The remarkably enzyme-rich venom of the Big Bend Scorpion (*Diplocentrus whitei*)

Gunnar S. Nystrom[a], Schyler A. Ellsworth[a], and Darin R. Rokyta[a,*]

[a] Department of Biological Science, Florida State University, Tallahassee, Florida 32306

[*] Corresponding author: drokyta@bio.fsu.edu

Keywords: Scorpion, venom, *Diplocentrus*, arylsulfatase B

Corresponding author:
Darin R. Rokyta
Florida State University
Department of Biological Science
319 Stadium Dr.
Tallahassee, FL USA 32306-4295
email: drokyta@bio.fsu.edu
phone: (850) 645-8812

Abstract

Scorpion venoms have long been studied for their peptide discovery potential, with modern high-throughput venom-characterization techniques paving the way for the discovery of thousands of novel putative toxins. Research into these toxins has provided insight into the pathology and treatment of human diseases, even resulting in the development of one compound with Food and Drug Administration (FDA) approval. Although most of this research has focused on the toxins of scorpion species considered medically significant to humans, the venom of harmless scorpion species possess toxins that are homologous to those from medically significant species, indicating that harmless scorpion venoms may also serve as valuable sources of novel peptide variants. Furthermore, as harmless scorpions represent a vast majority of scorpion species diversity, and therefore venom toxin diversity, venoms from these species likely contain entirely new toxin classes. We sequenced the venom-gland transcriptome and venom proteome of two male Big Bend scorpions (*Diplocentrus whitei*), providing the first high-throughput venom characterization for a member of this genus. We identified a total of 82 toxins in the venom of *D. whitei*, 25 of which were identified in both the transcriptome and proteome, and 57 of which were only identified in the transcriptome. Furthermore, we identified a unique, enzyme-rich venom dominated by serine proteases and the first arylsulfatase B toxins identified in scorpions.

1 Introduction

As the cause of over 1.2 million stings in humans annually (Chippaux and Goyffon, 2008), scorpions and their venoms have been under scientific investigation for more than 200 years (Lourenço, 2014). The use of modern transcriptomic and proteomic techniques for characterizing scorpion venoms has led to the discovery of thousands of novel putative toxins, including antimicrobial (AMPs) and anticancer peptides, cysteine rich secretory proteins (CRISPs), ion-channel modulating toxins, non-disulfide bridge peptides (NDBPs), peptidases, proteases, phospholipases (PLA2s), as well as many with unknown functions (Quintero-Hernández et al., 2015; Santibáñez-López et al., 2016; Rokyta and Ward, 2017; Romero-Gutierrez et al., 2017; Batista et al., 2018; Cid-Uribe et al., 2018; Ward et al., 2018b; Grashof et al., 2019; Valdez-Velázquez et al., 2020). Ion-channel toxins, AMPs, and/or NDBPs typically comprise the more abundant and diverse venom proteins, with most proteases, peptidases, and other enzymes and proteins observed at lower abundances and diversity (Cid-Uribe et al., 2020), although this may not always be the case (Santibáñez-López et al., 2017; de Oliveira et al., 2018). Research on the function and biochemistry of these novel toxins has helped improve our understanding of the pathology and development of treatments for multiple human diseases, such as autoimmune disorders (Valverde et al., 2004; Huang et al., 2017; Tanner et al., 2017), cancers (Rui et al., 2005; Yamada et al., 2021), and vascular diseases (Song et al., 2005). To date, one scorpion venom-derived compound has received FDA approval (Tozuleristide; Yamada et al., 2021). Tozuleristide, or Tumor Paint®, is a fluorescent tumor imaging agent that was developed from a small-conductance Cl^--channel inhibiting toxin found in *Leiurus quinquestriatus* venom (Veiseh et al., 2007).

Although the use of combined proteomic and transcriptomic approaches have improved the speed and feasibility of high-throughput scorpion venom characterizations, these characterizations have been performed on less than 1% of all scorpion species (Ward et al., 2018a). Furthermore, research on characterizing scorpion venom components has disproportionately focused on the species considered medically significant, or harmful to humans. Of the approximately 2,200 species of scorpions, only about 104 are considered to be medically significant, yet the venom from more than half of these medically significant species has been at least partially characterized (Ward et al., 2018a). However, partial and high-throughput venom characterizations of harmless scorpion species have identified toxin families homologous to those with more dangerous stings (Quintero-Hernández et al., 2015; Santibáñez-López et al., 2017; Rokyta and Ward, 2017; Cid-Uribe et al., 2018; Ward and Rokyta, 2018), indicating that harmless scorpion venoms could serve as comparable, rich sources of novel biologically active peptides.

We therefore performed a high-throughput, proteomic-driven characterization of *Diplocentrus whitei* venom, representing the first high-throughput venom characterization for a scorpion of the genus *Diplocentrus*. Members of this genus are endemic to North and Central America and are the most diverse genus in the family Diplocentridae (Sissom and Fet, 2000; Santibáñez-López et al., 2014). *Diplocentrus whitei* (Gervais, 1884),

in particular, are distributed across southern parts of Brewster and Presidio counties in Texas and the adjacent regions of northern Mexico (Sissom and Fet, 2000). As fossorial scorpions, *D. whitei* tend to concentrate in rocky areas where the soil type supports their ability to burrow (Francke, 2019).

Biochemical characterization of the venom glands of two scorpions from Diplocentridae have been performed (Grashof et al., 2019; Rojas-Azofeifa et al., 2019). For example, transcriptomic characterization of *Nebo hierichonticus* venom-glands by Grashof et al. (2019) revealed an abundance and diversity of K^+-channel modulating toxins and bradykinin potentiating peptides. In addition, venom from the crab scorpion (*Didymocentrus krausi*), which exhibits cytotoxic and hemolytic effects on mammalian cell lines and myonecrotic effects on mouse gastrocnemius muscle, contains several types of proteases, non-disulfide bridge peptides, and other putative toxins (Rojas-Azofeifa et al., 2019). Although no high-throughput venom characterizations have been performed on *Diplocentrus*, two 1,4-benzoquinone compounds have been isolated from *Diplocentrus mellici* from Mexico and have shown effective bactericidal activity against *Staphyloccocus aureus* and *Mycobacterium tuberculosis* (Carcamo-Noriega et al., 2019). We sequenced the venom-gland transcriptome and venom proteome of two male *D. whitei* from the southern U.S. to screen for novel scorpion venom toxins and provide the first venom characterization for a *Diplocentrus* scorpion.

2 Materials and Methods

2.1 Sample collection

We collected *D. whitei* from Presidio County, Texas in summer of 2018 by searching along the sides of roads with UV lights after dark. Scorpions were maintained at Florida State University and sexed by counting pectine teeth under the microscope (females: 14–18, males: 16–20; Stockwell and Nilsson, 1987; Stockwell and Baldwin, 2001). We performed all the following venom proteomic and venom-gland transcriptomic analyses on two individual male *D. whitei* (C0687 and C0689). Venom was extracted from *D. whitei* by anesthetizing scorpions in CO_2 and electrostimulating the telson, as previously described in other scorpion species (Rokyta and Ward, 2017; Ward et al., 2018b). Lyophilized venom was stored at $-80°C$ until use in the following proteomic analyses. Venom glands were dissected under stereoscopic microscope four days after venom extraction, transferred to 100 μL of RNAlater, kept overnight at 4°C, and then stored at $-80°C$ until RNA extraction. To preserve scorpion specimens we placed each specimen in 95% ethanol and stored at $-80°C$.

2.2 Venom proteomics

Total protein content of *D. whitei* venom samples was quantified with a Nanodrop 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA). To gen-

erate a chromatographic profile of *D. whitei* venom, venom samples were run on a Shimadzu Prominence reversed-phase high-performance liquid chromatography (RP-HPLC) system. Approximately 15 μg of venom was injected onto an Aeris 3.6 μm C18 column (Phenomenex, 125 Torrance, CA). Samples were run at a flow rate of 0.2 mL/min using a standard solvent system of solution A (0.1% trifluoroacetic acid in water) and solution B (0.06% trifluoroacetic acid in acetonitrile) with the following 140-minute gradient and column wash: 10% B for five minutes, gradual increase to 55% B over 110 minutes, increase to 75% B over five minutes, five minutes at 75% B, and finally a 15-minute column wash step at 100% B.

To prepare samples for quantitative mass spectrometry (LC-MS/MS), we submitted 11 μg of dried whole venom to the Florida State University's Department of Biological Science Core Facilities for trypsin digestion. Venom samples were prepared for reduction and denaturation by adding 150 μL of 100mM ammonium bicarbonate and incubating for 20 minutes at room temperature. Samples were then reduced by adding 30 μL of 10mM dithiothreitol and incubated for 10 minutes in the dark at room temperature. Next samples were denatured by heating to 60°C for one hour. After denaturation, we added 30 μL of 50mM Iodoacetamide (alkylating agent) and incubated samples for 30 minutes in the dark at room temperature. After adding 150 μL of ammonium bicarbonate we started digestion using 1 μL of trypsin (Promega Cat. No. V5111). Samples were allowed to incubate for approximately 18 hours at 37°C before adding about 18 μL of 1% Trifluoroacetic acid to terminate digestion. Digested venom samples were dried with a SpeedVac and LC-MS/MS was completed on each venom sample in triplicate by the College of Medicine Translational Science Laboratory at Florida State University, as previously described (Ward et al., 2018b).

The resulting LC-MS/MS data were analyzed using Proteome Discover (version 2.5), custom FASTA databases, percolator for peptide and protein validation, and SequestHT as the search engine with the following settings: dynamic modifications, Trypsin as the enzyme name, fragment mass tolerance of 0.2 Da, carbamidomethyl +57.021 Da(C), minimum peptide length of 6, maximum peptide length of 144, maximum missed cleavage of 2, maximum delta Cn of 0.05, oxidation +15.995 Da(M), and a precursor mass tolerance of 10 ppm. Next, we confirmed protein and peptide identities in each venom sample using Scaffold (version 5.1.2; Proteome Software Inc., Portland, OR, USA) with the minimum number of recognized peptides set at 1 and the peptide and protein false discovery rates set a 1.0%. Finally, we calculated estimates of peptide abundances for each unique peptide in each of the three LC-MS/MS replicate per individual, as described by Ward et al. (2018b).

2.3 Transcriptome sequencing

We extracted RNA from *D. whitei* venom-glands by removing glands from RNAlater and performing a TRIzol-chloroform (Invitrogen) extraction, as previously described (Rokyta et al., 2012; Ward et al., 2018b; Ward and Rokyta, 2018). We quantified and

quality checked total RNA content of our samples using the Qubit RNA Broad-range kit (Thermo Fisher Scientific) and an RNA 6000 Pico Bioanalyzer chip (Agilent Technologies), per the manufacturer's instructions, respectively. Next, we isolated the mRNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs) and fragmented it for 15.5 minutes to generate average fragment sizes of approximately 370 base pairs. We prepared sequencing-ready cDNA libraries from isolated mRNA using a NEBNext Ultra RNA Library Prep Kit, High Fidelity 2× Hot Start PCR Mix, AMPure XP beads (Agencourt) for purification the PCR reaction, and Illumina Multiplex Oligos as our unique sequencing indices (New England Biolabs). We quantified and quality checked our cDNA libaries using KAPA PCR (performed by the Molecular Cloning Facility at Florida State University) and a High Sensitivity DNA Bioanalyzer chip (Aglient Technologies). Quality checked libraries were pooled with other sequencing libraries and sequenced with 150PE on an Illumina NovaSeq 6000 system at the Florida State University College of Medicine Translational Science Laboratory.

2.4 Transcriptome assembly and analysis

Using the raw 150PE sequencing reads, we analyzed and assembled transcriptomes as previously described (Rokyta and Ward, 2017; Holding et al., 2018; Ward et al., 2018b), with some slight modifications. Raw reads were filtered and quality controlled to identify and remove sample cross contamination using custom python scripts and FastQC (version 0.11.5). Next we quality trimmed and merged the resulting filtered reads using Trim Galore! (version 0.4.4; Krueger, 2015) and PEAR (version 0.9.6; Zhang et al., 2014), respectively. Using a multi-assembly approach to maximize our ability to identify unique toxin transcripts, we performed *de novo* transcriptome assembly using DNAStar NGen (version 12.3.1), Extender (version 1.04; Rokyta et al., 2012), and Trinity (version 2.4.0; Grabherr et al., 2011). We ran DNAStar NGen and Trinity (kmer size of 31) using both the merged and unmerged reads and considered all reads as unpaired. We ran Extender using just the merged reads and the following parameters: minimum phred of 30, overlap of 20 nucleotides, and replicates of 20.

Using the assembled transcriptomes and a custom python script to filter out contigs from our three assemblers, we annotated putative toxins in *D. whitei* based on homology to other known toxins from the Uniprot (UPT; downloaded April 13, 2018) toxin database. We only performed homology-based searches on contigs that displayed a match of at least 90% of the total length of a curated toxin from the UPT toxin database. Next, we took the open reading frame (ORF) present in the primary BLAST hit for searched contigs and checked for signal peptides using SignalP (version 4.1; Petersen et al., 2011) under the sensitive settings and checked for a valid stop codon. All sequences that contained a valid stop, a signal peptide, and were $> 90\%$ of the total length of the primary BLAST hit, were kept and named for the toxin with the primary BLAST hit from the UPT dataset. To maximize our ability to identify venom components and complement our homology-based approach to annotation, we also performed a proteomic-driven toxin

annotation. Using the getorf function in Emboss (version 6.6.0.0; Rice et al., 2000), we identified all available ORFs across our contigs. Next, we took our *D. whitei* venom LC-MS/MS data and searched it against custom databases for each assembly using Proteome Discoverer and Scaffold. Implementing custom python scripts, we annotated putative sequences by (1) identifiying all available ORFs with proteomic evidence, and (2) validating that each of these ORFs had a valid stop codon and signal peptide. All validated contigs were kept and named according to the primary hit from a BLAST search of the UPT toxin database. Putative toxin sequences from the proteomic-driven annotations were combined within individuals and clustered with cd-hit-est (version 4.6; Li and Godzik, 2006) at 100% sequence identity. We then aligned merged reads against this combined putative toxin set using bwa (version 0.7.12; Li, 2013) to identify and discard chimeric sequences. Using a 151 base pair sliding window, we screened read distributions and removed any sequences that did not show any coverage within windows. Remaining sequences with read distributions that differed by > 20 fold were individually hand checked for potential chimeric properties. All remaining putative toxins within individuals were clustered at 99% using cd-hit-est before combining putative toxins from both individual *D. whitei* and clustering within species at 98%. Nontoxin transcripts were annotated from contigs generated from the Trinity assembly using BUSCO (version 5.1.2; Seppey et al., 2019) under the genomics settings and using the Arachnida Odb10 database (downloaded August 2020). Single copy BUSCO matches for each individual were then fed into custom scripts to name and verify that each transcript had a valid start position and valid stop. To generate the consensus transcriptome for *D. whitei*, we combined this putative toxin set with the transcripts from the homology-based annotation approach above and clustered at 98% using cd-hit-est. To calculate individual transcript abundances, we used RSEM with bowtie2 (version 2.3.0; Langmead and Salzberg, 2012) alignments against coding sequences in the consensus transcriptome. Finally, the estimated abundances for the proteome and transcriptome were clr-transformed (Aitchison, 1986) and used to test for protein and transcript abundance relationships within and between individuals.

2.5 Multiple protein sequence alignment

All multiple protein sequence alignments of *D. whitei* venom toxins were performed using Clustal Omega (version 1.2.3; Sievers et al., 2011; Sievers and Higgins, 2018) with the default settings in Geneious Prime (version 2022.2.2; Kearse et al., 2012). Alignments were visualized using the ggmsa package (Zhou et al., 2022) in R version 4.2.1. Protein domains were identified by searching amino acid sequences against the NCBI Conserved Domain Database (Marchler-Bauer et al., 2010, 2015, 2017; Lu et al., 2020).

2.6 Data Availability

All raw RP-HPLC data is available in Supplementary Table 1. All raw transcriptome reads can be found in the National Center for Biotechnology Information (NCBI) Se-

quence Read Archive (SRA) under the BioProject number PRJNA340270, BioSamples SAMN27783483 (C0687) and SAMN27783484 (C0689), and SRA accession numbers SRR18927020 (C0687) and SRR18927019 (C0689). The assembled transcripts were deposited in the NCBI Transcriptome Shotgun Assembly (TSA) database at DDBJ/EMBL/GenBank under the accession number GJYU00000000. The version in this paper represents the first version (GJYU01000000). The mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Vizcaíno et al., 2016) with the dataset identifier PXD033911 and 10.6019/PXD033911.

3 Results and Discussion

3.1 The transcriptomic basis for *D. whitei* venom

After Illumina quality filtering of our raw venom-gland RNA-seq data, we generated 20,416,197 raw read pairs for *D. whitei* individual C0687 and 11,771,573 raw reads pairs for individual C0689. After read trimming and merging, we generated 15,179,881 and 9,098,631 merged reads for C0687 and C0689, respectively. Using our multi-assembly (*i.e.*, Extender, DNAStar NGen, and Trinity) and mass spectrometery-directed approaches to annotating the venom-gland transcriptome and venom proteome, we identified a total of 1,929 protein-encoding sequences from *D. whitei*, including 30 proteins detected in the venom-proteome of at least one individual. The unique peptide counts and percent MS/MS coverage for the proteins detected via LC-MS/MS are reported in Table 1 and their amino acid sequence, top nr protein BLAST, and PFam matches are reported in Supplementary Table 2. Of the 1,929 protein-encoding sequences, we classified 1,838 as nontoxins and 87 as putative toxins that were identified in the venom-gland transcriptome by homology to known scorpion venom toxins and/or venom proteome of at least one individual. However, five of the putative toxins were proteins that showed no clear signal peptide, contributed less than a combined 0.6% abundance in both the proteome and transcriptome of each individual, and coded for proteins with likely roles in cell growth and maintenance (*i.e.*, chromodomain helicase DNA binding protein, Dbl homology domain protein, glutamine-dependent NAD(+) synthetase, mediator of RNA polymerase II transcription, and vacuolar protein sorting-associated protein). We therefore classified them as nontoxins, resulting in 1,847 nontoxins and 82 putative toxins. Of these 82 toxins, 25 were proteomically confirmed in the venom of at least one *D. whitei* individual and 57 were identified only on the basis of homology to known animal toxins.

The nontoxin transcripts were responsible for 136,924.63 transcripts per million (TPM) and 163,573.07 TPM in C0687 and C0689, respectively. The proteomically confirmed toxin transcripts were responsible for 506,742.05 and 514,679.40 TPM in C0687 and C0689, respectively. The homology-only toxin transcripts were responsible for 356,333.16 and 321,747.48 TPM in C0687 and C0689, respectively. We observed a strong agreement in nontoxin transcript (Spearman's rank correlation $\rho = 0.97$, Pearson's rank

correlation coefficient $R = 0.96$, and $R^2 = 0.93$; Figure 1, left), protoeomically confirmed toxin transcript (Spearman's rank correlation $\rho = 0.94$, Pearson's rank correlation coefficient $R = 0.93$, and $R^2 = 0.87$; Figure 1, middle) and homology-only toxin transcript abundances (Spearman's rank correlation $\rho = 0.82$, Pearson's rank correlation coefficient $R = 0.77$, and $R^2 = 0.59$; Figure 1, right) between *D. whitei* individuals, indicating a high degree of similarity between venom-gland transcriptomic profiles. Only four proteomically confirmed toxin sequences (*i.e.* CathepsinD-1, CathepsinD-2, CathepsinD-3, and GC-1) and 13 homology-only toxin sequences (*i.e.* HistP-1, NDBP-8, 9, and 11, PeptidaseM2-1, SP-6, and TIL-2–4, TIL-6, and TIL-11–13) fell outside a 99% confidence interval of differences between the two nontoxin measures, representing toxins expressed at unusually different abundances between individuals. However, these differences clearly did not have a large impact on the observed strong toxin transcript abundance correlations between individuals.

3.2 Novel arylsulfatase B toxins in *D. whitei* venom

We identified two arylsulfatase B toxins (ARSBs) in the venom of *D. whitei* (ARSB-1 and ARSB-2; Table 2), the first ARSB toxins identified in a scorpion venom. ARSBs were the most abundant toxin class in the venom proteome of one individual (52.2%; C0687) and the second most abundant toxin class in the venom proteome of the other (24.4%; C0689) with ARSB-1 being the most abundant protein in both venom proteomes (Figure 2, Table 2). However, they were observed at lower abundances in the transcriptomes of both C0687 (6.2%) and C0689 (5.5%). In humans, ARSBs are proteins that metabolize glycosaminoglycans (GAGs) in the lysosomes of several tissues, including colonic epithelium, liver, renal, and pancreatic tissues (Kovacs et al., 2019). More specifically, they metabolize the chondroitin sulfate/dermatin sulfate GAGs (CSGAGs) by hydrolyzing the 4-*O*-sulfate group from *N*-acetyl-D-galactosamine residues at the non-reducing end of CSGAGs (Matalon et al., 1974; Peters et al., 1990). CSGAGs provide adhesive support to bones and tissues and likely play a role in growth factor signaling, wound repair, cell division, and central nervous system development (Sugahara et al., 2003).

In animal venoms, arylsulfatases have been observed in the black-necked spitting cobra (*Naja nigricollis*; Nok et al., 2003), the ectoparasitoid wasp *Nasonia vitripennis* (De Graaf et al., 2010), and in the salivary glands of the giant triton snail (*Charonia tritonis*), which are thought to produce venom and/or sulfuric acid (Bose et al., 2017). Nok et al. (2003) isolated the arylsulfatase from *N. nigricollis* venom and showed that the enzyme had the ability to hydrolyze the sulfated GAG, chondroiton-4-sulfate. They suggested that because chondroiton-4-sulfate provides adhesive support to connective tissue, ligaments, tendons, and the aorta, *N. nigricollis* arylsulfatases may compromise the lubricating role of GAGs, upsetting homeostasis. Compromising the adhesive support of GAGs could precede the toxic effects of other venom components, thereby amplifying overall *N. nigricollis* venom toxicity. In *C. tritonis*, Bose et al. (2017) suggested that ARSBs may be involved in prey digestion because of (1) the likelihood that arylsulfatases

metabolize saponins (Pesentseva et al., 2012), which are glycosides produced by plants and some animals (*e.g.* Echinoderms) as chemical defenses against predation, and (2) that they have been observed at high levels in digestive organs from other carnivorous molluscs (Corner et al., 1960). In *D. whitei* venom, ARSBs may metabolize GAGs as suggested in *N. nigricollis* and *C. tritonis* and may therefore interfere with the normal physiological processes (*e.g.* cell signaling, connective tissue support, etc.) of prey and/or aid in digestion. Further studies on the enzymatics of these ARSBs is needed to confirm hypothesized functions. However, their moderate abundance in the venom-gland transcriptomes and high abundance in venom proteomes provides strong evidence that ARSBs represent real venom toxins that likely play a major role in overall *D. whitei* venom function.

Both ARSB-1 and ARSB-2 from *D. whitei* had molecular weights of approximately 62 kDa and a 19 amino acid long signal peptide (Table 2). In addition, both ARSBs contained one alkaline phosphatase and sulfatase (ALP) superfamily domain, or more specifically, an N-acetylgalactoseamine 4-sulfatase (4-S) domain, and one domain of unknown function 4976 (DUF4976). Proteins with a DUF4976 typically have lengths of around 530 amino acids and several have been identified as arylsulfatases, although the function of these proteins is unknown (Lu et al., 2020). These proteins are primarily found in bacteria of the genus *Bacteroides*, with the DUF4976 appearing downstream of a sulfatase domain (Donaldson et al., 2020; Lu et al., 2020). ARSB-1 and ARSB-2 displayed a 51.2% and 53.6% match to a nontoxic genomic ARSB sequence from the Arizona bark scorpion (*Centruroides sculpturatus*; i5K Consortium, 2013) in the NCBI non-redundant (nr) protein database and a 56.7% and 56.3% match to a transcribed RNA sequence from the scorpion *Euscorpius sicanus* (1K Insect Transcriptome Evolution project, or 1KITE; https://1kite.cngb.org/) in the NCBI Transcriptome Shotgun Assembly (TSA), respectively. Although the top scorpion homologs from *C. sculpturatus* in the nr database did contain a signal peptide and a 4-S domain, they did not contain clear evidence for a DUF4976 via our NCBI conserved domain database. Furthermore, nr database BLAST searches of the top non-*Centruroides* homologs, such as that from the horseshoe crab (*Limulus polyphemus*), revealed that other nontoxic arylsulfatase homologs from other invertebrates do contain evidence for a DUF4976. However, our multiple protein sequence alignment of the two *D. whitei* ARSBs and their top protein BLAST matches from the NCBI database, including the top scorpion ARSB homologs from *C. sculpturatus*, showed a higher degree of sequence conservation in the 4-S and DUF4976 domain regions (Figure S1), relative to the signal peptide and non-domain containing regions of the proteins. Therefore, although our domain search did not reveal clear evidence for a DUF4976 domain in the top scorpion ARSB homologs, this high degree of sequence homology suggests the DUF4976 domain is likely still present in the *C. sculpturatus* ARSB homologs.

3.3 High abundance of serine proteases in *D. whitei* venom

Serine proteases (SPs) are widespread across animal venoms and have been observed in snakes (Tasoulis and Isbister, 2017), spiders (Veiga et al., 2000; Khamtorn et al., 2020), hymenopterans (Han et al., 2008; De Graaf et al., 2010), centipedes (Undheim et al., 2014), and scorpions (Rokyta and Ward, 2017; Ward et al., 2018b; So et al., 2021). SPs found in snake venom likely possess proinflammatory and cytotoxic activity (Menaldo et al., 2013; Nalbantsoy et al., 2017). In scorpions, serine proteases are thought to possess gelatinolytic (Almeida et al., 2002) and fibrinogenolytic activity (Brazón et al., 2014). We identified six SPs in the venom-gland transcriptomes of *D. whitei* (Table 2). However, only two SPs were proteomically confirmed (SP-4 and SP-5; Table 2). Serine proteases were the third most abundant toxin class in the transcriptome of both individuals, the second most abundant toxin class in the proteome of C0687 (behind ARSBs), and the most abundant toxin class in the proteome of C0689 (Figure 2). They accounted for 14.4% and 17.2% of the total toxin transcript abundance in the transcriptome and 18.9% and 28.8% of the total toxin abundance in the venom proteome for C0687 and C0689, respectively. Serine proteases have not previously been observed at such high abundances in scorpion venoms (Cid-Uribe et al., 2020). In addition, SPs were observed at much lower abundances in the venom proteomes of *C. hentzi* ($< 1.3\%$) and *H. spadix* ($< 6.1\%$), both of which were analyzed with nearly identical toxin annotation and quantification strategies (Rokyta and Ward, 2017; Ward et al., 2018b).

The two proteomically confirmed SPs in *D. whitei* venom, SP-4 and SP-5, were also by far the most abundant SPs in the transcriptomes of both individuals (Table 2). Both SP-4 and SP-5 had molecular weights of 31 kDa and contained a trypsin-like serine protease superfamily domain. SP-4 had a signal peptide of 25 amino acids in length while SP-5 had a signal peptide of 17 amino acids in length. SP-4 and SP-5 showed 58.2% and 48.2% similarity to SP-3 from *Hadrurus spadix* venom (Rokyta and Ward, 2017), respectively. As other trypsin-like SPs from scorpion venoms, such as those from *Tityus bahiensis* and *Tityus serrulatus*, are thought to possess gelatinolytic activity (Almeida et al., 2002), SPs from *D. whitei* may serve a similar functional role. Almeida et al. (2002) suggested that proteolytic enzymes, including SPs, may serve as spreading factors and could facilitate the spread of other toxins by increasing the permeability of the affected tissue. Conversely, SPs could also be important for the post-translational modification and processing of other *D. whitei* venom proteins. Enzymatic verification of these toxins is necessary to confirm their function in *D. whitei* venom. However, their high abundance in both *D. whitei* venom proteomes and venom-gland transcriptomes provides strong evidence that SPs play a significant role in venom function.

3.4 Other proteomically confirmed toxins

3.4.1 Acetylcholinesterase

Acetylcholinesterases (AChEs), which have been identified in snakes venoms (Frobert et al., 1997; Ahmed et al., 2021), are thought to rapidly disrupt neurotransmission by hydrolysis of the neurotransmitter acetylcholine (Colovic et al., 2013). We identified one acetylcholinesterase (AChE-1) in *D. whitei* venom. AChE-1 contained a 19 amino acid signal peptide, six cysteine residues, and had an estimated molecular weight of 61 kDa (Table 2). AChE-1 also contained an alpha/beta hydrolase superfamily domain. However, the most significant database matches for AChE-1 were to nontoxic homologs. AChE-1 was observed at low abundances in both transcriptomes ($> 1\%$ in both cases; Figure 2), but the moderate abundance in the venom proteomes of both C0687 (3.7%) and C0689 (12.7%) suggests AChE-1 represents a real venom toxin with a non-trivial role in venom function. Although AChEs have not been identified in scorpions venoms, AChE-1 from *D. whitei* venom may possess neurotransmission disruption capabilities similar to those observed in snake venoms, but further studies are needed to confirm this hypothesis.

3.4.2 Nucleotidase

Nucleotidases (NUCs) are hydrolytic enzymes that have been observed in the venom of other animals, including snakes (Dhananjaya and D'Souza, 2010) and scorpions (Cid-Uribe et al., 2020). As they breakdown nucleic acid containing substrates, such as ATP, NUCs from snake venoms may play a role in prey immobilization via depletion of ATP (Dhananjaya and D'Souza, 2010). However, no NUCs identified in scorpion venoms have received functional characterization. We identified one NUC in *D. whitei* venom (NUC-1) that had a 17 amino acid long signal peptide, 7 cysteine residues, and a molecular weight of approximately 64 kDa (Table 2). NUC-1 also showed a 71.8% match to a nucleotidase identified from the venom of *H. spadix* (Rokyta and Ward, 2017) in the TSA database. NUC-1 was responsible for 1.8% and 1.7% of the total toxin transcript abundance in the transcriptome and 5.9% and 7.4% of the total toxin abundance in the proteome for individuals C0687 and C0689 (Figure 2), respectively.

3.4.3 Peroxidase

We also identified one peroxidase (Peroxidase-1) that was proteomically confirmed in both individuals (Table 2). The top nr and TSA database matches for Peroxidase-1 were to nontoxic homologs, which could suggest that Peroxidase-1 represents a protein that was accidentally introduced into the venom during extraction. However, the relatively moderate to high abundance of Peroxidase-1 in the transcriptome (C0687 = 2.5% , C0689 = 3.1%; Figure 2) and proteome (C0687 = 7.2%, C0689 =11.7%; Figure 2) of both individuals could indicate otherwise. Peroxidase-1 was found to contain a 21 amino acid signal peptide, 33 cysteine residues, and a molecular weight of approximately 76 kDa.

This protein also contained an animal haem peroxidase superfamily domain. Animal peroxidases use hydrogen peroxide to catalyze oxidative reactions (Furtmüller et al., 2006), suggesting they may be involved in the oxidative stress response. As peroxidases have not been observed in scorpion venoms, the role that Peroxidase-1 plays, if any, in *D. whitei* venom function is unclear. Other oxidative enzymes, such as the L-amino acid oxidases (LAOOs), have been found to play a significant functional role in snake venoms (Guo et al., 2012). LAAOs are flavoenzymes that catalyze the removal of an amine group from L-amino acids, resulting in the production of the harmful reactive oxygen species, H_2O_2 (Guo et al., 2012). These toxins have shown effective cytotoxic and proinflammatory activities, and the ability to induce cell apoptosis (Alves et al., 2008; Zhang and Wu, 2008; Wei et al., 2009). Similarly, Peroxidase-1 in *D. whitei* venom may serve to catalyze the production of toxic reactive oxygen species in prey or predators, but further enzymatic studies would be needed to confirm this hypothesis.

3.4.4 Lower abundance toxins

We also identified six other putative venom toxins at lower proteomic and transcriptomic abundances (Table 2), including three CathepsinD toxins (CathepsinD-1, CathepsinD-2, and CathepsinD-3; 41–43 kDa and 16–22 amino acid signal peptides), one cysteine-rich secretory peptide (CRISP-1; 51 kDa with a 20 amino acid signal peptide), one hyaluronidase (HYAL-1; 46 kDa with a 20 amino acid signal peptide), and one transferrin (Transferrin-1; 77 kDa with a 19 amino acid signal peptide), all of which had database matches to previously identified scorpion venom toxins. CathepsinD-3, and HYAL-1 were only identified in the proteome of C0689 (Table 3). We also found one transforming growth factor-beta-induced protein ig-h3 (TGFBI-1; Table 2), which was only expressed in the proteome of individual C0687 (Table 3). TGFBIs are secretory extracellular matrix proteins that may contain N-terminal signal peptides (Runager et al., 2008; Ween et al., 2012), indicating that TGFBI-1 may not be a toxic component of *D. whitei* venom. We also identified one venom protein that displayed evidence for a somatomedin B domain and had a 38.6% match to a nontoxic G protein-coupled receptor homolog (GPCR-1; Table 2) from *C. hentzi* (Ward et al., 2018b). Although not common, GPCRs have been observed to contain signal peptides (Schülein et al., 2012), suggesting that GPCR-1 is likely not a putative toxin in *D. whitei* venom. We also found one glucosylceramidase (GC-1; Table 2) in *D. whitei* venom. Glucosylceramidases are hydrolytic enzymes that play an essential role in sphingolipid metabolism (Astudillo et al., 2016). However, the role that GC-1 may play in *D. whitei* venom function is unclear. TGFBI-1, GPCR-1, and GC-1 may also play a role in venom-gland cell maintenance and could have leaked into the venom during the extraction process.

Finally, we identified three proteomically confirmed venom proteins for which we could not designate a functional classification (VP-3, VP-4, and VP-7; Table 2). VP-3 was only identified in the proteome of C0689 (Table 3), had a relatively small molecular weight (13 kDa), and an 18 amino acid long signal peptide. VP-4 had a molecular

weight of 19 kDa and a 19 amino acid long signal peptide. The last VP, VP-7, was only detected in the proteome of C0687 (Table 3), had a molecular weight of 16 kDa, and a 26 amino acid long signal peptide. All VPs showed significant database matches to other uncharacterized scorpion venom proteins.

3.5 Toxins identified by homology

We identified several classes of scorpion venom toxins at high relative abundances in *D. whitei* venom-gland transcriptomes that had either greatly reduced expression in or were absent from venom proteomes (Figure 2, Table 2), including antimicrobial peptides with (AMPs) and without disulfide bridges (NDBPs), ion-channel toxins, and La1-like peptides. The low expression and/or absence of AMPs, NDBPs, ion-channel toxins, and La1-like peptides in the venom proteome was not surprising as small molecular weight toxins have been difficult to proteomically detect in other scorpion venoms (Zhang et al., 2015; Rokyta and Ward, 2017; Ward et al., 2018b; Romero-Gutiérrez et al., 2018), potentially because they require significant post-translational modifications to produce the mature peptide. Furthermore, Rokyta and Ward (2017) suggested that because peptides digested with trypsin show varying propensities to mass spectrometry detection, some peptides may have been more difficult to detect than others. Although more research is necessary to make sense of toxin signal discrepancies between venom proteomic and venom-gland transcriptomic abundances, these detectability challenges emphasize importance of using joint transcriptomic and proteomic approaches for scorpion venom gene characterizations.

3.5.1 Antimicrobial peptides with and without disulfide bridges

Antimicrobial peptides, or host defense peptides, are widespread in scorpion venoms and typically classified as those with disulfide bridges and those without disulfide bridges (Harrison et al., 2014; Cid-Uribe et al., 2020). We identified three antimicrobial peptides with disulfide bridges (AMP-1–3) and 11 without disulfide bridges (NDBP-1–11) in *D. whitei* (Table 2). Of the three observed AMPs, only two were identified in the proteome of at least one individual (AMP-1 and AMP-2; Table 2). These AMPs were responsible for 11.9% and 7.7% of the total toxin transcript abundance in the venom-gland transcriptomes of C0687 and C0689, respectively (Figure 2). AMPs also contributed 4.2% and 2.2% of the total toxin abundance in venom proteomes from C0687 and C0689, respectively. Each of the three AMPs contained a signal peptide of 19–20 amino acids long, six cysteine residues, and showed homology to previously described scorpion venom AMPs in the TSA database. AMP-1 and AMP-3 had 57.3% and 48.1% matches to the Scorpine-like AMP-1 from *H. spadix*, respectively, and AMP-2 had a 55.8% match to two different Scorpine-like AMPs (AMP-4 and AMP-7) from *H. spadix*.

Of the NDBPs identified in *D. whitei*, all 11 were identified in the venom-gland transcriptome of both *D. whitei* (Table 2), but only NDBP-2 was detected in the proteome of C0687 (Table 3). However, NDBPs were the most abundant toxin in the transcriptome

of both C0687 and C0689 and were responsible for 25.9% and 19.4% of the total toxin transcript abundance, respectively. NDBP-2 was responsible for 1.5% of the total toxin abundance in the proteme of C0687. As AMPs and NDBPs are typically found at higher diversities and abundances in the non-Buthidae scorpions (Cid-Uribe et al., 2020), their significant diversity and abundance in the transcriptome of *D. whitei* is not surprising. All 11 NDBPs had a precursor length of 65–88 amino acids (Table 2), but each also contained a 22 amino acid signal peptide, which is likely cleaved during proteolytic processing. Many NDBPs possess a processing signal and a propeptide that are also cleaved during post-translational processing, resulting in mature peptides of only 13–56 amino acids (Zeng et al., 2005; Almaaytah and Albalas, 2014). As previously discussed, the discrepancy between transcriptome and proteome abundance is likely due to significant post-translational processing and the small size of the mature NDBPs.

Using the propeptide cleavage site predictor, ProP (version 1.0; Duckert et al., 2004), we predicted propeptide cleavage signals in eight *D. whitei* NDBPs (NDBP1, NDBP-4, NDBP-6, NDBP-7, NDBP-8, NDBP-9, NDBP-10, and NDBP-11). Zeng et al. (2005) classified scorpion venom NDBPs into six distinct subfamilies based on their pharmacological activity, length, and sequence similarity. Although NDBPs were more recently classified into five groups by Almaaytah and Albalas (2014), this newer classification only accomodates those that are functionally characterized. As we did not functionally characterize any of the NDBPs from *D. whitei*, we classified *D. whitei* NDBPs by homology to the closest subfamily in the Zeng et al. (2005) classification. We identified four NDBPs with closest nr and/or TSA database matches to scorpion venom NDBPs from Zeng et al. (2005) subfamily-3 (NDBP-2, NDBP-3, NDBP-6, and NDBP-9) and seven NDBPs with closest matches to scorpion venom toxins from subfamily-4 (NDBP-1, NDBP-4, NDBP-5, NDBP-7, NDBP-8, NDBP-10, and NDBP-11).

Both AMPs and NDBPs have shown effective antimicrobial activity against pathogenic microorganisms (Conde et al., 2000; Torres-Larios et al., 2000; Uawonggul et al., 2007; Trentini et al., 2017; Jiménez-Vargas et al., 2021). AMPs, in particular, are also thought to interact with ion channels (Harrison et al., 2014) with those from *Tityus discrepans* having shown effective modulatory activity against insect and mammal sodium channels (Peigneur et al., 2012). A tendency to target ion channels suggests that AMPs likely play a role in prey subjugation and/or predator deterrence. Conversely, NDBPs have received significant attention because of their relatively small size and diversity of potential functions, including anticancer activity (Almaaytah et al., 2013), antimicrobial and hemoloytic activity (Torres-Larios et al., 2000; Moerman et al., 2002; Trentini et al., 2017; Jiménez-Vargas et al., 2021), bradykinin-potentiating activity (Ferreira et al., 1993; Meki et al., 1995), and immune modulating activity (Moerman et al., 2003; Willems et al., 2004). AMPs and NDBPs have also been suggested to regulate the scorpion's external microbiome and internal venom-gland microbiome. For instance, some scorpion species are thought to spray themselves with their own venom to help remove bacteria and fungi (Torres-Larios et al., 2000) or to clean their own wounds (Gao et al., 2007). Self-spraying was suggested as a mechanism for AMPs in *Hadrurus aztecus*

venom because of the species' burrowing tendencies and therefore constant exposure to soil microbes (Torres-Larios et al., 2000). In the venom from another burrowing scorpion, *H. spadix*, AMPs and NDBPs were one of the most diverse and abundant toxin families (Rokyta and Ward, 2017). As a burrowing species, *D. whitei* could also be using their venom as part of a self-spraying mechanism, but further studies are needed to test for this behavior. Using an infection model, Gao et al. (2007) detected antibacterial activity in the venom of *Buthus martensii* after injection of *Escherichia coli* and *Micrococcus luteus* into the venom glands. As previously characterized NDBPs from *B. martensii* were shown to possess significant antibacterial activity, Gao et al. (2007) also suggested that these peptides could protect the venom-gland from infection. However, whether AMPs and NDBPs in *D. whitei* venom also play a role in protecting the venom-gland from infection is unclear.

3.5.2 Ion-channel toxins

We identified seven toxins with homology to known ion-channel toxins from scorpion venoms (Table 2), including four alpha K^+-channel toxins (αKTx), one beta K^+-channel toxin (βKTx), one kappa K^+-channel toxin (κKTx), and one Ca^{2+}-channel toxin (CaTx). K^+-channel toxins, in particular, are one of the more diverse and abundant toxin classes in scorpion venoms (Quintero-Hernández et al., 2013). Overall, ion-channel toxins were the second most abundant toxin class in the transcriptomes of both C0687 and C0689 and were responsible for 15.3% and 17.8% of the total toxin transcript abundance, respectively (Figure 2). No ion-channel toxins were identified in either venom proteome, likely because of mass spectrometry detectability challenges associated with their small size.

The αKTxs are the most diverse family and comprise peptides of approximately 23–42 amino acids longs with 3–4 disulfide bridges. These toxins are thought to interact with K^+-channels either extracellularly or via blocking the channel pore (Quintero-Hernández et al., 2013). Of the four αKTx channel toxins we identified, two of them (αKTx-1 and αKTx-2) were responsible for the majority of the total αKTx abundance in *D. whitei* venom glands (Table 2). αKTx-1 and αKTx-2 had signal peptides of 25 amino acids and eight cysteine residues. The remaining two αKTxs, αKTx-3 and αKTx-4, had 25 and 32 amino acid long signal peptides and six and eight cysteine residues, respectively. The βKTxs are are long chain toxins (50–75 amino acids) that have shown various K^+-channel blocking and inhibition abilities (Diego-García et al., 2008; Quintero-Hernández et al., 2013). We identified one βKTx (βKTx-1; Table 2), which had a 21 amino acid long signal peptide and six cysteine residues. κKTxs have also been identified in scorpion venoms and comprise peptides of two α-helices connected by two disulfide bridges (Quintero-Hernández et al., 2013). These peptides may interact with and inhibit K^+-channels with comparable mechanisms to the αKTxs (Srinivasan et al., 2002; Quintero-Hernández et al., 2013). The κKTx we identified in *D. whitei* venom glands (κKTx-1; Table 2) had a 17 amino acid long signal peptide and four cysteine residues. This κKTx matched to only

one sequence from the nr and NCBI databases, which was a previously identified κKTx from *Pandinus cavimanus* (κ-KTxpcavC10; Diego-García et al., 2012). However, a multiple protein sequence alignment of the κKTxs from *D. whitei* and *P. cavimanus*, along with three other scorpion venom κKTxs identified from the literature showed sequence conservation in the signal peptide and the last 30 amino acids (Figure S2).

Finally, we also identified one CaTx (CaTx-1; Table 2) in *D. whitei*. CaTxs have been observed in other scorpion venoms and are thought to interact with calcium-voltage-gated, voltage-independent, and ligand-activated channels (Quintero-Hernández et al., 2013), including ryanodine receptors (Schwartz et al., 2009). CaTx-1 had a 26 amino acid long signal peptide and six cysteine residues. As CaTxs have shown promise for use in systematics of the non-Buthid scorpions (Santibáñez-López et al., 2018), we performed a multiple protein sequence alignment using CaTx-1 from *D. whitei* and the top nr and TSA database matches. As expected, this alignment displayed strong sequence conservation in the Toxin 27 Domain (scorpion calcine) region of the protein (Figure S3) via the NCBI conserved domain database search. All ion-channel toxins showed significant matches to previously described scorpion venom toxins in the TSA and/or nr databases. The αKTxs were responsible for the majority of the total toxin transcript abundance among ion-channel toxins in the transcriptomes (C0687 = 13.1%, C0689 = 16.3%) followed by the βKTx (C0687 = 1.3%, C0689 = 0.8%), κTX (C0687 = 0.9%, C0689 = 0.7%), and CaTx (C0687 = 0.03%, C0689 = 0.02%).

Although we identified several KTxs, we did not find any Na^+-channel toxins (NaTxs) in *D. whitei* venom glands. KTxs are widespread across scorpion venoms, but NaTxs are found in significantly higher diversities and abundances in those from the Buthidae family (Cid-Uribe et al., 2020), which includes almost all scorpions with medically relevant stings (Ward et al., 2018a). NaTxs have not been observed at significant expression levels in venom from other members of the Diplocentridae family (Grashof et al., 2019; Rojas-Azofeifa et al., 2019), although five putative NaTx transcripts were identified in *N. heirichonticus*. However, the significant expression of other ion-channel toxins, including KTxs, in venom from this family, indicates that KTxs likely play an important role in the overall function of Diplocentrid scorpion venoms.

3.5.3 La1-like peptides

La1-like peptides, named after the scorpion they were first discovered in (*Liocheles australasiae*; Miyashita et al., 2007), are peptides that range in length from 73–116 amino acids, have four disulfide bridges, and typically contain an SVWC, or Single von Willebrand factor type C, domain (Cid-Uribe et al., 2020). Although they have been identified in the venom of other scorpions, their function remains largely unknown. We found three La1 toxins in the transcriptome of both *D. whitei* individuals (Table 2), all of which were found in the proteome (La1-1, La1-2, La1-3) of at least one individual. However, La1-2 was only detected in the proteome of C0687. These toxins had signal peptides of 19–37 amino acids long, eight cysteine residues, and all contained a SVWC domain. Fur-

thermore, all toxins displayed database matches to previously identified scorpion venom toxins. La1 toxins were the fourth most abundant toxin class in the transcriptomes of both individuals (C0687 = 14.1%, C0689 = 16.2%; Figure 2). However, similar to the AMPs, La1 toxins were expressed at much lower abundances in the venom proteomes of both individuals (C0687 = 1.7%, C0689 = 2.6%).

3.5.4 Tryspin inhibitor-like peptides

We detected 14 trypsin inhibitor-like (TIL) peptides in the venom-gland transcriptome of *D. whitei* (Table 2). However, TIL-11 and TIL-12 were only detected in the transcriptome of individual CO687. As a toxin family, TILs were detected at very low abundances, contributing $> 0.3\%$ of total toxin transcript abundance in both individuals. All detected TILs had signal peptides of 16–24 amino acids longs, 10 cysteine residues, and evidence for a trypsin inhibitor-like cysteine-rich domain. Protease inhibitors are widespread in scorpion venom-gland transcriptomes, having been reported in all scorpion families with transcriptomic data (Cid-Uribe et al., 2020). Protease inhibitors, including TILs, may improve the effectiveness of venom by inhibiting prey and predator extracellular enzymes from degrading injected venom proteins (Hakim et al., 2016). TILs may also inhibit venom proteins stored in the venom glands to prevent toxins from acting on the host. However, without further analyses of scorpion venom TIL function in *D. whitei* the exact role that these toxins play is unclear.

3.5.5 Other low abundance toxins

We described 12 other putative *D. whitei* toxins only by their homology to other known, putative scorpion venom toxin classes (Table 2), including one carbonic anhydrase (CarbAn-1; 16 amino acid signal petide and 2 cysteine residues), one cysteine peptidase (CP-1; 18 amino acid signal peptide with 7 cysteine residues), one histidine phosphatase (HistP-1; 21 amino acid signal peptide with 6 cysteine residues), three insulin-like growth factor-binding proteins (IGFBP-1–3; 16–17 amino acid signal peptides with 10–12 cysteine residues), two Kunitz-type protease inhibitors (KUN-1 and KUN-2; 21 amino acid signal peptide with six and 12 cysteine residues, respectively), one lysozyme-C (LysC-1; 20 amino acid signal peptide with eight cysteine residues), one metalloprotease (MP-1; 20 amino acid signal peptide with 37 cysteine residues), two peptidase family M2 angiotensin converting enzymes (PeptidaseM2-1 and PeptidaseM2-2; 19–24 amino acid signal peptides and 11–14 cysteine residues), and one phospholipase A2 (PLA2-1; 20 amino acid signal peptide with 12 cysteine residues). Several of these low abundance putative toxins (*i.e.* HistP-1, IGFBP-1–3, and LysC-1) may represent non-toxic venom components or contaminants encountered during the venom extraction process.

We also identified four other low abundant proteins with database matches to non-toxic homologs (Table 2), including one biotinidase (Biotinidase-1; 18 amino acid signal peptide with 13 cysteine residues), one C-reactive protein (CReactive-1; 22 amino acid signal peptide with six cysteine residues), and one peptidylglycine alpha-hydroxylating

monooxygenase (PaHM-1; 21 amino acid signal peptide with eight cysteine residues), suggesting they may represent nontoxic components expressed in *D. whitei* venom glands.

Finally, we identified five uncharacterized venom proteins (VP-1, VP-2, VP-5, VP-6, and VP-8) at low abundance only in venom-gland transcriptomes (Table 2). VP-1 had an 18 amino acid signal peptide and three cysteine residues. VP-2 had a 21 amino acid signal peptide and zero cysteine residues. VP-5 contained a 16 amino acid long signal peptide, 7 cysteine residues, and evidence for a N-terminal nucleophile hydrolase domain. VP-6 and VP-8 each had eight cysteine residues and 22 and 18 amino signal peptides, respectively. Each of these VPs, except VP-5, had matches to previously identified scorpion venom proteins in either the nr and/or TSA databases, all of which were matches to venom proteins with no existing functional characterizations.

3.6 Venom proteomic abundances display strong agreement between individuals

Of the 25 toxins identified in the venom proteome of at least one individual *D. whitei*, 22 were observed in C0687, 20 were observed in C0689, and 17 were shared between both individuals (Table 2). Most scorpion venom characterizations that used high throughput venom proteomic analyses reported between 23 and 84 unique proteins (Cid-Uribe et al., 2020), although see Zhang et al. (2015), which reported 16 unique proteins in *Androctonus bicolor* venom. Our estimate for the number of unique proteomically confirmed venom toxins in *D. whitei* falls within this range. We also observed a strong agreement in the venom proteomic abundances of the 17 proteomically shared toxins between individuals (Spearman's rank correlation $\rho = 0.67$, Pearson's rank correlation coefficient $R = 0.71$, and $R^2 = 0.51$; Figure 3). The eight toxins responsible for presence-absence differences in venom proteomes were identified at relatively low abundances in the respective individuals (Table 3). Furthermore, our RP-HPLC analysis showed few major differences between venom chromatographic profiles (Figure 4), providing convincing evidence for a strong similarity between *D. whitei* venom proteomic profiles.

3.7 Venom gene transcript and proteomic abundances show weak agreement within individuals

We found weak correlations between mRNA transcript and venom proteomic abundances within both individual C0687 (Spearman's rank correlation $\rho = 0.43$, Pearson's rank correlation coefficient $R = 0.43$, and $R^2 = 0.18$; Figure 5) and C0689 (Spearman's rank correlation $\rho = 0.54$, Pearson's rank correlation coefficient $R = 0.57$, and $R^2 = 0.33$; Figure 5). This weak correlation between mRNA transcript and venom proteomic abundances was not surprising considering we found significant discrepancies between toxin class abundances in the transcriptome and the proteome, such as how αKTxs, AMPs, and La1s had much higher abundances in the transcriptome. This weak agreement between mRNA transcript and proteomic abundances has been observed in other scorpion

species, with some even showing less agreement (Rokyta and Ward, 2017; Ward et al., 2018b). Furthermore, as previously discussed, we may have been unable to easily detect small proteins in our venom proteomes (*e.g.* AMPs, La1, etc.), which could have contributed to an underestimation of the abundances of these small toxins in the proteome and, therefore, the weak agreement between transcript and protein abundances.

4 Conclusions

We generated the first high-throughput venom-gland transcriptomic and venom proteomic characterization for a harmless scorpion in the genus *Diplocentrus*. Of the 82 toxins identified in *D. whitei* venom, 57 were identified only by homology to known animal toxins while 25 were protoemically confirmed in at least one of the two individuals. We identified two novel ARSB toxins, which were either the most highly expressed or the second most highly expressed toxin families in the venom proteome of both *D. whitei* individuals. We also observed serine proteases and other enzymatic components (*i.e.* AChE, NUC, and Peroxidase) at high abundances in the venom of *D. whitei*, revealing a unique enzyme-rich scorpion venom. Our venom characterizations also revealed several scorpion venom toxin classes at high abundances in the venom-gland transcriptome that were not easily detected in venom proteomes (AMPs, NDBPs, ion-channel toxins, and La-1 like peptides), a trend that has been observed in venom characterizations on other scorpion species. Venom proteomic abundance comparisons showed a strong agreement between *D. whitei* venom proteomes, while venom-gland transcript and venom proteomic abundances showed a weaker agreement within individuals. Although scorpion venom characterization research has focused on those species considered medically significant, our identification of novel enzymatic toxins from *D. whitei* venom builds upon the work of previous studies that emphasize the importance of studying the venom of harmless scorpion species.

Acknowledgments

Funding for this work was provided by the National Science Foundation (NSF DEB 1638902). We thank Erich Hofmann, Rhett Rautsaw, and Jason Strickland for their assistance with collecting *D. whitei* specimens. We thank Carl Whittington (Florida State University Department of Biological Science) for his assistance with proteomics. We also thank Kenneth Wray for providing specimen photos. Finally, we thank Rakesh Singh and Cynthia Vied from the Florida State University College of Medicine's Translational Science Laboratory for their assistance with proteomics and transcriptomics, respectively.

References

Ahmed, M., W. Ahmad, N. Mushtaq, R. A. Khan, and M. R. C. Schetinger, 2021. Reptile venom acetylcholinesterases. Pp. 445–452, *in* Handbook of Venoms and Toxins of Reptiles. CRC Press.

Aitchison, J., 1986. The statistical analysis of compositional data. Chapman and Hall, London.

Almaaytah, A. and Q. Albalas, 2014. Scoprion venom peptides with no disulfide bridges. Peptides 51:35–45.

Almaaytah, A., S. Tarazi, N. Mhaidat, Q. Al-Balas, and T. L. Mukattash, 2013. Mauriporin, a novel cationic α-helical peptide with selective cytotoxic activity against prostate cancer cell lines from the venom of the scorpion *Androctonus mauritanicus*. Int. J. Pept. Res. Ther. 19:281–293.

Almeida, F., A. Pimenta, S. De Figueiredo, M. Santoro, M. Martin-Eauclaire, C. Diniz, and M. De Lima, 2002. Enzymes with gelatinolytic activity can be found in *Tityus bahiensis* and *Tityus serrulatus* venoms. Toxicon 40:1041–1045.

Alves, R. M., G. A. Antonucci, H. H. Paiva, A. C. O. Cintra, J. J. Franco, E. P. Mendonça-Franqueiro, D. J. Dorta, J. R. Giglio, J. C. Rosa, A. L. Fuly, et al., 2008. Evidence of caspase-mediated apoptosis induced by L-amino acid oxidase isolated from *Bothrops atrox* snake venom. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 151:542–550.

Astudillo, L., N. Therville, C. Colacios, B. Ségui, N. Andrieu-Abadie, and T. Levade, 2016. Glucosylceramidases and malignancies in mammals. Biochimie 125:267–280.

Batista, C., J. Martins, R. Restano-Cassulini, F. I. Coronas, F. Zamudio, R. Procópio, and L. Possani, 2018. Venom characterization of the Amazonian scorpion *Tityus metuendus*. Toxicon 143:51–58.

Bose, U., T. Wang, M. Zhao, C. Motti, M. Hall, and S. F. Cummins, 2017. Multiomics analysis of the giant triton snail salivary gland, a crown-of-thorns starfish predator. Sci. Rep. 7:1–14.

Brazón, J., B. Guerrero, G. D'Suze, C. Sevcik, and C. L. Arocha-Piñango, 2014. Fibrin (ogen) olytic enzymes in scorpion (*Tityus discrepans*) venom. Comp. Biochem. Physiol. B. Biochem. Mol. Bio. 168:62–69.

Carcamo-Noriega, E. N., S. Sathyamoorthi, S. Banerjee, E. Gnanamani, M. Mendoza-Trujillo, D. Mata-Espinosa, R. Hernández-Pando, J. I. Veytia-Bucheli, L. D. Possani, and R. N. Zare, 2019. 1, 4-benzoquinone antimicrobial agents against *Staphylococcus*

aureus and *Mycobacterium tuberculosis* derived from scorpion venom. Proc. Natl. Acad. Sci. U.S.A. 116:12642–12647.

Chippaux, J.-P. and M. Goyffon, 2008. Epidemiology of scorpionism: a global appraisal. Acta. Trop. 107:71–79.

Cid-Uribe, J. I., C. E. Santibáñez-López, E. P. Meneses, C. V. Batista, J. M. Jiménez-Vargas, E. Ortiz, and L. D. Possani, 2018. The diversity of venom components of the scorpion species *Paravaejovis schwenkmeyeri* (Scorpiones: Vaejovidae) revealed by transcriptome and proteome analyses. Toxicon 151:47–62.

Cid-Uribe, J. I., J. I. Veytia-Bucheli, T. Romero-Gutierrez, E. Ortiz, and L. D. Possani, 2020. Scorpion venomics: a 2019 overview. Expert Rev. Proteom. 17:67–83.

Colovic, M. B., D. Z. Krstic, T. D. Lazarevic-Pasti, A. M. Bondzic, and V. M. Vasic, 2013. Acetylcholinesterase inhibitors: pharmacology and toxicology. Curr. Neuropharmacol. 11:315–335.

Conde, R., F. Z. Zamudio, M. H. Rodríguez, and L. Possani, 2000. Scorpine, an anti-malaria and anti-bacterial agent purified from scorpion venom. FEBS Lett. 471:165–168.

i5K Consortium, 2013. The i5k initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J. Hered. 104:595–600.

Corner, E., Y. Leon, and R. Bulbrook, 1960. Steroid sulphatase, arylsulphatase and β-glucuronidase in marine invertebrates. J. Mar. Biolog. Assoc. 39:51–61.

De Graaf, D. C., M. Aerts, M. Brunain, C. A. Desjardins, F. J. Jacobs, J. H. Werren, and B. Devreese, 2010. Insights into the venom composition of the ectoparasitoid wasp *Nasonia vitripennis* from bioinformatic and proteomic studies. Insect Mol. Biol. 19:11–26.

Dhananjaya, B. L. and C. J. D'Souza, 2010. The pharmacological role of nucleotidases in snake venoms. Cell Biochem. Funct. 28:171–177.

Diego-García, E., Y. Abdel-Mottaleb, E. Schwartz, R. R. de la Vega, J. Tytgat, and L. Possani, 2008. Cytolytic and k+ channel blocking activities of β-ktx and scorpine-like peptides purified from scorpion venoms. Cell. Mol. Life Sci. 65:187–200.

Diego-García, E., S. Peigneur, E. Clynen, T. Marien, L. Czech, L. Schoofs, and J. Tytgat, 2012. Molecular diversity of the telson and venom components from *Pandinus cavimanus* (Scorpionidae latreille 1802): transcriptome, venomics and function. Proteomics 12:313–328.

Donaldson, G. P., W.-C. Chou, A. L. Manson, P. Rogov, T. Abeel, J. Bochicchio, D. Ciulla, A. Melnikov, P. B. Ernst, H. Chu, et al., 2020. Spatially distinct physiology of *Bacteroides fragilis* within the proximal colon of gnotobiotic mice. Nat. Microbiol. 5:746–756.

Duckert, P., S. Brunak, and N. Blom, 2004. Prediction of proprotein convertase cleavage sites. Protein Eng. Des. Sel. 17:107–112.

Ferreira, L., E. Alves, and O. Henriques, 1993. Peptide t, a novel bradykinin potentiator isolated from *Tityus serrulatus* scorpion venom. Toxicon 31:941–947.

Francke, O., 2019. Scorpions (Arachnida: Scorpiones) from the Cuatro Ciénegas basin. Pp. 53–59, *in* Animal Diversity and Biogeography of the Cuatro Ciénegas Basin. Springer.

Frobert, Y., C. Créminon, X. Cousin, M.-H. Rémy, J.-M. Chatel, S. Bon, C. Bon, and J. Grassi, 1997. Acetylcholinesterases from Elapidae snake venoms: biochemical, immunological and enzymatic characterization. Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol. 1339:253–267.

Furtmüller, P. G., M. Zederbauer, W. Jantschko, J. Helm, M. Bogner, C. Jakopitsch, and C. Obinger, 2006. Active site structure and catalytic mechanisms of human peroxidases. Arch. Biochem. Biophys. 445:199–213.

Gao, B., C. Tian, and S. Zhu, 2007. Inducible antibacterial response of scorpion venom gland. Peptides 28:2299–2305.

Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, 2005. Protein identification and analysis tools on the ExPASy server. Springer.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Grashof, D. G., H. M. Kerkkamp, S. Afonso, J. Archer, D. J. Harris, M. K. Richardson, F. J. Vonk, and A. van der Meijden, 2019. Transcriptome annotation and characterization of novel toxins in six scorpion species. BMC Genom. 20:1–10.

Guo, C., S. Liu, Y. Yao, Q. Zhang, and M.-Z. Sun, 2012. Past decade study of snake venom l-amino acid oxidase. Toxicon 60:302–311.

Hakim, M., T. Xiao-Peng, Y. Shi-Long, L. Qiu-Min, and L. Ren, 2016. Protease inhibitor in scorpion (*Mesobuthus eupeus*) venom prolongs the biological activities of the crude venom. Chin. J. Nat. Med. 14:607–614.

Han, J., D. You, X. Xu, W. Han, Y. Lu, R. Lai, and Q. Meng, 2008. An anticoagulant serine protease from the wasp venom of *Vespa magnifica*. Toxicon 51:914–922.

Harrison, P. L., M. A. Abdel-Rahman, K. Miller, and P. N. Strong, 2014. Antimicrobial peptides from scorpion venoms. Toxicon 88:115–137.

Holding, M. L., M. J. Margres, A. J. Mason, C. L. Parkinson, and D. R. Rokyta, 2018. Evaluating the performance of de novo assembly methods for venom-gland transcriptomics. Toxins 10:249.

Huang, J., S. Han, Q. Sun, Y. Zhao, J. Liu, X. Yuan, W. Mao, B. Peng, W. Liu, J. Yin, et al., 2017. Kv1. 3 channel blocker (ImKTx88) maintains blood–brain barrier in experimental autoimmune encephalomyelitis. Cell Biosci. 7:1–13.

Jiménez-Vargas, J. M., S. Ramírez-Carreto, G. Corzo, L. D. Possani, B. Becerril, and E. Ortiz, 2021. Structural and functional characterization of NDBP-4 family antimicrobial peptides from the scorpion *Mesomexovis variegatus*. Peptides 141:170553.

Kearse, M., R. Moir, A. Wilson, S. Stones-havas, S. Sturrock, S. Buxton, and A. Cooper, 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Khamtorn, P., P. Rungsa, N. Jangpromma, S. Klaynongsruang, J. Daduang, T. Tessiri, and S. Daduang, 2020. Partial proteomic analysis of brown widow spider (*Latrodectus geometricus*) venom to determine the biological activities. Toxicon, X 8:100062.

Kovacs, Z., I. Jung, and S. Gurzu, 2019. Arylsulfatases A and B: From normal tissues to malignant tumors. Pathol. Res. Pract. 215:152516.

Krueger, F., 2015. Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Buisulfite-Seq) libraries. URL `https://github.com/FelixKrueger/TrimGalore`.

Langmead, B. and S. L. Salzberg, 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 .

Li, W. and A. Godzik, 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Lourenço, W. R., 2014. A historical approach to scorpion studies with special reference to the 20th and 21st centuries. J. Venom. Anim. Toxins Incl. Trop. Dis. 20:1–9.

Lu, S., J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, et al., 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 48:D265–D268.

Marchler-Bauer, A., Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, et al., 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45:D200–D203.

Marchler-Bauer, A., M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, et al., 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43:D222–D226.

Marchler-Bauer, A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, et al., 2010. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39:D225–D229.

Matalon, R., B. Arbogast, and A. Dorfman, 1974. Deficiency of chondroitin sulfate n-acetylgalactosamine 4-sulfate sulfatase in maroteaux-lamy syndrome. Biochem. Biophys. Res. Commun. 61:1450–1457.

Meki, A.-R. M., A. Y. Nassar, and H. Rochat, 1995. A bradykinin-potentiating peptide (peptide k12) isolated from the venom of Egyptian scorpion *Buthus occitanus*. Peptides 16:1359–1365.

Menaldo, D. L., C. P. Bernardes, J. C. Pereira, D. S. Silveira, C. C. Mamede, L. Stanziola, F. de Oliveira, L. S. Pereira-Crott, L. H. Faccioli, and S. V. Sampaio, 2013. Effects of two serine proteases from *Bothrops pirajai* snake venom on the complement system and the inflammatory response. Int. Immunopharmacol. 15:764–771.

Miyashita, M., J. Otsuki, Y. Hanai, Y. Nakagawa, and H. Miyagawa, 2007. Characterization of peptide components in the venom of the scorpion *Liocheles australasiae* (Hemiscorpiidae). Toxicon 50:428–437.

Moerman, L., S. Bosteels, W. Noppe, J. Willems, E. Clynen, L. Schoofs, K. Thevissen, J. Tytgat, J. Van Eldere, J. Van Der Walt, et al., 2002. Antibacterial and antifungal properties of α-helical, cationic peptides in the venom of scorpions from southern Africa. Eur. J. Biochem. 269:4799–4810.

Moerman, L., F. Verdonck, J. Willems, J. Tytgat, and S. Bosteels, 2003. Antimicrobial peptides from scorpion venom induce Ca2+ signaling in HL-60 cells. Biochem. Biophys. Res. Commun. 311:90–97.

Nalbantsoy, A., B.-F. Hempel, D. Petras, P. Heiss, B. Göçmen, N. Iğci, M. Z. Yildiz, and R. D. Süssmuth, 2017. Combined venom profiling and cytotoxicity screening of the Radde's mountain viper (*Montivipera raddei*) and mount bulgar viper (*Montivipera bulgardaghica*) with potent cytotoxicity against human A549 lung carcinoma cells. Toxicon 135:71–83.

Nok, A. J., M. Abubakar, A. Adaudi, and E. Balogun, 2003. Aryl sulfatase from *Naja nigricolis* venom: Characterization and possible contribution in the pathology of snake poisoning. J. Biochem. Mol. Toxicol. 17:59–66.

de Oliveira, U. C., M. Y. Nishiyama Jr, M. B. V. Dos Santos, A. d. P. Santos-da Silva, H. d. M. Chalkidis, A. Souza-Imberg, D. M. Candido, N. Yamanouye, V. A. C. Dorce, and I. d. L. M. Junqueira-de Azevedo, 2018. Proteomic endorsed transcriptomic profiles of venom glands from *Tityus obscurus* and *T. serrulatus* scorpions. PLOS ONE 13:e0193739.

Peigneur, S., C. Sevcik, J. Tytgat, C. Castillo, and G. D'Suze, 2012. Subtype specificity interaction of bactridines with mammalian, insect and bacterial sodium channels under voltage clamp conditions. FEBS J. 279:4025–4038.

Pesentseva, M., V. Sova, A. S. Sil, A. Kicha, A. S. Sil, T. Haertlé, T. Zvyagintseva, et al., 2012. A new arylsulfatase from the marine mollusk *Turbo chrysostomus*. Chem. Nat. Compd. 48:853–859.

Peters, C., B. Schmidt, W. Rommerskirch, K. Rupp, M. Zühlsdorf, M. Vingron, H. Meyer, R. Pohlmann, and K. von Figura, 1990. Phylogenetic conservation of arylsulfatases. cDNA cloning and expression of human arylsulfatase B. J. Biol. Chem. 265:3374–3381.

Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen, 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8:785–786.

Quintero-Hernández, V., J. M. Jiménez-Vargas, G. B. Gurrola, H. H. Valdivia, and L. D. Possani, 2013. Scorpion venom components that affect ion-channels function. Toxicon 76:328–342.

Quintero-Hernández, V., S. Ramírez-Carreto, M. T. Romero-Gutiérrez, L. L. Valdez-Velázquez, B. Becerril, L. D. Possani, and E. Ortiz, 2015. Transcriptome analysis of scorpion species belonging to the *Vaejovis* genus. PLOS ONE 10:e0117188.

Rice, P., I. Longden, and A. Bleasby, 2000. EMBOSS: the European molecular biology open software suite.

Rojas-Azofeifa, D., M. Sasa, B. Lomonte, E. Diego-García, N. Ortiz, F. Bonilla, R. Murillo, J. Tytgat, and C. Díaz, 2019. Biochemical characterization of the venom

of central american scorpion *Didymocentrus krausi* francke, 1978 (Diplocentridae) and its toxic effects in vivo and in vitro. Comp. Biochem. Physiol. C Toxicol. Pharmacol. 217:54–67.

Rokyta, D. R., A. R. Lemmon, M. J. Margres, and K. Aronow, 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genom. 13:312.

Rokyta, D. R. and M. J. Ward, 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.

Romero-Gutiérrez, M. T., C. E. Santibáñez-López, J. M. Jiménez-Vargas, C. V. F. Batista, E. Ortiz, and L. D. Possani, 2018. Transcriptomic and proteomic analyses reveal the diversity of venom components from the Vaejovid scorpion *Serradigitus gertschi*. Toxins 10:359.

Romero-Gutierrez, T., E. Peguero-Sanchez, M. A. Cevallos, C. V. Batista, E. Ortiz, and L. D. Possani, 2017. A deeper examination of *Thorellius atrox* scorpion venom components with omic technologies. Toxins 9:399.

Rui, Y., P. Fang, L. Hui, C. Zhijian, L. Wenxin, M. Xin, and J. Dahe, 2005. Functional analysis of a gene encoding a chlorotoxin–like peptide derived from scorpion toxin. Chin. J. Biochem. Mol. Bio. 21:19–23.

Runager, K., J. J. Enghild, and G. K. Klintworth, 2008. Focus on molecules: Transforming growth factor beta induced protein (TGFBIp). Exp. Eye Res. 87:298.

Santibáñez-López, C. E., J. I. Cid-Uribe, C. V. Batista, E. Ortiz, and L. D. Possani, 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., J. I. Cid-Uribe, F. Z. Zamudio, C. V. Batista, E. Ortiz, and L. D. Possani, 2017. Venom gland transcriptomic and venom proteomic analyses of the scorpion *Megacormus gertschi* Díaz-Najera, 1966 (Scorpiones: Euscorpiidae: Megacorminae). Toxicon 133:95–109.

Santibáñez-López, C. E., O. F. Francke, L. Prendini, et al., 2014. Phylogeny of the north american scorpion genus *Diplocentrus* peters, 1861 (Scorpiones: Diplocentridae) based on morphology, nuclear and mitochondrial DNA. Arthropod Syst. Phylogeny 72:257–279.

Santibáñez-López, C. E., R. Kriebel, J. A. Ballesteros, N. Rush, Z. Witter, J. Williams, D. A. Janies, and P. P. Sharma, 2018. Integration of phylogenomics and molecular modeling reveals lineage-specific diversification of toxins in scorpions. PeerJ 6:e5902.

Schülein, R., C. Westendorf, G. Krause, and W. Rosenthal, 2012. Functional significance of cleavable signal peptides of G protein-coupled receptors. Eur. J. Cell Biol. 91:294–299.

Schwartz, E. F., E. M. Capes, E. Diego-Garcïa, F. Z. Zamudio, O. Fuentes, L. D. Possani, and H. H. Valdivia, 2009. Characterization of hadrucalcin, a peptide from *Hadrurus gertschi* scorpion venom with pharmacological activity on ryanodine receptors. Br. J. Pharmacol. 157:392–403.

Seppey, M., M. Manni, and E. M. Zdobnov, 2019. Busco: assessing genome assembly and annotation completeness. Pp. 227–245, *in* Gene prediction. Springer.

Sievers, F. and D. G. Higgins, 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 27:135–145.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.

Sissom, W. D. and V. Fet, 2000. Family Diplocentridae Karsch, 1880. Catalog of the Scorpions of the World (1758–1998) (Fet, V., WD Sissom, G. Lowe, & ME Braunwalder, eds.). The New York Entomological Society, New York Pp. 329–354.

So, W. L., T. C. Leung, W. Nong, W. G. Bendena, S. M. Ngai, and J. H. Hui, 2021. Transcriptomic and proteomic analyses of venom glands from scorpions *Liocheles australasiae*, *Mesobuthus martensii*, and *Scorpio maurus palmatus*. Peptides 146:170643.

Song, Y.-M., X.-X. Tang, X.-G. Chen, B.-B. Gao, E. Gao, L. Bai, and X.-R. Lv, 2005. Effects of scorpion venom bioactive polypolypeptides on platelet aggregation and thrombosis and plasma 6-keto-PG F1α and TXB2 in rabbits and rats. Toxicon 46:230–235.

Srinivasan, K. N., V. Sivaraja, I. Huys, T. Sasaki, B. Cheng, T. K. S. Kumar, K. Sato, J. Tytgat, C. Yu, B. C. C. San, et al., 2002. κ-hefutoxin1, a novel toxin from the scorpionheterometrus fulvipes with unique structure and function: Importance of the functional diad in potassium channel selectivity. J. Biol. Chem. 277:30040–30047.

Stockwell, S. A. and A. S. Baldwin, 2001. A new species of Diplocentrus (Scorpiones, Diplocentridae) from Texas. J. Arachnol. 29:304–311.

Stockwell, S. A. and J. A. Nilsson, 1987. A new species of Diplocentrus Peters from Texas (Scorpiones, Diplocentridae). J. Arachnol. Pp. 151–156.

Sugahara, K., T. Mikami, T. Uyama, S. Mizuguchi, K. Nomura, and H. Kitagawa, 2003. Recent advances in the structural biology of chondroitin sulfate and dermatan sulfate. Curr. Opin. Struct. Biol. 13:612–620.

Tanner, M. R., R. B. Tajhya, R. Huq, E. J. Gehrmann, K. E. Rodarte, M. A. Atik, R. S. Norton, M. W. Pennington, and C. Beeton, 2017. Prolonged immunomodulation in inflammatory arthritis using the selective Kv1. 3 channel blocker HsTX1 [R14A] and its PEGylated analog. Clin. Immunol. 180:45–57.

Tasoulis, T. and G. K. Isbister, 2017. A review and database of snake venom proteomes. Toxins 9:290.

Torres-Larios, A., G. B. Gurrola, F. Z. Zamudio, and L. D. Possani, 2000. Hadrurin, a new antimicrobial peptide from the venom of the scorpion *Hadrurus aztecus*. Eur. J. Biochem. 267:5023–5031.

Trentini, M. M., R. C. das Neves, B. d. P. O. Santos, R. A. DaSilva, A. C. Souza, M. R. Mortari, E. F. Schwartz, A. Kipnis, and A. P. Junqueira-Kipnis, 2017. Non-disulfide-bridge peptide 5.5 from the scorpion *Hadrurus gertschi* inhibits the growth of *Mycobacterium abscessus* subsp. *massiliense*. Front. Microbiol. 8:273.

Uawonggul, N., S. Thammasirirak, A. Chaveerach, T. Arkaravichien, W. Bunya-tratchata, W. Ruangjirachuporn, P. Jearranaiprepame, T. Nakamura, M. Matsuda, M. Kobayashi, et al., 2007. Purification and characterization of heteroscorpine-1 (hs-1) toxin from *Heterometrus laoticus* scorpion venom. Toxicon 49:19–29.

Undheim, E. A. B., A. Jones, K. R. Clauser, J. W. Holland, S. S. Pineda, G. F. King, and B. G. Fry, 2014. Clawing through evolution: toxin diversification and convergence in the ancient lineage Chilopoda (centipedes). Mol. Bio. Evol. 31:2124–2148.

Valdez-Velázquez, L. L., J. Cid-Uribe, M. T. Romero-Gutierrez, T. Olamendi-Portugal, J. M. Jimenez-Vargas, and L. D. Possani, 2020. Transcriptomic and proteomic analyses of the venom and venom glands of *Centruroides hirsutipalpus*, a dangerous scorpion from Mexico. Toxicon 179:21–32.

Valverde, P., T. Kawai, and M. A. Taubman, 2004. Selective blockade of voltage-gated potassium channels reduces inflammatory bone resorption in experimental periodontal disease. J. Bone Miner. Res. 19:155–164.

Veiga, S. S., R. B. da Silveira, J. L. Dreyfuss, J. Haoach, A. M. Pereira, O. C. Mangili, and W. Gremski, 2000. Identification of high molecular weight serine-proteases in *Loxosceles intermedia* (brown spider) venom. Toxicon 38:825–839.

Veiseh, M., P. Gabikian, S.-B. Bahrami, O. Veiseh, M. Zhang, R. C. Hackman, A. C. Ravanpay, M. R. Stroud, Y. Kusuma, S. J. Hansen, et al., 2007. Tumor paint: a chlorotoxin: Cy5. 5 bioconjugate for intraoperative visualization of cancer foci. Cancer Res. 67:6882–6888.

Vizcaíno, J. A., A. Csordas, N. Del-Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, et al., 2016. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 44:D447–D456.

Ward, M. J., S. A. Ellsworth, and G. S. Nystrom, 2018a. A global accounting of medically significant scorpions: Epidemiology, major toxins, and comparative resources in harmless counterparts. Toxicon 151:137–155.

Ward, M. J., S. A. Ellsworth, and D. R. Rokyta, 2018b. 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. Toxicon 142:14–29.

Ward, M. J. and D. R. Rokyta, 2018. Venom-gland transcriptomics and venom proteomics of the giant Florida blue centipede, *Scolopendra viridis*. Toxicon 152:121–136.

Ween, M. P., M. K. Oehler, and C. Ricciardelli, 2012. Transforming growth factor-beta-induced protein (TGFBI)/(βig-H3): a matrix protein with dual functions in ovarian cancer. Int. J. Mol. Sci. 13:10461–10477.

Wei, J.-F., H.-W. Yang, X.-L. Wei, L.-Y. Qiao, W.-Y. Wang, and S.-H. He, 2009. Purification, characterization and biological activities of the L-amino acid oxidase from *Bungarus fasciatus* snake venom. Toxicon 54:262–271.

Willems, J., L. Moerman, S. Bosteels, E. Bruyneel, F. Ryniers, and F. Verdonck, 2004. Parabutoporin—an antibiotic peptide from scorpion venom—can both induce activation and inhibition of granulocyte cell functions. Peptides 25:1079–1084.

Yamada, M., D. M. Miller, M. Lowe, C. Rowe, D. Wood, H. P. Soyer, K. Byrnes-Blake, J. Parrish-Novak, L. Ishak, J. M. Olson, et al., 2021. A first-in-human study of BLZ-100 (tozuleristide) demonstrates tolerability and safety in skin cancer patients. Contemp. Clin. Trials Commun. 23:100830.

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

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

Zhang, L., W. Shi, X.-C. Zeng, F. Ge, M. Yang, Y. Nie, A. Bao, S. Wu, and E. Guoji, 2015. Unique diversity of the venom peptides from the scorpion *Androctonus bicolor* revealed by transcriptomic and proteomic analysis. J. Proteom. 128:231–250.

Zhang, L. and W. Wu, 2008. Isolation and characterization of ACTX-6: a cytotoxic L-amino acid oxidase from *Agkistrodon acutus* snake venom. Nat. Prod. Res. 22:554–563.

Zhou, L., T. Feng, S. Xu, F. Gao, T. T. Lam, Q. Wang, T. Wu, H. Huang, L. Zhan, L. Li, et al., 2022. ggmsa: a visual exploration tool for multiple sequence alignment and associated data. Brief. Bioinform. .

Figures and Figure Legends

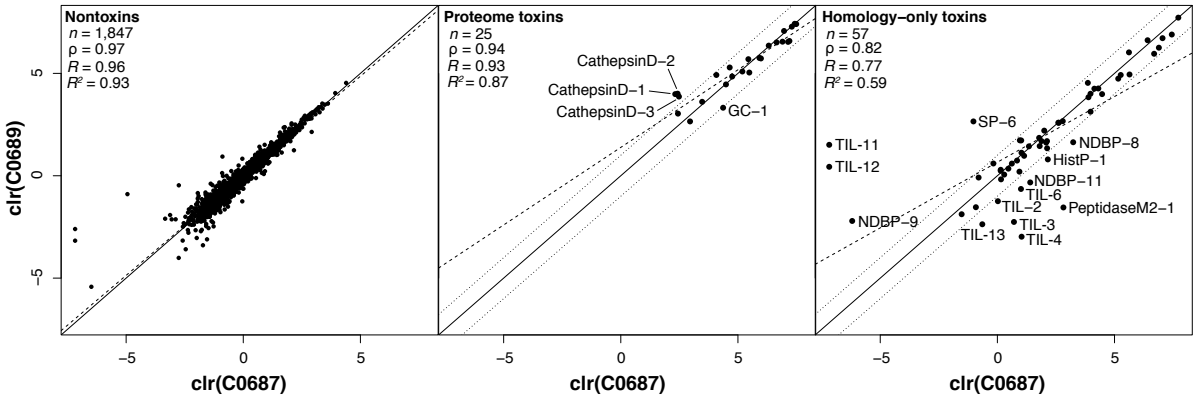

Figure 1. A venom-gland nontoxin (left), proteomically confirmed toxin (middle), and homology-only toxin (right) transcript abundance comparison between *D. whitei* individuals (C0687 and C0689) all showed strong agreement. Solid lines represent a correlation coefficient of one, while the longer dashed lines represent the lines of best fit. Labeled transcripts are those that fall outside of the 99$^{\text{th}}$ percentile of differences (region between the shorter dashed lines) between the two nontoxin measures and represent toxins with unusually different expression levels between individuals. Abbreviations: clr—centered logratio transformation, GC—glucosylceramidase, HistP—histidine phosphatase, n—number of transcripts, NDBP—non-disulfide bridge peptide, ρ—Spearman's rank correlation coefficient, R—Pearson's correlation coefficient, R^2—coefficient of determination, SP—serine protease, and TIL—trypsin inhibitor-like.

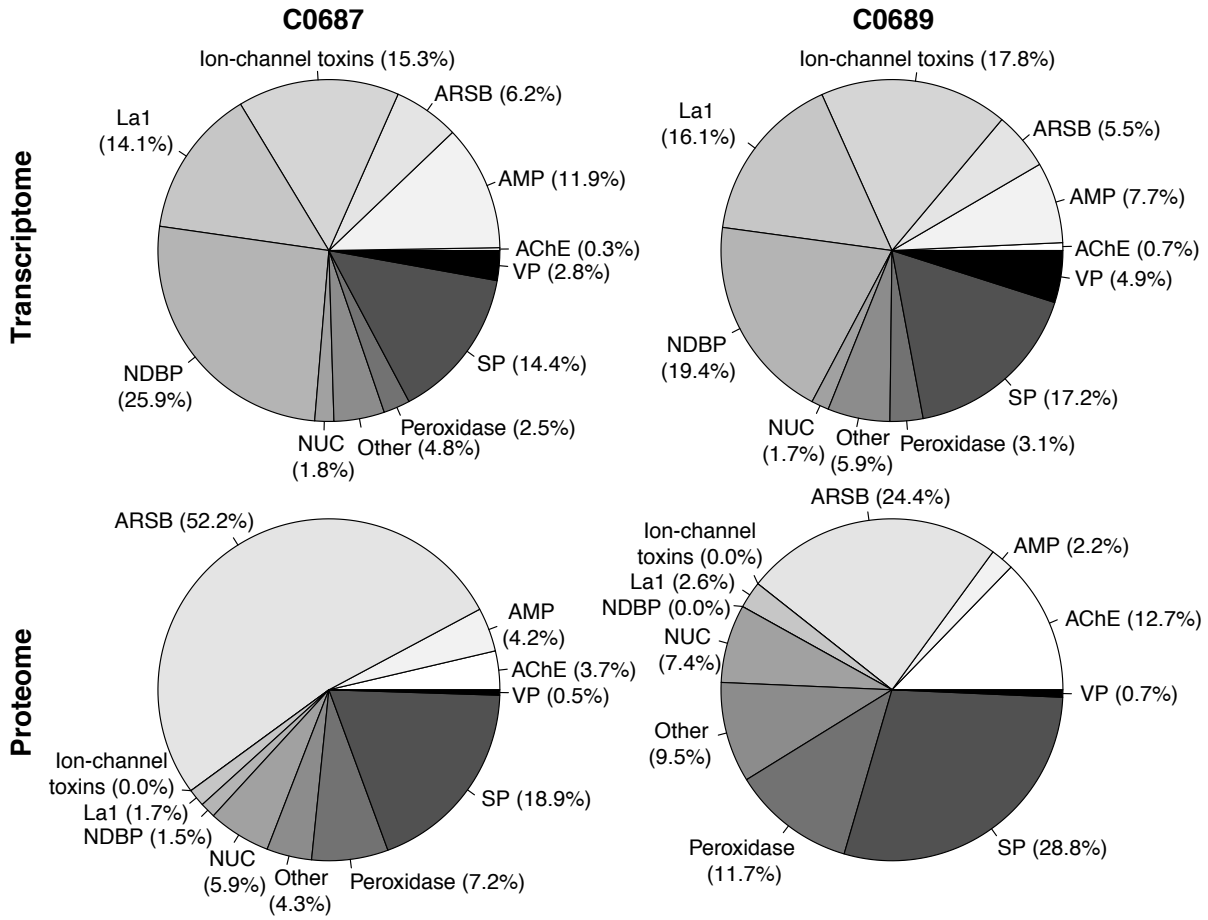

Figure 2. *Diplocentrus whitei* venom-gland transcriptomic (top) and venom proteomic (bottom) abundances for the most abundant toxin classes in individual C0687 (left) and C0689 (right) reveal an enzyme-rich venom dominated by novel ARSBs and serine proteases. We observed a weak agreement between toxin class abundances in the transcriptome and proteome within individuals. Abbreviations: AChE—acetylcholinesterase, αKTx—α-potassium channel toxin, AMP—antimicrobial peptide, ARSB—arylsulfatase B, βKTx—β-potassium channel toxin, CarbAn—Carbonic anhydrase, CaTx—calcium channel toxin, CP—cysteine peptidase, CReactive—C-reactive protein, CRISP—cysteine-rich secretory protein, GC—glucosylceramidase, GPCR—G protein-coupled receptor, HistP—histidine phosphatase, HYAL—hyaluronidase, IGFBP—insulin-like growth factor-binding protein, κKTx—κ-potassium channel toxin, KUN—kunitz-type protease inhibitor, La1—La1-like peptide, LysC—lysozyme C, MP—metalloprotease, NDBP—non-disulfide bridge peptide, NUC—nucleotidase, PaHM—peptidylglycine alpha-hydroxylating monooxygenase, PeptidaseM2—peptidase family M2 angiotensin converting enzyme, PLA2—phospholipase A2, SP—serine protease, TGFBI–transforming growth factor-beta-induced protein, TIL—trypsin inhibitor-like, VP—uncharacterized venom protein.

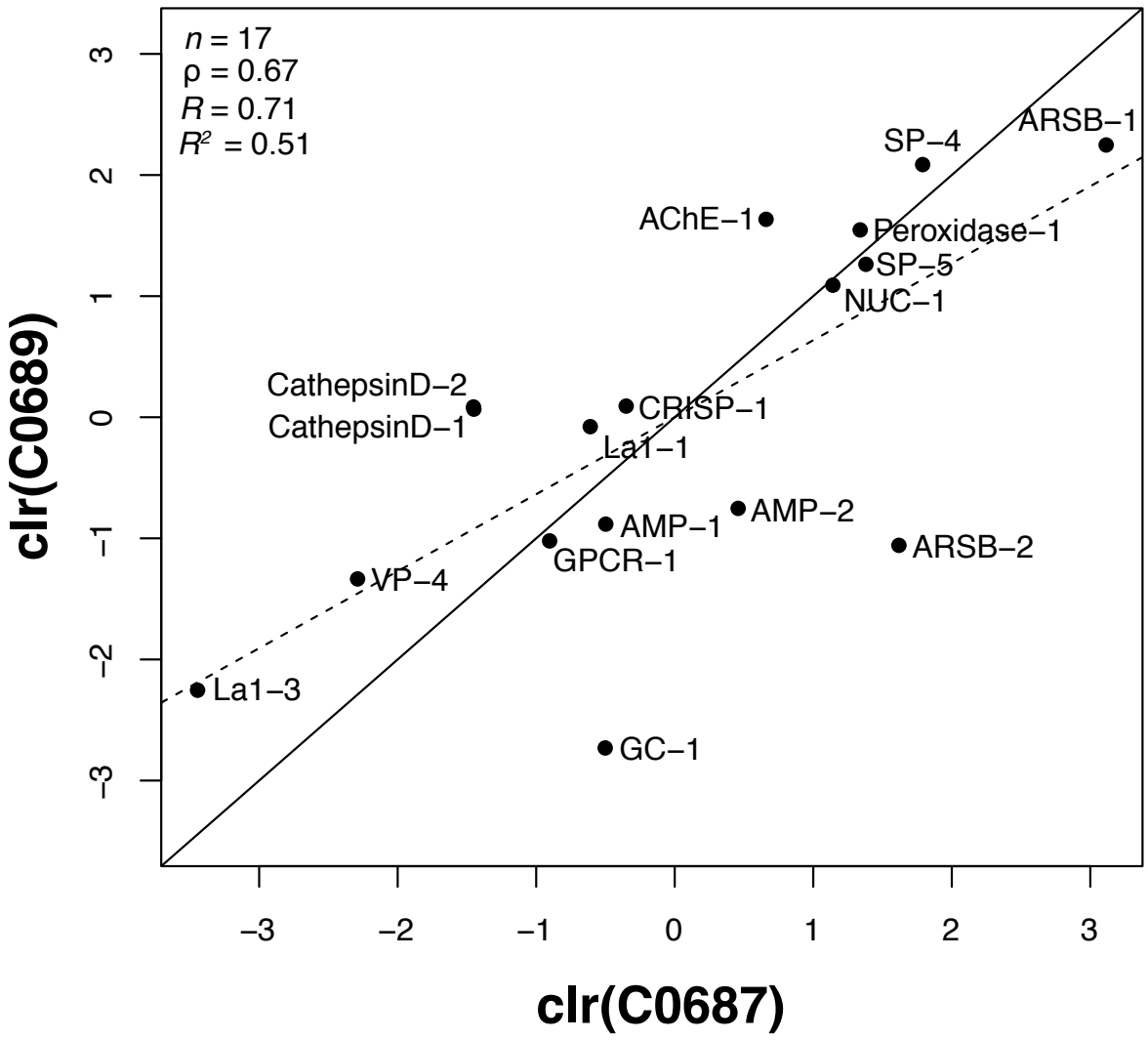

Figure 3. Venom proteomic abundance comparison between *D. whitei* individuals (C0687 and C0689) showed a strong agreement for proteins detected in the venom of both individuals. Table 3 shows presence/absence differences between the two *D. whitei* proteomes. Solid line represents a correlation coefficient of one, while the dashed line represents the line of best fit. 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, AChE—acetylcholinesterase, AMP—antimicrobial peptide, ARSB—arylsulfatase B, CRISP—cysteine-rich secretory protein, GC—Glucosylceramidase, GPCR—G protein-coupled receptor, La1—La1-like peptide, NUC—nucleotidase, SP—serine protease, VP—uncharacterized venom protein.

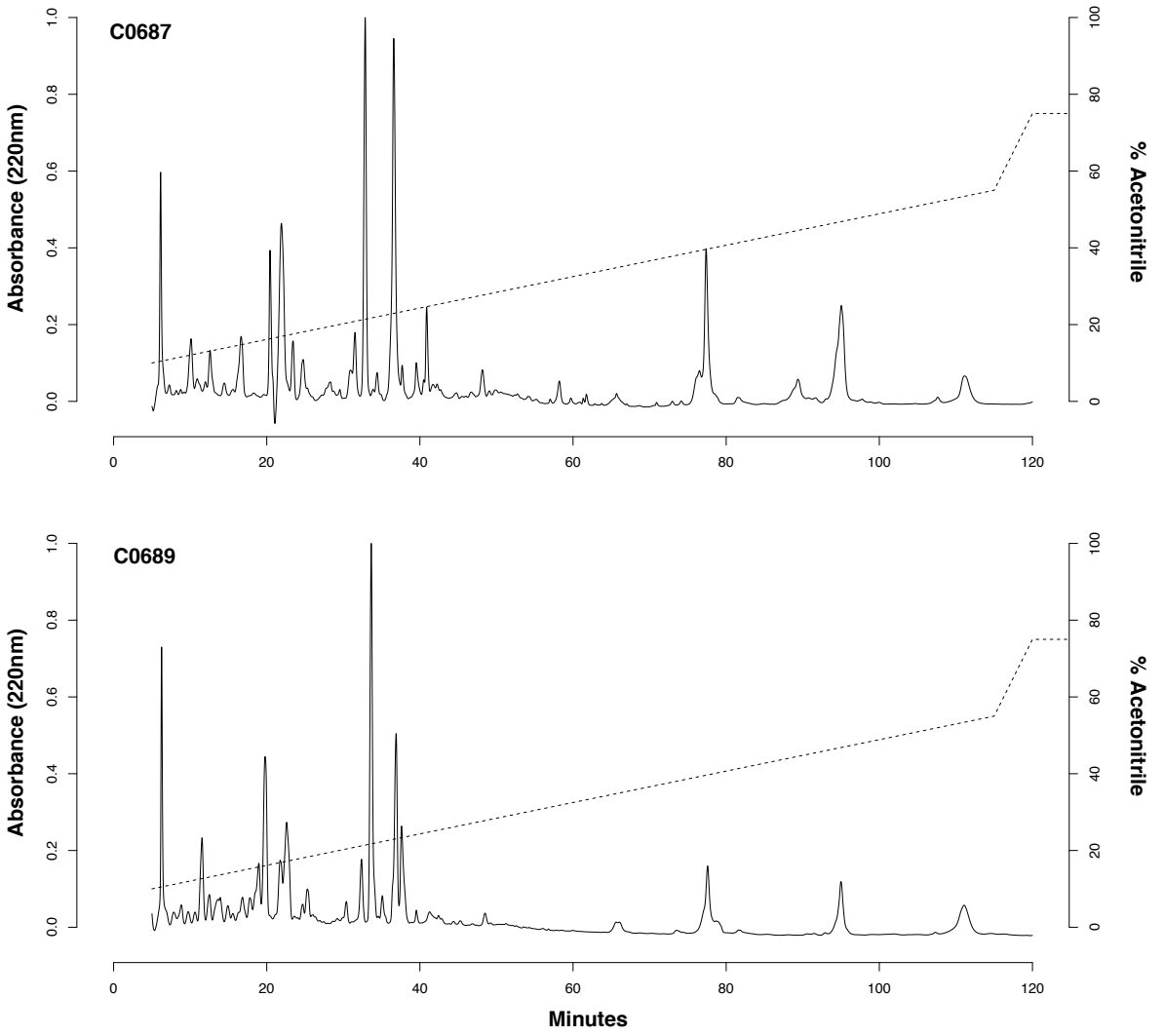

Figure 4. *Diplocentrus whitei* venom RP-HPLC profiles for individuals C0687 (top) and C0689 (bottom) show the relatively similar venoms between individuals, with most venom components eluting after shorter retention times. RP-HPLC chromatographic peaks were normalized to the absorbance of the highest peak in in each profile.

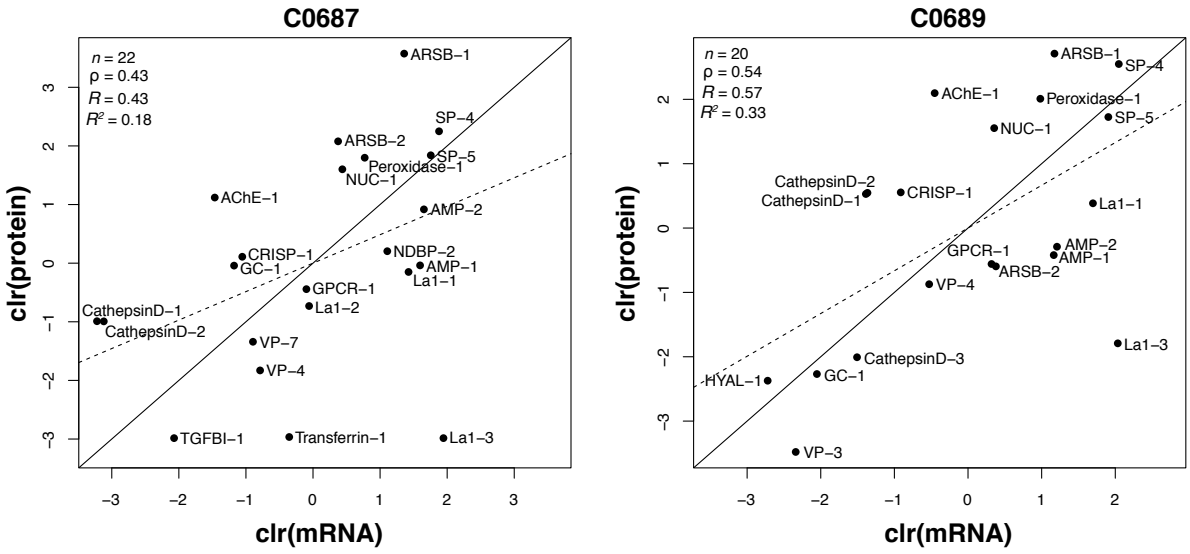

Figure 5. Venom-gland transcript and venom proteomic abundances showed agreement within *D. whitei* individuals C0687 (left) and C0689 (right). Solid lines represent a correlation coefficient of one, while the dashed lines represent the line of best fit. Abbreviations: clr—centered logratio transformation, n—number of proteins, ρ—Spearman's rank correlation coefficient, R—Pearson's correlation coefficient, R^2—coefficient of determination, AChE—acetylcholinesterase, AMP—antimicrobial peptide, ARSB—arylsulfatase B, CRISP—cysteine-rich secretory protein, GC—Glucosylceramidase, GPCR—G protein-coupled receptor, HYAL—hyaluronidase, La1—La1-like peptide, NUC—nucleotidase, SP—serine protease, TGFBI–transforming growth factor-beta-induced protein, VP—uncharacterized venom protein.

Tables and Table Legends

Table 1. Proteins identified via LC-MS/MS in *D. whitei* venom.

Protein	Unique peptide count (C0687)	%MS/MS coverage (C0687)	Unique peptide count (C0689)	%MS/MS coverage (C0689)
AChE-1	18	37.50%	9	21.60%
AMP-1	2	24.00%	1	12.50%
AMP-2	3	34.00%	2	25.50%
ARSB-1	19	43.60%	20	43.60%
ARSB-2	2	2.37%	11	30.10%
CathepsinD-1	4	10.20%	2	7.35%
CathepsinD-2	4	10.20%	2	11.80%
CathepsinD-3	1	7.91%	–	–
ChromodomHeliDNAbindProt-1	1	1.55%	–	–
CRISP-1	6	15.90%	2	5.53%
DblhomologyDHdomain-1	–	–	1	1.37%
GC-1	2	6.18%	3	12.00%
GlutamineNADsynthetase-1	–	–	1	2.80%
GPCR-1	3	12.10%	3	12.10%
HYAL-1	1	5.03%	–	–
La1-1	5	38.80%	3	21.40%
La1-2	–	–	1	16.30%
La1-3	1	9.48%	1	9.48%
MediatorRNApolyIITranscri-1	1	0.00%	–	–
NDBP-2	–	–	1	14.80%
NUC-1	17	47.50%	14	42.60%
Peroxidase-1	14	26.30%	10	18.80%
SP-4	5	22.00%	4	22.00%
SP-5	13	33.80%	9	25.30%
TGFBI-1	–	–	1	2.90%
Transferrin-1	–	–	1	3.68%
VacuolarProtSorting-1	–	–	1	1.36%
VP-3	1	8.55%	–	–
VP-4	1	6.79%	1	6.79%
VP-7	–	–	1	10.90%

Unique peptide counts and percent MS/MS coverage for each protein identified in *D. whitei* venom were extracted using Scaffold (version 5.1.2). Abbreviations: AChE—acetylcholinesterase, AMP—antimicrobial peptide, ARSB—arylsulfatase B, ChromodomHeliDNAbindProt—chromodomain helicase DNA binding protein, CRISP—cysteine-rich secretory protein, DblhomologyDHdomain—Dbl homology domain protein, GC—Glucosylceramidase, GlutamineNADsynthetase—glutamine-dependent NAD(+) synthetase, GPCR—G protein-coupled receptor, HYAL—hyaluronidase, La1—La1-like peptide, MediatorRNApolyIITranscri—mediator of RNA polymerase II transcription, NUC—nucleotidase, SP—serine protease, TGFBI–transforming growth factor-beta-induced protein, VacuolarProtSorting—vacuolar protein sorting-associated protein, and VP—uncharacterized venom protein.

Table 2. Putative toxins identified in the venom-gland transcriptome and venom proteome of *Diplocentrus whitei*.

Toxin	Signal Peptide (aa)	Precursor (aa)	Cysteine Residues	MW (kDa)	C0687 TPM	C0689 TPM	C0687 fmol	C0689 fmol
AChE-1	19	546	6	61	2323.83	6220.36	821.75	4173.68
αKTx-1	25	61	8	–	88874.29	102381.89	–	–
αKTx-2	25	61	8	–	23943.00	33942.62	–	–
αKTx-3	25	60	6	–	102.52	255.40	–	–
αKTx-4	32	75	8	–	33.15	82.54	–	–
AMP-1	19	96	6	11	49607.17	31333.38	258.36	336.86
AMP-2	19	106	6	12	52532.32	32749.28	671.73	383.61
AMP-3	20	58	6	–	328.35	173.81	–	–
ARSB-1	19	546	7	62	39134.79	31642.65	9571.76	7710.69
ARSB-2	19	548	7	62	14614.10	14283.14	2144.08	282.72
Biotinidase-1	18	501	13	–	529.66	604.65	–	–
βKTx-1	21	79	6	–	11066.17	6374.79	–	–
CarbAn-1	16	287	2	–	2923.60	3218.37	–	–
CathepsinD-1	16	381	7	42	401.61	2442.89	99.69	870.39
CathepsinD-2	16	381	7	41	444.36	2496.63	99.42	886.20
CathepsinD-3	22	392	6	43	470.33	2166.75	–	68.87
CaTx-1	26	66	6	–	239.19	193.01	–	–
CP-1	18	333	7	–	541.33	603.83	–	–
CReactive-1	22	237	6	–	2098.23	1034.73	–	–
CRISP-1	20	452	16	51	3502.64	3923.42	299.32	892.75
GC-1	20	518	6	59	3099.26	1256.52	257.30	53.06
GPCR-1	21	348	24	40	9111.18	13460.57	172.24	293.29
HistP-1	21	381	6	–	338.57	100.72	–	–
HYAL-1	20	398	13	46	755.97	643.83	–	47.83
IGFBP-1	17	97	12	–	45.27	57.78	–	–
IGFBP-2	16	105	12	–	233.57	288.58	–	–
IGFBP-3	17	109	10	–	45.32	60.57	–	–
κKTx-1	17	68	4	–	7606.11	6220.95	–	–
KUN-1	21	77	6	–	8.55	6.92	–	–
KUN-2	21	150	12	–	52.60	47.99	–	–
La1-1	24	98	8	11	41769.00	53362.42	230.96	753.40
La1-2	19	104	8	11	9484.44	7006.32	129.18	–
La1-3	37	116	8	13	70296.73	74769.92	13.57	85.48
LysC-1	20	144	8	–	2091.91	2484.91	–	–
MP-1	20	617	37	–	108.79	254.40	–	–
NDBP-1	22	66	0	–	67277.90	44883.17	–	–
NDBP-2	22	88	0	9	30429.28	30528.32	329.08	–
NDBP-3	22	82	0	–	38742.98	23598.12	–	–
NDBP-4	22	84	0	–	45318.74	37548.51	–	–
NDBP-5	22	65	0	–	31633.52	17536.82	–	–
NDBP-6	22	82	0	–	2468.35	3188.28	–	–
NDBP-7	22	87	0	–	6837.51	5195.65	–	–
NDBP-8	22	82	0	–	999.52	232.83	–	–
NDBP-9	22	71	0	–	0.08	4.98	–	–
NDBP-10	22	81	0	–	15.70	9.75	–	–
NDBP-11	22	84	0	–	159.76	32.72	–	–
NUC-1	17	573	7	64	15572.34	13954.52	1330.93	2422.56
PaHM-1	21	345	8	–	640.76	652.33	–	–
PeptidaseM2-1	19	620	14	–	658.79	9.63	–	–
PeptidaseM2-2	24	629	11	–	3431.65	2443.25	–	–
Peroxidase-1	21	666	33	76	21729.01	26114.88	1620.83	3823.35
PLA2-1	20	226	12	–	326.78	246.07	–	–
SP-1	24	286	10	–	109.96	139.66	–	–
SP-2	16	425	18	–	151.10	190.48	–	–
SP-3	20	281	10	–	90.97	95.17	–	–
SP-4	25	282	10	31	65824.23	75812.45	2544.78	6558.77
SP-5	17	281	9	31	58192.71	65770.81	1689.70	2878.78
SP-6	18	300	9	–	14.14	644.11	–	–

TGFBI-1	20	758	16	86	1267.07	1678.12	13.57	–
TIL-1	19	96	10	–	72.95	82.15	–	–
TIL-2	22	91	10	–	40.16	13.08	–	–
TIL-3	20	88	10	–	79.78	4.75	–	–
TIL-4	20	89	10	–	110.56	2.31	–	–
TIL-5	16	89	10	–	290.51	410.71	–	–
TIL-6	24	88	10	–	107.58	23.76	–	–
TIL-7	24	87	10	–	123.06	119.98	–	–
TIL-8	24	89	10	–	17.64	41.53	–	–
TIL-9	22	90	10	–	100.77	55.37	–	–
TIL-10	24	88	10	–	62.86	64.16	–	–
TIL-11	20	89	10	–	–	205.36	–	–
TIL-12	22	90	10	–	–	70.18	–	–
TIL-13	24	92	10	–	20.66	4.24	–	–
TIL-14	23	88	10	–	45.93	38.02	–	–
Transferrin-1	19	707	28	77	7053.43	7373.41	13.83	–
VP-1	18	108	3	–	10810.94	18769.69	–	–
VP-2	21	135	0	–	1918.62	2095.26	–	–
VP-3	18	117	9	13	450.98	942.40	–	15.81
VP-4	19	162	12	19	4568.14	5770.86	43.06	214.24
VP-5	16	416	7	–	258.05	246.87	–	–
VP-6	22	102	8	–	315.39	229.29	–	–
VP-7	26	137	6	16	4107.13	8975.55	70.19	–
VP-8	18	117	8	–	1869.31	4230.78	–	–

Cysteine residues were determined using ExPASy ProtParam (Gasteiger et al., 2005) with signal peptides excluded. Molecular weight (MW) mass spectrometry estimates are provided for proteomically confirmed toxins. Abbreviations: AChE—acetylcholinesterase, αKTx—α-potassium channel toxin, AMP—antimicrobial peptide, ARSB—arylsulfatase B, βKTx—β-potassium channel toxin, CarbAn—Carbonic anhydrase, CaTx—calcium channel toxin, CP—cysteine peptidase, CReactive—C-reactive protein, CRISP—cysteine-rich secretory protein, GC—Glucosylceramidase, GPCR—G protein-coupled receptor, HistP—histidine phosphatase, HYAL—hyaluronidase, IGFBP—insulin-like growth factor-binding protein, κKTx—κ-potassium channel toxin, KUN—Kunitz-type protease inhibitor, La1—La1-like peptide, LysC—lysozyme C, MP—metalloprotease, NDBP—non-disulfide bridge peptide, NUC—nucleotidase, PaHM—peptidylglycine alpha-hydroxylating monooxygenase, PeptidaseM2—Peptidase family M2 angiotensin converting enzyme, PLA2—phospholipase A2, SP—serine protease, TGFBI–transforming growth factor-beta-induced protein, TIL—trypsin inhibitor-like peptide, VP—uncharacterized venom protein.

Table 3. Presence/absence differences between venom proteomes.

	C0687			C0689			Average	
Protein	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	C0687	C0689
CathepsinD-3	–	–	–	1.2674	–	2.5237	–	1.2637
HYAL-1	–	–	–	–	0.9782	1.6825	–	0.8869
La1-2	3.4586	3.0117	2.6397	–	–	–	3.0367	–
NDBP-2	8.07	9.0352	6.1593	–	–	–	7.7548	–
TGFBI-1	–	–	0.8799	–	–	–	0.2933	–
Transferrin-1	–	1.0039	–	–	–	–	0.3346	–
VP-3	–	–	–	–	–	0.84125	–	0.2804
VP-7	1.1529	1.0039	2.6397	–	–	–	1.5988	–

Quantities provided in fmol. Abbreviations: HYAL—hyaluronidase, La1—La1-like peptide, NDBP—non-disulfide bridge peptide, TGFBI–transforming growth factor-beta-induced protein, VP—uncharacterized venom protein.

Supplemental Figures and Supplemental Figure Legends

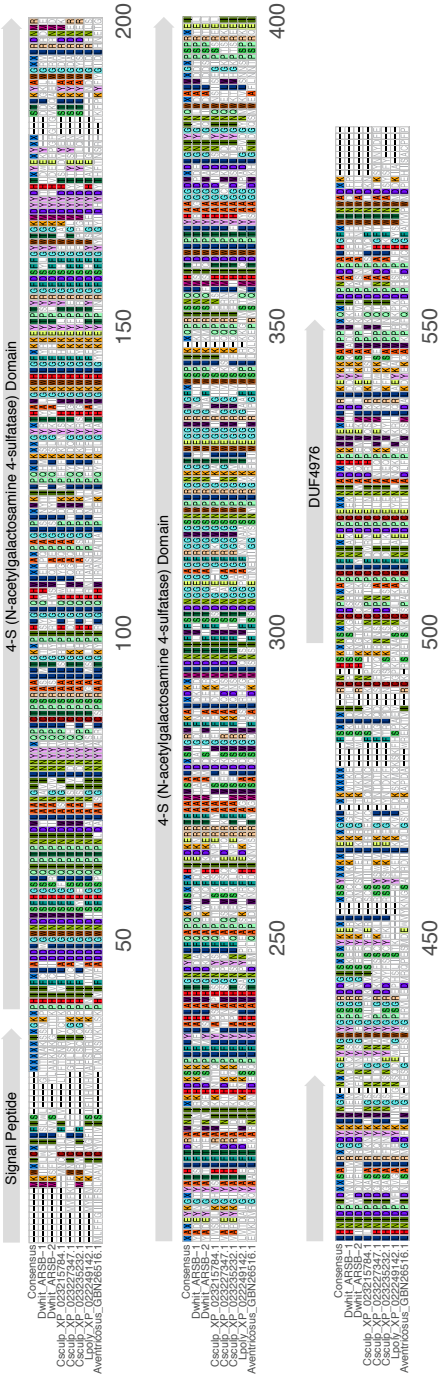

Figure S1. Multiple sequence alignment of putative ARSB toxins from *D. whitei* venom and their top NCBI protein BLAST matches. All ARSB sequences contained a signal peptide and evidence for a 4-S (N-acetylgalactosamine 4-sulfatase) domain, but only those from *D. whitei*, the Atlantic horseshoe crab (*Limulus polyphemus*), and the orb weaving spider (*Araneus ventricosus*), showed evidence for a DUF4976. However, *C. sculpturatus* ARSBs displayed a high degree of sequence homology in the DUF4976 region. Colors represent individual amino acid identities. Abbreviations: ARSB—Arylsulfatase B, Aventricosus—*Araneus ventricosus*,Csculp—*Centruroides sculpturatus*, Dwhit—*Diplocentrus whitei*, and Lpoly—*Limulus polyphemus*.

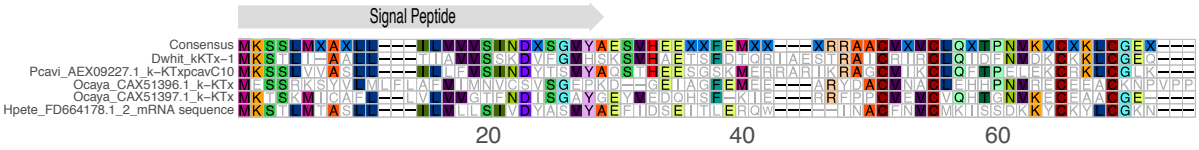

Figure S2. Multiple sequence alignment of putative κKTx toxins from *D. whitei* venom, the top NCBI protein BLAST match, and other published κKTxs identified from scorpion venoms. All κKTx sequences contained a signal peptide, but no additional protein domains. Colors represent individual amino acid identities. Abbreviations: Dwhit—*Diplocentrus whitei*, Hpete—*Heterometrus petersii*, Ocaya—*Opisthacanthus cayaporum*, and Pcavi—*Pandimus cavimanus*.

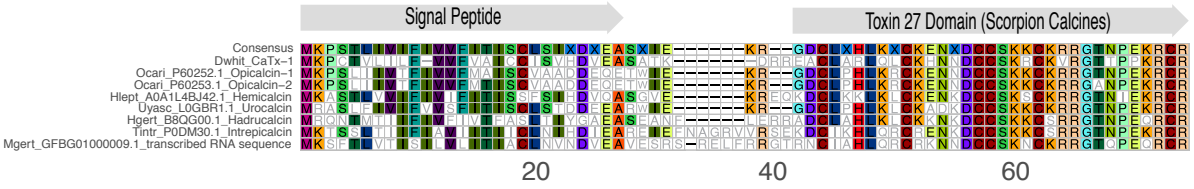

Figure S3. Multiple sequence alignment of the putative CaTx toxin from *D. whitei* venom and the top BLAST matches to the TSA and nr databases. All CaTx sequences contained a signal peptide and displayed clear evidence for a Toxin 27 (scorpion calcine) domain. Colors represent individual amino acid identities. Abbreviations: CaTx—Calcium channel toxin (scorpion calcine), Dwhit—*Diplocentrus whitei*, Hgert—*Hadrurus gertschi*, Hlept—*Hemiscorpius lepturus*, Mgert—*Megacormus gertschi*, Ocari—*Opistophthalmus carinatus*, Tintr— *Thorellius intrepidus*, and Uyasc— *Urodacus yaschenkoi*.