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

Xie, B.

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

Bing Xie

Xie, Bing
Convergent molecular evolution of toxins in the venom of advanced snakes (Colubroidea)
PhD thesis, Leiden University, the Netherlands
The research described in thesis was funded by the China Council Scholarship (No. 201708440368).
Cover: It contains three elements: the snake fang and venom, the phylogenetic tree and the three-dimension of the snake peptide, which correspond to the content of this thesis. Original illustration with permission of Dr. Bryan G. Fry, University of Queensland.
An electronic version of this thesis can be downloaded from:
openaccess.leidenuniv.nl
Lay-out: Bing Xie
Print: PRINTSUPPORT4U | www.printsupport4u.nl

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

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 1 maart 2022
klokke 10:00 uur

door

Bing Xie
geboren te Zaozhuang, China
in 1989

Promotor:

Prof. dr. Michael K. Richardson

Co-promotor:

Dr. Bryan G. Fry (University of Queensland)

Promotiecommissie:

Prof. dr. Gilles P. van Wezel

Dr. Hans W. Slabbekoorn

Dr. Karen. de Morais-Zani (Instituto Butantan)

Prof. dr. Vera van Noort

Prof. dr. Nicholas R. Casewell (Liverpool School of Tropical Medicine)

Table of Contents

Chapter 1. Introduction and outline of this thesis

Introduction

Early snakes likely possessed mixed serous-mucous oral glands inherited from the last common ancestor of Toxicoferan reptiles (Fry, et al. 2006). Extant snakes have evolved a number of different gland morphologies from this ancestral state (Fry, et al. 2008; Fry, et al. 2013). Some basal snake genera such as *Cylindrophis* and *Eryx* retain some serous gland tissue which likely produces appreciable quantities of protein (Phisalix and Caius 1918; Fry, et al. 2013). In contrast, in the derived basal snake lineages which have secondarily evolved powerful constriction as a novel form of prey capture, the glands have evolved towards primarily secreting mucous in order to lubricate the large furred or feathered prey, in order to facilitate their ingestion (Fry, et al. 2013). However, in the constricting snakes trace levels of proteins are still secreted as an evolutionary relic. These proteins are evolved from ancestral toxins and can be detected by SDS-PAGE gel electrophoresis or via PCR amplification of the encoding genes, and remain sufficiently similar to other snake toxins to produce false positives in antibody-based snake venom detection kits (Jelinek, et al. 2004; Fry, et al. 2013).

The explosive radiation of the advanced snakes (superfamily Colubroidea (Hsiang, et al. 2015) was associated with the partitioning of the mixed glands into two discrete glands, one devoted to venom production, the other for mucous (Fry, et al. 2008; Jackson, et al. 2017). The venom gland has subsequently evolved into an extraordinary diversity of morphological forms (Fry, et al. 2008; Fry, Sunagar, et al. 2015; Jackson, et al. 2017). The venom systems of the lamprophiid lineage (including the genera *Atractaspis* and *Homoroselaps*), elapids, and viperids are homoplastic in that they have convergently evolved hollow fangs linked via tube-like ducts to the venom glands which are enclosed by powerful compressor muscles to increase the speed and efficiency of venom delivery. Similar compressor muscles have also evolved in at least three other lineages (*Brachyophis revoili, Dispholidus typus,* and *Gonionotophis capensis*) without the additional refinement of syringe-like venom delivering dentition (Fry, et al. 2008; Fry, Sunagar, et al. 2015).

Numerous morphological and developmental studies have ascertained that the venom producing glands of all advanced snakes are homologous structures that develop from the primordium at the posterior end of the dental lamina (Kochva 1963a; Kochva 1963b; Kochva and Gans 1965; Fry, et al. 2008; Vonk, et al. 2008; Jackson, et al. 2017). Despite this demonstrated homology, the gland of rear-fanged snakes is often distinguished in the literature from that of front-fanged snakes through the use of the term 'Duvernoy's gland'. The use of this term perpetuates a historical mistake that was made in the original designation of the gland by Taub in 1967 (Taub 1967). At that time, Taub agreed with earlier work that the post-orbital gland of rear-fanged snakes produced venom, citing studies from the early 1900s (Alcock and Rogers 1902; Phisalix and Caius 1918), but considered the glands of viperids and elapids to be non-homologous to each other, and thus assumed the gland of the rear-fanged snakes were also not homologous to elapids or viperids. Crucially, the phrase 'Duvernoy's gland' was not even suggested to highlight this erroneous non-homology, but was simply suggested as a replacement for the name 'parotid gland' which was also occasionally assigned to this structure in rear-fanged snakes. It is now considered

well-established that the venom glands of all colubroid snakes are homologous (Fry, et al. 2008). In fact, elapid and viperid glands are more closely related to the glands of rear-fanged snakes than they are to each other. Thus, the use of the term 'venom gland' for the homologous glands of all advanced snakes is the more appropriate than to refer to a paraphyletic array of morphologies as 'Duvernoy's gland'.

Just as the venom glands of advanced snakes are homologous, so are the venom delivering teeth (fangs). Developmental uncoupling of the posterior sub-region of the tooth forming epithelium facilitated evolution of a wide range of highly modified posterior teeth, with tremendous diversity both in size and morphology (Fry, Sunagar, et al. 2015; Cleuren, et al. 2021). Enlarged rear-fangs—which have evolved on a myriad of occasions—and the three independent evolutions of hollow front fangs, all evolved from the same posterior teeth (Vonk, et al. 2008). These teeth evolved to be farther forward in the mouth of the three front-fanged lineages due to shortening of the maxillary bone and the loss of more anterior teeth (Vonk, et al. 2008).

In a similar fashion to the homology of the morphological aspects of the venom system, modern evidence has accumulated for the homology venom gland toxins expressed across the advanced snakes. The discovery of three-finger toxins (3FTx) for the first time outside of elapid snakes (Fry, Lumsden, et al. 2003a) stimulated a phylogenetic analysis of all known toxin types, revealing the non-monophyly of a myriad of toxin types relative to the organismal relationships (Fry and Wüster 2004). Several toxin families were found to be shared across all advanced snakes including 3FTx, acetylcholinesterase, C-type natriuretic peptides (CNP), kallikrein enzymes, hyaluronidase, kunitz, lectins, and snake venom metalloproteases (SVMP). As the species and their venom secretion and delivery system have diversified, so too have the venom proteins themselves. Accelerated evolution and rapid neofunctionalization are common traits of snake venom gene families (Fry, Wüster, et al. 2003; Sunagar, et al. 2013; Junqueira-de-Azevedo, et al. 2016; Dashevsky, et al. 2018; Dashevsky and Fry 2018; Barua and Mikheyev 2020; Dashevsky, et al. 2021). These genes are frequently duplicated, which can lead to functional and structural diversification (Moura-da-Silva, et al. 1996; Slowinski, et al. 1997; Afifiyan, et al. 1999; Kordiš and Gubenšek 2000), as well as faster rates of sequence evolution (Nakashima, et al. 1993; Kini and Chan 1999). This variety could be due to selection for the ability to kill and digest prey (Daltry, et al. 1996), or it could be the outcome of a predator–prey arms race (e.g., (Poran, et al. 1987; Heatwole and Poran 1995).

Despite the evolutionary novelty of snake venom proteins, a comprehensive reconstruction of the molecular evolutionary history of these major shared toxin types has not been undertaken. This has been in part due to the relatively low number of sequences available from rear-fanged species. Elapid and viperid snake species have been the focus of much more research because of their medical importance (Saviola, et al. 2014; Jackson, et al. 2019). To carry out these broad analyses, we obtained venom gland transcriptomes from eight rear-fanged species spanning the families Colubridae subfamilies Dipsadinae (*Helicops leopardinus, Heterodon nasicus*) and Natricinae (*Rhabdophis subminiatus*), Homalopsidae (*Homalopsis buccata*), and the Lamprophiidae subfamily Psammophiinae (*Malpolon monspessulanus, Psammophis schokari, Psammophis subtaeniatus,* and *Rhamphiophis oxyrhynchus*) as well as two viperid species spanning that family's phylogenetic range (*Pseudocerastes urarachnoides* and *Vipera*

transcaucasiana). With these sequences, we were able to reconstruct the molecular evolutionary history of the shared toxins and map their diversification patterns relative to the organismal relationships and functional changes in the venom. This allows us to evaluate the relative order of evolutionary events such as the diversification of Colubroidea, the partitioning of the venom glands into discrete mucous and protein-secreting units, diversifications in toxin families, and structural and functional novelties in toxin sequences.

Outline of this thesis

In **Chapter 1**, we introduce the background of this thesis. We describe how this thesis involves the *de novo* sequencing and of venom gland tissue from multiple snake species, including rarely-studied rear-fanged snakes. Because of the large amount of new data generated and analysed, we devote each of chapters 2-5 to a different family of toxins for which we found particularly important results.

In **Chapter 2**, we prepare the ground for future chapters by addressing some problematic issues in RNA-seq based transcriptomics, when used to study venom toxin evolution. This approach has been widely used in the study of the evolution, ecology, function, and pharmacology of animal venoms. 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 sequenced eight species of RFS and two FFS, and applied several commonly used *de novo* assembly methods to recover the authentic venom profiles 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 transcripts abundance are demonstrated here. The authentic toxin transcripts are then used in Chapter 3-5 to reveal the full-scale molecular evolutionary history and the patterns of selection.

In **Chapter 3**, we analyse the evolution of Kunitz-type toxins. These toxins, found in reptile venoms have exhibited extensive diversity of structures and functions, from enzyme inhibitors to channel-blocking neurotoxins. However, their detailed evolutionary trends and patterns remain a mystery. We therefore conducted a large-scale phylogenetic and selection analysis. This revealed that the kunitz-type toxins evolved by gene duplication and rapid diversification and showed that: (1) the main ancestral function plasmin inhibitors in Viperidae are under neutral selection while in non-vipers they are under purifying selection; (2) neurotoxic toxins are only found in non-viper clades and different neurotoxic types clustered in separate distinct clades under positive selection.

In **Chapter 4**, we examine C-type lectins, one of the largest toxin families in mammals and reptiles. These toxins are known to have various functions including defence against predator. Since the Viperidae split off from the remaining caenophidian snakes, a novel heterodimeric lectin type evolved by the loss of carbohydrate-binding ability. However, the evolutionary trends and patterns of C-type lectins remain a mystery. We therefore conducted a large-scale phylogenetic and selection analysis. We recovered multiple variants from non-viperid snakes that possessed the diagnostic cysteine of the dimeric lectin

form, but forms with and without the glutamine motif form were present. We also found that the α-subunit and β-subunit were subject to different selection pressures.

In **Chapter 5**, we study SVMP (snake venom metalloproteinase) toxins. These toxins serve as a model system for looking at the evolutionary mechanisms that lead to changes in protein function and structure. Extensive gene duplication and domain loss has occurred. And the ancestral P-III SVMPs have indicated a much more complex structure and functional diversity than P-I and P-II SVMPs. P-IIId SVMP and truncated SVMPs are recently emerged novel traits, but the evolution of P-IIId SVMP and the truncated SVMPs remain a mystery. The aim of this study was to investigate the evolutionary process that resulted in the structural and functional diversification within the P-IIId subfamily and the truncated SVMP propeptide type. We found structural convergences in including the evolution of cysteines for form heteromeric complexes within SVMP, and *de novo* evolution of new toxin families within the propeptide region occurring in the SVMP gene.

In **Chapter 6**, we summarize and discuss the results of Chapters 2-5.

In **Chapter 7**, we provide supplementary materials for this thesis.

References

Afifiyan F, Armugam A, Tan CH, Gopalakrishnakone P, Jeyaseelan K. 1999. Postsynaptic α-neurotoxin gene of the spitting cobra, *Naja naja sputatrix*: structure, organization, and phylogenetic analysis. Genome research, 9(3), 259-266.

Alcock AW & Rogers L. 1902. On the toxic properties of the saliva of certain 'non-poisonous' Colubrines. Proceedings of the Royal Society of London, 70(459-466), 446-454.

Barua A & Mikheyev AS. 2020. Toxin expression in snake venom evolves rapidly with constant shifts in evolutionary rates. Proceedings of the Royal Society B, 287(1926), 20200613.

Cleuren SG, Hocking DP & Evans AR. 2021. Fang evolution in venomous snakes: adaptation of 3D tooth shape to the biomechanical properties of their prey. Evolution. In Press

Daltry JC, Wüster W, Thorpe RS. 1996. Diet and snake venom evolution. Nature, 379(6565), 537-540.

Dashevsky D, Debono J, Rokyta D, Nouwens A, Josh P, Fry BG. 2018. Three-finger toxin diversification in the venoms of cat-eye snakes (Colubridae: *Boiga*). Journal of Molecular Evolution, 86(8), 531-545.

Dashevsky D, Fry BG. 2018. Ancient diversification of three-finger toxins in *Micrurus* coral snakes. Journal of Molecular Evolution, 86(1), 58-67.

Dashevsky D, Rokyta D, Frank N, Nouwens A, Fry BG. 2021. Electric Blue: Molecular Evolution of Three-Finger Toxins in the Long-Glanded Coral Snake Species *Calliophis bivirgatus*. Toxins, 13(2), 124.

Fry BG, Lumsden NG, Wüster W, Wickramaratna JC, Hodgson WC, Kini MR. 2003. Isolation of a neurotoxin (α-colubritoxin) from a nonvenomous colubrid: evidence for early origin of venom in snakes. Journal of Molecular Evolution, 57(4), 446-452.

Fry BG, Scheib H, van der Weerd L, Young B, McNaughtan J, Ramjan SR. 2008. Evolution of an arsenal: structural and functional diversification of the venom system in the advanced snakes (Caenophidia). Molecular & Cellular Proteomics, 7(2), 215-246.

Fry BG, Sunagar K, Casewell NR, Kochva EL, Roelants K, Scheib H. 2015. The origin and evolution of the Toxicofera reptile venom system. In: *Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery*. New York: Oxford University Press. p.1-31

Fry BG, Undheim EA, Ali SA, Jackson TN, Debono J, Scheib H. 2013. Squeezers and leaf-cutters: differential diversification and degeneration of the venom system in toxicoferan reptiles. Molecular & Cellular Proteomics, 12(7), 1881-1899.

Fry BG, Vidal N, Norman JA, Vonk FJ, Scheib H, Ramjan SR. 2006. Early evolution of the venom system in lizards and snakes. Nature, 439(7076), 584-588.

Fry BG, Wüster W. 2004. Assembling an arsenal: origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences. Molecular Biology and Evolution, 21(5), 870-883.

Fry BG, Wüster W, Ryan R, Sheik F, Jackson T, Martelli P. 2003. Analysis of Colubroidea snake venoms by liquid chromatography with mass spectrometry: evolutionary and toxinological implications. Rapid Communications in Mass Spectrometry, 17(18), 2047-2062.

Heatwole H, Poran NS. 1995. Resistances of sympatric and allopatric eels to sea snake venoms. Copeia, 136-147.

Hsiang AY, Field DJ, Webster TH, Behlke AD, Davis MB, Racicot RA. 2015. The origin of snakes: revealing the ecology, behavior, and evolutionary history of early snakes using genomics, phenomics, and the fossil record. BMC evolutionary biology, 15(1), 1-22.

Jackson TN, Jouanne H, Vidal N. 2019. Snake venom in context: neglected clades and concepts. Frontiers in Ecology and Evolution, 7, 332.

Jackson TN, Young B, Underwood G, McCarthy CJ, Kochva E, Vidal N. 2017. Endless forms most beautiful: the evolution of ophidian oral glands, including the venom system, and the use of appropriate terminology for homologous structures. Zoomorphology, 136(1), 107-130.

Jelinek GA, Tweed C, Lynch D, Celenza T, Bush B, Michalopoulos N. 2004. Cross reactivity between venomous, mildly venomous, and non-venomous snake venoms with the Commonwealth Serum Laboratories Venom Detection Kit. Emergency Medicine, 16(5-6), 459-464.

Junqueira-de-Azevedo IL, Campos PF, Ching AT, Mackessy SP. 2016. Colubrid venom composition: an-omics perspective. Toxins, 8(8), 230.

Kini RM, Chan YM. 1999. Accelerated evolution and molecular surface of venom phospholipase A 2 enzymes. Journal of Molecular Evolution, 48(2), 125-132.

Kochva, E. 1963. Development of the venom gland and trigeminal muscles in *Vipera palaestinae*. Cells Tissues Organs, 52(1-2), 49-89.

Kochva, E. 1963. The phylogenetic significance of the venom apparatus in snakes. American Zoologist, 3(4). 487.

Kochva E, Gans C. 1965. The venom gland of *Vipera palaestinae* with comments on the glands of some other viperines. Cells Tissues Organs, 62(3), 365-401.

Kordiš D, Gubenšek F. 2000. Adaptive evolution of animal toxin multigene families. Gene, 261(1), 43-52.

Moura-da-Silva AM, Theakston RD, Cramptonl JM. 1996. Evolution of disintegrin cysteine-rich and mammalian matrix-degrading metalloproteinases: gene duplication and divergence of a common ancestor rather than convergent evolution. Journal of Molecular Evolution, 43(3), 263-269.

Nakashima K, Ogawa T, Oda N, Hattori M, Sakaki Y, Kihara H. 1993. Accelerated evolution of *Trimeresurus flavoviridis* venom gland phospholipase A2 isozymes. Proceedings of the National Academy of Sciences 90(13), 5964-5968.

Phisalix M, Caius R. 1918. L'extension de la fonction venimeuse dans l'ordre entière des ophidiens et son existence chez des familles ou elle n'avait pas été soupçonnée jusqu'içi. Journal de Physiologie et de Pathologie générale, 17, 923-964.

Poran NS, Coss RG, Benjamini E. 1987. Resistance of California ground squirrels (*Spermophilus beecheyi*) to the venom of the northern Pacific rattlesnake (*Crotalus viridis oreganus*): a study of adaptive variation. Toxicon, 25(7), 767-777.

Saviola AJ, Peichoto ME, Mackessy SP. 2014. Rear-fanged snake venoms: an untapped source of novel compounds and potential drug leads. Toxin Reviews, 33(4), 185-201.

Slowinski JB, Knight A, Rooney AP. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. Molecular phylogenetics and evolution, 8(3), 349-362.

Sunagar K, Jackson TN, Undheim EA, 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.

Taub AM. 1967. Comparative histological studies on Duvernoy's gland of colubrid snakes. Bulletin of the AMNH; v. 138, article 1.

Vonk FJ, Admiraal JF, Jackson K, Reshef R, de Bakker MA, Vanderschoot K. 2008. Evolutionary origin and development of snake fangs. Nature, 454(7204), 630-633.

Abbreviations

The following abbreviations for toxins are used in this thesis:

3ftx	Three-finger toxin
AChE	Acetylcholinesterase
BPP	Bradykinin-potentiating peptide
C3/CVF	Complement C3 of cobra venom factor
CNP	C-type natriuretic peptide
CTL	C-type lectin
CRISP	Cysteine-rich secretory protein
ESP	Epididymal secretory protein
HYAL	Hyaluronidase
KTT	Kunitz type toxin
LAAO	L-amino acid oxidase
NGF	Nerve growth factor
PDE	Phosphodiesterase
PLA2_II_E	Phospholipase_A2_type_II_E
PLB	Phospholipase_B
RAP	Renin aspartate protease
RNase	Ribonuclease
SVMP	Snake venom metalloproteinase
SVSP	Snake venom serine protease
VEGF	Vascular endothelial growth factor

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.

Chapter 3. Evolutionary Novelties in the Kunitz-type Toxins

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

Kunitz-type toxins from snake venoms have undergone extraordinary structural and functional molecular adaptation. These toxins may have functions ranging from enzyme inhibitors to channel-blocking neurotoxins. However, their detailed evolutionary history is not known. We therefore conducted a large-scale phylogenetic and selection analysis to illustrate this issue. Phylogenetic analysis and analysis of selection revealed that the kunitz-type toxins evolved by gene duplication and rapid diversification. We show that the main ancestral function of kunitz-type toxins in Viperidae (plasmin inhibitors) is under neutral selection, while in non-vipers it is under purifying selection. We also show that neurotoxic kunitz-type toxins are found only in non-vipers; the various neurotoxic types clustered in distinct clades under positive selection. These results provide a detailed roadmap for future work to elucidate predator-prey evolutionary arms races, as well as documenting rich biodiscovery resources for lead compounds in the drug design and discovery pipeline.

Keywords: Kunitz, protease inhibitor, neurotoxin, evolution

Introduction

Toxins of the Kunitz type are interesting. According to their functions, they may be divided into two groups. Then there are the inhibitors of serine proteases and neurotoxins that block channels, such as dendrotoxins (DTX) and the B chain of bungarotoxin, constituting the second kind (BTX-b). The neurotoxins have a kunitz-like domain, although their protease inhibitory action is minimal. They are an example of dual convergence: not only have they been separately recruited into the venom arsenal of different species, but they are also convergently apotypic in multiple lineages for the same neurotoxic and coagulopathic properties (Beress 1982; Antuch, et al. 1993; Minagawa, et al. 1997; Harvey and Robertson 2004; Bayrhuber, et al. 2005; Honma and Shiomi 2006; Fry, Scheib, van der Weerd, Young, McNaughtan, Ramjan, Vidal, Poelmann and Norman 2008; Koludarov, et al. 2012).

Kunitz-type protease inhibitors attach to the active site of serine proteases predominantly through an exposed binding loop in a canonical configuration. P3, P2, P1, P1, P1, P2, and P3 (nomenclature according to Schechter and Berger) are the six residues that potentially interact with the enzyme in this area. A catalytic trio of residues on the enzyme's pocket (typically Ser, His, and Asp) is responsible for amide bond hydrolysis. In textilinin-1, this hexapeptide is Pro-Cys-Arg-Val-Arg-Phe (PCRVRF) (Flight, et al. 2009), whereas in aprotinin it is Pro-Cys-Lys-Ala-Arg-lle (PCKARI) (Wells and Strickland 1994). The P1 residue, whose side chain protrudes into the specificity pocket of the protease, is Arg in textilinin-1 and Lys in aprotinin.

Despite the large size of the kunitz-type protease inhibitors, they can be divided into three categories based on their target: trypsin inhibitors (positively charged residues Lys/Arg preferred at P1), elastase inhibitors (small hydrophobic residues Ala/Val at P1), and chymotrypsin inhibitors (large hydrophobic residues Phe/Trp/Tyr/Leu/Val at P1) (Laskowski Jr and Kato 1980; Hedstrom 2002; Wang, et al. 2012). Met, Asn and His at the P1 site residues were also reported for chymotrypsin inhibiting (Laskowski Jr and Kato 1980; Scheidig, et al. 1997; Chang, et al. 2001; Chen, et al. 2001; Guo, et al. 2013). The interaction inhibitor: trypsin is essentially independent of the nature of the basic residue at P1 position, with no substantial changes in the association energies with trypsin following the mutation K15R in BPTI, according to analysis of different mutants of BPTI and other trypsin inhibitors (Navaneetham, et al. 2010). Inhibition of kallikrein prefers Arg over Lys (Fieldler 1987; Grzesiak, et al. 2000), but plasmin inhibition is increased by Lys at the P1 site (Van Nostrand, et al. 1995). A hydrophobic amino acid residue (Ala, Gly, or Phe) binds to the P1′ site of venom trypsin/chymotrypsin inhibitors, with Ala being the most prevalent. AvKTI (Araneus ventricosus Kunitz-type serine protease inhibitor), which has the basic amino acid Lys (K) within the P1 site, has dual antitrypsin and antichymotrypsin action (Wan, et al. 2013). A snake venom serine protease inhibitor with Lys (K) at the P1 site and dual inhibitory action against trypsin and chymotrypsin was shown to have a similar effect (Guo, et al. 2013). Furthermore, the P1-P1 residues in AvKTI are Lys (K) and Ala (A), as shown for bovine pancreatic trypsin inhibitor (BPTI), which not only inhibits trypsin but also chymotrypsin, plasmin, and kallikrein. Kunitz peptides that inhibit plasmin are invariably dual-functional, with a P1 site of Arg (R) inhibiting both trypsin and plasmin. P1 and P1 are important residues. P1 fits into the pocket of human plasmin and interacts with six plasmin residues.

In this study, we applied the Bayesian method to reconstruct a large-scale phylogeny so as to reveal the evolution history of the snake venom kunitz-type toxins, particularly the evolutionary traits of the neurotoxic types. Analysis of selection was also utilized to evaluate the selection pressure on different toxin types.

Materials and Methods

Sequence Alignments and Phylogenetic Reconstruction

Protein sequences for all toxin sequences were retrieved from the UniProt database (https://www.uniprot.org) and NCBI database (http://www.ncbi.nlm.nih.gov), then combined with the toxin transcripts from our assembly and annotation (Chapter 2). Partial sequences, sequences with suspect assembling errors were excluded. For the blocks of sequence in between these sites, the sequences were aligned using a mix of manual alignment of the conserved cysteine locations and alignment using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) method (Edgar 2004) implemented in AliView (Larsson 2014). Manual refinement of the alignment was also involved because there are structural differences within different toxin families. The phylogenetic trees for different toxin families were reconstructed with MrBayes 3.2 (Ronquist, et al. 2012) based on the amino acid sequence alignment. The settings for MrBayes can be found in the Supplementary File 2. The output trees from MrBayes were midpoint rooting, then further edited and annotated with iTol (Letunic and Bork 2007).

Tests for Selection

Coding DNA sequences, which are corresponding to the toxin sequences used for phylogenetic analysis, were retrieved from GenBank (Benson, et al. 2012) and our assembly. Using AliView and the MUSCLE method, the sequences were trimmed to only containing codons that translate to the mature protein, then translated, aligned, and reverse translated. Clades were created based on taxonomy and structural differences (functional domains/motifs, for example). The resultant codon alignments were used to create phylogenetic trees for each clade using the same methods outlined in the ‘Phylogenetic Reconstruction' section. All following studies were conducted using these tree topologies.

Calculating the ratio of nonsynonymous nucleotide substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous sites (dS) (ω=dN/dS) for each codon in the alignment might reveal if a gene is undergoing rapid evolution or stays functionally restricted. Codons developing with $\omega > 1$ are thought to have evolved under positive selection (functional diversity), whereas codons evolving with $\omega < 1$ are thought to have evolved under purifying selection. Sites with a value of $\omega = 1$ are believed to evolve in a neutral manner. In order to find the most likely groups on which positive selection has been working, we conducted a series of experiments integrating data from site-based and lineage-specific studies.

Due to their various emphases, we employed many of the selection tests developed in HyPhy v 2.220150316 beta (Pond and Muse 2005) to study the patterns of selection acting on distinct toxin families. The Analyze Codon Data analysis in HyPhy produces overall alignment values, whereas the FUBAR technique assesses the intensity of persistent positive or negative selection on individual amino

acids (Murrell, et al. 2013). The Mixed Effects Model of Evolution (MEME) approach, on the other hand, finds particular locations that have been exposed to positive selection in the past (Murrell, et al. 2012).

Protein Modelling

To map residues evolving under positive selection in three-dimensional (3D) structures, sample sequences from the RCSB PDB database (Rose, et al. 2010) were used to create bespoke models for each clade belonging to various toxin families (Table 1). Alignments of each clade were trimmed to match these PDB structures. To render and colour the 3D structure of the proteins, we utilized the UCSF Chimera program v 1.10.2 with attribute files generated from FUBAR and MEME results. For FUBAR, we used the value from the beta-alpha column which is a measure of the difference between the rates of non-synonymous (beta) and synonymous (alpha) mutations. For MEME, since MEME estimates two rates of positive selection and gives each a probability, we take the weighted average of those two and then subtract alpha to arrive at a similar value to the one we used for FUBAR.

Table 1: Custom models for protein modelling.

Clade	Sequence for 3D modeling	PDB ID
Viperidae	Vipera_ammodytes_ammodytes_P00992	6a5i
DTX	Dendroaspis_polylepis_P00981	1dtk
BTX	Bungaru_ multicinctus_Q1RPT0	1bun
non vipers	Bungarus_multicinctus_Q1RPT0	1bun
Viper type plasmin inhibitor	Vipera_ammodytes_ammodytes_P00992	6a5i
non-viper type plasmin inhibitor	Pseudonaja_textilisQ90WA1	5zj3

Results and Discussion

Diverse kunitz peptides have been characterized as neurotoxins (Harvey and Karlsson 1980; Bohlen, et al. 2011; Baconguis, et al. 2014; Possani, et al. 1992), and our phylogenetic analysis combined with differences in sequence, structure, and function suggest that the evolution of this derived activity has occurred on four independent occasions (Figure 1). The new toxins include monomeric toxins and members of toxin complexes. Dendrotoxins are monomeric toxins from *Dendroaspis* venoms that selectively block $K_V1.1$ voltage-gated potassium channels (Harvey and Karlsson 1980). Kunitz peptides that are subunits of complex neurotoxins may be associated through non-covalent interactions (MitTx and taicatoxin) or covalently linked (β-bungarotoxin). Intriguingly, all such multimers are heteromeric and include PLA_2 toxins. MitTx is a complex of one kunitz subunit and two PLA_2 subunits that activates acid-sensing ion channel ASIC1 to cause intense pain as part of the defensive arsenal of *Micrurus tener*

(Bohlen, et al. 2011; Baconguis, et al. 2014). Taicatoxin was discovered in the venom of *Oxyuranus scutellatus* and is a complex toxin consisting of one 3FTx subunit, one PLA_2 subunit, and 4 kunitz subunits that blocks cardiac voltage-dependent L-type calcium channels (Ca_V) (Possani, et al. 1992). β-bungarotoxins are voltage-gated potassium channels (K_V) blocking heterodimers consisting of a kunitz peptide disulphide-linked to a PLA_2 subunit via a newly evolved cysteine not found in other kunitz peptides, linked to a matching novel cysteine in the PLA_2 subunit that is also not found in other PLA_2 toxins. Our phylogenetic analysis indicates that the characteristic cysteine in β-bungarotoxin kunitz peptides evolved independently on two different occasions (Figure 1). In each case, the cysteine is in the same position, suggesting strong structural selection due to inter-chain structural constraints. However, consistent with the phylogenetic placement into two distinct clades, each type differs in the flanking amino acids. Intriguingly, there appears to have been a secondary loss of this trait occurring in the one of the kunitz peptide clades β-bungarotoxins, with the sequences C5H0E4 (Uniprot ID) and B4ESA2 (Uniprot ID) lacking the diagnostic and structurally necessary cysteine (Figure 1).

Other derived kunitz peptides have the pathophysiological action of inhibiting the clotting regulatory enzyme plasmin, which breaks down blood clots in the body. Unsurprisingly, plasmin inhibitors have been isolated and characterized from venoms which are powerfully procoagulant. The venoms allow the snakes to subjugate their prey by triggering the rampant production of endogenous thrombin, leading to the formation of enough blood clots to induce debilitating and lethal strokes. Such toxins have been well-described for the *Daboia* genus within the Viperidae family, and the *Oxyuranus/Pseudonaja* clade within the Elapidae family (Figures 2 and 3). While *Daboia* venoms produce procoagulant toxicity through the activation of Factor X and *Oxyuranus* and *Pseudonaja* through the activation of prothrombin, both converge on the same functional outcome: the production of high levels of endogenous thrombin which convert fibrinogen to fibrin. Phylogenetic analysis (Figure 1) reveals that they show convergent neofunctionalization of the kunitz peptide such that they inhibit plasmin, thereby prolonging the half-life of the blood clots formed by the venom. Both species also show convergent modification of the same key residue into an arginine, which has been shown to be critical for activity. Both mutants of plasmin inhibitors with only this amino acid changed and native isoforms lacking this arginine from *Pseudonaja* venom were found to not affect plasmin (Flight, et al. 2005; Flight, et al. 2009). Intriguingly, other phylogenetically distinct sequences contain this mutation (Figure 3), the majority of which are in species with procoagulant venoms, including *Oxyuranus* variants that are phylogenetically distinct from the functionally characterized plasmin inhibitors, suggesting that this genus may have evolved plasmin-inhibiting kunitz peptides on multiple occasions. However, functional studies are needed to confirm that these other arginine-containing peptides are indeed plasmin inhibitors.

Selection analyses revealed very different rates of molecular evolution for this toxin type within the snake families (Figure 4 and Table 2), strongly suggesting that there is a multiplicity of undocumented novel activities yet to be discovered across the full range of this toxin class. This is consistent of the documentation of three evolutions of neurotoxin function within just the elapid snakes. The differential rates between the monomeric dendrotoxins (ω=2.10) and the disulphide-linked β-bungarotoxin subunits (ω=1.23) is consistent with the structural constraints imposed upon the β-bungarotoxin subunits by being

not only part of a multi-subunit complex, but a disulphide-linked one at that. However, despite these structural constraints, the β-bungarotoxin subunits display evidence of individual sites under positive selection. In contrast to the neurotoxins, but consistent with the high structural conservation of the enzymatic pathophysiological target, both independent lineages of plasmin inhibitors are under negative purifying selection pressures (Figure 4 and Table 2).

Table 2: Molecular evolutionary rates of kunitz peptides (See Figure 13 for structural models).

Toxin group	ω	FUBAR(-)[a]	FUBAR(+)[b]	MEME[c]	FUBAR & MEME[d]
Viperidae clade	0.96	1	8	8	4
Non-viperid clade	1.18	5	10	13	6
Neurotoxins--Dendrotoxins	2.10	0	6	7	3
Neurotoxins--Bungarotoxins	1.23	15	25	30	22
Plasmin inhibitors *Daboia*	0.80	1	0	2	0
Plasmin inhibitors *Oxyuranus/ Pseudonaja*	0.62	2	0	0	0

[a] Number of codons under negative selection according to FUBAR

[b] Number of codons under positive selection according to FUBAR

[c] Number of codons under episodic diversifying selection according to MEME

[d] Number of codons that fit criteria [b] and [c]

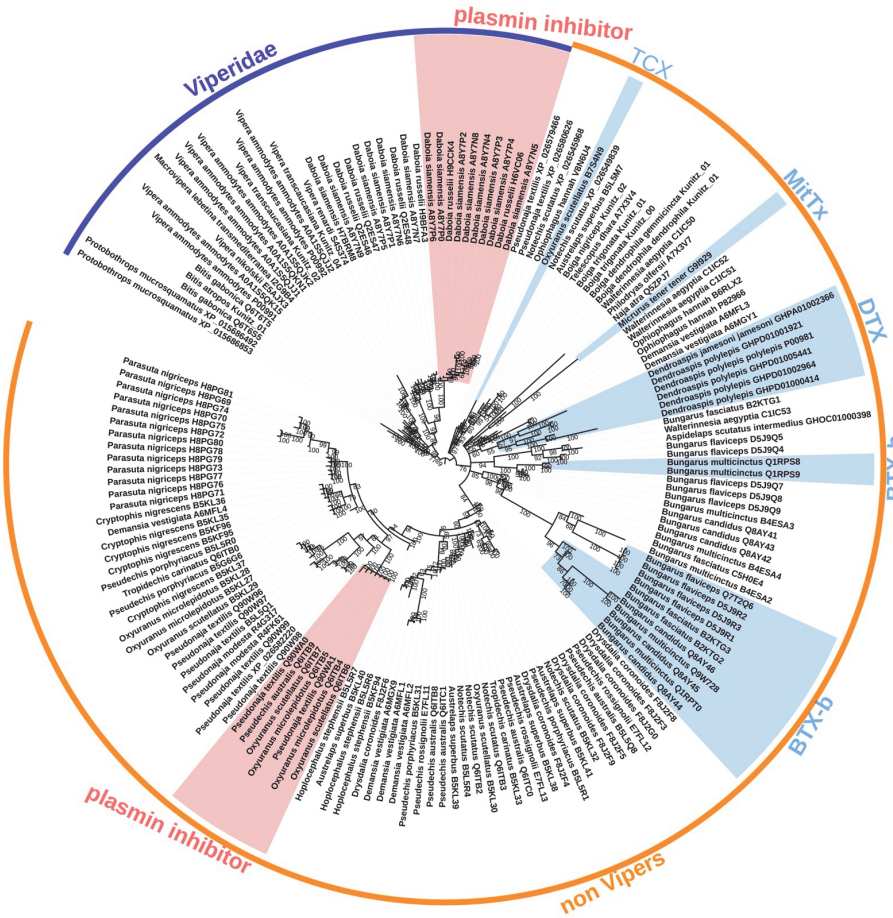


Figure 1: Molecular phylogenetic reconstruction of kunitz peptide toxins. The convergent evolution of coagulotoxins are shaded in red, while the convergent evolution of neurotoxins is highlighted in blue. TCX= taicatoxin. DTX=Dendrotoxin. BTX-b= β-bungarotoxins (characterised by the presence of a novel cysteine). Sequence alignment used for constructing phylogenetic tree can be viewed in Supplementary File 3. For tree output file for kunitz toxins, see Supplementary File 4.

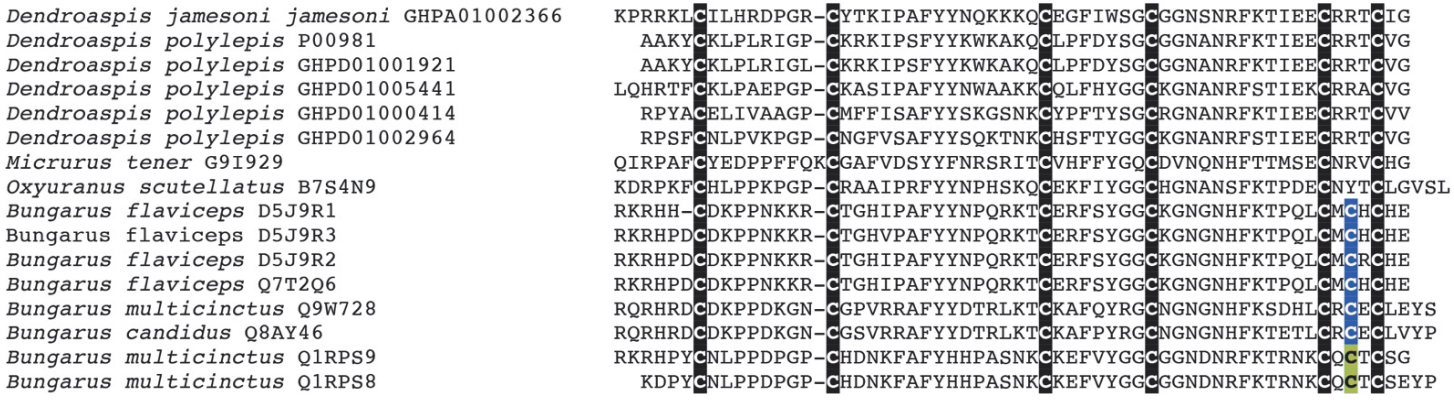


Figure 2: Sequence alignment of representative derived neurotoxic forms of kunitz peptides. Shown are: *Dendroaspis* $K_V1.1$ voltage-gated potassium channel blocking dendrotoxins; *Micrurus* acid-sensing ion channel ASIC1 activating MitTx; *Oxyuranus* cardiac voltage-dependent L-type calcium channels (Ca_V) blocking taicatoxin; and *Bungarus* voltage-gated potassium channels (K_V) blocking bungarotoxins. The convergent evolution of interchain cysteines in the *Bungarus* sequences are indicated by different highlight colours.

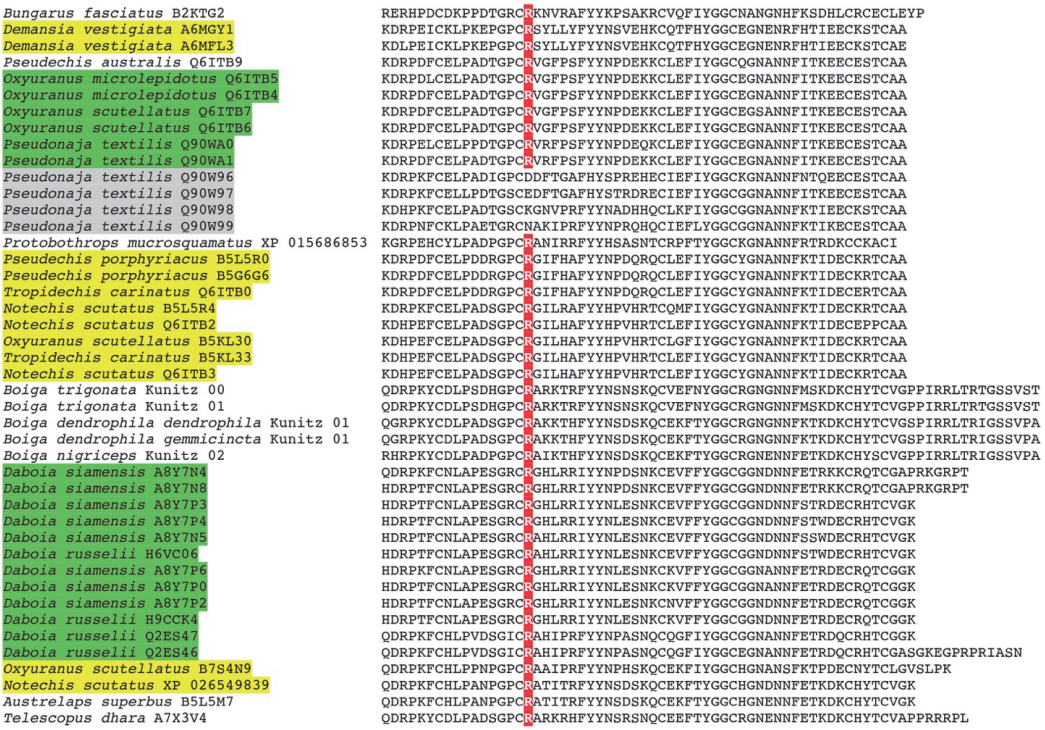


Figure 3: Sequence alignment of representative plasmin-inhibiting derived forms of kunitz peptide toxins, with the functionally important arginine residue shaded in red. Procoagulant species with confirmed plasmin inhibiting activating are shaded in green. Procoagulant species which contain derived kunitz forms with the functionally important arginine but with phylogenetically distinct sequences that have not been functionally confirmed as plasmin-inhibiting are highlighted in yellow. *Pseudonaja* variants which lack the diagnostic arginine and have been bioactivity tested confirm the lack of plasmin inhibition activity are highlighted in gray.

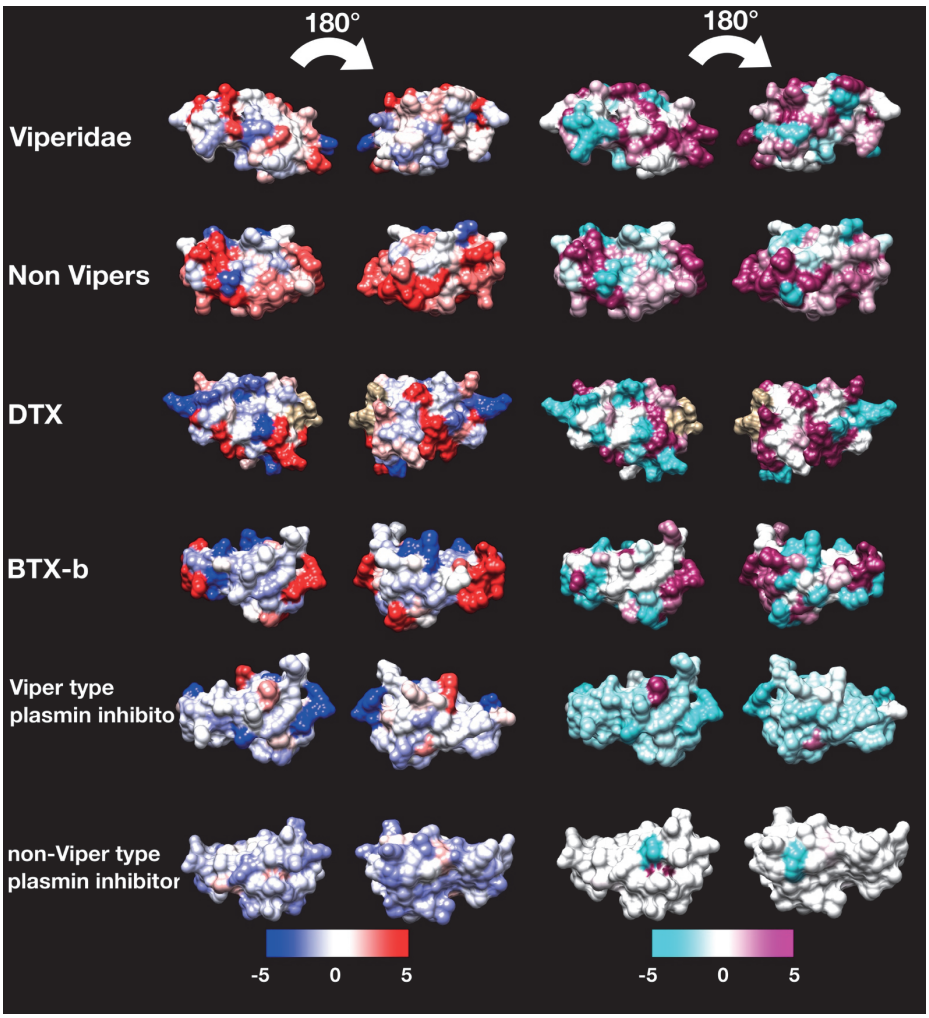


Figure 4: Molecular modelling of kunitz toxins showing sites under selection by FUBAR (left) and MEME (right) colour coded to show sites that are negatively, neutrally, or positively selected. See Table 2 for values. Protein models show front and back views colored according to FUBAR's estimated strength of selection (β-α, left) and MEME's significance levels (right). Table 1 contains the information regarding template choice for each toxin subclass.

References

Antuch W, Berndt KD, Chavez MA, Delfin J, Wuthrich K. 1993. The NMR solution structure of a Kunitz-type proteinase inhibitor from the sea anemone *Stichodactyla helianthus*. European Journal of Biochemistry 212:675-684.

Baconguis I, Bohlen CJ, Goehring A, Julius D, Gouaux E. 2014. X-ray structure of acid-sensing ion channel 1–snake toxin complex reveals open state of a Na^+-selective channel. Cell 156:717-729.

Bayrhuber M, Vijayan V, Ferber M, Graf R, Korukottu J, Imperial J, Garrett JE, Olivera BM, Terlau H, Zweckstetter M. 2005. Conkunitzin-S1 is the first member of a new Kunitz-type neurotoxin family: structural and functional characterization. Journal of Biological Chemistry 280:23766-23770.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. Nucleic Acids Research 41:D36-D42.

Beress L. 1982. Biologically active compounds from coelenterates. Pure and Applied Chemistry 54:1981-1994.

Bohlen CJ, Chesler AT, Sharif-Naeini R, Medzihradszky KF, Zhou S, King D, Sánchez EE, Burlingame AL, Basbaum AI, Julius D. 2011. A heteromeric Texas coral snake toxin targets acid-sensing ion channels to produce pain. Nature 479:410-414.

Chang L-s, Chung C, Huang H-B, Lin S-r. 2001. Purification and characterization of a chymotrypsin inhibitor from the venom of *Ophiophagus hannah* (King Cobra). Biochemical and Biophysical Research Communications 283:862-867.

Chen C, Hsu C-H, Su N-Y, Lin Y-C, Chiou S-H, Wu S-H. 2001. Solution structure of a Kunitz-type chymotrypsin inhibitor isolated from the elapid snake *Bungarus fasciatus*. Journal of Biological Chemistry 276:45079-45087.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Fiedler F. 1987. Effects of secondary interactions on the kinetics of peptide and peptide ester hydrolysis by tissue kallikrein and trypsin. European Journal of Biochemistry 163:303-312.

Flight S, Johnson L, Trabi M, Gaffney P, Lavin M, de Jersey J, Masci P. 2005. Comparison of textilinin-1 with aprotinin as serine protease inhibitors and as antifibrinolytic agents. Pathophysiology of Haemostasis and Thrombosis 34:188-193.

Flight SM, Johnson LA, Du QS, Warner RL, Trabi M, Gaffney PJ, Lavin MF, De Jersey J, Masci PP. 2009. Textilinin-1, an alternative anti-bleeding agent to aprotinin: importance of plasmin inhibition in controlling blood loss. British Journal of Haematology 145:207-211.

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.

Grzesiak A, Krokoszynska I, Krowarsch D, Buczek O, Dadlez M, Otlewski J. 2000. Inhibition of six serine proteinases of the human coagulation system by mutants of bovine pancreatic trypsin inhibitor. Journal of Biological Chemistry 275:33346-33352.

Guo C-t, McClean S, Shaw C, Rao P-f, Ye M-y, Bjourson AJ. 2013. Trypsin and chymotrypsin inhibitor peptides from the venom of Chinese *Daboia russellii siamensis*. Toxicon 63:154-164.

Harvey A, Karlsson E. 1980. Dendrotoxin from the venom of the green mamba, *Dendroaspis angusticeps*. Naunyn-Schmiedeberg's Archives of Pharmacology 312:1-6.

Harvey A, Robertson B. 2004. Dendrotoxins: structure-activity relationships and effects on potassium ion channels. Current Medicinal Chemistry 11:3065-3072.

Hedstrom L. 2002. Serine protease mechanism and specificity. Chemical reviews 102:4501-4524.

Honma T, Shiomi K. 2006. Peptide toxins in sea anemones: structural and functional aspects. Marine Biotechnology 8:1-10.

Koludarov I, Sunagar K, Undheim EA, Jackson TN, Ruder T, Whitehead D, Saucedo AC, Mora GR, Alagon AC, King G. 2012. Structural and molecular diversification of the Anguimorpha lizard mandibular venom gland system in the arboreal species *Abronia graminea*. Journal of Molecular Evolution 75:168-183.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276-3278.

Laskowski Jr M, Kato I. 1980. Protein inhibitors of proteinases. Annual Review of Biochemistry 49:593-626.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

Minagawa S, Ishida M, Shimakura K, Nagashima Y, Shiomi K. 1997. Isolation and amino acid sequences of two Kunitz-type protease inhibitors from the sea anemone *Anthopleura aff. xanthogrammica*. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 118:381-386.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Molecular Biology and Evolution 30:1196-1205.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetic 8:e1002764.

Navaneetham D, Sinha D, Walsh PN. 2010. Mechanisms and specificity of factor XIa and trypsin inhibition by protease nexin 2 and basic pancreatic trypsin inhibitor. The Journal of Biochemistry 148:467-479.

Pond SL, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. In: Statistical Methods in Molecular Evolution: Springer. p. 125-181.

Possani LD, Martin BM, Yatani A, Mochca-Morales J, Zamudio FZ, Gurrola GB, Brown AM. 1992. Isolation and physiological characterization of taicatoxin, a complex toxin with specific effects on calcium channels. Toxicon 30:1343-1364.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539-542.

Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlić A, Quesada M, Quinn GB, Westbrook JD. 2010. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Research 39:D392-D401.

Scheidig AJ, Hynes TR, Pelletier LA, Wells JA, Kossiakoff AA. 1997. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of alzheimer's amyloid β-protein precursor (APPI) and basic pancreatic trypsin inhibitor (BPTI): Engineering of inhibitors with altered specificities. Protein Science 6:1806-1824.

Van Nostrand WE, Schmaier AH, Siegel RS, Wagner SL, Raschke WC. 1995. Enhanced plasmin inhibition by a reactive center lysine mutant of the Kunitz-type protease inhibitor domain of the amyloid β-protein precursor. Journal of Biological Chemistry 270:22827-22830.

Wan H, Lee KS, Kim BY, Zou FM, Yoon HJ, Je YH, Li J, Jin BR. 2013. A spider-derived Kunitz-type serine protease inhibitor that acts as a plasmin inhibitor and an elastase inhibitor. PLoS One 8:e53343.

Wang H, Wang L, Zhou M, Yang M, Ma C, Chen T, Zhang Y, Zeller M, Hornshaw M, Shaw C. 2012. Functional peptidomics of amphibian skin secretion: a novel Kunitz-type chymotrypsin inhibitor from the African hyperoliid frog, *Kassina senegalensis*. Biochimie 94:891-899.

Wells JM, Strickland S. 1994. Aprotinin, a Kunitz-type protease inhibitor, stimulates skeletal muscle differentiation. Development 120:3639-3647.

Chapter 4. Evolution of C-type Lectin Toxins

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

C-type lectins are one of the largest protein families in mammals and reptiles and they have shown various functions including defence mechanisms against predators. Since the Viperidae split off from the remaining caenophidian snakes, a novel heterodimeric lectin type evolved through loss of the carbohydrate-binding activity. The detailed evolutionary history of these toxins is unknown. We therefore conducted large-scale transcriptome sequencing, phylogenetic analysis and selection analysis to address this issue. Our results showed that the heterodimeric lectins form three clusters: two in the Viperidae and one in the Colubridae. All three clusters are under strong selection. We discuss the evolutionary implications of our findings. We also argue that these toxins may have considerable potential in drug design and development, in biochemical assays and in drug discovery.

Keywords: C-type lectins, basal/ancestral form, dimeric form, evolution

Introduction

The first major evolutionary modification in snake-venom C-type lectin was the mutation of the EPN motif in the loop to QPD, which led to a change in specificity from mannose to galactose and thus to mediation of the alternative erythrocyte-agglutination pathway (Drickamer 1992). Apotypic heterodimeric lectin forms from snake venom (snaclec) that lack carbohydrate-binding ability but instead possess numerous neofunctionalizationed activities. They are covalently linked by a single disulfide bond, and each subunit contains three internal disulfide bonds (Andrews, et al. 1989; Usami, et al. 1993). Heterodimeric lectins represent an extreme structural derivation and form a complex with P-IIId type SVMP (Chapter 5). Examples of this apotypic form are RVV-X and carinactivase-I, which activate factor X and prothrombin, respectively (Morita 2005). The linkage between the snaclec domain and the metalloprotease/cysteine-rich and disintegrin domains involves an additional disulfide bond in the chain homologous with the alpha chain of other snaclecs.

A previous study showed two forms that lie basal to the alpha and beta viperid venom sequences in the phylogenetic tree (Fry, Jackson, et al. 2015). While viper venoms are extremely rich in heterodimeric forms, it remains unclear if those forms are present in other snake lineages. Heterodimeric lectins have been sequenced from the transcriptome of the RFS snake *Philodryas olfersii* (Ching, et al. 2006). These may represent the plesiotypic state of the heterodimers. As these sequences are known only from transcriptome sequencing, both their bioactivity and their heterodimer structure remain to be elucidated. Conversely, while a heterodimer has in fact been isolated from the venom of *Ophiophagus hannah* (ophioluxin)(Du, et al. 2002), it is known only from very small sequence fragments, and thus the phylogenetic affinity of each chain remains enigmatic.

Although structurally identical, snake venom C-type lectins proteins have a range of physiological roles that are reliant on binding to platelet cell surface receptors or coagulation factors, as mentioned above. It would be interesting to learn more about the evolutionary mechanisms that led to this functional difference. cDNA sequences of different C-type lectins and C-type lectin-like proteins from snake venoms were used to create a phylogenetic tree (Fry, et al. 2008) using the Bayesian approach. The tree revealed that these proteins split into three groups: C-type lectins, C-type lectin-like proteins, and C-type lectin-like proteins' A (or a) and B (or b) chains. The A and B chains clearly evolved from an ancient C-type lectin before snake species diversification. The evolutionary tree built using amino acid sequences shows a similar branching structure (Tani, et al. 2002).

In this chapter, we used phylogenetic reconstruction and sequence analysis to reconstruct the evolutionary history of snake venom C-type lectin toxins. Particular emphasis was given to the heterodimeric types. Analysis of selection was also utilized to evaluate the selection pressure on different phylogenetic clusters of the toxins. We argue that C-type lectins may be potential new leads in drug development.

Materials and Methods

Sequence Alignments and Phylogenetic Reconstruction

Protein sequences for all toxin sequences were retrieved from the UniProt database (https://www.uniprot.org) and NCBI database (http://www.ncbi.nlm.nih.gov), then combined with the toxin transcripts from our assembly and annotation (Chapter 2). Partial sequences, sequences with suspect assembling errors were excluded. For the blocks of sequence in between these sites, the sequences were aligned using a mix of manual alignment of the conserved cysteine locations and alignment using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) method (Edgar 2004) implemented in AliView (Larsson 2014). Manual refinement of the alignment was also involved because there are structural differences within different toxin families. The phylogenetic trees for different toxin families were reconstructed with MrBayes 3.2 (Ronquist, et al. 2012) based on the amino acid sequence alignment. The settings for MrBayes can be found in the Supplementary File 2. The output trees from MrBayes were midpoint rooting, then further edited and annotated with iTol (Letunic and Bork 2007).

Tests for Selection

Coding DNA sequences, which are corresponding to the toxin sequences used for phylogenetic analysis, were retrieved from GenBank (Benson, et al. 2012) and our assembly. Using AliView and the MUSCLE method, the sequences were trimmed to only containing codons that translate to the mature protein, then translated, aligned, and reverse translated. Clades were created based on taxonomy and structural differences (functional domains/motifs, for example). The resultant codon alignments were used to create phylogenetic trees for each clade using the same methods outlined in the 'Phylogenetic Reconstruction' section. All following studies were conducted using these tree topologies.

Calculating the ratio of nonsynonymous nucleotide substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous sites (dS) (ω=dN/dS) for each codon in the alignment might reveal if a gene is undergoing rapid evolution or stays functionally restricted. Codons developing with $\omega > 1$ are thought to have evolved under positive selection (functional diversity), whereas codons evolving with $\omega < 1$ are thought to have evolved under purifying selection. Sites with a value of $\omega = 1$ are believed to evolve in a neutral manner. In order to find the most likely groups on which positive selection has been working, we conducted a series of experiments integrating data from site-based and lineage-specific studies.

Due to their various emphases, we employed many of the selection tests developed in HyPhy v 2.220150316 beta (Pond and Muse 2005) to study the patterns of selection acting on distinct toxin families. The Analyze Codon Data analysis in HyPhy produces overall alignment values, whereas the FUBAR technique assesses the intensity of persistent positive or negative selection on individual amino acids (Murrell, et al. 2013). The Mixed Effects Model of Evolution (MEME) approach, on the other

hand, finds particular locations that have been exposed to positive selection in the past (Murrell, et al. 2012).

Protein Modelling

To map residues evolving under positive selection in three-dimensional (3D) structures, sample sequences from the RCSB PDB database (Rose, et al. 2010) were used to create bespoke models for each clade belonging to various toxin families (Table 1). Alignments of each clade were trimmed to match these PDB structures. To render and colour the 3D structure of the proteins, we utilized the UCSF Chimera program v 1.10.2 with attribute files generated from FUBAR and MEME results. For FUBAR, we used the value from the beta-alpha column which is a measure of the difference between the rates of non-synonymous (beta) and synonymous (alpha) mutations. For MEME, since MEME estimates two rates of positive selection and gives each a probability, we take the weighted average of those two and then subtract alpha to arrive at a similar value to the one we used for FUBAR.

Table 1: Custom models for protein modelling.

Toxin Groups	Sequence for 3D modeling	PDB ID
ancestral	Erythrolamprus_poecilogyrus_A7X3Z7	5f2q
viper dimeric alpha	Bothrops_jararaca_Q56EB1	5f2q
non-viper dimeric	Philodryas_olfersii_Q09GK0	1v7p
viper dimeric beta	Echis_multisquamatus_Q7T2Q0	1fvu

Results and Discussion

The basal form of lectin toxins in reptile venoms is a single-chain form which may form non-covalently linked complexes and contains diagnostic tripeptide functional motif (Figures 1 and 2) (Walker, et al. 2004; Arlinghaus, et al. 2015). Consistent with previous analyses (Fry, et al. 2008), our phylogenetic results suggest that the earliest functional motif is the amino acids EPN (glutamic acid + proline + asparagine). Our results also indicate that the QPD (glutamine + proline + aspartic acid) motif has arisen on two convergent occasions, once in the last common ancestor of the advanced snakes, and again in the last common ancestor of the Australian radiation of elapids.

In addition, other mutations in the functional motif were documented across a myriad of lineages: EPG (glutamic acid + proline + glycine) in *Parasuta nigriceps* within the Elapidae; EPK (glutamic acid + proline + lysine) in *Heterodon nasicus* within the Colubridae; KPK (lysine + proline + lysine) in *Tropidolaemus subannulatus* within the Viperidae; KPN (lysine + proline + asparagine) in *Homalopsis buccata* within the Homalopsidae; KPS (lysine + proline + serine) in *Micrurus corallinus* within the Elapidae; KRN (lysine + arginine + asparagine) in *Leioheterodon madagascarensis* within the Lamprophiidae; LTD (leucine + threonine + aspartic acid) in *Bitis gabonica* within the Viperidae; and QPN (glutamine + proline + asparagine) in *Vipera transcaucasiana* within the Viperidae (Figures 1 and 2).

To-date only viperid venom variants of the QPD form have had their bioactivity tested. They were shown agglutinate erythrocytes and promote edema by increasing vascular permeability (Guimarães-Gomes, et al. 2004; Panunto, et al. 2006; Lin, et al. 2007). The impacts of the extreme diversifications of the key functional motif shown in this study upon the functions of the toxins are entirely unknown, and require further research. The overall ω value for these toxins was only 0.72, but there were 14 sites identified as positively selected by FUBAR & MEME, which is an indication that the variation which occurs in these toxins is tightly constrained and only occurs at a relatively small subset of positions (Table 2).

In addition to the ancestral single-chain form, a disulphide-linked dimer composed of two different lectins has long been known from viperid venoms (sometimes referred to as snaclecs, or C-type lectins), with variants producing a wide diversity of coagulotoxic effects including inhibition of the clotting factor vWF and clotting enzymes such as Factors IXa and XIa (Arlinghaus, et al. 2015). In addition to the diagnostic newly evolved cysteines that facilitate the inter-chain disulphide bond leading to the dimeric tertiary structure, this type is also molecularly distinct because they have lost the functional motif (Figure 1). The α and β chains have the interchain cysteine in the same position which suggests that either these toxins evolved from a single gene that produced a homomeric ancestral toxin and subsequently underwent duplication and sub-functionalization or that structural constraints led to the novel cysteine mutation occurring at the same location in two different genes. The α chain is readily distinguished from the β chain by a characteristic glutamine motif present immediately before this diagnostic cysteine (Figure 2).

Not only did we recover multiple variants from non-viperid snakes that possessed the diagnostic cysteine of the dimeric lectin form, but also forms with and without the glutamine motif form were present (Figure 2). This suggests that the evolution of the dimeric lectin toxins preceded the divergence of Viperidae from other advanced snakes. However, the rates of evolution are reflective of the explosive diversification of this toxin type within the viperids (Figure 3 and Table 2). A previous study found differential rates of evolution between the α-subunit and β-subunits based on one species (*Crotalus helleri*) (Sunagar, et al. 2014). Consistent with that study, we found that the α-subunit and β-subunit were subject to different selection pressures when analysed across all viperid species in this study. The α-subunit had an overall ω value of 1.40 with 49 sites shown to be positively selected by FUBAR & MEME, while the β-subunit had an overall ω value of 1.27 with 38 sites shown to be positively selected by FUBAR & MEME (Table 2). In contrast, the non-viperid dimeric forms (excluding the unique diversification within *Helicops*) had an overall neutral ω of 0.97 but with 10 sites shown as positively selected by FUBAR & MEME (Table 2).

The comparative analysis of 3D modelling (Figure 3) on different clades showed that those residues under positive selection are located on different position of the surface for the viperid α-subunit versus β-subunits, and the other forms as well, indicative different selective forces and the potential for the discovery of novel activities within the non-viperid dimeric and monomeric forms. Structure-function studies on these toxins may therefore be particularly interesting. The unique form present in the venom of *Helicops leopardinus,* which has novel insertions in the key functional region, including the evolution of novel cysteines may also be of particular interest for these future research efforts (Figure 1).

Table 2: Molecular evolutionary rates of lectins (See Figure 3 for modelling).

Toxin group	ω	FUBAR(-)[a]	FUBAR(+)[b]	MEME[c]	FUBAR & MEME[d]
Ancestral (monomeric)	0.72	44	14	31	14
Derived (dimeric)--Non-vipers excluding *Helicops*	0.97	6	16	21	10
Derived (dimeric)--*Helicops*	1.16	0	6	7	3
Derived (dimeric)--Viperidae-alpha	1.40	23	54	60	49
Derived (dimeric)--Viperidae-beta	1.27	17	45	54	38

[a] Number of codons under negative selection according to FUBAR

[b] Number of codons under positive selection according to FUBAR

[c] Number of codons under episodic diversifying selection according to MEME

[d] Number of codons that fit criteria [b] and [c]

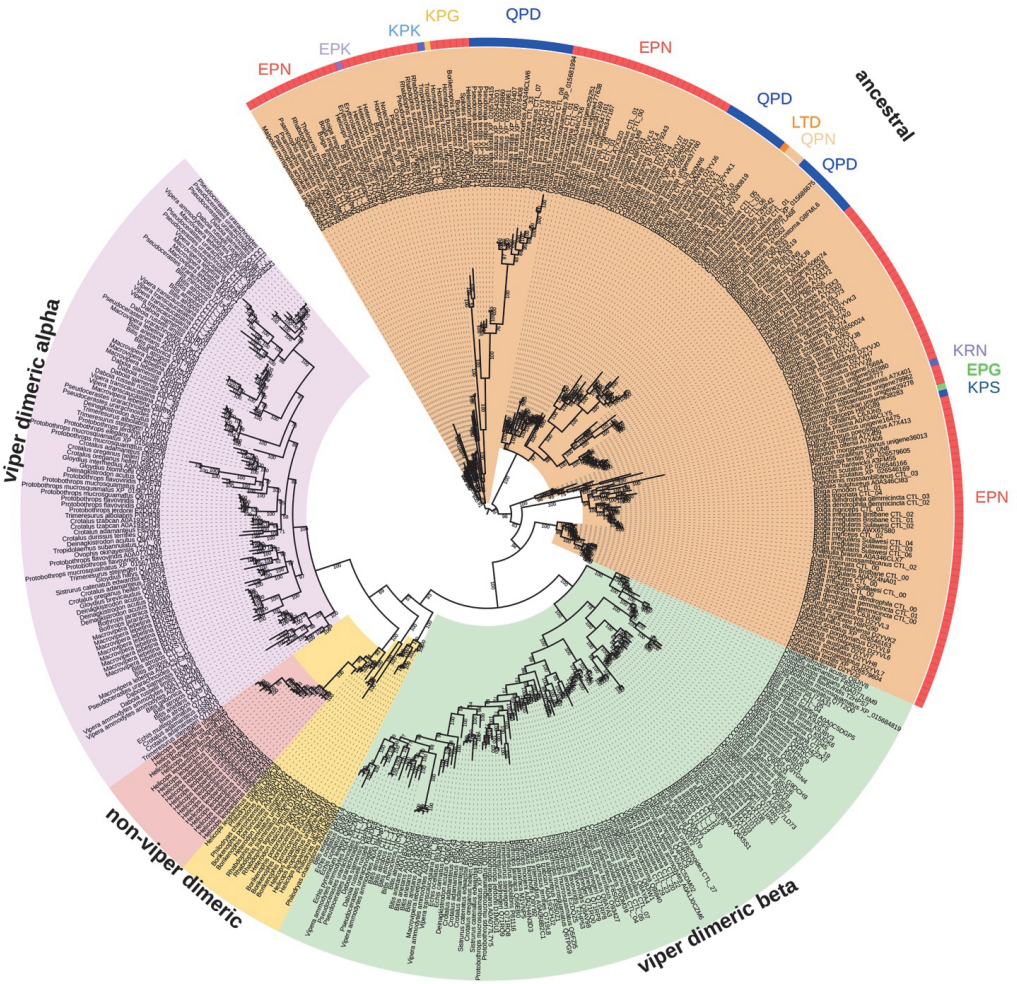


Figure 1: Molecular phylogenetic reconstruction of C-type lectin toxins. The ancestral monomeric form is shaded annotated with variations in the key functional amino acid triad. Sequence alignment for constructing phylogenetic tree can be viewed in Supplementary File 5. For tree output file for lectin toxins, see Supplementary File 6.

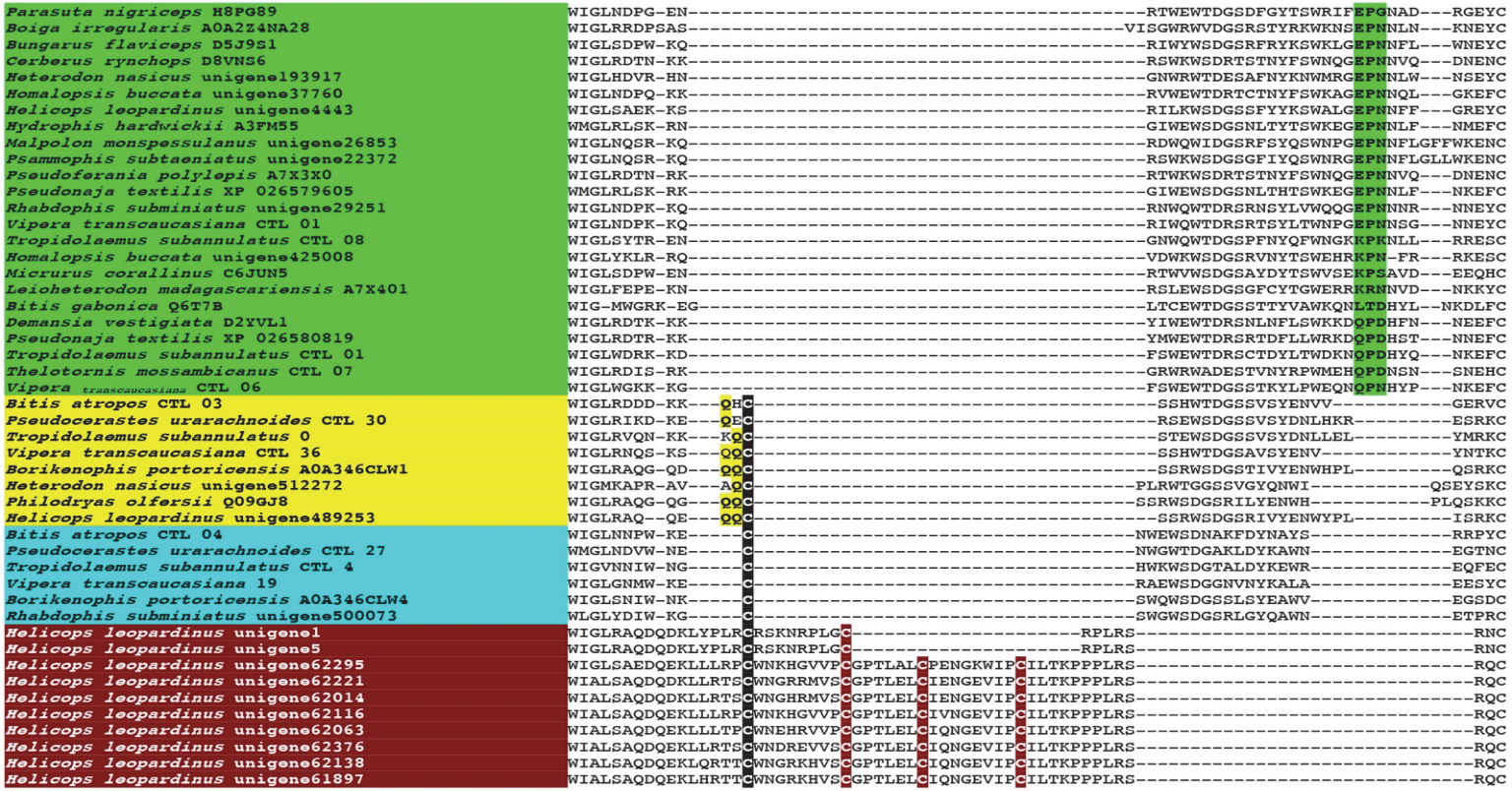


Figure 2: Partial amino acid sequence alignment of representative C-type lectin toxins in the region that contains the key functional amino acid triad (shaded in green). The alpha chain of the derived cysteine-linked heterodimeric form is shaded in yellow, including the glutamine motif that diagnoses the alpha chain, while the beta chain is highlighted in blue. The newly evolved dimer-forming cysteine is shaded in black. The sequence alignment also shows the insertion characteristic of an extremely derived form known currently only from *Helicops leopardinus,* which contains unique cysteines shaded in burgundy.

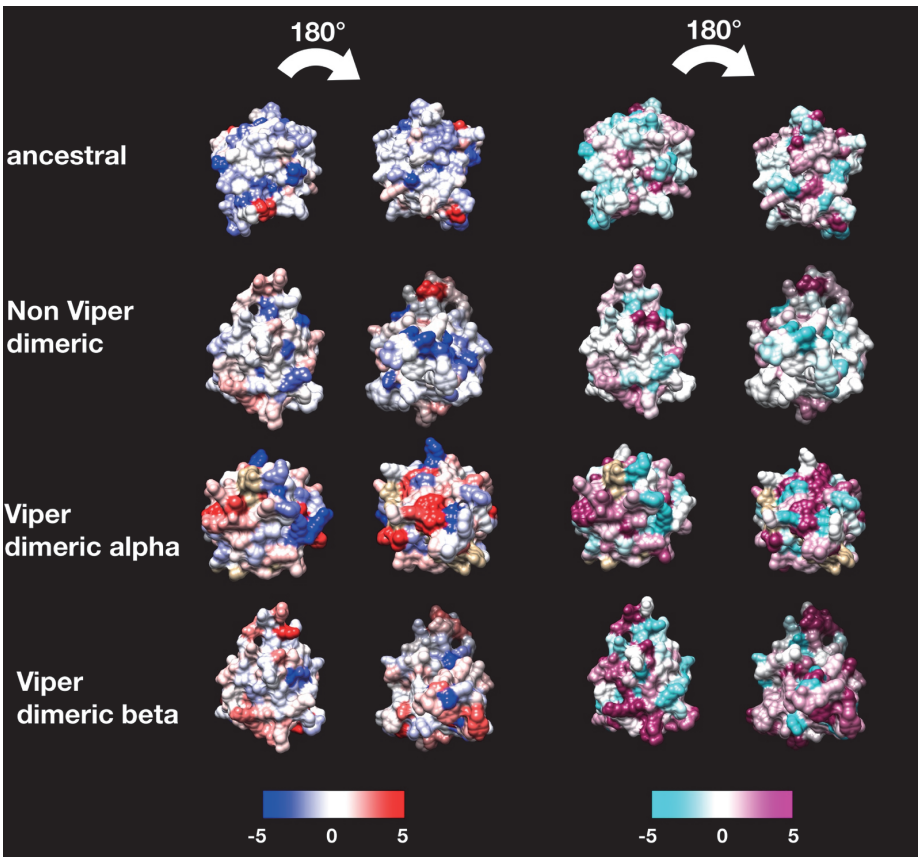


Figure 3: Molecular modelling of C-type lectin toxins showing sites under selection by FUBAR (left) and MEME (right) colour coded to show sites that are negatively, neutrally, or positively selected. See Table 2 for values. Protein models show front and back views colored according to FUBAR's estimated strength of selection (β-α, left) and MEME's significance levels (right). Table 1 contains the information regarding template choice for each toxin subclass.

References

Andrews RK, Booth WJ, Gorman JJ, Castaldi PA, Berndt MC. 1989. Purification of botrocetin from Bothrops jararaca venom. Analysis of the botrocetin-mediated interaction between von Willebrand factor and the human platelet membrane glycoprotein Ib-IX complex. Biochemistry 28:8317-8326.

Arlinghaus FT, Fry BG, Sunagar KK, Jackson TNW, Eble JA, Reeks T, Clemetson KJ. 2015. Lectin proteins. In. Venomous reptiles and their toxins: evolution, pathophysiology and biodiscovery. New York: Oxford University Press. p. 299-311.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. Nucleic Acids Research 41:D36-D42.

Ching AT, Rocha MM, Leme AFP, Pimenta DC, Maria de Fátima DF, Serrano SM, Ho PL, Junqueira-de-Azevedo IL. 2006. Some aspects of the venom proteome of the Colubridae snake *Philodryas olfersii* revealed from a Duvernoy's (venom) gland transcriptome. FEBS letters 580:4417-4422.

Drickamer K. 1992. Engineering galactose-binding activity into a C-type mannose-binding protein. Nature 360:183-186.

Du X-Y, Clemetson JM, Navdaev A, Magnenat EM, Wells TN, Clemetson KJ. 2002. Ophioluxin, a convulxin-like C-type lectin from *Ophiophagus hannah* (King cobra) is a powerful platelet activator via glycoprotein VI. Journal of Biological Chemistry 277:35124-35132.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Fry BG, Jackson TNW, Takacs Z, Reeks T, Sunagar K. 2015. C-type natriuretic peptides. In: Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery. New York: Oxford University Press. p. 318-326.

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.

Guimarães-Gomes V, Oliveira-Carvalho AL, de LM Junqueira-de-Azevedo I, Dutra DL, Pujol-Luz M, Castro HC, Ho PL, Zingali RB. 2004. Cloning, characterization, and structural analysis of a C-type lectin from *Bothrops insularis* (BiL) venom. Archives of Biochemistry and Biophysics 432:1-11.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276-3278.

Letunic I, Bork P. 2007. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

Lin L-P, Lin Q, Wang Y-Q. 2007. Cloning, expression and characterization of two C-type lectins from the venom gland of *Bungarus multicinctus*. Toxicon 50:411-419.

Morita T. 2005. Structures and functions of snake venom CLPs (C-type lectin-like proteins) with anticoagulant-, procoagulant-, and platelet-modulating activities. Toxicon 45:1099-1114.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Molecular Biology and Evolution 30:1196-1205.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetic 8:e1002764.

Panunto PC, Da Silva MA, Linardi A, Buzin MP, Melo SE, Mello SM, Prado-Franceschi J, Hyslop S. 2006. Biological activities of a lectin from *Bothrops jararacussu* snake venom. Toxicon 47:21-31.

Pond SLK, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. In. Statistical Methods in Molecular Evolution: Springer. p. 125-181.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539-542.

Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlić A, Quesada M, Quinn GB, Westbrook JD. 2010. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Research 39:D392-D401.

Sunagar K, Undheim EA, Scheib H, Gren EC, Cochran C, Person CE, Koludarov I, Kelln W, Hayes WK, King GF. 2014. Intraspecific venom variation in the medically significant Southern Pacific Rattlesnake (*Crotalus oreganus helleri*): biodiscovery, clinical and evolutionary implications. Journal of Proteomics 99:68-83.

Tani A, Ogawa T, Nose T, Nikandrov NN, Deshimaru M, Chijiwa T, Chang C-C, Fukumaki Y, Ohno M. 2002. Characterization, primary structure and molecular evolution of anticoagulant protein from *Agkistrodon actus* venom. Toxicon 40:803-813.

Usami Y, Fujimura Y, Suzuki M, Ozeki Y, Nishio K, Fukui H, Titani K. 1993. Primary structure of two-chain botrocetin, a von Willebrand factor modulator purified from the venom of *Bothrops jararaca*. Proceedings of the National Academy of Sciences 90:928-932.

Walker JR, Nagar B, Young NM, Hirama T, Rini JM. 2004. X-ray crystal structure of a galactose-specific C-type lectin possessing a novel decameric quaternary structure. Biochemistry 43:3783-3792.

Chapter 5. Evolution of Novel Structures and Functions in Snake Venom Metalloproteinases

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

Snake Venom Metalloproteinases (SVMPs) are toxins found in snake venom that may be used to study the development of protein structure and function. SVMP has undergone extensive gene duplication and domain loss over its evolution. However, nothing is known about the development of the P-IIId SVMP and shortened variants of SVMPs. This research looked at the evolutionary mechanism that led to structural and functional diversity within the P-IIId subfamily and the shortened SVMP type. The discovery of a subset of amino acid positions that are targets of positive Darwinian selection, resulting in rapid structural and functional diversity within and within disintegrin subfamilies, is shown here. We can establish genetic and temporal characteristics throughout the evolutionary pathway of distinct P-III SVMP lineages by clustering within the phylogeny and connecting positively selected sites to structural and functional areas. We also discovered that the shortened SVMP propeptide undergoes a rapid evolution.

Keywords: PIII-d SVMP, SVMP propeptide, novel domain, evolution

Introduction

Snake metalloproteinases venoms (SVMPs), are discovered in a variety of sophisticated snake lineages, showing hemorrhagic properties in snake prey. Snake venom glands on the base of the advanced snake (caenophidian) radiation is considered to have recruited SVMPs (Gutiérrez, et al. 2008; Casewell, et al. 2009; Wagstaff, et al. 2009; Jiang, et al. 2011; Petras, et al. 2011; Ching, et al. 2012). They are frequently reported to be the major venom components in viperid snake venom, although they are generally considerably less important in the venom of other snake families (Gutiérrez, et al. 2008; Wagstaff, et al. 2009; Casewell, et al. 2011; Jiang, et al. 2011; Petras, et al. 2011; Ching, et al. 2012).

SVMPs are classified into three classes according to their domain structures in the C-terminal region (Casewell, et al. 2015b): PI SVMP contains (in downstream order) three domains: signal peptide+ propeptide + metalloprotease domains; PII SVMP contains four domains: signal peptide+ propeptide + metalloprotease + disintegrin domains; and ancestral PIII SVMP contains five domains: signal peptide+ propeptide + metalloprotease + disintegrin-like + cysteine-rich domains.

P-III SVMP have been isolated from a wide lineage of snakes, including both three FFS lineages and some of RFS lineages (Fry, et al. 2008; Casewell, et al. 2011). To date, all SVMP sequences recovered from non-viper lineages belong to the P-III SVMP type and form mono clade on the base of the SVMP toxin radiation with undergoing considerable structural and functional alteration over evolutionary time (Fry, et al. 2008; Casewell, et al. 2011). Apotypic P-III SVMP subclasses include those that remain intact (P-IIIa), those that proteolytically process the disintegrin-like and cysteine-rich domains (P-IIIb), those that form intact dimeric structures (P-IIIc), and those that bond covalently with C-type lectin venom components (P-IIId) (Fox and Serrano 2008).

Following the split of the viperid snakes from the remaining caenophidian snakes, gene duplication resulted in considerable diversification of P-III SVMPs within the former, with multiple P-III isoforms typically retained in the venom of any one species (Casewell, et al. 2009; Wagstaff, et al. 2009). In light of the conservation of cysteines in the P-III plesiotypic SVMPs, amino acid alterations following the recruitment of the SVMP scaffold into venom result in the acquisition and removal of additional cysteine residues crucial for enabling structural changes and posttranslational modifications of apotypic SVMPs. Some P-III SVMPs have evolved additional procoagulant functions by activating other components of the clotting cascade, such as Factor X (Kisiel, et al. 1976; Hofmann and Bon 1987; Takeya, et al. 1992; Siigur, et al. 2004). By now, the majority of the P-III SVMP sequences remain structurally uncharacterized also with their evolutionary history.

There are two known examples of SVMP genes that are extensively truncated, resulting in the expression of proteins containing only parts of the propeptide domain (Fry, et al. 2008; Casewell, et al. 2011; Brust, et al. 2013). These atypical SVMPs have been identified from the Lamprophiid genus *Psammophi*s and the viperid snake genus *Echis*, and they lack the metalloprotease domain (and additional C-terminal domains) and zinc-binding motif that characterize the adamalysins. The propeptide-only expression in *Psammophis* and *Echis* are convergent derivations (Brust, et al. 2013). The *Psammophis* monodomain

form had a lot of variance in sequence, but the *Echis* monodomain pre-propeptide form was virtually similar to the pre-propeptide region produced in the multidomain gene.

Although these genes have yet to be identified as translated and secreted in *Echis* venom, the resultant toxins are present in *Psammophis* venom (Fry, Lumsden, et al. 2003b; Brust, et al. 2013), where they exhibit an entirely novel neurotoxic activity: inhibition of postsynaptic α7 nicotinic acetylcholine receptors (Brust, et al. 2013). Notably, domain loss in *Psammophis* has resulted in increased selection pressure, which has driven a rapid rate of mutations in the propeptide domain and resulted in protein neofunctionalization, analogous to the processes observed in the evolution of P-I and P-II SVMPs (Casewell, et al. 2011; Brust, et al. 2013). Bioassays on two post-translationally cleaved new proline-rich peptides from the *P. mossambicus* propeptide domain revealed that they had been neofunctionalizationed for selective inhibition of human a7 neuronal nicotinic acetylcholine receptors, according to Fry (Brust, et al. 2013).

The goal of this study was to see if (1) the multi domains within P-III SVMP have had the same evolutionary history, and if so, what impact changes in molecular structure have had on the rate of evolution and neofunctionalization; and (2) whether the truncated SVMP type is widespread within the genus *Psammophis*, and if so, what phylogenetic history it has. As a result, we use Bayesian inference analyses on multiple domain partitions of an extensive P-III SVMP data set to trace the evolutionary history of the ancestral P-III SVMPs and truncated SVMPs, as well as their constituent domains, before using adaptive molecular evolution tests on points of the tree where domain alterations were inferred to have occurred. The results of this study's studies demonstrate the uniqueness of SVMP subtype development with neofunctionalization.

Materials and Methods

Sequence Alignments and Phylogenetic Reconstruction

Protein sequences for all toxin sequences were retrieved from the UniProt database (https://www.uniprot.org) and NCBI database (http://www.ncbi.nlm.nih.gov), then combined with the toxin transcripts from our assembly and annotation (Chapter 2). Partial sequences, sequences with suspect assembling errors were excluded. For the blocks of sequence in between these sites, the sequences were aligned using a mix of manual alignment of the conserved cysteine locations and alignment using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) method (Edgar 2004) implemented in AliView (Larsson 2014). Manual refinement of the alignment was also involved because there are structural differences within different toxin families. The phylogenetic trees for different toxin families were reconstructed with MrBayes 3.2 (Ronquist, et al. 2012) based on the amino acid sequence alignment. The settings for MrBayes can be found in the Supplementary File 2. The output trees from MrBayes were midpoint rooting, then further edited and annotated with iTol (Letunic and Bork 2007).

Tests for Selection

Coding DNA sequences, which are corresponding to the toxin sequences used for phylogenetic analysis, were retrieved from GenBank (Benson, et al. 2012) and our assembly and annotation (Chapter 2). Using

AliView and the MUSCLE method, the sequences were trimmed to only contain codons that translate to the mature protein, then translated, aligned, and reverse translated. Clades were created based on taxonomy and structural differences (functional domains/motifs, for example). The resultant codon alignments were used to create phylogenetic trees for each clade using the same methods outlined in the above 'Phylogenetic Reconstruction' section. All following studies were conducted using these tree topologies.

Calculating the ratio of nonsynonymous nucleotide substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous sites (dS) (ω=dN/dS) for each codon in the alignment might reveal if a gene is undergoing rapid evolution or stays functionally restricted. Codons developing with ω >1 are thought to have evolved under positive selection (functional diversity), whereas codons evolving with ω<1 are thought to have evolved under purifying selection. Sites with a value of ω=1 are believed to evolve in a neutral manner. In order to find the most likely groups on which positive selection has been working, we conducted a series of experiments integrating data from site-based and lineage-specific studies.

Due to their various emphases, we employed many of the selection tests developed in HyPhy v 2.220150316 beta (Pond and Muse 2005) to study the patterns of selection acting on distinct toxin families. The Analyze Codon Data analysis in HyPhy produces overall alignment values, whereas the FUBAR technique assesses the intensity of persistent positive or negative selection on individual amino acids (Murrell, et al. 2013). The Mixed Effects Model of Evolution (MEME) approach, on the other hand, finds particular locations that have been exposed to positive selection in the past (Murrell, et al. 2012).

Protein Modelling

To map residues evolving under positive selection in three-dimensional (3D) structures, sample sequences from the RCSB PDB database (Rose, et al. 2010) were used to create bespoke models for each clade belonging to various toxin families (Table 1). Alignments of each clade were trimmed to match these PDB structures. To render and colour the 3D structure of the proteins, we utilized the UCSF Chimera program v 1.10.2 with attribute files generated from FUBAR and MEME results. For FUBAR, we used the value from the beta-alpha column which is a measure of the difference between the rates of non-synonymous (beta) and synonymous (alpha) mutations. For MEME, since MEME estimates two rates of positive selection and gives each a probability, we take the weighted average of those two and then subtract alpha to arrive at a similar value to the one we used for FUBAR.

Results and Discussion

Snake Venom Metalloproteinases (SVMPs)

While SVMPs have long been known as one of the dominant venom types in viperid venoms, increasing evidence is emerging of their importance in the venoms of other families (Kamiguti, et al. 2000; Peichoto, et al. 2007; Fry, et al. 2008; Casewell, et al. 2015a; Debono, et al. 2017; Modahl, et al. 2018; Debono, et al. 2020). The basal SVMP structural form (P-III) is a final processed protein consisting of three domains: protease + disintegrin + cysteine-rich. Domain-deletion forms are largely known only from viperid

venoms including the P-II (protease + disintegrin, with the cysteine-rich domain deleted), and P-I (protease only, with both the disintegrin and cysteine-rich domains deleted) (Casewell, et al. 2015a). Intriguingly the P-I derived condition appears to have evolved convergently in the dipsadinae lineage within the Colubridae rear-fanged snake family (Campos, et al. 2016).The P-III form, however, remains a major constituent of viperid venoms and other than select dipsadinae lineages, it is the only form present in non-viperid snakes. Consistent with the structural and functional diversification of SVMP within the viperids, phylogenetic analysis in this study revealed evidence of extensive gene duplication in the last common ancestor of the viperid snakes (Figure 1). In contrast, for the colubrid and elapid snakes, the sequences broadly follow organismal relationships, with diversification events largely confined to within a particular lineage.

Table 1: Custom models for protein modelling.

Clade	Sequence for 3D modeling	PDB ID
Elapidae	Cerberus_rynchops_D8VNS0	2dw2
Colubriae	Naja_atra_D5LMJ3	3k7l
Viperidae	Trimeresurus_stejnegeri_Q2LD49	3k7l

Within the P-III SVMP enzymatic toxin class, there have been convergent structural derivations characterized by the evolution of a new cysteine that allows these toxins to form covalently linked multimers with lectin dimers. SVMP with this novel cysteine are termed P-IIId. Phylogenetic analysis suggests that the P-IIId type have evolved on at least three occasions: *Bothrops*; the last common ancestor of the genera *Daboia, Macrovipera, Montivipera* and *Vipera*; and in the genus *Echis* (Figure 1). *Echis* venoms contain two distinct forms of P-IIId which confirms previous hypotheses that were based on sequence similarity but lacked phylogenetic analyses (Casewell, et al. 2009). Sequence analysis in our study shows that while the cysteines have evolved in homologous regions of the SVMP scaffold, suggesting structural constraints in the formation of a multimeric complex, they differ slightly in position and, consistent with independent evolutions, differ in flanking residues (Figure 2).

In addition to structural diversifications, SVMPs have acquired a number of novel functions, the most common of which is procoagulant activity (Casewell, et al. 2015a). Identifying sequences in our phylogeny that have demonstrated procoagulant effects suggest that, within the viperids, the procoagulant trait has independently evolved within the viperine subfamily (such as *Echis, Daboia, Macrovipera, Pseudocerastes,* and *Vipera*) and the crotaline subfamily (*Bothrops*) (Figure 1). Clotting factor activation has been documented in the additional crotaline genera *Calloselasma* and *Crotalus* (Debono, et al. 2019; Seneci, et al. 2021), but the toxins responsible have not been sequenced. Consequently, their phylogenetic affinity to the *Bothrops*-type procoagulant P-III SVMP are unknown, and it cannot be determined whether procoagulant SVMPs have evolved once or several times in the pit vipers. Once the sequences become available, this will be resolved by whether the toxins form a

monophyletic group with the *Bothrops* toxins or if they form distinct clades. In addition to evolving at least twice within the viperids, the procoagulant SVMP trait evolved independently again in the last common ancestor of the colubrid genera *Dispholidus* and *Thelatornis* (Debono, et al. 2017) and also in the elapid genus *Micropechis* (Gao, et al. 2002). If P-III SVMP are responsible for the procoagulant activity shown for *Atractaspis* venoms (Oulion, et al. 2018), then this would represent another convergent evolution of this trait. Similarly, the toxins responsible for the procoagulant toxicity of the *Rhabdophis* genus have not been identified (Iddon and Theakston 1986; Komori, et al. 2017), but if the *Rhabdophis* procoagulant effect is due to a SVMP, this would almost certainly represent another instance of functional convergence considering the tens of millions of years of separation between this genus and the other procoagulant lineages.

The overall ω value for all lineages was consistently higher for the cysteine-rich domains than for the disintegrin or protease domain. This suggests that the cysteine-rich domain is crucial for target binding prior to the interaction of the catalytic site located on the protease domain, and therefore this is a critical domain for the evolution of neofunctionalization. Analysis of selection (Table 2) and 3D modelling (Figure 3) showed that more than half of the positively selected sites detected were confined to the protease domain; of the remaining variations, more were found in the cysteine-rich domains than in the disintegrin-like domains. Again, this pattern suggests a bias in positive selection toward the protease domain, consistent with this domain being the subunit responsible for the enzymatic activity.

Table 2: Molecular evolutionary rates of SVMP (See Figure 3 for modelling).

Clade	Domains	ω	FUBAR (-)[a]	FUBAR (+)[b]	MEME[c]	FUBAR & MEME[d]
Colubridae	Full length secreted form	1.19	74	103	46	33
	Peptidase domain	1.54	19	53	50	39
	Disintegrin domain	0.72	14	12	12	10
	Cys-rich domain	1.64	18	30	36	24
Elapidae	Full length secreted form	1.23	36	75	86	56
	Peptidase domain	1.47	13	25	35	21
	Disintegrin domain	0.86	5	8	9	7
	Cys-rich domain	1.76	7	29	36	21
Viperidae	Full length secreted form	1.37	79	132	159	124
	Peptidase domain	1.44	22	57	78	55
	Disintegrin domain	0.80	23	21	21	18
	Cys-rich domain	1.63	27	41	50	38

[a] Number of codons under negative selection according to FUBAR
[b] Number of codons under positive selection according to FUBAR
[c] Number of codons under episodic diversifying selection according to MEME
[d] Number of codons that fit criteria [b] and [c]

Figure 1: Molecular phylogenetic reconstruction of SVMP toxins with the lineage specific amplification of particular forms shaded in pink (Colubridae), blue (Elapidae), or yellow (Viperidae). Gray shading shows the convergent evolutions of procoagulant functionally derived forms. Colours on the outside of the ring designate the convergent evolutions of the P-IIId structurally derived forms. Sequence alignment for constructing phylogenetic tree can be viewed in Supplementary File 7. For tree output file for SVMP toxins, see Supplementary File 8.

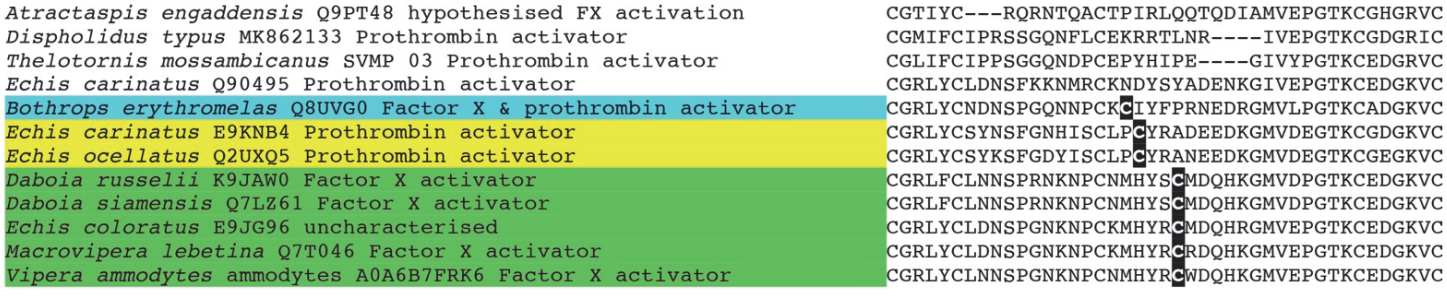


Figure 2: Partial amino acid sequence alignment of representative SVMP with coloured shading of species names indicating the three convergent evolutions of interchain cysteines, diagnostic of the P-IIId structurally derived forms. Other procoagulant functionally derived forms (species names not shaded) are included for comparison. Cysteines forming links to lectin dimers are highlighted in black background.

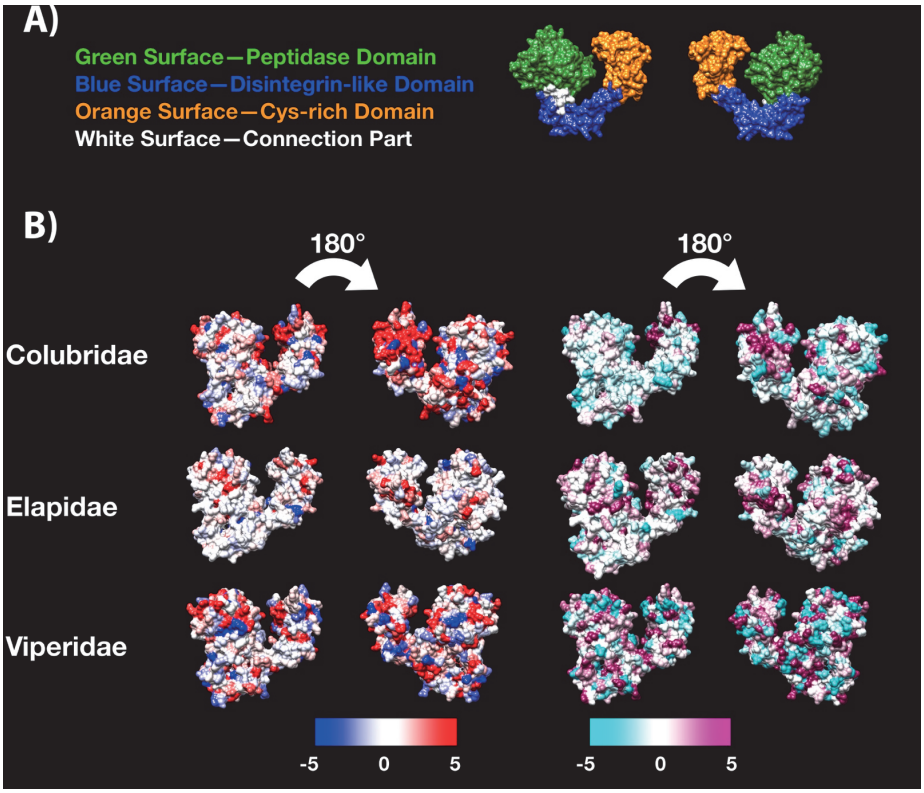


Figure3: 3D models of PIII SVMP. A) location of the three domains which make up the SVMP toxin type. B) Molecular modelling of SVMP showing sites under selection by FUBAR (left) and MEME (right) colour coded to show sites that are negatively, neutrally, or positively selected. See Table 1 for values. Protein models show front and back views colored according to FUBAR's estimated strength of selection (β-α, left) and MEME's significance levels (right). Table 2 contains the information regarding template choice for each toxin subclass.

SVMP propeptide domain novel toxins

In addition to the structural variations noted in the above section, on at least two independent occasions the propeptide domain of SVMP genes have been recruited as toxins in their own right, without accompanying expression of any of the three domains making up the P-III enzyme. This was first noted in *Echis* venoms, where the truncation is formed by stop codons terminating otherwise unremarkable sequences (Casewell, et al. 2011). More intriguing toxins are found in the venoms psammophiine snakes which were first noted in the species *Psammophis mossambica* (Fry, et al. 2008), where the propeptide domain was selectively expressed. Unlike the *Echis* forms, there was explosive diversification of these novel toxins: 26 variants were discovered in this species alone, including forms with novel cysteines which could potentially form disulphide bonds. Subsequent testing of two of these toxins revealed them

to be novel neurotoxins (Brust, et al. 2013). The activity of the other variants is unknown. In this study, this novel toxin class was shown to be present with staggering sequence diversity across the psammophiine snakes, including not only the additional *Psammophis* species we sequenced but also the *Malpolon* and *Rhamphiophis* species (Figures 4 and 5). Sequence analysis revealed that the first half of the toxins are homologous to the propeptide region of typical SVMP P-III genes. However, there is then an abrupt shift in sequence patterns, which is consistent with a frame-shift mutation providing the starting substrate for the evolution of this novel toxin class. The subsequent evolution resulted in such sequence diversity that calculating rates of was evolution impossible due to the unalignable diversity in the second half of the peptides. Such incredible diversity suggests there may be extensive neofunctionalization beyond the previously characterized neurotoxicity. Therefore, this toxin class represents a particularly rich area for future research, especially as most of these toxins are either short linear or with a single disulphide-bond, which would allow for efficient synthesis.

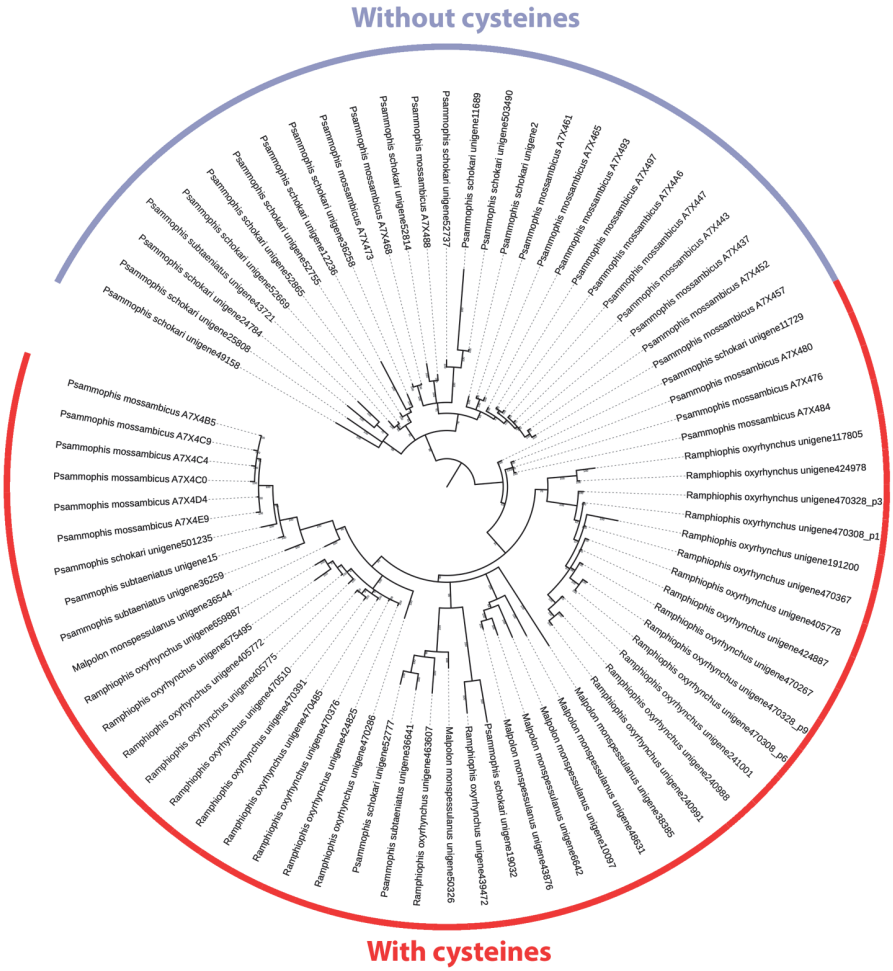


Figure 4: Molecular phylogenetic reconstruction of the psammophiine, lineage-specific derivation of the SVMP propeptide domain into a novel toxin family, with the subsequent explosive diversification of the cysteine-linked forms. Sequence alignment for constructing phylogenetic tree can be viewed in Supplementary File 9. For tree output file for SVMP propeptide domain toxins, see Supplementary File 10.

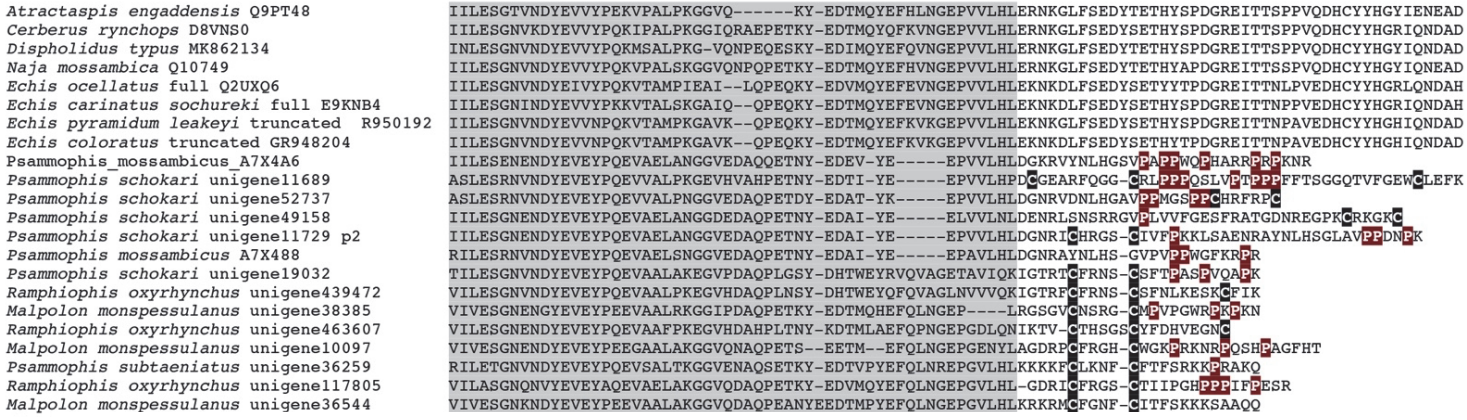


Figure 5: Alignment of representative SVMP propeptide domains. Region shaded in gray is that which is homologous across all representatives, with the unshaded region indicating the location of the putative frameshift mutation that led to the evolution of the new toxin family in psammophiine snakes. Cysteines and Prolines are indicated in black and red boxes.

References

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. Nucleic Acids Research 41:D36-D42.

Brust A, Sunagar K, Undheim EA, Vetter I, Yang DC, Casewell NR, Jackson TN, Koludarov I, Alewood PF, Hodgson WC. 2013. Differential evolution and neofunctionalization of snake venom metalloprotease domains. Molecular & Cellular Proteomics 12:651-663.

Campos PF, Andrade-Silva D, Zelanis A, Paes Leme AF, Rocha MMT, Menezes MC, Serrano SM, Junqueira-de-Azevedo LM. 2016. Trends in the evolution of snake toxins underscored by an integrative omics approach to profile the venom of the colubrid *Phalotris mertensi*. Genome Biology and Evolution 8:2266-2287.

Casewell NR, Harrison RA, Wüster W, Wagstaff SC. 2009. Comparative venom gland transcriptome surveys of the saw-scaled vipers (Viperidae: *Echi*s) reveal substantial intra-family gene diversity and novel venom transcripts. BMC Genomics 10:1-12.

Casewell NR, Sunagar K, Takacs Z, Calvete JJ, Jackson TNW, Fry BG. 2015. Snake venom metalloprotease enzymes. In: Venomous, Reptiles and Their Toxins. Evolution, Pathophysiology and Biodiscovery. New York: Oxford University Press. p. 347-363.

Casewell NR, Wagstaff SC, Harrison RA, Renjifo C, Wüster W. 2011. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. Molecular Biology and Evolution 28:2637-2649.

Ching AT, Paes Leme AF, Zelanis A, Rocha MM, Furtado MdFtD, Silva DbA, Trugilho MR, da Rocha SL, Perales J, Ho PL. 2012. Venomics profiling of *Thamnodynastes strigatus* unveils matrix metalloproteinases and other novel proteins recruited to the toxin arsenal of rear-fanged snakes. Journal of Proteome Research 11:1152-1162.

Debono J, Bos MH, Coimbra F, Ge L, Frank N, Kwok HF, Fry BG. 2019. Basal but divergent: Clinical implications of differential coagulotoxicity in a clade of Asian vipers. Toxicology in vitro 58:195-206.

Debono J, Dashevsky D, Nouwens A, Fry BG. 2020. The sweet side of venom: Glycosylated prothrombin activating metalloproteases from *Dispholidus typus* (boomslang) and *Thelotornis mossambicanus* (twig snake). Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 227:108625.

Debono J, Dobson J, Casewell NR, Romilio A, Li B, Kurniawan N, Mardon K, Weisbecker V, Nouwens A, Kwok HF. 2017. Coagulating colubrids: Evolutionary, pathophysiological and biodiscovery implications of venom variations between boomslang (*Dispholidus typus*) and twig snake (*Thelotornis mossambicanus*). Toxins 9:171.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Fox JW, Serrano SM. 2008. Exploring snake venom proteomes: multifaceted analyses for complex toxin mixtures. Proteomics 8:909-920.

Fry BG, Lumsden NG, Wüster W, Wickramaratna JC, Hodgson WC, Kini RM. 2003. Isolation of a neurotoxin (α-colubritoxin) from a nonvenomous colubrid: evidence for early origin of venom in snakes. Journal of Molecular Evolution 57:446-452.

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.

Gao R, Kini RM, Gopalakrishnakone P. 2002. A novel prothrombin activator from the venom of *Micropechis ikaheka*: isolation and characterization. Archives of Biochemistry and Biophysics 408:87-92.

Gutiérrez JM, Sanz L, Escolano J, Fernández J, Lomonte B, Angulo Y, Rucavado A, Warrell DA, Calvete JJ. 2008. Snake venomics of the Lesser Antillean pit vipers *Bothrops caribbaeus* and *Bothrops lanceolatus*: correlation with toxicological activities and immunoreactivity of a heterologous antivenom. Journal of Proteome Research 7:4396-4408.

Hofmann H, Bon C. 1987. Blood coagulation induced by the venom of *Bothrops atrox*. 2. Identification, purification, and properties of two factor X activators. Biochemistry 26:780-787.

Iddon D, Theakston R. 1986. Biological properties of the venom of the red-necked keel-back snake (*Rhabdophis subminiatus*). Annals of Tropical Medicine & Parasitology 80:339-344.

Jiang Y, Li Y, Lee W, Xu X, Zhang Y, Zhao R, Zhang Y, Wang W. 2011. Venom gland transcriptomes of two elapid snakes (*Bungarus multicinctus* and *Naja atra*) and evolution of toxin genes. BMC Genomics 12:1-13.

Kamiguti AS, Theakston RDG, Sherman N, Fox JW. 2000. Mass spectrophotometric evidence for P-III/P-IV metalloproteinases in the venom of the Boomslang (*Dispholidus typus*). Toxicon 38:1613-1620.

Kisiel W, Hermodson MA, Davie EW. 1976. Factor X activating enzyme from Russell's viper venom: isolation and characterization. Biochemistry 15:4901-4906.

Komori Y, Hifumi T, Yamamoto A, Sakai A, Ato M, Sawabe K, Nikai T. 2017. Comparative Study of Biological Activities of Venom from Colubrid Snakes *Rhabdophis tigrinus* (Yamakagashi) and Rhabdophis lateralis. Toxins 9:373.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276-3278.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

Modahl CM, Frietze S, Mackessy SP. 2018. Transcriptome-facilitated proteomic characterization of rear-fanged snake venoms reveal abundant metalloproteinases with enhanced activity. Journal of Proteomics 187:223-234.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Molecular Biology and Evolution 30:1196-1205.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics 8:e1002764.

Oulion B, Dobson JS, Zdenek CN, Arbuckle K, Lister C, Coimbra FC, Op den Brouw B, Debono J, Rogalski A, Violette A. 2018. Factor X activating *Atractaspis* snake venoms and the relative coagulotoxicity neutralising efficacy of African antivenoms. Toxicology letters 288:119-128.

Peichoto M, Teibler P, Mackessy S, Leiva L, Acosta O, Gonçalves L, Tanaka-Azevedo A, Santoro M. 2007. Purification and characterization of patagonfibrase, a metalloproteinase showing α-fibrinogenolytic and hemorrhagic activities, from *Philodryas patagoniensis* snake venom. Biochimica et Biophysica Acta -General Subjects 1770:810-819.

Petras D, Sanz L, Segura Á, Herrera M, Villalta M, Solano D, Vargas M, León G, Warrell DA, Theakston RDG. 2011. Snake venomics of African spitting cobras: toxin composition and assessment of congeneric cross-reactivity of the pan-African EchiTAb-Plus-ICP antivenom by antivenomics and neutralization approaches. Journal of Proteome Research 10:1266-1280.

Pond SLK, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. In. Statistical methods in molecular evolution: Springer. p. 125-181.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539-542.

Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlić A, Quesada M, Quinn GB, Westbrook JD. 2010. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Research 39:D392-D401.

Seneci L, Zdenek CN, Chowdhury A, Rodrigues CF, Neri-Castro E, Bénard-Valle M, Alagón A, Fry BG. 2021. A clot twist: extreme variation in coagulotoxicity mechanisms in mexican neotropical rattlesnake venoms. Frontiers in Immunology 12:552.

Siigur E, Aaspõllu A, Trummal K, Tõnismägi K, Tammiste I, Kalkkinen N, Siigur J. 2004. Factor X activator from *Vipera lebetina* venom is synthesized from different genes. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1702:41-51.

Takeya H, Nishida S, Miyata T, Kawada S, Saisaka Y, Morita T, Iwanaga S. 1992. Coagulation factor X activating enzyme from Russell's viper venom (RVV-X). A novel metalloproteinase with disintegrin (platelet aggregation inhibitor)-like and C-type lectin-like domains. Journal of Biological Chemistry 267:14109-14117.

Wagstaff SC, Sanz L, Juarez P, Harrison RA, Calvete JJ. 2009. Combined snake venomics and venom gland transcriptomic analysis of the ocellated carpet viper, *Echis ocellatus*. Journal of Proteomics 71:609-623.

Chapter 6. Conclusions

The findings of this study represent a significant advance in our knowledge of the broad-scale molecular evolution of snake toxin families. We have revealed novel patterns of expression of basal toxin types, including previously unrecognized instances of molecular and structural convergence. The new toxin encoding sequences from RFS and FFS that we have included in our analyses proved particularly valuable demonstration of the distribution of novel toxin classes derived from the propeptide domains of pre-existing toxin genes. These results provide a framework to help guide future bioactivity testing work and further evolutionary studies. Research into the evolutionary and selective forces that result in the instances of explosive diversification or molecular convergence will provide crucial insights into how venom evolves.

One interesting question raised by these data is why, despite all these toxin families being present in the last common ancestor of the advanced snakes, particular descendant lineages have specialized in the production and refinement of certain toxin families and isoforms. For example, kunitz peptides are fairly common in viperid venoms and have diversified to an extreme degree in the elapid family, but are largely absent in other lineages. It is unclear whether these lineage-specific differences are the product of chance or if they were constrained by the ecological contexts in which the progenitors of these snake families employed their venoms.

Other toxin families—such as the SVMPs—show broadly similar levels of duplication across multiple families. This mirrors the pattern of neofunctionalization where the P-III SVMPs have repeatedly evolved into potent procoagulant factor activating toxins. However, only in the viperids has the P-IIId multimeric form evolved, where a SVMP is covalently linked to a lectin dimer (which already possesses another interchain disulphide bond). The viperids show by far the greatest diversity of lectin dimers which may have provided a greater range of molecular opportunities for the P-IIId trait to evolve.

From a broad perspective, almost all toxin families we examined demonstrate a phylogenetic pattern of large clades belonging to snakes of the same family. This suggests that these toxins had not yet diversified in the common ancestor of Colubroidea. The only exception was the lectins, which suggests that this common ancestor likely possessed multiple copies of this toxin already including dimeric forms. Duplication of toxin genes in snakes is often associated with higher abundance of that toxin family in the final venom composition (Margres, et al. 2017; Jackson and Koludarov 2020), so this may constitute preliminary evidence that the common ancestor of Colubroidea would have possessed a primarily lectin-based venom to accompany the other innovations in the venom system such as partitioned oral glands and perhaps modified dentition. The pattern we see in the other families indicates that the vast majority of the variation in terms of composition, unconventional structures, and novel functions that we observe in extant snake venoms arose after the divergence between the families and during the diversification and specialization of those lineages.

While we have discussed many toxins which have rapidly diversified, this phenomenon is most extreme in the propeptide regions of the natriuretic and SVMP genes. Typically, the propeptide region is post-translationally cleaved and does not play a role in envenomation. However, in both these families, new

mutations have caused part of the propeptide region to be translated into protein to form entirely new toxins with novel functions. For the natriuretic peptides, the newly evolved toxins include repeating series of bradykinin potentiating peptides which increase the hypotensive effect of the venoms (Fry, Jackson, et al. 2015). The viperid genera *Azemiops* and *Tropidolaemus* have separately evolved novel neurotoxins derived from within the natriuretic propeptide domain, which share the unusual feature of creating multiple peptides that are translated from a single transcript and then separated during post-translational modification despite their independent origins. Another peptide type was first documented in *Dendroaspis* venom and the molecular evolutionary history remained enigmatic, but our analyses indicate these are members of yet another novel toxin class arising from the natriuretic gene propeptide region. The most explosive diversification of all toxin classes was that of the newly evolved toxin family that evolved in the SVMP propeptide domain within psammophiine snakes. The staggering sequence and structural diversity of these toxins makes it likely that other toxic activities in addition to the already documented novel neurotoxic forms (Brust, et al. 2013) will be documented as more bioactivity testing is undertaken.

Our analyses show that these shared colubroid toxin families exhibit remarkable instances of convergent evolution in terms of pathophysiological function and protein structure. For example, within two potently procoagulant lineages (the *Daboia* genus within the viperid snakes and the *Oxyuranus/Pseudonaja* clade within the elapid snakes), plasmin inhibiting kunitz peptides have evolved which would potentiate the procoagulant effects by increasing the half-life of the blood clots formed due to the inhibition of the blood clot destroying enzyme plasmin. Other taxa possess the arginine residue that is crucial for these plasmin inhibitors and may represent further instances of convergence if functional research confirms this hypothesized activity. Similarly, within the SVMP neofunctionalizationed procoagulant variants, which activate Factor X or prothrombin, have arisen on multiple independent occasions. The lectins may potentially contain further examples of functional convergence given the multiple origins of the QPD motif at a key functional location, but these have not been tested.

One of the most striking cases of structural convergence is the previously mentioned P-IIId derived form of the SVMP which form a covalent linkage to a lectin dimer. The novel cysteine crucial for the formation of these toxin complexes was shown to have evolved on three separate occasions within the viperids as structural modifications of forms that were themselves functionally derived (procoagulant). The selection pressures leading to this convergence have not been explored and the functional impacts are similarly uncharacterised.

The kunitz peptides have been the substrate for both levels of convergence. Structurally three out of the four neurotoxin types (MitTx, taicatoxin, and bungarotoxins) converge in their formation of heteromers with PLA_2 subunits, but diverge structurally in this regard by being non-covalently linked (MitTx and taicatoxin) or covalently linked (β-bungarotoxin), and also in the number of PLA_2 subunits associated with (MitTx = 2, taicatoxin = 1, β-bungarotoxin = 1). All four of the neurotoxic kunitz peptides converge in being ion-channel toxins, with dendrotoxins and β-bungarotoxins further converging on the same target (K_V channels). While taicatoxins affect a different ion channel type (L-type calcium channels) than dendrotoxins and β-bungarotoxins, they converge with bungarotoxin in the PLA_2 toxin facilitating a

secondary action that results in the net functional outcome in blocking the release acetylcholine, leading to flaccid paralysis. In contrast, dendrotoxins act upon the voltage-gated potassium channels to facilitate acetylcholine release, leading to spastic paralysis. β-bungarotoxin demonstrates another instance of convergent molecular evolution in the novel cysteines which allow for the formation of these complexes have evolved twice in the exact same location. The convergent evolution of novel cysteines at the same residue on two occasions within both β-bungarotoxin and the 3FTx dimers is strongly indicative of structural constraints in the formation of these dimers.

This study gives a broad overview of the diversity in the toxin families which are homologous in Colubroid venoms. It is this diversity that produces the wide range of clinical effects and variable responses to antivenom that contribute to the global problem of snakebite. However, such molecular diversity also provides fertile ground in the search for novel molecules as lead compounds for the discovery of new tools and medications. This diversity has also allowed these toxins to converge repeatedly on similar sequences, structures, and functions. This widespread convergence suggests that certain pathophysiological activities and certain configurations of proteins may be evolutionary 'good tricks' (Dennett and Dennett 1996) that are similarly effective across multiple taxa and may solve evolutionary problems that venoms encounter such as prey resistance.

References

Brust A, Sunagar K, Undheim EA, Vetter I, Yang DC, Casewell NR, Jackson TN, Koludarov I, Alewood PF, Hodgson WC. 2013. Differential evolution and neofunctionalization of snake venom metalloprotease domains. Molecular & Cellular Proteomics 12:651-663.

Dennett DC, Dennett DC. 1996. Darwin's Dangerous Idea: Evolution and the Meanins of Life: Simon and Schuster.

Fry BG, Jackson TN, Takacs Z, Reeks T, Sunagar K. 2015. C-type natriuretic peptides. In: Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery. New York: Oxford University Press. p. 318-326.

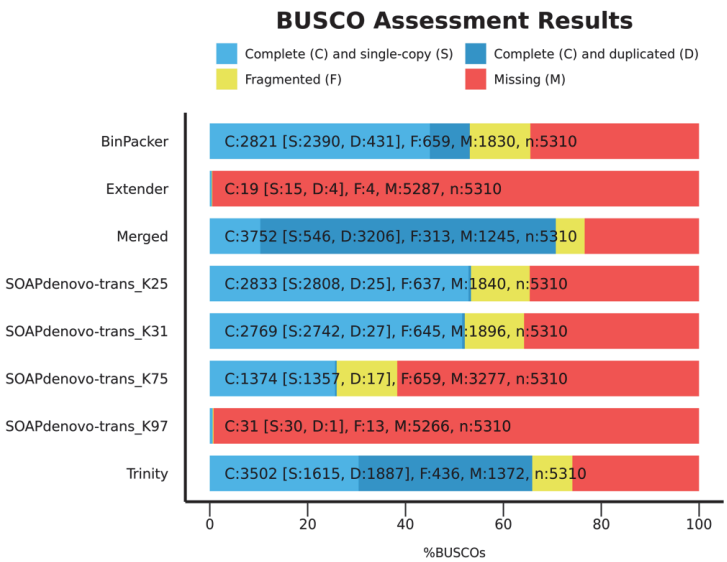
Jackson TN, Koludarov I. 2020. How the toxin got its toxicity. Frontiers in Pharmacology 11:1893.

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 206:1569-1580.

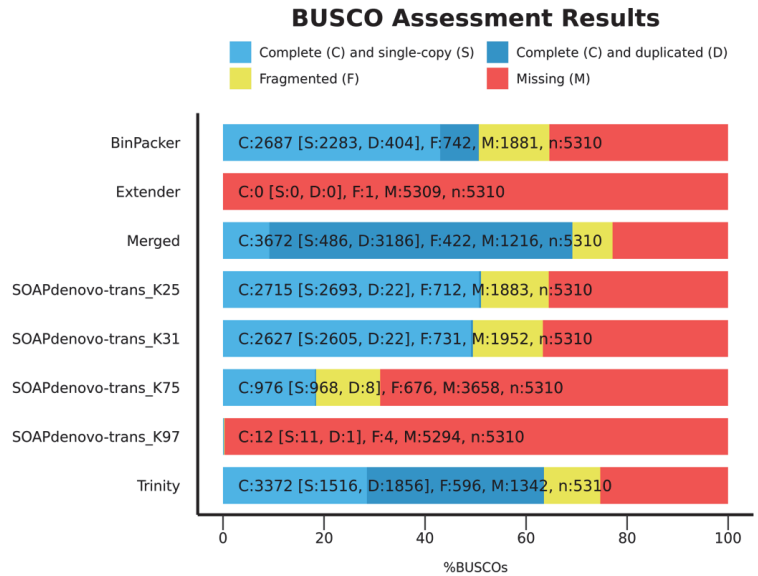
Chapter 7. Supplementary Materials

This chapter is published as Supplementary Material 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.

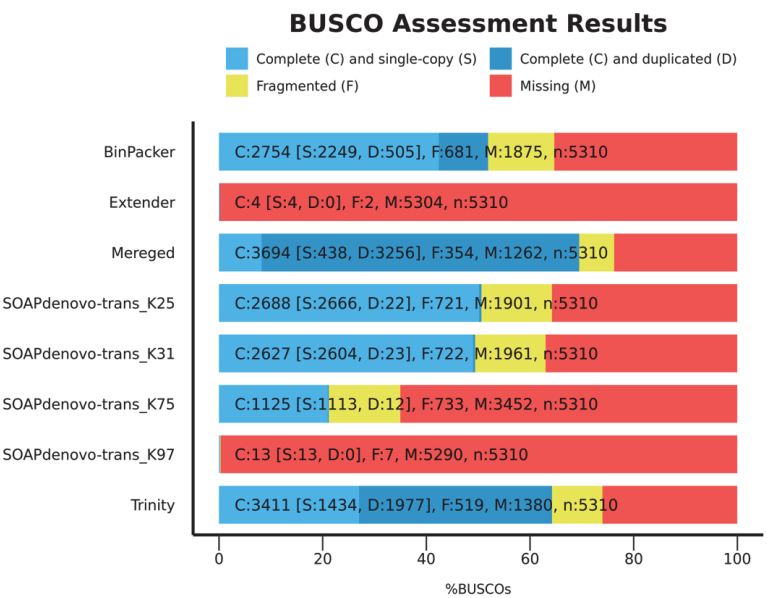
https://doi.org/10.1186/s12915-021-01208-9



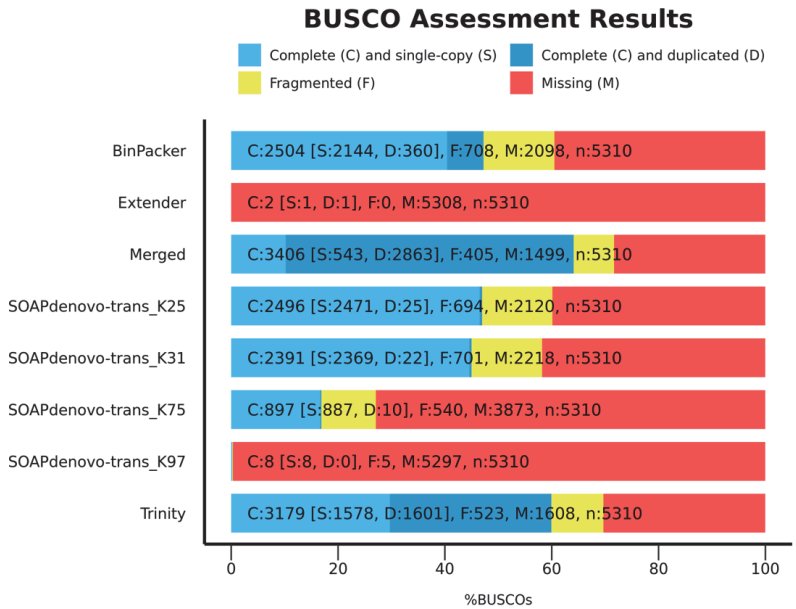
Supplementary Figure 1A: BUSCO completeness analyses of venom gland transcriptome assemblies for *Helicops leopardinus*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



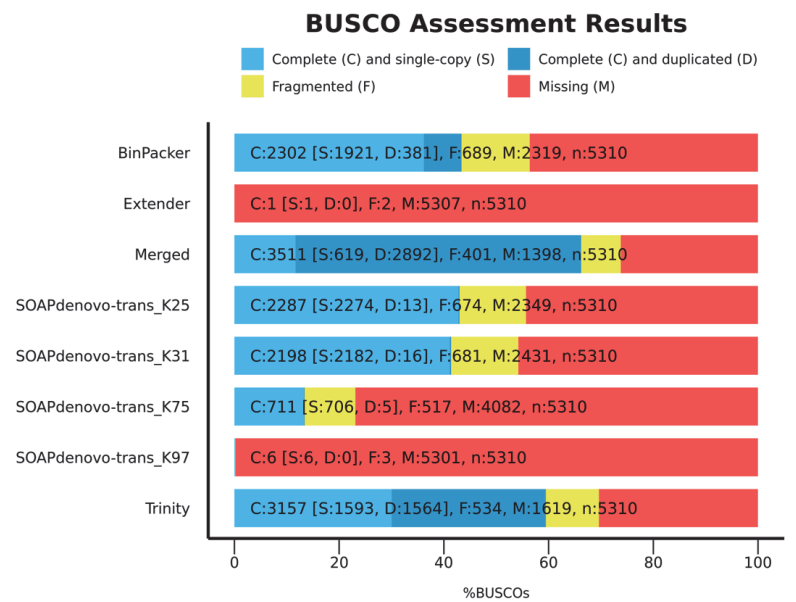
Supplementary Figure 1B: BUSCO completeness analyses of venom gland transcriptome assemblies for *Rhabdophis subminiatus*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



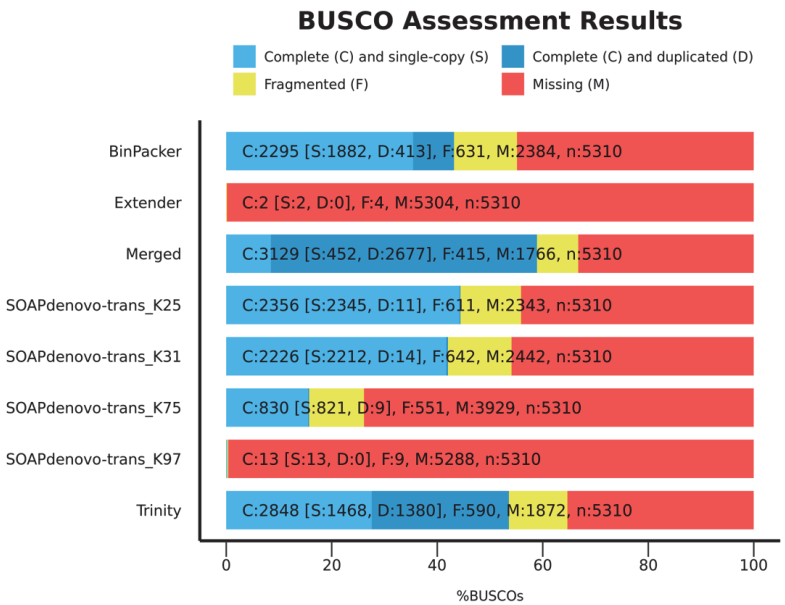
Supplementary Figure 1C: BUSCO completeness analyses of venom gland transcriptome assemblies for *Heterodon nasicus*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



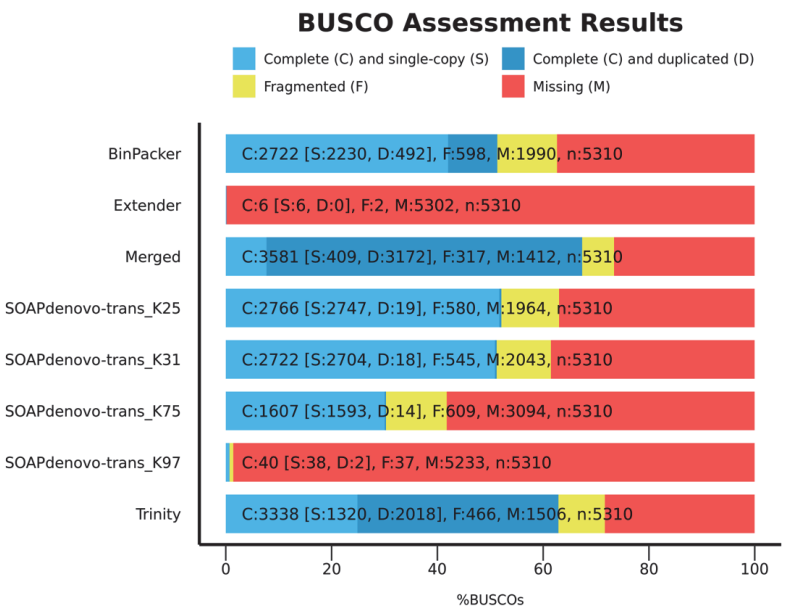
Supplementary Figure 1D: BUSCO completeness analyses of venom gland transcriptome assemblies for *Malpolon monspessulanus*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



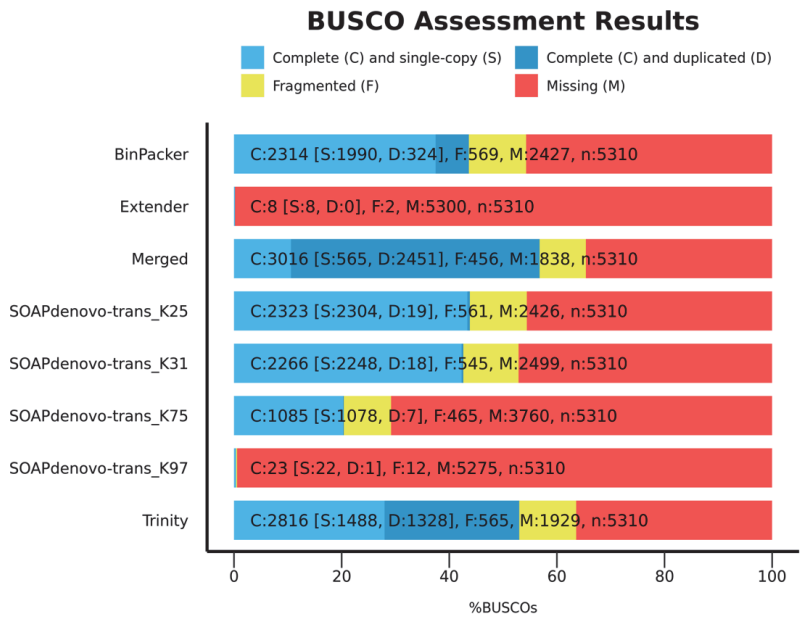
Supplementary Figure 1E: BUSCO completeness analyses of venom gland transcriptome assemblies for *Psammophis schokari*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



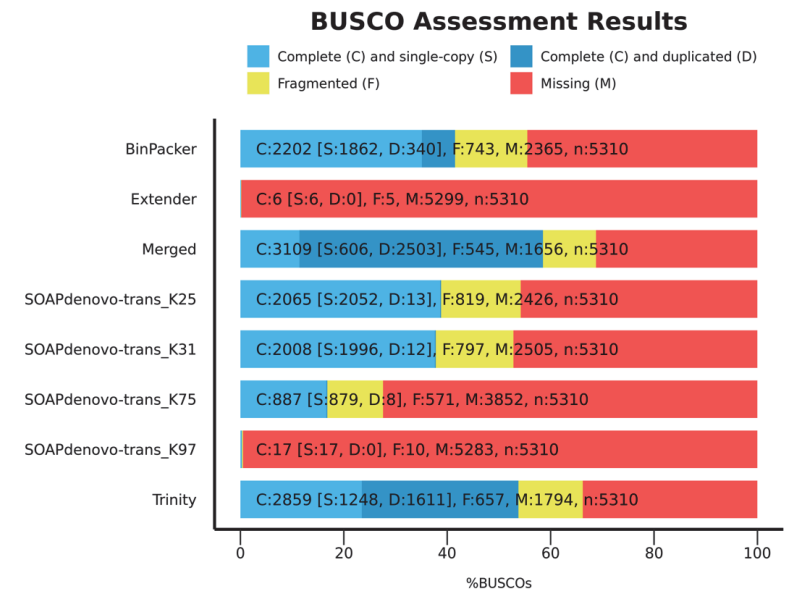
Supplementary Figure 1F: BUSCO completeness analyses of venom gland transcriptome assemblies for *Psammophis subtaeniatus*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



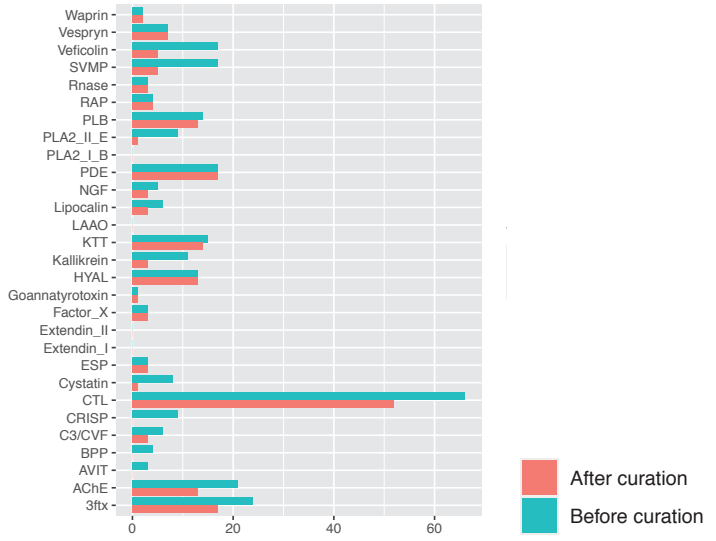
Supplementary Figure 1G: BUSCO completeness analyses of venom gland transcriptome assemblies for *Homalopsis buccata*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



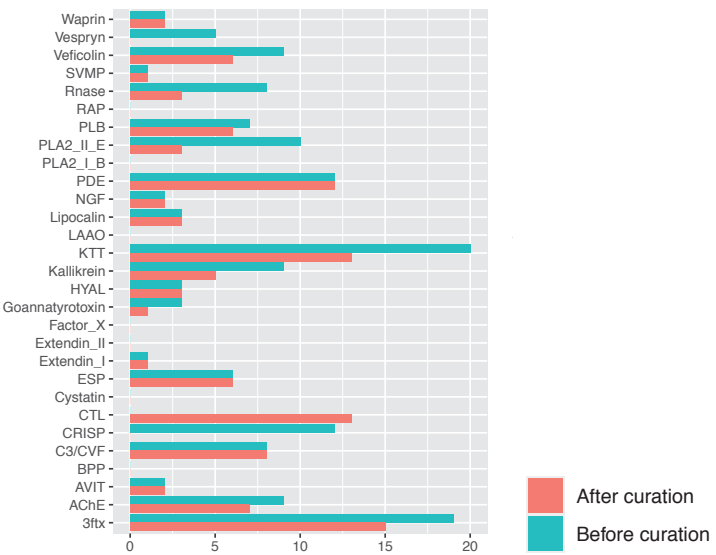
Supplementary Figure 1H: BUSCO completeness analyses of venom gland transcriptome assemblies for *Pseudocerastes urarachnoides*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



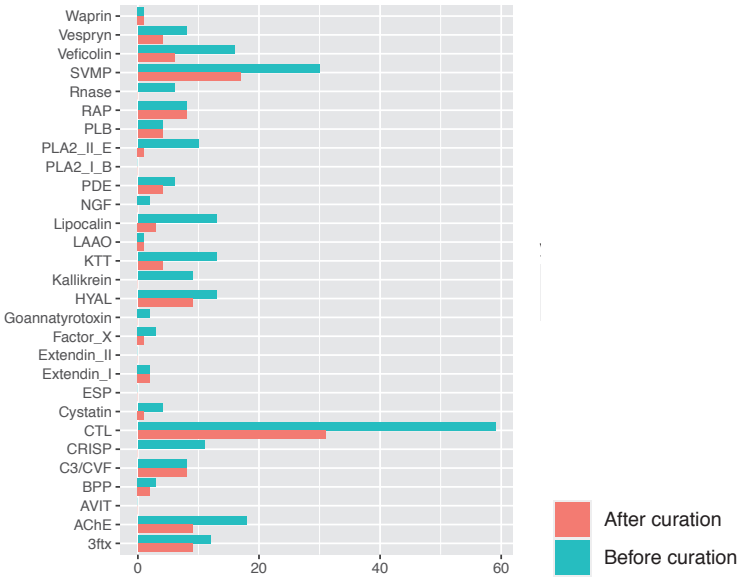
Supplementary Figure 1I: BUSCO completeness analyses of venom gland transcriptome assemblies for *Vipera transcaucasiana*. Snake contigs were matched against 5310 orthologous loci defined within Tetrapoda.



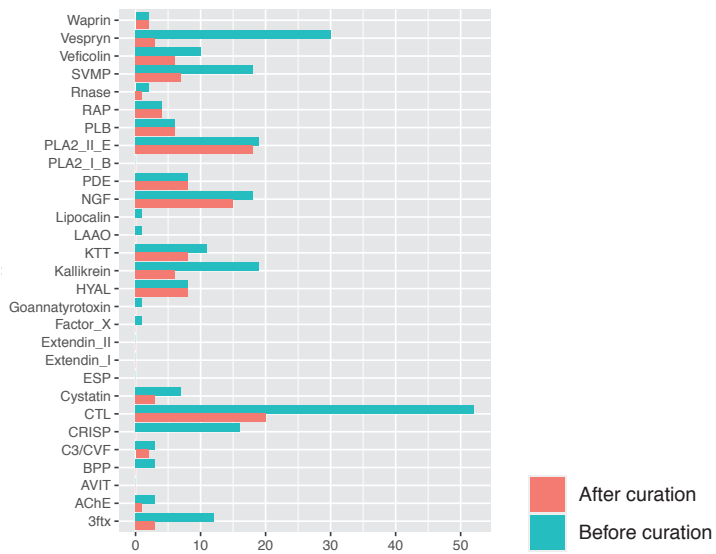
Supplementary Figure 2A: The statistics of toxins from *Helicops leopardinus* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



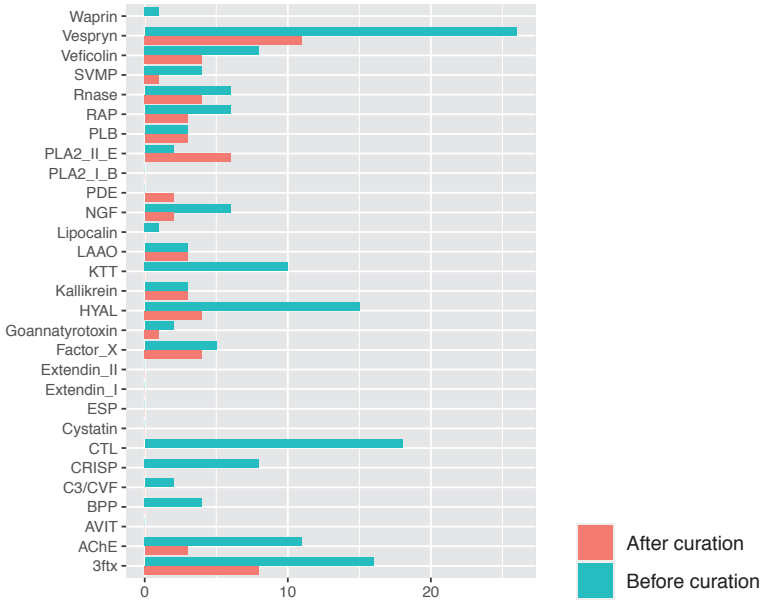
Supplementary Figure 2B: The statistics of toxins from *Rhabdophis subminiatus* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



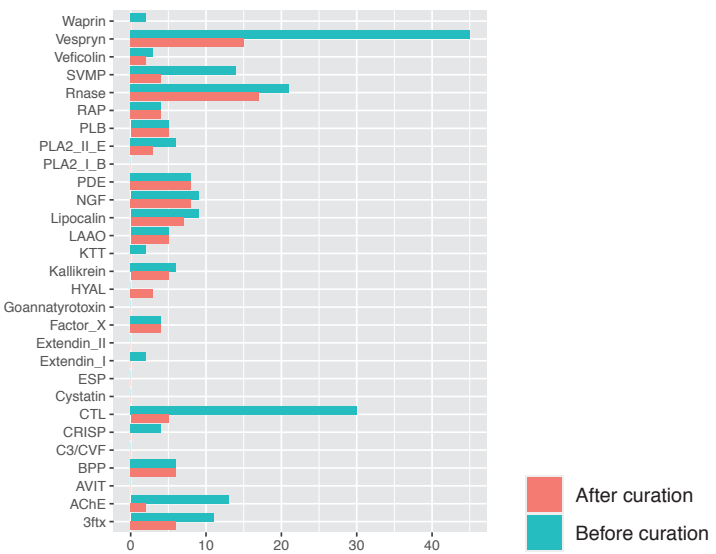
Supplementary Figure 2C: The statistics of toxins from *Heterodon nasicus* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



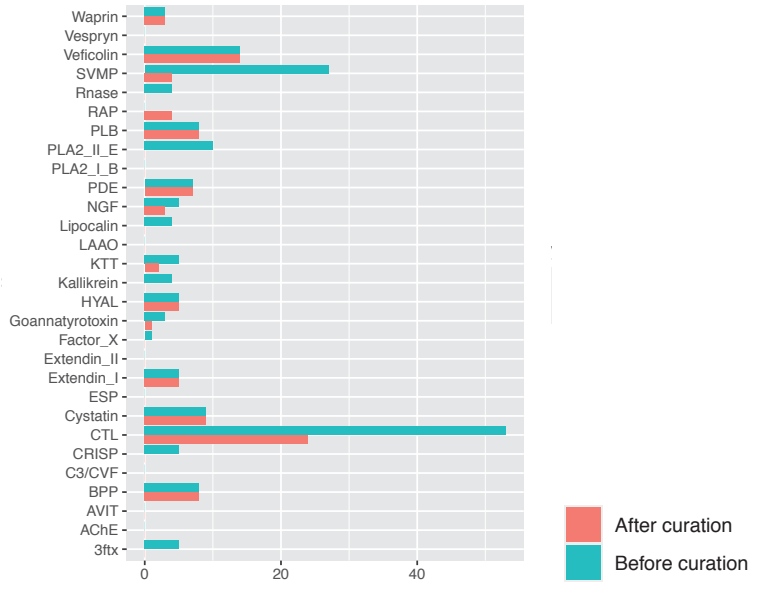
Supplementary Figure 2D: The statistics of toxins from *Malpolon monspessulanus* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



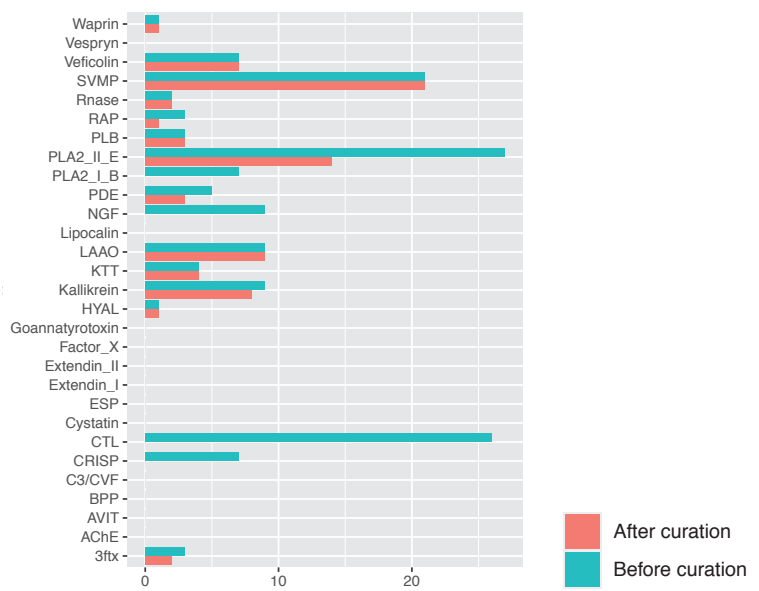
Supplementary Figure 2E: The statistics of toxins from *Psammophis schokari* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



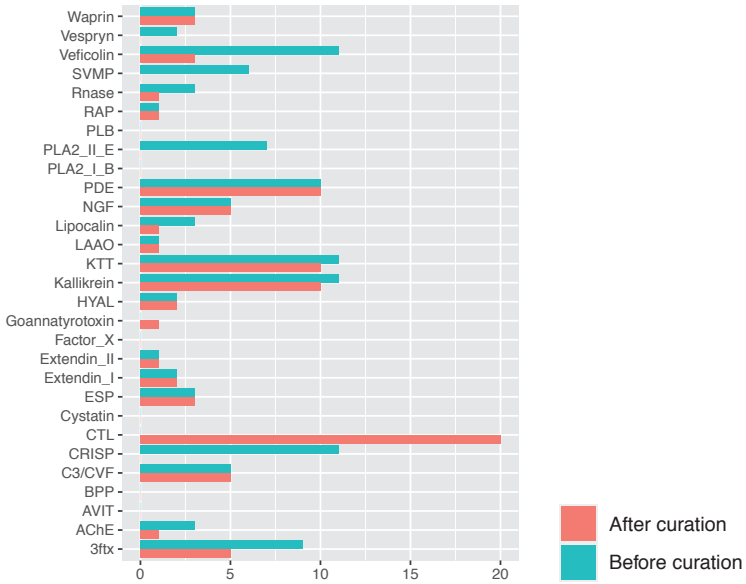
Supplementary Figure 2F: The statistics of toxins from *Psammophis subtaeniatus* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



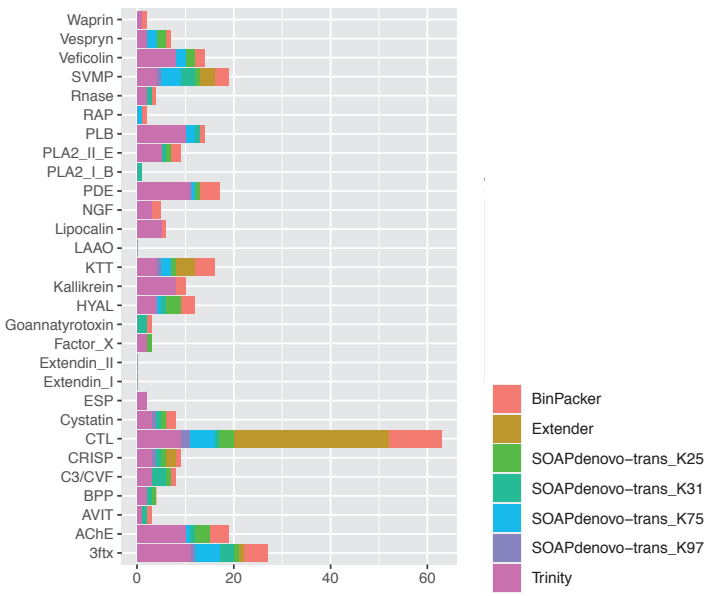
Supplementary Figure 2G: The statistics of toxins from *Homalopsos buccata* before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



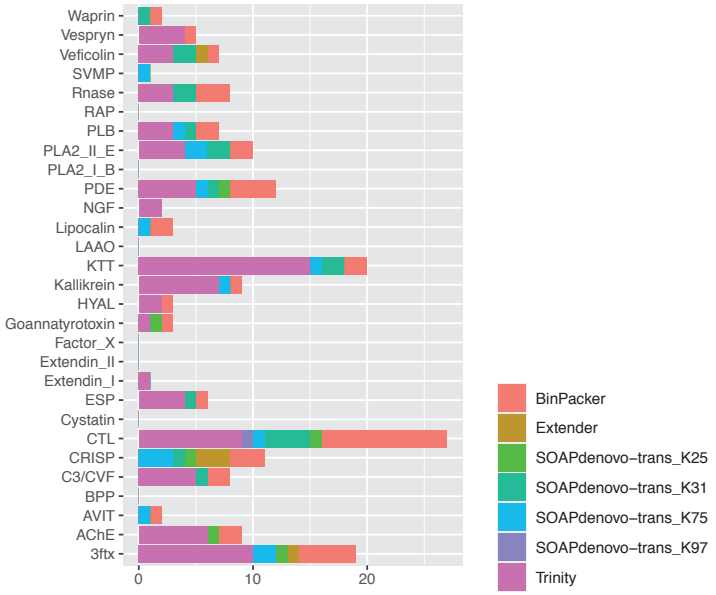
Supplementary Figure 2H: The statistics of toxins before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



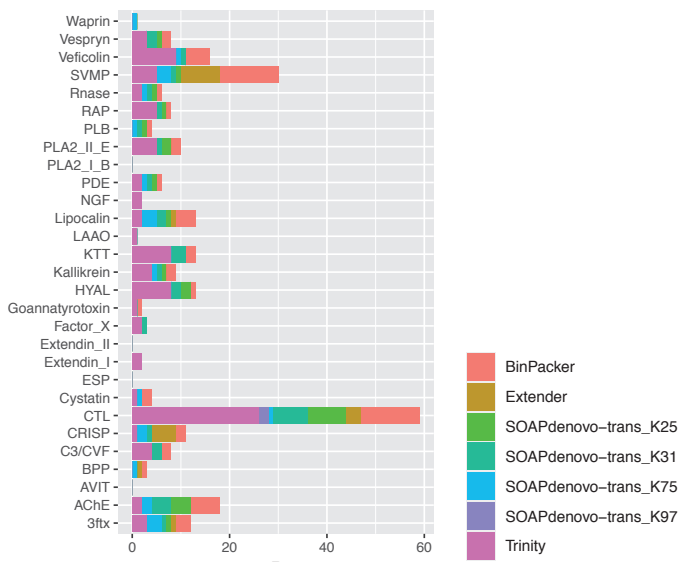
Supplementary Figure 2I: The statistics of toxins before and after curation. As can be seen, curation leads to a huge decline in the number of both toxin families diversities and toxin transcripts. Some toxin families are discarded as a whole. X axis is toxin family name and Y axis is the number of the corresponding toxin transcripts.



Supplementary Figure 3A: Toxin distribution of *Helicops leopardinus* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

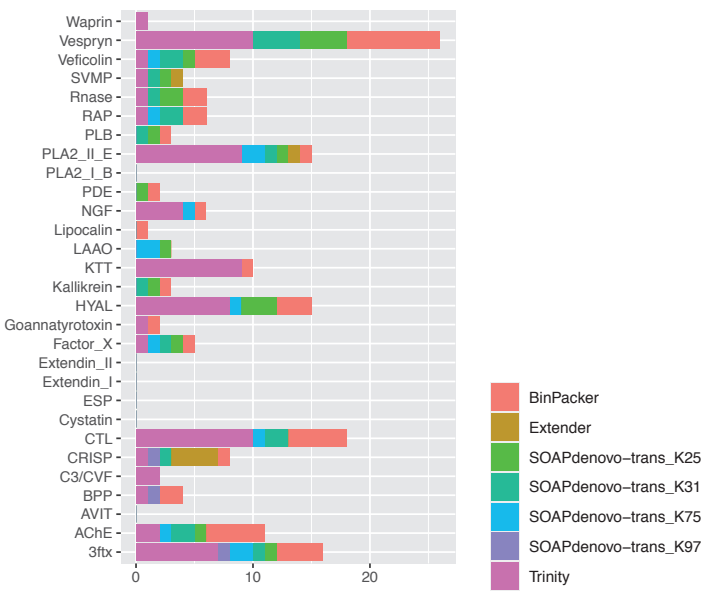


Supplementary Figure 3B: Toxin distribution of *Rhabdophis subminiatus* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

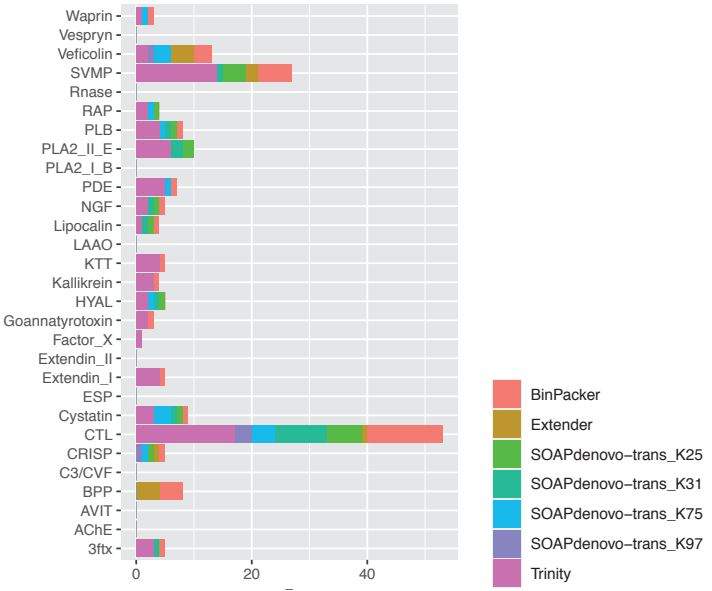


Supplementary Figure 3C: Toxin distribution of *Heterodon nasicus* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

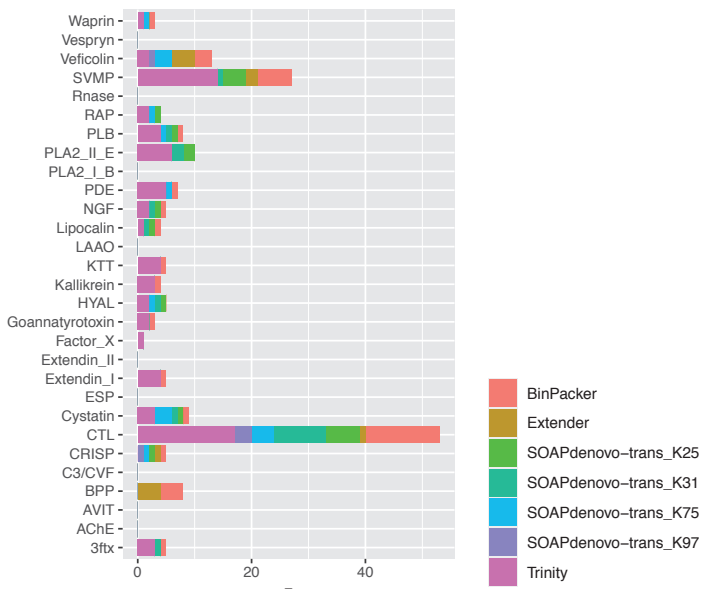
Supplementary Figure 3D: Toxin distribution of *Malpolon monspessulanus* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



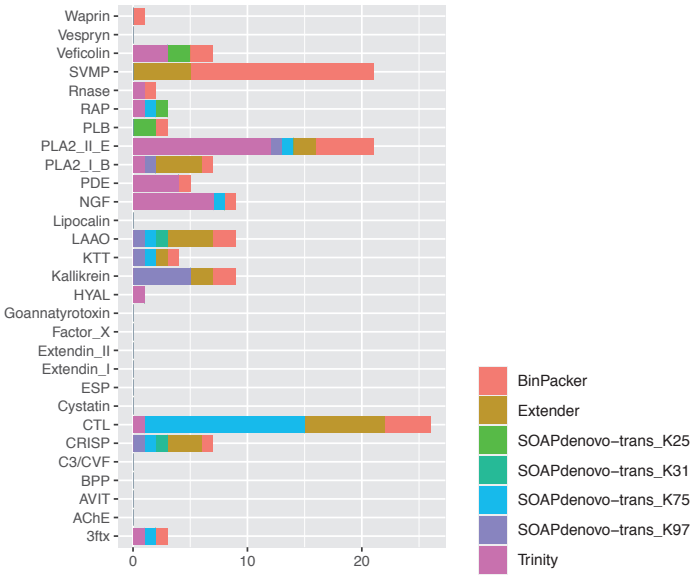
Supplementary Figure 3E: Toxin distribution of *Psammophis schokari* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



Supplementary Figure 3F: Toxin distribution of *Psammophis subtaeniatus* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

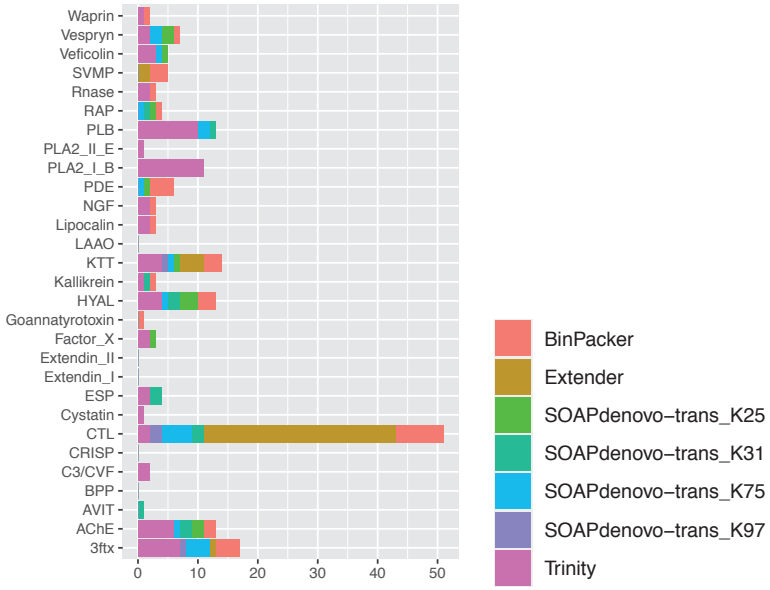


Supplementary Figure 3G: Toxin distribution of *Homalopsis buccata* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

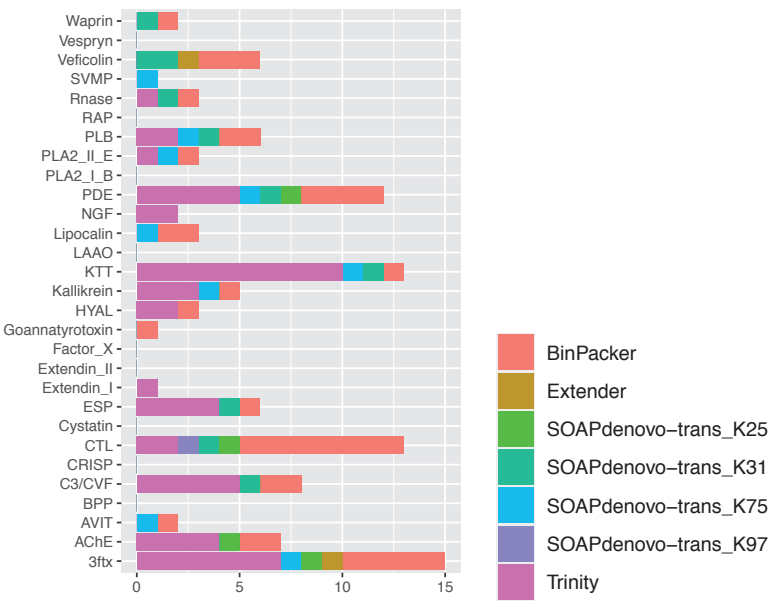


Supplementary Figure 3H: Toxin distribution of *Pseudocerastes urarachnoides* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

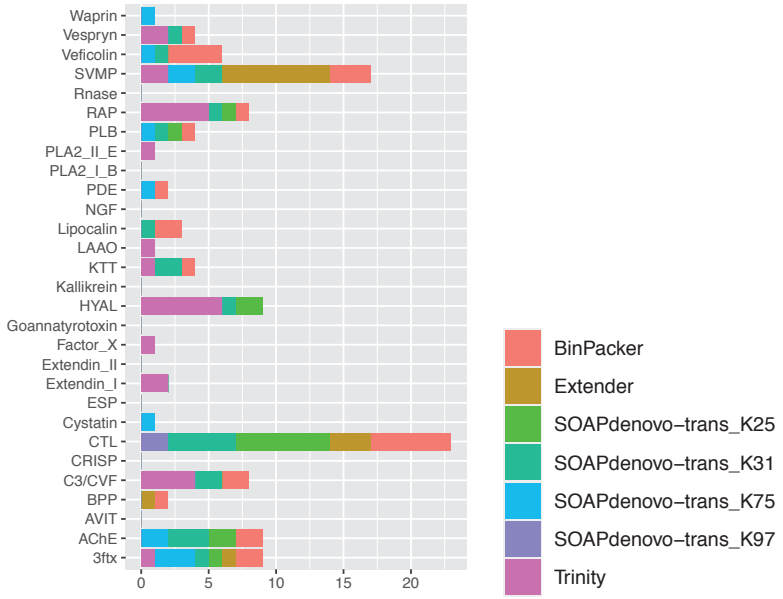
Supplementary Figure 3I: Toxin distribution of *Vipera transcaucasiana* before curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



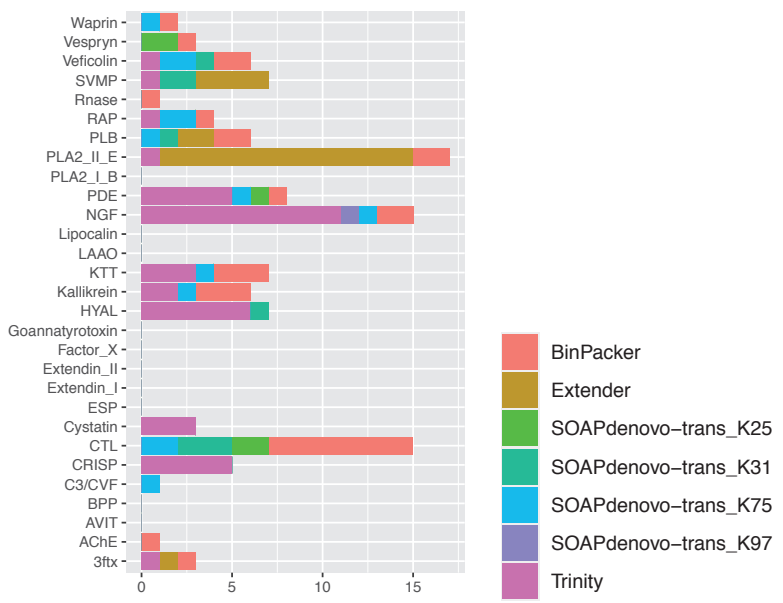
Supplementary Figure 4A: Toxin distribution of *Helicops leopardinus* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



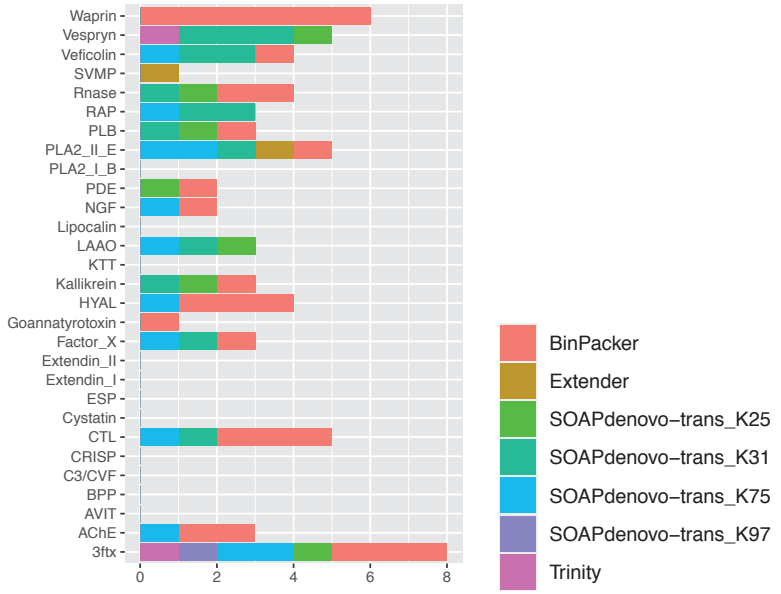
Supplementary Figure 4B: Toxin distribution of *Rhabdophis subminiatus* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



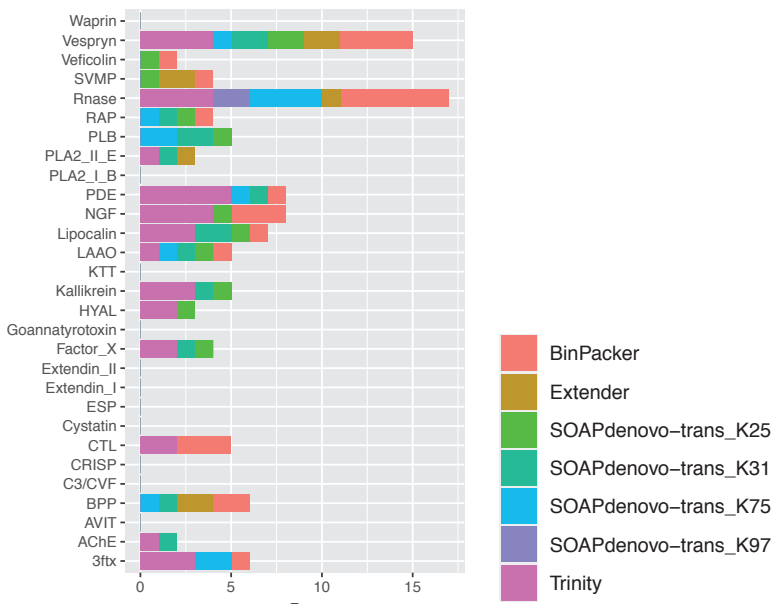
Supplementary Figure 4C: Toxin distribution of *Heterodon nasicus* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



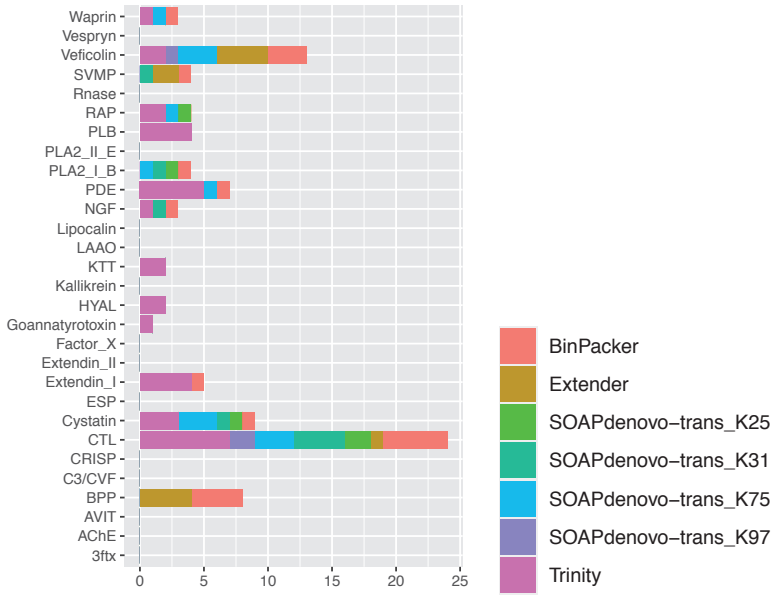
Supplementary Figure 4D: Toxin distribution of *Malpolon monspessulanus* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



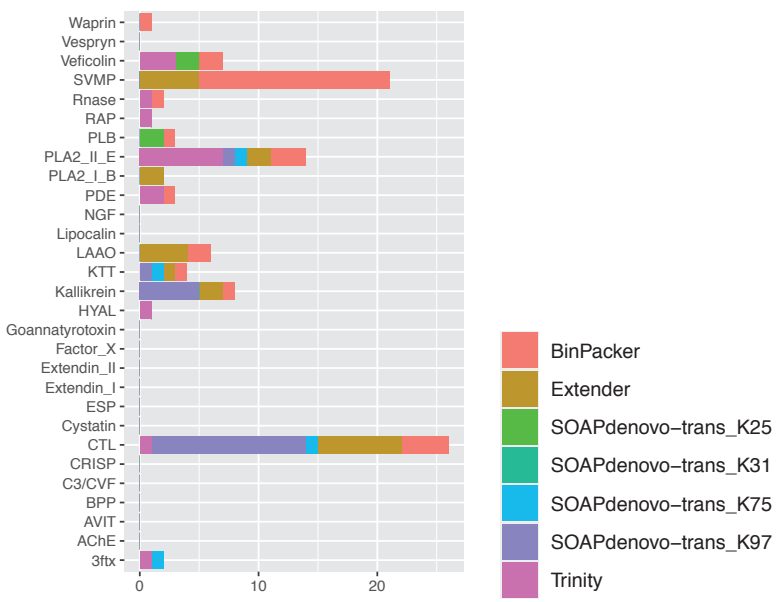
Supplementary Figure 4E: Toxin distribution of *Psammophis schokari* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



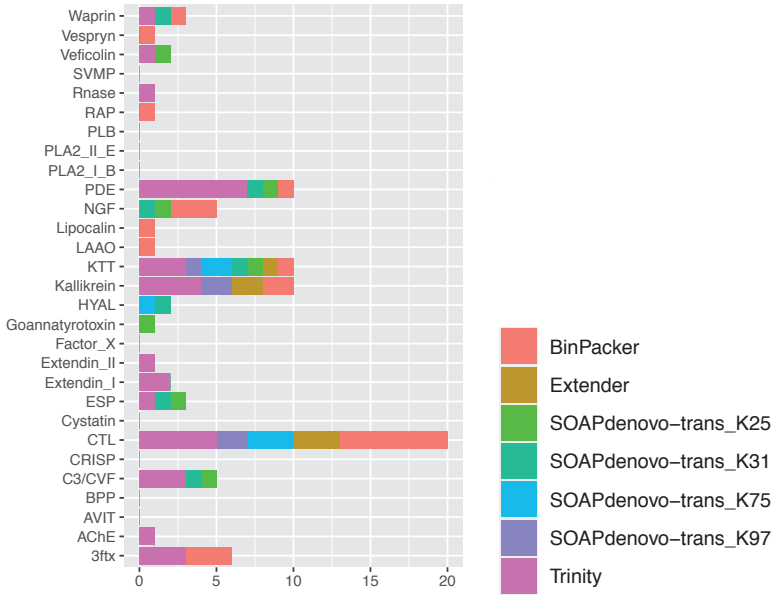
Supplementary Figure 4F: Toxin distribution of *Psammophis subtarniatus* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



Supplementary Figure 4G: Toxin distribution of *Homalopsis buccata* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



Supplementary Figure 4H: Toxin distribution of *Pseudocerastes urarachnoides* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.



Supplementary Figure 4I: Toxin distribution of *Vipera transcaucasiana* after curation. For each toxin family, the numbers of toxin transcripts recovered by different assembly methods are indicated by different colours.

Supplementary File 2: The settings for MrBayes

begin mrbayes;

log start replace;

set autoclose = no nowarn=no;

lset applyto = (all) nst = 6 rates = invgamma;

prset applyto = (all) aamodelpr = mixed;

unlink revmat = (all) shape = (all) pinvar = (all) statefreq = (all) tratio = (all);

showmodel;

mcmc ngen = 15000000 printfreq = 1000 samplefreq = 100 nchains = 4 temp = 0.2 checkfreq = 50000 diagnfreq = 1000 stopval = 0.01 stoprule = yes;

sumt relburnin = yes burninfrac = 0.25 contype = halfcompat;

sump relburnin = yes burninfrac =0.25;

outgroup 1;

log stop;

end;

For the following **Supplementary Files**, they can be viewed via

https://figshare.com/articles/dataset/Supplementary_Files/15085353

Supplementary_File_1_inhouse_toxin_database.fasta

Supplementary_File_3_kunitz_alignment.fasta

Supplementary_File_4_kunitz_alignment.con.tre

Supplementary_File_5_lectin_alignment.fasta

Supplementary_File_6_lectin_tree.con.tre

Supplementary_File_7_SVMP_alignment.fasta

Supplementary_File_8_SVMP_tre.con.tre

Supplementary_File_9_SVMP_propeptide_alignment.fasta

Supplementary_File_10_SVMP_propeptide_tree.con.tre

Nederlandse samenvatting

De explosieve straling en diversificatie van de geavanceerde slangen (superfamilie Colubroidea) werd geassocieerd met veranderingen in alle aspecten van het gedeelde gifsysteem. Morfologische veranderingen omvatten de verdeling van de gemengde voorouderlijke klieren in twee afzonderlijke klieren die respectievelijk bestemd zijn voor de productie van gif of slijm, evenals veranderingen in de locatie, grootte en structurele elementen van de gif-leverende tanden. Er is ook bewijs voor homologie tussen gifkliertoxinen die tot expressie worden gebracht in de geavanceerde slangen. Ondanks de evolutionaire nieuwigheid van slangengif, zijn diepgaande reconstructies van de moleculaire evolutionaire geschiedenis van toxines echter meestal beperkt tot die typen die aanwezig zijn in slechts twee slangenfamilies met voortanden, Elapidae en Viperidae. Om een breder begrip te krijgen van toxines die worden gedeeld door bestaande slangen, hebben we hier eerst de transcriptomen van acht taxonomisch diverse soorten met achtertanden en vier belangrijke adderachtige soorten geanalyseerd en de belangrijkste toxinetypen geanalyseerd die door de geavanceerde slangen worden gedeeld.

Transcriptomen werden geconstrueerd voor de volgende families en soorten: Colubridae - *Helicops leopardinus*, *Heterodon nasicus*, *Rhabdophis subminiatus*; Homalopsidae - *Homalopsis buccata*; Lamprophiidae - *Malpolon monspessulanus*, *Psammophis schokari*, *Psammophis subtaeniatus*, *Rhamphiophis oxyrhynchus*; en Viperidae - *Bitis atropos*, *Pseudocerastes urarachnoides*, *Tropidolaeumus subannulatus*, *Vipera transcaucasiana*. Deze sequenties werden gecombineerd met die uit beschikbare databases van andere soorten om een robuuste reconstructie mogelijk te maken van de moleculaire evolutionaire geschiedenis van de belangrijkste toxineklassen die aanwezig zijn in het gif van de laatste gemeenschappelijke voorouder van de geavanceerde slangen, en dus aanwezig zijn in de volledige diversiteit van colubroid slangengif. Naast differentiële evolutiesnelheden in toxineklassen tussen de slangenlijnen, onthulden deze analyses meerdere gevallen van voorheen onbekende gevallen van structurele en functionele convergenties. Structurele convergenties omvatten: 1. de evolutie van nieuwe cysteïnes om heteromere complexen te vormen, zoals binnen kunitz-peptiden (de bèta-bungarotoxine-eigenschap die zich bij minstens twee gelegenheden ontwikkelt) en binnen SVMP-enzymen (de P-IIId-eigenschap die zich bij minstens drie gelegenheden ontwikkelt); 2. en de C-terminale staart die evolueert bij twee verschillende gelegenheden binnen de C-type natriuretische peptiden, om structurele en functionele analogen van de ANP/BNP-staartaandoening te creëren. Er werd ook aangetoond dat de de novo-evolutie van nieuwe post-translationeel vrijgemaakte toxinefamilies binnen het propeptide-gebied van het natriuretisch peptidegen minstens vijf keer plaatsvond, met nieuwe functies variërend van inductie van hypotensie tot postsynaptische neurotoxiciteit. Functionele convergenties omvatten het volgende: meerdere gevallen van SVMP neofunctionaliseerde in procoagulant-gif in activatoren van de stollingsfactoren protrombine en Factor X; meerdere gevallen in procoagulant-giffen waar kunitz-peptiden neofunctionaliseerden tot remmers van het stolselvernietigende enzym plasmine, waardoor de halfwaardetijd van de stolsels, gevormd door de stollingsactiverende enzymatische toxinen, werd verlengd; en meerdere keren dat kunitz-peptiden

neofunctionaliseerden tot neurotoxinen die inwerken op presynaptische doelen, waaronder tweemaal binnen *Bungarus*-gif.

We vonden nieuwe convergenties in zowel structurele als functionele evolutie van slangentoxines. Deze resultaten bieden een gedetailleerde routekaart voor toekomstig werk om evolutionaire wapenwedlopen tussen roofdieren en prooien op te helderen, differentiële klinische pathologieën vast te stellen en rijke bronnen voor bio-ontdekking voor loodverbindingen in de pijplijn voor het ontwerpen en ontdekken van geneesmiddelen te documenteren.

Curriculum vitae

Bing Xie was born in Zaozhuang, Shandong Province, China. After completing his secondary education in Zaozhuang, he went to the city of Tai`an to attend Shandong Agriculture University for a four-year bachelor program. He gained two majors there: Agronomy and Computer Science. Since then, he became increasingly interested in the interdisciplinary field of bioinformatics. Then, he moved to Chongqing to pursue his master study at Southwest University in bioinformatics. After two years of study, he was employed by the Beijing Genomics Institute (BGI) as a bioinformatician. During his stay at the BGI, he was involved in the research of FishT1K and marine venom with multi 'omics' methods, including genomics, transcriptomics and proteomics. After BGI, he went to pursue his PhD studies in Leiden. He was supervised by Prof. M. K. Richardson in Leiden University and Dr. B. G. Fry in Queensland University. Both promotors are experts in the field of venomics (venom studies). In his PhD work, he aimed at studying the evolution of toxins, and with a longer-term objective of developing animal venom toxins into novel therapeutics, by means of bioinformatics methods and functional assays. He also looked into the differential expression patterns of developmental genes in model animals. He is currently a postdoctoral fellow in the European Molecular Biology Laboratory – European Bioinformatic Institute (EMBL-EBI), Cambridge, UK, one of the world's leading bioinformatics institutes, to continue his passion for using bioinformatics to address major biological questions.

List of publications

Xie B, Dashevsky D, Rokyta D, Ghezellou P, Fathinia B, Shi Q, Richardson M, Fry B. 2022. Dynamic genetic differentiation drives the widespread structural and functional convergent evolution of snake venom proteinaceous toxins. BMC Biology, 20(4).

Long C, Wu F, Lu Q, **Xie B,** Shen C, Li J, Deng Y, Liang P, Yu Y, Lai R. A strategy for efficient preparation of genus-specific diagnostic antibodies for snakebites. Frontiers Immunology, 2021, p.4621

de Bakker MA, van der Vos W, de Jager K, Chung WY, Fowler DA, Dondorp E, **Xie B**, ..., Richardson MK. 2021. Selection on phalanx development in the evolution of the bird wing. Molecular Biology and Evolution 38 (10), 4222-4237

Huang Y, Bian C, Liu Y, You X, Yu H, Yi Y, **Xie B,** Shi Q. 2020. Fish Genomics. Encyclopedia of Marine Biotechnology (Book), 3, 1843-1866.

Ibrahim M, **Xie B,** Richardson MK. 2020. The growth of endothelial-like cells in zebrafish embryoid body culture. Experimental Cell Research. 392(2), 112032.

Xie B, Yu H, Kerkkamp H, Wang M, Richardson M, Shi Q. 2019. Comparative transcriptome analyses of venom glands from three scorpionfishes. Genomics, 111(3), 231-241.

Derez CM, Arbuckle K, Ruan Z, **Xie B,** Huang Y, Dibben L, Shi Q, Vonk FJ, Fry BG. 2018. A new species of bandy-bandy (Vermicella: Serpentes: Elapidae) from the Weipa region, Cape York, Australia. Zootaxa, 4446 (1), 001-012.

Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur RR, Li C, Becker L, Bellora N, Zhao X, Li X, Wang M, Fang C, **Xie B,** Zhou Z, Huang H, Chen S, Venkatesh B, Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences, 115(24), 6249-6254.

Xie B, Huang Y, Baumann K, Fry BG, Shi Q. 2017. From marine venoms to drugs: Efficiently supported by a combination of transcriptomics and proteomics. Marine Drugs*, 15*(4), 103.

Debono J, **Xie B,** Violette A, Fourmy R, Jaeger M, Fry BG. 2017. Viper venom botox: the molecular origin and evolution of the waglerin peptides used in anti-wrinkle skin cream. Journal of Molecular Evolution, 84(1), 8-11.

Yang DC, Deuis JR, Dashevsky D, Dobson J, Jackson TN, Brust A, **Xie B,** Koludarov I, Debono J, Hendrikx I. 2016.The snake with the scorpion's sting: Novel three-finger toxin sodium channel activators from the venom of the long-glanded blue coral snake (*Calliophis bivirgatus*). Toxins, 8(10), 303.

Xie B, Li X, Lin Z, Ruan Z, Wang M, Liu J, Tong T, Li J, Huang Y, Wen B, Shi Q. 2016. Prediction of toxin genes from Chinese yellow catfish based on transcriptomic and proteomic sequencing. International Journal of Molecular Sciences, 17(4), 556.