Comparative analyses of venom-associated genes from an Old World viper, *Daboia russelii*

Neeraja M Krishnan[1] and Binay Panda[1]*

[1]Ganit Labs, Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology, Biotech Park, Electronic City Phase I, Bangalore 560100

* Corresponding author: binay@ganitlabs.in

Abstract

Molecular basis of toxin gene diversity among snakes is poorly understood. Lack of whole genome sequence information for most snakes makes studies on toxin genes and their orthologous counterparts difficult. One of the challenges in studying snake genomes is the acquisition of biological material from live animals, especially from the venomous ones. Additionally, in certain geographies, Government permission is required to handle live snakes making the process cumbersome and time-consuming. Here, we report comparative sequence analyses of toxin genes from Russell's viper (*Daboia russelii*) using whole-genome sequencing data obtained from the skin exuviate. In addition to the comparative analyses of 46 toxin-associated proteins, we present evidence of unique sequence motifs in five key toxin-associated protein domains; nerve growth factor (NGF), platelet derived growth factor (PDGF), Kunitz/Bovine pancreatic trypsin inhibitor (Kunitz BPTI), cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins (CAP) and cysteine-rich secretory protein (CRISP). We compared the venom-associated domains from Russell's viper with those from both venomous and non-venomous vertebrates and invertebrates. The *in silico* study on structures identified V11 and T35 in the NGF domain; F23 and A29 in the PDGF domain; N69, K2 and A5 in the CAP domain; and Q17 in the CRISP domain to be responsible for differences in the largest pockets across the protein domain structures in New World vipers, Old World vipers and elapids. Similarly, residues F10, Y11 and E20 appear to play an important role in the protein structures across the kunitz protein domain of viperids and

elapids. Our study sheds light on the uniqueness of these key toxin-associated proteins and their evolution in vipers.

Data deposition: Russell's viper sequence data is deposited in the NCBI SRA database under the accession number SRR5506741 and the GenBank accession numbers for the individual venom-associated genes is provided in Table S1.

Keywords: Toxin-associated genes, Russell's viper, Old World vipers, New World vipers, elapids

Introduction

Snake venom genes and their products offer an excellent model system to study gene duplication, evolution of regulatory DNA sequences, and biochemical diversity and novelty of venom proteins. Additionally, snake venoms have tremendous potential in developing new drugs and bioactive compounds (Vonk et al. 2011). Previous studies have highlighted the importance of gene duplications and/or sub-functionalization (Malhotra et al. 2010; Rokyta et al. 2011; Hargreaves et al. 2014) and transcriptional/post-transcriptional mechanisms (Casewell et al. 2014) contributing towards snake venom diversity. Venom studies so far have extensively used peptides/protein data alongside individual gene sequences or sequences of particular family members to study variations on gene structure and sequence composition. Till date, whole genome sequences of seven snake species, king cobra *Ophiophagus hannah* (Vonk et al. 2013); Burmese

python *Python bivitattus* (Castoe et al. 2013); rattlesnake *Crotalus atrox* (Dowell et al. 2016), Florida pygmy rattlesnake *Sistrurus miliarius barbouri* (Vicoso et al. 2013); garter snake *Thamnophis elegans* (Vicoso et al. 2013); five-pacer viper *Deinagkistrodon acutus* (Yin et al. 2016); and corn snake *Pantherophis guttatus* (Ullate-Agote et al. 2014), have either been published or their sequence information available in the public domain. In addition, genome-sequencing efforts are either underway or the sequences of venom-associated genes have been deposited in the databases for few others (Kerkkamp et al. 2016). Out of the sequenced genomes, only a few have been annotated or the annotations have been made public, a key requirement for comparative analysis of genes. This, along with the lack of availability of whole genome sequences and/or complete coding sequences for most snakes has made studies on toxin gene orthologies and gene variation among venomous snakes limiting.

Four snakes, Russell's viper (*Daboia russelii),* saw-scaled viper (*Echis carinatus*), spectacled cobra (*Naja naja*), and common krait (*Bungarus caeruleus*) are responsible for most snakebite-related mortality in India (Mohapatra et al. 2011; Whitaker 2015). Russell’s viper is a Old World viper, member of the taxon Viperidae and subfamily Viperinae and is responsible for large numbers of snakebite incidents and deaths in India. Very little is known about the diversity of venom-associated genes from any viper, including the only Old World viper where complete genome sequence information is available (European adder, *Vipera berus berus*, https://www.ncbi.nlm.nih.gov/bioproject/170536). Lack of any complete

genome annotation from this viper and other snake species reduces the scope of a detailed comparative study on toxin-associated genes. Such a study involving various groups of venomous and non-venomous snakes, in addition to other venomous vertebrates and invertebrates, will facilitate our understanding on the evolution of these genes, their diversity, and function.

One of the challenges in studying the genomes of venomous animals is related to sample acquisition. Additionally, in India, Government permission is required to catch snakes and extract blood samples from them (all snakes are protected in India under the Indian Wildlife Protection Act, 1972). Even with permission, there is a chance to adversely affect the animals during sample acquisition. This may be circumvented by the use of skin exuviate (shed skin) that does not require handling or drawing blood or taking any tissue from the animals. However, working with DNA isolated from shed skin has its own challenges. Microbial contamination, lack of full-length DNA in the exuviate cells and computational challenges in dealing with short stretches of DNA are some of the bottlenecks for working with DNA from exuviate skins.

In the current study, we explored the possibility of getting toxin gene information from low-coverage whole-genome sequencing data using skin exuviate from an Old World viper, Russell's viper, and performed comparative analysis on the annotated 51 venom-associated genes representing all the major venom-associated protein families (Fry 2005) for which coding sequences were available from a New World viper, a pit viper, *Protobothrops mucrosquamatus*. We focused our analyses on five key toxin-associated protein domains; nerve growth factor (NGF), platelet derived growth factor (PDGF), Kunitz/Bovine pancreatic trypsin inhibitor (Kunitz BPTI),

cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins (CAP) and cysteine-rich secretory protein (CRISP) in Russell's viper and discovered key residues that are changed across Old World vipers, New World vipers and elapids that might have contributed towards the evolution of venom in vipers.

Materials and Methods

Russell's viper skin exuviate and DNA isolation

Freshly shed skin exuviate of Russell's viper from Bangalore, India was a gift from Mr. Gerry Martin. The skin exuviate for the entire snake was obtained, cleaned thoroughly with 70% ethanol and nuclease-free water 3 times each, dried thoroughly and frozen until the time of extraction of DNA. Genomic DNA was extracted following the protocol of Fetzner (Fetzner 1999) with modifications.

Sequencing, read processing and assembly

Illumina paired-end read libraries were prepared following the manufacturer instructions using amplification free genomic DNA library preparation kit and sequenced using Illumina HiSeq2500 instrument. Archael, bacterial and human sequence contamination were removed from the Russell's viper sequence using DeConSeq (Schmieder and Edwards 2011) using curated and representative genomes (https://www.ncbi.nlm.nih.gov/genome/browse/reference/). Furtherm ore, the sequenced reads were post-processed to remove unpaired reads and quality analysis was performed using FastQC v0.1

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The rd_len_cutoff option was exercised during the read assembly step to trim off the low-quality bases, since the per-base quality was found to drop below 28 after the initial 50-70 bases of the read. The Russell's viper read libraries were assembled using SOAPdenovo2 (r240) (Luo et al. 2012).

Identifying toxin-associated genes, coding regions, and gene structures

The DNA sequences for 51 out of 54 venom-associated genes (Fry 2005) from *Protobothrops mucrosquamatus* were downloaded (Table 1). These were used to fish genomic scaffolds bearing highly similar sequences in Russell's viper genome assembly, using a BLAST with an E-value threshold of 10^{-3}. These scaffolds were then anchored to the respective coding sequences from *Protobothrops mucrosquamatus* using a discontiguous megaBlast, to determine the correct frame of translation and extract the complete amino acid coding sequence (CDS) corresponding to Russell's viper. We obtained the exon-intron structures for all the toxin-associated genes in Russell's viper by aligning the CDS with gene sequences using discontiguous megaBlast and plotted using the tool GSDS2.0 (Hu et al. 2015). The sequences for the Russell's viper venom-associated genes and their GenBank accession numbers are provided in Table S1.

Comparative analyses of toxin-associated protein domains

The amino acid sequences of all the Russell's viper's toxin-associated genes were subjected to domain search using Pfam (Finn et al. 2016) (Table S2). All domain sequences was then serially aligned using blastp to protein sequences from 18 snake species (Table S3). Five domains (NGF, PDGF,

Kunitz BPTI, CAP and CRISP) from four genes (NGF, VEGF, CRISP/Serotriflin, and Kunitoxin), which showed variability across different snake groups and where sequence information were available beyond the whole genome sequences, were used for expansive comparative analyses (Table S4) using sequences from Viperids (taxid: 8689), elapids (taxid: 8602), Colubrids (taxid: 8578), Boids (taxid: 8572), Acrochordids (taxid: 42164), Pythonids (taxid: 34894), lizards (squamates (taxid: 8509) minus snakes (taxid: 8570)), Crocodylia (taxid: 51964) and Testudines (taxid: 8459).

3D structure prediction of the chosen domains

Consensus sequences were determined from NGF, PDGF, Kunitz BPTI, CAP and CRISP domain alignments using Simple Consensus Maker (https://www.hiv.lanl.gov/content/sequence/CONSENSUS/SimpCon.html) for New World vipers (NWV), Old World vipers (OWV) and elapids. The consensus sequences were submitted to the Phyre2 (Kelley et al. 2015) protein fold recognition server using the standard mode (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). The best 3D model was further investigated by Phyre2 (Kelley et al. 2015) to analyze the structural model using various open source tools.

Results

Skin exuviate yielded fairly good quality DNA for genome sequencing and near complete coding sequences for toxin-associated genes

Genomic DNA isolated from the skin exuviate of Russell's viper was fairly intact with most of the DNA in the size range of more than 5kbp range

(File S1). Sequenced short reads were assembled and then used to fish all the 51 toxin-associated genes in Russell's viper (see Materials and Methods). Next, we obtained the exon-intron structures for all the toxin-associated genes in Russell's viper by aligning the CDS with gene sequences using discontiguous megaBlast and plotted using the tools GSDS2.0 (Hu et al. 2015) (File S2). We found that the average length of exons in Russell's viper toxin-associated genes is around 190 nucleotides (nt), matching well with the lengths of other vertebrate exons (Gelfman, et al. 2012).

Comparative protein domain analyses

Among the toxin-associated genes, a larger pool of sequences are available only for NGF, PDGF domain of VEGF, Kunitz_BPTI domain of Kunitoxin, CRISP and CAP domains in CRISP and Serotriflin proteins, from various snake groups (Colubridae, Boidae, Pythonidae and Acrochordidae), non-snake reptilian groups (lizards, crocodiles and Testudines), venomous invertebrates (wasps, spiders and scorpions) and venomous vertebrates (fishes and mammals). Therefore, these domains were compared with those from Russell's viper. Comparative domain analysis was performed for all toxin-associated genes (File S3) across 18 snake species where sequence information is available (Table S3). In the case of five domains: CAP and CRISP domains of CRISP and serotriflin genes (L and AL), Kunitz BPTI of kunitoxin (S), NGF (T) and PDGF of VEGFA (AP-AR) and VEGFF (AU), we found that the maximum number of species aligned to their domain sequences. Some protein domains, the CRISP, Kunitz BPTI, guanylate CYC, PDGF of VEGFF and WAP, show long stretches of mismatches (File S3)

compared with Russell's viper sequence. Out of these, only NGF and PDGF domains of VEGF carried amino acid changes specific to the New World vipers, that were completely absent in any other group used for comparison, including in lizards, crocodiles, and turtles (File S4). Specific changes in these proteins and their implications are discussed below.

Russell's viper NGF is a single exon gene with a 745nt transcript coding for a 244 amino acid protein consisting of a single NGF domain (fig. 1A). The NGF domain bears 28% sequence conservation across all the five vertebrate phyla, namely, fishes, amphibians, reptiles, aves and mammals distributed along the length of the domain (fig. 1B). Thirty-six percent out of these residues are conserved across other venomous vertebrates (fishes, squamates and mammals) and venomous invertebrates (scorpions and wasps) (fig. 1C). Thirteen percent of Russell's viper NGF domain residues are variable with respect to the domain sequence in at least one among the NGF sequences in the groups of New World vipers, Old World vipers and elapids (fig. 1D). Although several amino acids in the NGF domain in New World vipers seem to have changed from the Russell's viper and other Old World vipers, their function probably remains unchanged. For example, phenylalanine (F) to isoleucine (I) at position 12 and serine (S) to Asparagine (N) at position 19 between the Old World vipers to the New World vipers does not change the function of the amino acids (from one hydrophobic amino acid to another and from one polar amino acid to another). However, there are others, for example, threonine (T) and glutamine (Q), at position 67 and 68 respectively in the NGF domain of the New World viper evolved only in that group. One of those, a polar amino acid glutamine at position 68, is a very

236 important residue as its corresponding amino acid in any of the other snakes,
237 except in colubrids, is a hydrophobic proline. This evolution of NGF sequence
238 might have implications on its function in the venom of the New World vipers.
239 In Russell's viper, the VEGFA gene comprises five exons coding for a
240 652nt long transcript and a protein with two domains: PDGF and VEGF-C (fig.
241 2A). The PDGF domain sequence exhibits conservation in 65% of its residues
242 across the three vertebrate phyla (reptiles, aves and mammals) (fig. 2B).
243 Since sequence information from fishes and amphibians were not available,
244 they could not be included in the comparison study. Out of the conserved
245 residues, 21% of those were also conserved in venomous vertebrates
246 (squamates and mammals) and venomous invertebrates (wasps). Fifteen
247 percent of the PDGF domain residues were variable in at least one of the
248 three snake groups: New World vipers, Old World vipers and elapids (fig.2C
249 and fig. 2D). Like the NGF domain, the evolution of the PDGF domain in New
250 World vipers at certain amino acids is striking. For example, in the New World
251 vipers, the position 67 is a polar amino acid tyrosine (Y) while in all other
252 reptiles, venomous invertebrates and mammals; this is primarily a
253 hydrophobic amino acid phenylalanine (F). This might bear implications on the
254 protein's structure and function.
255 Kunitoxin in Russell's viper is a 3.1kb gene comprising two exons, with
256 a transcript length of 270nt that codes for a 44 amino acids long single Kunitz
257 BPTI domain (fig. 3A). About 29% of the protein domain residues are
258 conserved across the four vertebrate phyla (amphibians, reptiles, aves and
259 mammals) (fig. 3B). Since sequence information from the Kunitz BPTI for
260 fishes was not available, they could not be included in the comparison. Out of

these conserved residues, 76% are conserved in venomous vertebrates (squamates and mammals) and venomous invertebrates (scorpions and wasps) (fig. 3C) and 56% of the domain residues are variable in at least one of three snake groups (New World vipers, Old World vipers, and elapids) (fig. 3D). Of the residues that are evolved in the New World vipers, the second residue, a positively charged one, alanine (A) is present only in the Old World vipers, which is replaced by a hydrophobic residue, proline (P), in the New World vipers and elapids. Residues 14-18 are very polymorphic in the New World vipers and elapids, but not so in the Old World vipers.

The CRISP gene in Russell's viper is a 25kb long gene, comprises of 8 exons coding for a 787nt transcript and two protein domains, CAP and CRISP (fig. 4A). The CAP domain exhibits conservation in 7% of its residues across all the five vertebrate phyla (fig. 4B). Forty-two percent of those residues are conserved across venomous vertebrates (amphibians, squamates and mammals) and venomous invertebrates (scorpions and wasps) (fig. 4C). In addition, there are five residues conserved across all the venomous animals (fig. 4C). Twenty-seven percent of the CAP domain and 15% of the CRISP domain residues are variable in at least one of the three snake groups (fig. 4D). There are several extra residues for the CAP domain in the New World vipers and elapids, but not in the Old World vipers. The conserved residues comprised mostly of Cystines and to a lesser extent Asparagines (fig. 4E) across venomous vertebrates (squamates and mammals) (fig. 4F). Sixty percent of the CRISP domain residues are variable in at least one viperids or elapid member with respect to the domain sequence of Russel's viper (fig. 4G).

Next, we explored the role of consensus domain sequences and their possible role of conserved amino acids in those key toxin-associated protein domains across New World vipers, Old World vipers and elapids. We constructed the 3D structure models using Phyre2, followed by Phyre2 investigation, for further analyses on the structural model. As evident from the analyses, amino acid residues 18-19 and 117 of the NGF domain reflected a difference in mutation sensitivity as detected by SusPect algorithm (Yates et al. 2014), especially in the elapids compared to the viperids (fig. 5). Residue 18 is Valine in the viperids and Isoleucine in the elapids; residue 19 is Serine in the Old World vipers and Asparagine in the New World vipers; and residue 117 is Threonine in the elapids and Serine in the New World vipers (fig. 5A). This might have implications in the structure of the protein as the largest pockets detected by fpocket algorithm appear to be vastly different among the New World vipers, Old World vipers and elapids for the NGF, PDGF, CAP and CRISP domains (fig. 5). The pockets appear small in all cases for the elapids, and largest in the case of Old World vipers (fig. 5). Minor differences in clashes were observed at residues 10,11 and 20 of the Kunitz domain and residue 38 of this domain showed a rotamer conflict in the case of the New World vipers (fig. 5C). Similarly, residue 46 of the CAP domain and residues 4 and 31 of the CRISP domain showed rotamer conflict for the viperids (fig. 5D and fig. 5E). The other protein quality and functional parameters were not affected across the 3D structure models for the three snake groups (File S5).

Discussion

311 Accessibility and affordability for high-throughput sequencing
312 technologies along with the availability of sophisticated computational tools to
313 assemble, annotate and interpret genomes is playing a powerful role in
314 deciphering gene functions and their role in evolution. Multiloci gene families
315 usually encode snake toxin genes that produce gene isoforms generated by
316 the process of gene duplications (Fry 2005; Casewell et al. 2013). Russell's
317 viper (*Daboai russelli*) is a member of the taxon *Viperidae* and
318 subfamily *Viperinae* and is one among four other venomous snakes
319 responsible for the most snakebite incidents and deaths in India (Mohapatra
320 et al. 2011; Whitaker 2015). Russell's viper is a member of the group Old
321 World vipers or true vipers or pitless vipers, characterized by the lack of heat-
322 sensing pit organs (Mallow et al. 2003). Snakes have evolved two main
323 venom types, neurotoxic (elapids that include cobras and kraits) or
324 haemotoxic (viperids that include vipers) (Fry et al. 2008). Russell's viper
325 venom is haemotoxic. There is significant variation in the venom compositions
326 of snakes, even at times within the same species as in the case of Russell's
327 viper (Jayanthi and Gowda 1988). The variation in the venom composition
328 within the same species is thought to be a result of adaptation in response to
329 the difference in diets (Daltry et al. 1996; Barlow et al. 2009; Casewell et al.
330 2013). The composition of venom of Russell's viper from different regions of
331 India is different (Jayanthi and Gowda 1988), making anti-venom produced
332 using snake venoms from a single geography against all Russell's viper bites
333 across the country ineffective. Currently, efforts are underway to collect
334 venoms of Russell's viper from different regions of India in order to
335 understand the difference in their venom composition (Rom Whitaker, and

Gerry Martin, *personal communications*). Currently, information on toxin-associated genes from any viper, especially Russell's viper is extremely limiting and therefore, comparative analyses of the toxin-associated genes will add value to our understanding of its venom. The only Old World viper, where complete genome sequence information is available is from a European adder, *Vipera berus berus* (https://www.ncbi.nlm.nih.gov/bioproject/170536). Although sequence information is there for this species, the annotation is not available.

The aim of the current study was two folds. First, as handling and getting biological material from snakes in India requires Government permission and specific expertise, we wanted to test whether good quality whole-genome sequence information can be obtained using skin exuviate. Second, we wanted to test whether we can assemble the toxin-associated genes using low-coverage sequencing data from a next-generation sequencing platform, and use it for comparative genomics study. On both the accounts, we found the results to be satisfactory. First, although shed skin is often contaminated with bacteria and other microorganisms, using freshly shed skin exuviate, we successfully isolated high molecular weight genomic DNA (File S1), which was subsequently used to generate genome sequencing data. This helps in obtaining biological material without having to require permission to capture venomous animals and also reduce stress in animals during handling. We believe this is an attractive option for generating snake genome data and studying molecular evolution in snakes. Second, from the sequence data, we succeeded in assembling near complete CDS for 51 out of 54 toxin-associated genes (Fry 2005). This highlights the utility of low-

coverage genome sequencing data in studying important genes and their role in evolution. As the lengths of the toxin-associated genes in Russell's viper were much longer than the CDS, the intronic sequences were assembled with gaps. This was primarily due to the low coverage sequencing data used for assembly and the lack of long-insert mate pair sequencing data in our repertoire. The mean length of exons for the toxin-associated genes in Russell's viper is 190 base pairs, which was much smaller compared to the average intron length; the exons were assembled accurately using short-read sequences. Another interesting observation from our data is that the AT to GC ratio in the CDS regions (cumulatively for all the 51 genes) of the toxin-associated genes in Russell's viper is 1:1 whereas it is skewed (the ratio is 1.5:1) for the full gene sequences.

Our study has two implications. First, it demonstrates the feasibility of *de novo* analyses of genes and gene families without prior sequence information and annotation, and second; this facilitates sequencing a large number of genes and gene families without going through the process of designing individual primers for amplification and Sanger sequencing. Despite these advantages, our study has certain limitations. First, like any other annotation-based study, it relies on the quality of existing/prior annotation of toxin-related genes. Although the chances are slim, as we used a closely related viper to get the toxin-associated gene sequences in Russell's viper, it is entirely possible that we might have missed some key genes specific to Russell's viper. Using a high-coverage sequencing data along with the use of *de novo* tools might help annotate unique genes in Russell's viper in the future. Second, our study does not touch the expression of the toxin genes in

386 Russell's viper and their significance compared to other venomous snakes. It
387 is possible that the unique spatial and temporal expression pattern of the
388 toxin-associated genes will explain their evolution better.

389 **Acknowledgements**

390 This work was supported by Department of Electronics and Information
391 Technology, Government of India [18(4)/2010-E-Infra., 31-03-2010]; and
392 Department of IT, BT and ST, Government of Karnataka, India [3451-00-090-
393 2-22]. We thank Kunal Dhas and Arun Hariharan for their help in laboratory
394 experiments.

395 **Literature Cited**
396
397 Barlow, A., C. E. Pook, R. A. Harrison, and W. Wuster. 2009. Coevolution of
398 diet and prey-specific venom activity supports the role of selection in
399 snake venom evolution. Proc Biol Sci 276:2443-2449.
400 Casewell, N. R., S. C. Wagstaff, W. Wuster, D. A. Cook, F. M. Bolton, S. I.
401 King, D. Pla, L. Sanz, J. J. Calvete, and R. A. Harrison. 2014. Medically
402 important differences in snake venom composition are dictated by
403 distinct postgenomic mechanisms. Proc Natl Acad Sci U S A 111:9205-
404 9210.
405 Casewell, N. R., W. Wuster, F. J. Vonk, R. A. Harrison, and B. G. Fry. 2013.
406 Complex cocktails: the evolutionary novelty of venoms. Trends Ecol
407 Evol 28:219-229.
408 Castoe, T. A., A. P. de Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K.
409 Fujita, R. P. Ruggiero, J. F. Degner, J. M. Daza, W. Gu, J. Reyes-
410 Velasco, K. J. Shaney, J. M. Castoe, S. E. Fox, A. W. Poole, D.
411 Polanco, J. Dobry, M. W. Vandewege, Q. Li, R. K. Schott, A. Kapusta,
412 P. Minx, C. Feschotte, P. Uetz, D. A. Ray, F. G. Hoffmann, R. Bogden,
413 E. N. Smith, B. S. Chang, F. J. Vonk, N. R. Casewell, C. V. Henkel, M.
414 K. Richardson, S. P. Mackessy, A. M. Bronikowski, M. Yandell, W. C.
415 Warren, S. M. Secor, and D. D. Pollock. 2013. The Burmese python
416 genome reveals the molecular basis for extreme adaptation in snakes.
417 Proc Natl Acad Sci U S A 110:20645-20650.
418 Daltry, J. C., W. Wuster, and R. S. Thorpe. 1996. Diet and snake venom
419 evolution. Nature 379:537-540.
420 Dowell, N. L., M. W. Giorgianni, V. A. Kassner, J. E. Selegue, E. E. Sanchez,
421 and S. B. Carroll. 2016. The Deep Origin and Recent Loss of Venom
422 Toxin Genes in Rattlesnakes. Curr Biol 26:2434-2445.

Fetzner, J. W., Jr. 1999. Extracting high-quality DNA from shed reptile skins: a simplified method. Biotechniques 26:1052-1054.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44:6.

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

Fry, B. G., H. Scheib, L. van der Weerd, B. Young, J. McNaughtan, S. F. Ramjan, N. Vidal, R. E. Poelmann, and J. A. Norman. 2008. Evolution of an arsenal: structural and functional diversification of the venom system in the advanced snakes (Caenophidia). Mol Cell Proteomics 7:215-246.

Gelfman S, Burstein, D., Penn, O., Savchenko, A. et al. 2012. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res 22: 35-50.

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

Heise, P. J., L. R. Maxson, H. G. Dowling, and S. B. Hedges. 1995. Higher-level snake phylogeny inferred from mitochondrial DNA sequences of 12S rRNA and 16S rRNA genes. Mol Biol Evol 12:259-265.

Hu, B., J. Jin, A. Y. Guo, H. Zhang, J. Luo, and G. Gao. 2015. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31:1296-1297.

Jayanthi, G. P. and T. V. Gowda. 1988. Geographical variation in India in the composition and lethal potency of Russell's viper (Vipera russelli) venom. Toxicon 26:257-264.

Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass and M. J. E. Sternberg. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10: 845–858.

Kerkkamp, H. M., R. M. Kini, A. S. Pospelov, F. J. Vonk, C. V. Henkel, and M. K. Richardson. 2016. Snake Genome Sequencing: Results and Future Prospects. Toxins (Basel) 8.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, and J. Wang. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.

Malhotra, A., S. Creer, C. E. Pook, and R. S. Thorpe. 2010. Inclusion of nuclear intron sequence data helps to identify the Asian sister group of New World pitvipers. Mol Phylogenet Evol 54:172-178.

Mallow, D., D. Ludwig, and G. Nilson. 2003. True vipers : natural history and toxinology of Old World vipers. Krieger Pub. Co., Malabar, Fla.

Mohapatra, B., D. A. Warrell, W. Suraweera, P. Bhatia, N. Dhingra, R. M. Jotkar, P. S. Rodriguez, K. Mishra, R. Whitaker, P. Jha, and C. Million Death Study. 2011. Snakebite mortality in India: a nationally representative mortality survey. PLoS Negl Trop Dis 5:e1018.

Rokyta, D. R., K. P. Wray, A. R. Lemmon, E. M. Lemmon, and S. B. Caudle. 2011. A high-throughput venom-gland transcriptome for the Eastern Diamondback Rattlesnake (Crotalus adamanteus) and evidence for pervasive positive selection across toxin classes. Toxicon 57:657-671.

Schmieder, R. and R. Edwards. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6:e17288.

Ullate-Agote, A., M. C. Milinkovitch, and A. C. Tzika. 2014. The genome sequence of the corn snake (Pantherophis guttatus), a valuable resource for EvoDevo studies in squamates. Int J Dev Biol 58:881-888.

Vicoso, B., J. J. Emerson, Y. Zektser, S. Mahajan, and D. Bachtrog. 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. PLoS Biol 11:e1001643.

Vonk, F. J., N. R. Casewell, C. V. Henkel, A. M. Heimberg, H. J. Jansen, R. J. McCleary, H. M. Kerkkamp, R. A. Vos, I. Guerreiro, J. J. Calvete, W. Wuster, A. E. Woods, J. M. Logan, R. A. Harrison, T. A. Castoe, A. P. de Koning, D. D. Pollock, M. Yandell, D. Calderon, C. Renjifo, R. B. Currier, D. Salgado, D. Pla, L. Sanz, A. S. Hyder, J. M. Ribeiro, J. W. Arntzen, G. E. van den Thillart, M. Boetzer, W. Pirovano, R. P. Dirks, H. P. Spaink, D. Duboule, E. McGlinn, R. M. Kini, and M. K. Richardson. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proc Natl Acad Sci U S A 110:20651-20656.

Vonk, F. J., K. Jackson, R. Doley, F. Madaras, P. J. Mirtschin, and N. Vidal. 2011. Snake venom: From fieldwork to the clinic: Recent insights into snake biology, together with new technology allowing high-throughput screening of venom, bring new hope for drug discovery. Bioessays 33:269-279.

Whitaker, R. 2015. Snakebite in India today. Neurol India 63:300-303.

Wüster, W., L. Peppin, C. E. Pook, and D. E. Walker. 2008. A nesting of vipers: Phylogeny and historical biogeography of the Viperidae (Squamata: Serpentes). Mol Phylogenet Evol 49:445-459.

Yates, C. M., I. Filippis, L. A. Kelley and M. J. Sternberg. 2014. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features J Mol Biol. 426(14): 2692-2701

Yin, W., Z. J. Wang, Q. Y. Li, J. M. Lian, et al. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. Nat Commun 7:13107.

512 **Table 1:** Venom-associated genes and their representative families used in

513 the current study.

514

Gene	Species with the available sequence information	Protein Family
ACHE	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	Acetylcholinesterase
ADAM11		ADAM (disintegrin/ metalloprotease)
ADAM17		
ADAM19		
ADAM23		
PROK1	*Protobothrops mucrosquamatus, Ophiophagus hannah* and *Python bivitattus*	AVIT (prokinectin)
PROK2	*Protobothrops mucrosquamatus and Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	
CPAMD8	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	Complement C3
crotasin	*Protobothrops mucrosquamatus*	Crotasin/ beta defensin
CST1	No sequence information is available in any of the four species	Cystatin
CST3	*Ophiophagus hannah*	
CST4	No sequence information is available in any of the four species	
CSTA	*Protobothrops mucrosquamatus* and *Thamnophis sirtalis*	
EDN1	*Protobothrops mucrosquamatus* and *Python bivittatus*	Endothelin
EDN3	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus*, *Thamnophis sirtalis* and	
F5		Factor V
F10	*Ophiophagus hannah*	Factor X
KLKB1	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus, Thamnophis sirtalis* and	Kallikrein
KLK14	*Ophiophagus hannah*	
kunitoxin	*Protobothrops mucrosquamatus, Python bivittatus* and *Ophiophagus hannah*	Kunitz-type protease inhibitor
LYNX1	*Ophiophagus hannah*	LYNX/SLUR
CLEC3A	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	Lectin
CLEC3B	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	
CLEC11A	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	
CLEC16A	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	
CLEC19A	*Protobothrops mucrosquamatus and Python bivittatus*	
NPR1	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	Natriuretic peptide
NPR2	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	
NPR3		
NGF	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus, Thamnophis sirtalis, Protobothrops flavoviridis, Crotalus horridus, Sistrurus miliarius barbouri* and *Boa constrictor*	Beta-nerve growth factor
PLAA	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	Phospholipase A (2)
PLA2R1		

PLA2G1B	*Python bivittatus, Thamnophis sirtalis* and *Protobothrops mucrosquamatus*	
PLA2G10	*Protobothrops mucrosquamatus, Protobothrops flavoviridis, Thamnophis sirtalis, Ophiophagus hanna* and *Python bivittatus*	
PLA2G12A	*Python bivittatus, Thamnophis sirtalis* and *Protobothrops mucrosquamatus*	
PLA2G12B	*Protobothrops mucrosquamatus, Ophiophagus hannah* and *Python bivittatus*	
PLA2G15	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus and Thamnophis sirtalis*	
PLA2G3	*Protobothrops mucrosquamatus, Python bivittatus* and *Ophiophagus hannah*	
PLA2G4A	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	
PLA2G4C	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	
PLA2G6	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	
PLA2G7		
SPSB4	*Protobothrops mucrosquamatus* and *Thamnophis sirtalis*	SPIa/Ryanodine
SPSB3	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	
SPSB1	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus* and *Thamnophis sirtalis*	
VEGFA1	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus, Thamnophis sirtalis, Crotalus horridus* and *Protobothrops flavoviridis*	Vascular endothelial growth factor (VEGF)
VEGFA2		
VEGFA3		
VEGFB	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus, Thamnophis sirtalis, Crotalus horridus, Protobothrops flavoviridis* and *Sistrurus miliarius barbouri*	
VEGFC	*Protobothrops mucrosquamatus, Python bivittatus* and *Thamnophis sirtalis*	
VEGFF	*Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus, Thamnophis sirtalis* and *Protobothrops flavoviridis*	
WAP	*Protobothrops mucrosquamatus, Python bivittatus, Thamnophis sirtalis* and *Ophiophagus hannah*	Whey acidic protein/secretory leukoproteinase inhibitor
WFIKKN1		
WFIKKN2		
CRISP	*Protobothrops flavoviridis, Protobothrops mucrosquamatus, Ophiophagus hannah, Python bivittatus, Thamnophis sirtalis, Crotalus horridus, Calloselasma rhodostoma, Sistrurus miliarius barbouri* and *Deinagkistrodon acutus*	CRISP

515

516

Figure Legends

Fig. 1: Comparative analyses of nerve growth factor (NGF). NGF gene, mRNA, and protein domains of Russell's viper (A) and its comparison with the consensus NGF sequences from all five vertebrate phyla (fishes, amphibians, reptiles, aves and mammals) (B), with venomous (V) vertebrates from multiple phyla of vertebrates and invertebrates (C), and from various reptilian subgroups (D). The shades of brown and grey in B and C represent conservation to various degrees and variability, respectively. Grey in D represents conserved residues, red represents variable residues in New World vipers (NWV), yellow and green represent conserved and variable residues in Old World vipers (OWV), and elapids respectively.

Fig. 2: Comparative analyses of vascular endothelial growth factor - A (VEGF-A). Organization of the gene, mRNA, and protein domains of Russell's viper PDGF domain (A) and its comparison with the consensus sequences from all five vertebrate phyla (fishes, amphibians, reptiles, aves and mammals) (B), from the venomous (V) vertebrates and invertebrates (C), and from various reptilian subgroups (D). The shades of brown and grey in B and C represent conserved and varying residues, respectively. Grey in D represents conserved residues, red represents variable residues in New World vipers (NWV), yellow and green represent conserved and variable residues in Old World vipers (OWV), and elapids respectively

Fig. 3: Comparative analyses of kunitoxin. Organization of the gene, mRNA, and protein domains of Russell's viper (A) and its comparison with the

consensus BPTI domain sequences from all five vertebrate phyla (fishes, amphibians, reptiles, aves and mammals) (B), from venomous (V) vertebrates and invertebrates (C), from various reptilian subgroups (D). The shades of brown and grey in B and C represent conserved and varying residues, respectively. Grey in D represents conserved residues, red represents variable residues in New World vipers (NWV), yellow and green represent conserved and variable residues in Old World vipers (OWV), and elapids respectively.

Fig. 4: Comparative analyses of CRISP. Organization of CRISP gene, mRNA, and protein domains of Russell's viper (A) and its comparison with the consensus CRISP sequences from all five vertebrate phyla (fishes, amphibians, reptiles, aves and mammals, B and E); from venomous animals (V) vertebrates (fishes, squamates and mammals) and invertebrates (scorpions and wasps, C and F); and from various reptilian subgroups (D and G). The shades of brown and grey in B, C, E and F represent conserved and varying residues, respectively. Grey in D and G represents conserved residues, red represents variable residues in New World vipers (NWV), yellow and green conserved and variable residues in Old World vipers (OWV), and elapids respectively.

Fig. 5: Three-dimensional protein structural models to access quality and functional differences in key venom-associated proteins (NGF, A; PDGF, B; Kunitz BPTI, C; CAP, D; and CRISP, E) across New World vipers (NWV), Old

World vipers (OWV) and elapids. The status of the parameters being investigated using Phyre2 are indicated in the color legends on the side.

Supplementary File legends

Table S1: Genbank accession numbers for Russell's viper venom-associated proteins.

Table S2: The amino acid sequences for the Russell's viper's toxin-associated genes.

Table S3: List of completely sequenced snake species.

Table S4: Accession IDs and taxonomic groups of species used in all comparative analyses.

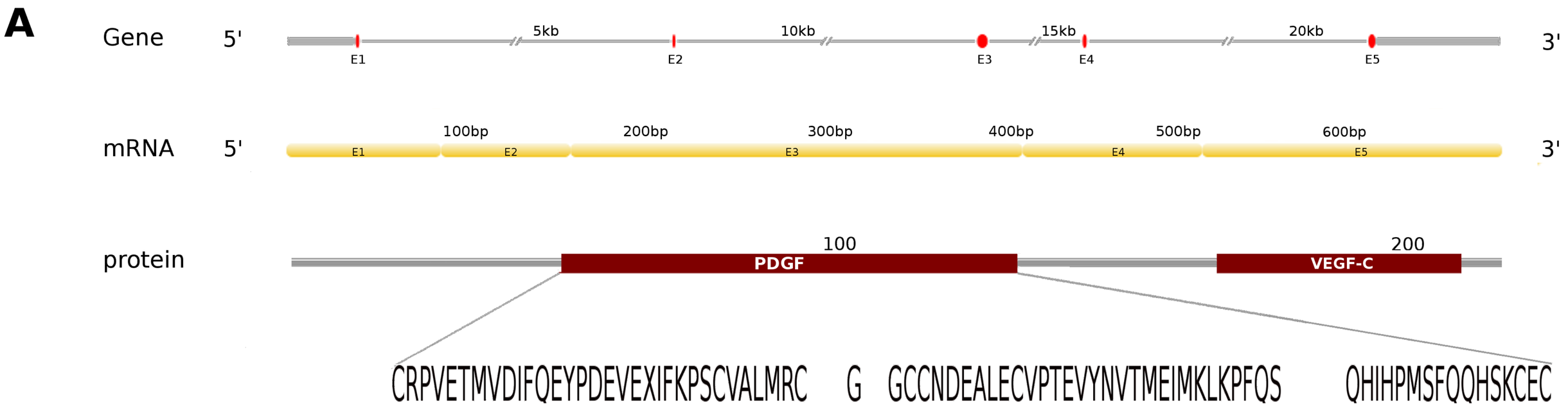
File S1: Agarose gel electrophoresis of genomic DNA extracted from skin exuviate of Russell's viper.

.

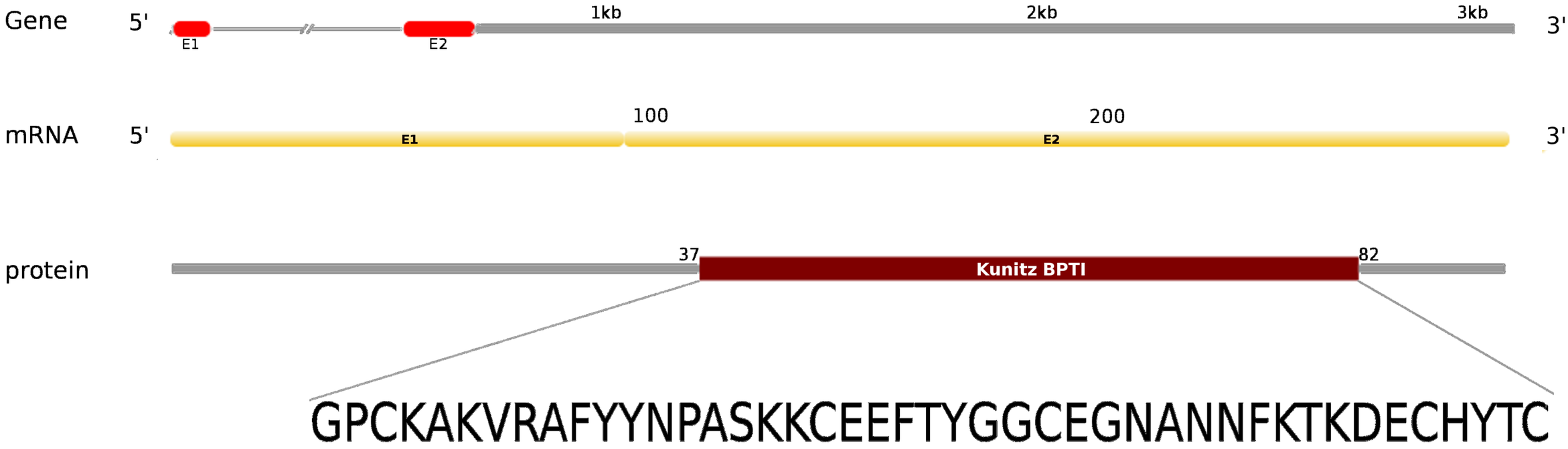
File S2: Exon-intron structures for all the toxin-associated genes in Russell's viper.

File S3: Graphical summary of comparative domain analysis performed for all toxin-associated genes across 18 completely sequenced snake species.

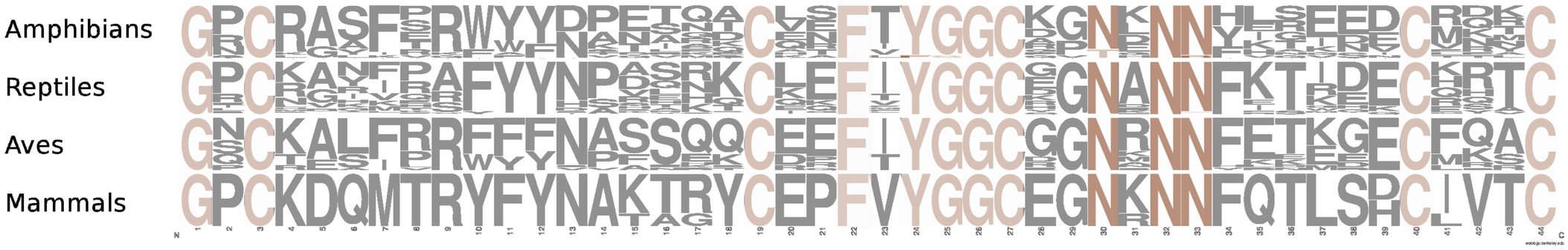
590 **File S4:** A larger pool of sequences available from various snakes and other
591 reptile groups (Colubridae, Boidae, Pythonidae, Acrochordidae, Lizards,
592 Crocodiles and Testudines) were compared with the sequence of Russell's
593 viper for NGF, PDGF, Kunitz BPTI, CAP and CRISP domains.
594
595 **File S5:** 3D protein structural model quality and function measuring parameter
596 comparison across New World vipers (NWV), Old World vipers (OWV) and
597 elapids in NGF, PDGF, Kunitz BPTI, CAP and CRISP domains. The status of
598 the parameters being investigated using Phyre2 are indicated in the color
599 legends on the side.



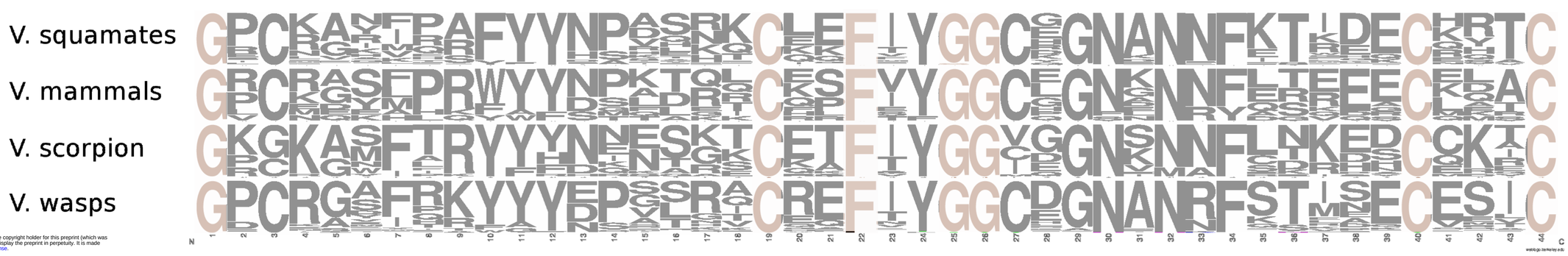
A



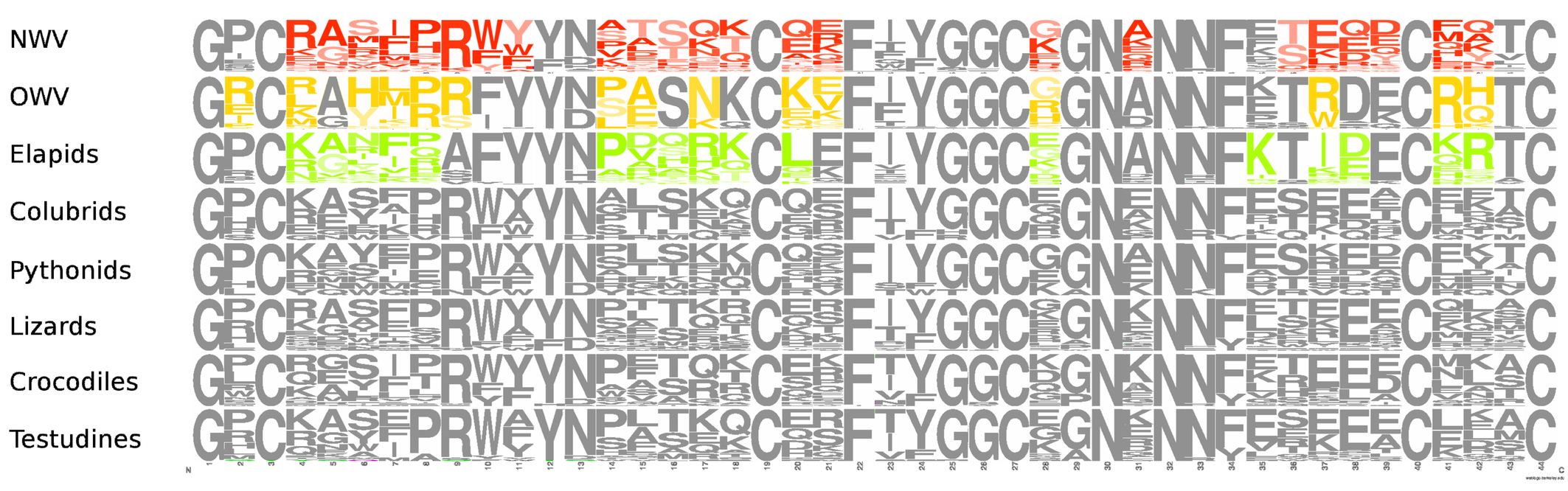
B

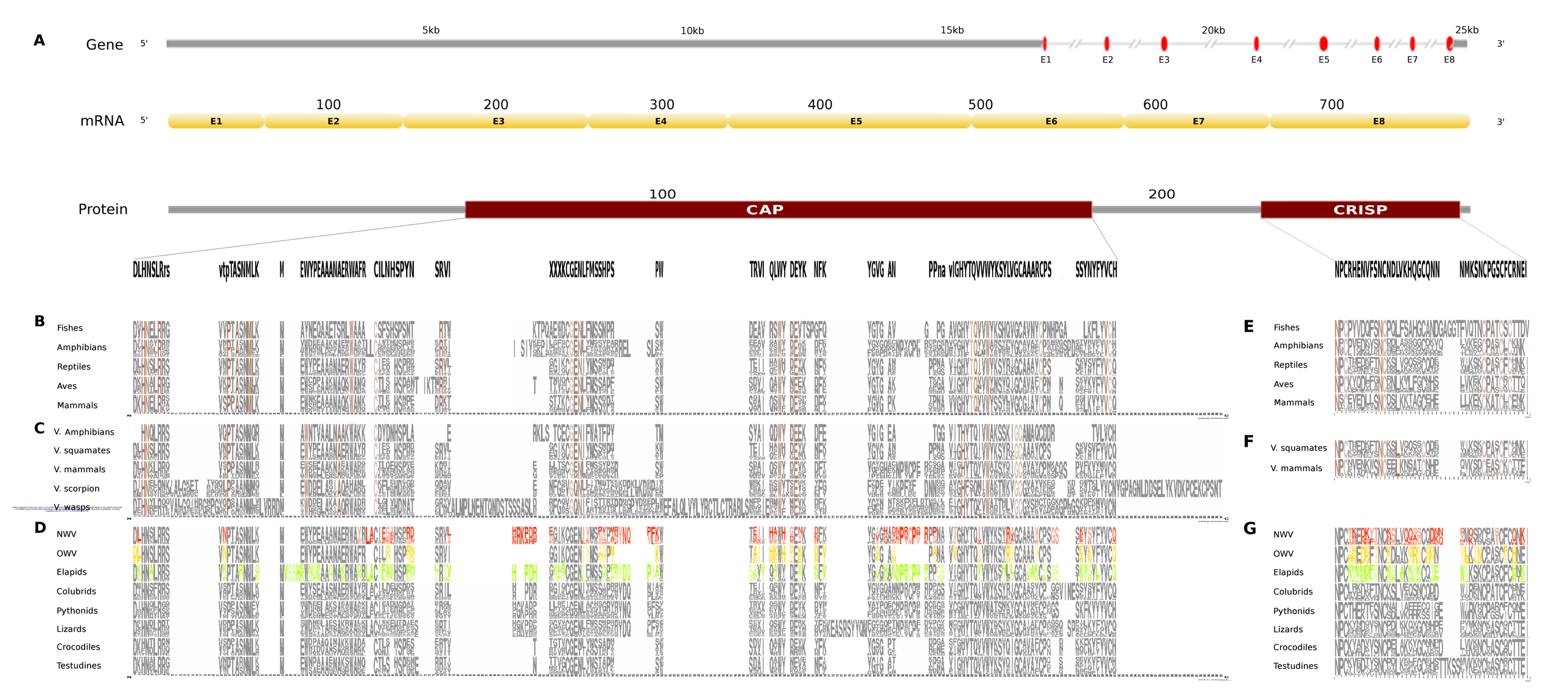


C



D




A
Gene
5'
3'
5kb
10kb
15kb
20kb
25kb
E1
E2
E3
E4
E5
E6
E7
E8
mRNA
100
200
300
400
500
600
700
Protein
CAP
CRISP
B
Fishes
Amphibians
Reptiles
Aves
Mammals
C
V. Amphibians
V. squamates
V. mammals
V. scorpion
V. wasps
D
NWV
OWV
Elapids
Colubrids
Pythonids
Lizards
Crocodiles
Testudines
E
F
G