Current Biology, Volume *28*

Supplemental Information

Extremely Divergent Haplotypes in Two Toxin Gene Complexes Encode Alternative Venom Types within Rattlesnake Species

Noah L. Dowell, Matt W. Giorgianni, Sam Griffin, Victoria A. Kassner, Jane E. Selegue, Elda E. Sanchez, and Sean B. Carroll

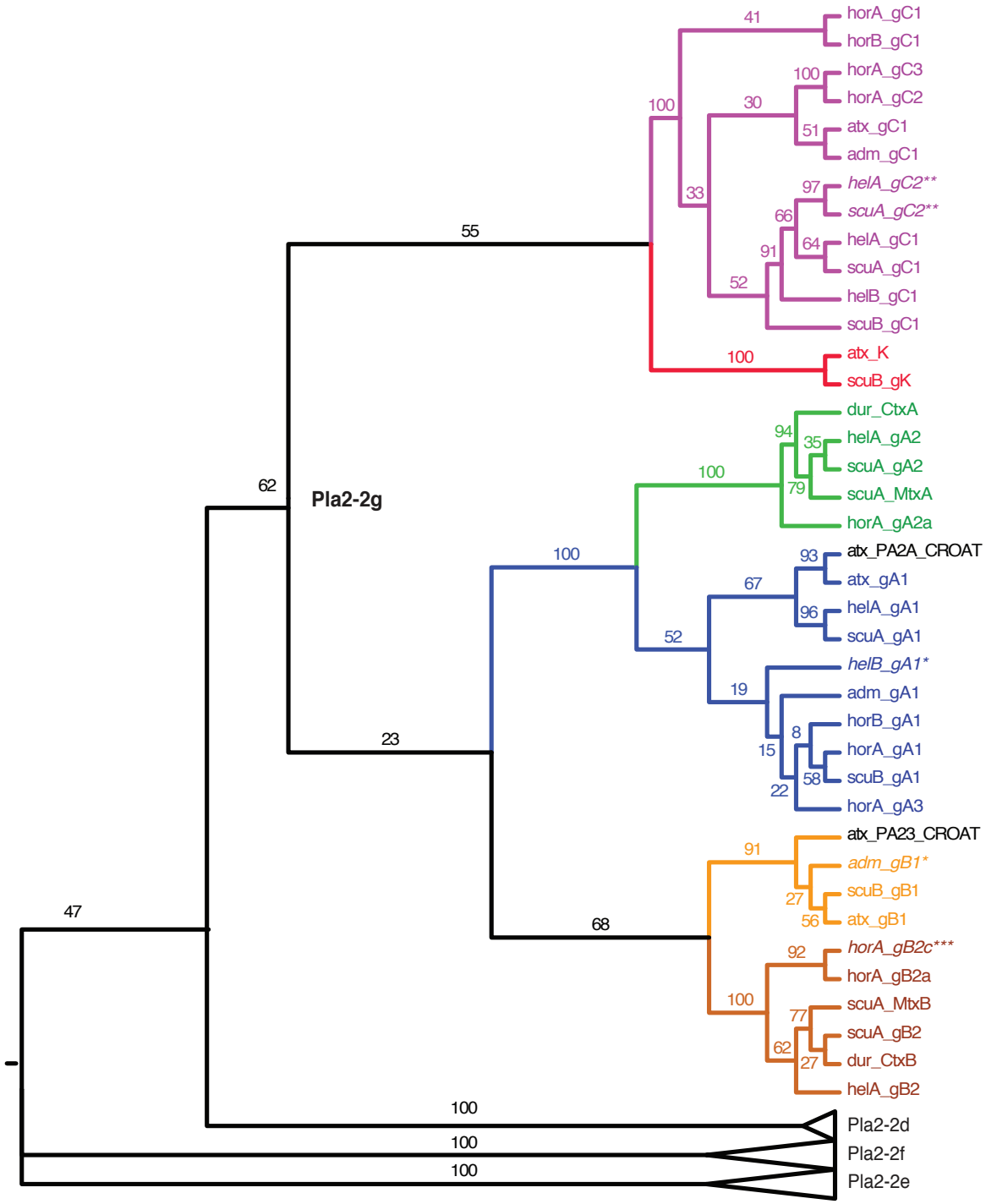


Figure S1. Phylogenetic analysis of PLA2 proteins. Related to Figures 2 and 3.

Protein phylogeny of full-length Pla2 proteins was used to identify gene orthologs in PLA2 complexes. The figure shows the gene members of the *C. scutulatus B* type complex, *Pla2-gB1* (scuB-gB1), *Pla2-gC1* (scuB-gC1) and *Pla2-gK* (scuB-gK) group with *C. atrox* (atx-gB1, -gC1,-gK) and *C. adamanteus* (adm-gC1, adm-gB1*; with the asterisk identifying putative pseudo-genes.) Note the *C. scutulatus* type B complex does not contain genes encoding the neurotoxin heterodimer that are present in the *C. scutulatus* type A complex found in this study (scuA_gB2 and scuA_-gA2) and identified previously (scuA_MtxB and scuA_MtxA and in *C. durissus terrificus*: dur_CtxB and dur_CtxA). Mutations in various genes suggest these are pseudogenes (italics) as they would result in untranslated or truncated products. In these instances, the exonic structure remains intact with some viable sequence information that was used to reconstruct the phylogenetic relationships of genes in the complex. Single asterisk (*): adm gB1 has a single nucleotide deletion in the second exon which would result in a truncated protein. We have “restored” the deletion to more accurately establish the phylogenetic relationship of the pseudogene. Double asterisk (**): scu gC2 and helA gC2 are missing the first exon. Triple asterisk (***): horA gB2c has a deletion which removes the first three amino acids, including the start codon. Quadruple asterisk (****): helB gA1 contains two frameshift mutations in the second exon which would yield a truncated protein. We have used sequence information from the three intact exons. Bootstrap values from 1,000 replicate trees are indicated on branches. Species included: *C. atrox* (atx for proteins presented in this study and database proteins: atx_PA2A_CROAT atx_PA23_CROAT), *C. scutulatus* A and B types (scuA & scuB), *C. helleri* A and B types (helA & helB), *C. adamanteus* (adm), *C. horridus* A and B types (horA & horB).

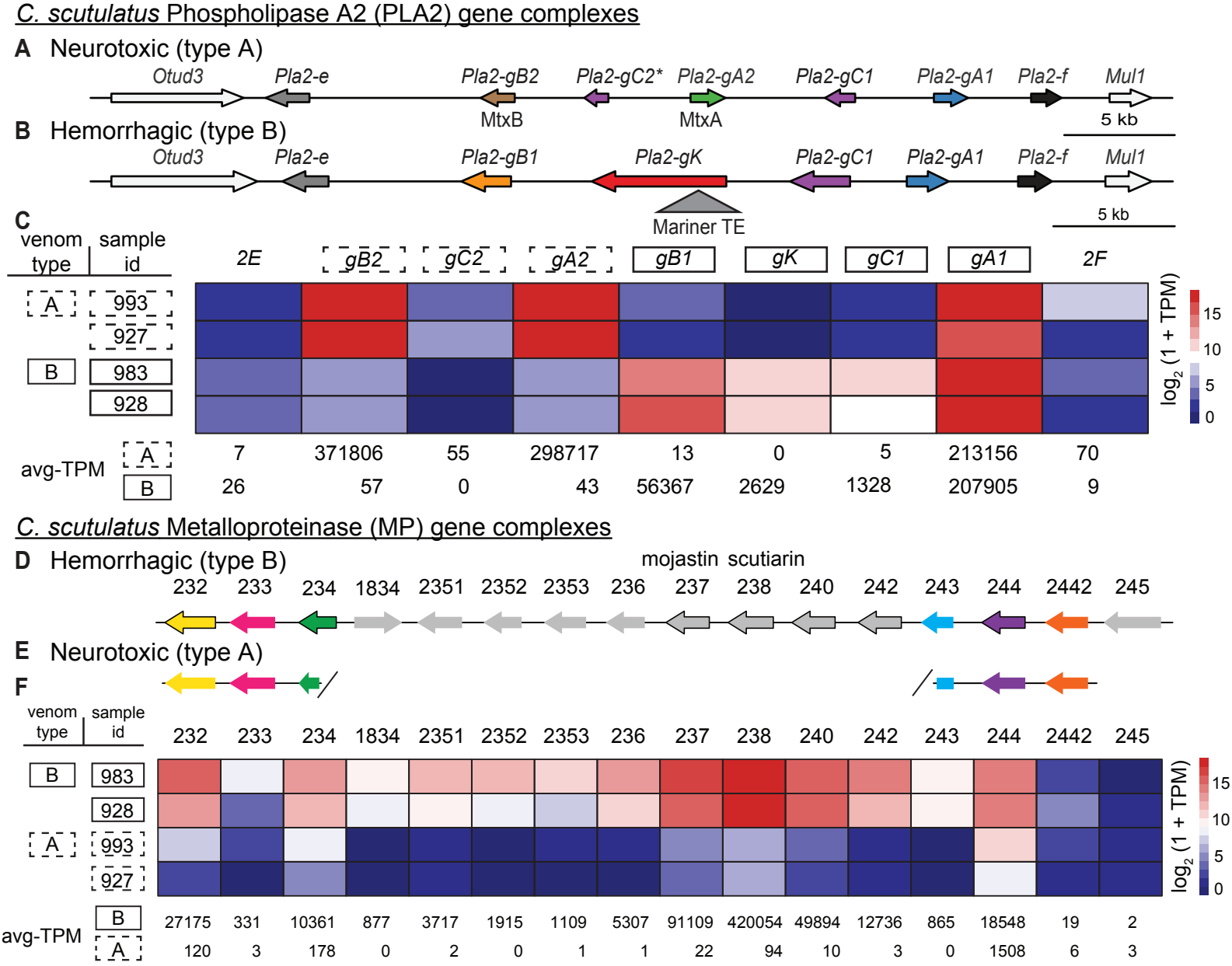


Figure S2. Venom gland gene expression at the PLA2 and MP complexes in *C. scutulatus*. Related to Figure 2.

(A) The *C. scutulatus* neurotoxic (type A) PLA2 gene complex. (B) The *C. scutulatus* hemorrhagic (type B) PLA2 gene complex. (C) A heatmap showing the relative level of venom gland gene expression (log base 2 transformed read counts) of the *Pla2* genes, with highly expressed genes in red and lowly expressed genes in blue. Neurotoxic (type A) specimens (993 and 927) and neurotoxin subunit genes are enclosed in dashed boxes. Hemorrhagic (type B) specimens (928 and 983) and genes are enclosed by solid boxes. Columns in the heatmap correspond to single a gene and are ordered by genomic location in the complex. The read counts (average transcripts per million; avg-TPM) for each venom type are listed below the heatmap. The neurotoxin-encoding genes (*Pla2-gB2* and *Pla2-gA2*) are highly expressed in type A, but not type B specimens. *Pla2-gB1* and *Pla2-gK* genes are expressed in animals with type B venom, however *Pla2-gK* expression is low relative to most venom *Pla2* genes. Pla2-gK protein has not been identified in previous studies of *C. scutulatus* type B venom composition. *Pla2-gA1* is highly expressed in both type A and type B animals. (D) The *C. scutulatus* hemorrhagic (type B) metalloproteinase (MP) gene complex. Medium to highly expressed genes are highlighted with a black outline of the gene arrow. (E) The *C. scutulatus* neurotoxic (type A) MP gene complex. The borders of a deleted segment are depicted as a forward slash. (F) A heatmap showing expression of MP complex genes in the venom gland (log base 2 transformed read counts). Columns correspond to a single gene and are ordered according the genomic position within the complex. The rows are individual specimens expressing either type B (solid boxes) or type A (dashed boxes) venom. Mean TPM values are shown below the heatmap for each venom type. The most highly expressed genes, *238*, *237* and *240*, with TPMs of 420054, 91109 and 49894, respectively, are located in the middle of the type B complex. Read counts for expression analysis were obtained by mapping venom gland reads to the annotated type B MP complex. Accurately mapping reads to highly duplicated regions of genomes is challenging and the type A TPM values for genes that are absent from the neurotoxic genome are likely artificially high due to the mis-mapping of reads originating from similar genes.

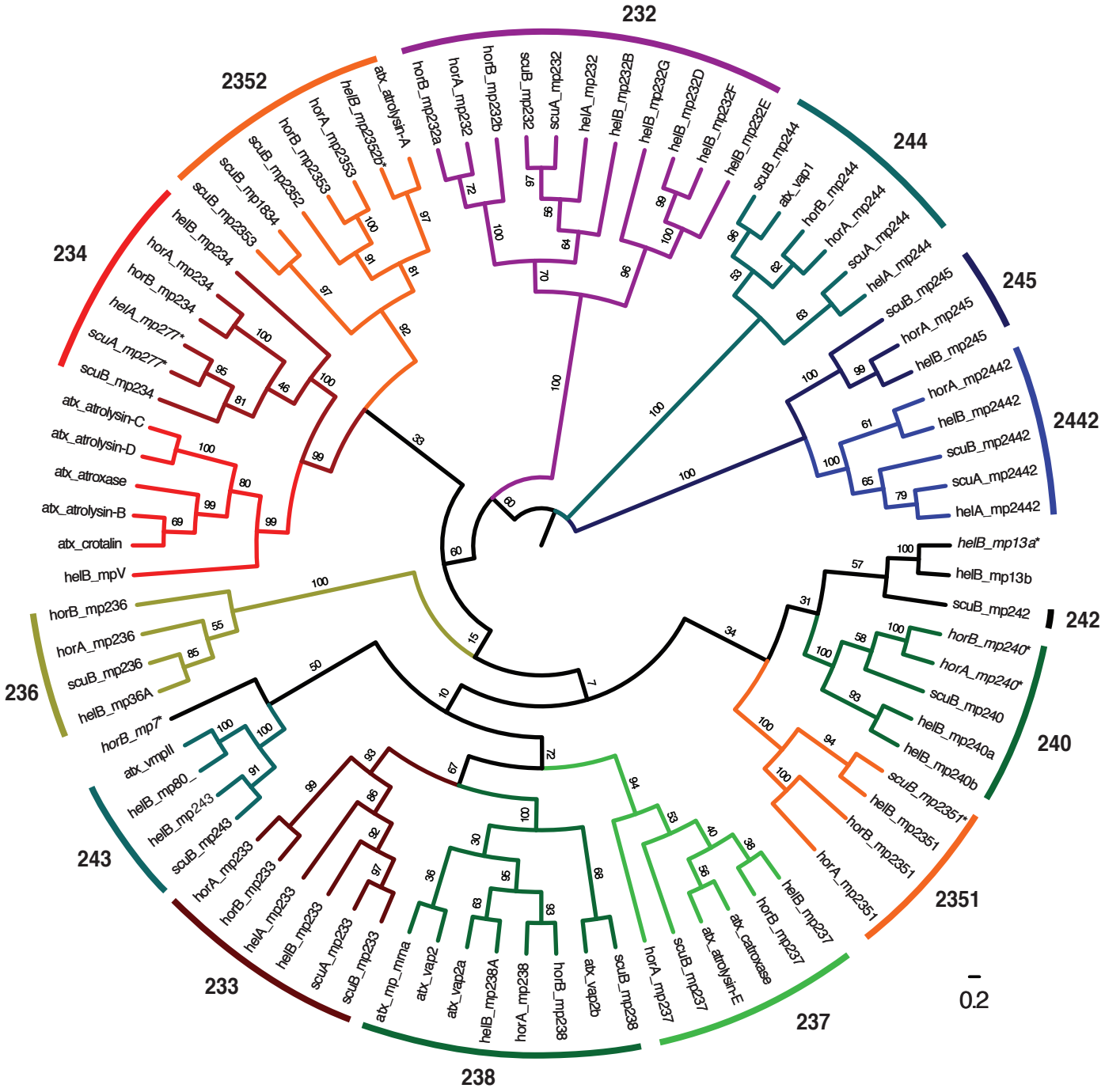


Figure S3. Phylogenetic analysis of metalloproteinase (MP) domains. Related to Figures 2 and 4. Protein phylogeny of the metalloproteinase domains (exons 7-12) was used to ascertain the orthologous genes in each species complex. We have also included venom expressed *C. atrox* proteins available in sequence databases (clustering with 234). Large gaps in the colored arcs on the perimeter highlight lineage specific MP proteins that cluster poorly with the *C. scutulatus* MP protein groups (helB_mp13a*, helB_mp13b, horB_mp7* and helB_mp-V). In several instances there are mutations in what appear to be pseudogenes (italics) as they would result in untranslated or truncated products. In these instances, where the exonic structure remains intact with viable sequence information, we use that sequence to reconstruct the phylogenetic relationships of genes and likely pseudogenes in the complex. Asterisk (*): helB mp2352b has a stop codon in exon 6; helB mp242a has three indels resulting in a frameshift mutation; horA and horB mp240 have deletions of exons 2-5; scuB mp2351 has a deletion of exons 13 and 14; horB mp7 has a small deletion in exon 10 which results in an early stop codon; scuA and helA mp277 is the fusion of mp234 and mp243 created by the genomic deletion and scuA and helA mp277 have deletions of exons 1 and 2. Note that the mp277 metalloproteinase domain clusters with members of 234 but the N terminal exons cluster with 243 (not shown) as expected from sequence alignment (see figure 4). Bootstrap values from 1,000 replicate trees are indicated on branches. Species included: *C. atrox* (atx), *C. scutulatus* A and B types (scuA & scuB), *C. helleri* A and B types (helA & helB), *C. adamanteus* (adm), *C. horridus* A and B types (horA & horB).

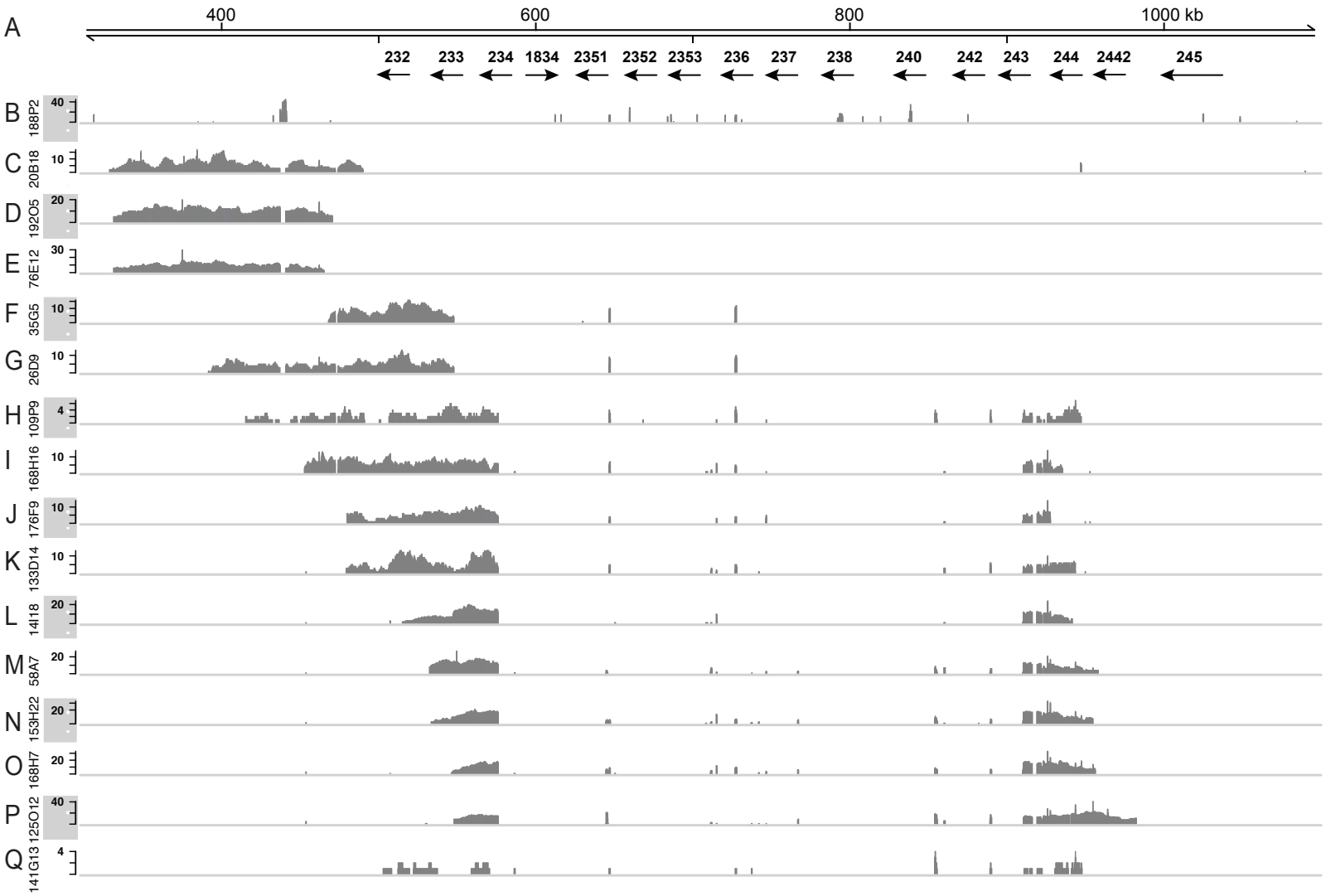


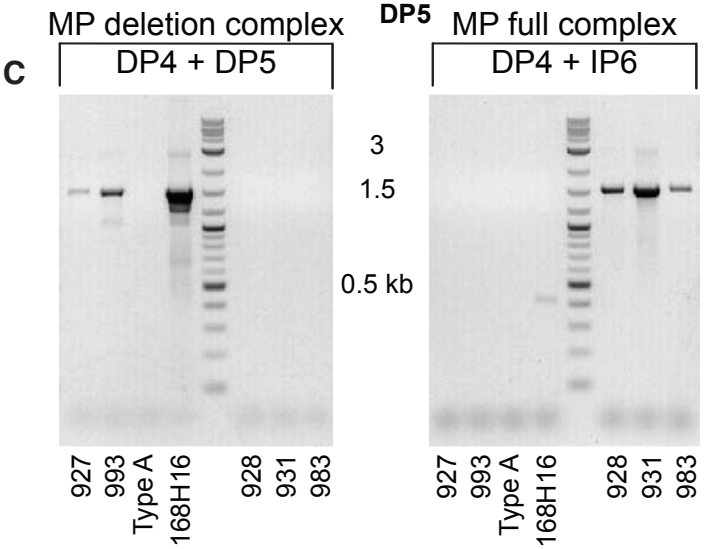
Figure S4. Alignment of individual *C. scutulatus* type A BAC clones of the MP complex. Related to Figure 2.

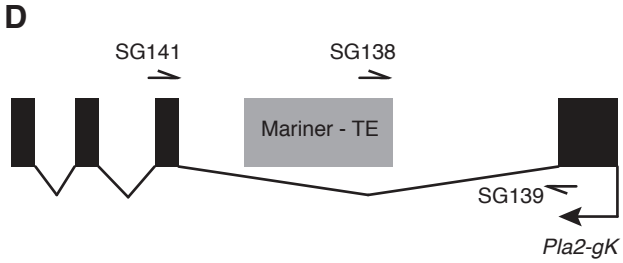
Numerous clones were isolated and sequenced to confirm the large deletion in type A individuals relative to type B individuals, and to determine whether the deleted sequence had been translocated to another site in the genome. (A) The relative positions and orientations (arrowheads) of the genes in the *C. scutulatus* type B MP complex. (B - Q) Sixteen individual *C. scutulatus* type A BAC clones were independently sequenced and assembled. Corrected reads were aligned to the type B MP complex and read coverage (grey histograms) for each unique clone is shown in each corresponding track. Single clone identifiers are on the y-axis. Note the sparse read coverage across the middle (> 300 kb) of the *C. scutulatus* type B MP gene complex. Spurious coverage peaks in this region are likely due to alignment of short reads to repetitive sequences. Ten unique clones (H - Q) have reads aligning to both sides of the genomic segment that spans from gene 234 to 243 thus offering conclusive evidence that the intervening genomic sequence has been deleted from the *C. scutulatus* type A haplotype. Therefore the ten clones (H – Q) *C. scutulatus* type A SVMP BAC clones that span the SVMP gene complex all contain the deletion haplotype.

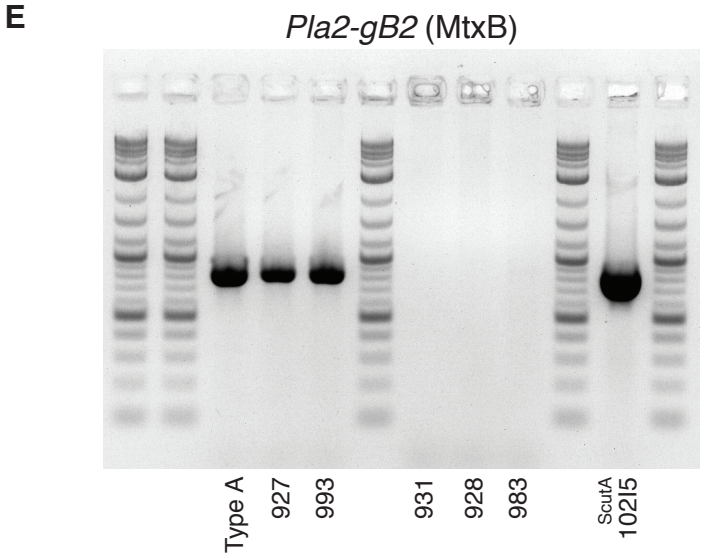
B

predicted product size (bp) for each complex

	full	deletion
DP4 + DP5	> 330000	1504
DP4 + IP6	1549	0







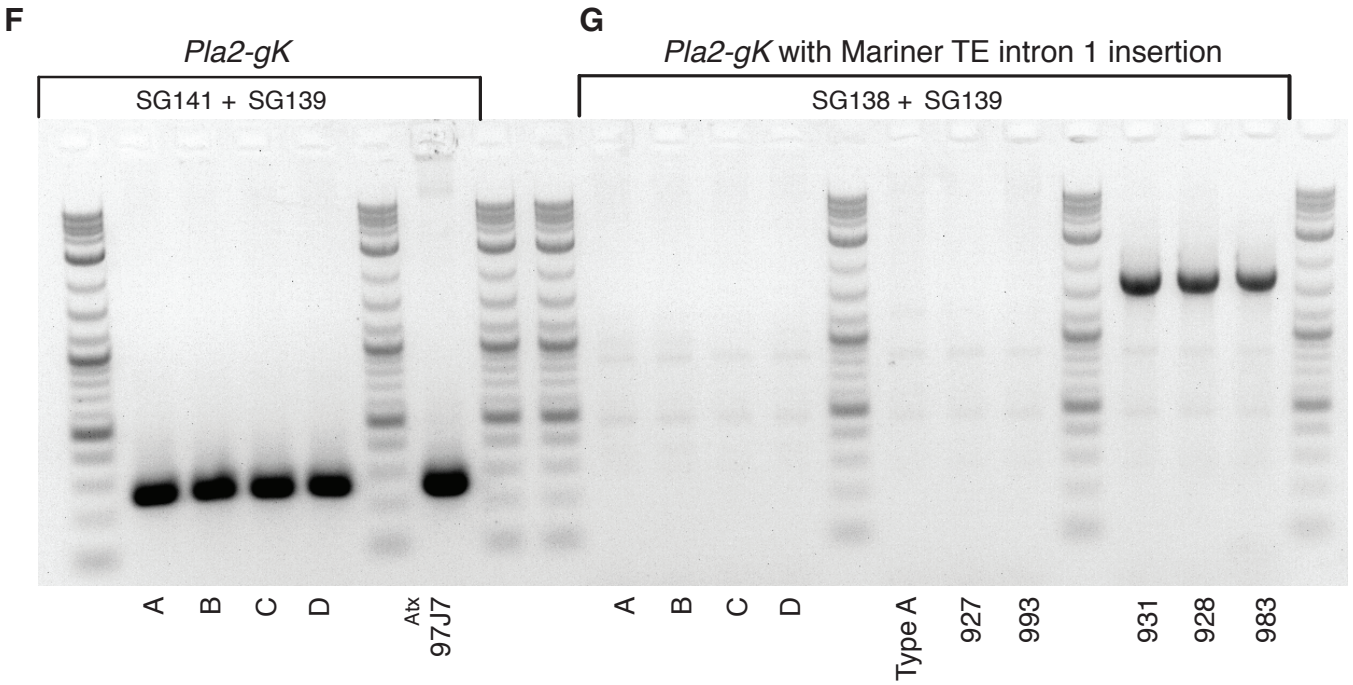


Figure S5. Genotyping *C. scutulatus* individuals for Pla2 and metalloproteinase (MP) haplotypes. Related to Figures 2 and 4.

(A) PCR primers (DP4, DP5, IP6; triangles) were designed to the *C. scutulatus* type B MP complex such that presence or absence of products indicates the full or deleted MP complex. (B) The combination of DP4 and IP6 yields a product of 1549 base pairs (bp) on the *C. scutulatus* type B full complex but yields no product when using genomic template from type A specimens because IP6 is designed for sequence adjacent to the deletion boundary (MP full complex gel). MP clone 168H16 is one of the ten deletion spanning clones isolated from the *C. scutulatus* type A whole genome BAC library (Figure S4-I). If the entire genomic region between *MP 234* and *MP 243* is deleted (as observed in clone 168H16) then primers DP4 and DP5 are predicted to yield a product of 1504 bp. (C) Using genomic template from *C. scutulatus* type A specimens 927 and 993 amplifies the DP4/DP5 1504 bp product. The "Type A" specimen genomic DNA performed poorly in PCR assays but the whole genomic type A BAC library (e.g. clone 168H16) was created using the DNA from this particular specimen. (D) To test for the presence/absence of *Pla2-gK* with/without the Mariner-TE inserted into intron one, PCR primers were designed to exons 1 and 2 (half arrows, SG139 and SG141, respectively) and to the Mariner TE (SG138). PCR primers were also designed to the *C. scutulatus Pla2-gB2* (MtxB) gene to test for the presence/absence of the neurotoxin haplotype. (E) All type A specimens (Type A, 927, 993) have the *Pla2-gB2* gene but all of the type B specimens (931, 928, 983) lack the *Pla2-gB2* gene. (F) The presence of *Pla2-gK* (without an insertion in intron 1) in four *C. atrox* specimens (A – D labels) is shown along with a BAC clone (91J7) known to contain *Pla2-gK*. (G) The *Pla2-gK* with a Mariner TE insertion is present in all three *C. scutulatus* type B specimens (931, 928, 983) but absent in type A specimens (Type A, 927, 993) and *C. atrox* specimens (A – D). For the specimens in this study there is no evidence of heterozygotes that possess the neurotoxic haplotype and the *Pla2-gK* haplotype. The 2-log DNA ladder is shown on all gels.

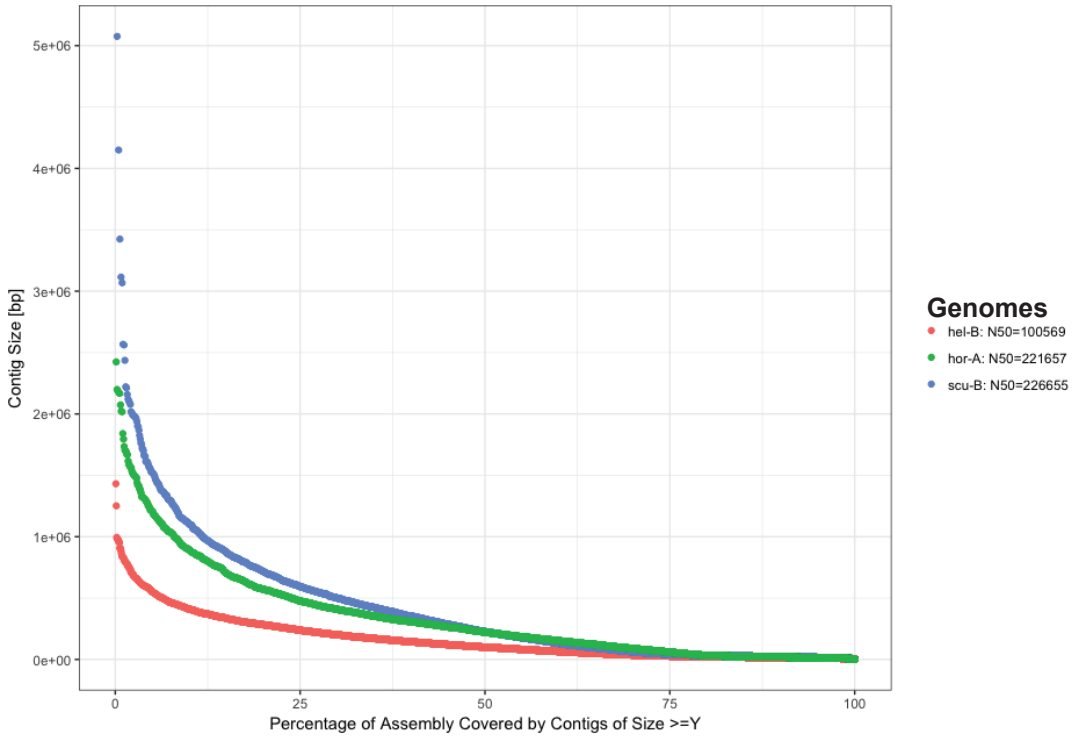


Figure S6. The contig lengths for three raw genome assemblies. Related to Figures 2 – 4.
The contig lengths (y-axis) that correspond to the percentage of the genome (x-axis) that is covered by contigs of that size are shown for three rattlesnake genome assemblies. *C. helleri* type B (red dots), *C. scutulatus* type B (blue dots) and *C. horridus* type A (green dots). The shortest contig lengths in base pairs (bp) covering 50% of the genome (N50) are 100569 for *C. helleri* (hel-B), 226655 for *C. scutulatus* (scu-B) and 221657 for *C. horridus* (hor-A). The N50 statistic reports the quality of an assembly in terms of contiguity.

Species	Specimen ID	Location	Venom type A = Neurotoxic B = Hemorrhagic
C. scutulatus	NA	Brewster Co. TX	A
C. scutulatus	927	San Bernardino Co., CA	A
C. scutulatus	993	Cochise Co., AZ	A
C. scutulatus	928	Pima Co., AZ	B
C. scutulatus	931	Pima Co., AZ	B
C. scutulatus	983	Pima Co., AZ	B
C. helleri	677	Riverside Co., CA	A
C. helleri	789	Born at NNTRC mother was from Riverside Co. CA (also a type B)	B
C. horridus	CH-A090108-06	Jasper Co., SC (parent origin)	A
C. horridus	CHO114KY	Floyd Co., KY	B

Table S1. Rattlesnake specimen location and venom type. Related to Figure 1.

C. scutulatus hemorrhagic (type B) and *C. helleri* neurotoxic (type A) specimens were identified by using the median lethal dose (LD_{50}) of venom on BALB/C mice. LD_{50} values were used to classify individuals as neurotoxic (type A) if the LD_{50} was between 0.3 - 0.9 mg/kg or as hemorrhagic (type B) if the LD_{50} was > 3.0 mg/kg [13]. The LD_{50} values for venom from the *C. scutulatus* type B specimens 928, 931, 983 were 3.25, 3.95, 4.7 mg/kg, respectively. The LD_{50} values for the *C. oreganus helleri* type A and B specimens were 0.56 and 3.0 mg/kg, respectively.

Species	Venom type	Seq library tissue	PacBio RS II sequencing material	Venom complex	Clones	Raw data accession
C. scutulatus	Neurotoxic (A)	Blood	BAC	PLA2	102I5	SRR3478362
C. scutulatus	Neurotoxic (A)	Blood	BAC	SVMP	125O12 133D14	SRR5858077 SRR5858076
C. helleri	Neurotoxic (A)	Blood	BAC	PLA2	85A17	SRR5858079
C. helleri	Neurotoxic (A)	Blood	BAC	SVMP	46A6 136F17	SRR5858078 SRR5858071
C. horridus	Hemorrhagic (B)	Blood	BAC	PLA2	190H17	SRR5858070
C. horridus	Hemorrhagic (B)	Blood	BAC	SVMP	27I3 145P9 132M5	SRR5858069 SRR5858068 SRR5858067
C. scutulatus	Hemorrhagic (B)	Blood	Whole genome	PLA2, SVMP	N/A	SRR6410430
C. helleri	Hemorrhagic (B)	Blood	Whole genome	PLA2, SVMP	N/A	SRR6410429
C. horridus	Neurotoxic (A)	Blood	Whole genome	PLA2, SVMP	N/A	SRR6410431

Table S2. Genome sequence data types assembled and annotated to characterize the PLA2 and SVMP gene complexes in three species of rattlesnake. Related to Figures 2 – 4.

Whole genome bacterial artificial chromosome (BAC) libraries were constructed from *C. scutulatus* type A, *C. helleri* type A and *C. horridus* type B genomic DNA that was isolated from whole blood. Clones for specific venom gene complexes (PLA2 or SVMP) were isolated from the library and sequenced using Pacific Biosciences technology (PacBio RS II). The raw sequencing data for each clone are available at the Short Read Archive database (Raw data accession numbers). Additionally, genomic DNA was extracted from the blood of *C. scutulatus* type B, *C. helleri* type B and *C. horridus* type A specimens and sequenced using Pacific Biosciences technology (PacBio RS II). Contigs containing the PLA2 and SVMP gene complexes were extracted from the raw assemblies and annotated. The raw data (reads) used to generate the assemblies are available at the Short Read Archive database (Raw data accession numbers).

Species	Venom	# raw reads	Raw seq. (bp)	Est. coverage (1.8 – 1.4 G bp genome)	# contigs	Contig N50 (kb)	Median contig length (kb)	Raw assembly length
C. scutulatus	Hemorrhagic (B)	4474901	47917343033	26 – 34X	32065	227	23.9	2030922379
C. helleri	Hemorrhagic (B)	9607654	66869924141	37 – 47X	45034	101	18.1	1878780749
C. horridus	Neurotoxic (A)	5322653	57746831678	32 – 41X	31766	222	19.2	1875843582

Table S3. Summary statistics for three rattlesnake genome assemblies. Related to Figures 2 – 4.
Genomic DNA was extracted from whole blood and large insert sequencing libraries constructed (see methods for details). The libraries were sequenced until ~ 35X genome coverage was obtained. The Crotalus genome size is estimated to be 1.4 to 1.8 Giga base pairs (G bp). The raw reads were error corrected and assembled (see methods) to yield contigs. The contigs containing the PLA2 and SVMP gene complexes were annotated in this study. Whole genome analysis is currently in progress.

species	contig ID	gene_IDs	gene status	exons_for_partial_genes	MEPRO domain
C.helleri-B	58718	MP232F	partial	MP232F:ex 6 - 11	yes
C.helleri-B	58940	MP232A	partial	MP232A:ex 14 -17	no
C.helleri-B	61437	MP232B	partial	MP232B:ex 9 - 11	yes
C.helleri-B	84783	MP232C	partial	MP232C:ex 12 -17	no
C.helleri-B	87611	MP232E	partial	MP232E:ex 6 - 12	yes
C.helleri-B	87610	MP232D	partial	MP232D:ex 1 - 10	yes
C.helleri-B	87610	MP232H	partial	MP232H:ex 12 -17	no
C.helleri-B	96453	MP232G	partial	MP232G:ex 9 - 13	yes
C.helleri-B	90179	MP233A	partial	MP233A:ex 5 - 14,17	yes
C.helleri-B	67486	MP233B	partial	MP233B:ex 1 - 2	no
C.helleri-B	67486	MP234	partial	MP234:ex 3 - 17	yes
C.helleri-B	64313	MP1834	partial	MP1834:ex 15 - 17	no
C.helleri-B	64313	MP2351	full	MP2351:ex 1 - 17	yes
C.helleri-B	64313	MP2352A	partial	MP2352A:ex 11 -17	no
C.helleri-B	68240	MP2352B	partial	MP2352B:ex 7 - 17	yes*
C.helleri-B	66196	MP-7	partial	MP-F:ex 12 - 17	no
C.helleri-B	66196	MP2352C	partial	MP2352C:ex 1	no
C.helleri-B	94164	MP-10m	partial	MP-10m:ex 9 - 11	no
C.helleri-B	94164	MP-Fm	partial	MP-Fm:ex 7 - 8	no
C.helleri-B	70926	MP236	partial	MP236:ex 9 - 14	yes
C.helleri-B	71561	MP237	partial	MP237:ex 11 - 14, 17	yes
C.helleri-B	70695	MP238A	partial	MP238A:ex 9 -15	yes
C.helleri-B	72085	MP238B	partial	MP238B:ex 1 - 6	no
C.helleri-B	60881	MP238C	partial	MP238C:ex 1	no
C.helleri-B	87327	MP-V	full		yes
C.helleri-B	87327	MP240A	partial	MP240A:ex 10 - 17	yes
C.helleri-B	60407	MP240B	partial	MP240B:ex 1 - 14	yes
C.helleri-B	60407	MP243	full		yes
C.helleri-B	60407	MP80	full		yes
C.helleri-B	60407	MP13A	partial	MP13A:ex 10 - 17	yes
C.helleri-B	73042	MP243	partial	MP243:ex 1 - 3	no
C.helleri-B	73042	MP244	partial	MP244:ex 15 - 17	no
C.helleri-B	60942	MP13B	partial	MP13B:ex 1 - 14	yes
C.helleri-B	60942	MP2442	full		yes
C.helleri-B	60942	MP245	full		yes

Table S4. Summary of *C. helleri* type B SVMP gene complex annotation. Related to Figure 4. For this study, our annotation and phylogenetic analysis efforts focused on the conserved metalloproteinase domain (exons 7 -12). However, there are some partial MP genes found across the twenty-two contigs. For example, there are three C-terminal exons for 1834 on contig 64313 and

they are adjacent the C-terminal exons of 2351 as they are in *C. scutulatus* type B. There are five *C. helleri* duplicated 232 genes with annotatable metalloproteinase domains but exon evidence for three additional 232 genes giving a total of 8 putative 232 genes. MP-V is a homolog of *C. atrox* atrolysin MP proteins (see Figure S3) and is not present in the *C. scutulatus* MP complex. MP-80 is a homolog of *C. atrox* SVMP II and is not present in the *C. scutulatus* MP complex. MP13a and 13b are homologs of scuB_mp242 but there is uncertainty regarding the exact orthology (see Figure S3).