Supplementary Information for

The origin and diversification of a novel protein family in venomous snakes

Matt W. Giorgianni¹, Noah L. Dowell¹, Sam Griffin², Victoria A. Kassner², Jane E. Selegue², Sean B. Carroll¹

Email: sbcarrol@umd.edu

This PDF file includes:

- Methods
- Figures S1 to S4
- Dataset legends
- SI References

Datasets are excel files accompanying this submission
Methods

Transposable element annotation. Transposable elements were identified using the Repeatmasker web server (A.F.A. Smit, R. Hubley and P. Green, unpublished data. Version: open-4.0.5, RMLib: 20140131, Dfam: 1.3). We used the cross_match search engine run in slow mode (higher sensitivity).

RNA isolation, sequencing, assembly. Venom gland RNA was isolated using the standard Trizol method. One μg of total RNA was provided to the University of Wisconsin - Madison DNA sequencing core for strand-aware Illumina TruSeq RNA-sequencing library preparation. The sequencing library was created using size selected RNA (180 bp). The venom gland libraries were sequenced with Illumina HiSeq2000 for 2 x 100 cycles producing 100bp paired end reads. We used Overlapper8 (12) to merge paired reads and generate reads of uniform quality. RNA quantification was carried out by mapping the merged reads against our list of predicted transcripts using with RSEM 1.3.1 (13), with the bowtie2 option.

Mass Spectrometry. 100mg of lyophilized venom from the same individual that the BAC library was generated from was sent to the Mass Spectrometry/Proteomics facility at the University of Wisconsin Biotechnology Center (Madison, WI) for trypsin digestion and mass spectrometry. Peptides were analyzed by nanoLC-MS/MS using the Agilent 1100 nanoflow system (Agilent) connected to a hybrid linear ion trap-orbitrap mass spectrometer (LTQ-Orbitrap Elite™, Thermo Fisher Scientific) equipped with an EASY-Spray™ electrospray source. Peptide identification was performed with Mascot (2.2.07). We used BLAST (blastp) with the list of peptides against our C. atrox SVMP proteins to identify peptides that match with 100% identity across the full length of the peptide.
Phylogenetic analysis. Sequences used in our protein phylogenies were deduced from hypothetical translation of genomic regions or from the UniProt and NCBI databases (see full list of sequences and accession numbers in supplement). To construct protein phylogenies we first performed multiple protein sequence alignment with the program MUSCLE (14). Phylogenetic trees were then inferred using maximum likelihood and the Dayhoff substitution model with the program PhyML (v3.3) (15). Bootstrap analysis to assess node support was based on 100 replicate trees. Trees were then formatted with FigTree (v1.4.4; http://tree.bio.ed.ac.uk/software/Figuretree/).

Genomic alignments. BLAST (blastn) was used to align full SVMP complexes from C. atrox and C. scutulatus. Homology blocks were made by scanning along each SVMP complex for contiguous alignments > 1kb and mapping the best match on the other SVMP complex. In most cases there is a clear, single best hit with an identity greater than 95%. Given the duplicated sequences within the complex, it is possible to recover a large number of paralogous segments as hits. This is particularly the case at a threshold of match identity below 90%. Sequence alignments were visualized with Seaview (v4.7) (16).
**Figure S1. BAC assembly of the gene complex.**

25 BAC clones were sequenced and assembled. Above the MP complex are the 15 overlapping BAC clones we used to assemble a large contiguous 1.3 MB complex. The contig labeled Multipool_006 represents three BAC clones that were pooled and sequenced together. Colored bars and shading show where clones overlap. Sequence identity in overlapping regions was >99.7% except for homopolymer indels, which are a known error in PacBio assemblies, and in the two areas marked by red bars with yellow highlighting. In these two regions nucleotide identity falls to 98%, which we have assumed to represent an allelic difference in the two BACs given that the adjacent regions are stretches of 100% sequence identity. Purple droplet represents the single break in our complex. The break occurs between mdc-8a and mdc-6e (both of which are incomplete), between a likely segmental duplication of three genes (mad-5a:mpf-1a:mdc-8a/ mad-5b:mpf-1b:mdc-8b). We consider these sequences to be part of the same complex and not allelic based upon the following criteria: 1) The gene content differs with an orphaned exon and the mad-6 gene located between mad-5a and mpf-1a, and a forward orientated mdc-6e gene 5' of mad-5b. 2) The sequence identity across the homologous stretches of these two segments is lower than regions we consider allelic (~ 95% compared with > 98%, see also Dataset S6). 3) The synteny of matched sequence blocks is not well conserved.
Figure S2. SVMP loci fusions create genes of hybrid origin.
Same phylogeny as Figure 2, with arrows that connect different branches on the tree to indicate genes with a potential hybrid origin. Evidence derives both from genomic alignment (see Figures 5, 6, S4) and protein alignment in which some exons have a higher sequence identity with proteins from other paralog groups than more closely related genes (see Table S2). The C. atrox mpf genes have high homology to the mdc-8 genes except from exon 14 to exon 17 (orange arrow). C. scutulatus mdc-6b is more similar to mdc-5 genes from other species than it is to other mdc-6 genes in exons 1-5 (green arrow). C. scutulatus mad-2 shares higher similarity with mdc-3 genes in exons 1-6 than other mad-2 homologs (teal arrow). mdc-4 and mad-6 genes share highly homologous exons 12 and 13 (blue arrow). C. horridus mpf-2 exons 1-5 derive from a mad-4 paralog (yellow arrow). See Dataset S2 for supporting alignment evidence.
**Figure S3. The ΔCys-4k deletion is shared by all mad (P-II) and mpo (p-I) genes.** Sequence alignment of C. atrox MP genes and a C. durissus mpo-1 gene. Top is a schematic of a typical P-III MP with boxes representing exons 14 and 16. The nucleotide sequences of those exons are shown below. The ~1.9kb stretch from intron 14 to intron 15 is abbreviated and the corresponding sequence is not shown. In mad, mpf, and mpo gene regions a ~26bp fragment replaces part of exons 14 and 16 and the 1.9kb segment between them. The bottom schematic shows a typical P-II MP for which the transcript runs into the inserted sequence that contains a stop codon. The P-II proteins end with the amino acid pair YG or HA. The shared sequences and structures among mad genes indicate that the P-II MPs share a common origin, and the shared sequences of C. atrox mpo-1 and C. durissus mpo-1 indicate that these P-I MPs share a common origin.
small deletions shared by mpf-1 and mdc-8

mad-specific deletions not found at mpf-1

Figure S4. mpf-1 and mdc-8 genes share genomic features.
Sequence alignment of portions of the 12th intron of each C. atrox MP gene. Two sequence fragments are aligned to show 6 small deletions (blue boxes) that are shared and unique between mpf-1a/b and mdc-8b/c. A third genomic fragment shows that mpf-1a/b do not share a 8nt deletion (orange box) that is unique to and shared by all mad genes with the exception of mad-6. These genomic landmarks highlight the clear genomic homology between mpf-1 and the mdc genes 5' to exon 14 and ΔCys-4k.
Dataset S1. List of protein sequences used in phylogeny (Fig. 2)

Dataset lists proteins by gene name as well as species names, source material, appropriate references as well as Accession numbers and the full protein sequence. Genomic sequence data from previous studies is indicated in the source column, all annotations and subsequently derived protein sequences from that data is indicated as being from this study in the reference column.

References
Dataset S1
Hite’94 (1)
Chiou’92 (2)
Scarborough ’93 (3)
Masuda ’08 (4)
Dowell ’18 (5)
Howard ’00 (6)
Vonk ’13 (7)
Guo ’07 (8)
Soto ’07 (9)
Jia’10 (10)
Amemiya’13 (11)

Dataset S2. Exon-exon blast evidence of hybrid origins

Exon-exon protein BLAST (blastp) alignment to support a hybrid origin of some MP genes. Exons from all MPs were aligned against each other. A limited list for four genes against a few relevant genes is shown here as a grid of percent sequence identities. Green highlights reflect the higher identity which switches from one gene to another. Exon identity is used to inform arrows found in SI Appendix, Fig. S2, particularly for mdc-4 which share highly exons 12 and 13 with mad-6.

Dataset S3. Uniquely mapping peptides from C. atrox venom to genes of the MP complex.

The list of 31 C. atrox MP genes (query_id) was aligned against a list of peptides (subject_ids) isolated from lyophilized whole venom from the same individual snake from which the BAC library was generated. Blast results were limited to sequences that match
with 100% identity over 100% of the peptide length and then tallied for the number of genes with which a given peptide sequence is matched (count). Peptide sequences that were found to match a single gene under these criteria are presented here and serve as the basis for Fig. 1B. In addition to unique matches, peptide sequences that match exactly two genes, where those two genes are a gene pair (e.g. MAD-5a/b) are also included. We’ve also included 4 peptides that match with 100% identity to MDC-3b across 29/30 or 30/31 amino acids, mismatched only at the n-terminal amino acid serine. The next best match to this peptide is 83% (22/31).

**Dataset S4. Summary of unique peptide matches.**

Totals of unique peptides per gene from the results in Dataset S3.

**Dataset S5. Alignment results to support BAC assembly.**

Limited results of BLAST alignments of BAC clones to show high sequence identity in overlaps. The majority of all differences are in the form of 1 nt indels in homopolymer stretches. There are two limited regions of lower identity (AO1 and AO2) that contain nearly all of the mismatch differences between BAC 36c20/64j14 and BAC 128n1/25j10. Refer to *SI Appendix*, Fig. S1 for schematic representation of BAC arrangement.

**Dataset S6. Alignment of BACs flanking gap.**

Overlap of BAC 92e13 and 117m6 shows lower identity and more discontinuous blocks of homology than we would expect for allelic regions.

**Dataset S7. Alignment results to support BAC assembly.**

List of primers used for screening the *C. atrox* BAC library.
SI References

1. L. A. Hite, L. G. Jia, J. B. Bjarnason, J. W. Fox, cDNA Sequences for Four Snake Venom Metalloproteinases: Structure, Classification, and Their Relationship to Mammalian Reproductive Proteins. *Arch. Biochem. Biophys.* (1994) https://doi.org/10.1006/abbi.1994.1026.

2. S. H. Chiou, C. C. Hung, K. F. Huang, Characterization of a protease with alpha- and beta-fibrinogenase activity from the Western diamondback rattlesnake, Crotalus atrox. *Biochem. Biophys. Res. Commun.* 187, 389–96 (1992).

3. R. M. Scarborough, *et al.*, Characterization of the integrin specificities of disintegrins isolated from American pit viper venoms. *J. Biol. Chem.* 268, 1058–1065 (1993).

4. S. Masuda, H. Hayashi, S. Araki, Two vascular apoptosis-inducing proteins from snake venom are members of the metalloprotease/disintegrin family. *Eur. J. Biochem.* 253, 36–41 (1998).

5. N. L. Dowell, *et al.*, Extremely Divergent Haplotypes in Two Toxin Gene Complexes Encode Alternative Venom Types within Rattlesnake Species. *Curr. Biol.* 28, 1016-1026.e4 (2018).

6. L. Howard, R. A. Maciewicz, C. P. Blobel, Cloning and characterization of ADAM28: evidence for autocatalytic pro-domain removal and for cell surface localization of mature ADAM28. *Biochem. J.* 348 Pt 1, 21–7 (2000).

7. F. J. Vonk, *et al.*, 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 (2013).

8. X. X. Guo, L. Zeng, W. H. Lee, Y. Zhang, Y. Jin, Isolation and cloning of a metalloproteinase from king cobra snake venom. *Toxicon* 49, 954–965 (2007).
9. J. G. Soto, et al., Molecular evolution of PIII-SVMP and RGD disintegrin genes from the genus Crotalus. *Gene* **389**, 66–72 (2007).

10. Y. Jia, J. C. Pérez, Molecular cloning and characterization of cDNAs encoding metalloproteinases from snake venom glands. *Toxicon* **55**, 462–9 (2010).

11. C. T. Amemiya, et al., The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).

12. D. R. Rokyta, A. R. Lemmon, M. J. Margres, K. Aronow, The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). *BMC Genomics* **13** (2012).

13. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinforma. Impact Accurate Quantif. Proteomic Genet. Anal. Res.*, **41**–74 (2014).

14. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

15. S. Guindon, et al., New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

16. M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–4 (2010).