam at known concentrations and mixed in specified proportions prior to digestion to yield the final desired concentrations of 2500 fmol of P31697 (Chaperone protein FimC), 250 fmol of P31658 (Protein deglycase 1), and 25 fmol of P00811 (Beta-lactamase ampC) per injection. The digested peptide mix was infused into samples prior to LC-MS/MS injection. An externally calibrated Thermo Q Exactive HF (high-resolution electrospray tandem mass spectrometer) was used in conjunction with Dionex UltiMate3000 RSLCnano System to perform LC-MS/MS analysis of each sample, using a 2 μL aliquot. Beginning with LC, samples were aspirated into a 50 μL loop and loaded onto the trap column (Thermo μ-Precolumn 5 mm, with nanoViper tubing 30 μm i.d.×10 cm), with a flow rate of 300 nL/min for separation on the analytical column (Acclaim pepmap RSLC 75 μMx 15 cm nanoviper). A 60 min linear gradient from 3% to 45% B was performed using mobile phases A (99.9% H_2O (EMD Omni Solvent) and 0.1% formic acid) and B (99.9% ACN and 0.1% formic acid). The LC eluent was directly nanosprayed into a Q Exactive HF mass spectrometer (Thermo Scientific). The Q Exactive HF was operated in a data-dependent mode and under direct control of the Thermo Excalibur 3.1.66 (Thermo Scientific) throughout the chromatographic separation. A data-dependent top-20 method was used for MS data acquisition, selecting the most abundant, not-yet-sequenced precursor ions from the survey scans (350–1700 m/z). Sequencing was performed using 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. A 15-s dynamic exclusion window was used and ions with a single charge, a charge more than seven, or an unassigned charge were excluded. All measurements were performed at room temperature, and each sample was run and measured in triplicate to facilitate lable-free quantification and account for any machine-related variability between samples. The resulting raw files were searched with Proteome Discoverer 1.4 using SequestHT as the search engine, custom-generated FASTA databases, and percolator to validate peptides. The SequestHT search parameters used were: 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.

Proteomic abundances for each individual (C0136 and C0148) were estimated by calculating separate conversion factors for each of three replicates, using the known concentrations of the *E. coli* internal standards and their observed quantitative values (i.e., normalized spectral counts) calculated by Scaffold. Conversion factors were calculated by finding the slope of the best fit line of the known internal standard concentrations and the observed normalized spectral counts, with an intercept at the origin. We then used these conversion factors to convert the normalized spectral counts for each venom protein in each replicate to a concentration value, and the final concentrations for each sample were averaged across the three replicates for each individual (C0136 and C0148).

2.4. Protein bioanalyzer electrophoresis

Two venom samples from separate venom extractions for each *C. hentzi* individual (C0136: V0136 and V0188; C0148: V0263 and V0552) were processed using the Agilent Protein 80 assay (Santa Clara, CA). The majority of venom proteins have molecular masses between 3 and 60 kDa, so the Protein 80 Kit was selected on the basis of its ability to separate proteins between 5 and 80 kDa (Zancolli et al., 2017). On the basis of protein concentrations quantified from the Qubit Protein Assay Kit (Thermo Fisher Scientific), approximately 4–6 μg of crude venom from each sample was reconstituted in PBS, reduced in dithiothreitol, and diluted to a concentration of 44.4–66.6 ng/μg in LC/MS water. Samples were run on the protein bioanalyzer chip according to the manufacturer's instructions. Each sample was prepared with an internal marker of known concentration to calculate the relative abundances of the detected proteins. The area under the peak of the internal standard was then compared to the area under the peaks from the sample to get relative concentrations. Two additional internal standards were present in the gel matrix within each run to facilitate comparison and alignment of different samples. The Agilent 2100 Expert

software (version B.02.09 (SR1), Santa Clara, CA, USA) used the internal standards to align the samples and the ladder.

2.5. Transcriptome assembly and analysis

Illumina quality filtering was implemented on the generated transcriptome sequencing data, resulting in filtered raw read-pairs for each individual. Because we had targeted an insert size of around 250 nucleotides and performed 150 paired-end sequencing, we expected most resulting read-pairs to show significant 3' overlaps. We therefore merged reads using PEAR version 0.9.6 (Zhang et al., 2014), and these merged reads were used for subsequent analyses. We generated our primary transcriptome assembly using DNAStar NGen version 12.3.1, using 10 million merged reads and the default transcriptome assembly settings, retaining only contigs with at least 200 reads. We used five search strategies for identifying and annotating proteins in the transcriptome, because we were not expecting to find many known homologs of toxins for this species in public databases. Two of the strategies used the whole-venom mass-spectrometry results and the generated protein databases by applying TransDecoder version 2.0.1 to our assembled transcriptomes. We first created a database using the TransDecoder-predicted protein sequences with a minimum length of 50 and searched our mass-spectrometry results against this database. We then filtered these results using Scaffold Viewer version 4.6.0. To accommodate possible short proteins in the venom, protein and peptide false-discovery rates were set to 1.0%, and the minimum number of peptides was set to one. In our second strategy, we wanted to ensure that small peptides were not being missed by the TransDecoder predictions, so we generated another database using all possible protein or peptide sequences of at least 50 amino-acids from all six possible reading frames. We filtered the results as before in Scaffold. For this second strategy, all contigs already annotated in the first strategy (i.e., those predicted by TransDecoder) were excluded. Our third strategy aimed to identify proteins from the transcriptome with homology to known toxins. To do this, 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) and attempted to annotate full-length putative toxins that showed a match against at least 80% of the length of a known toxin. For our fourth strategy, 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 generate a general database of toxins and nontoxins expressed in the venom-glands. Only transcripts comprised of at least 1000 reads with a match of at least 95% of the length of a known protein were considered. In our fifth strategy, we used Extender (Rokyta et al., 2012) to assemble the transcriptome from 1000 random reads to better ensure that no high-abundance transcripts were missed. Reads were only used if they had phred qualities of $\geq$30 at all positions and an exact match of 120 nucleotides for extension. We then searched the resulting contigs against the UniProt animal toxins database with blastx. By combining the results from all five search strategies described above, we generated a consensus transcriptome for each of the two sequenced individuals. To remove any exact duplicates, we used cd-hit-est version 4.6 (Li and Godzik, 2006) on the coding sequences. We then used bowtie2 version 2.2.7 (Langmead and Salzberg, 2012) to align merged reads and screen for chimeric sequences or other coverage anomalies. The final representative transcriptome for *C. hentzi* was generated by merging the results from the two individuals. Transcripts were clustered using cd-hit-test using only their coding sequences and a global sequence identity of 0.98. Transcript abundances were estimated on the basis of bowtie (Langmead et al., 2009) version 1.1.2 alignments, using RSEM (Li et al., 2011) version 1.2.28, and alignments were based on all merged reads for each individual. We used the centered logratio transform (Aitchison, 1986) on all of our transcriptome and proteome abundances as previously described (Rokyta et al., 2015); this transform is equivalent to a log transform for linear analyses and does not affect rank-based analyses. The presence of signal peptides was verified with SignalP version 4.1 using the default settings (Petersen et al., 2011). In the few cases where a signal peptide was not detected in an identified putative toxin using default settings, the sensitive option in SignalP was also tested.

Because our two *C. hentzi* RNA-seq libraries were prepared and sequenced alongside RNA-seq libraries from other species, we performed a final quality-control step on our assembled transcripts to ensure the purity of our final transcriptome. Using PEAR version 0.9.6 (Zhang et al., 2014), we merged the reads from each library that was prepared alongside and/or sequenced in the same HiSeq lane as our focal individuals. We then used bowtie2 version 2.2.7 (Langmead and Salzberg, 2012) to align these merged reads against our *C. hentzi* coding sequences. Transcripts were removed as contaminants if they showed >100× higher coverage for another library relative to the highest-coverage of the two *C. hentzi* libraries, had read coverage over the entire length of the coding sequence, and had no homozygous variants relative to the consensus sequence.

2.6. Data availability

The raw transcriptome reads were submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject PRJNA340270, BioSamples SAMN07655950 (C0136) and SAMN07655951 (C0148), and SRA accessions SRR6041834 (C0136) and SRR6041835 (C0148). 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 PXD007788 and 10.6019/PXD007788. 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 GFWZ00000000. The version described in this paper is the first version, GFWZ01000000.

3. Results and discussion

3.1. Venom-gland transcriptomes

After Illumina quality filtering, we generated 18,977,620 raw read pairs for the male (C0136). A total of 16,240,246 of these reads were merged, and ten million merged reads were assembled into 3880 contigs using NGen. Using blastx hits to the UniProt toxins database, we annotated 141 coding sequences. We annotated an additional 342 coding sequences on the basis of blastx hits to the NCBI nr database, 35 coding sequences using TransDecoder predictions with MS-directed analysis, and another 17 coding sequences using all ORFs in our MS-directed analysis. After a blastx search against the UniProt toxins database, the extender assembly produced 45 annotated coding sequences. All annotated sequences were combined and screened for chimeras and duplicates, leaving 398 identified unique coding sequences.

For individual C0148 (female), we generated 20,032,620 read pairs passing the Illumina quality filter. Based on their 3' overlaps, 17,678,591 pairs were merged, and ten million of the merged reads were assembled into 3470 contigs using NGen. Using blastx hits to the UniProt database, we annotated 131 coding sequences. We annotated an additional 259 coding sequences based on blastx hits

to the NCBI nr database, 35 coding sequences using TransDecoder predictions with MS-directed analysis, and another 14 coding sequences using all ORFs in our MS-directed analysis. After a blastx search against the UniProt toxins database, the extender assembly produced 56 annotated coding sequences. All annotated sequences were combined and screened for chimeras and duplicates, generating 322 identified unique coding sequences.

The annotated transcripts from individual C0136 (398 total) and C0148 (322 total) were combined, and duplicates were removed, producing a final *C. hentzi* transcriptome of 477 unique protein-coding transcripts. These transcripts were used for all subsequent transcript-abundance estimates and LC-MS/MS analyses for both individuals and were divided into three classes. The first were those that were proteomically confirmed in the venom, the second were those that were identified as toxins on the basis of homology, and third were the nontoxins. For simplicity, we will use the term "toxin" to refer to transcripts with corresponding proteins detected in the venom proteome, as well as those identified by homology to an Arachnida toxin using the UniProt toxins database. These transcripts have high likelihoods of encoding true toxic venom components, but have not had toxic functions confirmed. The proteomically confirmed toxins consisted of 59 transcripts encoding proteins that were detected in the proteome of one or both *C. hentzi* individuals (Table 1). These transcripts were considered to have a high likelihood of encoding toxins because they were present in the venom. Some (i.e., those detected at low levels), however, may be nontoxins that leaked into the venom as a result of cell rupture of gland tissue during venom extraction. The proteomically confirmed toxins accounted for 674,761.00 and 721,833.00 transcripts per million (TPM) of the mapped reads in C0136 and C0148, respectively. The homology-based toxins consisted of 63 transcripts that were highly expressed in the transcriptome (Table 1). These transcripts encode proteins with homology to known animal toxins in other species or with homology to proteins or peptides identified in the proteome of *C. hentzi* (i.e., K^+-channel toxins). Our approach does not allow for proteomic detection of small peptides that have undergone extreme post-translational proteolytic processing, and both K^+-channel toxins and antimicrobial peptides have shown detectability challenges in other scorpion venoms (Rokyta and Ward, 2017). Large proteins (>200 amino-acid precursors) that were not detected in the proteome and had no UniProt toxin matches from class Arachnida were excluded from the homology-based toxins. Of the mapped reads, the homology-based toxins accounted for 135,506.24 TPM in C0136 and 139,219.42 TPM in C0148. The nontoxins consisted of 355 transcripts that were not assigned to either the proteomically confirmed or homology based toxin classes. These transcripts likely encode proteins that are essential to cell function and toxin production but have low likelihoods of encoding toxic proteins or peptides. The nontoxins accounted for 189,732.80 TPM of the mapped reads in C0136 and 138,947.59 TPM in C0148.

3.2. Ion-channel toxins

Ion-channel toxins are among the most diverse groups of scorpion toxins, both in terms of function and relative abundance (de la Vega et al., 2010). Peptides belonging to this group of toxins are characterized by their modification of the Na^+- and K^+-channel gating mechanism and their interference in the function of ligand-activated ryanodine Ca^{2+}-channels (Quintero-Hernández et al., 2013). We identified 36 putative Na^+-channel toxins and 32 K^+-channel toxins in the venom-gland transcriptome of *C. hentzi*, accounting for 680,048.79 TPM in C0136 and 748,598.52 TPM in C0148 (i.e., 83.9% and 87.0% of the total toxin transcriptional output, respectively; Fig. 2 and Table 1). We did not identify any Ca^{2+}-channel toxins in the transcriptome of *C. hentzi*.

The majority of ion-channel toxin transcripts identified were Na^+-channel toxins (NaTxs), with 449,727.74 TPM in C0136 and 457,853.38 TPM in C0148 (Table 1). The NaTxs were by far the most abundant group of toxins identified in the *C. hentzi* transcriptome, accounting for 55.5% of the total toxin transcriptional output in C0136 and 53.2% in C0148 (Fig. 2). The structural integrity of these toxins relies on the formation of 3–4 disulfide bonds, and all of the NaTxs identified in the *C. hentzi* transcriptome contained the conserved cysteine residues necessary to form these bonds. These NaTxs are generally 60–76 amino-acid residues in length and molecular weights range from 6.9 to 8.5 kDa (Possani et al., 1999). All of the NaTxs identified contained a 17–23 amino-acid signal peptide. Scorpion NaTxs are currently classified into two main toxin types, the α-toxins and the β-toxins, determined by their binding sites and physiological properties (Wheeler et al., 1983; Possani et al., 1999). On the basis of sequence alignments in the NCBI nr database, 17 of the 36 identified NaTxs belonged to the αNaTx class with 35–85% sequence identity, and 19 belonged to the βNaTx class with 30–86% sequence identity (Table 1). Within the α and β designations, Possani et al. (1999) used more detailed sequence, structure, and function information to further classify scorpion NaTxs into ten groups. Our NaTxs were assigned to these groups on the basis of sequence identity using the nr database where possible (Table 1). The identified NaTx groups include group 2 (NaTx-1, a mammal-specific β-toxin), group 4 (NaTx-8, NaTx-9, NaTx-17, NaTx-22 and NaTx-30, which are insect-specific β-toxins that act as depressants), group 8 (NaTx-13, NaTx-14, NaTx-23 and NaTx-26, which are weakly-active α-toxins), and group 9 (NaTx-10, NaTx-11 and NaTx-16, which is a new-world β-toxin specific to insects and crustaceans). The αNaTxs bind to the extracellular surface at receptor site 3 and inhibit the fast inactivation process (Wheeler et al., 1983; Possani et al., 1999; Quintero-Hernández et al., 2013). All but three of the αNaTxs identified in the transcriptome were detected proteomically. The three that were not proteomically detected (NaTx-15, NaTx-33 and NaTx-35) were also not highly expressed in the venom-gland transcriptome, accounting for a total 1714.51 TPM in C0136 and 943.22 TPM in C0148. The βNaTxs bind to receptor site 4 and cause a negative shift in membrane potential (Wheeler et al., 1983; Possani et al., 1999; Quintero-Hernández et al., 2013). This group includes NaTx-19, which was the most abundant transcript identified in both C0136 and C0148 with 231,053.10 TPM and 171,091.23 TPM, respectively. Six of the 19 βNaTxs identified in the transcriptome were not detected proteomically (NaTx-8, NaTx-10, NaTx-16, NaTx-20, NaTx-21, and NaTx-32), nor were they highly expressed in the venom-gland transcriptome (Table 1).

The K^+-channel toxins (KTxs) were the second most abundant class of toxins identified in the transcriptome of *C. hentzi*, accounting for 230,321.05 TPM and 290,745.14 TPM in C0136 and C0148, respectively. Four families of KTxs are currently recognized: α-, β-, γ-, and κKTxs (Tytgat et al., 1999; de la Vega and Possani, 2004; Quintero-Hernández et al., 2013). We identified 20 αKTxs, eight γKTxs, and two βKTxs. We were unable to classify two additional KTxs (Table 1). The αKTx family is the largest of the four KTx families, and members are characterized by small 23–42 amino-acid peptides (3–5 kDa) that contain 3–4 stabilizing disulfide bridges (de la Vega and Possani, 2004). αKTxs have been shown to block both the pore and extracellularly interact with K^+-channels (de la Vega and Possani, 2004; Quintero-Hernández et al., 2013). Only four of the 20 αKTxs identified in the transcriptome of *C. hentzi* were detected proteomically (αKTx-1, αKTx-6, αKTx-8, and αKTx-17). Among those not detected in the proteome was αKTx-13, which has the highest TPM of any of the αKTxs identified in both C0136 (24,576.22 TPM) and C0148 (10,254.59 TPM). The lack of proteome

Table 1
Putative toxins identified in the venom-gland transcriptome of *Centruroides hentzi*.

Putative toxin	Signal peptide	Precursor (aa)	C0136 TPM	C0148 TPM	C0136 fmol	C0148 fmol	Notes
αKTx-1	Yes	59	4051.69	4074.36	108.88	159.79	–
αKTx-2	Yes	59	1395.61	757.18	–	–	–
αKTx-3	Yes	63	1678.85	2979.88	–	–	–
αKTx-4	Yes	62	1076.41	3.09	–	–	–
αKTx-5	Yes	64	1846.18	2337.65	–	–	–
αKTx-6	Yes	56	6186.68	6197.05	7.27	37.42	–
αKTx-7	Yes	62	9328.13	7966.19	–	–	–
αKTx-8	Yes	61	2687.81	1755.51	6.12	–	–
αKTx-9	Yes	62	4791.34	8417.96	–	–	–
αKTx-10	Yes	60	1880.24	3049.49	–	–	–
αKTx-11	Yes	62	360.28	–	–	–	–
αKTx-12	Yes	53	569.69	1008.62	–	–	–
αKTx-13	Yes	58	24576.22	10254.59	–	–	–
αKTx-14	Yes	65	514.58	406.28	–	–	–
αKTx-15	Yes	59	121.84	80.70	–	–	–
αKTx-16	Yes	60	143.04	54.57	–	–	–
αKTx-17	Yes	59	700.02	786.37	–	60.99	–
αKTx-18	Yes	62	18.35	5737.60	–	–	–
αKTx-19	Yes	62	10.91	8157.58	–	–	–
αKTx-20	Yes	59	–	243.81	–	–	–
AMP-1	Yes	74	4999.11	6793.67	–	–	NDBP-4
AMP-2	Yes	69	10482.71	8788.21	–	–	NDBP-4
AMP-3	Yes	73	10367.92	10968.54	–	–	NDBP-4
AMP-4	Yes	73	19161.08	19404.13	–	–	NDBP-4
AMP-5	Yes	80	16855.95	20349.41	–	–	NDBP-4
Chitinase	Yes	71	2916.99	2362.85	691.95	244.87	CBM-14 superfamily
CRISP-1	Yes	418	799.86	1219.24	–	–	SCP superfamily
CRISP-2	Yes	398	47.13	47.09	–	–	SCP superfamily
CRISP-3	Yes	399	939.04	398.88	151.00	–	SCP superfamily
Defensin	Yes	61	404.53	28.58	–	–	Defensin-2 superfamily
FK506	Yes	207	138.90	223.14	4.08	–	FKBP-C superfamily
γKTx-1	Yes	62	3178.53	2767.21	306.65	491.15	–
γKTx-2	Yes	62	1033.30	129.18	224.62	157.70	–
γKTx-3	Yes	62	444.41	644.46	49.12	–	–
γKTx-4	Yes	64	69.54	58.08	–	–	–
γKTx-5	Yes	69	576.42	337.54	–	–	–
γKTx-6	Yes	66	236.64	142.95	16.20	–	–
γKTx-7	Yes	63	220.63	364.67	–	–	–
γKTx-8	Yes	62	16.46	4673.16	–	–	–
GPR	Yes	330	39.49	9.99	158.68	–	–
Headcase	No	426	127.53	73.98	–	4.20	Headcase superfamily
HYAL	Yes	401	451.44	15.91	–	–	Glyco-hydro-56 superfamily
IGFBP-1	Yes	108	1056.34	899.68	16.30	24.39	–
IGFBP-2	Yes	254	142.45	76.15	–	–	–
IGFBP-3	Yes	109	1626.96	1208.00	–	–	–
KTx-1	Yes	65	28527.05	96.22	359.85	–	–
KTx-2	Yes	85	154.93	609.20	–	–	–
KTx-3	Yes	89	53546.75	85480.12	87.49	145.06	βKTx
KTx-4	Yes	92	80378.52	131173.87	721.16	704.76	βKTx
KUN-1	Yes	87	2513.47	3489.88	184.78	220.21	serine protease inhibitor
KUN-2	Yes	89	171.63	21.88	–	–	serine protease inhibitor
KUN-3	Yes	79	98.31	46.69	–	–	serine protease inhibitor
KUN-4	Yes	80	3672.13	5313.78	14.28	227.91	serine protease inhibitor
LAP-1	Yes	93	4679.41	6134.59	304.59	717.64	–
LAP-2	Yes	75	476.74	1090.43	–	–	–
MonoO	Yes	347	121.22	122.76	30.85	–	Cu^{2+} monooxegenase
MP-1	Yes	400	3246.92	677.78	–	–	ZnMc superfamily
MP-2	Yes	397	2629.40	740.06	476.72	22.85	ZnMc superfamily
MP-3	Yes	408	57.71	34.67	–	–	ZnMc superfamily
MP-4	Yes	406	22.97	9.19	–	–	ZnMc superfamily
MP-5	Yes	396	8771.40	3169.07	405.54	194.56	ZnMc superfamily
MP-6	Yes	400	1907.06	1771.85	260.62	595.99	ZnMc superfamily
NaTx-1	Yes	83	7210.41	7451.20	77.15	456.10	βNaTx, group 2
NaTx-2	Yes	102	6456.33	14477.04	12.12	898.64	αNaTx
NaTx-3	Yes	85	2988.65	4147.08	62.97	305.29	βNaTx
NaTx-4	Yes	83	7022.83	9341.90	207.92	694.13	αNaTx
NaTx-5	Yes	99	3353.39	5226.49	68.12	139.71	βNaTx
NaTx-6	Yes	98	3907.27	1778.63	118.91	–	αNaTx
NaTx-7	Yes	98	3089.04	10351.64	–	595.18	βNaTx
NaTx-8	Yes	100	1248.76	2170.33	–	–	βNaTx, group 4
NaTx-9	Yes	98	1465.59	2936.17	110.45	404.39	βNaTx, group 4
NaTx-10	Yes	104	4948.72	3715.95	–	–	βNaTx, group 9
NaTx-11	Yes	84	4393.83	8151.3	72.66	134.71	βNaTx, group 9

(*continued on next page*)

Table 1 (*continued*)

Putative toxin	Signal peptide	Precursor (aa)	C0136 TPM	C0148 TPM	C0136 fmol	C0148 fmol	Notes
NaTx-12	Yes	85	1117.24	7.86	83.28	–	αNaTx
NaTx-13	Yes	85	3466.39	21927.08	225.40	1843.18	αNaTx, group 8
NaTx-14	Yes	87	13443.07	20137.46	345.64	1482.54	αNaTx, group 8
NaTx-15	Yes	89	1367.61	344.94	–	–	αNaTx
NaTx-16	Yes	84	566.85	242.67	–	–	βNaTx, group 9
NaTx-17	Yes	85	15650.30	9707.42	313.77	574.26	βNaTx, group 4
NaTx-18	Yes	89	50270.54	50921.62	1563.98	833.31	αNaTx
NaTx-19	Yes	83	231053.10	171091.23	5663.20	4132.55	βNaTx
NaTx-20	Yes	85	196.56	–	–	–	βNaTx
NaTx-21	Yes	84	161.25	370.76	–	–	βNaTx
NaTx-22	Yes	83	218.12	6.70	33.68	–	βNaTx, group 4
NaTx-23	Yes	82	40865.21	52238.25	1780.03	3895.17	αNaTx, group 8
NaTx-24	Yes	88	5060.89	4906.03	417.59	719.61	αNaTx
NaTx-25	Yes	83	3412.17	4751.87	–	437.05	βNaTx
NaTx-26	Yes	93	370.13	1596.75	–	247.26	αNaTx, group 8
NaTx-27	Yes	93	5932.37	5727.14	421.53	775.29	αNaTx
NaTx-28	Yes	106	3588.63	4122.86	120.95	242.08	βNaTx
NaTx-29	Yes	88	4877.43	4766.65	115.25	319.45	αNaTx
NaTx-30	Yes	91	171.02	1008.84	–	220.77	βNaTx, group 4
NaTx-31	Yes	85	329.65	2375.00	–	775.41	βNaTx
NaTx-32	Yes	87	216.52	639.59	–	–	βNaTx
NaTx-33	Yes	97	230.61	341.36	–	–	αNaTx
NaTx-34	Yes	87	726.03	1158.86	–	359.95	αNaTx
NaTx-35	Yes	87	116.29	256.92	–	–	αNaTx
NaTx-36	Yes	90	20234.94	29457.79	256.57	1628.77	αNaTx
PLA2	Yes	232	34.65	3.43	–	–	PLA2 like superfamily
SP-1	Yes	265	3631.92	3933.22	188.38	165.12	Tryp SPc superfamily
SP-2	Yes	309	276.58	135.95	13.39	–	Tryp SPc superfamily
SP-4	Yes	369	60.65	21.82	–	–	Tryp SPc superfamily
Synapt25	No	202	191.83	79.15	–	6.30	SNARE/SNAP superfamily
Transferrin	Yes	712	90.51	49.04	–	6.30	Transferrin superfamily
VP-1	Yes	125	955.26	45.29	–	–	–
VP-4	Yes	111	6586.73	3537.74	226.57	308.52	–
VP-5	Yes	134	1577.43	69.00	–	–	SVWC superfamily
VP-6	Yes	82	8638.71	1318.89	29.21	–	–
VP-8	Yes	109	1068.88	129.09	–	–	IGFBP superfamily
VP-9	Yes	125	867.69	209.61	–	–	–
VP-10	Yes	81	1220.67	–	–	–	TIL superfamily
VP-12	Yes	90	459.44	53.42	–	–	TIL superfamily
VP-13	Yes	120	304.23	28.18	184.24	–	flagellin-C superfamily
VP-15	Yes	100	235.49	21.55	–	–	–
VP-16	No	84	377.43	708.00	–	–	–
VP-17	Yes	103	94.87	14.91	–	–	SVWC superfamily
VP-18	Yes	95	155.72	4.37	–	–	IGFBP superfamily
VP-19	Yes	109	65.56	0.29	–	–	–
VP-20	Yes	123	101.41	7.26	–	–	–
VP-21	Yes	123	3882.14	5022.16	102.76	97.28	–
VP-22	Yes	103	414.85	1187.35	–	–	SVWC superfamily
VP-23	Yes	81	–	84.10	–	–	TIL superfamily

Abbreviations: AMP—antimicrobial peptide, CRISP—cysteine-rich secretory protein, GPR—G-protein coupled receptor, HYAL—hyaluronidase, IGFBP—insulin like growth factor binding protein, KUN—kunitz-type protease inhibitor, KTx—K^+-channel toxin, LAP—lipolysis-activating peptide, MP—metalloproteinase, NaTx—Na^+-channel toxin, SP—serine proteinase, VP—venom protein.

detection was not surprising considering their small precursor size and known detectability challenges of these toxins in other scorpion venoms (Rokyta and Ward, 2017). The γKTxs have been shown to target hERG channels, which are associated with the cell cycle and proliferation of several cancer types (Quintero-Hernández et al., 2013), and are generally 40–43 amino-acids in length, weighing 4–5 kDa (de la Vega and Possani, 2004). All of the γKTxs identified in the transcriptome of *C. hentzi* contained eight cysteine residues, and all contained a 19–20 amino-acid signal peptide. βKTxs are between 60 and 65 amino-acids in length and approximately 7 kDa based on previously reported sequences (de la Vega and Possani, 2004) run through ExPASy ProtParam (Gasteiger et al., 2005).

3.3. Antimicrobial peptides

We identified five transcripts encoding proteins homologous to scorpion antimicrobial peptides (AMPs) in the venom-gland transcriptome of *C. hentzi*, which accounted for 61,866.77 TPM (7.6%) and 66,303.96 TPM (7.7%) of the total toxin transcription in C0136 and C0148, respectively (Fig. 2). None of the AMPs identified in the transcriptome were detected proteomically, possibly due to their high levels of post-transcriptional modification resulting in final peptide lengths of 13–78 amino-acids (Harrison et al., 2014). Scorpion venom AMPs can be divided into those with and without cysteine residues, which, when present, give rise to the formation of disulfide bridges. Cysteine-containing AMPs typically have 3 or 4 disulfide bridges, and some have shown to interact with Na^+ (Díaz et al., 2009) and K^+-channels (Bontems et al., 1991). Cysteine-containing AMPs have been identified in other buthids (Bontems et al., 1991; Díaz et al., 2009); this group of AMPs was not, however, detected in the transcriptome of *C. hentzi*. The five identified AMP transcripts did not contain cysteine residues and belong to the non-disulfide bridge peptides (NDBPs). NDBPs have been previously classified into six subfamilies based on pharmacological activity, peptide length, and structural similarity (Zeng et al., 2005).

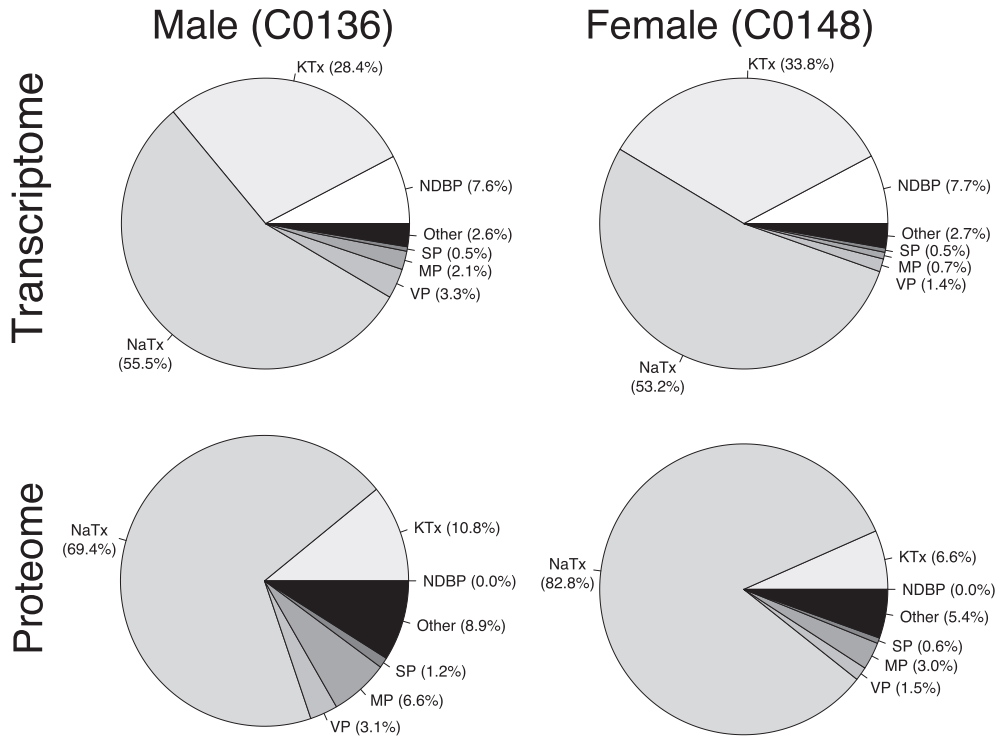

Fig. 2. Class-level abundance comparisons were similar across individuals in both venom-gland transcriptomes and venom proteomes, but transcriptome and proteome abundances did not agree well within individuals. Both KTxs and NDBPs show considerably less representation in the proteomes of each individual than would be predicted by their presence in the transcriptomes. Many KTxs and NDBPs are known to undergo extensive post-translational proteolytic processing and have previously shown detectability challenges in other scorpion venoms (see text Sections 3.2 and 3.3). Transcriptome abundances were based on transcripts per million (TPM) and percentages refer only to reads mapped to putative toxins (total toxin transcriptional output). Proteome abundances were expressed as molar percentages. Abbreviations: CRISP—cysteine-rich secretory protein, GPR—G-protein coupled receptor, HYAL—hyaluronidase, IGFBP—insulin like growth factor binding protein, KUN—kunitz-type protease inhibitor, KTx—K^+-channel toxin, LAP—lipolysis-activating peptide, MP—metalloproteinase, NaTx—Na^+-channel toxin, NDBP—non-disulfide bridge containing antimicrobial peptide, SP—serine proteinase, VP—venom protein.

We identified one of the six subfamilies in the transcriptome of *C. hentzi*, NDBP-4, and each representative contained a 21–22 amino-acid signal peptide. NDBP-4 AMPs are considered medium-length or intermediate chain AMPs, with final lengths of 18–29 amino-acids, weighing 2–4 kDa (Zeng et al., 2005; Harrison et al., 2014). This group is characterized by antimicrobial and hemolytic activity (Zeng et al., 2005). Four of the NBDP-4 transcripts identified in *C. hentzi* (AMP-1, AMP-2, AMP-3 and AMP-4) shared over 63% sequence identity with TsAP-2 from the Brazilian yellow scorpion (*Tityus serrulatus*) which shows high-potency against Gram-positive bacteria, high hemolytic activity, and inhibits growth of multiple human cancer cell lines (Guo et al., 2013). The fifth NBDP-4 transcript, AMP-5, shared 73% sequence identity to Androcin 18-1 from the black fat-tailed scorpion (*Androctonus bicolor*), but its function has not been well characterized (Zhang et al., 2015).

3.4. Proteases

We identified nine proteases in the transcriptome of *C. hentzi*, belonging to two main classes: the metalloproteases (MPs) and serine proteases (SPs; Table 1). Venom proteases are extremely active in posttranslational modifications of other venom toxins, and they can also exhibit their own toxic activity (Carmo et al., 2014; Serrano, 2013). Six of the nine proteases identified in *C. hentzi* were classified as metalloproteases (MPs), accounting for 2.1% (16,635.46 TPM) of the total toxin transcriptional output in C0136 and 0.7% (6402.62 TPM) in C0148 (Fig. 2). Three of the MPs detected were identified on the basis of homology alone (MP-1, MP-3, and MP-4), and the remaining three (MP-2, MP-5 and MP-6) were detected proteomically. All six of the MPs identified matched with 39–61% identity to similar metalloproteases, referred to as metalloserrulases (TsMs), identified in the venom-gland transcriptome and proteome of the Brazilian yellow scorpion (*T. serrulatus*; Carmo et al., 2014). The metalloserrulases classified by Carmo et al. (2014) in the venom of *T. serrulatus* belonged to two families: the metzincin family and the gluzincin family. In *T. serrulatus*, nine of the ten metalloserrulases were classified as being in the metzincin family, which is classified by three domains (signal peptide, propeptide and metalloprotease domain), eight conserved cysteine residues (suggesting that these proteins are stabilized by four disulfide bridges), and molecular weights between 22 and 28 kDa (Carmo et al., 2014). All six metalloserrulases identified in the *C. hentzi* venom-gland transcriptome met this criteria, suggesting they may also be classified as belonging to the metzincin family. Three serine proteases were identified in the venom-gland transcriptome of *C. hentzi*, two of which were also detected proteomically (SP-1 and SP-2). Serine proteases generally have higher molecular weights than most venom proteins, ranging in size from 26 to 67 kDa (Serrano and Maroun, 2005). All of the SPs identified are members of the TrypSpc superfamily and contain an 18–22 amino-acid signal peptide. The SPs only accounted for 0.5% (3969.15 TPM in C0136 and 4090.99 TPM in C0148) of the total toxin transcriptional output in each individual (Fig. 2).

3.5. Putative toxins with functionally characterized homologs

We identified four Kunitz-type protease inhibitors (KUNs), all of which belong to the KU superfamily of serine protease inhibitors. These can function as trypsin inhibitors as well as act in the blockage of K^+-channels (Chen et al., 2012; Santibáñez-López et al., 2017). Each of the four KUNs had an 18–20 amino-acid signal peptide, and three of the four (KUN-2, KUN-3 and KUN-4) contained trypsin interaction sites. This group accounts for 0.8% (6455.54 TPM) and 1.0% (8872.23 TPM) of the total toxin transcriptional output in C0136 and C0148, respectively. Two of the four KUNs (KUN-1 and KUN-4) were detected proteomically in both individuals. The strongest nr match to KUN-1 was a 41% identity to a Kunitz-type protease inhibitor identified in the genome of the sea anemone, *Aiptasia* (Baumgarten et al., 2015), and its exact function in scorpion venom is unknown. KUN-4 matched with a 56% identity to a serine protease inhibitor identified in the Chinese swimming scorpion (*Lychus mucronatus*), which shows complete inhibition of trypsin activity as well as inhibitory effects on Kv1.3, Kv1.2 and Kv1.1 K^+-channels (Ruiming et al., 2010; Chen et al., 2012). KUNs identified in other scorpion venoms have molecular weights ranging from 7 to 13 kDa (Bringans et al., 2008).

We identified three cysteine-rich secretory proteins (CRISP-1, CRISP-2 and CRISP-3), although only CRISP-3 was detected in the proteome of the male *C. hentzi*, C0136 (Table 2). All three CRISPs contained a signal peptide and an SCP domain, although we did not find homology between the three sequences in alignments. CRISP-3 exhibited 57% identity to a generic venom toxin identified in the venom of the Iranian scorpion (*Hemiscorpius lepturus*; Kazemi-Lomedasht et al., 2017). Three insulin-like growth factor binding protein (IGFBP) transcripts were detected (2825.75 TPM in C0136 and 2183.83 TPM in C0148), all of which contained signal peptides, but only one (IGFBP-1) was detected in the proteome. IGFBP-1 and IGFBP-2 each contained an IB domain, and IGFBP-2 also contained IG (immunoglobin) and KAZAL (Kazal type serine protease inhibitor) domains, suggesting that the IGFBPs may have KUN toxin-like activity.

We identified one chitinase that was proteomically confirmed in the venoms of both C0136 and C0148. This chitinase had a 16 amino-acid signal peptide and a chitin binding Peritrophin-A (CBM-14) domain. We identified one putative defensin with a 23 amino-acid signal peptide and an arthropod defensin domain (defensin-2). This defensin was 60% identical to defensin-1, identified in the transcriptome of the black fat-tailed scorpion (*Androctonus bicolor*; Zhang et al., 2015), although it was not detected in the proteome of *C. hentzi*. One venom hyaluronidase (HYAL) was identified, which contained a signal peptide and glycoside hydrolase family 56 domain, but it was not detected in the proteome of *C. hentzi*. We also identified one phospholipase A2 (PLA2), which contained a 16 amino-acid signal peptide and a PLA2-bee-venom-like domain. The PLA2 was one of the least abundant toxins detected in the *C. hentzi* transcriptome, with 34.65 TPM in C0136 and 3.43 TPM in C0148, and it was not confirmed proteomically.

Six low-abundance putative toxins identified had their closest database matches to nontoxic homologs. Of these six, three were proteomically detected in C0136 (FK506, GPR, and monoO) but not C0148, and the remaining three (headcase, Synapt25, and transferrin) were proteomically detected in C0148, but not C0136 (Table 2). The FK506 binding transcript contained a signal peptide and multiple domains including the FKBP-type peptidyl-prolyl cis-trans isomerase and EF-hand domain-pair domains. The G-protein coupled receptor (GPR) did contain a 16 amino-acid signal peptide, but did not contain any conserved structural domains. The headcase transcript contained both headcase protein and headcase protein homolog domains, but no signal peptide. The monooxegenase (monoO) had both N- and C-terminal copper type II ascorbate-dependent monooxygenase domains. The synaptosomal-associated protein transcript (Synapt25) contained soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor (SNARE), synaptosome-associated protein (SNAP), and SNARE helical region (tSNARE) domains. But no signal peptide. The

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

Protein	C0136			C0148			Average	
	rep 1	rep 2	rep 3	rep 1	rep 2	rep 3	C0136	C0148
αKTx-8	–	18.36	–	–	–	–	6.12	–
αKTx-17	–	–	–	100.90	–	82.07	–	60.99
CRISP-3	218.24	146.86	87.89	–	–	–	151.00	–
FK506	–	12.24	–	–	–	–	4.08	–
γKTx-3	72.75	55.07	19.53	–	–	–	49.12	–
γKTx-6	36.37	12.24	–	–	–	–	16.20	–
GPR	240.07	152.97	83.01	–	–	–	158.68	–
Headcase	–	–	–	–	12.61	–	–	4.20
KTx-1	436.49	379.38	263.68	–	–	–	359.85	–
MonoO	43.65	24.48	24.41	–	–	–	30.85	–
NaTx-6	174.59	104.02	78.12	–	–	–	118.91	–
NaTx-7	–	–	–	645.78	592.65	547.11	–	595.18
NaTx-12	87.30	79.55	83.01	–	–	–	83.28	–
NaTx-22	50.92	30.59	19.53	–	–	–	33.68	–
NaTx-25	–	–	–	484.33	460.25	366.56	–	437.05
NaTx-26	–	–	–	269.07	264.80	207.90	–	247.26
NaTx-30	–	–	–	255.62	220.66	186.02	–	220.77
NaTx-31	–	–	–	941.79	788.07	596.37	–	775.41
NaTx-34	–	–	–	430.52	359.37	289.96	–	359.95
SP-2	21.82	18.36	–	–	–	–	13.39	–
Synapt25	–	–	–	–	18.91	–	–	6.30
Transferrin	–	–	–	–	18.91	–	–	6.30
VP-6	50.92	36.71	–	–	–	–	29.21	–
VP-13	283.72	171.33	97.66	–	–	–	184.24	–

Quantities are given in fmol. Abbreviations: CRISP—cysteine-rich secretory protein, GPR—G-protein coupled receptor, KTx—K^+-channel toxin, NaTx—Na^+-channel toxin, rep—replicate, SP—serine proteinase, VP—venom protein.

transferrin transcript contained a signal peptide and a transferrin family of the type 2 periplasmic-binding protein superfamily domain.

3.6. Putative toxins without functionally characterized homologs

Two lipolysis-activating peptides (LAPs) were detected in the transcriptome, but only one of these (LAP-1) was detected in the proteome of *C. hentzi*. The top nr blast hit for LAP-1 was 41% identical to an LAP identified in the venom-gland transcriptome of *L. mucronatus*, which may be involved in the modulation of Na^+-channels and may also block K^+-channels. Its exact function is unknown (Ruiming et al., 2010).

We identified 18 proteins and peptides that we were unable to classify functionally and we therefore generically labeled as "venom proteins" (VPs). These unclassified toxins accounted for a total of 3.3% and 1.4% of the total putative toxin transcriptional output in C0136 and C0148, respectively (27,006.51 TPM in C0136 and 12,441.21 TPM in C0148). Of the 18 VPs, only four were detected proteomically in C0136 (VP-4, VP-6, VP-13 and VP-21), and two of these were also detected in the proteome of C0148 (VP-4 and VP-21; Table 1). All but one of the 18 VPs (VP-16) contained a signal peptide, and their precursor lengths ranged from 81 to 134 amino-acid residues. Among these putative toxins, we identified four groups with recognized superfamily domains. Group I (VP-5, VP-17, and VP-22) contained a single von Willebrand factor type C (SVWC) domain. Within this group, VP-17 and VP-22 shared the highest degree of sequence identity to each other, each with a 22 amino-acid signal peptide. Group II (VP-8 and VP-18) contained an insulin-like growth factor binding protein (IGFBP) domain. VP-18 was closer in sequence identity to the other identified IGFBP proteins in the venom-gland transcriptome of *C. hentzi* than was VP-8. Group III (VP-10, VP-12 and VP-23) contained a trypsin inhibitor-like cysteine rich (TIL) domain. Both VP-10 and VP-23 contained an 18 amino-acid signal peptide and shared higher sequence identity to each other than to VP-12, which contained a 21 amino-acid signal peptide. Group IV (VP-13) contained a flagellin domain and was detected in the proteome of C0136. The remaining nine VPs did not contain any characterized superfamily domain, and their closest homologous matches were to other unknown hypothetical or generic venom proteins.

3.7. Transcript and protein abundances across individuals

We found high correlation between the mRNA abundances of nontoxin-encoding proteins between the two individuals (Spearman's rank correlation $\rho = 0.92$, Pearson's rank correlation coefficient $R = 0.91$, and $R^2 = 0.83$; Fig. 3). The proteomically confirmed toxins were also well correlated (Spearman's rank correlation $\rho = 0.81$, Pearson's correlation coefficient $R = 0.81$, and $R^2 = 0.66$), but the homology-only toxins were much less consistent between the two individuals (Spearman's rank correlation $\rho = 0.54$, Pearson's correlation coefficient $R = 0.41$, and $R^2 = 0.17$). The results of the nontoxin-encoding mRNA expression levels indicate that any divergence between the two putative toxin classes is biological rather than technical, but whether these differences were due to individual variation or were sex-related (De Sousa et al., 2010; Miller et al., 2016; Rodríguez-Ravelo et al., 2015; Uribe et al., 2017) is unknown. The divergence in the proteome and homology-only toxin transcripts was mainly due to a small number of transcripts with unusually different expression levels between individuals. These transcripts, that fell on or beyond the 99th percentile of differences between the nontoxin measures of the two individuals, were considered outliers in the data sets (Fig. 3). Six of the 59 proteome toxin transcripts were outliers, most of which belonged to the Na^+-channel toxins, and were among those that were proteomically detected in one individual and completely absent from the other (Table 2). Fifteen of the 63 homology-only toxin transcripts were outliers and mostly belonged to the K^+-channel toxins and generic venom proteins (VPs).

The LC-MS/MS results showed fair agreement between the two individuals (Spearman's rank correlation $\rho = 0.54$, Pearson's correlation coefficient $R = 0.57$, and $R^2 = 0.33$), considering only 35 of the 59 proteomically detected toxins were confirmed in the venom of both C0136 and C0148 (Fig. 4). The most divergent toxins that were proteomically detected in both individuals were MP-2, which was abundant in C0136 and barely detected in C0148, and NaTx-2, which was abundant in C0148 and barely detected in C0136 (Table 1). The venom proteome of C0136 had 14 proteins that were not detected in C0148, and the venom proteome of C0148 had ten proteins that were not detected in C0136 (Table 2). Many of these presence/absence disagreements involved proteins that were detected at very low levels in one or the other individual (i.e., FK506 in C0136, and Headcase in C0148), however, a handful of putative toxins were present in fairly high levels in the proteome of one individual and completely absent from the other. The most abundant protein that was detected in the proteome of C0136 but not detected in the proteome of C0148 was KTx-1, with an average of 359.85 fmol between the three LC-MS/MS replicates (Table 2). KTx-1 was the tenth most abundant of the 49 toxins detected in the proteome of C0136. The most abundant protein that was detected in the venom proteome of C0148 but not C0136 was NaTx-31, with an average of 775.41 fmol between the three LC-MS/MS replicates. This protein was the eighth most abundant of the 45 toxins detected in the proteome of C0148.

Six of the transcripts that were detected as outliers in the transcript-abundance analysis (KTx-1, NaTx-12, NaTx-22, NaTx-30, NaTx-31, and VP-13, Fig. 3) also showed proteomic presence/absence differences between individuals (Table 2). This makes sense considering the extreme differences in transcript abundances between the two individuals for these toxins. For example, KTx-1, which was identified as an outlier in the transcript abundance analysis, has a TPM of 28,527.05 in C0136 and 96.22 TPM in C0148. Some of the proteins showing presence/absence differences in the proteome were not detected as outliers in the transcriptome. For example, NaTx-25 was detected in fairly high concentration (437.05 fmol) in the proteome of C0148 while completely absent in the proteome of C0136 (Table 2). In the transcriptome, the TPM values for NaTx-25 were similar across individuals with 3412.17 TPM in C0136 and 4751.87 TPM in C0148 (Table 1). NaTx-13 was detected in the proteome of both individuals (C0136: 225.40 fmol, C0148: 1843.18 fmol) and also detected as an outlier in the transcriptome (C0136: 3466.99 TPM, C0148: 21,927.08 TPM). Because the nontoxins were tightly correlated with just a few outliers, our results imply some level of biological variation between the two individuals, but, to be considered a significant outlier, the difference in mRNA (TPM) and protein expression (fmol) must be fairly extreme.

3.8. Transcript versus protein abundance estimates

Our comparisons of transcriptomic (Fig. 3) and proteomic (Fig. 4) abundances across individuals were positively correlated, and we found a similarly strong relationship when comparing transcript and protein abundance estimates, although the correlation was stronger in the female than the male (Fig. 5). For C0136 (male), we found $\rho = 0.48$, $R = 0.58$, and $R^2 = 0.31$. For C0148 (female), we found $\rho = 0.64$, $R = 0.77$, and $R^2 = 0.60$. Casewell et al. (2014) argued that major discrepancies between venom-gland transcriptomes and venom proteomes in snakes can be attributed

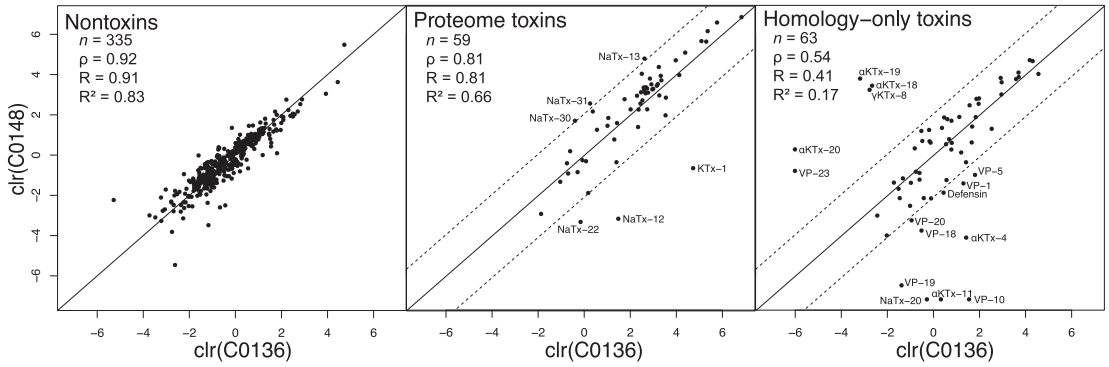

Fig. 3. A venom-gland transcript abundance comparison between a male (C0136) and female (C0148) *C. hentzi* generally showed strong agreement. Transcript levels were highly correlated between the venom-gland transcriptomes of the two individuals for the nontoxins and toxins confirmed proteomically, but less so for the putative toxins identified on the sole 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 line therefore represent toxins with unusually different expression levels relative to the nontoxins and are considered outliers. Abbreviations: clr—centered logratio transformation, *n*—number of transcripts, *ρ*—Spearman's rank correlation coefficient, *R*—Pearson's correlation coefficient, R^2—coefficient of determination, KTx—K^+-channel toxin, NaTx—Na^+-channel toxin, VP—venom protein (unknown function).

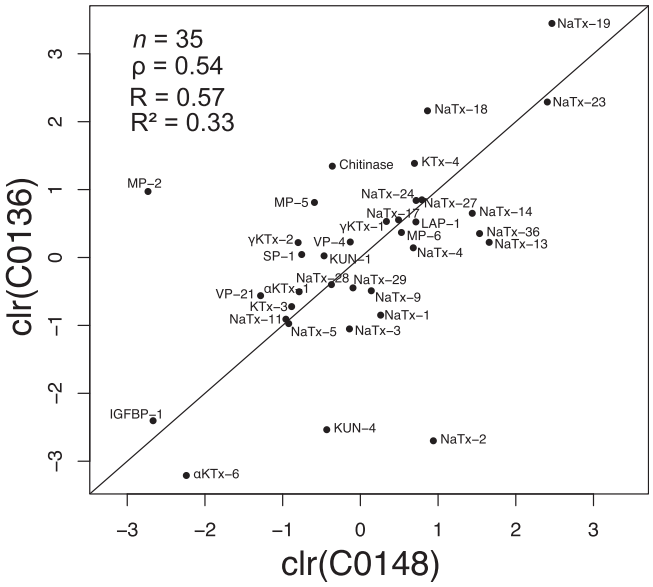

Fig. 4. A venom proteomic comparison between male (C0136) and female (C0148) *C. hentzi* showed fair agreement across individuals for proteins detected in both venom proteomes. Table 2 shows the proteomic presence/absence differences between the two individuals. Abbreviations: clr—centered logratio transformation, *n*—number of proteins, *ρ*—Spearman's rank correlation coefficient, *R*—Pearson's correlation coefficient, R^2—coefficient of determination, IGFBP—insulin like growth factor binding protein, KUN—kunitz-type protease inhibitor, KTx—K^+-channel toxin, MP—metalloproteinase, NaTx—Na^+-channel toxin, SP—serine proteinase, VP—venom protein (unknown function).

to post-transcriptional regulatory mechanisms such as microRNAs. However, Rokyta et al. (2015) argued that a failure to find agreement between different sets of measurements must be assumed to have been due to methodologial/technical issues or biases, unless those sources can be absolutely ruled out as potential causes of the disagreement. To determine whether the transcriptome/proteome discrepancy simply resulted from biases related to the sizes of the proteins as found by Rokyta and Ward (2017), we compared the differences between the protein and transcript abundance estimates to the length of their corresponding coding sequences (CDS length, Fig. 6). We found that C0136 had a significantly positive correlation ($p = 1.9 \times 10^{-3}$) and C0148 did not have a significant correlation ($p = 0.56$). The significantly positive correlation implies that, for smaller proteins, the protein abundance estimates were lower than expected based on the transcriptome, and, for larger proteins, the protein abundance estimates were higher than expected (Fig. 6). Unfortunately, this pattern was not uniform, so a correction for length would not eliminate all of the discrepancies between the transcriptome and proteome. In C0148, however, the lack of significance between the protein and transcript abundance estimates may explain why the protein abundance estimates were closer to expectations based on the transcriptome.

3.9. Protein bioanalyzer profiles

In the comparison of two separate venom samples for each individual (C0136: V0136 and V0188; C0148: V0263 and V0552) run on an Agilent protein bioanalyzer chip, we found strong agreement in biological replicate venom samples in the male (C0136), but not as strong agreement in the female (C0148, Fig. 7). The Bioanalyzer reported total relative sample concentration for each sample as: V0136: 977.9 ng/μl, V0188: 5908.1 ng/μl, V0263: 2086.1 ng/μl, V0552: 1834.4 ng/μl. Though the two venom samples run for C0136 appear different, these differences can be attributed to the substantial difference in concentration between the two samples, but the overall profile shape and relative peak abundances are nearly identical in both the profile (Fig. 7, A) and side-by-side gel image (Fig. 7, C). The two venom samples for C0148 were much closer in concentration, and differences in relative abundances can still be seen in both the profile (Fig. 7, B) and side-by-side gel images (Fig. 7, C). The differences observed between the two C0148 samples could be a result of the timing of venom extractions. Venom samples for C0148 were taken several months apart (V0263 in June and V0552 in September), whereas the samples for C0136 were take only one month apart (V0136 in May and V0188 in June), indicating possible phenotypic plasticity in venom expression (Gangur et al., 2017). We also see a greater relative abundance in the larger proteins (26–46 kDa) compared to the smaller proteins (6.5–15 kDa) in the male (C0136) versus the female (C0148). These expression differences are in agreement with MPs (22–67 kDa) being both more highly expressed in the transcriptome and detected at higher concentrations in the proteome of C0136 relative to C0148 (Table 1). All venom samples analyzed on the protein bioanalyzer chips showed much higher relative abundances of proteins in the 6.5–15 kDa range, and these sizes are consistent with Na^+-channel toxins, ranging from 6.9 to 8.5 kDa (Possani et al., 1999). We did not expect to see representation of K^+-channel toxins, which are

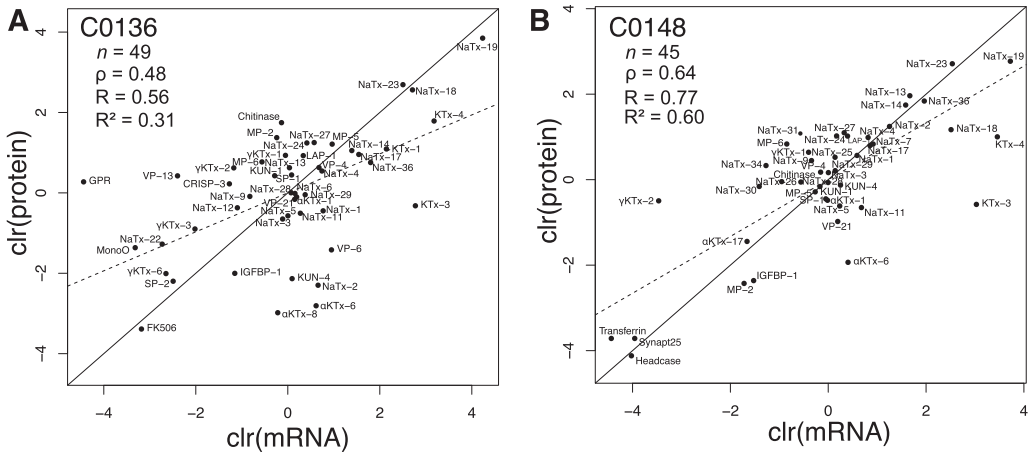

Fig. 5. Transcript and protein abundances were positively correlated, but the agreement across data sets was weaker in the male (A) compared the female (B). Abbreviations: clr—centered logratio transformation, *n*—number of transcripts, *ρ*—Spearman's rank correlation coefficient, *R*—Pearson's correlation coefficient, R^2—coefficient of determination, CRISP—cysteine-rich secretory protein, GPR—G-protein coupled receptor, HYAL—hyaluronidase, IGFBP—insulin like growth factor binding protein, KUN—kunitz-type protease inhibitor, KTx—K^+-channel toxin, MP—metalloproteinase, NaTx—Na^+-channel toxin, SP—serine proteinase, VP—venom protein (unknown function).

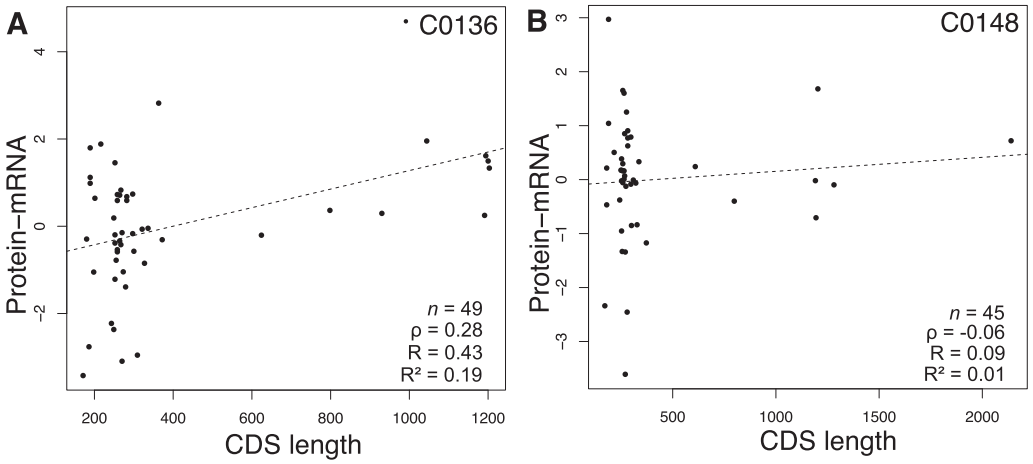

Fig. 6. We found a significant positive correlation between the discrepancy between protein and transcript abundance estimates and the sizes of the proteins in the male (C0136: $p = 1.9\times10^{-3}$) but not the female (C0148: $p = .56$). Relative to transcript abundance estimates, smaller proteins tended to have lower-than-expected protein abundances and larger proteins tended to have higher-than-expected protein abundances in the male (A), while protein expression in the female (B) met abundance expectations based on the transcriptome. Abbreviations: CDS—coding sequence, *n*—number of transcripts, *ρ*—Spearman's rank correlation coefficient, *R*—Pearson's correlation coefficient, R^2—coefficient of determination.

reported to be 4–5 kDa (de la Vega and Possani, 2004), or antimicrobial peptides, which are reported to be 2–4 kDa (Zeng et al., 2005; Harrison et al., 2014), as these fall below the range of the assay (5–80 kDa) or may be overpowered by the internal markers (1.6 kDa and 95 kDa) or system peak (3.5 kDa).

3.10. Comparison to other scorpion species

Species of *Centruroides* are known for producing venoms rich in Na^+ and K^+-channel toxins (Simard et al., 1992; Possani et al., 1999, 2000; Corona et al., 2001, 2002; Nastainczyk et al., 2002; de la Vega and Possani, 2004, 2005; Tan et al., 2006), and the venom-gland transcriptome and proteome of *C. hentzi* were consistent with this expectation (Fig. 2 and Table 1). These ion-channel toxins are primarily responsible for inflicting pain and neurotoxic physiological responses that often require medical attention (Rowe and Rowe, 2008; Quintero-Hernández et al., 2013). The Arizona bark scorpion, *C. sculpturatus*, has gained attention for its Na^+-channel toxins and their coevolutionary relationship with grasshopper mice (*Onychomys* spp.), which are known predators of the deadly scorpion (Rowe and Rowe, 2008; Rowe et al., 2013). Partial venom characterization in *C. tecomas* (Valdez-Velazquéz et al., 2013, 2016), *C. noxius* (Rendón-Anaya et al., 2012), and *Centruroides suffusus suffusus* (Martin et al., 1987; Espino-Solis et al., 2011) also revealed a plethora of Na^+ and K^+-channel toxins, as well as other proteins and peptides in the same families as those found in *C. hentzi* (i.e., AMPs, MPs, and venom proteins with unknown function). *Centruroides sculpturatus*, *C. noxius*, *C. tecomas*, and *C. s. suffusus* have been reported to cause extremely painful and sometimes deadly reactions in human encounters (Rendón-Anaya et al., 2012; Skolnik and Ewald, 2013; Valdez-Velazquéz et al., 2013; Strommen and Shirazi, 2015; Vega and Lia, 1966), yet their close relatives, *C. hentzi* and *C. vittatus*, cause only slight pain and discomfort (More et al., 2004; Stevenson et al., 2012). In the venom-gland transcriptome and proteome of *C. hentzi*, the ion-channel toxins were by far the most prominent toxins identified, including a mammal-specific Na^+-channel toxin (NaTx-1) and several other NaTxs that may be active toward mammals, insects, or both. The contrast of

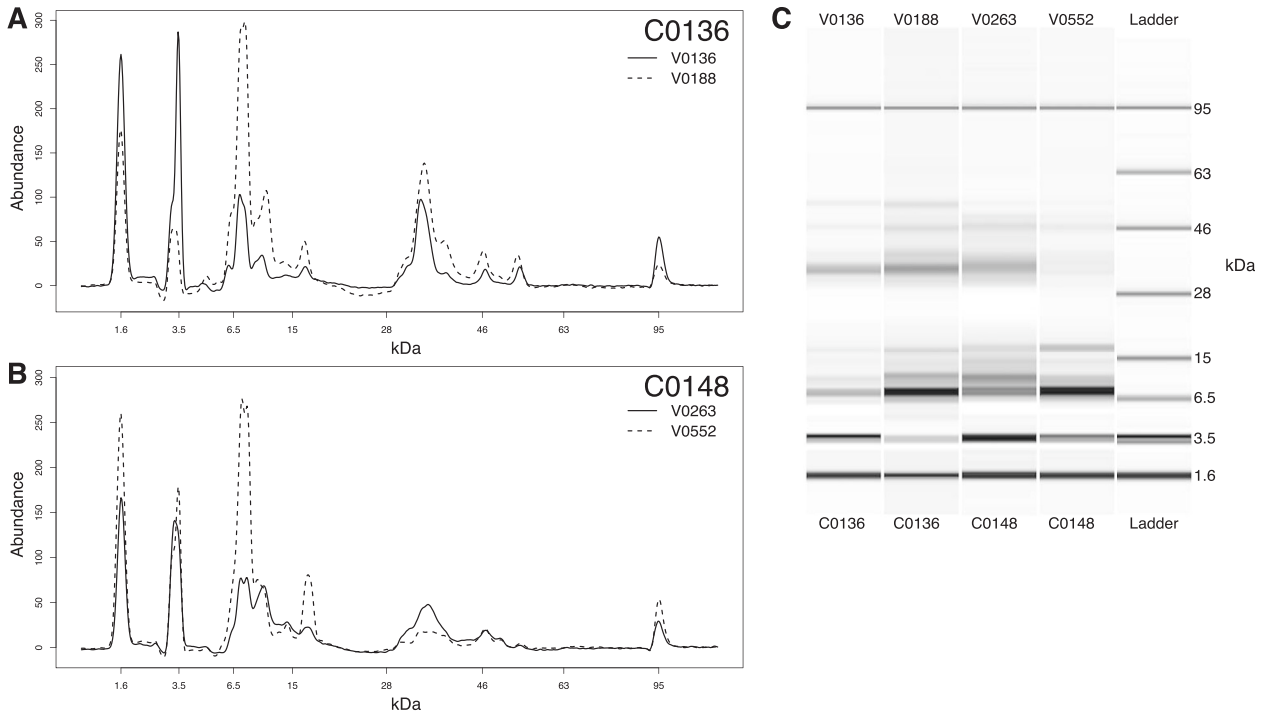

Fig. 7. Venom protein profiles using the Agilent Protein 80 bioanalyzer assay revealed strong agreement between biological replicates in the male (A) but not the female (B) *C. hentzi*. Side-by-side gel comparisons for each sample (C) also illustrate the similarities and differences in biological replicates. Total sample concentration reported by bioanalyzer for each sample was: V0136: 977.9 ng/μl, V0188: 5908.1 ng/μl, V0263: 2086.1 ng/μl, V0552: 1834.4 ng/μl. Internal markers and system peaks are represented at 1.6, 3.5 and 95 kDa in each profile. Peaks that fall in the 6.5—15 kDa range are likely the highly expressed Na^+-channel toxins, peaks that fall in the 28—46 kDa range are likely metalloproteases, and those that are above 46 kDa are likely serine proteases.

physiological symptoms elicited by different *Centruroides* species is perplexing considering the similarities in overall venom composition. In performing blastp analysis in the NCBI nr database, all but three of the ion-channel toxins identified in *C. hentzi* shared some percentage of sequence identity to a more harmful member of the *Centruroides* genus, though the majority of these fell below 70% sequence identity. The mammal specific NaTx-1 matched with 86% sequence identity to the well-characterized beta-neurotoxin, CssIX from *C. s. suffusus*, which has been shown to be lethal at low doses in mice (Espino-Solis et al., 2011). Comparative analyses of less-harmful venom and those with higher toxicity could potentially reveal the specific toxins, post-translational modifications, or sequence differences responsible for causing the wide discrepancy in physiological symptoms, as well as provide additional insight into the complex evolutionary dynamics between predator and prey.

Ion-channel toxins were also the most abundant and diverse group of toxins identified in the venom of the black-back scorpion, *Hadrurus spadix* (Rokyta and Ward, 2017), a member of the Caraboctonidae family. The majority of the ion-channel toxins identified in *H. spadix* belonged to the αKTx family. In contrast to *C. hentzi* and other members of the Buthidae family that exhibit a high abundance of Na^+-channel toxins, no Na^+-channel toxins were found in the venom-gland transcriptome or proteome of *H. spadix*. The *H. spadix* venom-gland transcriptome also contained a large number of AMPs, which were much more abundant and diverse than those found in *C. hentzi*. The AMPs identified in *H. spadix* included several cysteine-containing AMPs, which were not identified in the venom-gland transcriptome or proteome of *C. hentzi*. Although over three times as many generic venom proteins (VPs) with unknown homologs were present in *H. spadix* compared to *C. hentzi*, the large number of VPs present in both venoms demonstrates the need to fully characterize venoms from a broader range of scorpion species and families than are currently present in the literature.

4. Conclusions

We found 59 proteomically confirmed toxins in the venom of the Hentz striped scorpion, *Centruroides hentzi*, along with 63 toxin transcripts that were identified on the basis of homology to known toxins in other species. The transcriptome was rich in Na^+ and K^+-channel toxins, and we also identified a handful of AMPs, SPs, MPs and KUNs. In this species alone, 18 putative peptides and proteins were unable to be classified by homology, representing a realm of unexplored toxin diversity in scorpion venoms. Of the 59 proteomically confirmed toxins, only 35 were detected in the venoms of both the male and female individual. The remaining 24 proteome toxins were present in one individual and completely absent from the other, implying a high level of expression variation between individuals. Sex-based variation in venom has been reported in other scorpion species (De Sousa et al., 2010; Miller et al., 2016; Rodríguez-Ravelo et al., 2015; Uribe et al., 2017), though whether the variation found in *C. hentzi* can be attributed to sex is unclear with a sample size of one individual per sex. We found better agreement in transcriptome and proteome expression in the female compared to the male, suggesting more proteins may undergo post-translational modification in the male venom, though these differences in agreement could also be an artifact of technical biases in our approach. We also reported scorpion venom bioanalyzer profiles using the Agilent Protein 80 bioanalyzer assay and found agreement between these profiles in comparison to transcriptome and proteome abundances. Using this data, we were able to broadly compare the venom components identified in the venom-gland of

C. hentzi with other members of the *Centruroides* genus and with the characterized venom of the black-back scorpion, *H. spadix*. We found that the harmless venom of *C. hentzi* shares many homologous toxins with its lethal relatives, despite the stark contrast of physiological symptoms between them. In comparison to *H. spadix*, we found that members of the *Centruroides* family, including *C. hentzi*, have venoms rich in both Na^+ and K^+-channel toxins, while the venom of *H. spadix* completely lacked Na^+-channel toxins. The venom variation we observe between individuals within species, genera, and families of scorpions, illustrates the importance of completing full scorpion venom characterizations using high-throughput transcriptomic and proteomic methods, and provides evidence that even less-harmful scorpions might also be a rich source of medically relevant components. As more venom-characterization and functional data become available, detailed comparisons of scorpion venom composition should be made to identify lethal and non-lethal components and further address the evolutionary question of why some closely-related species are harmful to humans, while others are not.

Ethical statement

Reporting standards: The authors declare that our manuscript describes original research and every effort was made to ensure the accuracy of the results and the account.

Data Access and Retention: The authors will make the raw data available as needed for review, and the data described has been deposited in the NCBI SRA, NCBI TSA, and ProteomeXchange databases and is scheduled to become available to the public upon publication. The data will be retained among the authors indefinitely.

Originality and Plagiarism: The authors declare that our manuscript is an original work with proper citations as needed.

Multiple, Redundant or Concurrent Publication: The authors declare that the data and work described in our manuscript has not and will not be submitted for consideration to another journal.

Acknowledgment of Sources: The authors have provided proper acknowledgment of sources to the best of their abilities.

Authorship of the Paper: The three authors of the manuscript all made significant contributions to the reported study, and no one making a significant contribution was excluded as an author. All co-authors have seen and read the submitted version of the manuscript and all have approved submission.

Hazards and Human or Animal Subjects: The described study did not involve the use of vertebrate animals.

Disclosure and Conflicts of Interest: The authors declare no conflicts of interest.

Fundamental errors in published works: If a fundamental error or inaccuracy is discovered in the results described in the manuscript, the authors will immediately notify the editor or publisher.

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 (CRC PG-036698). We thank Rakesh Singh of the Florida State University College of Medicine Translational Science Laboratory for advice and assistance with proteomic analyses. We also thank Pierson Hill for his assistance in collecting specimens, and Michael Hogan for his assistance in photographing them.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.toxicon.2017.12.042.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.
Alvarenga, É.R., Mendes, T.M., Magalhães, B.F., Siqueira, F.F., Dantas, A.E., Barroca, T.M., Horta, C.C., Kalapothakis, E., 2012. Transcriptome analysis of the *Tityus serrulatus* scorpion venom gland. Open J. Genet. 2, 210–220.
Baumgarten, S., Simakov, O., Esherick, L.Y., Liew, Y.J., Lehnert, E.M., Michell, C.T., Li, Y., Hambleton, E.A., Guse, A., Oates, M.E., et al., 2015. The genome of *Aiptasia*, a sea anemone model for coral symbiosis. Proc. Natl. Acad. Sci. Unit. States Am. 112, 11893–11898.
Bontems, F., Roumestand, C., Gilquin, B., Menez, A., Toma, F., 1991. Refined structure of charybdotoxin: common motifs in scorpion toxins and insect defensins. Science 254, 1521–1523.
Bringans, S., Eriksen, S., Kendrick, T., Gopalakrishnakone, P., Livk, A., Lock, R., Lipscombe, R., 2008. Proteomic analysis of the venom of *Heterometrus longimanus* (Asian black scorpion). Proteomics 8, 1081–1096.
Carmo, A., Oliveira-Mendes, B., Horta, C., Magalhães, B., Dantas, A., Chaves, L., Chavez-Olortegui, C., Kalapothakis, E., 2014. Molecular and functional characterization of metalloserrulases, new metalloproteases from the *Tityus serrulatus* venom gland. Toxicon 90, 45–55.
Casewell, N.R., Wagstaff, S.C., Wüster, W., Cook, D.A.N., Bolton, F.M.S., King, S.I., Pla, D., Calvete, J.J., Harrison, R.A., 2014. Medically important differences in snake venom composition are dictated by distinct postgenomic mechanisms. Proceed. Natnl. Acad. Sci. USA 111, 9205–9210.
Chen, Z.-Y., Hu, Y.-T., Yang, W.-S., He, Y.-W., Feng, J., Wang, B., Zhao, R.-M., Ding, J.-P., Cao, Z.-J., Li, W.-X., et al., 2012. Hg1, novel peptide inhibitor specific for Kv1. 3 channels from first scorpion Kunitz-type potassium channel toxin family. J. Biol. Chem. 287, 13813–13821.
Chippaux, J.-P., Goyffon, M., 2008. Epidemiology of scorpionism: a global appraisal. Acta Trop. 107, 71–79.
Corona, M., Valdez-Cruz, N., Merino, E., Zurita, M., Possani, L., 2001. Genes and peptides from the scorpion *Centruroides sculpturatus* (Ewing), that recognize Na^+-channels. Toxicon 39, 1893–1898.
Corona, M., Gurrola, G.B., Merino, E., Cassulini, R.R., Valdez-Cruz, N.A., García, B., Ramírez-Domínguez, M.E., Coronas, F.I., Zamudio, F.Z., Wanke, E., et al., 2002. A large number of novel Ergtoxin-like genes and ERG K^+-channels blocking peptides from scorpions of the genus *Centruroides*. FEBS Lett. 532, 121–126.
Díaz, P., D'suze, G., Salazar, V., Sevcik, C., Shannon, J.D., Sherman, N.E., Fox, J.W., 2009. Antibacterial activity of six novel peptides from *Tityus discrepans* scorpion venom. A fluorescent probe study of microbial membrane Na^+ permeability changes. Toxicon 54, 802–817.
Diego-García, E., Peigneur, S., Clynen, E., Marien, T., Czech, L., Schoofs, L., Tytgat, J., 2012. Molecular diversity of the telson and venom components from *Pandinus cavimanus* (Scorpionidae Latreille 1802): transcriptome, venomics and function. Proteomics 12, 313–328.
Diego-García, E., Caliskan, F., Tytgat, J., 2014. The Mediterranean scorpion *Mesobuthus gibbosus* (Scorpiones, Buthidae): transcriptome analysis and organization of the genome encoding chlorotoxin-like peptides. BMC Genom. 15, 295.
Espino-Solis, G.P., Estrada, G., Olamendi-Portugal, T., Villegas, E., Zamudio, F., Cestele, S., Possani, L.D., Corzo, G., 2011. Isolation and molecular cloning of beta-neurotoxins from the venom of the scorpion *Centruroides suffusus suffusus*. Toxicon 57, 739–746.
Fet, V., Gantenbein, B., Gromov, A., Lowe, G., Lourenço, W.R., 2003. The first molecular phylogeny of Buthidae (Scorpiones). Euscorpius 2003, 1–10.
Gangur, A.N., Smout, M., Liddell, M.J., Seymour, J.E., Wilson, D., Northfield, T.D., 2017. Changes in predator exposure, but not in diet, induce phenotypic plasticity in scorpion venom. Proc. R. Soc. B 284, 20171364. The Royal Society.
Garcia-Calvo, M., Leonard, R., Novick, J., Stevens, S., Schmalhofer, W., Kaczorowski, G., Garcia, M., 1993. Purification, characterization, and biosynthesis of margatoxin, a component of *Centruroides margaritatus* venom that selectively inhibits voltage-dependent potassium channels. J. Biol. Chem. 268, 18866–18874.
Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein Identification and Analysis Tools on the ExPASy Server. Springer.
Guo, X., Ma, C., Du, Q., Wei, R., Wang, L., Zhou, M., Chen, T., Shaw, C., 2013. Two peptides, TsAP-1 and TsAP-2, from the venom of the Brazilian yellow scorpion, *Tityus serrulatus*: evaluation of their antimicrobial and anticancer activities. Biochimie 95, 1784–1794.
Harrison, P.L., Abdel-Rahman, M.A., Miller, K., Strong, P.N., 2014. Antimicrobial peptides from scorpion venoms. Toxicon 88, 115–137.
He, Y., Zhao, R., Di, Z., Li, Z., Xu, X., Hong, W., Wu, Y., Zhao, H., Li, W., Cao, Z., 2013. Molecular diversity of the Chaerilidae venom peptides reveals the dynamic evolution of scorpion venom components from Buthidae to non-Buthidae. J. Proteom. 89, 1–14.
Kang, A.M., Brooks, D.E., 2017. Nationwide scorpion exposures reported to US poison control centers from 2005 to 2015. J. Med. Toxicol. 13, 158–165.

Kazemi-Lomedasht, F., Khalaj, V., Bagheri, K.P., Behdani, M., Shahbazzadeh, D., 2017. The first report on transcriptome analysis of the venom gland of Iranian scorpion, *Hemiscorpius lepturus*. Toxicon 125, 123–130.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359.
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M., Cheung, V.G., 2011. Widespread RNA and DNA sequence differences in the human transcriptome. Science 333, 53–58.
Luna-Ramírez, K., Quintero-Hernández, V., Juárez-González, V.R., Possani, L.D., 2015. Whole transcriptome of the venom gland from *Urodacus yaschenkoi* scorpion. PLos One 10, e0127883.
Ma, Y., Zhao, R., He, Y., Li, S., Liu, J., Wu, Y., Cao, Z., Li, W., 2009. Transcriptome analysis of the venom gland of the scorpion *Scorpiops jendeki*: implication for the evolution of the scorpion venom arsenal. BMC Genom. 10, 290.
Ma, Y., Zhao, Y., Zhao, R., Zhang, W., He, Y., Wu, Y., Cao, Z., Guo, L., Li, W., 2010. Molecular diversity of toxic components from the scorpion *Heterometrus petersii* venom revealed by proteomic and transcriptomic analysis. Proteomics 10, 2471–2485.
Ma, Y., He, Y., Zhao, R., Wu, Y., Li, W., Cao, Z., 2012. Extreme diversity of scorpion venom peptides and proteins revealed by transcriptomic analysis: implication for proteome evolution of scorpion venom arsenal. J. Proteom. 75, 1563–1576.
Martin, M., Perez, L. G. y, El Ayeb, M., Kopeyan, C., Bechis, G., Jover, E., Rochat, H., 1987. Purification and chemical and biological characterizations of seven toxins from the mexican scorpion, *Centruroides suffusus suffusus*. J. Biol. Chem. 262, 4452–4459.
Mille, B.G., Peigneur, S., Diego-García, E., Predel, R., Tytga, J., 2014. Partial transcriptomic profiling of toxins from the venom gland of the scorpion *Parabuthus stridulus*. Toxicon 83, 75–83.
Miller, D.W., Jones, A.D., Goldston, J.S., Rowe, M.P., Rowe, A.H., 2016. Sex differences in defensive behavior and venom of the striped bark scorpion *Centruroides vitattus* (Scorpiones: Buthidae). Integr. Comp. Biol. 56, 1022–1031.
More, D., Nugent, J., Hagan, L., Demain, J., Schwertner, H., Whisman, B., Freeman, T., 2004. Identification of allergens in the venom of the common striped scorpion. Ann. Allergy, Asthma Immunol. 93, 493–498.
Nastainczyk, W., Meves, H., Watt, D., 2002. A short-chain peptide toxin isolated from *Centruroides sculpturatus* scorpion venom inhibits ether-à-go-go-related gene K^+ channels. Toxicon 40, 1053–1058.
de Oliveira, U.C., Candido, D.M., Dorce, V.A.C., de Loiola Meirelles Junqueira-de Azevedo, I., 2015. The transcriptome recipe for the venom cocktail of *Tityus bahiensis* scorpion. Toxicon 95, 52–61.
Ortiz, E., Gurrola, G.B., Schwartz, E.F., Possani, L.D., 2015. Scorpion venom components as potential candidates for drug development. Toxicon 93, 125–135.
Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8, 785–786.
Pipelzadeh, M.H., Jalali, A., Taraz, M., Pourabbas, R., Zaremirakabadi, A., 2007. An epidemiological and a clinical study on scorpionism by the Iranian scorpion *Hemiscorpius lepturus*. Toxicon 50, 984–992.
Polis, G., Sissom, W., 1990. Life history. Biol. Scorp. 161–223.
Possani, L., Martin, B., Svendsen, I., Rode, G., Erickson, B., 1985. Scorpion toxins from *Centruroides noxius* and *Tityus serrulatus*. primary structures and sequence comparison by metric analysis. Biochem. J. 229, 739–750.
Possani, L.D., Becerril, B., Delepierre, M., Tytgat, J., 1999. Scorpion toxins specific for Na^+-channels. FEBS J. 264, 287–300.
Possani, L.D., Merino, E., Corona, M., Bolivar, F., Becerril, B., 2000. Peptides and genes coding for scorpion toxins that affect ion-channels. Biochimie 82, 861–868.
Quintero-Hernández, V., Jiménez-Vargas, J.M., Gurrola, G.B., Valdivia, H.H., Possani, L.D., 2013. Scorpion venom components that affect ion-channels function. Toxicon 76, 328–342.
Quintero-Hernández, V., Ramírez-Carreto, S., Romero-Gutiérrez, M.T., Valdez-Velázquez, L.L., Becerril, B., Possani, L.D., Ortiz, E., 2015. Transcriptome analysis of scorpion species belonging to the *Vaejovis* genus. PLos One 10, e0117188.
Rendón-Anaya, M., Delaye, L., Possani, L.D., Herrera-Estrella, A., 2012. Global transcriptome analysis of the scorpion *Centruroides noxius*: new toxin families and evolutionary insights from an ancestral scorpion species. PLos One 7, e43331.
Rodríguez-Ravelo, R., Batista, C.V., Coronas, F.I., Zamudio, F.Z., Hernández-Orihuela, L., Espinosa-López, G., Ruiz-Urquiola, A., Possani, L.D., 2015. Comparative proteomic analysis of male and female venoms from the cuban scorpion *Rhopalurus junceus*. Toxicon 107, 327–334.
Rokyta, D.R., Ward, M.J., 2017. Venom-gland transcriptomics and venom proteomics of the blackback scorpion (*Hadrurus spadix*) reveal detectability challenges and an unexplored realm of animal toxin diversity. Toxicon 128, 23–37.
Rokyta, D.R., Lemmon, A.R., Margres, M.J., Aronow, K., 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genom. 13, 312.
Rokyta, D.R., Margres, M.J., Calvin, K., 2015. Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. Genes|Genomes|Genetics 5, 2375–2382.
Rowe, A.H., Rowe, M.P., 2008. Physiological resistance of grasshopper mice (*Onychomys* spp.) to Arizona bark scorpion (*Centruroides exilicauda*) venom. Toxicon 52, 597–605.
Rowe, A.H., Xiao, Y., Rowe, M.P., Cummins, T.R., Zakon, H.H., 2013. Voltage-gated sodium channel in grasshopper mice defends against bark scorpion toxin. Science 342, 441–446.
Ruiming, Z., Yibao, M., Yawen, H., Zhiyong, D., Yingliang, W., Zhijian, C., Wenxin, L., 2010. Comparative venom gland transcriptome analysis of the scorpion *Lychas mucronatus* reveals intraspecific toxic gene diversity and new venomous components. BMC Genom. 11, 452.
Santibáñez-López, C.E., Cid-Uribe, J.I., Batista, C.V., Ortiz, E., Possani, L.D., 2016. Venom gland transcriptomic and proteomic analyses of the enigmatic scorpion *Superstitionia donensis* (Scorpiones: Superstitioniidae), with insights on the evolution of its venom components. Toxins 8, 367.
Santibáñez-López, C.E., Cid-Uribe, J.I., Zamudio, F.Z., Batista, C.V., Ortiz, E., Possani, L.D., 2017. Venom gland transcriptomic and venom proteomic analyses of the scorpion *Megacormus gertschi* Díaz-Najera, 1966 (Scorpiones: Euscorpiidae: Megacorminae). Toxicon 133, 95–109.
Schwartz, E.F., Diego-Garcia, E., de la Vega, R.C.R., Possani, L.D., 2007. Transcriptome analysis of the venom gland of the Mexican scorpion *Hadrurus gertschi* (Arachnida: Scorpiones). BMC Genom. 8, 119.
Serrano, S.M., 2013. The long road of research on snake venom serine proteinases. Toxicon 62, 19–26.
Serrano, S.M., Maroun, R.C., 2005. Snake venom serine proteinases: sequence homology vs. substrate specificity, a paradox to be solved. Toxicon 45, 1115–1132.
Shelley, R.M., Sissom, W.D., 1995. Distributions of the scorpions *Centruroides vittatus* (Say) and *Centruroides hentzi* (Banks) in the United States and Mexico (Scorpiones, Buthidae). J. Arachnol. 100–110.
Simard, J., Meves, H., Watt, D., 1992. Neurotoxins in Venom from the North American Scorpion, Centruroides Sculpturatus (Ewing). Natural Toxins: Toxicology, Chemistry and Safety Alaken, Inc., Fort Collins, CO, pp. 236–263.
Skolnik, A.B., Ewald, M.B., 2013. Pediatric scorpion envenomation in the United States: morbidity, mortality, and therapeutic innovations. Pediatr. Emerg. Care 29, 98–103.
Soleglad, M.E., Fet, V., 2003. High-level systematics and phylogeny of the extant scorpions (Scorpiones: Orthosterni). Euscorpius 2003, 1–56.
De Sousa, L., Borges, A., Vásquez-Suárez, A., den Camp, H.J.O., Chadee-Burgos, R.I., Romero-Bellorín, M., Espinoza, J., De Sousa-Insana, L., Pino-García, O., 2010. Differences in venom toxicity and antigenicity between females and males *Tityus nororientalis* (Buthidae) scorpions. J. Venom Res. 1, 61.
Stahnke, H., Calos, M., 1977. A key to the species of the genus *Centruroides marx* (Scorpionida: Buthidae). Clave para las especies del género *Centruroides marx* (Scorpionida: Buthidae). Entomol. News 88, 111–120.
Stevenson, D.J., Greer, G., Elliott, M.J., 2012. The Distribution and Habitat of Centruroides Hentzi, vol. 11. Southeastern Naturalist.
Stockmann, R., Ythier, E., 2010. Scorpions of the World. NAP Editions.
Strommen, J., Shirazi, F., 2015. Methamphetamine Ingestion Misdiagnosed as Centruroides Sculpturatus Envenomation. Case reports in emergency medicine 2015.
Tan, P.T., Veeramani, A., Srinivasan, K.N., Ranganathan, S., Brusic, V., 2006. Scorpion2: a database for structure–function analysis of scorpion toxins. Toxicon 47, 356–363.
Tytgat, J., Chandy, K.G., Garcia, M.L., Gutman, G.A., Martin-Eauclaire, M.-F., van der Walt, J.J., Possani, L.D., 1999. A unified nomenclature for short-chain peptides isolated from scorpion venoms: α-KTx molecular subfamilies. Trends Pharmacol. Sci. 20, 444–447.
Uribe, J.I.C., Vargas, J.M.J., Batista, C.V.F., Zuñiga, F.Z., Possani, L.D., 2017. Comparative proteomic analysis of female and male venoms from the Mexican scorpion *Centruroides limpidus*: novel components found. Toxicon 125, 91–98.
Valdez-Velazquéz, L.L., Quintero-Hernández, V., Romero-Gutiérrez, M.T., Coronas, F.I.V., Possani, L.D., 2013. Mass fingerprinting of the venom and transcriptome of venom gland of scorpion Centruroides tecomanus. PLos One 8, e66486.
Valdez-Velazquéz, L., Romero-Gutierrez, M., Delgado-Enciso, I., Dobrovinskaya, O., Melnikov, V., Quintero-Hernández, V., Ceballos-Magaña, S., Gaitan-Hinojosa, M., Coronas, F., Puebla-Perez, A., et al., 2016. Comprehensive analysis of venom from the scorpion *Centruroides tecomanus* reveals compounds with antimicrobial, cytotoxic, and insecticidal activities. Toxicon 118, 95–103.
Vega, F., Lia, J., 1966. Epidemiological considerations on scorpion stings in the city of Durango. Rev. Invest. Salud Publica 26.
de la Vega, R.C.R., Possani, L.D., 2004. Current views on scorpion toxins specific for K^+-channels. Toxicon 43, 865–875.
de la Vega, R.C.R., Possani, L.D., 2005. Overview of scorpion toxins specific for Na^+ channels and related peptides: biodiversity, structure–function relationships and evolution. Toxicon 46, 831–844.
de la Vega, R.C.R., Schwartz, E.F., Possani, L.D., 2010. Mining on scorpion venom biodiversity. Toxicon 56, 1155–1161.
Vizcaíno, J.A., Csordas, A., del Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.W., Wang, R., Hermjakob, H., 2016. 2016 update of the PRIDE database and related tools. Nucleic Acids Res. 44, D447–D456.
Wang, G., Strichartz, G., 1983. Purification and physiological characterization of neurotoxins from venoms of the scorpions *Centruroides sculpturatus* and *Leiurus quinquestriatus*. Mol. Pharmacol. 23, 519–533.
Wheeler, K.P., Watt, D.D., Lazdunski, M., 1983. Classification of na channel receptors specific for various scorpion toxins. Pflügers Archiv. Europ.J. Physiol. 397, 164–165.

Zancolli, G., Sanz, L., Calvete, J.J., Wüster, W., 2017. Venom on-a-chip: a fast and efficient method for comparative venomics. Toxins 9, 179.
Zeng, X.-C., Corzo, G., Hahin, R., 2005. Scorpion venom peptides without disulfide bridges. IUBMB Life 57, 13–21.
Zhang, J., Kobert, K., Flouri, T., Stamatakis, A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614–620.
Zhang, L., Shi, W., Zeng, X.-C., Ge, F., Yang, M., Nie, Y., Bao, A., Wu, S., Guoji, E., 2015. Unique diversity of the venom peptides from the scorpion *Androctonus bicolor* revealed by transcriptomic and proteomic analysis. J. Proteom. 128, 231–250.