Chapter 6

Exploring Toxin Evolution: Venom Protein Transcript Sequencing and Transcriptome-Guided High-Throughput Proteomics

Cassandra M. Modahl, Jordi Durban, and Stephen P. Mackessy

Abstract

Studying animal toxin evolution requires sequences of these proteins and peptides, and transcript sequences allow for the construction of cladograms and evaluation of selection pressures from nonsynonymous and synonymous nucleotide mutation ratios. In addition, these translated sequences can be useful as custom databases for peptide identifications within venoms and for better proteomic quantification. Obtaining these transcripts is achieved by sequencing cDNA originating from venom gland tissue or venom. This chapter provides the methodology for (1) targeted sequencing of transcripts from a single venom protein family (RNA isolation and 3′RACE [rapid amplification of cDNA ends]), (2) generation of a venom gland transcriptome with next-generation sequencing (NGS) technology (de novo transcriptome assembly, toxin transcript identification, quantification, and positive selection analysis), and (3) combined high-throughput proteomics to identify secreted venom components. Transcriptomics has become fundamental for studying toxin evolution, but it creates many challenges for scientists who are unfamiliar with working with RNA, managing large NGS datasets and executing the required programs, particularly considering that there is an overabundance of available software in this field and not all perform optimally for venom gland transcriptome assembly. This chapter provides one pipeline for the integration of both low- and high-throughput transcriptomics with proteomics to characterize venoms.

Key words Venomics, Transcriptomics, Proteomics, Toxin evolution, 3′RACE, Next-generation sequencing, Bioinformatics

1 Introduction

The definition of a venom is "a secretion, delivered from one animal to another through the infliction of a wound, that contains molecular compounds (mainly peptides and proteins) to disrupt normal physiological or biochemical processes" [1]. Animal venoms are an ideal model of adaptive molecular evolution, where phenotypes can be directly linked to genetic change over time, whether in the form of rapid gene gain and loss [2–5] or nucleotide substitutions within gene sequences that alter toxin protein products [6–8]. Toxin gene

Avi Priel (ed.), *Snake and Spider Toxins: Methods and Protocols*, Methods in Molecular Biology, vol. 2068,
https://doi.org/10.1007/978-1-4939-9845-6_6, © Springer Science+Business Media, LLC, part of Springer Nature 2020

duplications result in multigene families evolving through a "birth and death" mode of evolution [9]. Analyses that involve toxin gene transcription are useful to evaluate selection pressures on these gene copies, addressing both toxin expression and diversity [10]. The field of transcriptomics offers many technologies and methodologies for these explorations.

Once toxin transcript sequences are obtained, the translated products can be predicted and structure modeling performed, as well as toxin peptides synthesized or proteins recombinantly produced for characterization. Therefore, toxin transcript sequences provide insight into not only sequence diversity, but also structure and potential function of the protein products. Further, the collection of translated sequences can be used as a custom database for proteomic identification and characterization of venoms, especially in cases where venoms contain unknown, hypervariable, or novel components that are not present in currently available databases. This integrated "omic" approach has been termed "venomics" [11, 12], and has been very successful at identifying and quantifying distinct proteoforms within a venom [13–15]. In addition to this chapter, Kaas and Craik [16] is a recommended review of this field.

There are two basic approaches to sequencing toxin transcripts: (1) sequencing the collection of expressed toxin transcripts within venom gland tissue (complete transcriptome) or (2) targeted amplification of toxin transcripts belonging to a select venom protein superfamily. This chapter provides a methodology for isolating total RNA from venom gland tissue or venom with yields useful for both target transcript amplification and next-generation sequencing (NGS) transcriptome assembly. For obtaining targeted venom protein transcripts, a protocol for 3′RACE (rapid amplification of cDNA ends), cloning, and Sanger sequencing is provided. For obtaining a complete RNA-seq transcriptome, a bioinformatics pipeline detailing read quality evaluation and processing, de novo transcriptome assembly, toxin transcript identification, gene expression quantification, protein sequence prediction, and positive selection analysis is given. Further, the use of de novo transcriptome assembly-predicted protein sequences as a custom reference for the integration of high-throughput proteomics to characterize animal venoms is discussed. These methods are applicable not only for scientists interested in venom gland transcriptome and venom proteome profiling, but also for investigations of transcriptomes/proteomes of various animal tissues.

2 Materials

2.1 RNA Isolation

1. TRIzol (Invitrogen®) or RNAzol (Sigma Aldrich®); the RNAzol protocol will be different than provided here, but it is ideal if a researcher wants to avoid the use of chloroform.

2.2 3′RACE (Rapid Amplification of cDNA Ends)

1. 3′RACE system for rapid amplification of cDNA ends (ThermoFisher Scientific®).
2. Venom protein superfamily-specific primer: Refer to **Note 1** for primer design.
3. Polymerase High Fidelity Supermix (ThermoFisher Scientific®) or any other proofreading polymerase mix.
4. Wizard SV gel and PCR cleanup system (Promega®) or any other PCR product gel purification kit.
5. pGEM-T Easy Vector System (Promega®) or a similar ligation/vector system.
6. *Escherichia coli* DH5α competent cells (ThermoFisher Scientific®) or any other competent cell line that can be used for subcloning.
7. LB broth.
8. Agar plates: 1 μL per 1 mL agar of 50 mg/mL X-gal in DMF or DMSO, 1 μL per 1 mL agar of 100 mg/mL ampicillin in ddH_2O, and 0.5 μL per/mL agar of 100 mM IPTG in ddH_2O, if using pGEM-T Easy Vector System and *E. coli* DH5α competent cells; refer to **Note 2** for agar additive preparation.
9. Quick Clean 5 M Miniprep kit (Genscript®) or similar plasmid purification kit.

2.3 Next-Generation Sequencing (NGS) Transcriptomics

2.3.1 NGS Library Preparation and Data Generation

1. TruSeq RNA Library Prep kit (Illumina®) for MiSeq, HiSeq, or NextSeq platforms, or a similar kit matching the technology to be used.
2. High-throughput computing resources are required for transcriptomic work. Usually a GNU/Linux workstation is used, as most software are for this platform. Multiple central processing units (CPUs) are ideal (at least 8), but in the case of transcriptome assembly lots of memory, both RAM and storage, is vital. In terms of storage, one lane of Illumina HiSeq data can be roughly 100–150 GB, and this can quickly be doubled as multiple files are generated during assembly. Additionally, large databases might need to be locally installed for BLAST+ searches. Transcriptome assembly software, such as Trinity [17], can be very memory intensive. Roughly 1G of RAM must be available for each one million reads for Trinity. A high-throughput computer with 256 GB RAM and at least 1 TB hard disk drive (HDD) storage are best. Gaining access to remote servers with this capacity is also an alternative, and most universities and research institutions have this available.

2.3.2 NGS Data Quality Checks

1. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [18].
2. Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) [19].
3. PEAR (https://sco.h-its.org/exelixis/web/software/pear/) [20].
4. FLASH (https://ccb.jhu.edu/software/FLASH/) [21].

2.3.3 De Novo Transcriptome Assembly

1. Trinity (https://github.com/trinityrnaseq/trinityrnaseq/wiki) [17].
2. Extender [22], not open source.
3. VTBuilder [23], not open source.
4. EvidentialGene (http://arthropods.eugenes.org/EvidentialGene/trassembly.html) [24].
5. Exonerate (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) [25].
6. CD-HIT (http://weizhongli-lab.org/cd-hit/) [26].

2.3.4 Toxin Gene Identification and Expression Quantification

1. BLAST+ (https://www.ncbi.nlm.nih.gov/books/NBK279690/) [27].
2. DIAMOND (https://ab.inf.uni-tuebingen.de/software/diamond) [28].
3. SignalP (http://www.cbs.dtu.dk/services/SignalP/) [29].
4. TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) [30].
5. RSEM (https://github.com/deweylab/RSEM) [31].
6. Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [32].

2.4 Toxin Selection

1. AliView (https://github.com/AliView/AliView) [33].
2. SeaView (http://doua.prabi.fr/software/seaview) [34].
3. Jalview (http://www.jalview.org/) [35].
4. PartitionFinder (http://www.robertlanfear.com/partitionfinder/).
5. Jmodeltest (https://github.com/ddarriba/jmodeltest2).
6. MEGA (https://www.megasoftware.net) [36].
7. PAML (http://abacus.gene.ucl.ac.uk/software/paml.html) [37].
8. DataMonkey server (http://www.datamonkey.org) [38].

2.5 High-Throughput Proteomics Integration

1. Scaffold (https://www.proteomesoftware.com/products/scaffold/) [39], licensed.
2. ProteinPilot (https://sciex.com/products/software/proteinpilot-software), licensed.

3. PEAKS (http://www.bioinfor.com/peaks-studio/) [40], licensed.
4. SearchGUI (http://compomics.github.io/projects/searchgui.html) [41].

3 Methods

3.1 RNA Isolation

This procedure is for isolating RNA from either venom or venom gland tissue. For RNA extraction from venom, the best results have been achieved using freshly extracted venom, but RNA has also been extracted from lyophilized venom after over 20 years of storage [42]. It is important to follow proper procedures when working with RNA to maximize yield and preserve RNA integrity (*see* **Note 3** for suggestions to optimize RNA work).

1. Add 100–500 μL of liquid venom or 2 mg of lyophilized venom (as low as 1 mg and up to 50 mg of lyophilized venom have been used successfully) to 1 mL TRIzol. If venom gland tissue is used, approximately 10–100 mg of tissue is added and homogenized in TRIzol (this can be done with sterile tissue grinders).
2. Incubate sample for 5 min at room temperature.
3. Add 200 μL of chloroform.
4. Cap tightly and shake for 15 s.
5. Incubate for 2–3 min at room temperature.
6. Centrifuge sample at 12,000 × *g* at 4 °C for 15 min. Remove the sample from the centrifuge, taking care not to disrupt the layers that have separated.
7. Remove the aqueous upper phase (should be about 50% of the total volume) by pipetting the solution out and into a new RNase-free microcentrifuge tube. Do not remove any of the organic layer or interphase layer—only the top layer.
8. Add 500 μL of 100% isopropanol to the aqueous layer in the new tube.
9. Incubate at room temperature for 10 min.
10. Centrifuge at 12,000 × *g* at 4 °C for 10 min.
11. Remove supernatant, leaving RNA pellet (might not be visible for venom, should be visible for tissue).
12. Wash pellet with 1 mL of 75% ethanol.
13. Centrifuge the tube at 7500 × *g* at 4 °C for 5 min and pour off supernatant.
14. Add 300 μL ice cold 100% ethanol and 40 μL 3 M sodium acetate.

15. Finger vortex and place in −20 °C overnight.
16. Centrifuge samples at 10,000 × *g* for 15 min at 4 °C.
17. Remove supernatant, invert over Kimwipe to remove all liquid, and air dry for 10 min.
18. Add 10–16 μL of RNase-free water and gently vortex. *See* **Note 4** for working with RNA from rear-fanged snake venoms that will need an additional next step.

3.2 3′RACE (Rapid Amplification of cDNA Ends): Targeting Specific Toxin Transcripts

3′RACE is usually performed using protocols and reagents that are supplied with kits. The 3′RACE kit sold by ThermoFisher Scientific® has been routinely used in our lab and the following protocol details the use of this kit, but is slightly modified from the kit manual (Fig. 1). Before beginning the procedure, make sure that a heat block has been set to 70 °C and a water bath has been set to 42 °C. A 37 °C incubator is also needed for *E. coli* growth.

1. Adaptor primers, 0.5 μL (provided with the ThermoFisher Scientific® 3′RACE kit), are combined with 1–5 μg of total RNA (if you are unsure of the concentration, use 5.5 μL), in a

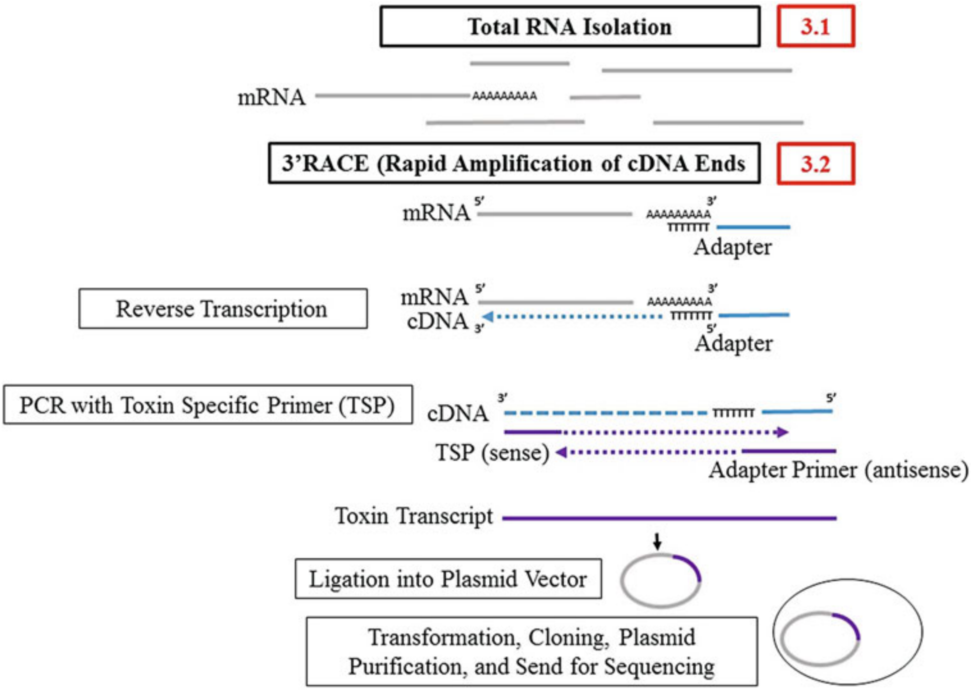
Total RNA Isolation
3.1
mRNA
AAAAAAAAA
3'RACE (Rapid Amplification of cDNA Ends
3.2
mRNA
5'
3'
AAAAAAAAA
TTTTTTT
Adapter
Reverse Transcription
mRNA
cDNA
5'
3'
3'
5'
Adapter
PCR with Toxin Specific Primer (TSP)
cDNA
3'
5'
TSP (sense)
Adapter Primer (antisense)
Toxin Transcript
Ligation into Plasmid Vector
Transformation, Cloning, Plasmid Purification, and Send for Sequencing

Fig. 1 Protocol overview for targeting specific toxin transcripts for sequencing. Protocol overview shows each step to be performed for targeted amplification of transcripts within a specific venom protein superfamily, as completed in a recent publication [42]. It is modified from the 3′RACE system for rapid amplification of cDNA ends kit manual, sold by ThermoFisher Scientific, and includes additional steps for Sanger sequencing preparation. Procedures discussed in the text are indicated by section numbers (red boxes)

total volume of 6 μL in a 0.5 mL RNase-free microcentrifuge tube.

2. Heat for 10 min at 70 °C and immediately chill on ice for 2 min.
3. Add the following to each tube (these can be mixed together in a master mix and 3.5 μL of the master mix used for each tube); all reagents are supplied with the kit:

   1 μL 10× PCR buffer (200 mM Tris–HCl, pH 8.4, 500 mM KCl)

   1 μL 25 mM $MgCl_2$

   1 μL 0.1 M DTT

   0.5 μL dNTP mix (10 mM each dNTP)
4. Mix components gently and centrifuge. Equilibrate each tube at 42 °C for 2–5 min.
5. Add 0.5 μL of SuperScript™ II Reverse Transcriptase (200 units/μL) to each tube (pipette this into the solution well).
6. Incubate at 42 °C for 50 min (can be done in a water bath or in a thermal cycler).
7. Terminate the reaction by incubating at 70 °C for 15 min.
8. Chill on ice and briefly centrifuge.
9. Optional: Add 0.5 μL of RNase H and incubate at 37 °C for 20 min to remove all traces of RNA in each sample. This step is required for RNA from venom of rear-fanged snakes.
10. The following should be added to a small 0.2 mL PCR tube:

    0.5 μL of sense primer (venom protein transcript specific, see Subheading 2.2)

    0.5 μL of antisense primer AUAP (Abridged Universal Amplification Primer; supplied by the kit, corresponds with kit adapters)

    1–2 μL of cDNA template, generated from reverse transcription above (works best if it is a 1:10 dilution)

    22–23 μL of Polymerase High Fidelity Supermix (this mix includes the polymerase, dNTPs, and buffer)

    Final total volume = 25 μL (can also be adjusted to have a final volume of 50 μL)
11. Tubes should be vortexed well and briefly centrifuged (quick spin).
12. Place tubes in the thermal cycler with the program below for touchdown PCR (*see* **Note 5**). Annealing temperature will vary depending on primers used, and refer to **Note 6** for PCR troubleshooting.

94 °C 5 minutes

$$\left.\begin{array}{l}94\ {}^\circ\text{C } 25 \text{ seconds}\\ 52\ {}^\circ\text{C } 30 \text{ seconds}\end{array}\right\} \text{7X}$$

68 °C 2 minutes

$$\left.\begin{array}{l}94\ {}^\circ\text{C } 25 \text{ seconds}\\ 48\ {}^\circ\text{C } 30 \text{ seconds}\end{array}\right\} \text{30X}$$

68 °C 2 minutes

68 °C 5 minutes

Hold at 4–10 °C (programing to hold at 10 °C is better for the instrument).

13. Remove tubes from the thermal cycler and either store at −20 °C or immediately run on a 1% agarose gel to view products.
14. Excise band of appropriate size (predicted from transcripts within the venom protein superfamily) from the 1% agarose gel, and isolate the DNA using a PCR product gel purification kit, such as Wizard SV gel and PCR cleanup system.
15. Perform ligation into cloning vector of choice. pGEM-T Easy Vector System or similar can be purchased and has all needed reagents for ligation. Add the following to a 0.5 mL nuclease-free tube:

    5 μL 2× Ligation buffer

    1 μL pGEM-T Easy Vector

    3 μL of PCR product DNA isolated from gel band

    1 μL of DNA ligase
16. Mix the added reagents by pipetting.
17. Incubate at 4 °C overnight.
18. Bacterial transformation is then performed with the vector ligation product. This procedure should be completed following instructions given for the chosen competent cells purchased. Agar plates with antibiotics or other additives, such as IPTG, should be prepared according to competent cells and vector being used. For *E. coli* DH5α competent cells, 5 μL of ligation product is added to 50 μL of competent cells kept on ice. Refer to **Note 7** for general bacterial work suggestions that should be followed from this point forward.

19. Flick side of tube to mix competent cells with ligation product. Do NOT vortex, as competent cells are very fragile.
20. Incubate tube on ice for 30 min.
21. Heat shock tube for 20 s at exactly 42 °C, and return to ice immediately.
22. Incubate on ice for 2 min, and then add 1 mL of LB broth.
23. Incubate for 60 min in a shaking 37 °C warm water bath.
24. Plate 200 μL onto an agar plate, spreading the bacteria with the use of sterile glass beads or loop. Make sure that the sample has dried onto the plate before overturning for incubation.
25. Turn the plate upside down and incubate at 37 °C overnight (about 16–18 h at most; otherwise plate could become overgrown).
26. Place plate at 4 °C the following day to stop *E. coli* growth. Pick *E. coli* colonies as soon as possible.
27. Pick *E. coli* colonies that demonstrate venom protein transcript insertion into vector; this is done by colony blue/white screening (LacZ gene selection) for the pGEM-T Easy Vector System. Make sure that selected colony is white in coloration in this case. Scoop the white colony up with a sterile pipette tip and place into 2 mL of LB + ampicillin broth (ampicillin is 1 μL per l mL broth). Each *E. coli* colony could be a different venom protein transcript isoform. At least ten colonies should be selected, but the greater number selected, the better chance of obtaining all transcript isoforms [42].
28. Shake at 37 °C overnight.
29. In the morning, purify the plasmids of each *E. coli* colony with the use of the Quick Clean 5 M Miniprep kit or similar plasmid purification kit.
30. Send plasmids for Sanger sequencing. Usually, only around 200 ng is needed. Sequencing primers assigned will be based upon the vector. For pGEM-T Easy Vector, T7 and SP6 can be used as sequencing primers.

3.3 Next-Generation Sequencing (NGS) Transcriptomics: Constructing De Novo Transcriptomes

Next-generation sequencing technologies have now made it more cost effective and less labor intensive to generate a venom gland transcriptome. There are several different sequencing technologies that fall under the broad term "next-generation sequencing" (NGS). These include cyclic reverse termination sequencing (Illumina®, which patented MiSeq, HiSeq, and NextSeq instruments), sequencing by ligation (Applied Biosystems ABI SOLiD® system), single-molecule real-time sequencing (Pacific Biosciences®), ion semiconductor sequencing (Ion Torrent®), and Oxford nanopore® technologies [43]. With the amount of sequence obtained from

NGS technologies, especially from short-read sequencers, a kilobase of sequence costs a fraction of a cent. The extensive number of overall sequences obtained results in the recovery of full-length transcripts after assembly, including even lowly expressed transcripts that were previously difficult to obtain with expressed sequence tags (ESTs) [44, 45].

3.3.1 NGS Library Preparation and Data Generation

Preparing cDNA libraries for NGS also requires isolating total RNA from venom gland tissue, usually at 4 days following venom extraction, when venom protein transcript expression is the highest [46]; extraneous muscle, blood, and/or connective tissues should be trimmed away from gland tissues before proceeding. Of particular importance is making certain that the tissue processed is of venom gland origin, given the sensitivity of NGS and the presence of venom protein homologs within other tissues [47–49]. The same protocol as detailed above can be used to isolate total RNA for NGS library preparation. High-quality (200 ng–1 μg) total RNA (refer to **Note 8** for RNA quality evaluation) is usually required as starting material for NGS library preparation kits.

Given the fact that over 90% of isolated RNA will be ribosomal RNA, it is important to avoid sequencing this RNA prior to the downstream bioinformatics analysis. Enriching messenger RNA is achieved either by using oligo d(T) beads or by selective removal of rRNA. Currently, rRNA depletion is biased toward model organisms (known rRNA sequences), and therefore is not a recommended procedure for non-model organism NGS library preparation.

Examples of NGS library preparation kits for MiSeq or HiSeq sequencing technologies include the TruSeq RNA Library Prep kit or NEBNext Ultra RNA Prep Kit for Illumina. It is important to use kits specific for the sequencing technology to be used. For Illumina® sequencing, these kits provide adaptors and primers needed for proper binding to sequencing flow cells and for barcoding if multiplexing (sequencing multiple samples on the same lane). These kits can be purchased and directions followed to construct in-house libraries, which usually can be completed within a day. On the other hand, RNA can be submitted to sequencing facilities/companies that will prepare the libraries for a fee.

When generating a complete transcriptome, several considerations should be taken into account:

1. Sequencing depth: the number of reads needed to achieve complete transcriptome complexity. The number has been suggested to be around 30–50 million reads for a de novo transcriptome assembly. However, considering that venom protein transcripts are usually highly expressed, 8 million reads has been suggested to assemble all abundant toxin gene transcripts [50].

2. Paired-end reads (PE) or single reads: sequencing a single end of a transcript fragment, or both ends (paired end). It is best for de novo transcriptome assemblies to have paired-end longer reads (>150 bp) since this additional information can be useful for assembly, and paired-end reads can be merged by such programs such as PEAR (Paired-End reAd mergeR) [20] or FLASH (Fast Length Adjustment of SHort reads) [21] to create overall longer reads that also improve assembly.
3. Strand information: strand origin of a read. In order to quantify gene expression accurately, it is important to retain the strand specificity of origin for each transcript. This will allow one to identify from which overlapping gene the RNA transcript has originated.

There are many steps to produce a high-quality assembly, but the assembly has many downstream applications (refer to https://omicstools.com/rna-seq-categogy), such as evaluating toxin gene expression, selection, or use of the predicted translated products as custom databases for protein identifications (Fig. 2), so accuracy should be a major goal. Command examples for some of the programs discussed in the preceding text are given (Box 1), but individual documentation for each program should be referenced.

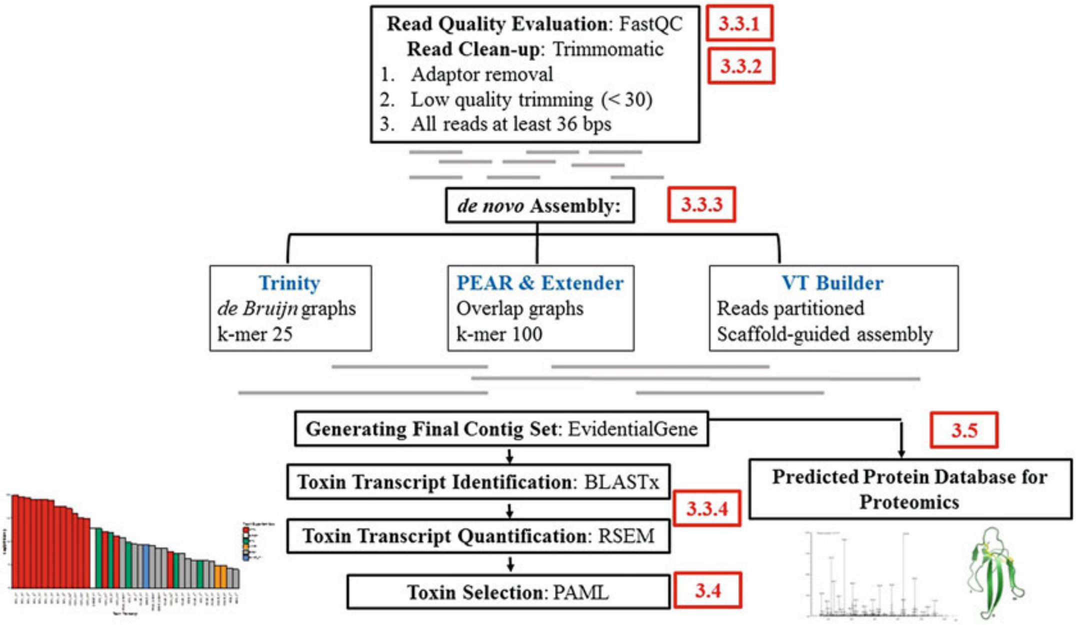
Read Quality Evaluation: FastQC
Read Clean-up: Trimmomatic
1. Adaptor removal
2. Low quality trimming (< 30)
3. All reads at least 36 bps
3.3.1
3.3.2
de novo Assembly:
3.3.3
Trinity
de Bruijn graphs
k-mer 25
PEAR & Extender
Overlap graphs
k-mer 100
VT Builder
Reads partitioned
Scaffold-guided assembly
Generating Final Contig Set: EvidentialGene
3.5
Toxin Transcript Identification: BLASTx
3.3.4
Predicted Protein Database for Proteomics
Toxin Transcript Quantification: RSEM
Toxin Selection: PAML
3.4

Fig. 2 Protocol overview for venom gland transcriptomics. Protocol overview shows each step to be performed for venom gland transcriptomic work, including the processing of next-generation sequencing reads, de novo transcriptome assembly, gene expression determination, toxin transcript identification, positive selection analysis, and integration of high-throughput proteomics with transcriptomics. Procedures discussed in the text are indicated by section numbers (red boxes)

Box 1 Abridged Pipeline Example Commands. A few command examples are given; documentation for each program should be referenced for all command arguments and parameters, and only examples are provided. All CPU/thread arguments should be modified based on computing resources:

```
##############################
### FASTQC example command ###
##############################
SYNOPSIS

Usage:
fastqc seqfile1 seqfile2 .. seqfileN
fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
[-c contaminant file] seqfile1 .. seqfileN

fastqc RAWDATA_PAIR_1.fastq.gz RAWDATA_PAIR_2.fastq.gz -o
OUTPUT_DIRECTORY

###################################
### TRIMMOMATIC example command ###
###################################
SYNOPSIS

Usage:
PE [-threads <threads>] [-phred33|-phred64] [-trimlog
<trimLogFile>] [-quiet] [-validatePairs] [-basein <inputBase> |
<inputFile1> <inputFile2>] [-baseout <outputBase> | <outputFile1P>
<outputFile1U> <outputFile2P> <outputFile2U>] <trimmer1>...
or:
SE [-threads <threads>] [-phred33|-phred64] [-trimlog
<trimLogFile>] [-quiet] <inputFile> <outputFile> <trimmer1>...

java -jar trimmomatic-0.35.jar PE -threads 4 -phred33 RAWDATA_PAIR_1.fastq.gz
RAWDATA_PAIR_2.fastq.gz OUTPUT_R1-paired.fastq
OUTPUT_R1-unpaired.fastq OUTPUT_R2-paired.fastq OUTPUT_R2-unpaired.fastq
ILLUMINACLIP:TruSeq3-PE-2.fa:2:40:15 SLIDINGWINDOW:4:15
LEADING:20 TRAILING:20 MINLEN:50 HEADCROP:9

############################
### PEAR example command ###
############################
SYNOPSIS

Usage:
pear <options>
```

(continued)

```
Standard (mandatory):
-f, --forward-fastq <str> Forward paired-end FASTQ file.
-r, --reverse-fastq <str> Reverse paired-end FASTQ file.
-o, --output <str> Output filename.

pear   -f   INPUT_R1-paired.fastq   -r   INPUT_R2-paired.fastq   -o
OUTPUT_NAME

#############################
### FLASH example command ###
#############################
SYNOPSIS

Usage:
flash [OPTIONS] MATES_1.FASTQ MATES_2.FASTQ
flash [OPTIONS] --interleaved-input (MATES.FASTQ | -)
flash [OPTIONS] --tab-delimited-input (MATES.TAB | -)

flash -o OUTPUT_PREFIX -t 5 INPUT_R1-paired.fastq INPUT_R2-paired.
fastq -r 140 -f 350 -s 50 -d OUTPUT_DIRECTORY

###############################
### TRINITY example command ###
###############################
SYNOPSIS

#Usage:
# --seqType <string> :type of reads: ('fa' or 'fq')
#
# --max_memory <string> :suggested max memory to use by #Trinity where
limiting can be enabled. (jellyfish, sorting, etc)
#provided in Gb of RAM, ie. '--max_memory 10G'
#
# If paired reads:
# --left <string> :left reads, one or more file names #(separated by
commas, no spaces)
# --right <string> :right reads, one or more file names #(separated by
commas, no spaces)
#
# Or, if unpaired reads:
# --single <string> :single reads, one or more file names, #comma-
delimited (note, if single file contains pairs, can use #flag: --
run_as_paired )
#
# Or,
```

(continued)

```
# --samples_file <string> tab-delimited text file #indicating
biological replicate relationships.
#ex.
#cond_A cond_A_rep1 A_rep1_left.fq A_rep1_right.fq
#cond_A cond_A_rep2 A_rep2_left.fq A_rep2_right.fq
#cond_B cond_B_rep1 B_rep1_left.fq B_rep1_right.fq
#cond_B cond_B_rep2 B_rep2_left.fq B_rep2_right.fq
#

Trinity --seqType fq --max_memory 50G --left INPUT_R1-paired.fastq.
gz --right INPUT_R2-paired.fastq.gz --CPU 6 --full_cleanup --min_-
contig_length 100 --verbose

###############################
### CD-HIT example command ###
###############################
SYNOPSIS

Usage:
cd-hit-est [Options]

cd-hit-est -i INPUT_SEQUENCE -o OUTPUT_SEQUENCE -c 1 -n 8

###############################
### BLAST+ example command ###
###############################
SYNOPSIS

Usage:
blastx [-h] [-help] [-import_search_strategy filename]
[-export_search_strategy filename] [-task task_name] [-db
database_name]
[-dbsize num_letters] [-gilist filename] [-seqidlist filename]
[-negative_gilist filename] [-entrez_query entrez_query]
[-db_soft_mask filtering_algorithm] [-db_hard_mask
filtering_algorithm]
[-subject subject_input_file] [-subject_loc range] [-query
input_file]
[-out output_file] [-evalue evalue] [-word_size int_value]
[-gapopen open_penalty] [-gapextend extend_penalty]
[-qcov_hsp_perc float_value] [-max_hsps int_value]
[-xdrop_ungap float_value] [-xdrop_gap float_value]
[-xdrop_gap_final float_value] [-searchsp int_value]
[-sum_stats bool_value] [-max_intron_length length] [-seg
SEG_options]
```

(continued)

```
[-soft_masking soft_masking] [-matrix matrix_name]
[-threshold float_value] [-culling_limit int_value]
[-best_hit_overhang      float_value]      [-best_hit_score_edge
float_value]
[-window_size int_value] [-ungapped] [-lcase_masking] [-query_loc
range]
[-strand strand] [-parse_deflines] [-query_gencode int_value]
[-outfmt format] [-show_gis] [-num_descriptions int_value]
[-num_alignments int_value] [-line_length line_length] [-html]
[-max_target_seqs   num_sequences]   [-num_threads   int_value]
[-remote]
[-comp_based_stats compo] [-use_sw_tback] [-version]

blastx -query INPUT_SEQUENCE -db nr -max_target_seqs 3 -num_threads
8 -outfmt '6 std stitle' -out Blastx_nr_outfmt6

################################
### RSEM example command ###
################################
SYNOPSIS

Usage:
rsem-prepare-reference      [options]      reference_fasta_file
(s) reference_name
rsem-calculate-expression      [options]      upstream_read_file
(s) reference_name sample_name
rsem-calculate-expression [options] --paired-end upstream_read_-
file(s) downstream_read_file(s) reference_name sample_name
rsem-calculate-expression [options] --alignments [--paired-end]
input reference_name sample_name

rsem-prepare-reference [options] INPUT_SEQUENCE INPUT_SEQUENCE.
rsem.ref
rsem-calculate-expression --paired-end -p 5 --bowtie2 INPUT_R1-
paired.fastq    INPUT_R2-paired.fastq    INPUT_SEQUENCE.rsem.ref
INPUT_SEQUENCE.rsem.results
```

3.3.2 NGS Data Quality Checks

The first step upon receiving sequencing reads is to conduct initial quality checks (QC). These QC results can be obtained by loading the read fastq files into the Java program FastQC [18]. This widely used quality control tool for high-throughput sequence data provides a modular set of analyses that can give an impression of potential problems during the library construction and the sequencing run. The following parameters need to be evaluated, and reads filtered to match these criteria, to be used reliably in the assembly:

1. Overall read quality should be greater than a quality score of 20.
2. Adapter contamination should be absent.
3. Proper read length should be at least 36 bp.

There are several available open-source tools that can be used to remove low-quality reads and adapter contamination, but Trimmomatic [19] is a commonly used software for this purpose and performs well in that a sliding window is used to evaluate base quality instead of just read quality averaged. Base quality is reported in a Phred-like score, which is the log value of the error probability (probability of incorrect base calling = $10^{-Q/10}$; Q = Phred score). A quality score (Q) of 20 indicates that there is a 1 in 100 chance that the base call is incorrect. Because low-quality bases are observed on read ends, when these are removed, a minimum length is also set to keep reads long enough to be informative for the assembly. The Trimmomatic package also contains common adaptor sequences that can be selected for removal. These quality-controlled and adaptor-removed filtered fastq files should then be checked again by FastQC before they are used as input for transcriptome assembly.

Paired-end reads can also be merged with programs such as PEAR [20] or FLASH [21] and then used as input into assemblers such as Extender, leveraging longer sequence lengths. However, some assemblers do require the paired-end read information for contig construction. Paired-end read merging can be used for assembling small transcripts, as some animal toxins can be quite small, such as those from arthropod venoms.

3.3.3 *De Novo Transcriptome Assembly*

Venom gland transcriptomes are notoriously difficult to assemble because of the abundance of transcript isoforms and the high levels of expression of these isoforms. However, it is important that toxin transcripts are properly assembled because there is exceptional functional diversity in many toxin families, and minor differences in sequence can greatly alter binding and overall activity.

Trinity [17] is currently one of the most popular de novo RNA-seq assemblers, with over 2500 citations. Trinity partitions RNA-seq reads into many independent *de Bruijn* graphs and with parallel computing reconstructs transcripts from these graphs. Three different software modules are used in Trinity contig construction: Inchworm, Chrysalis, and Butterfly. Inchworm assembles reads into unique sequences using a k-mer-base approach, where each read is partitioned into smaller nucleotide strings of k length. Next, Chrysalis clusters related reads and constructs a *de Bruijn* graph for each cluster of related sequences. Finally, Butterfly analyzes the *de Bruijn* graphs and read pairings to report all plausible transcript sequences. Assembly run times are quite quick, usually completed within 24 h (approximately one-half to 1 h per million reads). The Trinity software package contains many useful Perl scripts, such as those for transcript quantification, differential expression, coding region identification, translation (Transdecoder; https://github.com/TransDecoder/TransDecoder/wiki), and annotation (Trinotate pipeline; https://trinotate.github.io/). However, it has also been noted that Trinity does not perform well in distinguishing between highly similar paralogous or homologous transcripts [51], and because this is often the case with toxins Trinity has been reported to miss toxin transcripts during assembly, or to assemble only partial sequences [52]. Trinity has also been reported to struggle with assembling highly expressed transcripts [53], which is also often the case for toxin genes expressed in the venom gland [54]. These limitations are likely due to the smaller k-mer size (a fixed k-mer of 25) used for Trinity assemblies, because small k-mers are better for assembling minimally expressed genes while larger k-mers perform better for abundantly expressed genes [55].

Extender [22], a Java program, was designed to improve upon the issues observed using Trinity, and other *de Bruijn* graph assemblers such as ABySS [56] and Velvet [57], by utilizing a hashtag table and extending contigs based upon long overlaps. Extender also has faster run times, comparable to Trinity, but has smaller RAM requirements. A larger k-mer size can be used for Extender assemblies and because of an overlap versus a *de Bruijn* graph algorithm, there are fewer alternative paths and therefore less assembly errors are introduced. Extender has been used for multiple venom gland assemblies and performs well when assembling highly expressed transcripts within a venom gland [58]. Reads are first merged with PEAR [20] or FLASH [21] and then used as input into Extender. Extender also performs best when a large number of reads are used, >30 million, but it does produce fewer overall contigs in comparison to Trinity, likely excluding complete transcript diversity.

Another assembler, VTBuilder [23], was also designed to address the issues observed with assembling multi-isoform transcriptomes, making it ideal for venom gland transcriptomes. The VTBuilder assembly algorithm is more similar to reference-guided genome assemblies. Reads are partitioned and a guide sequence is generated from these reads. Reads are then mapped as scaffold-like alignments and reconstructed as contigs representing the transcript isoform diversity present. Unfortunately, the current VTBuilder version only allows up to 5 M reads to be used for assemblies and only works effectively with read lengths equal to or greater than 250 bp. With shorter reads, it has been noted as having performance equal to if not lower in comparison to Trinity when assembling snake venom gland transcriptomes and an RNA spike-in (RNA transcripts of known sequence and quantity used as a control) [13].

Overall, given that each assembler has its own advantages and disadvantages, using multiple assemblers might be the best approach to achieve total and accurate transcript diversity. This is quickly becoming the preferred method of transcriptome assembly, considering that a transcriptome is a heterogeneous mixture of transcripts of different sizes, GC content, complexity regions, expression levels, etc., and one assembler algorithm is likely not best for every transcript. There is also an advantage to generating multiple assemblies with different parameters, such as k-mer values, because the optimal k-mer value for an assembly will depend on the read length, sequencing depth, and read error rate [55], especially in cases where transcript abundances differ tremendously, as mentioned above. A disadvantage to using many different k-mer values is that this has been found to increase the number of fusion/chimeric transcripts when compared to single k-mer methods [59]. Therefore, multiple assemblers and multiple parameters should be explored in addition to quality control checks.

There are some pipelines that include such programs as CAP3 [60] that have been used to merge assembled contigs from multiple assemblers into a final transcriptome set. This DNA sequence assembly program constructs multiple sequence alignments between contigs and then generates a consensus sequence. It can end up merging contigs from separate isoforms, emphasizing again the importance of proper assemble quality control checks. Programs such as TransRate (http://hibberdlab.com/transrate/) can assess the quality of a transcriptome assembly [61]. In order to evaluate assembly performance, several metrics such as N50, average contig length, total assembled nucleotides, maximum contig length, total number of contigs, and number of singletons have largely been taken into consideration [62]. However, which metrics actually reveal assembly quality is unclear, and standard quality metrics commonly used are repurposed from genome assembly.

Further, because of the redundancy of using multiple assemblers, both redundancy removal and selection of the truest set of transcripts will be required. There are several redundancy removal software available, such as the CD-HIT suite software [26] or Exonerate [25], and script pipelines like those provided by EvidentialGene [24] identify high-quality transcripts. The EvidentialGene script pipeline has been shown to perform optimally when dealing with multiple transcriptome assemblies that include duplicated gene copies, and this is a feature of venom gland tissue transcriptomes. Moreover, the EvidentialGene pipeline has been found to be ideal for working with multiple transcript isoforms because transcripts are pooled into one super-set of sequences and then the "best" set of transcripts from this set is selected based on the coding sequence and protein length, emphasizing transcript coding potential.

3.3.4 Toxin Gene Identification and Expression Quantification

The best way to evaluate the quality of the final overall assembly is by the identification of full-length transcripts for toxins known to be present within the venom. In this sense, as mentioned above, the Trinity software package provides a Perl script based on sequence homology that could be used in order to decipher which toxin-identified transcripts expand throughout the entire length of protein sequence. Hence, evaluating the quality of coding sequences, such as if a full-length transcript starts with a methionine and ends with a stop codon, is better than relying on a value like “N50,” which is not very relevant to transcriptome assemblies, because a higher N50 value and the presence of many long contigs can be the result of misassemblies. However, it should also be noted that excluding partially assembled transcripts can lead to underestimation of venom complexity, as partial transcripts can contain valid variants.

BLAST+ (Basic Local Alignment Search Tool) [27], which is run from the command line of a computer (accessed through the terminal for Unix-like operating systems), is commonly used for toxin annotation and is based upon database searches. The databases used include the nonredundant protein database available on NCBI (National Center for Biotechnology Information) or the UniProt database. Custom databases, such as a collection of venom protein sequences, can be created and have been found to be equally successful at the identification of toxin sequences, as long as there exists homology to known toxins. There are also a few specific toxin databases that have been assembled (reviewed in [16]). It should be noted that it is possible to find toxin identities using BLASTn that might be missed using BLASTx or BLASTp. This was observed in the case of the *Boiga irregularis* venom gland transcriptome, where Trinity assembled many partial transcripts that showed untranslated region transcript bias and were unable to be identified with BLASTx, but were identified as toxin transcripts with BLASTn [14].

In this sense, given the fact that mobile elements such as saurian SINEs and LINEs have been largely characterized in all major lineages of squamate reptiles, it is best to mask repeat nucleotide sequences with Repeat Masker (addressing http://www.repeatmasker.org). The program makes use of Repbase (http://www.girinst.org/repbase/), a comprehensive database of repetitive element consensus sequences, reducing running times of the BLAST annotation process.

BLAST+ can have very long run times, and with a full transcriptome (20,000 plus contigs) and using a single workstation it can easily run for a month (if not longer) to generate results. A way to speed this up, besides assigning more processing cores to allow for parallel computing, is to split up files and run them separately, and this is recommended. The program Diamond [28] has a much faster algorithm, faster than the stand-alone BLAST+ by about 20,000 times, and is highly recommended for BLASTx or BLASTp searches.

Regarding the annotation process, within databases submissions are sometimes given the identification of "hypothetical protein," "transcribed mRNA," or even a mis-annotated description; some may not have complete identities when they are submitted, and some might even be partial sequences. Therefore, it is best to report at least the top three BLAST hits in case the top hit given has one of these non-informative labels or is incomplete. A filtering round using a list of keywords (including the acronyms of all known toxin protein families described so far) to distinguish putative snake venom toxins from non-toxin (ribosomal, mitochondrial, nuclear, etc.) proteins should be carried out over the BLAST hit results. The main issue with using previous toxin datasets on an identity search is that only toxin sequences similar to known toxins are identified. Other programs, such as HMMER [63] with the Pfam database [64] or InterPro [65], are sequence analysis programs that use hidden Markov models to identify domains for unknown proteins, and these can be useful to find unknown or novel toxins.

Venom components are secreted cell products and therefore a signal peptide sequence should be present. This is a common criterion used to identify potential toxins and is accomplished by evaluating translated transcripts for signal peptides with SignalP [29]. SignalP can be downloaded and run from a command line for large FASTA files with many sequences. Protein sequences can also be evaluated for transmembrane domains, which are suggestive of non-secreted cell products, and this is done through the use of the program tmHMM [30] that employs hidden Markov models to identify membrane-bound protein regions. It is likely that if a protein has membrane-bound regions, it is *not* a venom component; however, there are no unequivocal certainties, because these proteins could be posttranslationally processed or in the case of a signal peptide there are other mechanisms of cellular export observed as well [66]. To identify a venom protein transcript confidently, venom gland transcriptomics must be combined with venom proteomics (though posttranscriptional regulation may result in no translated product).

Transcript abundances are usually determined based on reads mapping to the de novo-assembled transcriptome and provide within-sample normalization for feature-length and library-size effects. They are reported as RPKM or FPKM (*r*eads/*f*ragments *p*er *k*ilobase of exon model per *m*illion mapped reads) [67] and TPM (transcripts per million), which is currently the most accepted quantification method. In order to estimate transcript abundances from full-length transcripts, several software packages have been developed. One of the most commonly used software packages for this is RSEM (RNA-Seq by expectation-maximization) [31]. This software package uses Bowtie/Bowtie2 [32] as the read aligner, utilizing a Burrows-Wheeler index to keep its memory requirements small. Because multiple transcript isoforms are present for many toxin genes, multi-mapping reads are frequently observed. One should note that for mapping programs like Bowtie2 (the read alignment program used for RSEM quantification), the search for alignments for a given read is randomized. This means that if Bowtie2 encounters a set of equally parsimonious alignments during mapping, one of these alignments is randomly picked. This allows for quick transcript quantification (RSEM run times are usual less than 48 h on a single workstation, depending on read numbers), but any transcript isoform quantification should be seen as a measure of relative abundance only.

3.4 Toxin-Positive Selection

Two primary modes of toxin evolution have been proposed: purifying and positive selection [68, 69]. It has been suggested that positive selection is the dominant driver of snake venom evolution [70], especially for highly expressed venom protein transcripts [13]. Additionally, it has been observed that abundant venom protein superfamilies experience weaker selective constraints because of multiple gene copies, allowing for the accumulation of deleterious mutations, and therefore also neutral evolution [71]. Even though there are multiple models that can be used to examine selection pressures, it must be noted that for large venom protein families that exhibit structural and functional diversity, toxin evolution can be complex.

The most common method of selection evaluation, and one of the easiest to perform, is analyzing toxin transcripts for positive selection. This method examines single-nucleotide polymorphisms (SNPs) within codons, identifying if nonsynonymous or synonymous substitutions are occurring more frequently between homologs. SNPs have been well documented in venom protein transcripts and linked to toxin functional diversification [72]. The ratio of nonsynonymous to synonymous substitutions, ω, can be used to determine if selection is acting on the overall protein and/or specific regions. Values of $\omega < 1$ are suggestive of negative purifying selection, $\omega = 1$ is suggestive of neutral evolution, and values $\omega > 1$ indicate positive selection.

There are several positive selection models that can be used. Branch models allow the ω ratio to vary among branches in a phylogeny to detect positive selection acting on particular lineages [73], and site models allow ω ratios to vary for sites (codons) [74]. There are also models that incorporate both branch and site evaluations, allowing ω to vary for both sites within the protein and across branches on the tree to detect positive selection affecting a few sites along particular lineages (foreground branches) [75, 76]. The most frequently used software for positive selection analysis is PAML (phylogenetic analysis by maximum likelihood) [37], specifically the codeml module. Usually a series of models within PAML are run, and model likelihood values are compared.

Toxin evolution evaluation has been incorporated into venom gland transcriptome assembly publications because of the need of transcript sequences to determine selection occurring for a toxin family. It is of interest to identify which codons experience increased mutation rates since positive selection has indeed been linked to toxin-active sites and molecular surface residues [6, 72]. To set up sequences for a codeml analysis, it is ideal if orthologous toxin sequences are used to compare sequence variation across species and identify which coding regions are more variable. However, identifying orthologous sequences can be particularly challenging with venom toxins. Large multi-isoform toxin families exist because gene duplications result in multiple paralogs, and different paralogs can be evolving under different selection pressures. Correct orthologous sequences between species must be identified from these gene families. Using BLAST identities, especially reciprocal BLAST outputs, potential orthologous toxin genes might be able to be identified.

Once a set of toxin sequences are chosen, PAML will need a nucleotide alignment file and tree file as input for codeml. The alignment will need to be in in PHYLIP format with sequence names identical to those present in the tree file. Each sequence also needs to have the same number of characters. The tree file will need to be in Newick format. Nucleotide models used for tree construction will not matter for PAML, but users should make sure that it is appropriate for their data set. PartitionFinder, Jmodeltest, or MEGA [36] can be used for model selection. Tree construction can be completed using either a maximum likelihood or a Bayesian approach. A suggested open-source pipeline to use is either Aliview [33] or Jalview [35] for the generation of a multiple sequence alignment with either a Clustal or a MUSCLE alignment algorithm, and SeaView [34] to construct a maximum likelihood tree once nucleotide model selection has been performed. The alignment and tree files will need to be designated in the codeml control file, as well as the resulting output file name and all models to be run for comparisons.

Some commonly used PAML models include M0 (one ratio), M1a (neutral), M2a (selection), M3 (discrete), M7 (beta), and M8 (beta&ω). Model M0 estimates a constant ω rate and is compared to model M3, which allows ω to vary across sites. M1a is a model of neutral evolution, where all sites are assumed to be under either negative or neutral selection and is compared to M2a, a model of positive selection. A Bayes empirical Bayes (BEB) approach is useful for identifying specific amino acids under positive selection by calculating the posterior probabilities of a particular amino acid belonging to a given selection class (neutral, conserved, or highly variable). These BEB calculations are performed with the M8 model, run in comparison to the M7 model. Once likelihood values are generated for each model, comparisons can be made with negative twice the difference in log likelihoods between each model compared to a χ^2 distribution. The length of time it takes to run PAML codeml is dependent on sequence number and models used, but it is usually completed within 24 h and can be executed easily on a desktop or laptop computer.

Another software that has been successfully used for toxin selection analysis is HyPhy [77]. HyPhy hypothesis testing using phylogenies is similar to PAML in that it carries out likelihood-based analyses on multiple alignments to find rates and patterns of sequence evolution. HyPhy can be executed from the DataMonkey server [38]. Tests for positive, negative, and episodic selection can all be performed on the DataMonkey server [78].

3.5 High-Throughput Proteomics Integration

Venom gland transcriptomes will then be used as databases for locus-specific matching of proteomic data. Although some top-down proteomics strategies are being developed for proteome profiling, characterization of venoms is usually completed with a bottom-up tandem mass spectrometry (MS/MS) approach, where proteins are first digested with proteases such as trypsin (most commonly used), chymotrypsin, or Glu-C, and then MS/MS produces spectra of fragmented singly charged peptide ions that can be matched to databases for protein identification (peptide mass fingerprinting) or can be used for de novo sequence determination [79]. Collision-induced dissociation (CID) is the most popular MS/MS technique for this type of analysis. This technique creates a series of backbone fragmentations at the peptide bond, resulting in b- and y-fragment ions, and using Mascot, SEQUEST, or other search engines, databases are searched to identify unknown proteins based on their peptide fragment spectra.

However, MS/MS peptide identification relying on available online protein sequence databases can overlook unique protein isoforms and/or be unsuccessful at recognizing novel toxins. Animal venoms can contain many different peptide and protein isoforms, and given that venoms experience high levels of variation even within species, such as ontogenetic [80–83] and regional venom variation [84–86], the use of public databases can be disadvantageous when attempting to characterize unexplored venoms. Venom compositional variation has direct implications for antiserum development and efficacy, and proper identification of toxin diversity is critical. Therefore, the use of an individual or species-specific transcriptome can greatly improve venom proteomic profiling.

There are several programs that allow for the input of a custom protein database, such as a translated venom gland transcriptome, as a FASTA file. Some of the more popular software that have this capability are listed in Subheading 2. Another important consideration when using custom databases, such as a species-specific transcriptome, is that there could be mis-assemblies or missing transcripts within these databases, and therefore searches against publicly available databases also are still advisable. Peptide to translated transcriptome matches assigned by these tools can also have false positives, and therefore a false discovery rate (FDR) metric is often used for confidence assessment [87]. False-positive screening is performed with the inclusion of a decoy database, where incorrect "decoy" sequences are added to the search space. This decoy database can be useful for the design of FDR filtering criteria [88].

An integrated transcriptomics and proteomics (venomics) approach is ideal for not only more accurate and complete identification of venom proteins, but also for better protein quantification [89]. There are several label-free methods of MS/MS quantification of venom components, such as normalized spectral abundance factors (NSAF) [89–91], which normalizes for protein length, or the use of an internal standard of known concentration that is then used to determine unknown concentrations of proteins based upon peptide intensities [92], similarly used for iBAQ [93]. The use of a species-specific or even individual-specific translated transcriptome database can aid in the quantification of venom components, such as providing exact protein sizes for NSAF calculations. Some proteomic programs can also generate their own quantification numbers, such as the emPAI (Exponentially Modified Protein Abundance Index) number [94] from ProteinPilot and Mascot. Additionally, other researchers have relied on the use of chromatogram peak areas for venom component quantification and perform a reversed-phase high-performance liquid chromatography (RP-HPLC) separation before the digestion and identification of proteins [95]. In cases where peaks consist of multiple proteins, gel densitometry is used to determine the abundance of different proteins within a single peak. It is also important to note that although the translated transcriptome is ideal as a species-specific database for MS/MS peptide identifications, there is not always a quantitative correspondence between the transcriptome and proteome.

Transcripts from an assembled transcriptome can be used to obtain the full amino acid sequence of a protein. Using proteomic methodologies (such as N-terminal sequencing and MS/MS de novo sequence determinations from many peptide fragments) to acquire full amino acid sequences of proteins can be labor intensive and expensive. Additionally, with these approaches, complete protein sequences are not guaranteed, as some proteins are N-terminally blocked, do not exhibit sequence for protease digestion, or do not ionize well for MS/MS.

A combined transcriptomic and proteomic approach is often necessary to identify toxins, but the presence of a transcript alone does not mean that it is a translated and secreted venom component [96]. Because the basic definition of a venom is as a secretion, it is therefore of great importance that venom proteomes are characterized, in addition to venom gland transcriptomes, to determine which transcripts belong to secreted venom components. Venom proteins originated from homologs that performed non-venom-related, physiological functions within tissues [97], and misidentification of these physiological proteins and peptides as toxins could distort our view of toxin evolution, especially when they are included in cladistics and selection analyses. This integration of transcriptomics and proteomics improves the accuracy of either approach used alone.

4 Notes

1. Venom protein-specific primer needs to be designed from venom protein transcripts. The best way to accomplish this, if the target sequence is unknown, is by performing a multiple sequence alignment with a collection of similar transcript sequences. Venom protein superfamilies tend to have conserved signal peptide regions and this region is ideal to design primers to target multiple venom protein transcripts within a single superfamily. It is best to incorporate some degenerate nucleotide bases, such as Y (designated for C or T nucleotides) and W (for A and T nucleotides), to improve amplification of all transcripts within a superfamily. Refer to specific instructions that companies have designated for ordering degenerate bases. Usually, 1–4 degenerate bases should be used; more degenerate bases will result in nonspecific binding and amplification. It is also best to run PCR products using agarose gel electrophoresis and excise the band belonging to the estimated transcript size, as this will also help avoid nonspecific transcripts. Modahl and Mackessy (2016) list several primers that have been successfully used to amplify multiple transcript isoforms within a single snake venom protein superfamily; this publication also has details regarding primer design and PCR for 3′RACE.
2. Make sure that X-gal, ampicillin, and IPTG are added after autoclaving agar, and when agar has cooled to approximately 50 °C.
3. RNA is degraded by RNases that occur in the environment, on skin, and in bacteria or mold that may be present on airborne dust particles. RNase contamination is prevented by always wearing gloves, only using plasticware that is labeled “RNase-free” (treat any glassware with RNase inhibitors), using filtered

pipette tips and micropipettes that are designated only for RNA work, and cleaning the work area with RNase inhibitors, such as RNase Away (ThermoFisher Scientific). Also, make sure that all reagents used are molecular grade and are only used for RNA work (this includes water, which must be treated beforehand with DEPC). It is better to be overly cautious when working to avoid environmental RNases than to be neglectful and end up with degraded RNA. Next-generation sequencing technology in particular requires high-quality RNA for library input, and some sequencing centers will even refuse to sequence RNA that falls below a RNA quality threshold. RNA is also unstable, and experiments should be planned to avoid multiple freeze-thaw cycles. RNA should be reverse-transcribed as quickly as possible to avoid degradation. Any long-term storage of RNA should be done at −80 °C and any tissue that will be used later for RNA isolation should also be stored at −80 °C but within a RNAlater stabilizing buffer for best preservation. If tissue samples will be used within 1--2 months and are small, such as venom glands from arthropods, they can be directly collected and stored in TRIzol. This is actually recommended considering that it can be hard to remove small samples from RNAlater. Isolated RNA should be kept on dry ice during any transport.

4. In the case of total RNA isolated from rear-fanged snake venom, a DNase I digestion (amplification grade; Invitrogen) must be performed to remove all traces of DNA before beginning the 3′RACE procedure. Venoms collected from rear-fanged snakes tend to have more DNA contamination that will interfere with later steps.

5. Touch-down PCR is used for this procedure. This means that the first set of repeated cycles has a higher annealing temperature to encourage specific primer binding, and the remaining repeated cycles have a lower annealing temperature to increase overall copy number. This is different than the nested PCR that is described in the manual for the 3′RACE system for rapid amplification of cDNA ends (ThermoFisher Scientific). The PCR method detailed in this chapter and modified from the ThermoFisher Scientific kit protocol has been shown to be successful [42].

6. Ways to troubleshoot PCR to improve amplification: (1) If you had a total reaction volume of 25 μL, sometimes doubling reagent volumes and increasing the total volume to 50 μL can improve amplification. (2) Lower the annealing temperature. However, a lower annealing temperature can result in an increase in nonspecific PCR products. (3) Increase the number of cycles. However, too many (>40) cycles increase the chance of polymerase errors. (4) Increase the time associated with the

68 °C extension, sometimes necessary with longer transcripts. (5) Too much cDNA template can inhibit PCR. Try 1:2 or 1:10 dilutions of cDNA template before it is added to the PCR.

7. Make sure that when bacterial work is completed, precautions are taken for all work to be conducted under sterile conditions. All microcentrifuge tubes and pipette tips should be autoclaved, as well as all prepared LB broth and agar. Any items that come in contact with the bacteria must be discarded as biohazard waste.

8. RNA quality can be determined using a Bioanalyzer. The RIN (RNA Integrity Number) is calculated on a Bioanalyzer by evaluating the ratio between the ribosomal RNA (rRNA) subunits 28S and 18S [98]; this is used to establish the extent of RNase sample degradation. A RIN of at least 7 or 8 is considered acceptable. Spectrophotometry ratios measured on a Nanodrop are also good evaluations of protein or chemical contamination of isolated RNA. The 260/280 absorbance ratio of RNA should be approximately 2.0 to be lacking significant protein contamination, and the 260/230 ratio should also approximate 2.0–2.2 to demonstrate the absence of residual phenol or guanidine that can be carried over from the RNA isolation protocol.

References

1. Mackessy SP (2010) The field of reptile toxinology: snakes, lizards and their venoms. In: Mackessy SP (ed) Handbook of venoms and toxins of reptiles. CRC Press/Taylor & Francis Group, Boca Raton, FL, pp 2–23
2. Gibbs HL, Rossiter W (2008) Rapid evolution by positive selection and gene gain and loss: PLA$_2$ venom genes in closely related *Sistrurus* rattlesnakes with divergent diets. J Mol Evol 66 (2):151–166
3. Dowell NL, Giorgianni MW, Kassner VA, Selegue JE, Sanchez EE, Carroll SB (2016) The deep origin and recent loss of venom toxin genes in rattlesnakes. Curr Biol 26 (18):2434–2445
4. Safavi-Hemami H, Lu A, Li Q, Fedosov AE, Biggs J, Corneli PS, Seger J, Yandell M, Olivera BM (2016) Venom insulins of cone snails diversify rapidly and track prey taxa. Mol Biol Evol 33(11):2924–2934
5. Gendreau KL, Haney RA, Schwager EE, Wierschin T, Stanke M, Richards S, Garb JE (2017) House spider genome uncovers evolutionary shifts in the diversity and expression of black widow venom proteins associated with extreme toxicity. BMC Genomics 18:14
6. Doley R, Mackessy SP, Kini RM (2009) Role of accelerated segment switch in exons to alter targeting (ASSET) in the molecular evolution of snake venom proteins. BMC Evol Biol 9:146
7. Li M, Fry BG, Kini RM (2005) Putting the brakes on snake venom evolution: the unique molecular evolutionary patterns of *Aipysurus eydouxii* (Marbled Sea snake) phospholipase A$_2$ toxins. Mol Biol Evol 22(4):934–941
8. Whittington AC, Mason AJ, Rokyta DR (2018) A single mutation unlocks cascading exaptations in the origin of a potent pitviper neurotoxin. Mol Biol Evol 35(4):887–898
9. Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 94(15):7799–7806
10. Margres MJ, Bigelow AT, Lemmon EM, Lemmon AR, Rokyta DR (2017) Selection to increase expression, not sequence diversity, precedes gene family origin and expansion in rattlesnake venom. Genetics 206 (3):1569–1580
11. Calvete JJ, Sanz L, Angulo Y, Lomonte B, Gutierrez JM (2009) Venoms, venomics, antivenomics. FEBS Lett 583(11):1736–1743
12. Calvete JJ (2014) Next-generation snake venomics: protein-locus resolution through

venom proteome decomplexation. Expert Rev Proteomics 11(3):315–329

13. Aird SD, Aggarwal S, Villar-Briones A, Tin MM, Terada K, Mikheyev AS (2015) Snake venoms are integrated systems, but abundant venom proteins evolve more rapidly. BMC Genomics 16:647
14. Pla D, Petras D, Saviola AJ, Modahl CM, Sanz L, Perez A, Juarez E, Frietze S, Dorrestein PC, Mackessy SP, Calvete JJ (2017) Transcriptomics-guided bottom-up and top-down venomics of neonate and adult specimens of the arboreal rear-fanged Brown Treesnake, *Boiga irregularis*, from Guam. J Proteome 174:71–84
15. Modahl CM, Frietze S, Mackessy SP (2018) Transcriptome-facilitated proteomic characterization of rear-fanged snake venoms reveal abundant metalloproteinases with enhanced activity. J Proteome 187:223–234
16. Kaas Q, Craik DJ (2015) Bioinformatics-aided venomics. Toxins 7(6):2159–2187
17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652
18. Andrews S, FastQC. A quality control tool for high throughput sequence data. http://www.bioinformaticsbabrahamacuk/projects/fastqc/
19. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120
20. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina paired-end reAd mergeR. Bioinformatics 30(5):614–620
21. Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21):2957–2963
22. Rokyta DR, Lemmon AR, Margres MJ, Aronow K (2012) The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genomics 13:312
23. Archer J, Whiteley G, Casewell NR, Harrison RA, Wagstaff SC (2014) VTBuilder: a tool for the assembly of multi isoform transcriptomes. BMC Bioinformatics 15:389
24. Gilbert D. Gene-omes built from mRNA seq not genome DNA [version 1; not peer reviewed]. F1000 Research. 5:1695 (poster) (https://doi.org/10.7490/f1000research.1112594.1)
25. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31
26. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152
27. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421–421
28. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60
29. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786
30. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580
31. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323
32. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9(4):357–359
33. Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30(22):3276–3278
34. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27(2):221–224
35. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191
36. Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33(7):1870–1874
37. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586–1591

38. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26(19):2455–2457

39. Searle BC (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics 10(6):1265–1269

40. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 11(4):M111.010587

41. Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 11(5):996–999

42. Modahl CM, Mackessy SP (2016) Full-length venom protein cDNA sequences from venom-derived mRNA: exploring compositional variation and adaptive multigene evolution. PLoS Negl Trop Dis 10(6):e0004587

43. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotech 26(10):1135–1145

44. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351

45. Parkinson J, Blaxter M (2009) Expressed sequence tags: an overview. Methods Mol Biol 533:1–12

46. Rotenberg D, Bamberger ES, Kochva E (1971) Studies on ribonucleic acid synthesis in the venom glands of *Vipera palaestinae* (Ophidia, Reptilia). J Biochem 121:609–612

47. Hargreaves AD, Swain MT, Hegarty MJ, Logan DW, Mulley JF (2014) Restriction and recruitment – gene duplication and the origin and evolution of snake venom toxins. Genome Biol Evol 8:2088–2095

48. Reyes-Velasco J, Card DC, Andrew AL, Shaney KJ, Adams RH, Schield DR, Casewell NR, Mackessy SP, Castoe TA (2015) Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. Mol Biol Evol 32(1):173–183

49. Junqueira-de-Azevedo IL, Bastos CM, Ho PL, Luna MS, Yamanouye N, Casewell NR (2015) Venom-related transcripts from *Bothrops jararaca* tissues provide novel molecular insights into the production and evolution of snake venom. Mol Biol Evol 32(3):754–766

50. Hargreaves AD, Mulley JF (2015) Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ 3:e1441

51. Nakasugi K, Crowhurst R, Bally J, Waterhouse P (2014) Combining transcriptome assemblies from multiple *de novo* assemblers in the Allo-tetraploid plant *Nicotiana benthamiana*. PLoS One 9(3):e91776

52. Macrander J, Broe M, Daly M (2015) Multi-copy venom genes hidden in *de novo* transcriptome assemblies, a cautionary tale with the snakelocks sea anemone *Anemonia sulcata* (pennant, 1977). Toxicon 108:184–188

53. Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, Altman NS, Pires JC, Leebens-Mack JH, dePamphilis CW (2016) Selecting superior *de novo* transcriptome assemblies: lessons learned by leveraging the best plant genome. PLoS One 11(1): e0146062

54. Aird SD, Watanabe Y, Villar-Briones A, Roy MC, Terada K, Mikheyev AS (2013) Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (*Ovophis okinavensis* and *Protobothrops flavoviridis*). BMC Genomics 14:790

55. Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart P (2012) Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics 13:92

56. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123

57. Zerbino DR (2010) Using the velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11: Unit 11.5

58. McGivern JJ, Wray KP, Margres MJ, Couch ME, Mackessy SP, Rokyta DR (2014) RNA-seq and high-definition mass spectrometry reveal the complex and divergent venoms of two rear-fanged colubrid snakes. BMC Genomics 15:1061

59. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P (2011) Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12 (14):S2

60. Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9 (9):868–877

61. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26(8):1134–1144

62. O'Neil ST, Emrich SJ (2013) Assessing *de novo* transcriptome assembly metrics for consistency and utility. BMC Genomics 14:465

63. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29–W37

64. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279–D285

65. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40:D306–D312

66. Rabouille C (2017) Pathways of unconventional protein secretion. Trends Cell Biol 27 (3):230–240

67. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7):621–628

68. Sunagar K, Moran Y (2015) The rise and fall of an evolutionary innovation: contrasting strategies of venom evolution in ancient and young animals. PLoS Genet 11(10):e1005596

69. Sunagar K, Undheim EA, Scheib H, Gren EC, Cochran C, Person CE, Koludarov I, Kelln W, Hayes WK, King GF, Antunes A, Fry BG (2014) Intraspecific venom variation in the medically significant southern Pacific rattlesnake (*Crotalus oreganus helleri*): biodiscovery, clinical and evolutionary implications. J Proteome 99:68–83

70. Rokyta DR, Wray KP, Lemmon AR, Lemmon EM, Caudle BS (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(5):657–671

71. Aird SD, Arora J, Barua A, Qiu L, Terada K, Mikheyev AS (2017) Population genomic analysis of a pitviper reveals microevolutionary forces underlying venom chemistry. Genome Biol Evol 9(10):2640–2649

72. Sunagar K, Jackson T, Undheim E, Ali S, Antunes A, Fry BG (2013) Three-fingered RAVERs: rapid accumulation of variations in exposed residues of snake venom toxins. Toxins 5(11):2172–2208

73. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148(3):929–936

74. Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. J Mol Evol 51:423–432

75. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22 (4):1107–1118

76. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22(12):2472–2479

77. Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676–679

78. Pond SL, Frost SD (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21 (10):2531–2533

79. Chapeaurouge A, Silva A, Carvalho P, McCleary RJR, Modahl CM, Perales J, Kini RM, Mackessy SP (2018) Proteomic deep mining the venom of the red-headed krait, *Bungarus flaviceps*. Toxins 10(9):E373

80. Mackessy SP (1988) Venom ontogeny in the pacific rattlesnakes *Crotalus viridis helleri* and *C. v. oreganus*. Copeia 1:92–101

81. Saldarriaga MM, Otero R, Núñez V, Toro MF, Díaz A, Gutiérrez JM (2003) Ontogenetic variability of *Bothrops atrox* and *Bothrops asper* snake venoms from Colombia. Toxicon 42 (4):405–411

82. Saviola AJ, Pla D, Sanz L, Castoe TA, Calvete JJ, Mackessy SP (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(R). J Proteome 121:28–43

83. Rokyta DR, Margres MJ, Ward MJ, Sanchez EE (2017) The genetics of venom ontogeny in the eastern diamondback rattlesnake (*Crotalus adamanteus*). PeerJ 5:e3249

84. Massey DJ, Calvete JJ, Sánchez EE, Sanz L, Richards K, Curtis R, Boesen K (2012)

Venom variability and envenoming severity outcomes of the *Crotalus scutulatus scutulatus* (Mojave rattlesnake) from Southern Arizona. J Proteome 75(9):2576–2587

85. Rokyta DR, Wray KP, Margres MJ (2013) The genesis of an exceptionally lethal venom in the timber rattlesnake (*Crotalus horridus*) revealed through comparative venom-gland transcriptomics. BMC Genomics 14:394
86. Margres MJ, Walls R, Suntravat M, Lucena S, Sanchez EE, Rokyta DR (2016) Functional characterizations of venom phenotypes in the eastern diamondback rattlesnake (*Crotalus adamanteus*) and evidence for expression-driven divergence in toxic activities among populations. Toxicon 119:28–38
87. Aggarwal S, Yadav AK (2016) False discovery rate estimation in proteomics. Methods Mol Biol 1362:119–128
88. Elias JE, Gygi SP (2010) Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol 604:55–71
89. Modahl CM, Mrinalini FS, Mackessy SP (2018) Adaptive evolution of distinct prey-specific toxin genes in rear-fanged snake venom. Proc Biol Sci 285(1884):20181003
90. Neilson KA, Keighley T, Pascovici D, Cooke B, Haynes PA (2013) Label-free quantitative shotgun proteomics using normalized spectral abundance factors. Methods Mol Biol 1002:205–222
91. Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, Conaway RC, Conaway JW, Florens L, Washburn MP (2006) Quantitative proteomic analysis of distinct mammalian mediator complexes using normalized spectral abundance factors. Proc Natl Acad Sci 103 (50):18928–18933
92. Rokyta DR, Margres MJ, Calvin K (2015) Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. G3 5(11):2375–2382
93. Fabre B, Lambour T, Bouyssié D, Menneteau T, Monsarrat B, Burlet-Schiltz O, Bousquet-Dubouch M-P (2014) Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. EuPA Open Proteom 4:82–86
94. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4(9):1265–1272
95. Calvete JJ (2013) Snake venomics: from the inventory of toxins to biology. Toxicon 75:44–62
96. Pahari S, Mackessy SP, Kini RM (2007) The venom gland transcriptome of the Desert Massasauga Rattlesnake (*Sistrurus catenatus edwardsii*): towards an understanding of venom composition among advanced snakes (Superfamily Colubroidea). BMC Mol Biol 8:115
97. Casewell NR, Huttley GA, Wuster W (2012) Dynamic evolution of venom proteins in squamate reptiles. Nat Commun 3:1066
98. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3–3