Convergent molecular evolution of toxins in the venom of advanced snakes (Colubroidea)

Xie, B.

Chapter 2. Transcriptome Assembling and Toxin Annotation from Pooled Venom Gland Samples

This chapter is published as part of:

Bing Xie, Daniel Dashevsky, Darin Rokyta, Parviz Ghezellou, Behzad Fathinia, Qiong Shi, Michael K. Richardson and Bryan G. Fry. Dynamic genetic differentiation drives the widespread structural and functional convergent evolution of snake venom proteinaceous toxins. *BMC Biology*, 2022, 20:4. https://doi.org/10.1186/s12915-021-01208-9

Abstract

In the study of the evolution, ecology, function and pharmacology of animal venoms, RNA-seq-based transcriptomics analysis is commonly used. However, the accuracy and completeness of venom profiles determined by transcriptomics are limited by the cross-contamination between samples and the performance of the transcriptome assembly. To solve these problems, we obtained and sequenced venom gland tissues from seven rear-fanged snake species and two front-fanged snake species, and then applied several commonly used *de novo* assembly methods to recover the venom profile followed by a strict criterion to discard chimeric transcripts. Evaluation of the pipelines and the software performance was carried out on the basis of the recovery of non-toxin and confidently-identified toxin transcripts. Serious misrepresentation of the diversity of the toxin families and their relative transcript abundances are demonstrated here. Our work demonstrated that the output from one assembler cannot represent the authentic venom profile. Instead, an effective method should apply different assemblers with various algorithmic strategies and strict quality-control measures. The choice of assembly method, rather than the combination of multiple *k*-mer sizes, is the most important factor in transcript recovery.

Keywords: snake, transcriptomics, toxin, assembly methods

Introduction

The composition of venom varies between species and the secretion of venom is directly regulated by differential gene expression in the venom gland (Rokyta, et al. 2015; Margres, et al. 2016). Those gene expression patterns and levels can be evaluated by RNA-seq transcriptomics. Data mining from transcriptomes also aids in the calculation of the rate of evolution, construction of the phylogenetic tree, genome annotation, and identification of toxin peptides with proteomics (Calvete 2014; Sunagar, et al. 2016). However, the robustness of those results derived from transcriptomics analysis is strongly related to the completeness and quality of the assembled transcripts.

The output of individual sequencing runs has increased considerably as a result of the development of next-generation sequencing (NGS), making samples multiplexing a regular sequencing protocol on sequencing platforms. This brings down the turnaround time and price for each gigabyte of data. For instance, in the Illumina Hiseq X Ten System, the output of a dual flow cell is 1.6-1.8Tb per with a run time of < 3 d (Illumina 2017). With this and other NGS platforms, the multiplexing of samples for sequencing becomes a common practice for genomic/transcriptomics studies (Craig, et al. 2008; Meyer and Kircher 2010; Smith, et al. 2010).

Read misassignment is usually caused by free-floating indexing primers in the final sequencing library for the latest sequencing platforms with patterned flow cells, such as the Hiseq X and NovaSeq platforms (Costello, et al. 2018; Larsson, et al. 2018). If sequencing libraries are not adequately kept and become fragmented, or if the final sequencing libraries have non-ligated indexing primers due to inadequate clean-up and size selection, then free-floating indexes can develop (Illumina 2017). Before the exclusion amplification on the flow cell, these free-floating primers can anneal to the pooled library molecules and be expanded by the DNA polymerase, resulting in a new library molecule with an incorrect index. Index-hopping is a term used by Illumina to describe the process of generating mis-assigned reads. The reported rate of read mis-assignment on Illumina platforms ranges from 0–10% (Kircher, et al. 2012; Nelson, et al. 2014; Wright and Vetsigian 2016; Sinha, et al. 2017; Costello, et al. 2018; Griffiths, et al. 2018; Owens, et al. 2018; Vodák, et al. 2018; Yao, et al. 2018). And this is particularly severe and prevalent when similar types of samples are pooled (e.g., a large number of individuals from the same population with a high degree of sequence similarity).

Even though dozens of bioinformatic methods have been developed for removing index-hopping and assembling (Wright and Vetsigian 2016; Larsson, et al. 2018; Owens, et al. 2018), assembling a good transcriptome can still be challenging. Although a reference genome can be used for transcriptome assembly, the lack of high-quality genomes for many venomous snakes has resulted in most studies relying on *de novo* assembling approaches. Especially when it comes to the recovery of toxin variants from venom glands, *de novo* assembling is the only method to excavate those transcripts. The major algorithm of these *de novo* assemblers is *de Bruijn* graph (Robertson, et al. 2010; Grabherr, et al. 2011; Bankevich, et al. 2012; Schulz, et al. 2012; Peng, et al. 2013; Xie, et al. 2014). A network of *k*-mer nodes is created and connected by edges representing *k*-mer similarity. Edges are traversed to recover the contigs, then transcripts and their various isoforms are identified. The *de Bruijn* graphs method can increase the computational efficiency, but it can also generate chimeric transcripts (Cahais, et al. 2012).

As a result, the assembled transcripts must be carefully annotated and curated to exclude the false positives.

Among previous venom gland transcriptome studies, *de Bruijn* based assembler – Trinity was the most widely employed (Haney, et al. 2014; Li, et al. 2014; Aird, et al. 2015; Luna-Ramirez, et al. 2015; Tan, et al. 2015; Zhang, et al. 2015; de Oliveira Júnior, et al. 2016; Santibáñez-López, et al. 2016; Amorim, et al. 2017; Kazemi-Lomedasht, et al. 2017; Martinson, et al. 2017; Tan, et al. 2017; Cusumano, et al. 2018). There are also some other assemblers successfully employed, such as BinPacker and Extender (Rokyta, et al. 2012; Barghi, et al. 2015; Brinkman, et al. 2015; Dhaygude, et al. 2017). However, empirical studies have shown that one single assembler can hardly recover a full transcriptome profile and assembler performance varies with different taxa and tissues (Holding, et al. 2018). Finally, dealing with the large difference in toxin gene expression as well as varied degrees of paralogy and toxin divergence, *de novo* assembly of venom-gland transcriptomes can be exceedingly difficult (Honaas, et al. 2016; Rana, et al. 2016; Cabau, et al. 2017). The main concern is that missing or biased transcripts will affect downstream analyses such as toxin gene expression levels, toxin diversity, and reconstruction of toxin evolution. In the rising body of transcriptomics investigations of animal venoms, these possible problems highlight the need for a rigorous evaluation of assembler performance for venom gland transcriptomics.

We used many alternative assembly approaches to create *de novo* assemblies of RNA-seq data from pooled FFS and RFS samples to: (1) assess the performance of each assembly using criteria for the number of high quality (non-chimeric) toxin genes assembled across eight species of rear-fanged snakes and two species of front-fanged snakes; (2) evaluate the strengths and weaknesses of each approach for the assembly of venom-gland transcriptomes across eight species of rear-fanged snakes and two species of front-fanged snakes. The considerable heterogeneity in toxin transcripts recovered by different assembly approaches is highlighted, and practical strategies for recovering entire, high-quality venom-gland transcriptomes for toxinology and toxin gene evolution research are provided.

Materials and Methods

Tissue collection and Transcriptome sequencing

Transcriptomes were constructed for the following families and species: Colubridae – *Helicops leopardinus, Heterodon nasicus, Rhabdophis subminiatus*; Homalopsidae – *Homalopsis buccata*; Lamprophiidae - *Malpolon monspessulanus, Psammophis schokari, Psammophis subtaeniatus, Rhamphiophis oxyrhynchus*; and Viperidae – *Pseudocerastes urarachnoides, Vipera transcaucasiana* (Table 1). Venom glands from euthanised captive specimens were obtained under University of Melbourne Animal Ethics Approval UM0706247-2005 and University of Queensland Animal Ethics Approval 2021/AE000075. Venom glands of all snakes were contributed by Dr. Bryan G. Fry from University of Queensland. Total RNA was extracted with Trizol (Invitrogen, Carlsbad, CA, USA) and purified using RNeasy Animal Mini Kit (Qiagen, Valencia, CA, USA).

Table 1: Snake Species studied in this thesis.

Latin names	English names	Super family	Family	Dentition
Helicops leopardinus	Leopard Keelback snake	Colubroidea	Colubridae	RFS
Rhabdophis subminiatus	Red-necked keelback snake	Colubroidea	Colubridae	RFS
Heterodon nasicus	Western hognose snake	Colubroidea	Colubridae	RFS
Malpolon monspessulanus	Montpellier snake	Colubroidea	Lamprophiidae	RFS
Psammophis schokari	Schokari sand racer	Colubroidea	Lamprophiidae	RFS
Psammophis subtaeniatus	Western Yellow-bellied Sand Snake	Colubroidea	Lamprophiidae	RFS
Rhamphiophis oxyrhynchus	Rufous beaked snake	Colubroidea	Lamprophiidae	RFS
Homalopsis buccata	Puff-faced water snake	Colubroidea	Homalopsidae	RFS
Pseudocerastes urarachnoides	Spider-tailed horned viper	Viperoidea	Viperidae	FFS
Vipera transcaucasiana	Armenian sand viper	Viperoidea	Viperidae	FFS

*The taxonomy and morphology of the fang for each species is given based on Taxonomy database in NCBI (https://www.ncbi.nlm.nih.gov/guide/taxonomy/).

In a nutshell, poly-A-containing mRNA molecules were isolated using poly-T oligo-attached magnetic beads, then separated from total RNA using Oligo (dT), and fragmented into minute fragments randomly using divalent cations at extreme temperatures. The first strand cDNAs were generated with reverse transcriptase and random hexamer primers, while the second strand cDNAs were created with the buffer, dNTPs, DNA polymerase I, and RNase H. (Takara Biotechnology, Beijing, China). Following synthesis, these cDNA fragments were ligated with adapters, purified, and PCR enrichment was used to create the final cDNA libraries. Qubit 2.0 and Agilent 2100 were used for preliminary quantification and detecting the insert size of the libraries, respectively, after the synthesis of cDNA libraries. The eligible cDNA libraries were sequenced after passing the screening through Illumina Hiseq X-ten platform at BGI (Shenzhen, China) with 150 bp paired-end reads.

de novo assembly

The majority of the contaminating readings were removed as the initial stage in our approach. This was accomplished by looking for *k*-mers (length set by *-k*, recommended value 57) in our focal read set that were also present at a greater level in another read set from the same lane (x-fold shift set by *-d*, recommended value 1000). Reads with a specified percentage of their sequence represented by such *k*-mers (set by *-p*, recommended value 0.25) were filtered out of the data set. Within the same sequencing lane, raw reads were examined for potential sample cross-leakage due to index mis-assignment. With Jellyfish v. 2.2.6 (Marçais and Kingsford 2011), counts of all 57-mers in raw readings for each sample in each lane were generated, and 57-mers with >1000 count differences between each pair of samples in

a lane were found. In this collection, reads that contained 57-mers for 25% or more of their length were eliminated from the sample with lower counts.

We used Fastp v. 0.20.0 (Chen, et al. 2018) for adapter and quality trimming. Paired forward and reverse reads were overlapped into longer single-end reads with PEAR v. 0.9.11 (Zhang, et al. 2014) as input for assembler Extender. Different assemblers have different strengths and weaknesses. Our strategy is to use several different assemblies and resolve the data later with our quality control methods. We chose four assemblers that have been widely used for *de novo* transcriptome assembling and assembled the identical short-read RNA-seq data with each assembler to compare assembly strategies.

SOAPdenovo-trans (Xie, et al. 2014) and Trinity (Grabherr, et al. 2011) were two of the assemblers that used versions of the *de Bruijn* graph technique to contig building (Haas, et al. 2013). We also used BinPacker (Liu et al. 2016), an assembler that uses coverage information to build splicing graphs and has been found to work well with multi-isoform data. Finally, we employed Extender (Rokyta, et al. 2012), an in-house assembler that picks seed reads at random and extends them outward based on matching overlap with other reads to construct contigs. In its approach to multi-isoform transcript assembly, the VTBuilder (Archer, et al. 2014) assembler uses a similar seed-and-extension technique. We did not use it since its current limit of five million input reads makes it inadequate for the current magnitude of RNA-seq datasets (average 5973 million reads per sample for our data). The format of input reads and, as a result, the overall read counts used for each assembly vary. In this study, BinPacker, SOAPdenovo-trans and Trinity processed with paired pairs, whereas Extender processed with the merged single pairs.

SOAPdenovo-trans is distinct in that it necessitates the usage of a configuration file. It's also unique in that the findings change significantly depending on *k*-mer size, therefore we test it using a variety of *k*-mer sizes. SOAPdenovo-trans v. 1.03 was run at five different *k*-mer sizes: $k = 25$, $k = 31$, $k = 75$, $k = 95$, and $k = 127$, with each run stored as its own assembly. The maximum and minimum read lengths were set to 500 and 200 bp, respectively, with a 250 bp average insert size. We used Trinity 2.5.1 with a minimum contig length of 150 bp and a *k*-mer of $k = 25$ for Trinity assembly. BinPacker v. 1.0 was also ran with $k = 25$ as the *k*-mer size. Finally, we used 2000 randomly selected seeds with a minimum quality of 30 at all base places to run our only seed and assemble strategy, Extender (Rokyta, et al. 2012). As long as the extension-overlap length was not exceeded, seeds were not allowed to share any *k*-mers (100 bp). To save the seed results, we needed at least two extensions in each direction. For a read to be considered for extension, we specified a minimum overlap of 100 bp and a minimum quality score of at least 20 at all base positions. To keep a seed, we permitted 20 duplicates per seed per direction and required that 20% of replicates per seed be expanded. In order to recover as many toxins as possible, we combined transcripts from all assemblers and then we used CD-HIT v. 4.7 (Fu, et al. 2012) to cluster the transcripts and remove the identical transcripts. For CD-HIT, the sequence identity threshold was set to 1 and word length was set to 11. We named this method as ‘Merged’ in this study.

Evaluation of assembly quality and non-toxins

We compared each assembly using a traditional assembly quality metrics to evaluate each assembly. In each assembly, we utilized the software BUSCO v. 4.1.3 to find single-copy, orthologous non-toxin regions (Simão, et al. 2015; Waterhouse, et al. 2018). Using tBLASTn (Gertz, et al. 2006), BUSCO compares assembled contigs to lineage specific subsets of the OrthoDB v. 10 database (Zdobnov, et al. 2017), followed by HMMER (Mistry, et al. 2013) classification of annotated contigs as complete and single copy; complete and duplicated; fragmented; or missing. Ortholog sets in the OrthoDB database comprise genes found as a single copy in the genomes of 90% of the species in the database, providing an evolutionary expectation of existence in an assembled gene set provided the assembly is complete. Although not all loci are expected to be present in a transcriptome study due to lack of expression in the target tissue, a BUSCO analysis will allow quantitative comparison of multiple transcriptome assemblies in terms of the overall number of complete and single-copy orthologous loci recovered from the OrthoDB reference database. We utilized the Tetrapoda ortholog set with 5310 loci for BUSCO analysis of the snake venom gland transcriptome assemblies. The criterion we used to assess non-toxin assembly quality was to see which assembler produced the most complete and single-copy matches to the OrthoDB loci.

Annotation of toxin genes and evaluation of the recovery of toxin genes

The completeness of the final transcript sets and the quality of the toxin contigs were used to assess the quality of the toxin transcript assemblies. We used a series of filtering processes to select contigs that were high-quality toxin transcripts to calculate the amount of high-quality toxin transcripts assembled by each phase of the assembly process: (1) annotation of toxin genes and (2) lack of signs of chimera; formation or fragmentation. We used the TransDecoder tool embedded in Trinity to extract open reading frames (ORFs) from the 'Merged' transcripts. Then we used Blast v. 2.10.1 to do the toxin annotation with our in-house toxin database (Supplementary File 1) to reduce the turnaround time and recover the interesting toxin genes for future evolutionary study.

We used a set of filtering criteria on our annotated toxin genes to come up with a final collection of unique and high-quality toxin sequences. First, we inspected the read coverage of our toxin ORFs. Since there were multi samples sequenced at the same time, the cross contamination can happen between each other. To investigate coverage map of every ORF and see how many reads mapped to it from each of the nine samples, we aligned our ORFs against the original reads from all nine species by BWA v. 0.7.17 (Li and Durbin 2009). We kept only the toxins that: (1) had coverage >0 across all bases in the coding region; (2) had coverage differentials of 100-fold or more across the length of the ORF; and (3) had coverage that varied consistently across its length (if it varied at all) because sharp discontinuities usually indicate chimeric assembly, cross contamination, or some other issue.

The final step was to manually check these remaining sequences for whether or not they really belong to the toxin family they should be assigned to. For this, we manually checked those remaining toxins against sequences on GenBank using the web version of BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi), to check if the annotation results from our in-house toxin database are the same as those annotation in GenBank. We kept those toxin genes for which both annotation results referred to the same toxin genes.

Results

Overall assembly quality and recovery of transcripts

The Illumina HiSeq X Ten sequencing platform generated 56.34~73.44 M raw reads from nine samples. Raw sequencing data were uploaded into the Short Read Archive (SRA) of the NCBI (https://www.ncbi.nlm.nih.gov/sra), retrieving accession numbers of SRR12802473~SRR12802481 (Table 2). After trimming out the low-quality reads, 55.17~71.18 M clean reads were generated from nine samples. Of these clean reads, the Q30 percentage in each library was approximately 90%, which indicated good quality sequencing (Table 3).

The assembly methods were measured by BUSCO (Simão, et al. 2015) to evaluate the completeness of the recovery of all transcripts. Although the results fluctuated significantly, there were clear and persistent trends in relative performance among individual transcriptomes. The 'Merged' approach produced the most complete, single-copy non-toxin transcripts of any method (give the value). This finding was consistent across all snake transcriptomes, with an average of 3430 (range: 3016–3752) complete and single copy non-toxin loci out of 5310 reference loci in snake transcriptomes assembled with 'Merged.'

Table 2: Snake venom gland samples and their raw Illumina data information.

Species	Total Reads (M)	Total Bases (G)	Q30(%)	GC(%)	SRA ID
Helicops leopardinus	59.670768	8.950615	88.30	47.01	SRR12802481
Rhabdophis subminiatus	61.908030	9.286205	87.71	46.84	SRR12802480
Heterodon nasicus	73.443828	11.016574	88.95	44.21	SRR12802479
Malpolon monspessulanus	59.597300	8.939595	88.54	46.28	SRR12802478
Psammophis schokari	71.464838	10.719726	87.30	46.78	SRR12802477
Psammophis subtaeniatus	66.870990	10.030648	88.21	47.21	SRR12802476
**Rhamphiophis oxyrhynchus*	177.324272	26.598641	87.77	44.67	SRR13234020
Homalopsis buccata	61.067040	9.160056	88.27	46.76	SRR12802475
Pseudocerastes urarachnoides	56.348378	8.452257	89.94	48.45	SRR12802474
Vipera transcaucasiana	62.804494	9.420674	87.97	47.67	SRR12802473

*Three samples were sequenced for *R. oxyrhynchus* and all raw data from three samples were merged as one.

Total Reads: number of reads before filtering, saved in M unit.
Total Bases: raw reads number multiply read length, saved in G unit.

Q30: percentages of bases whose correct base recognition rates are greater than 99.9% in total bases
GC content: (G & C base count) / (Total base count)

Table 3: Statistics for cleaned data for each snake species venom gland.

Species	Total Reads (M)	Total Bases (G)	Q30(%)	GC(%)	Reads passed filters(%)
Helicops leopardinus	57.518468	8.528288	90.05	46.93	96.39
Rhabdophis subminiatus	59.717436	8.795656	89.61	46.69	96.46
Heterodon nasicus	71.177292	10.574868	90.44	44.12	96.91
Malpolon monspessulanus	57.588762	8.533593	90.17	46.19	96.63
Psammophis schokari	68.781210	10.097262	89.56	46.66	96.24
Psammophis subtaeniatus	64.644424	9.517410	90.00	47.07	96.67
Rhamphiophis oxyrhynchus	174.167248	25.865585	88.28	44.56	98.22
Homalopsis buccata	58.809498	8.740258	90.02	46.69	96.30
Pseudocerastes urarachnoides	55.165528	8.150225	91.33	48.45	97.90
Vipera transcaucasiana	60.575680	8.951936	89.81	47.67	96.45

Total Reads: number of reads after filtering, saved in M unit.
Total Bases: clean reads number multiply read length, saved in G unit.

Q30: percentages of bases whose correct base recognition rates are greater than 99.9% in total bases
GC content: (G & C base count) / (Total base count).

Some other patterns also emerged. First, within the SOAPdenovo-trans (Xie, et al. 2014), the number of orthologous loci retrieved by BUSCO decreased as the *k*-mer size increased. (Supplementary Figures 1A-I). No transcripts were recovered using *k*-mer 127 (data not shown). Second, Trinity (Haas, et al. 2013) was the second best method for the recovery of non-toxins in all snakes, and it had similar performance to the 'Merged' method. Third, BinPacker (Liu, et al. 2016), SOAPdenovo-trans_K25 and SOAPdenovo-trans_K31 performed comparatively poorly in all snakes, recovering 50% of the loci. Finally, Extender (Rokyta, et al. 2012) and SOAPdenovo-trans_K97 were largely ineffective at recovering non-toxin loci, recovering between zero and 38 complete, single-copy loci (Supplementary Figures 1A-I). Overall, within any individual sample, the rank performance of the assemblers was: 'Merged' > Trinity > BinPacker > SOAPdenovo_trans_K25 > SOAPdenovo_trans_K31 > SOAPdenovo_trans_K75 > SOAPdenovo_trans_K97 > Extender.

Assessment of the recovery of snake toxins

There are large differences between toxin transcripts before and after curation (Supplementary Figure 2A-I). Some toxin families recovered in some samples are all identified as chimeric transcripts and discarded completely. A clear trend in *H. nasicus*, *M. monspessulanus*, *P. schokari*, *P. subtaeniatus*, *H. buccata*, and *P. urarachmoides* indicates that CTL family and Waprin family tend to generate a large number of chimeric transcripts.

There were 1814 transcripts recovered; however, 728 (40.13%) of them were discarded as chimeric, of which, Trinity and BinPacker generated 289 (15.93%) and 171 (9.42%) chimeric (Supplementary Figure 3A-I). The recovery of qualifying toxin transcripts varied greatly across assemblers. In SOAPdenovo-trans assemblies, the toxin gene family and *k*-mer size were obvious determinants of relative performance (Supplementary Figure 4A-I). Among toxin transcripts confidently identified, a clear trend emerged, with the recovery of CTL transcripts being among the highest among the snakes studied. The exceptions were *P. schokari* and *P. subtaeniatus*. In *H. leopardinus, H. nasiucus*, *H. buccata* and *P. urarachnoides*,

CTL was recoverd with highest transcripts. And in *R. subminiatus, M. monspessulanus* and *V. transcaucasiana*, CTL is the second highest. In *P. schokari*, the toxin of highest number is 3ftx and in *P. subtaeniatus*, the highest is Rnase. For CTL in different snakes, different assemblers have different contributions. In *R. subminiatus, M. monspessulanus, P. schokari, P. subtaeniatus,* and *V. transcaucasiana*, BinPacker performed best. While in *H. leopardinus*, *H. nasiucus*, *H. buccata*, and *P. urarachnoides*, the best performers are Extender, SOAPdenovo-trans_K25, Trinity and SOAPdenovo-trans_K97, respectively.

The remarkable difference in the transcripts assembled was further underlined by our accounting of the transcripts retrieved by each assembler in each of the transcriptomes. Trinity recovered the only transcript of Extendin_II from *V. transcaucasiana* amongst all transcriptomes. In general, for all toxin families, Trinity performed much better than the other assemblers in most toxin families except in *P. schokari*, where BinPacker performs better. Trinity did not recover all toxin families but the other assemblers can recover what Trinity did not. The rank performance of each assembler varies between species (Supplementary Figure 4A-I).

The best performing assembler was Trinity. Among 984 good transcripts recovered by all assemblers, 301 (30.5% of the total) were contributed by Trinity. The second-best performer was BinPacker which recovered 254 good transcripts (25.8%). The remaining assemblers–Extender, SOAPdenovo-trans_K75, SOAPdenovo-trans_K31, SOAPdenovo-trans_K25 and SOAPdenovo-trans_K97 recovered 125 (12.7%), 101 (10.3%), 94 (9.6%), 70 (7.1%) and 39 (4.0%) respectively. The best assembler, Trinity, outperformed BinPacker by 47 transcripts.

Discussion

The annotation of 10 assembled venom gland transcriptomes recovered 23 toxin families (Table 4). Here, we have compared the performance of the most widely-used assemblers of transcriptome data and find striking difference in their performance. The performance was assessed by the ability of the assemblers to recover confident toxin families and good transcripts from a series of snake venom gland transcriptomes from nine different species. Our results are significant because, with the rapid development of NGS technology, the choice of method for transcriptome assembly is increasing.

In most previous studies of animal venoms, only one assembly method was utilized to recover the toxin genes (Haney, et al. 2016; Kazemi-Lomedasht, et al. 2017; Tan, et al. 2017; Cusumano, et al. 2018). This prompted concerns that technique biases might result in poor recovery of the entire collection of transcripts and specific genes.

Our results confirm these concerns, because we have demonstrated significant differences in the performance of different assemblers. And their performance is random with regards to the recovery of transcripts belonging to various toxin gene families. Our findings demonstrate that those assemblers which perform well in traditional RNA-seq studies may not necessarily perform well in the recovery of toxin genes (or at all when it comes to specific toxin gene families). Fluctuations in assembler performance can happen when a large number of chimeric transcripts are produced. Our findings indicate

that the pipeline with ‘Merged’ methods for assembling, followed by careful curation, can circumvent these challenges.

Both the Trinity and BinPacker assemblers retrieved a large number of BUSCO orthologous non-toxin loci, with chimeric transcripts accounting for a large percentage of them. Extender's seed-and-extend strategy, on the other hand, was solely successful for toxin gene recovery when used in this study. As a consequence, Extender only found a few non-toxin BUSCO loci. Clearly, assessing the toxin gene quality necessitates a toxin-focused methodology as well as some prior understanding of animal venom biology. Running BWA on those preliminary toxin genes as a manual examination of coverage profiles, as part of our chimera-filtering phase, can be an efficient way to examine individual toxin transcripts for evidence of chimerism.

Previously, the effects of *k*-mer size on the recovery of quality transcripts after de novo assembly were investigated, although not particularly for toxin genes: due to the lack of some low abundance transcripts, greater *k*-mer sizes resulted in fewer transcripts overall (Schulz, et al. 2012). The transcripts that are retrieved, on the other hand, are less likely to be misassembled (Singhal 2013). Larger *k*-mer sizes result in the recovery of less full, single-copy BUSCO loci in our venom gland transcriptomes, confirming earlier findings. However, using SOAPdenovo-trans, certain toxin transcripts and particular toxin families in some samples tended to be retrieved at larger *k*-mer sizes. Because even rare toxins are frequently strongly expressed relative to non-toxin loci, the benefits of utilizing small *k*-mer sizes for assembling rare transcripts may not apply as well to toxin loci in venom gland tissue. Despite this, toxin contigs were occasionally retrieved when small *k*-mer sizes were used, but not when bigger *k*-mer sizes were used. As a result, it's still worth thinking about whether various *k*-mer settings should be merged in a final assembly (e.g., mixing SOAPdenovo-trans results with varied *k*-mer sizes). Previously, Schulz et al. (2012) used a mix of various *k*-mer sizes during assembly and found that it was successful in recovering a pretty comprehensive collection of toxin loci. However, our findings coincide with those of Rana et al. (2016) in that the choice of assembly technique, rather than the combination of multiple *k*-mer sizes, is the most important factor in transcript recovery.

Table 4: Toxin types recovered per species.

	Homalopsidae	Colubridae			Lamprophiidae				Viperidae	
	Homalopsis buccata	*Heterodon nasuta*	*Helicops leopardinus*	*Rhabdophis subminiatus*	*Psammophis sochureki*	*Psammophis sudanensis*	*Malpolon monspessulanus*	*Rhamphiophis oxyrhynchus*	*Pseudocerastes urarachnoides*	*Vipera transcaucasiana*
3ftx		X	X	X	X	X	X		X	X
AChE		X	X	X	X					
C3/CVF		X		X			X	X		
CNP				X						X
CRISP		X	X	X	X	X	X	X		
Cystatin	X	X	X	X		X	X		X	X
Extendin_I										X
Factor_X								X		
HYAL			X					X	X	
Kallikrein									X	X
Kunitz		X	X	X			X	X		X
LAAO									X	X
CTL	X	X	X	X		X	X	X	X	X
Lipocalin		X	X	X						X
NGF							X	X		X
PDE	X							X	X	
PLA2_II_E						X	X	X	X	X
PLB		X					X	X		
Rnase	X	X			X	X			X	X
SVMP				X			X	X	X	X
Veficolin	X	X					X	X		
Vespryn					X					
Waprin								X		

X indicates the recovery of the toxin family.

Our evaluation of assembly completeness for each sample showed that performance of the assemblers was random, which was quite opposite to the finding of Holding, Margres et al. (2018) that some assemblers performed well in the recovery of specific toxin families. In our study, the assembly method in Trinity produced the most chimeras, which was likely confused by many equally plausible routes. This is due to Trinity's poor performance in clustering many isoforms into a single transcript (Macrander, et al. 2015). When examining all nine species samples investigated here, the obvious message of these data is that no one assembler retrieved all toxin loci present, and no assembler demonstrated any bias toward certain toxin families. To get a comprehensive collection of toxin transcripts, reliable venom gland transcriptome research should integrate the quality-filtered output of various assembling techniques. The recovery of a high-quality transcriptome assembly by the clustering and merging of assemblies recovered using different approaches has been proven to work in other systems (Nakasugi, et al. 2014), and our findings imply that this strategy might work here as well. Clustering transcripts based on sequence identity and/or inferred homology may be done in a number of ways (Fu, et al. 2012). Although creating a bioinformatic pipeline for venom gland transcriptomes may appear to be beneficial, the substantial diversity in transcript quality and recovery we found across species, as well as the apparent inaccuracy of many quality measures, suggest that such an endeavor is premature.

References

Aird SD, Aggarwal S, Villar-Briones A, Tin MM-Y, Terada K, Mikheyev AS. 2015. Snake venoms are integrated systems, but abundant venom proteins evolve more rapidly. BMC Genomics 16:1-20.

Amorim FG, Morandi-Filho R, Fujimura PT, Ueira-Vieira C, Sampaio SV. 2017. New findings from the first transcriptome of the *Bothrops moojeni* snake venom gland. Toxicon 140:105-117.

Archer J, Whiteley G, Casewell NR, Harrison RA, Wagstaff SC. 2014. VTBuilder: a tool for the assembly of multi isoform transcriptomes. BMC Bioinformatics 15:1-11.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:455-477.

Barghi N, Concepcion GP, Olivera BM, Lluisma AO. 2015. High conopeptide diversity in *Conus tribblei* revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. Marine Biotechnology 17:81-98.

Brinkman DL, Jia X, Potriquet J, Kumar D, Dash D, Kvaskoff D, Mulvenna J. 2015. Transcriptome and venom proteome of the box jellyfish *Chironex fleckeri*. BMC Genomics 16:1-15.

Cabau C, Escudié F, Djari A, Guiguen Y, Bobe J, Klopp C. 2017. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ 5:e2988.

Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Molecular Ecology Resources 12:834-845.

Calvete JJ. 2014. Next-generation snake venomics: protein-locus resolution through venom proteome decomplexation. Expert Review of Proteomics 11:315-329.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884-i890.

Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T, Mason T. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19:1-10.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5:887-893.

Cusumano A, Duvic B, Jouan V, Ravallec M, Legeai F, Peri E, Colazza S, Volkoff A-N. 2018. First extensive characterization of the venom gland from an egg parasitoid: structure, transcriptome and functional role. Journal of Insect Physiology 107:68-80.

de Oliveira Júnior NG, da Rocha Fernandes G, Cardoso MH, Costa FF, de Souza Cândido E, Neto DG, Mortari MR, Schwartz EF, Franco OL, De Alencar SA. 2016. Venom gland transcriptome analyses of two freshwater stingrays (Myliobatiformes: *Potamotrygonidae*) from Brazil. Scientific Reports 6:1-14.

Dhaygude K, Trontti K, Paviala J, Morandin C, Wheat C, Sundström L, Helanterä H. 2017. Transcriptome sequencing reveals high isoform diversity in the ant *Formica exsecta*. PeerJ 5:e3998.

Fry BG, Scheib H, van der Weerd L, Young B, McNaughtan J, Ramjan SR, Vidal N, Poelmann RE, Norman JA. 2008. Evolution of an arsenal: structural and functional diversification of the venom system in the advanced snakes (Caenophidia). Molecular & Cellular Proteomics 7:215-246.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-3152.

Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biology 4:1-14.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644-652.

Griffiths JA, Richard AC, Bach K, Lun AT, Marioni JC. 2018. Detection and removal of barcode swapping in single-cell RNA-seq data. Nature Communications 9:1-6.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8:1494-1512.

Haney RA, Ayoub NA, Clarke TH, Hayashi CY, Garb JE. 2014. Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. BMC Genomics 15:366.

Haney RA, Clarke TH, Gadgil R, Fitzpatrick R, Hayashi CY, Ayoub NA, Garb JE. 2016. Effects of gene duplication, positive selection, and shifts in gene expression on the evolution of the venom gland transcriptome in widow spiders. Genome Biology Evolution 8:228-242.

Holding ML, Margres MJ, Mason AJ, Parkinson CL, Rokyta DR. 2018. Evaluating the performance of de novo assembly methods for venom-gland transcriptomics. Toxins 10:249.

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:e0146062.

Illumina I. 2017. Effects of index misassignment on multiplexing and downstream analysis. URL: www. illumina. com.

Kazemi-Lomedasht F, Khalaj V, Bagheri KP, Behdani M, Shahbazzadeh D. 2017. The first report on transcriptome analysis of the venom gland of Iranian scorpion, *Hemiscorpius lepturus*. Toxicon 125:123-130.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research 40:e3-e3.

Larsson AJ, Stanley G, Sinha R, Weissman IL, Sandberg R. 2018. Computational correction of index switching in multiplexed sequencing libraries. Nature Methods 15:305-307.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754-1760.

Li R, Yu H, Xue W, Yue Y, Liu S, Xing R, Li P. 2014. Jellyfish venomics and venom gland transcriptomics analysis of *Stomolophus meleagris* to reveal the toxins associated with sting. Journal of Proteomics 106:17-29.

Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, Chen P, Huang X. 2016. BinPacker: packing-based de novo transcriptome assembly from RNA-seq data. PLoS Computational Biology 12:e1004772.

Luna-Ramirez K, Quintero-Hernandez V, Juarez-Gonzalez VR, Possani LD. 2015. Whole transcriptome of the venom gland from *Urodacus yaschenkoi* scorpion. PloS One 10:e0127883.

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.

Margres MJ, Wray KP, Seavy M, McGivern JJ, Herrera ND, Rokyta DR. 2016. Expression differentiation is constrained to low-expression proteins over ecological timescales. Genetics 202:273-283.

Martinson EO, Kelkar YD, Chang C-H, Werren JH. 2017. The evolution of venom by co-option of single-copy genes. Current Biology 27:2007-2013. e2008.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27:764-770.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010:pdb. prot5448.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Research 41:e121-e121.

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:e91776.

Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PloS One 9:e94249.

Owens GL, Todesco M, Drummond EB, Yeaman S, Rieseberg LH. 2018. A novel post hoc method for detecting index switching finds no evidence for increased switching on the Illumina HiSeq X. Molecular Ecology Resources 18:169-175.

Peng Y, Leung HC, Yiu S-M, Lv M-J, Zhu X-G, Chin FY. 2013. IDBA-tran: a more robust *de novo de Bruijn* graph assembler for transcriptomes with uneven expression levels. Bioinformatics 29:i326-i334.

Rana SB, Zadlock IV FJ, Zhang Z, Murphy WR, Bentivegna CS. 2016. Comparison of *de novo* transcriptome assemblers and *k*-mer strategies using the killifish, *Fundulus heteroclitus*. PLoS One 11:e0153104.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ. 2010. *De novo* assembly and analysis of RNA-seq data. Nature Methods 7:909-912.

Rokyta DR, Lemmon AR, Margres MJ, Aronow K. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genomics 13:312.

Rokyta DR, Margres MJ, Calvin K. 2015. Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. G3: Genes, Genomes, Genetics 5:2375-2382.

Santibáñez-López CE, Cid-Uribe JI, Batista CV, Ortiz E, Possani LD. 2016. Venom gland transcriptomic and proteomic analyses of the enigmatic scorpion *Superstitionia donensis* (Scorpiones: *Superstitioniidae*), with insights on the evolution of its venom components. Toxins 8:367.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086-1092.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.

Singhal S. 2013. *De novo* transcriptomic analyses for non-model organisms: An evaluation of methods across a multi-species data set. Molecular Ecology Resources 13:403-416.

Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM. 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. BioRxiv:125724.

Smith AM, Heisler LE, St. Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Research 38:e142-e142.

Sunagar K, Morgenstern D, Reitzel AM, Moran Y. 2016. Ecological venomics: How genomics, transcriptomics and proteomics can shed new light on the ecology and evolution of venom. Journal of Proteomics 135:62-72.

Tan CH, Tan KY, Fung SY, Tan NH. 2015. Venom-gland transcriptome and venom proteome of the Malaysian king cobra (*Ophiophagus hannah*). BMC Genomics 16:687.

Tan KY, Tan CH, Chanhome L, Tan NH. 2017. Comparative venom gland transcriptomics of *Naja kaouthia* (monocled cobra) from Malaysia and Thailand: elucidating geographical venom variation and insights into sequence novelty. PeerJ 5:e3142.

Vodák D, Lorenz S, Nakken S, Aasheim LB, Holte H, Bai B, Myklebost O, Meza-Zepeda LA, Hovig E. 2018. Sample-index misassignment impacts tumour exome sequencing. Scientific Reports 8:1-6.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution 35:543-548.

Wright ES, Vetsigian KH. 2016. Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics 17:1-7.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660-1666.

Yao Y, Zia A, Wyrożemski Ł, Lindeman I, Sandve GK, Qiao S-W. 2018. Exploiting antigen receptor information to quantify index switching in single-cell transcriptome sequencing experiments. PloS One 13:e0208484.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Research 45:D744-D749.

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

Zhang Z, Zhang X, Hu T, Zhou W, Cui Q, Tian J, Zheng Y, Fan Q. 2015. Discovery of toxin-encoding genes from the false viper *Macropisthodon rudis*, a rear-fanged snake, by transcriptome analysis of venom gland. Toxicon 106:72-78.