Toxicon 216 (2022) 92–106

Contents lists available at ScienceDirect

Toxicon

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

Origins, genomic structure and copy number variation of snake venom myotoxins

Siddharth S. Gopalan [a], Blair W. Perry [a,b], Drew R. Schield [c], Cara F. Smith [d,e], Stephen P. Mackessy [d], Todd A. Castoe [a,*]

[a] Department of Biology, 501 S. Nedderman Dr., The University of Texas Arlington, Arlington, TX, 76019, USA
[b] School of Biological Sciences, Washington State University, Pullman, WA, 99164, USA
[c] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, 80309, USA
[d] School of Biological Sciences, 501 20th Street, University of Northern Colorado, Greeley, CO, 80639, USA
[e] Department of Biochemistry and Molecular Biology, 12801 East 17th Avenue, University of Colorado Denver, Aurora, CO, 80045, USA

ARTICLE INFO

Handling Editor: Dr. Ray Norton

Keywords:
β-defensin
Iso-seq
Hi-C proximity Ligation sequencing
crotamine
lncRNA
venom regulation

ABSTRACT

Crotamine, myotoxin a and homologs are short peptides that often comprise major fractions of rattlesnake venoms and have been extensively studied for their bioactive properties. These toxins are thought to be important for rapidly immobilizing mammalian prey and are implicated in serious, and sometimes fatal, responses to envenomation in humans. While high quality reference genomes for multiple venomous snakes are available, the loci that encode myotoxins have not been successfully assembled in any existing genome assembly. Here, we integrate new and existing genomic and transcriptomic data from the Prairie Rattlesnake (*Crotalus viridis* viridis) to reconstruct, characterize, and infer the chromosomal locations of myotoxin-encoding loci. We integrate long-read transcriptomics (Pacific Bioscience's Iso-Seq) and short-read RNA-seq to infer gene sequence diversity and characterize patterns of myotoxin and paralogous β-defensin expression across multiple tissues. We also identify two long non-coding RNA sequences which both encode functional myotoxins, demonstrating a newly discovered source of venom coding sequence diversity. We also integrate long-range mate-pair chromatin contact data and linked-read sequencing to infer the structure and chromosomal locations of the three myotoxin-like loci. Further, we conclude that the venom-associated myotoxin is located on chromosome 1 and is adjacent to non-venom paralogs. Consistent with this locus contributing to venom composition, we find evidence that the promoter of this gene is selectively open in venom gland tissue and contains transcription factor binding sites implicated in broad *trans*-regulatory pathways that regulate snake venoms. This study provides the best genomic reconstruction of myotoxin loci to date and raises questions about the physiological roles and interplay between myotoxin and related genes, as well as the genomic origins of snake venom variation.

1. Introduction

Identifying the sequence, structure, and genomic locations of genes that encode snake venom proteins can provide valuable insight for reproducing these proteins *in vivo*, for understanding *cis*-regulatory sequences that regulate their expression, and for understanding the genetic variation and origins of these loci. The accumulation of increasingly high-quality reference genomes for venomous snakes has substantially expanded such knowledge (Margres et al., 2021; Schield et al., 2019; Suryamohan et al., 2020), although the genic structure and chromosomal location of some critical components of snake venoms remain poorly resolved. Because many loci that make up snake venom are derived from multi-gene families, which often occur in tandemly duplicated arrays (Hargreaves et al., 2014; Schield et al., 2019; Suryamohan et al., 2020; Wong and Belov, 2012), even highly contiguous genomes fail to assemble and annotate completely all venom-encoding loci and their relevant *cis*-regulatory sequences. Several confounding factors, including transposable element insertions within and around venom families (Cardoso et al., 2010; Dowell et al., 2018; de LM Junqueira-de-Azevedo and Ho, 2002), and short coding regions that can be associated with heterozygous structural variation, pose additional challenges to their assembly.

* Corresponding author. Department of Biology, 501 S. Nedderman Dr., University of Texas at Arlington, Arlington, TX, 76010, USA.
E-mail address: todd.castoe@uta.edu (T.A. Castoe).

https://doi.org/10.1016/j.toxicon.2022.06.014
Received 26 April 2022; Received in revised form 21 June 2022; Accepted 27 June 2022
Available online 9 July 2022

It is important to point out that the term 'myotoxin' has been used in the literature to describe multiple unrelated snake venom proteins. These include polypeptides from snake venoms with similar effects on smooth muscle, such as phospholipase A_2s from *Bothrops* and other crotaline snakes (Gutiérrez and Lomonte, 1995; Lomonte et al., 1990; Maraganore et al., 1984; Yoshizumi et al., 1990), and cardiotoxins from *Naja* (Fletcher et al., 1996; Lachumanan et al., 1998; Ownby et al., 1993), although the sequences themselves are unrelated to rattlesnake myotoxins. Here, we use the term 'myotoxin' exclusively to refer to the short, 42 to 45 residues long, positively charged, cysteine-rich peptides found in rattlesnake venoms. To date, no myotoxin loci have been unambiguously identified in an assembled genome (Margres et al., 2021; Rádis-Baptista et al., 2003; Schield et al., 2019).

Myotoxins interact with voltage-gated potassium channels and sarcoplasmic reticulum calcium pumps to induce both short-and long-term paralysis by inhibiting hind leg movement, as well as causing tissue necrosis through non-enzymatic processes, which together rapidly immobilize prey (Mebs and Ownby, 1990; Ownby, 1998; Rizzi et al., 2007; Utaisincharoen et al., 1991; Yount et al., 2009). Because of its capacity to permeabilize and cross cell membranes (Hayashi et al., 2008; Rádis-Baptista and Kerkis, 2011), myotoxins are frequently studied for their pleiotropic pharmacological and bioactive properties, including as anti-cancer therapeutics (Campeiro et al., 2018; Hayashi et al., 2012; Kerkis et al., 2014), trans-membrane molecule transporter (de Carvalho Porta et al., 2020), and broad-spectrum antimicrobial (Costa et al., 2014; Oguiura et al., 2011; Passero et al., 2007). In human envenomation, myotoxins are likely responsible for spastic muscle contractions and fasciculations (Keyler et al., 2020; Mackessy et al., 2003). Rattlesnake myotoxin expression and the degree to which it comprises a major fraction of venom can vary substantially within and between species (Bober et al., 1988; Hofmann et al., 2018; Margres et al., 2017; Rokyta et al., 2011, 2012; Schield et al., 2019). This includes being either present or undetectable in venom protein profiles (Boldrini-França et al., 2010; Schenberg, 1959; Straight et al., 1991). Myotoxin content also exhibits ontogenetic shifts in some lineages, with expression generally being higher in adults (Colis-Torres et al., 2021; Durban et al., 2017; Hofmann et al., 2018; Saviola et al., 2015), and myotoxin-encoding gene copy-number polymorphism has been reported in *C. d. terrificus* and *C. adamanteus* (Margres et al., 2017; Oguiura et al., 2009).

Most current knowledge of myotoxin sequence diversity comes from homologous peptides identified from the venom of several rattlesnake species including *C. durissus* (Laure, 1975), *C. viridis* (Cameron and Tu, 1977; Griffin and Aird, 1990), *C. oreganus* (Bieber et al., 1987; Maeda et al., 1978), *C. adamanteus* (Samejima et al., 1991), and *C. horridus* (Allen et al., 1996). To date, the only description of the genic architecture of any myotoxin is that of crotamine from *C. d. terrificus*, which is approximately 1.8 kb in length and contains three exons, separated by a long and a short intron (Oguiura et al., 2005; Rádis-Baptista et al., 2003). Exon 1 primarily encodes a signal peptide, while the mature toxin is encoded by exons 2 and 3. Accordingly, exon 1 is typically highly conserved, while exons 2 and 3 exhibit the most variation across species (Oguiura et al., 2005). The mature, secreted peptide is between 42 and 45 residues, making myotoxins among the smallest of all snake venom peptides (Almeida et al., 2017).

Rattlesnake myotoxins are closely related, both structurally and phylogenetically, to β-defensins – a vertebrate gene family of cell penetrating peptides that are a conserved component of the innate immune system (Rádis-Baptista and Kerkis, 2011; Shafee et al., 2016; Wong and Belov, 2012; Xiao et al., 2004). The structure of crotamine has demonstrated that, despite low sequence similarity with mammalian β-defensins, three-dimensional conservation indicates shared functions that involves interaction with and destabilization of cell membrane phospholipids (Coronado et al., 2013; Whittington et al., 2008; Yount et al., 2009). The presence of multiple myotoxin variants and copy number variation in some rattlesnake species (Margres et al., 2017; Oguiura et al., 2009) is consistent with expectations that rattlesnake myotoxin arrays also share a similar tandemly duplicated structure to that of β-defensins (Schutte et al., 2002; Xiao et al., 2004; Zou et al., 2007) and other major snake venom gene families.

Ambiguous inferences of myotoxin identity have previously made characterizing the genes that encode them, their chromosomal location, and other features of their architecture a challenge. An early study used fluorescence *in situ* hybridization (FISH) of probes designed from Neotropical Rattlesnake (*C. d. terrificus*) crotamine sequences and identified a single putative locus on chromosome 2 (Rádis-Baptista et al., 2003). Recently, a chromosome-level assembly of *C. v. viridis* (hereafter referred to as CroVir3.0; NCBI: GCA_003400415.2; Schield et al., 2019) identified several major venom gene clusters, but failed to find an unambiguous and complete myotoxin locus. Myotoxin loci were similarly absent from the chromosome-level Tiger Rattlesnake (*C. tigris*) assembly (NCBI: GCF_016545835; Margres et al., 2021). In both assemblies, the genomic location of myotoxin was inferred to be on chromosome 1. In either case, inferences of the chromosomal location of the myotoxin locus derived from FISH and genomic studies are contradictory, suggesting two competing chromosomal locations. It has therefore remained a largely open question where the myotoxin genes are located in viper genomes, and how their genomic architecture may resemble, or differ from, that observed in other major gene families.

Here, we focus on myotoxins in *C. v. viridis*, a species for which myotoxins are the single most abundant venom component, at least in populations that have been previously studied (Saviola et al., 2015; Schield et al., 2019). Extensive genomic resources are available for this species, making it a valuable model for studying snake venom in a variety of contexts (Perry et al., 2020, 2022; Schield et al., 2019, 2020). Here, we conduct analyses of existing and newly generated genomic datasets, along with detailed assessment of unassembled genomic read data, to resolve the genomic sequence of myotoxin-encoding loci and to characterize the structure and chromosomal location of these loci in the *C. v. viridis* genome. We also incorporate data from Pacific Biosciences' long-read isoform sequencing (hereafter 'Iso-Seq') and Illumina short read RNA-seq to estimate transcript abundance across venom gland and non-venom gland tissues, as well as Dovetail Genomics' Omni-C chromatin conformation sequencing to infer the location of myotoxin in the genome. Despite being absent from prior genome assemblies, we find evidence that multiple myotoxin-encoding loci are present in the *C. v. viridis* genome, and infer that these loci are located on chromosomes 1, 2 and 4. We also identify key regulatory features of the myotoxin gene based on chromatin accessibility data and identify long non-coding RNAs (lncRNA) related to myotoxins, and conduct phylogenetic analyses to understand the relationship between newly described β-defensin-like genes from *C. v. viridis* (Cvv BDLs) and myotoxin. Finally, we conduct analyses to infer copy number variation at the myotoxin locus using population sampling of re-sequenced genomes. This detailed characterization of the genomic structure, alternative myotoxin-encoding genes, regulatory architecture, and copy number variation of myotoxins provides valuable and diverse insight into this enigmatic toxin gene family.

2. Methods and materials

2.1. Long read Iso-Seq data generation and analysis

All tissues utilized in this study were sampled in accordance with protocol 2004D-SM-S-23 (S.P Mackessy) approved by the University of Northern Colorado Institutional Animal Care and Use Committee and under scientific collecting permits from Colorado Parks and Wildlife (21HP0974 to S.P Mackessy). We sampled right venom gland tissue from an adult female *C. v. viridis* from the same population as the CroVir3.0 genome animal (Weld Co., Colorado; see Schield et al., 2019). Prior to extraction, tissue was snap-frozen in liquid nitrogen for storage at −80 °C. Total RNA was extracted from the sample using Trizol reagent (Life Technologies, Carlsbad, CA, USA, No. 15596–026) and was

sequenced on a PacBio Sequel II System (Novogene, CA, USA). Raw long reads were quality filtered using the Pacific Bioscience's SMRT Link v10.1 Continuous Long Read Iso-Seq workflow using default settings. We used the reference-free collapsing method CD-HIT-EST (Huang et al., 2010) using default settings (90% threshold).

Myotoxin a containing sequences were identified from Iso-Seq data using blast (Altschul et al., 1997). To assess whether identified myotoxins play a functional role in both the reference (Schield et al., 2019) and Iso-Seq transcriptomes (i.e., are a major component of venom), one day post-extraction venom gland reads were mapped to both transcriptomes using kallisto v0.46.2 (Bray et al., 2016). In both datasets, the highest expressed transcript were myotoxin a sequences from their respective transcriptomes (Supplementary Fig. S1). For additional details, see the Supplementary Methods.

2.2. Generation of 10x genomics linked-read assembly

Linked-read assembly data was generated in a separate study (Schield et al., In Press) and was used here to identify scaffolds containing myotoxin exons. In brief, liver tissue from an adult female C. v. viridis from the same population as the CroVir3.0 genome animal (Weld Co., Colorado) was sampled and snap-frozen in liquid nitrogen for storage prior to DNA extraction. High molecular-weight DNA was extracted and a 10x Genomics Chromium library was prepared for linked read sequencing on an Illumina NovaSeq 6000 using 150 bp paired-end reads. The final assembly contained 148,816 scaffolds and had a scaffold N50 of 37.83 kb. Using the reference myotoxin transcript as a blast query (see Section 2.1), a single 1.3 kb myotoxin containing scaffold was identified using blast that contained exons 1 and 2 identical to the reference transcript but lacking exon 3. This scaffold was used for the extension protocol described below.

2.3. Extension of the myotoxin gene

Reads from four previously generated short-insert paired-end libraries (Schield et al., 2019) were quality trimmed using Trimmomatic v0.39 with the settings LEADING:20 TRAILING:20 MINLEN:32 AVGQUAL:30 (Bolger et al., 2014). The bbduk.sh and reformat.sh scripts from the BBtools suite (https://sourceforge.net/projects/bbmap/) were used to capture reads *in silico*. PriceTI (Ruby et al., 2013) was then used to perform sequence extension. The resulting 2076 bp sequence, hereafter called the 'myotoxin 10x-extended scaffold' (Fig. 1A; Supplementary

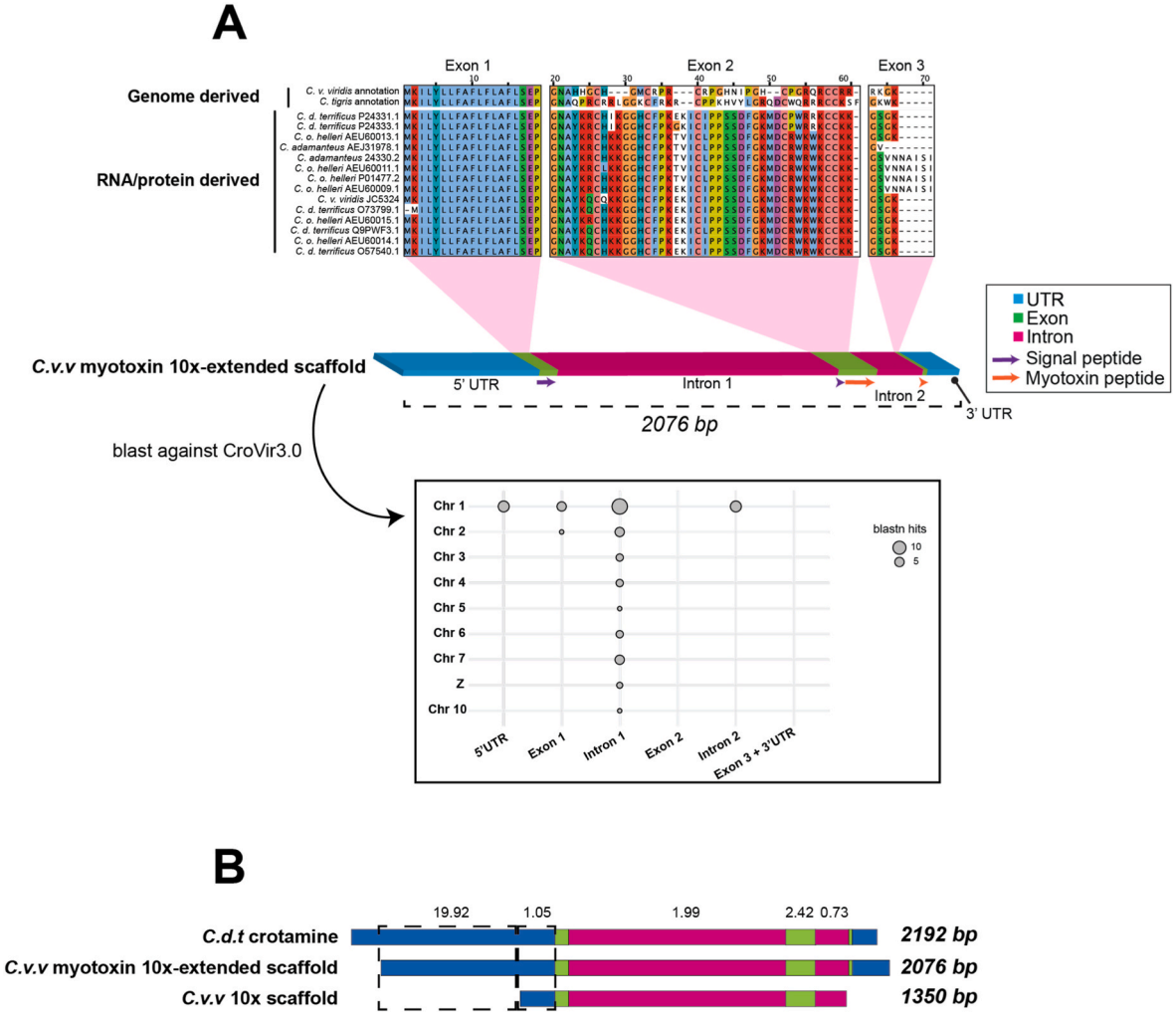

Fig. 1. Myotoxin is not present in current rattlesnake assemblies. A) A comparison of myotoxin coding sequences from Crotalus species indicates that annotated sequences from the genomes of *C. tigris* and *C. v. viridis* are not myotoxins, but rather closely related sequences. Blastn of myotoxin subsequences finds several regions of homology on chromosome 1, and other regions throughout the CroVir3.0 with homology to intron 1. B) A comparison of three independently derived inferences of myotoxin loci from a truncated 10x Genomics scaffold containing exons 1 and 2, but not 3, the paired-end read extended *C. v. viridis* myotoxin locus, and *C. d. terrificus* crotamine. Numbers above regions of the alignment indicate average pairwise Kimura-adjusted nucleotide distances, represented as differences per 100 nucleotides. Because the 5′ UTR of the 10x Genomics scaffold was truncated, two separate distance comparisons were calculated. Where distances are not noted, no nucleotide differences exist across sequences.

Fig. S2B), was aligned to the *C. d. terrificus* crotamine gene (NCBI: AF223946.1) and inspected for major inconsistencies (Fig. 1B). Because of its similarity to both the *C. d. terrificus* sequence and myotoxin Iso-Seq model 1 from the Iso-Seq set, which was the highest expressed transcript in one day post-extraction RNA-seq reads (Supplementary Fig. S1), we use the term 'venom-associated myotoxin' to refer to this sequence.

2.4. Analysis of short-read mRNA-seq for multiple tissues

RNA-seq libraries from multiple *C. v. viridis* tissues were previously generated (Schield et al., 2019, BioProject PRJNA477004) and were used here to explore the expression of myotoxin models across a diverse set of tissues. The myotoxin 10x-extended scaffold described above, as well as myotoxin Iso-Seq models 2 and 3, were added to the existing CroVir3.0 genome and annotation as they represent distinct loci in the genome. RNA-seq reads were mapped to the genome using STAR v.2.7.8a (Dobin et al., 2013) and reads mapped to individual features were counted using featureCounts v.2.0.1 (Liao et al., 2014). All counts were normalized using DESeq2 v1.34.0 (Love et al., 2014), and expression heatmaps were created using pheatmap (Kolde, 2015) in R v4.1.2 (R Core Team, 2021). BAM files were then filtered using the 'NH:1' flag to retain only uniquely mapped reads and BAM density plots were created using the karyoploteR package (Gel and Serra, 2017) with splice junctions inferred from STAR.

2.5. Integration of existing and new long-range contact-based mate-pair data

Here, we used Hi-C, Chicago and Omni-C libraries (Supplementary Table S2) to identify chromosomal contact points of myotoxin in the CroVir3.0 assembly. Reads from all libraries were trimmed using Trimmomatic (Bolger et al., 2014) with settings described for analyses above. Reads matching 31-mers of myotoxin exon two were captured from these datasets following the protocols described above. Matching reads were then mapped to the genome using bwa-mem v0.7.17 with the option -5SP -T0 to skip mate rescue and pairing (Li, 2013). Read density was visualized in R using karyoploteR. Cvv BDL exons were masked from the CroVir3.0 *C. v. viridis* genome assembly using bedtools maskfasta (Quinlan and Hall, 2010).

Using the Omni-C data, a more formal approach was used to determine contact frequency between the myotoxin 10x-extended scaffold (see section 2.3) and the CroVir3.0 genome assembly. For more details on the methods used, see the Supplementary Methods.

2.6. Identifying and characterizing Cvv BDLs in the C. viridis genome assembly

In brief, blastn (Altschul et al., 1997) and blat (Kent, 2002) were used to identify sequences homologous to myotoxin in the *C. v. viridis* genome using the reference myotoxin transcript sequence (Schield et al., 2019) as the initial query. We next used Hmmer v.3.3.2 (Finn et al., 2011) to search a six-frame translation of the *C. v. viridis* genome using the Pfam β-defensin profile (PF00711) and a custom protein profile generated from the above described BDLs. β-defensin nucleotide sequences from colubrids (de Oliveira et al., 2018) and other vipers (Correa and Oguiura, 2013), as well as myotoxin sequences from rattlesnake species (Rádis-Baptista et al., 1999; 2003; Rokyta et al., 2011; Supplementary Table S1) were aligned to Cvv BDLs 1–6. A phylogeny was then estimated using IQ-TREE (Nguyen et al., 2015). For additional details, see the Supplementary Methods.

2.7. Characterization of myotoxin-like RNAs

To identify open reading frames in myotoxin Iso-Seq models 2 and 3, we used ORFfinder (www.ncbi.nlm.nih.gov/orffinder) with default settings. Blastp was used to identify sequences homologous to ORFs. The single best-hit transposable element (TE) was identified using blastn, searching ORF sequences against a reference library of squamate TEs (Pasquesi et al., 2018). We then used a one-way tblastn search to identify homologous coding ORF sequences in the CroVir3.0 assembly with a stringent e-value cutoff of $1e^{-30}$. In the case of the vomeronasal-like ORFs, the four non-identical ORFs were queried together, and where overlap occurred in the subject, the result with the higher bit score was retained. Putative signal peptides were identified from myotoxin Iso-Seq models using SignalP 6.0, with settings for "Eukarya" and "Slow" model to predict cleavage sites more accurately (Teufel et al., 2022). ORF coding potential for myotoxin Iso-Seq models 2 and 3 was calculated using CPC2 (Kang et al., 2017).

2.8. Generation of genome resequencing data for multiple individual C. v. viridis, and estimation of myotoxin genomic copy number variation and expression

We generated high-coverage whole genome resequencing data for 7 adult *C. v. viridis* individuals sampled from three populations (South [Texas, New Mexico], Mid [Colorado], and North [Montana, Colorado]; Supplementary Table S3) following previously detailed methods (Schield et al., 2020, 2021). In brief, we assessed copy number in two ways. We first used CNV-seq (Xie and Tammi, 2009) to determine if there was significant copy number variation between individuals. We then used custom scripts to estimate absolute copy number at myotoxin loci (github.com/drewschield/venom_population_genomics). For additional details on copy number variation analyses, see the Supplementary Methods.

To assess myotoxin expression in individuals, matching right venom gland RNA-seq samples for the individuals for which high-coverage genomes were generated were extracted at the same time as the other RNA-seq extractions performed for this study and following the same protocols (see section 2.1). RNA-seq reads were mapped to the Iso-Seq transcriptome using kallisto v0.46.2 (Bray et al., 2016) and abundance files were imported and scaled by library size and read length in R (R Core Team, 2021) using tximport (Soneson et al., 2015).

2.9. Structure modelling

Homology modelling was performed with the *C. d. terrificus* crotamine (PDB: 1Z99, Fadel et al., 2005) template using the MODELLER package (Eswar et al., 2008) within UCSF Chimera (Pettersen et al., 2004). RMSD values were calculated in Chimera by superimposing structures to the template using the 'best-aligned chain' option and adjusting sequences to include the same number of atoms across structures. Predictive modelling for signal peptides was performed using homomer complex prediction with AlphaFold2 (Jumper et al., 2021), using ColabFold (https://github.com/sokrypton/ColabFold). Surface electrostatic properties for soluble proteins were calculated using the Adaptive Poisson-Boltzmann Solver software pipeline (Jurrus et al., 2018), and surface hydrophobicity and electrostatic charge were modelled on top of structures using Chimera.

2.10. Assessing chromatin state and transcription factor footprints using ATAC-seq data

To better characterize transcription factors involved in regulating myotoxin expression in venom tissue, ATAC-seq reads derived from venom gland and skin tissue were mapped to the myotoxin 10x-extended scaffold. Venom ATAC-seq read libraries were generated for a separate study (see Perry et al., 2022), and skin ATAC-seq were generated for this study following the same protocol. Briefly, data were processed following the Harvard Informatics ATAC-seq best-practices pipeline (github.com/harvardinformatics/ATAC-seq) and wiggletools 1.2.1 (github.com/Ensembl/WiggleTools) was used to calculate the mean normalized ATAC-seq density across samples. A 100 bp region

approximately 90 bp upstream of the start codon was identified as an ATAC-seq peak in venom tissue and was scanned for transcription factor binding sites using a filtered JASPAR motif matrix of binding sites for transcription factors upregulated during venom production (Perry et al., 2022) using Ciiider (Gearing et al., 2019). Predicted transcription factor binding sites were plotted onto the myotoxin 10x-extended scaffold using karyoploteR (Gel and Serra, 2017).

3. Results

3.1. Rattlesnake genomes lack myotoxin loci

Currently, the most complete nucleotide sequence of a myotoxin gene is from *C. d. terrificus* (Rádis-Baptista et al., 2003), although transcripts and peptides from several other species and subspecies have also been identified. A myotoxin transcript was also identified from a *de novo* Illumina RNA-seq assembly of the venom gland of a *C. v. viridis* (Schield et al., 2019). A preliminary blast search of CroVir3.0 using the *C. v. viridis* transcript and *C. d. terrificus* crotamine yielded no significant hits. Although CroVir3.0 includes an annotated myotoxin locus on chromosome 1 based on the best-blast search hit of a whole coding transcript, the CroVir3.0 genome assembly does not contain the canonical myotoxin exon 2 sequence. Comparing the currently annotated myotoxins from CroVir3.0 and *C. tigris* to myotoxin sequences from other rattlesnake species derived from proteomic and transcriptomic data indicates that neither locus is correctly annotated (Fig. 1A).

3.2. Identification of a myotoxin-containing scaffold from 10x genomics linked-read genome sequencing

Given preliminary evidence rattlesnake genome assemblies lack an intact myotoxin, we generated new linked-read sequencing data from 10x Genomics to develop an independent inference of genome assembly from an individual from the same population as the CroVir3.0 genome animal (Weld Co., Colorado). These data were assembled, and a 1.3 kb myotoxin-containing scaffold that shared homology to *C. d. terrificus* crotamine was isolated. Using this scaffold, we performed a round of scaffold extension using raw Illumina paired-end genomic reads to generate a 2 kb extended scaffold (see section 2.3), which was confirmed to share a high sequence similarity with crotamine at both introns and exons (Fig. 1B). This scaffold (hereafter 'myotoxin 10x-extended scaffold') represents the first assembled myotoxin locus in the genome of *C. v. viridis*.

3.3. Comparing the myotoxin 10x-extended scaffold to the CroVir3.0 genome assembly

Previous studies have come to contradictory conclusions regarding the myotoxin locus or loci (Margres et al., 2021; Rádis-Baptista et al., 2003; Schield et al., 2019). In addition to using existing sequences to investigate the chromosomal location of myotoxin in *C. v. viridis*, we searched for patterns of homology using subsequences of the myotoxin 10x-extended scaffold (introns, exons, UTRs) to the existing CroVir3.0 assembly to further investigate its absence in the genome (Fig. 1A). Although no regions of homology to exon 2 were found, we did find several hits to intron 1 across multiple chromosomes, most of which were on regions of chromosome 1. Blast hits to the 5' UTR, exon 1 and intron 2 were also primarily to regions of chromosome 1.

The myotoxin signal peptide, with the exception of 3 residues on exon 2, is encoded entirely on exon 1, and is highly conserved across several myotoxins and myotoxin-related genes (Correa and Oguiura, 2013). As a result, it can be used as a marker for confirming the presence of myotoxins and paralogous genes. Searching for exon 1 returned hits to regions on CroVir3.0 chromosome 2, a chromosome of interest given evidence from previous FISH experiments for a presumed myotoxin on chromosome 2 of *C. d. terrificus* (Rádis-Baptista et al., 2003). There are five regions of CroVir3.0 chromosome 2 with homology to myotoxin intron 1 and one sequence highly similar to the myotoxin signal peptide (6 mismatches, e-value $1.84e^{-16}$; Supplementary Fig. S3; Supplementary Table S4). While not definitive, this raises the possibility that FISH probes designed using the entirety of the myotoxin gene, including introns and signal peptide, as was the case in the original experiment by Rádis-Baptista et al. (2003), could potentially lead to off-target hybridization caused by non-specificity of the probe. For example, these probes may have led to hybridization with off-target loci, or to loci that share some homology (e.g., with the signal peptide exon or intron 1) with a subsequence of the probe.

3.4. Evidence for myotoxin-like RNAs with unique properties encoded on chromosomes 2 and 4

To investigate potential splice and sequence variation of myotoxin transcripts, we generated a set of *C. v. viridis* venom gland transcripts using Pacific Bioscience's continuous long-read isoform-sequencing (Iso-Seq). From this transcript set, we identified three different myotoxin-related RNAs from *C. v. viridis* that included exons 2 and 3, which encode the mature peptide (Fig. 2A). One of these myotoxin Iso-Seq transcripts (myotoxin Iso-Seq model 1) is a coding sequence match to a previously identified transcript derived from an Illumina-based short-read mRNA-seq data from the *C. v. viridis* venom gland transcriptome (Schield et al., 2019) and also represents the most abundant myotoxin-like transcript sampled from 1 day post-extraction venom gland tissue (Supplementary Fig. S1). The two other Iso-Seq myotoxin-like RNAs (myotoxin Iso-Seq models 2 and 3) lack a canonical signal peptide, yet have high coding potential (Fig. 2B and C; Supplementary Table S5). General properties of these two sequences strongly suggest that they belong to a class of non-coding RNA with coding potential, often called 'mRNA-like lncRNA' or 'ppcRNA' (see section 4.6). Notably, the nucleotide sequences of exons 2 and 3 are identical across all three Iso-Seq models, though upstream regions in each of the three models are highly divergent (Fig. 2A), suggesting that models 2 and 3 likely represent independent loci. Interestingly, predicted structure modelling of the longest alternative signal peptide (from myotoxin Iso-Seq model 2) demonstrates a high level of structural similarity to the canonical myotoxin signal peptide (Fig. 2D). Though the peptides are of different lengths, hydrophobic core regions are present in both structures, as well as an identical number of coils in the alpha-helix region.

Several aspects of the structure of myotoxin Iso-Seq models 2 and 3 differentiate them from all previously described venom genes. For example, the intronic region of myotoxin Iso-Seq model 2 contains open reading frames (ORFs) that share homology with a region of a type-2 vomeronasal receptor (V2R) from *C. adamanteus* (NCBI: KAG6539740) and which have been duplicated resulting in five copies in tandem (Fig. 2B, Supplementary Fig. S4). Additionally, a 19 residue long coding sequence near the 5' end of the transcript is predicted to act as a signal peptide (P[cleavage site] = 0.97, L[signal peptide] = 0.99; Supplementary Fig. S5A). Using tblastn, we identified a single region on chromosome 2 with a high density of sequences homologous to the V2R ORFs (67 sequences across 1.8 Mb, mean e-value $< 1e^{-30}$, mean subject length = 254 bp). Interestingly, we also found this region had an abundance of LINE elements compared to the rest of the assembled genome and found that a subsequence of a *C. v. viridis* CR1 LINE is homologous to the V2R ORFs (Fig. 2B; Supplementary Figs. S4, S6).

Like myotoxin Iso-Seq model 2, model 3 also contained a 5' ORF, with homology to a histocompatibility antigen (HCA) subunit from *C. tigris* (NCBI: XP_039204351; Supplementary Fig. S4). Similar again to model 2, the first 23 residues of this ORF is predicted to be a signal peptide (P[cleavage site] = 0.95, L[signal peptide] = 0.99; Supplementary Fig. S5A). Although the predicted signal peptides in myotoxin Iso-Seq models 2 and 3 differ in their cleavage site, their sequences are highly similar, varying only in the length of the hydrophobic core. A tblastn search for homologous ORFs in CroVir3.0 resulted in two

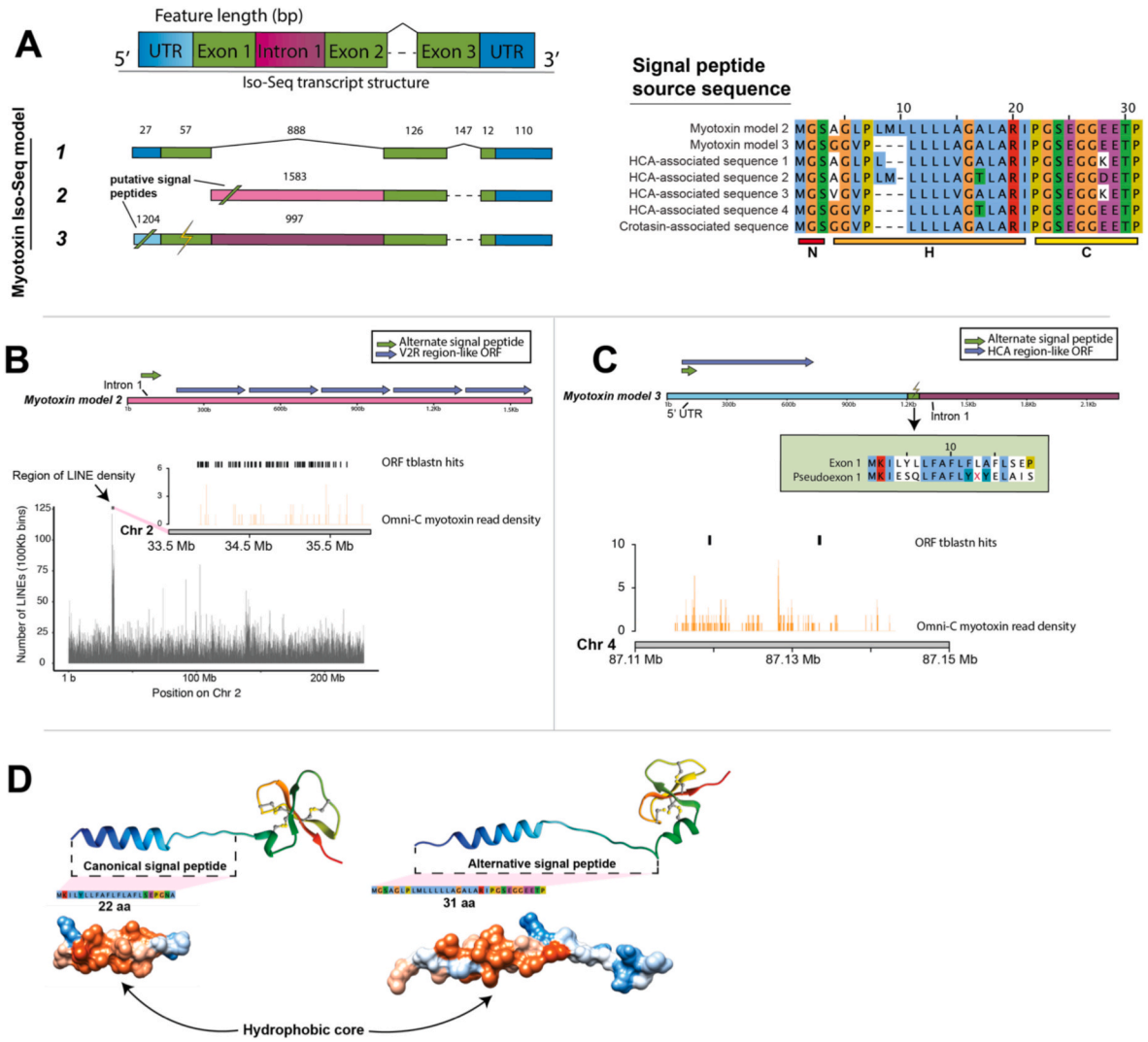

Fig. 2. Characterization of alternative myotoxin transcripts. A) Structure of the myotoxin Iso-Seq models are shown with lengths in base pairs of each feature noted above. Coloured shading is used to represent sequences which are highly divergent. Splice junctions are denoted with an upside down 'V', and fused exons with a dashed line. Figure is not to scale. Iso-Seq sequences which contain alternative signal peptides are shown in the alignment to the right, with N (N-terminal), H (hydrophobic) and C (C-terminal) residues annotated by SignalP (Teufel et al., 2022). **B)** Myotoxin Iso-Seq model 2 contains five V2R ORFs and a putative signal peptide in the intron 1 region. ORF tblastn hits are shown on a region of chromosome 2, which also corresponds to a local pileup of Omni-C myotoxin-derived reads, as well as a peak in annotated squamate LINEs. C) Myotoxin Iso-Seq model 3 contains a pseudoexon likely derived from a once functional myotoxin signal peptide, as well as an alternate signal peptide housed within a histocompatibility antigen ORF. ORF tblastn hits on chromosome 4 correspond with a pileup of Omni-C myotoxin derived reads. D) Predicted structures of the canonical myotoxin signal peptide and the alternative signal peptide from myotoxin Iso-Seq model 2 show conservation of alpha-helix coil structure, important for cell surface membrane interaction. The hydrophobic core (shown in red shading) of the alternate signal peptide is slightly longer which suggests differences in the rate of myotoxin peptides produced by the two RNAs.

significant blast hits to a region on chromosome 4 (e-value $< 1e^{-30}$). This transcript also contains a sequence approximately 1000 bp upstream of the start of exon 2 which appears to be derived from a canonical myotoxin exon 1. Relaxed constraint on the exon is evident, which has allowed a nonsense mutation to occur at amino acid position 13, rendering it non-functional pseudoexon (Fig. 2C). The putative signal peptides identified in myotoxin Iso-Seq models 2 and 3 were also found in several other transcripts from our Iso-Seq set (Fig. 2A, Supplementary Table S6), including other HCA-associated sequences and a crotasin-like coding sequence (identified by blast similarity to crotasin; NCBI: AF250212). Crotasin is a putative β-defensin primarily expressed in the pancreas and closely related to crotamine (Fry, 2005; Rádis-Baptista et al., 2004; Yount et al., 2009).

3.5. β-Defensin-like genes form a tandem array on C. v. viridis chromosome 1

We identified 12 previously unannotated sequences on chromosome 1, and one on chromosome 2 with homology to human β-defensins, which we refer to as *C. v. viridis* β-defensin-like sequences (Cvv BDLs). All 13 Cvv BDLs have a conserved six cysteine motif, important for disulfide bridge formation in vertebrate defensins and myotoxins (Fadel et al., 2005; Ganz and Lehrer, 1994; Oguiura et al., 2011; Whittington et al., 2008; Zou et al., 2007). The 12 Cvv BDL sequences on chromosome 1 are organized in two discrete clusters of 6 Cvv BDLs each (Fig. 3B, Supplementary Fig. S7; Supplementary Table S4). Within the first of these clusters, tandem duplication is apparent and supported by the presence of two identical copies of Cvv BDL 3 located approximately 800 bp apart in the CroVir3.0 assembly. A second cluster of Cvv BDLs (β-defensin-like 7 through 12) was identified approximately 3 Mb

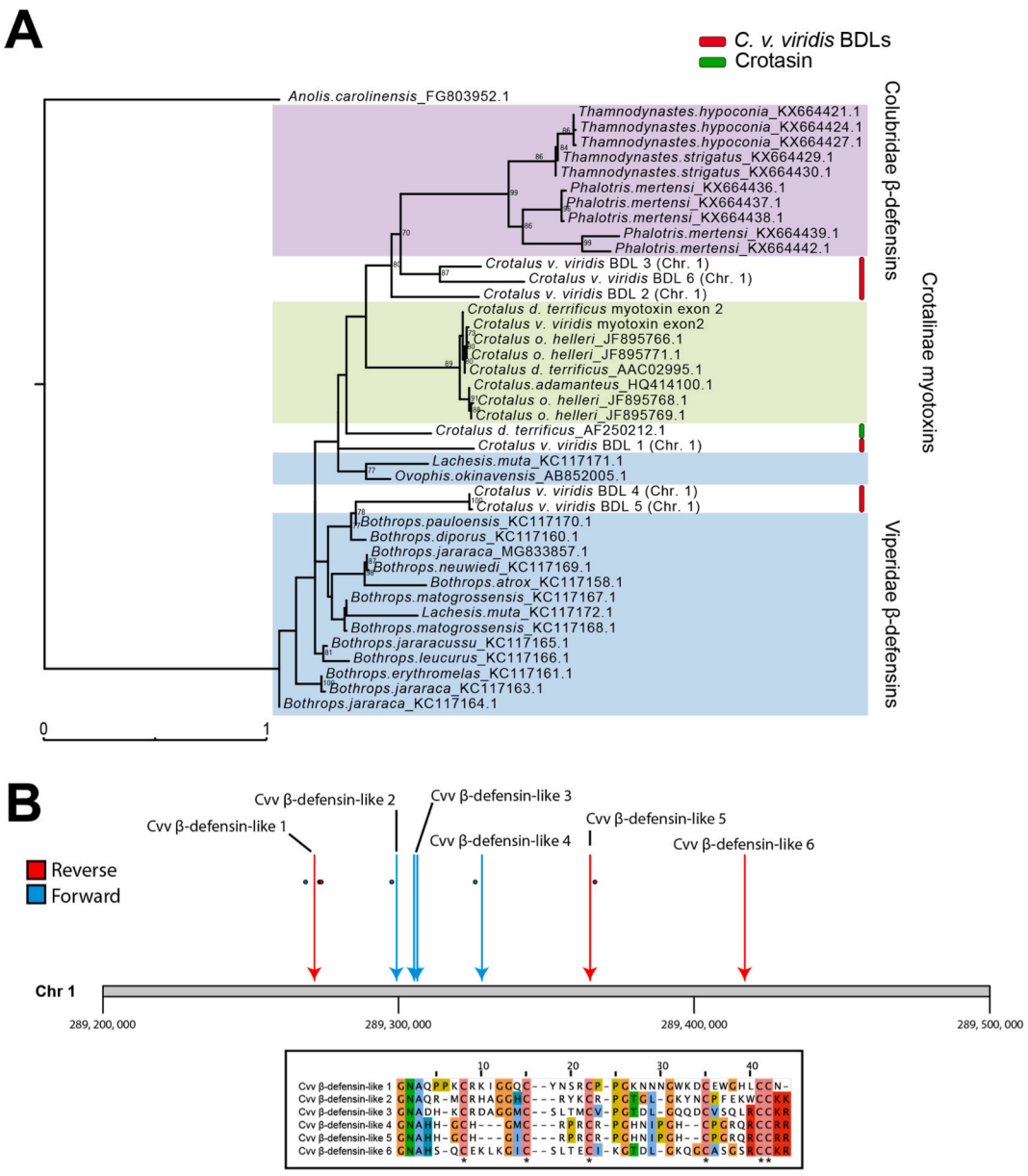

Fig. 3. β-defensin-like genes form an array on chromosome 1. **A)** A rooted phylogeny of snake β-defensins, myotoxins and Cvv BDLs indicates Cvv BDLs are similar to β-defensins from both rear-fanged and front-fanged snakes. Cvv BDLs are highlighted with red bars and C. d. terrificus crotasin with a green bar. Node labels represent ultrafast bootstrap with 1000 replicates supported in over 70% of trees. Tree is rooted using a β-defensin sequence from Anolis carolinensis with homology to crotamine (Dalla Valle et al., 2012). **B)** 6 β-defensin-like genes are found on chromosome 1 with signal peptides (shown as coloured dots). Exon 2 sequences are shown in the outlined box below with the conserved six-cysteine pattern labelled with asterisks.

upstream of the first cluster on chromosome 1. A comparison of the sequences from the two clusters indicates major divergence and lack of a conserved signal peptide, possibly a result of *in silico* gene loss obscuring intermediate duplicates. A reconstruction of the gene tree of snake β-defensins, myotoxins and Cvv BDLs supports these novel sequences belonging to the β-defensin family (Fig. 3A). While β-defensins from colubrids and rattlesnake myotoxins cluster into discrete clades, Cvv BDLs are distributed across the phylogeny with other non-myotoxin β-defensins. We did not include Cvv BDLs 7–12 in phylogenetic analysis due to an inability to determine homologous sites accurately (see section 2.6).

The predicted three-dimensional structures of Cvv BDL peptides are very similar to that of crotamine (mean RMSD 1.64 Å, $\sigma = 0.39$), despite low primary sequence identity (Supplementary Fig. S8). Though all peptides are stabilized by three disulfide bonds, Cvv BDLs are predicted to lack a secondary alpha helix present in myotoxins. This suggests there is a structural difference between toxic and non-toxic peptides, differentiating those used for venom (myotoxins) and those for innate immunity (β-defensins).

3.6. Evidence that the venom-associated myotoxin locus is located within a β-defensin cluster on chromosome 1

Considering established findings that myotoxins are descendants of β-defensin genes, and that several β-defensin genes are found on primarily on chromosome 1, we hypothesized that the myotoxin locus is likely adjacent. To test this, we used three types of long-range mate-pair chromatin contact datasets to help highlight the chromosomal location of the canonical venom-associated myotoxin locus (Supplementary Figs. S2A, S9; see section 2.5). Using this approach, we identified a locus of interest within the first β-defensin-like cluster in the genome (containing Cvv BDLs 1–6). To reduce spurious genomic mappings (i.e.,

myotoxin exon reads aligning to Cvv BDL exons), we performed the mapping step again with Cvv BDL exon 2 sequences masked, which did not result in any major read mapping pileups elsewhere in the genome (Supplementary Fig. S9; Supplementary Table S2). Instead, reads mostly mapped to non-coding regions flanking exon two, such as introns and UTRs, reinforcing the sequence similarity of these elements across β-defensin-like genes and myotoxins. An additional filtering step to retain mappings with 0 edit distance (identical matches between read and subject) did not result in major pileups elsewhere, though the magnitude of some peaks were reduced (Supplementary Fig. S10). A more formal analysis using inferences of chromosomal contact from Omni-C data was also performed to identify regions of the myotoxin gene which were physically associated with regions of the genome (Supplementary Figs. S11, S12). This analysis again reinforced that the region on chromosome 1 containing Cvv BDLs 1–6 was frequently in contact the myotoxin gene, which strongly supports the adjacency of the myotoxin locus with paralogous BDLs.

3.7. Multi-tissue mRNA analysis reveals myotoxin expression outside the venom gland

To investigate expression patterns of myotoxin and myotoxin-like RNAs in different tissues, we used a previously generated Illumina-based, multi-tissue RNA-seq dataset of 32 RNA-seq libraries generated from 18 different tissues and tissue states from *C. v. viridis* (Schield et al., 2019). We estimated RNA expression using genome annotations that included the myotoxin 10x-extended scaffold (representing the corresponding gene for the Iso-Seq transcript model 1 RNA), myotoxin Iso-Seq models 2 and 3, as well as all newly described Cvv BDLs. We found all myotoxin Iso-Seq models were expressed in venom gland tissues even after accounting for multi-mapped reads (Fig. 4A see section 2.4). Myotoxin Iso-Seq model 1 was the highest expressed transcript in venom tissues, and along with other highly expressed venom transcripts, also exhibited significant upregulation in venom tissues compared to other body tissues ($\log_2$ fold change >0, BH-adjusted p-value < 0.05; Fig. 4B), highlighting its importance in venom expression. Myotoxin Iso-Seq model 1 was also moderately expressed in several body tissues at levels higher than that of other venom genes and other highly expressed Cvv BDLs (Fig. 4C, Supplementary Fig. S13). Cvv BDL 8 and 9 especially are highly expressed in nearly all sampled non-venom tissues, especially the kidney and liver. In other mammals including humans, β-defensin expression in these organs has been reported (Froy et al., 2005; Schröder and Harder, 1999; Zhao et al., 1996), underscoring the likely analogous role played by CVV BDL 8 and 9 here.

RNA-seq expression support the usage of the alternative signal peptide of myotoxin Iso-Seq model 2 in lung tissue, a sample in which expression of this sequence was greater than either Iso-Seq models 1 or 3 (Fig. 4A, D). A junction spanning exon 1 and the remaining two exons was supported by the presence of the canonical GT-AG splice motif on either end of sequence (Burset et al., 2000). This finding represents new evidence of myotoxin-related RNAs being expressed in tissues not normally associated with venom secretion and the utilization of non-canonical signal peptides, with potentially different physiological properties and functions.

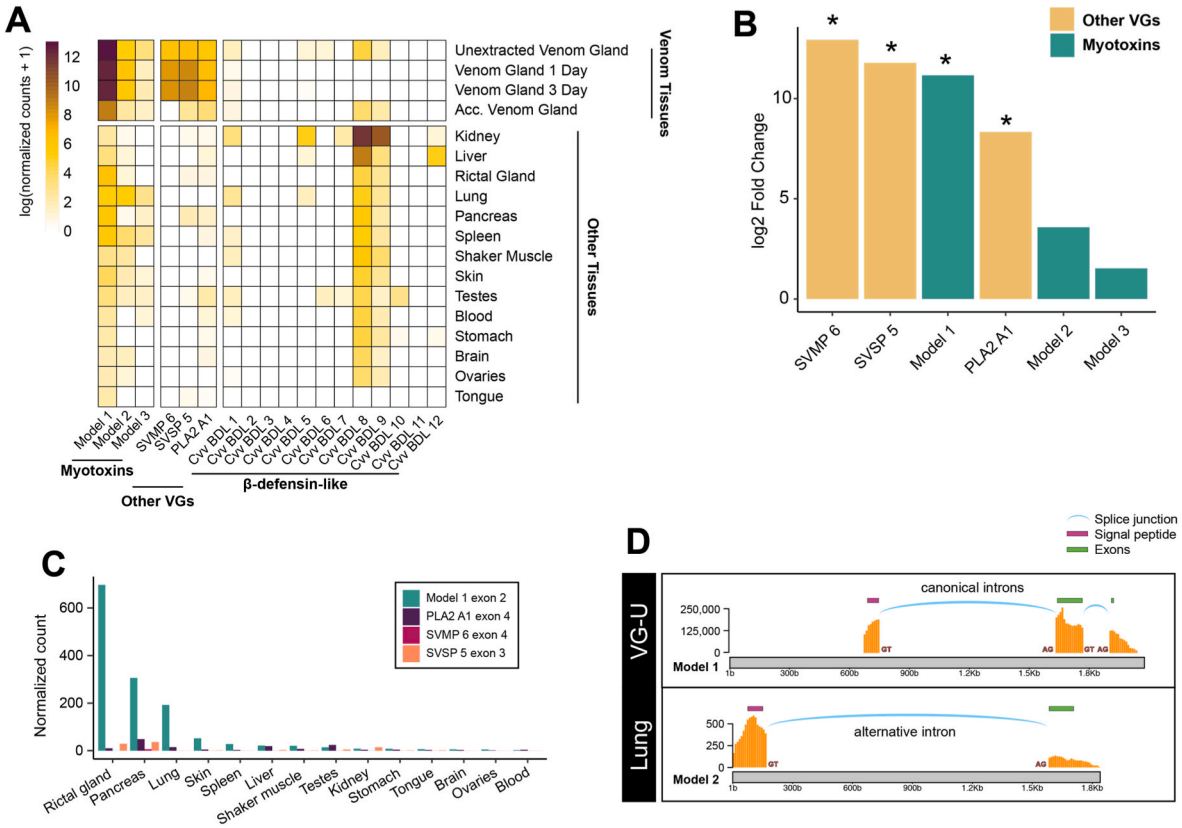

Fig. 4. RNA-seq analysis of alternative myotoxin Iso-Seq models. A) Normalized expression heatmap of three myotoxin Iso-Seq models, the three most highly expressed venom genes in *C. v. viridis* (Dalla Valle et al., 2012), and 12 novel *C. v. viridis* β-defensin-like genes. **B)** Log2 fold change of venom associated genes shows upregulation between venom- and non-venom tissues. BH-corrected p-values < 0.05 indicated with asterisk. **C)** DESeq2 normalized counts of RNA reads mapped to the longest exons of venom associated genes in non-venom tissues. **D)** A non-canonical myotoxin RNA (Iso-Seq model 2) is spliced in lung tissue. Top panel shows canonical splicing of myotoxin in unextracted venom gland tissue. The canonical splice junction GT-AG is shown in red.

3.8. Myotoxin 5' UTRs contain several venom-critical transcription factor binding sites

In a previous study, Assay for Transposase-Accessible Chromatin (ATAC)-seq chromatin accessibility data for *C. v. viridis* venom glands were generated (Perry et al., 2022), and skin tissue ATAC-seq was generated for this study following the same protocols. Here, we used this data to explore differences in chromatin structure and infer relevant transcription factor (TF) binding sites for the reconstructed myotoxin locus generated from paired-end reads (myotoxin 10x-extended scaffold, see section 2.3). We find evidence for a promoter region with open chromatin in venom gland tissue approximately 90 bp upstream of the start codon in this sequence (Fig. 5A). Within the open chromatin region of the promoter in venom gland tissue, we find several transcription

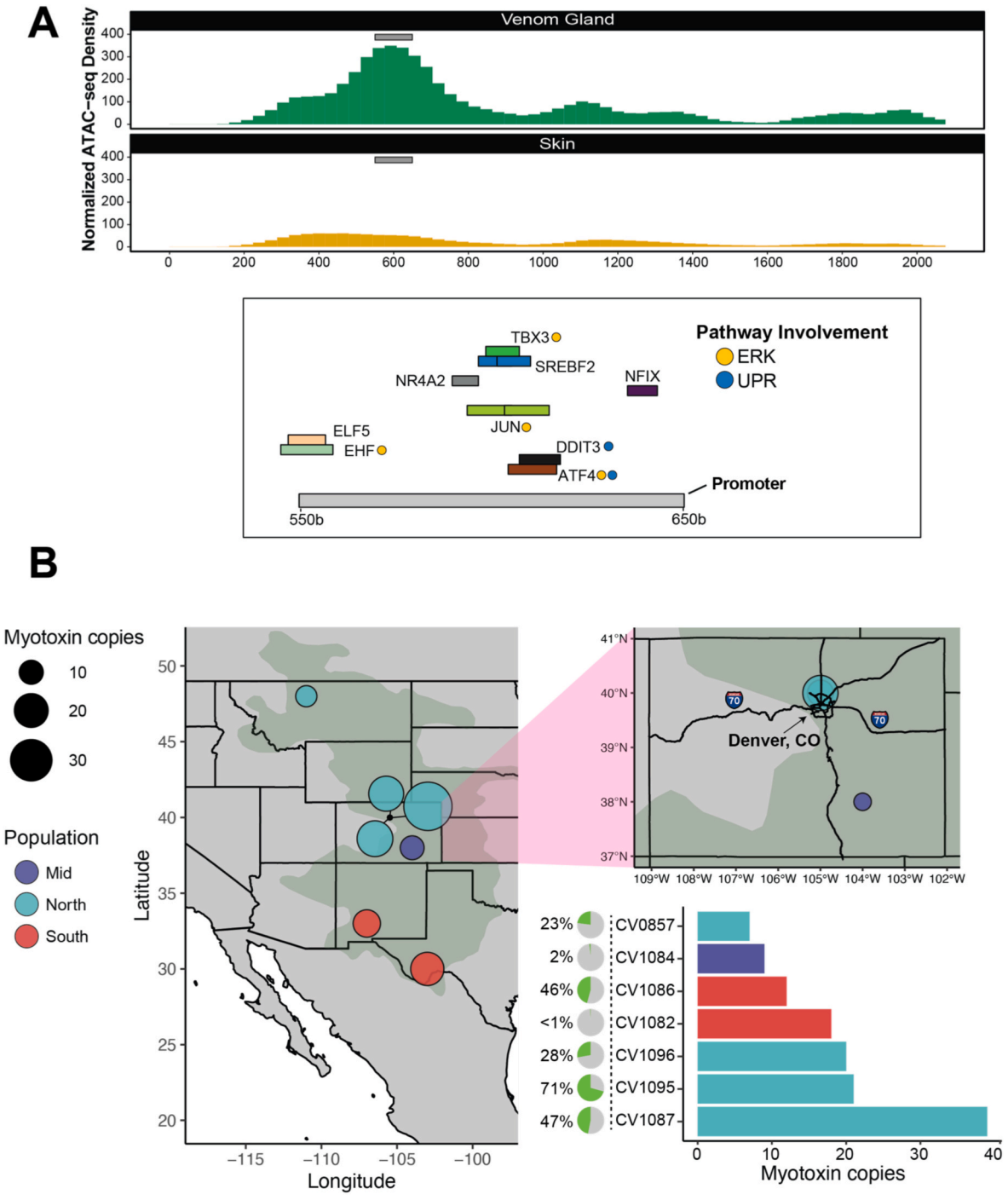

Fig. 5. Myotoxin transcription factor binding site and copy number variation analyses. A) ATAC-seq data from venom and skin tissues mapped to the myotoxin 10x-extended scaffold highlight differential chromatin accessibility at the myotoxin promoter region (shown as a grey bar). The promoter contains several binding sites for transcription factors known to be important for venom regulation, as well as several which are directly involved in the ERK and UPR pathways (Perry et al., 2022). **B)** Myotoxin copy number varies greatly across the sampled range of *C. v. viridis* (N = 7). The inset figure shows copy number proxy is highest in individuals (N = 3, points collapsed) in the Denver, Colorado area north of highway I-70, and decreases in populations south and further north of this landmark. CV1087 (Weld Co., Colorado) acted as the reference specimen for CNV-seq analysis. Bar plot shows number of myotoxin copies in each genome as well as the respective fraction of venom gland RNA reads represented by myotoxin (in green) in pie charts to the left.

factor (TF) binding sites that were previously identified as key factors in the regulation of venom in *C. v. viridis*, including ELF5, EHF, NR4A2, TBX3, SREBF2, JUN, DDIT3, ATF4 and NFIX (Fig. 5A; Perry et al., 2022). Consistent with conclusions from Perry et al. (2022) that UPR- and ERK-related TFs coordinate snake venom gene expression, multiple TFs predicted to bind the myotoxin promoter are associated with ERK and UPR signaling pathways (Fig. 5A). Together, these analyses suggest that myotoxin is regulated by the same overarching pathways as other major venom gene families in snakes, and that *cis*-regulatory elements driving myotoxin expression have been similarly co-opted to be responsive to ERK and UPR signaling.

3.9. Myotoxin gene copy number and expression estimation and variation across individuals

Quantitative changes in expression can be achieved simply by tandem duplication, which can increase expression severalfold (Loehlin and Carroll, 2016). Under expectations that this process drives expression changes in myotoxin, we examined evidence that myotoxin expression has been influenced by copy-number variation (CNV) at its locus by analyzing whole genome resequencing data from adult *C. v. viridis* individuals spanning its range (N = 7). We find evidence of significant CNV (p-value < 0.05) at the canonical myotoxin locus in five pairwise comparisons between individuals from different geographic populations (Fig. 5B). Significantly lower copy number is apparent for individuals from southern (Texas, New Mexico) and far northern (Montana) populations when compared to the reference central Colorado population (Supplementary Table S3). Additionally, geographic distance, and the associated genetic distances, alone does not predict difference in myotoxin copy number between individuals (Supplementary Fig. S14). In contrast to myotoxin, significant CNV at any Cvv BDL gene was found in only 2 comparisons.

Because CNV-seq does not compute absolute values for copy number, we also compared relative read depth for the three myotoxin models to calculate a copy number proxy (see section 2.8) We find that absolute copy number varies substantially and generally correlates with results from CNV-seq. Venom-associated myotoxin copy number is highest in individuals near Denver, Colorado north of highway I-70 and decreases in southern and more northern populations (Fig. 5B). Copy number estimations for the two myotoxin-like RNAs (myotoxin Iso-Seq model 2 and 3) are consistent with there being one or no copies in each individual, suggesting they belong to a different, rapidly evolving class of RNA.

We used individually matched venom gland RNA-seq to compare the relative contribution of myotoxin to the whole expression profile which we compared to absolute copy number. We find that myotoxin copy number is not correlated with myotoxin expression. The fraction of venom gland RNA-seq reads is highly variable across individuals (<1%–71%), even for those sampled from the same location. When expressed, myotoxin makes up a significant fraction of the venom gland RNA-seq reads per individual, upwards of a quarter of all transcripts. We also find preliminary evidence of major differences in myotoxin composition between northern and southern sides of Denver, Colorado, perhaps representing a barrier to gene flow.

4. Discussion

4.1. A new genomic context for understanding viperid venom myotoxins

Our findings provide multiple sources of evidence that the primary myotoxin venom toxin is encoded on chromosome 1 in rattlesnakes and helps clarify previous contradictory inferences of its genomic location. Our results also highlight the expectation that modern genome assemblies are likely to fail in completely assembling this locus due to its structural complexity and multi-copy nature. We estimate that the venom-associated myotoxin is commonly present at up to 30 or more copies in some individual *C. viridis*, although this does not necessarily correlate with a concomitant increase in gene expression in venom glands, at least in our limited sampling of individuals. The expression of myotoxin in several body tissues, combined with structural similarity to physically adjacent and highly expressed β-defensins on chromosome 1, suggest that this myotoxin may also play minor secondary functional roles outside of venom. We also discovered additional novel myotoxin-encoding RNA (which we believe represent lncRNAs, as discussed below) that are predicted to produce peptides identical to the canonical myotoxin, except for alternate signal peptides distinct from the canonical venom myotoxin which may affect the rate of secretion or play other roles (von Heijne, 1990). These lncRNA also appear to encode non-toxin ORFs related to vomeronasal receptors and histocompatibility antigens, suggesting a broader relationship between non-toxin polycistronic RNA and venom genes, though the exact functions of these sequences remain unknown.

4.2. Identification of the primary venom myotoxin on chromosome 1

We find multiple corroborative lines of evidence indicating that the primary venom myotoxin locus is found adjacent to six Cvv BDLs (*C. v. viridis* β-defensin-like peptides) on chromosome 1. It is notable that this inferred location corresponds with the general location of annotations of this locus in two available rattlesnake genomes (Margres et al., 2021; Schield et al., 2019), although these inferences were based solely on shared homology between myotoxin blast queries and assembled BDLs in those genome assemblies (as these genome assemblies lack an actual myotoxin locus). While we find no evidence that myotoxins are present in any form on chromosome 2 in *C. v. viridis*, the possibility remains that recent, lineage-specific translocations could have dissociated myotoxin from non-venom paralogs in some species. Indeed, we do find a Cvv BDL on chromosome 2, which lends some support to this explanation.

In addition to extremely high expression levels in the venom gland, we find evidence that this venom-association myotoxin is also expressed at lower levels in the rictal gland, pancreas, lung, skin, and spleen, and at levels comparable to highly expressed Cvv BDLs in the pancreas and lung. This observed expression pattern could be explained by 'leaky expression' (de LM Junqueira-de-Azevedo et al., 2015; Ramsköld et al., 2009), where a tolerable level of non-functional expression outside the primary organ occurs as a result of changes in the location of expression over time. Moderate levels of expression in non-venom tissues supports the model of subfunctionalization from an ancestral, ubiquitously expressed Cvv BDL, followed by a restriction of expression to the venom gland, which, given its rattlesnake specificity (Bober et al., 1988), has presumably occurred within the last ~12 million years (Castoe et al., 2009; Reyes-Velasco et al., 2015). Interestingly, the antimicrobial activity of myotoxin against several pathogens is comparable to that of human β-defensins *in-vitro* (Yount et al., 2009), supporting the hypothesis that myotoxin may play a similar physiological role (analogous to Cvv BDLs 8 and 9) in non-venom tissues. This multi-tissue expression pattern further suggests that myotoxin-expressing venomous snakes are largely immune to the autotoxicity of myotoxin, which is also consistent with the hypothesis that these peptides represent a mammalian-targeted component of venom (Mackessy and Saviola, 2016)

4.3. Myotoxin loci pose challenges for conventional genome assembly

While medical and pharmacological relevance of rattlesnake myotoxins has been long recognized, little was previously known about the genes that encode them or their genomic location. We show that modern genome assemblies are prone to excluding these loci in final assemblies primarily due to a high degree of similarity between the non-coding regions of myotoxins and β-defensins, and likely due to the recent evolution of the array that also appears to encode many duplicated copies of this locus in *C. v. viridis*. Though the partially assembled Cvv BDL array presents as two discrete clusters in the *C. v. viridis* CroVir3.0 genome

assembly, we expect these clusters are likely incomplete and are larger than reconstructed in the genome assembly. An estimate using mean gene densities within each cluster and extrapolating across the intervening sequence suggests as many as 70 myotoxin-β-defensin paralogs exist on chromosome 1 that have not been assembled. In this assembly, the paralogs found on either end of each array are most divergent, which we posit is a result of physically intermediate copies having been collapsed into one of the two clusters during assembly. Further complications for accurate assembly of these loci arise from their very short gene length which may compound assembly errors (Meader et al., 2010; Phillippy et al., 2008). By integrating multiple '-omics' approaches that bypass read assembly, we find evidence for multiple loci that encode myotoxin and myotoxin-related RNA located on three different macrochromosomes in the *C. v. viridis* genome.

Though multiple myotoxin protein forms are found in rattlesnake venoms, we identified only a single major transcript of myotoxin in *C. v. viridis*, based on our long-read transcriptome Iso-Seq data. Prior studies have described a secondary myotoxin using protein fractionation from *C. v. viridis* venom, referred to as "myotoxin 2" (Griffin and Aird, 1990), although it is present at very low concentrations compared to the primary myotoxin, called "myotoxin a" (Saviola et al., 2015). It remains unknown if there is a secondary locus that encodes myotoxin 2, or if this represents an alternate transcript or protein cleavage product, or perhaps a slightly distinct duplicated paralog of the primary myotoxin a locus in some individuals.

4.4. Myotoxin copy number and expression variation in rattlesnakes

Many other snake venom gene families show evidence for gene duplication, especially in cases where those families make up a substantial fraction of the venom (Margres et al., 2017; Oguiura et al., 2009; Schield et al., 2019; Suryamohan et al., 2020). Supported by independent analyses of copy number variation (CNV), we find evidence that the primary venom myotoxin locus in *C. v. viridis* similarly exhibits CNV consistent with previous inferences of copy number variation in myotoxin-encoding loci in *C. d. terrificus* and *C. adamanteus* (Margres et al., 2017; Oguiura et al., 2009). For example, we find that individual *C. v. viridis* from populations in central Colorado (near Denver, north of interstate 70) possess the greatest number of copies, with copy number decreasing in more northern and southern populations. In contrast to this primary multi-copy myotoxin locus, we find little evidence of CNV across individuals for the nearby Cvv BDL loci, which have broadly similar genomic structure, suggesting that CNV is particularly elevated for the myotoxin venom locus that is not observed in adjacent non-venom BDL paralogs.

Interestingly, copy number appears to be decoupled from venom transcript abundance across the small set of *C. v. viridis* individuals examined, which contrasts with previous findings of a correlation between copy number and protein expression in previous studies of myotoxin a in a different rattlesnake species (Margres et al., 2017). This finding in *C. v. viridis* suggests that differential regulation of the myotoxin locus in the venom gland, perhaps by ncRNA or chromatin accessibility, or possibly due to cis-regulatory element variation not detected here, also contribute substantially to myotoxin expression in venom. From our population-scale sampling, we find that myotoxin comprises a wide range of overall venom gland transcript abundance across individuals, from as little as <1% of the venom gland transcriptome to over 70% in some individuals, which at the high end represents one of the most highly expressed (by percentage) single transcripts reported in any vertebrate tissue. The extreme degree to which observed expression varies between individuals suggests that the rapid accumulation of gene copies might not necessarily be adaptive in all cases and could instead be a byproduct of stochastic processes such as drift (Aird et al., 2017; Nozawa et al., 2007). This may occur if a lethal dose is achieved at a certain fraction of expression, beyond which additional expression has neutral or minimal effects on venom function and lethality, which may also explain why individuals from the same population can vary in the degree of myotoxin expression, where prey composition is likely not a factor.

4.5. Venom myotoxins share conserved regulatory architecture with other venom genes

Recent studies of snake venom regulation have characterized pathways and transcription factors responsible for orchestrating venom expression but have lacked incorporation of myotoxin because this locus was not characterized at the time of this work (Perry et al., 2022). Using ATAC-seq chromatin accessibility data, we identified the myotoxin promoter approximately 90 bp upstream of the start codon and show that the chromatin accessibility of this promoter is elevated during venom production (post extraction) in venom gland tissue. Using this ATAC-seq data, we identified putative binding sites in this open chromatin region for TFs that directly linked to ERK-signaling, and other TFs involved in the UPR (unfolded protein response) – the two pathways identified as central regulators of snake venom expression (Perry et al., 2022). This finding implies an evolutionarily recent recruitment of these pathways for the regulation of myotoxin that directly link it to the broad, conserved regulatory architecture governing other snake venom gene families. However, the details of other *cis*-regulatory features that contribute to myotoxin regulation, such as the location and evolutionary origins of enhancers, remain unknown due to the poor resolution of contiguous sequences in myotoxin-encoding regions. Future studies using long-read sequencing would help address these remaining questions by building more contiguous genomic alignments for these regions that would be important for further resolving regulatory features associated with myotoxins.

4.6. Myotoxin-encoding lncRNA located on chromosomes 2 and 4

To date, no studies have identified long non-coding RNAs (lncRNA) associated with snake venom genes. LncRNA form an independent class of ncRNA and are key components for the regulation of eukaryotic gene expression. Along with being both transcriptional activators and repressors, lncRNA can also directly modify chromatin and scaffold protein complexes (Chen, 2016; Sun et al., 2018). From the venom gland Iso-Seq data, we identified two additional myotoxin-encoding RNA transcripts (myotoxin Iso-Seq models 2 and 3) that we predict are located on chromosomes 2 and 4, respectively. Surprisingly, exons 2 and 3 of these myotoxin Iso-Seq models are identical at the nucleotide level to the venom-associated myotoxin locus on chromosome 1, suggesting either strong purifying selection maintains this sequence homology across genetically unlinked loci, or possibly that ectopic gene conversion may have played a role in maintaining sequence identity. Strong purifying selection is consistent with the otherwise highly conserved nature of myotoxin orthologs within and between species. For example, only three nonsynonymous substitutions across 195 positions differentiate the *C. v. viridis* myotoxin Iso-Seq model 1 from *C. d. terrificus* crotamine (Supplementary Fig. S15), and myotoxin paralogs from different polymorphic individuals of the same species are frequently identical (Margres et al., 2017), suggesting that functional myotoxin sequences are generally subject to very strong constraint. Because the mature peptide is encoded solely on exons 2 and 3, all myotoxin Iso-Seq models would theoretically produce identical secreted peptides after cleavage of the signal peptide sequence. The discovery of these additional myotoxin-related loci on distinct chromosomes, and evidence of their broad patterns of expression across tissues, raises the question of what the functions of these RNA may be.

The ORF content and low expression of myotoxin Iso-Seq models 2 and 3 suggest that they belong to a subclass of lncRNA with coding potential (Wu et al., 2014), contrary to the suggestion of them being 'non-coding'. Indeed, recent evidence has accumulated describing small, upstream ORFs transcribed from lncRNA that encode physiologically

relevant micropeptides with a range of functions (Andrews and Rothnagel, 2014; Bazzini et al., 2014; Cai et al., 2021; Payre and Desplan, 2016; Razooky et al., 2017; Robinson et al., 2020; Zeng and Hamada, 2018). Some lncRNAs can be polyadenylated and spliced like mRNA, suggesting they can be recovered and sequenced from standard poly(A) enrichment library preparation protocols, such as those used in this study (Guttman et al., 2009; Sun et al., 2018). While both myotoxin lncRNAs lack canonical signal peptides, novel ORFs that appear to encode signal peptides have arisen in their place. A comparison of the predicted structures of the novel and canonical signal peptides illustrates remarkable similarity between these uniquely derived structures (Fig. 2D). Despite the primary sequences being unrelated, the structural similarity of these peptides implies the potential retention of their secretory capacity, as each would produce identical secondary structures with hydrophobic cores needed for membrane docking (von Heijne, 1990). It is possible these two lncRNAs encode functionally distinct myotoxins using alternative signal peptides which are expressed in specific tissues or tissue states, which may or may not be relevant to venom. We also found a homologous signal peptide sequence encoded on the 3′ end of a transposable element (TE) from a Squamate TE library (Pasquesi et al., 2018; Supplementary Table S3), which raises the possibility that the origins of this alternative peptide sequence may TE insertion related. The putative location of myotoxin Iso-Seq model 2 on chromosome 2 corresponds to the largest LINE element hotspot (in terms of regional density) in the genome (Fig. 2B), lending some support to this theory.

Given the notable absence of nonsense mutations and high coding potential (Supplementary Table S5) of the non-myotoxin ORFs encoded on these putative lncRNAs, it is likely that multiple micropeptides may also be produced from these sequences, though this has yet to been confirmed empirically by proteomic data. Eukaryotic polycistronic RNA were first described in flour beetles (*Tribolium* sp.) and were proposed to belong to a new class of RNA termed polycistronic peptide coding RNA (ppcRNA), whose 'hallmark' is the repetition of a coding peptide (Savard et al., 2006). Consistent with this description, myotoxin Iso-Seq model 2 contains a repeated ORF, although this feature is not shared with myotoxin Iso-Seq model 3 (Fig. 2B). Because of the apparent inconsistency in the naming convention of these types of RNAs and their structural description, we opt to refer to them as lncRNA though their exact designation and function remain unclear. Interestingly, these lncRNA do not appear to be associated with the extreme structural and copy number variation observed for the venom-associated myotoxin, and they are expressed at particularly low levels in venom gland tissue, suggesting alternative roles not related directly to encoding venom toxins. Their origins are also not yet clear, but a possible explanation for how such unrelated peptide-encoding ORFs may have originated on different chromosomes may involve the activity of retroelements, leading to the duplication of retro-transcribed mRNA copies of the myotoxin toxin-encoding locus.

It important to note that high-throughput annotation pipelines used to characterize lncRNA based on coding potential alone will discard sequences with the unique features described here for these lncRNAs. This highlights the importance of performing a broader and more exhaustive appraisal of different classes of lncRNA in future transcriptomic studies to fully characterize ncRNA diversity, particularly those involved in the regulation of venom. While other non-coding RNA, specifically miRNA, have been implicated in venom regulation previously (Durban et al., 2013, 2017), this example represents the first lncRNA associated with snake venom systems. Our findings add to the impressive functional repertoire of lncRNA, not only as chromatin-modifying and transcript-binding sequences akin to other classes of ncRNA, but also as a potential distinct source of coding sequence variation for a fitness-critical adaptation.

5. Conclusions

Despite ongoing advancements in the contiguity and completeness of genome assemblies, myotoxins have not been characterized or correctly assembled in existing snake genomes. As a result, our understanding of the genomic structure, variation, and regulation of these genes has lagged far behind other more well-studied snake venom gene families. Here, we make progress bridging this gap by identifying the sequence of myotoxin-containing loci, and identifying their chromosomal locations, and relationships to adjacent β-defensin genes. We also conduct the first broad analyses of patterns of myotoxin gene expression across tissues, provide new evidence for extreme copy number variation across populations, and identify cis-regulatory elements and associated transcription factors involved in the expression of the primary venom myotoxin locus. Our findings also demonstrate that myotoxin is regulated by the same core trans-regulatory architecture as other snake venom gene families. Considering myotoxins are thought to be primarily a rattlesnake-specific venom toxin family, they provide a notable example of recent functional convergence of ERK and UPR pathway co-option to regulate a venom gene family, compared to other venom gene clusters that adopted this regulatory architecture more anciently in the common ancestor of viperids or earlier. We also find intriguing but still poorly understood relationships between the primary venom-encoding myotoxin and remarkably similar yet distinct lncRNA, which may play roles in venom regulation or other physiological processes. This finding represents the first evidence that lncRNAs may play roles in venom regulation and highlights the degree to which both small and long ncRNAs have generally received minimal attention in the regulation of venom, despite their demonstrated importance in eukaryotic gene regulation. Additional detailed reconstruction of myotoxin loci and careful reannotation, perhaps using 3rd generation long-read sequencing approaches, would be valuable for understanding the evolutionary genomic origins and mechanisms leading to the neofunctionalization of β-defensins and other myotoxin-related paralogs.

Author statement

Siddharth S. Gopalan: Conceptualization, Formal analysis, Writing – original draft. **Blair W. Perry**: Conceptualization, Formal analysis, Writing – review & editing. **Drew R. Schield**: Conceptualization, Writing – review & editing. **Cara F. Smith**: Writing – review & editing. **Stephen P. Mackessy**: Writing – review & editing. **Todd A. Castoe**: Conceptualization, Supervision, Writing – original draft.

Data availability

All raw data generated for and used in this study have been submitted to the NCBI Sequence Read Archive under the BioProject accession PRJNA837532. The assembled *C. v. viridis* venom gland Iso-Seq transcriptome has additionally been submitted to the NCBI Transcriptome Shotgun Assembly under accession GJZK00000000. Accessions for all data used in this study can be found in Supplementary Table 7.

Ethical statement

All tissues utilized in this study were sampled in accordance with protocol 2004D-SM-S-23 (S.P Mackessy) approved by the University of Northern Colorado Institutional Animal Care and Use Committee and under scientific collecting permits from Colorado Parks and Wildlife (21HP0974 to S.P Mackessy).

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Acknowledgments

Support for this work was provided by National Science Foundation (NSF) grants IOS-655735 to TAC, and DEB-1655571to TAC and SPM.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.toxicon.2022.06.014.

References

Aird, S.D., Arora, J., Barua, A., Qiu, L., Terada, K., Mikheyev, A.S., 2017. Population genomic analysis of a pitviper reveals microevolutionary forces underlying venom chemistry. Genome Biol. Evol. 9, 2640–2649.

Allen, H.R., Merchant, M.L., Tucker, R.K., Fox, J.W., Geren, C.R., 1996. Characterization and chemical modification of E toxin isolated from timber rattlesnake (*Crotalus horridus horridus*) venom. J. Nat. Toxins 5.

Almeida, J.R., Resende, L.M., Watanabe, R.K., Carregari, V.C., Huancahuire-Vega, S., da S Caldeira, C.A., Coutinho-Neto, A., Soares, A.M., Vale, N., de C, G., 2017. Snake venom peptides and low mass proteins: molecular tools and therapeutic agents. Curr. Med. Chem. 24, 3254–3282.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andrews, S.J., Rothnagel, J.A., 2014. Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. 15, 193–204.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., 2014. Identification of small ORF s in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 33, 981–993.

Bieber, A.L., McParland, R.H., Becker, R.R., 1987. Amino acid sequences of myotoxins from Crotalus viridis concolor venom. Toxicon 25, 677–680.

Bober, M.A., Glenn, J.L., Straight, R.C., Ownby, C.L., 1988. Detection of myotoxin a-like proteins in various snake venoms. Toxicon 26, 665–673.

Boldrini-França, J., Corrêa-Netto, C., Silva, M.M.S., Rodrigues, R.S., De La Torre, P., Pérez, A., Soares, A.M., Zingali, R.B., Nogueira, R.A., Rodrigues, V.M., Sanz, L., Calvete, J.J., 2010. Snake venomics and antivenomics of *Crotalus durissus* subspecies from Brazil: assessment of geographic variation and its implication on snakebite management. J. Proteomics 73, 1758–1776.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527.

Burset, M., Seledtsov, I.A., Solovyev, V.V., 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28, 4364–4375.

Cai, T., Zhang, Q., Wu, B., Wang, J., Li, N., Zhang, T., Wang, Z., Luo, J., Guo, X., Ding, X., 2021. LncRNA-encoded microproteins: a new form of cargo in cell culture-derived and circulating extracellular vesicles. J. Extracell. Vesicles 10, e12123.

Cameron, D.L., Tu, A.T., 1977. Characterization of myotoxin a from the venom of prairie rattlesnake (*Crotalus viridis viridis*). Biochemistry 16, 2546–2553.

Campeiro, J.D., Marinovic, M.P., Carapeto, F.C., Dal Mas, C., Monte, G.G., Carvalho Porta, L., Nering, M.B., Oliveira, E.B., Hayashi, M.A.F., 2018. Oral treatment with a rattlesnake native polypeptide crotamine efficiently inhibits the tumor growth with no potential toxicity for the host animal and with suggestive positive effects on animal metabolic profile. Amino Acids 50, 267–278.

Cardoso, K.C., Da Silva, M.J., Costa, G.G., Torres, T.T., Del Bem, L.E.V., Vidal, R.O., Menossi, M., Hyslop, S., 2010. A transcriptomic analysis of gene expression in the venom gland of the snake *Bothrops alternatus* (urutu). BMC Genom. 11, 605.

Castoe, T.A., Daza, J.M., Smith, E.N., Sasa, M.M., Kuch, U., Campbell, J.A., Chippindale, P.T., Parkinson, C.L., 2009. Comparative phylogeography of pitvipers suggests a consensus of ancient Middle American highland biogeography. J. Biogeogr. 36, 88–103.

Chen, L.-L., 2016. Linking long noncoding RNA localization and function. Trends Biochem. Sci. 41, 761–772.

Colis-Torres, A., Neri-Castro, E., Strickland, J.L., Olvera-Rodríguez, A., Borja, M., Calvete, J., Jones, J., Parkinson, C.L., Bañuelos, J., López de León, J., Alagón, A., 2021. Intraspecific Venom Variation of Mexican West Coast Rattlesnakes (*Crotalus basiliscus*) and its Implications for Antivenom Production. Biochimie.

Core Team, R., 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. http://www.R-project.org/.

Coronado, M.A., Gabdulkhakov, A., Georgieva, D., Sankaran, B., Murakami, M.T., Arni, R.K., Betzel, C., 2013. Structure of the polypeptide crotamine from the Brazilian rattlesnake Crotalus durissus terrificus. Acta Crystallogr. D Biol. Crystallogr. 69, 1958–1964.

Correa, P.G., Oguiura, N., 2013. Phylogenetic analysis of β-defensin-like genes of *Bothrops, Crotalus* and *Lachesis* snakes. Toxicon 69, 65–74.

Costa, B.A., Sanches, L., Gomide, A.B., Bizerra, F., Dal Mas, C., Oliveira, E.B., Perez, K.R., Itri, R., Oguiura, N., Hayashi, M.A.F., 2014. Interaction of the rattlesnake toxin crotamine with model membranes. J. Phys. Chem. B 118, 5471–5479.

Dalla Valle, L., Benato, F., Maistro, S., Quinzani, S., Alibardi, L., 2012. Bioinformatic and molecular characterization of beta-defensins-like peptides isolated from the green lizard *Anolis carolinensis*. Dev. Comp. Immunol. 36, 222–229.

de Carvalho Porta, L., Fadel, V., D'Arc Campeiro, J., Oliveira, E.B., Godinho, R.O., Hayashi, M.A.F., 2020. Biophysical and pharmacological characterization of a full-length synthetic analog of the antitumor polypeptide crotamine. J. Mol. Med. 98, 1561–1571.

de LM Junqueira-de-Azevedo, I., Bastos, C.M.V., Ho, P.L., Luna, M.S., Yamanouye, N., Casewell, N.R., 2015. Venom-related transcripts from *Bothrops jararaca* tissues provide novel molecular insights into the production and evolution of snake venom. Mol. Biol. Evol. 32, 754–766.

de LM Junqueira-de-Azevedo, I., Ho, P.L., 2002. A survey of gene expression and diversity in the venom glands of the pitviper snake *Bothrops insularis* through the generation of expressed sequence tags (ESTs). Gene 299, 279–291.

de Oliveira, Y.S., Corrêa, P.G., Oguiura, N., 2018. Beta-defensin genes of the Colubridae snakes *Phalotris mertensi, Thamnodynastes hypoconia*, and *T. strigatus*. Toxicon 146, 124–128.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Dowell, N.L., Giorgianni, M.W., Griffin, S., Kassner, V.A., Selegue, J.E., Sanchez, E.E., Carroll, S.B., 2018. Extremely divergent haplotypes in two toxin gene complexes encode alternative venom types within rattlesnake species. Curr. Biol. 28, 1016–1026 e4.

Durban, J., Pérez, A., Sanz, L., Gómez, A., Bonilla, F., Rodríguez, S., Chacón, D., Sasa, M., Angulo, Y., Gutiérrez, J.M., 2013. Integrated "omics" profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, *Crotalus simus simus*. BMC Genom. 14, 1–17.

Durban, J., Sanz, L., Trevisan-Silva, D., Neri-Castro, E., Alagón, A., Calvete, J.J., 2017. Integrated venomics and venom gland transcriptome analysis of juvenile and adult Mexican rattlesnakes *Crotalus simus, C. tzabcan*, and *C. culminatus* Revealed miRNA-modulated Ontogenetic Shifts. J. Proteome Res. 16, 3370–3390.

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., Sali, A., 2008. Protein structure modeling with MODELLER. In: Structural Proteomics. Springer, pp. 145–159.

Fadel, V., Bettendorff, P., Herrmann, T., de Azevedo Jr., W.F., Oliveira, E.B., Yamane, T., Wüthrich, K., 2005. Automated NMR structure determination and disulfide bond identification of the myotoxin crotamine from *Crotalus durissus terrificus*. Toxicon 46, 759–767.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37.

Fletcher, J.E., Hubert, M., Wieland, S.J., Gong, Q.-H., Jiang, M.-S., 1996. Similarities and differences in mechanisms of cardiotoxins, melittin and other myotoxins. Toxicon 34, 1301–1311.

Froy, O., Hananel, A., Chapnik, N., Madar, Z., 2005. Differential expression of rat β-defensins. IUBMB Life 57, 41–43.

Fry, B.G., 2005. From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res. 15, 403–420.

Ganz, T., Lehrer, R.I., 1994. Defensins. Curr. Opin. Immunol. 6, 584–589.

Gearing, L.J., Cumming, H.E., Chapman, R., Finkel, A.M., Woodhouse, I.B., Luu, K., Gould, J.A., Forster, S.C., Hertzog, P.J., 2019. CiiiDER: a tool for predicting and analysing transcription factor binding sites. PLoS One 14, e0215495.

Gel, B., Serra, E., 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33, 3088–3090.

Griffin, P.R., Aird, S.D., 1990. A new small myotoxin from the venom of the prairie rattlesnake (*Crotalus viridis viridis*). FEBS Lett. 274, 43–47.

Gutiérrez, J., Lomonte, B., 1995. Phospholipase A2 myotoxins from *Bothrops* snake venoms. Toxicon 33, 1405–1424.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227.

Hargreaves, A.D., Swain, M.T., Hegarty, M.J., Logan, D.W., Mulley, J.F., 2014. Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. Genome Biol. Evol. 6, 2088–2095.

Hayashi, M.A., Nascimento, F.D., Kerkis, A., Oliveira, V., Oliveira, E.B., Pereira, A., Radis-Baptista, G., Nader, H.B., Yamane, T., Kerkis, I., 2008. Cytotoxic effects of crotamine are mediated through lysosomal membrane permeabilization. Toxicon 52, 508–517.

Hayashi, M.A., Oliveira, E.B., Kerkis, I., Karpel, R.L., 2012. Crotamine: a novel cell-penetrating polypeptide nanocarrier with potential anti-cancer and biotechnological applications. In: Nanoparticles in Biology and Medicine. Springer, pp. 337–352.

Hofmann, E.P., Rautsaw, R.M., Strickland, J.L., Holding, M.L., Hogan, M.P., Mason, A.J., Rokyta, D.R., Parkinson, C.L., 2018. Comparative venom-gland transcriptomics and venom proteomics of four Sidewinder Rattlesnake (*Crotalus cerastes*) lineages reveal little differential expression despite individual variation. Sci. Rep. 8, 1–15.

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure

prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L.E., Brookes, D.H., Wilson, L., Chen, J., Liles, K., 2018. Improvements to the APBS biomolecular solvation software suite. Protein Sci. 27, 112–128.
Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., Gao, G., 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45, W12–W16.
Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.
Kerkis, I., Hayashi, M.A., Prieto da Silva, A.R., Pereira, A., De Sa Junior, P.L., Zaharenko, A.J., Rádis-Baptista, G., Kerkis, A., Yamane, T., 2014. State of the art in the studies on crotamine, a cell penetrating peptide from South American rattlesnake. BioMed Res. Int. 2014, 1–9.
Keyler, D.E., Saini, V., O'Shea, M., Gee, J., Smith, C.F., Mackessy, S.P., 2020. *Crotalus oreganus concolor*: envenomation case with venom analysis and a diagnostic conundrum of myoneurologic symptoms. Wilderness Environ. Med. 31, 220–225.
Kolde, R., 2015. Pheatmap: Pretty Heatmaps. R Package version 1.0.12. https://CRAN.R-project.org/package=pheatmap.
Lachumanan, R., Armugam, A., Tan, C.-H., Jeyaseelan, K., 1998. Structure and organization of the cardiotoxin genes in *Naja naja sputatrix*. FEBS Lett. 433, 119–124.
Laure, C.J., 1975. Die primärstruktur des crotamins. Hoppe Seylers Z Physiol Chem 356, 213–215.
Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. ArXiv Prepr. ArXiv13033997.
Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.
Loehlin, D.W., Carroll, S.B., 2016. Expression of tandem gene duplicates is often greater than twofold. Proc. Natl. Acad. Sci. 113, 5988–5992.
Lomonte, B., Gutiérrez, J., Furtado, M., Otero, R., Rosso, J.-P., Vargas, O., Carmona, E., Rovira, M.E., 1990. Isolation of basic myotoxins from *Bothrops moojeni* and *Bothrops atrox* snake venoms. Toxicon 28, 1137–1146.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 1–21.
Mackessy, S.P., Saviola, A.J., 2016. Understanding Biological Roles of Venoms Among the Caenophidia: the Importance of Rear-Fanged Snakes.
Mackessy, S.P., Williams, K., Ashton, K.G., 2003. Ontogenetic variation in venom composition and diet of *Crotalus oreganus concolor*: a case of venom paedomorphosis? Copeia 769–782, 2003.
Maeda, N., Tamiya, N., Pattabhiraman, T.R., Russell, F.E., 1978. Some chemical properties of the venom of the rattlesnake, *Crotalus viridis helleri*. Toxicon 16, 431–441.
Maraganore, J.M., Merutka, G., Cho, W., Welches, W., Kezdy, F.J., Heinrikson, R.L., 1984. A new class of phospholipases A2 with lysine in place of aspartate 49. Functional consequences for calcium and substrate binding. J. Biol. Chem. 259, 13839–13843.
Margres, M.J., Bigelow, A.T., Lemmon, E.M., Lemmon, A.R., Rokyta, D.R., 2017. Selection to increase expression, not sequence diversity, precedes gene family origin and expansion in rattlesnake venom. Genetics 206, 1569–1580.
Margres, M.J., Rautsaw, R.M., Strickland, J.L., Mason, A.J., Schramer, T.D., Hofmann, E.P., Stiers, E., Ellsworth, S.A., Nystrom, G.S., Hogan, M.P., 2021. The Tiger Rattlesnake genome reveals a complex genotype underlying a simple venom phenotype. Proc. Natl. Acad. Sci. 118.
Meader, S., Hillier, L.W., Locke, D., Ponting, C.P., Lunter, G., 2010. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res. 20, 675–684.
Mebs, D., Ownby, C.L., 1990. Myotoxic components of snake venoms: their biochemical and biological activities. Pharmacol. Ther. 48, 223–236.
Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.
Nozawa, M., Kawahara, Y., Nei, M., 2007. Genomic drift and copy number variation of sensory receptor genes in humans. Proc. Natl. Acad. Sci. 104, 20421–20426.
Oguiura, N., Boni-Mitake, M., Rádis-Baptista, G., 2005. New view on crotamine, a small basic polypeptide myotoxin from South American rattlesnake venom. Toxicon 46, 363–370.
Oguiura, N., Collares, M.A., Furtado, M.F.D., Ferrarezzi, H., Suzuki, H., 2009. Intraspecific variation of the crotamine and crotasin genes in *Crotalus durissus* rattlesnakes. Gene 446, 35–40.
Oguiura, N., Boni-Mitake, M., Affonso, R., Zhang, G., 2011. In vitro antibacterial and hemolytic activities of crotamine, a small basic myotoxin from rattlesnake *Crotalus durissus*. J. Antibiot. (Tokyo) 64, 327–331.
Ownby, C.L., 1998. Structure, function and biophysical aspects of the myotoxins from snake Venoms. J. Toxicol. Toxin Rev. 17, 213–238.
Ownby, C.L., Fletcher, J.E., Colberg, T.R., 1993. Cardiotoxin 1 from cobra (*Naja naja atra*) venom causes necrosis of skeletal muscle in vivo. Toxicon 31, 697–709.
Pasquesi, G.I., Adams, R.H., Card, D.C., Schield, D.R., Corbin, A.B., Perry, B.W., Reyes-Velasco, J., Ruggiero, R.P., Vandewege, M.W., Shortt, J.A., 2018. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat. Commun. 9, 1–11.
Passero, L.F.D., Tomokane, T.Y., Corbett, C.E.P., Laurenti, M.D., Toyama, M.H., 2007. Comparative studies of the anti-leishmanial activity of three *Crotalus durissus* ssp. venoms. Parasitol. Res. 101, 1365–1371.
Payre, F., Desplan, C., 2016. Small peptides control heart activity. Science 351, 226–227.
Perry, B.W., Schield, D.R., Westfall, A.K., Mackessy, S.P., Castoe, T.A., 2020. Physiological demands and signaling associated with snake venom production and storage illustrated by transcriptional analyses of venom glands. Sci. Rep. 10, 1–10.
Perry, B.W., Gopalan, S.S., Pasquesi, G.I.M., Schield, D.R., Westfall, A.K., Smith, C.F., Koludarov, I., Chippindale, P.T., Pellegrino, M.W., Chuong, E.B., Mackessy, S.P., Castoe, T.A., 2022. Snake venom gene expression is coordinated by novel regulatory architecture and the integration of multiple co-opted vertebrate pathways. Genome Res. https://doi.org/10.1101/gr.276251.121.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612.
Phillippy, A.M., Schatz, M.C., Pop, M., 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 9, 1–13.
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.
Rádis-Baptista, G., Kerkis, I., 2011. Crotamine, a small basic polypeptide myotoxin from rattlesnake venom with cell-penetrating properties. Curr. Pharm. Des. 17, 4351–4361.
Rádis-Baptista, G., Oguiura, N., Hayashi, M.A.F., Camargo, M.E., Grego, K.F., Oliveira, E.B., Yamane, T., 1999. Nucleotide sequence of crotamine isoform precursors from a single South American rattlesnake (*Crotalus durissus terrificus*). Toxicon 37, 973–984.
Rádis-Baptista, G., Kubo, T., Oguiura, N., Svartman, M., Almeida, T.M., Batistic, R.F., Oliveira, E.B., Vianna-Morgante, Â.M., Yamane, T., 2003. Structure and chromosomal localization of the gene for crotamine, a toxin from the South American rattlesnake, *Crotalus durissus terrificus*. Toxicon 42, 747–752.
Rádis-Baptista, G., Kubo, T., Oguiura, N., Da Silva, A.P., Hayashi, M.A.F., Oliveira, E.B., Yamane, T., 2004. Identification of crotasin, a crotamine-related gene of Crotalus durissus terrificus. Toxicon 43, 751–759.
Ramsköld, D., Wang, E.T., Burge, C.B., Sandberg, R., 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5, e1000598.
Razooky, B.S., Obermayer, B., O'May, J.B., Tarakhovsky, A., 2017. Viral infection identifies micropeptides differentially regulated in smORF-containing lncRNAs. Genes 8, 206.
Reyes-Velasco, J., Card, D.C., Andrew, A.L., Shaney, K.J., Adams, R.H., Schield, D.R., Casewell, N.R., Mackessy, S.P., Castoe, T.A., 2015. Expression of venom gene homologs in diverse Python tissues suggests a new model for the evolution of snake venom. Mol. Biol. Evol. 32, 173–183.
Rizzi, C.T., Carvalho-de-Souza, J.L., Schiavon, E., Cassola, A.C., Wanke, E., Troncone, L.R.P., 2007. Crotamine inhibits preferentially fast-twitching muscles but is inactive on sodium channels. Toxicon 50, 553–562.
Robinson, E.K., Covarrubias, S., Carpenter, S., 2020. The how and why of lncRNA function: an innate immune perspective. Biochim. Biophys. Acta BBA-Gene Regul. Mech. 1863, 194419.
Rokyta, D.R., Wray, K.P., Lemmon, A.R., Lemmon, E.M., Caudle, S.B., 2011. A high-throughput venom-gland transcriptome for the Eastern Diamondback Rattlesnake (Crotalus adamanteus) and evidence for pervasive positive selection across toxin classes. Toxicon 57, 657–671.
Rokyta, D.R., Lemmon, A.R., Margres, M.J., Aronow, K., 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genom. 13, 1–23.
Ruby, J.G., Bellare, P., DeRisi, J.L., 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 Genes Genomes Genet. 3, 865–880.
Samejima, Y., Aoki, Y., Mebs, D., 1991. Amino acid sequence of a myotoxin from venom of the eastern diamondback rattlesnake (*Crotalus adamanteus*). Toxicon 29, 461–468.
Savard, J., Marques-Souza, H., Aranda, M., Tautz, D., 2006. A segmentation gene in *Tribolium* produces a polycistronic mRNA that codes for multiple conserved peptides. Cell 126, 559–569.
Saviola, A.J., Pla, D., Sanz, L., Castoe, T.A., Calvete, J.J., Mackessy, S.P., 2015. Comparative venomics of the Prairie Rattlesnake (*Crotalus viridis viridis*) from Colorado: identification of a novel pattern of ontogenetic changes in venom composition and assessment of the immunoreactivity of the commercial antivenom CroFab®. J. Proteomics 121, 28–43.
Schenberg, S., 1959. Geographical pattern of crotamine distribution in the same rattlesnake subspecies. Science 129, 1361–1363.
Schield, D.R., Card, D.C., Hales, N.R., Perry, B.W., Pasquesi, G.M., Blackmon, H., Adams, R.H., Corbin, A.B., Smith, C.F., Ramesh, B., 2019. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. Genome Res. 29, 590–601.
Schield, D.R., Pasquesi, G.I., Perry, B.W., Adams, R.H., Nikolakis, Z.L., Westfall, A.K., Orton, R.W., Meik, J.M., Mackessy, S.P., Castoe, T.A., 2020. Snake recombination landscapes are concentrated in functional regions despite PRDM9. Mol. Biol. Evol. 37, 1272–1294.
Schield, D.R., Perry, B.W., Nikolakis, Z.L., Mackessy, S.P., Castoe, T.A., 2021. Population genomic analyses confirm male-biased mutation rates in snakes. J. Hered. 112, 221–227.
Schröder, J.-M., Harder, J., 1999. Human beta-defensin-2. Int. J. Biochem. Cell Biol. 31, 645–651.
Schutte, B.C., Mitros, J.P., Bartlett, J.A., Walters, J.D., Jia, H.P., Welsh, M.J., Casavant, T.L., McCray, P.B., 2002. Discovery of five conserved β-defensin gene clusters using a computational search strategy. Proc. Natl. Acad. Sci. 99, 2129–2133.
Shafee, T.M.A., Lay, F.T., Hulett, M.D., Anderson, M.A., 2016. The defensins consist of two independent, convergent protein superfamilies. Mol. Biol. Evol. 33, 2345–2356.
Soneson, C., Love, M.I., Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4.

Straight, R.C., Glenn, J.L., Wolt, T.B., Wolfe, M.C., 1991. Regional differences in content of small basic peptide toxins in the venoms of *Crotalus adamanteus* and *Crotalus horridus*. Comp. Biochem. Physiol. B 100, 51–58.
Sun, Q., Hao, Q., Prasanth, K.V., 2018. Nuclear long noncoding RNAs: key regulators of gene expression. Trends Genet. TIG 34, 142–157.
Suryamohan, K., Krishnankutty, S.P., Guillory, J., Jevit, M., Schröder, M.S., Wu, M., Kuriakose, B., Mathew, O.K., Perumal, R.C., Koludarov, I., 2020. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. Nat. Genet. 52, 106–117.
Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gíslason, M.H., Pihl, S.I., Tsirigos, K.D., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat. Biotechnol. 1–3.
Utaisincharoen, P., Baker, B., Tu, A.T., 1991. Binding of myotoxin a to sarcoplasmic reticulum calcium-ATPase: a structural study. Biochemistry 30, 8211–8216.
von Heijne, G., 1990. The signal peptide. J. Membr. Biol. 115, 195–201.
Whittington, C.M., Papenfuss, A.T., Bansal, P., Torres, A.M., Wong, E.S., Deakin, J.E., Graves, T., Alsop, A., Schatzkamer, K., Kremitzki, C., 2008. Defensins and the convergent evolution of platypus and reptile venom genes. Genome Res. 18, 986–994.
Wong, E.S.W., Belov, K., 2012. Venom evolution through gene duplications. Gene 496, 1–7.
Wu, Z., Liu, X., Liu, L., Deng, H., Zhang, J., Xu, Q., Cen, B., Ji, A., 2014. Regulation of lncRNA expression. Cell. Mol. Biol. Lett. 19, 561–575.
Xiao, Y., Hughes, A.L., Ando, J., Matsuda, Y., Cheng, J.-F., Skinner-Noble, D., Zhang, G., 2004. A genome-wide screen identifies a single β-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins. BMC Genom. 5, 1–11.
Xie, C., Tammi, M.T., 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinf. 10, 1–9.
Yoshizumi, K., Liu, S.-Y., Miyata, T., Saita, S., Ohno, M., Iwanaga, S., Kihara, H., 1990. Purification and amino acid sequence of basic protein I, a lysine-49-phospholipase A2 with low activity, from the venom of *Trimeresurus flavoviridis* (Habu snake). Toxicon 28, 43–54.
Yount, N.Y., Kupferwasser, D., Spisni, A., Dutz, S.M., Ramjan, Z.H., Sharma, S., Waring, A.J., Yeaman, M.R., 2009. Selective reciprocity in antimicrobial activity versus cytotoxicity of hBD-2 and crotamine. Proc. Natl. Acad. Sci. 106, 14972–14977.
Zeng, C., Hamada, M., 2018. Identifying sequence features that drive ribosomal association for lncRNA. BMC Genom. 19, 41–49.
Zhao, C., Wang, I., Lehrer, R.I., 1996. Widespread expression of beta-defensin hBD-1 in human secretory glands and epithelial cells. FEBS Lett. 396, 319–322.
Zou, J., Mercier, C., Koussounadis, A., Secombes, C., 2007. Discovery of multiple beta-defensin like homologues in teleost fish. Mol. Immunol. 44, 638–647.