Supplemental Materials

Table of Contents

Supplemental Methods

Tissue sampling

Venom was manually extracted from three adult Prairie Rattlesnakes (*Crotalus viridis viridis*) one day prior to sacrifice. Venom gland, pancreas, stomach, and/or liver tissue samples were dissected out and snap frozen in liquid nitrogen following humane sacrifice of the individual via deep anesthesia with Isoflurane followed by decapitation (See Supp. Table S1 for additional detail). All sampled animals were collected from the same population in order to control for genetic background in subsequent analyses. All animals were housed and sampled at the University of Northern Colorado under approved and registered IACUC protocols.

Refining venom gene annotations

During early exploratory analysis of ChIP-seq and ATAC-seq data described below, we noticed patterns of elevated read mapping density (i.e., ChIP-seq and ATAC-seq peaks) in several intergenic regions of the SVSP gene array that closely resembled patterns associated with annotated SVSP genes. Following the approach described in (Schield et al. 2019), we used FGENESH+ (Solovyev 2004) and known guide protein sequences to identify two previously unannotated SVSP genes in these regions, which we include in analyses below. Following the naming convention of the nine originally annotated SVSPs, these new SVSP genes were named SVSP 10 and SVSP 11.

RNA isolation, sequencing and analyses

Total RNA was extracted from snap frozen tissues with Trizol reagent (Invitrogen). All tissues were subsampled to produce three technical replicates. Poly-A selected mRNA libraries were sequenced on an Illumina NovaSeq using 150bp paired-end reads. Raw RNA-seq reads were quality trimmed using default settings with Trimmomatic v0.39 (Bolger et al. 2014). RNAseq reads were mapped to the annotated *Crotalus viridis* genome (NCBI: GCA_003400415.2) using STAR v2.7.3a (Dobin et al. 2013) and raw gene expression counts were estimated using featureCounts v1.6.3 (Liao et al. 2013). Pairwise comparisons between venom and non-venom tissues were conducted using DEseq2 v1.30.1 (Love et al. 2014) in R (R Core Team 2013), with independent hypothesis-weighting p-value correction via the IHW package v1.18.0 (Ignatiadis et al. 2016) using baseMean expression from DEseq2 as the covariate. Annotated venom genes from the *Crotalus viridis* genome publication (Schield et al. 2019) were considered to be “relevant” venom genes if they were found to be significantly upregulated in the venom gland (IHW p-value < 0.05) in pairwise comparisons of venom gland versus non-venom tissues. For subsequent analyses aimed at comparing relative gene expression between genes in the venom gland, we normalized gene expression counts in the three venom gland replicates to transcripts per million (TPM), and used the median TPM measure across replicates for each gene as its expression in subsequent analyses. Visualization of gene expression in the context of venom gene family arrays conducted using the gggenes package v0.4.1 (github.com/wilkox/gggenes) in R.

Hi-C Sequencing and Analysis

Hi-C data for a *Crotalus viridis* venom gland at one day post-extraction was generated previously (see (Schield et al. 2019) for details; NCBI BioProject: PRJNA413201). Raw Illumina paired-end Hi-C reads were mapped to the rattlesnake reference genome using the Juicer pipeline (Durand et al. 2016), and Hi-C contact maps were generated using KR normalization at 10kb and 5kb resolution. Topologically-associated chromatin domains (TADs) and sub-TADs were determined at 10kb resolution using rGMAP v1.3.1 (Yu et al. 2017) in R. Chromatin loops were identified

using the HICCUPS algorithm in Juicer v1.9.9 (Durand et al. 2016) with default settings. Hi-C contact heatmaps were generated using the Sushi package v1.28.0 (Phanstiel et al. 2014) in R.

Chromatin Immunoprecipitation (ChIP) data generation and analysis

ChIP-seq libraries were generated for post-extraction (1DPE) venom gland tissue by Active Motif (Carlsbad, CA) for bound CTCF and histone modifications H3K4me3 and H3K27ac. Basic ChIP-seq data processing was performed by Active Motif using their standard analysis pipeline. In brief, libraries were sequenced on an Illumina NextSeq 500 using 75-nt reads and mapped to the UTA_CroVir_3.0 genome assembly (GCA_003400415.2) using BWA v0.6.1 (Li and Durbin 2009) with default settings. Reads that failed to pass Illumina's purity filter, aligned with greater than 2 mismatches, were not uniquely mapped, or were identified as PCR duplicates were removed for all subsequent analyses. Aligned reads were extended in silico at the 3' end using Active Motif's in-house software, and fragment densities were determined for 32-nt bins across the genome. Intervals of enriched ChIP-seq fragment density were determined using MACS2 v2.1.0 (Zhang et al. 2008). Super enhancers were determined by merging enriched H3K27ac intervals if their inner distance was equal to or less than 12,500 bp, and classifying merged regions with the top 5% strongest enrichment as super enhancers.

To infer sites of bound CTCF and putative CTCF-bound chromatin loops, we used MEME v5.1.1 (Bailey et al. 2009) to reconstruct the Prairie Rattlesnake CTCF binding motif (Supp. Fig. S22) from CTCF ChIP-seq peak sequences. We then scanned all CTCF ChIP-seq peaks for this binding motif, and peaks with a binding motif were considered "verified" ChIP-seq peaks. We then used the pairtobed tool in bedtools v2.29.2 (Quinlan and Hall 2010) to intersect chromatin loops with verified CTCF ChIP-seq peaks, and considered a chromatin loop to be CTCF-bound if a verified CTCF ChIP-seq peak was identified within 10kb of both ends of the loop.

ATAC-seq data generation and analysis

ATAC-seq libraries were generated using the same post-extraction venom gland samples used to generate mRNA-seq. ATAC-seq libraries were prepared from snap-frozen venom gland tissue by Active Motif (Carlsbad, CA) and sequenced on an Illumina NextSeq 500 using 42-bp paired-end reads. Reads were then mapped to the UTA_CroVir_3.0 genome assembly (GCA_003400415.2) using BWA v0.7.17 with default settings. PCR duplicates were removed using MarkDuplicates in Picard Tools v2.22.6 (https://broadinstitute.github.io/picard/), and non-unique alignments and improperly paired reads were removed using samtools v1.9 (Li et al. 2009). Samtools flagstat was used to find the sample with the lowest number of properly paired reads, and the randsample tool within MACS2 v2.2.7.1 was then used to randomly downsample the other two BAM files to the same total number of paired reads present in the smallest sample. ATAC-seq peaks were called using MACS2 callpeak with a q-value cutoff of 0.001. A merged set of narrow ATAC-seq peak regions from all samples was generated using bedtools merge, and individual peak files were then intersected with these merged regions to assess the degree of overlapping peaks between samples. A set of ATAC-seq peaks for use in downstream analyses was constructed by taking merged peak regions for which an ATAC-seq peak overlapping that region was present in at least two of the replicate datasets. Raw ATAC-seq read coverage for each sample was quantified with a bin size of 32 and smoothing length of 96 and then converted to a raw count matrix using deepTools v3.1.3 (Ramírez et al. 2016) bamCoverage and multiBigWigSummary, respectively. The EdgeR v3.32.1 package in R was used to calculate size factors for each library, and these size factors were then used as scale factors in deepTools bamCoverage to generate bigwig files of normalized ATAC-seq density for each sample. Wiggletools 1.2.1 (github.com/Ensembl/WiggleTools) was used to calculate the mean normalized ATAC-seq read density across the three replicates.

ATAC-seq footprinting analysis was conducted using TOBIAS v.0.12.4 (Bentsen et al. 2020). For each sample, Tn5 insertion site bias was corrected using TOBIAS ATACorrect and footprint scores were calculated using TOBIAS ScoreBigwig. TOBIAS BINDetect was then used to calculate a footprint score "binding" threshold for each sample with the default p-value of 0.001, defined as the lowest footprint score with which a TFBS was classified as "bound" by the program. These footprint cutoffs were used to assess evidence for TFBS binding as described below.

Identifying candidate transcription factors

We constructed a candidate set of TFs with evidence for activity in the venom gland during venom production through independent analysis of gene expression and ChIP-seq data. To first identify annotated TFs in the Prairie Rattlesnake genome, we downloaded Uniprot (UniProt Consortium 2019) gene lists that were annotated with one or more of the following gene ontology terms: DNA binding transcription factor activity, protein binding transcription factor activity, and transcription factor co-regulatory activity. These gene lists were used to parse the Prairie Rattlesnake gene annotation for known TFs using previously published rattlesnake-to-human orthology tables (Perry et al. 2020). Resulting rattlesnake TFs were cross-referenced with the results of differential gene expression analyses described above to identify TFs with evidence of upregulation in the venom gland compared to non-venom tissues (IHW p-value < 0.05). Separately, we identified TF genes that are associated with super-enhancers (SEs), regions of elevated H3K27ac ChIP-seq read density (described above). A TF gene was considered to be SE-associated if it overlapped with an annotated SE region, or was the nearest gene to a SE if that SE does not otherwise directly overlap with any annotated genes.

These two independently derived candidate TF sets were merged to form one master set of candidate TFs for subsequent analyses. To characterize candidate TFs, we used WebGestalt 2019 (Liao et al. 2019) to identify GO Terms and KEGG Pathways with overrepresentation in our candidate set compared to a background of all TFs annotated in the rattlesnake genome and default parameters. To assess known involvement of our candidate TFs with ERK, a central regulatory molecule within the ERK/MAPK signaling pathway previously implicated in venom gene regulation, we used STRINGdb v11.0 (Szklarczyk et al. 2019) to identify interactions between candidate TFs and ERK/MAPK1, filtering to include only interactions from curated databases or that were experimentally determined. Resulting interactions were visualized using Cytoscape v3.8.2 (Shannon et al. 2003). A custom binding site motif database for candidate TFs was then created by filtering the JASPAR 2020 Core Vertebrates Non-Redundant TFBS motif database to only include motifs corresponding to our candidate TFs.

Identifying promoters and relevant promoter regions for manually-annotated venom genes.

For all genes except those belonging to the venom families discussed below, the promoter region of a given gene was defined as 1kb upstream of the transcription start site (TSS) through the TSS. Genes within the SVMP, SVSP, and PLA_2 venom gene clusters were originally annotated manually using FGENESH+ (Solovyev 2004) and known guide protein sequences (see (Schield et al. 2019)). Because FGENESH+ does not take into account gene expression data, it does not identify the TSS and instead attempts to identify a likely TATA box based on nucleotide sequence. Thus the "TSS" position labeled by FGENESH+ may not actually represent the true TSS. In order to focus on a region most likely to represent the actual TSS and adjacent sequence for these genes, we defined **p**romoter **A**TAC-seq **p**eak (PAP) regions by taking 1kb in either direction from the FGENESH+ "TSS" location (2kb region total) and identifying ATAC-seq peak regions that overlap with this window using bedtools intersect. In the event that more than one ATAC-seq peak was found within one of these regions, the most downstream peak (relative to the associated gene) was taken to be the PAP. Nucleotide sequences for promoters and PAPs were then extracted from the Prairie Rattlesnake genome using bedtools getfasta, aligned using

MAFFT v7.475 (Katoh and Standley 2013) with flags –reorder, --adjustdirectionaccurately, --allowshift, --unalignlevel 0.8, --maxiterate 0, and –globalpair and visualized using the msa package v1.22.0 (Bodenhofer et al. 2015) in R.

Identification of putative enhancer regions (PERs) and PER-gene interactions.

We used the Activity-by-Contact (ABC) model v0.2(Fulco et al. 2019) to identify putative enhancer regions (PERs) and infer PER-gene regulatory interactions in the post-extraction venom gland. Candidate enhancer regions were first determined by filtering post-extraction ATAC-seq peaks to exclude regions containing anomalously high read density (normalized density > 1000); these anomalous regions are likely to skew enhancer-gene predictions due to their extreme magnitudes. Any resulting overlapping peaks were merged into a single region. ABC was then run using an ABC score threshold of 0.05, expression cutoff of 100 TPM, and otherwise default parameters on these candidate enhancer regions using H3K27ac ChIP-seq density, ATAC-seq density, KR normalized Hi-C data at 5kb resolution, and median gene expression (TPM) of the three post-extraction venom gland samples. Venom PERs (vPERs) were defined as PER regions inferred to interact with one or more annotated venom genes. Nucleotide sequences for PERs were then extracted from the Prairie Rattlesnake genome using bedtools getfasta. Enhancer-gene interactions were plotted using ggplot2 v3.3.3 (Wickham 2011) and ggforce v0.3.2 (github.com/thomasp85/ggforce/) packages in R.

To simplify downstream analyses by determining one representative, or “core,” enhancer sequence per venom gene family, we manually curated vPER alignments within each family using the following criteria. We removed sequences that did not appear to align well to any other sequences, trimmed off extraneous sequences at the beginning and end of sequences that were not present in any other sequences, and cut out indel regions for which only a single vPER contained sequence. Sequence curation was conducted in Jalview v2.11.1.4 (Waterhouse et al. 2009), and the consensus sequence for each curated alignment was taken as the "core" vPER sequence of a given family. TFBS scans in CIIIDER were conducted using our candidate TF set to confirm that TFBS with evidence of being bound in the original vPER sequences were present within the core sequences for each family.

Transcription factor binding site (TFBS) prediction, enrichment analyses, and TFBS alignment.

Transcription factor binding site prediction and enrichment analyses were conducted using CIIIDER v0.9 (Gearing et al. 2019) with the default deficit threshold of 0.15, a gene coverage p-value cutoff of 0.05, and using the custom motif database generated above for candidate TFBS. For TFBS enrichment analyses in venom promoters, sequences for a given family were used as the target sequences and compared to promoter sequences for all non-venom genes as a background. Similarly, for vPER TFBS enrichment analyses, all non-venom PERs were used as the background. Overlap of enriched TFBS between venom gene families was plotted using the ggVennDiagram v0.5.0 package (github.com/gaospecial/ggVennDiagram) in R. Tobias ScoreBed was used to annotate TFBS positions with the footprint scores for the three ATAC-seq datasets determined above. A given TFBS was considered “bound” if the footprint scores for that position in at least two of the three replicates exceeded the “bound” threshold determined above by Tobias BINDetect.

Venom promoter and vPER sequences were aligned using MAFFT v7.475 (Katoh and Standley 2013) with flags –reorder, --adjustdirectionaccurately, --allowshift, --unalignlevel 0.8, --maxiterate 0, and –globalpair and visualized using the msa package v1.22.0 (Bodenhofer et al. 2015) in R. Locations of enriched TFBS identified in unaligned sequences via CIIIDER were then converted to their corresponding positions in the MAFFT-aligned sequences using a custom Python script (See Data Availability). This custom Python script also calculates a simple “consensus score” for each alignment, defined as the maximum percent of sequences with an identical nucleotide at a

given position in the alignment, not including alignment-introduced gaps. TFBS alignments were visualized in R using ggplot2 v3.3.3 (Wickham 2011).

Novel TFBS motif searches in venom regulatory sequences

We used de novo motif identification analyses in elevated ATAC-seq footprint regions to identify any novel TFBS motifs that would not be otherwise detected by our candidate approach described above. We confined these searches to regions with evidence of being bound by a transcription factor by only searching regions with an ATAC-seq footprint score greater than the "bound" threshold determined during the BINDetect step of the ATAC-seq footprint analysis in at least two of the three ATAC-seq replicates.

Novel motifs were identified and annotated using MEME v5.3.3 (Bailey and Elkan 1994) and TomTom v5.3.3 (Gupta et al. 2007) within the online MEME-ChIP tool v5.3.3 (Machanick and Bailey 2011). MEME was run in Differential Enrichment mode using a background of all "bound" footprint regions in enhancers or promoters not associated with venom genes, and was set to identify at most 25 motifs. MEME motifs with an e-value < 0.05 were considered significant, and these motifs were compared to motifs in the JASPAR 2020 non-redundant vertebrate motif database using TomTom with a permissive e-value cutoff of 50 to assess similarity with known binding sites.

Comparisons of venom regulatory sequences with those of non-venom paralogs.

Non-venom paralog genes were identified previously (Schield et al. 2019). To assess whether any vPER sequences identified for the three major venom gene families were also present near a family's non-venom paralogs, we used BLASTN to identify similar sequences genome-wide using as a query the sequence the core vPER sequence of each family. We then surveyed non-venom paralogs and adjacent regions for significant vPER BLAST hits (e < 0.000001).

To compare venom gene and non-venom paralog promoters, we scanned non-venom paralog promoters for candidate TFBS using CIIIDER with default settings and our custom candidate TF motif database, and filtered the results to include TFBS inferred to be bound in promoters of the corresponding venom gene family. We also tested for enrichment of any candidate TFBS in non-venom paralog promoters for each family compared to a background of all promoters (excluding all venom gene and non-venom paralog promoters).

Identifying potential conserved vPER sequences in other venomous snake species.

To investigate whether vPER sequences are conserved in other venomous snakes, we used BLASTN (Altschul et al. 1990) to search all snake nucleotide sequences on NCBI (via the online BLAST platform) and BLASTN in BLAST+ v2.6.0 to search a set of existing snake genome assemblies - *Naja naja* (NCBI: GCA_009733165.1 (Suryamohan et al. 2020)), *Deinagkistrodon acutus* (Yin et al. 2016), *Thamnophis sirtalis* (NCBI: GCA_001077635.2 (Perry et al. 2018)*), Protobothrops flavoviridis* (NCBI: GCA_003402635.1 (Shibata et al. 2018)), and *Python bivittatus* (NCBI: GCA_000186305.2 (Castoe et al. 2013)). We used as the query sequence the core sequence for the SVMP and PLA_2 families (SVSP was excluded from these analyses for reasons discussed below). The at most five best hits from each species with were selected based on e-value and bit-scores. Alignments were generated using MAFFT with parameters described above. For PLA_2 vPER BLAST searches against all snake nucleotide sequences on NCBI, a subset of returned hits were small (i.e., covered less than 25% of the query sequence) and only hit to regions on the extremities of the query vPER sequence with no similarity to the center of the vPER sequence where functionally-relevant TFBS are inferred to be located; these sequences were manually removed from alignments. TFBS inferred to be bound in SVMP and PLA_2

enhancers were scanned in the resulting BLAST hit sequences using CIIIDER with default parameters and visualized in R using ggplot2. An approximated phylogeny for lineages represented in these analyses was downloaded from TimeTree (Kumar et al. 2017).

For SVMP vPER BLAST hits, TFBS positions were classified into three categories: shared, viper-specific, and elapid-specific. A given TFBS position was considered viper-specific if that TFBS was inferred in the same position in the majority of BLAST hits to viperid species, but not present in the majority of BLAST hits to elapid species (and vice versa for elapid-specific TFBS positions). TFBS positions present in the majority of both elapid and viperid BLAST hit sequences were considered "shared" TFBS positions.

Analyses of transposable elements (TEs) associated with SVSP regulatory sequences.

Given the fact that BLASTs of SVSP vPER sequences yielded a large number of hits throughout the Prairie Rattlesnake genome, we investigated whether this could be explained by an association between these sequences and transposable elements (TEs). Using TE annotations from (Pasquesi et al. 2018), we used Giggle v0.6.3 (Layer et al. 2018) to test whether SVSP regulatory regions (promoters and vPERs) were significantly enriched for overlap with any particular TE (one-tailed Fisher's exact test; $p < 0.05$). This analysis identified a DNA/hAT-Tip100 element (Cv1-hAT-Tip100) that was enriched in the SVSP regulatory regions and generally common on chromosome 10. A genome-wide consensus sequence for this element was generated using mafft by providing the DNA hAT-Tip100 consensus from the repeat element library as reference. This preliminary alignment was then manually curated by removing the DNA hAT-Tip100 reference, re-aligning the genomic copies, and removing major regions with limited coverage by using the Gblocks server v0.91b (Talavera and Castresana 2007). The final consensus sequence was derived by using the Unipro UGENE software (Rose et al. 2019). This consensus sequence was then used to calculate sequence divergence (pairwise-pi) for all Cv1-hAT-Tip100. For this calculation, we excluded alignment positions where the highest nucleotide frequency exceeded 0.7. Using these pairwise-pi values, we estimated TE age as $pi \div 2 \times 2.4 \times 10^9$ following (Pasquesi et al. 2018). For Cv1-hAT-Tip100 copies within the SVSP region, including those in regulatory and "other" intergenic sequences, we used CIIIDER to identify TFBS identified above as bound in SVSP promoters and/or enhancers. These sequences and TFBS positions were then aligned using mafft and the custom Python script described above, and plotted in R using ggplot2.

To secondarily characterize TE content in the SVMP and PLA2 venom gene regions, we used the coverage tool within bedtools to calculate the percent of bases annotated as repetitive in 10kb windows. Windows overlapping with venom gene regions were compared to all other windows on the corresponding chromosome. Separately, we used bedtools intersect to identify annotated repeats within each venom gene region, filtered to retain the top ten most abundant repeat element types per venom gene region, and plotted the results using pHeatmap in R. Finally, Giggle was run as described above for SVMP and PLA2 regulatory regions (promoters and vPERs), and elements with significant overlap (one-tailed Fisher's exact test; $p < 0.05$) were plotted using ggplot2 in R.

Identification of exonic debris in the PLA_2 gene cluster.

We used BLAST feature of ncbi-blast v.2.7.1+ suite (tblastx, e-value of 0.01, default restrictions on word count and gaps) to perform initial search of the exons against a pre-compiled database of exons previously successfully used to annotate Pla_2GIIe family of genes in vertebrate animals (Koludarov et al. 2020). We then manually assessed each result and established exon boundaries using Geneious v11 (www.geneious.com). The resulting exon debris locations were overlapped with BLASTN hits of the PLA_2 core vPER sequence to the PLA_2 region to assess the relationship between exon debris

Supplemental References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–W208.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T. 2020. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**: 1–11.

Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. 2015. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**: 3997–3999.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Castoe TA, de Koning AP, Hall KT, Card DC, Schield DR, Fujita MK, Ruggiero RP, Degner JF, Daza JM, Gu W, et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc Natl Acad Sci USA* **110**: 20645–20650. http://www.ncbi.nlm.nih.gov/pubmed/24297902.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**: 95–98.

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA. 2019. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**: 1664–1669.

Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, Gould JA, Forster SC, Hertzog PJ. 2019. CiiiDER: A tool for predicting and analysing transcription factor binding sites. *PLoS One* **14**: e0215495.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: 1–9.

Ignatiadis N, Klaus B, Zaugg JB, Huber W. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods* **13**: 577.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Koludarov I, Jackson TNW, Suranse V, Pozzi A, Sunagar K, Mikheyev AS. 2020. Reconstructing the evolutionary history of a functionally diverse gene 2 family reveals complexity at the genetic origins of novelty. *bioRxiv* 583344.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**: 1812–1819.

Layer RM, Pedersen BS, DiSera T, Marth GT, Gertz J, Quinlan AR. 2018. GIGGLE: a search engine for large-scale integrated genome analysis. *Nat Methods* **15**: 123.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao Y, Smyth GK, Shi W. 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* **47**: W199–W205.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**: 1696–1697.

Pasquesi GIM, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, Reyes-Velasco J, Ruggiero RP, Vandewege MW, Shortt JA. 2018. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat Commun* **9**: 1–11.
Perry BW, Card DC, McGlothlin JW, Pasquesi GIM, Adams RH, Schield DR, Hales NR, Corbin AB, Demuth JP, Hoffmann FG. 2018. Molecular adaptations for sensing and securing prey and insight into amniote genome diversity from the garter snake genome. *Genome Biol Evol* **10**: 2110–2129.
Perry BW, Schield DR, Westfall AK, Mackessy SP, Castoe TA. 2020. Physiological demands and signaling associated with snake venom production and storage illustrated by transcriptional analyses of venom glands. *Sci Rep* **10**: 1–10.
Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi. R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**: 2808–2810.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
R Core Team. 2013. R: A language and environment for statistical computing.
Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–W165.
Rose R, Golosova O, Sukhomlinov D, Tiunov A, Prosperi M. 2019. Flexible design of multiple metagenomics classification pipelines with UGENE. *Bioinformatics* **35**: 1963–1965.
Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, Adams RH, Corbin AB, Smith CF, Ramesh B. 2019. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* **29**: 590–601.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.
Shibata H, Chijiwa T, Oda-Ueda N, Nakamura H, Yamaguchi K, Hattori S, Matsubara K, Matsuda Y, Yamashita A, Isomoto A. 2018. The habu genome reveals accelerated evolution of venom protein genes. *Sci Rep* **8**: 11300.
Solovyev V. 2004. Statistical approaches in eukaryotic gene prediction. *Handb Stat Genet*.
Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schröder MS, Wu M, Kuriakose B, Mathew OK, Perumal RC, Koludarov I. 2020. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nat Genet* 1–12.
Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**: D607–D613.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564–577.
UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**: D506–D515.
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.
Wickham H. 2011. ggplot2. *Wiley Interdiscip Rev Comput Stat* **3**: 180–185.
Yin W, Wang Z, Li Q, Lian J, Zhou Y, Lu B, Jin L, Qiu P, Zhang P, Zhu W. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nat Commun* **7**.
Yu W, He B, Tan K. 2017. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nat Commun* **8**: 535.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: 1–9.

Supplemental Tables

Supplemental Table S1. RNA-seq sample information, including sample name, species, sex, tissue, total RNA-seq reads, and the number and percent of uniquely mapped RNA-seq reads.

Sample Name	Species	Sex	Tissue	# Total RNA-seq Reads	# Uniquely Mapped RNA-seq Reads (Millions)	% Uniquely Mapped RNA-seq Reads
Lvr_11	*Crotalus viridis viridis*	F	Liver	25,253,136	16,639,532	65.90%
Lvr_4	*Crotalus viridis viridis*	M	Liver	20,389,359	15,785,368	77.40%
RVG_1	*Crotalus viridis viridis*	M	Venom Gland	18,348,330	13,090,067	71.30%
Panc_1	*Crotalus viridis viridis*	M	Pancreas	23,596,362	18,822,720	79.80%
Pnc_11	*Crotalus viridis viridis*	F	Pancreas	24,652,951	20,187,639	81.90%
Pnc_4	*Crotalus viridis viridis*	M	Pancreas	22,605,467	17,949,996	79.40%
RVG_11	*Crotalus viridis viridis*	F	Venom Gland	25,165,353	16,928,839	67.30%
RVG_4	*Crotalus viridis viridis*	M	Venom Gland	19,401,217	14,419,759	74.30%
Stom_1	*Crotalus viridis viridis*	M	Stomach	17,152,252	12,642,498	73.70%

Supplemental Table S2 (see Supplemental_Table_S2.xslx). Candidate transcription factors (TFs) for potential roles in venom regulation. Candidate approach from which a candidate TF was identified is shown, as well as normalized gene expression counts across all samples, membership of TFs in relevant functional groups (i.e., interactions with ERK, previously implicated in venom regulation), and differential expression analysis results (bold IHW p-values are significant; $p < 0.05$).

Supplemental Table S3 (see Supplemental_Table_S3.xslx). Transcription factor binding site (TFBS) enrichment analysis details. Analysis statistics for TFBS enrichment analyses of venom gene promoters and enhancers compared to all non-venom gene promoters and enhancers, respectively.

Supplemental Table S4 (see Supplemental_Table_S4.xslx). De novo motif search results for promoters and enhancers, including significantly enriched motifs (e-value < 0.05) and permissive motif characterization results (e < 50) using TomTom.

Supplemental Table S5 (see Supplemental_Table_S5.xslx). Full analysis details and statistics for all putative enhancer-gene pairs predicted by ABC analysis.

Supplemental Table S6 (see Supplemental_Table_S6.xslx). BLAST results of SVMP and PLA2 core enhancer sequences against all snake sequences on NCBI and available snake reference genomes.

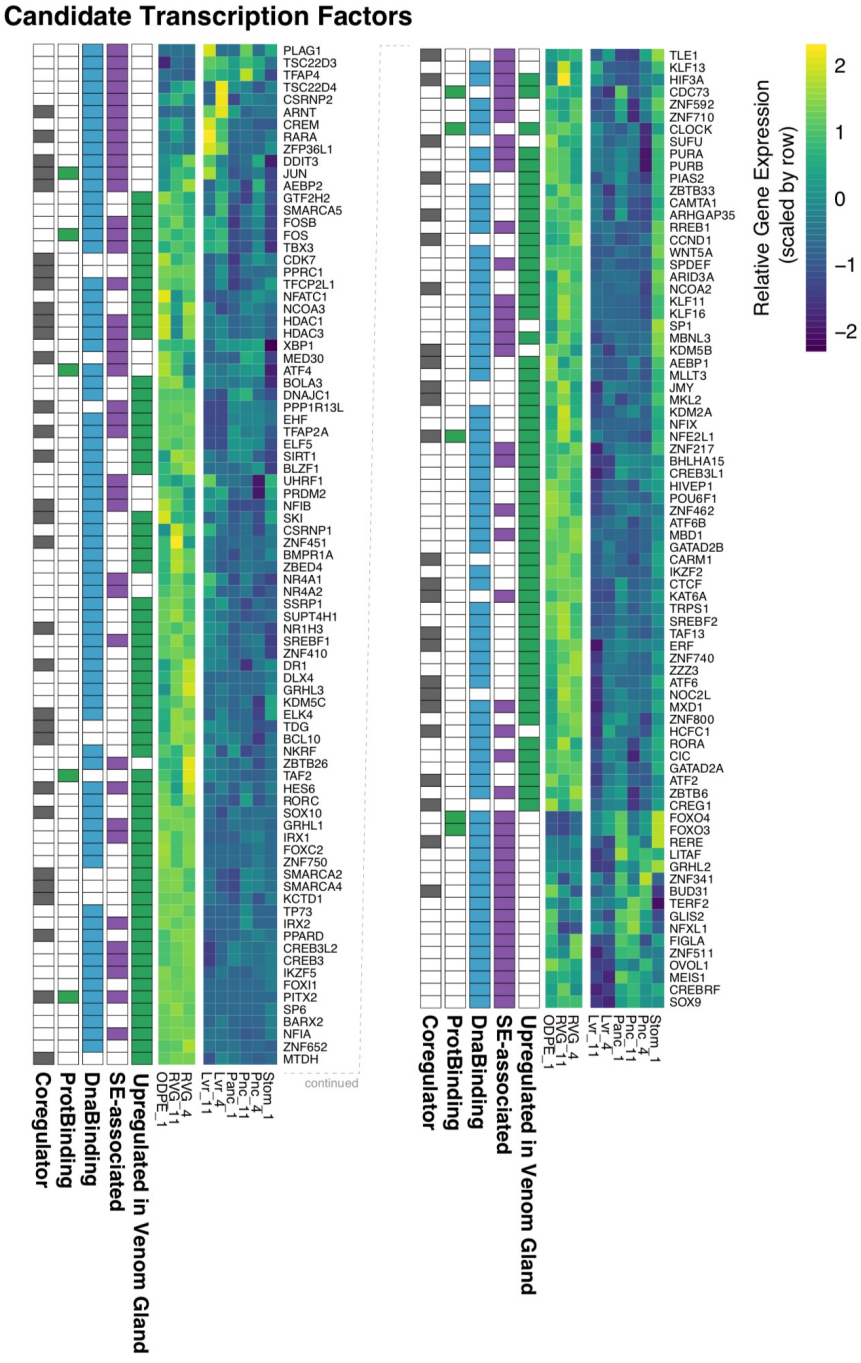


Figure S1. Gene expression heatmap showing expression of candidate TFs with upregulated expression in the venom gland compared to non-venom tissues and/or association with super-enhancers (SEs). Gene expression "heat" is scaled by row to illustrate differences between tissues and time points. Annotations on the left of the heatmap show TF types (DNA Binding, Protein Binding, and/or Co-regulator). For annotation columns, a colored box indicates membership to a given functional group/analysis.

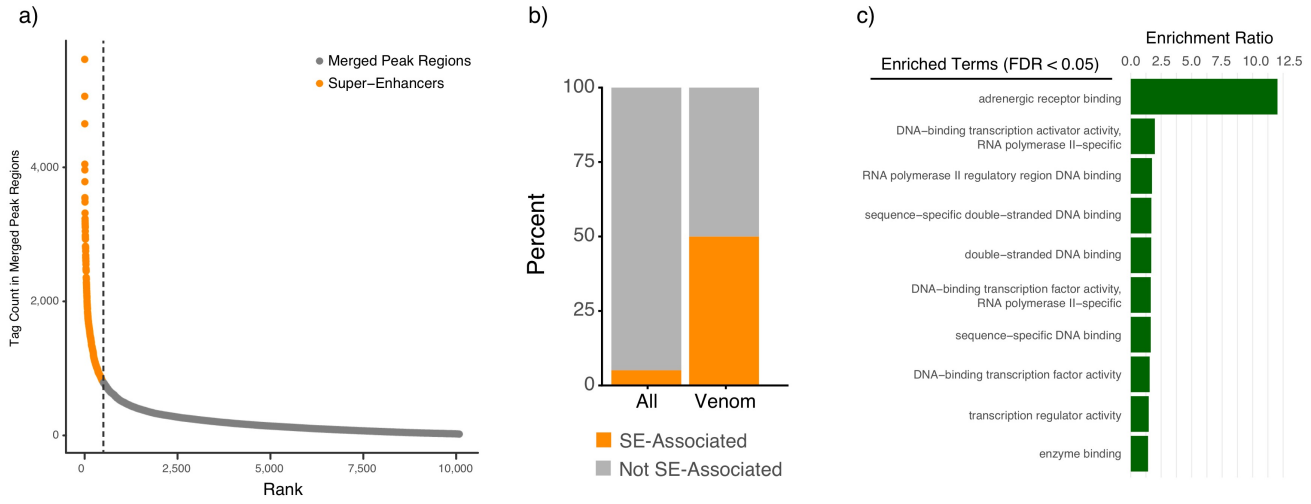


Figure S2. Super-enhancers in the post-extraction venom gland. a) Rank-intensity plot of merged H3K27ac ChIP-seq peak regions used to define super-enhancers (regions with top 5% highest ChIP-seq intensity). b) Proportion of genes associated with (within or nearest-to) super-enhancers compared between the all annotated genes ('All') and venom genes ('Venom'). c) Gene ontology overrepresentation analysis results of SE-associated genes compared to a background of all annotated genes in the Prairie Rattlesnake genome. All terms shown are significantly overrepresented in SE-associated genes (FDR < 0.05).

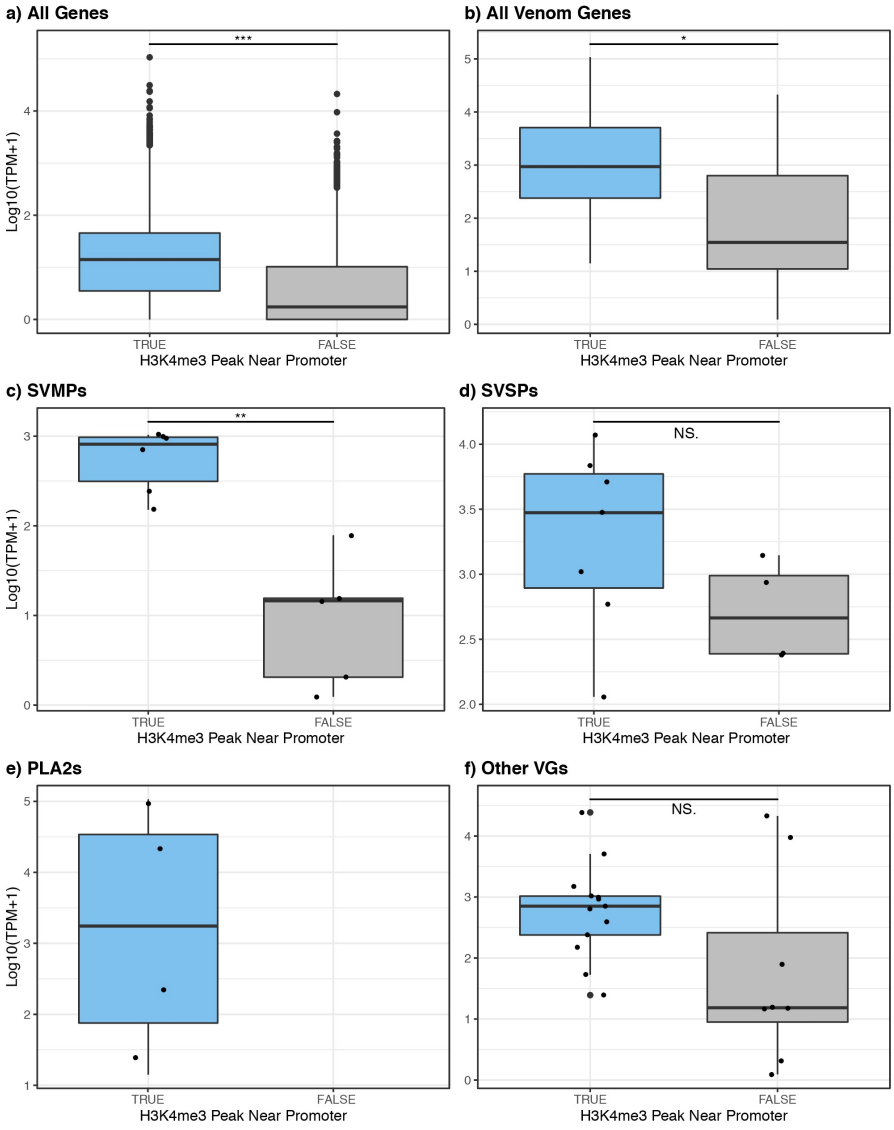


Figure S3. Comparisons of normalized gene expression at 1 day post-extraction (1DPE) between genes with and without an H3K4me4 ChIP-seq peak within 1kb of the promoter for a) all genes, b) all venom genes lumped together, c) SVMP, d) SVSP, e) PLA2, and f) "other" venom genes (*: p-value < 0.05, ***: p-value < 0.001, NS: not significance; Student's t-test).

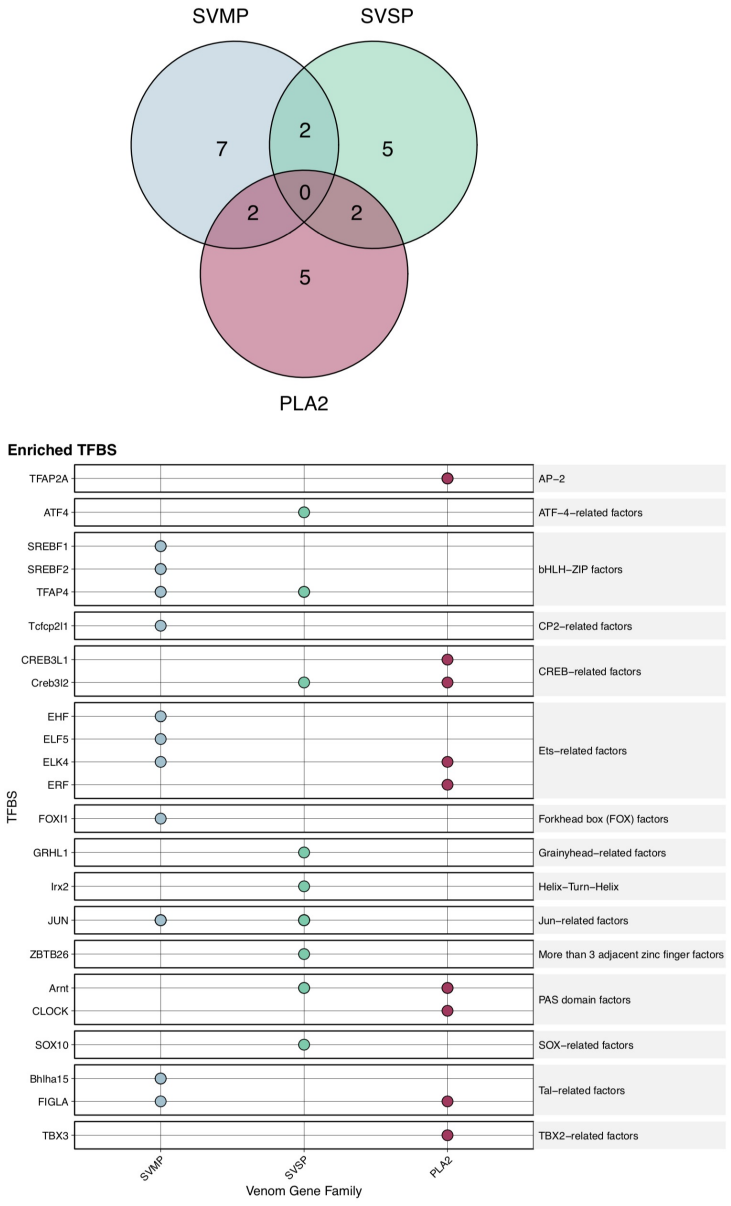


Figure S4. Significantly enriched TFBS in promoters of the three major venom gene families. The venn diagram (top) shows shared enriched TFBS between SVMP, SVSP, and PLA2 venom gene families. In the bottom plot, dots indicate enrichment in the promoter ATAC-seq peaks of a given venom gene family ($p < 0.05$).

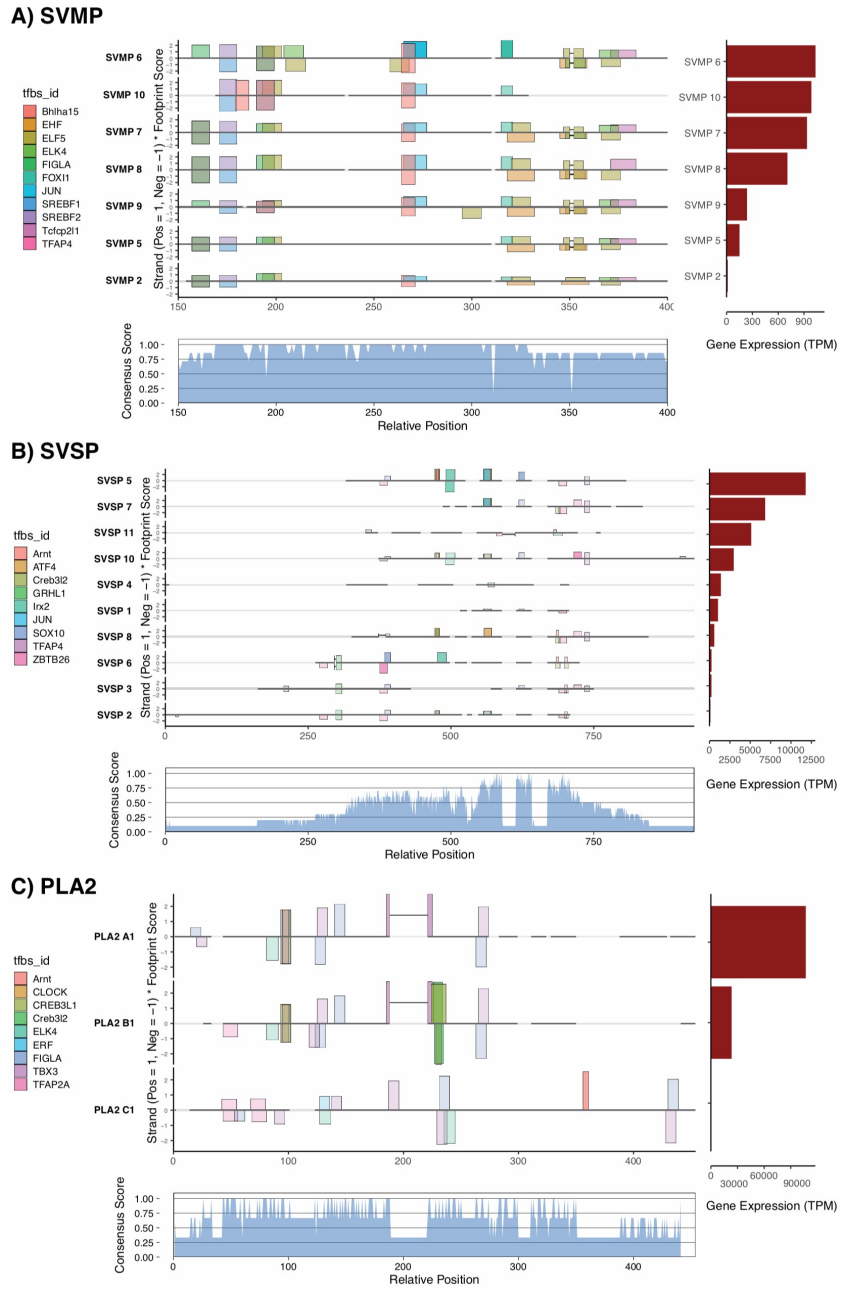


Figure S5. Promoter alignments of major venom gene clusters with TFBS inferences. Colored vertical bars on each promoter indicate presence of a TFBS for an enriched TF (bar size scaled by ATAC-seq footprint score, and the TFBS orientation indicated by its position above or below the line). Alignment gaps shown by faded regions of center lines, and colored lines connecting TFBS bars indicate TFBS that span gaps. Gene expression is shown to the right, and conservation of aligned sequences are shown below.

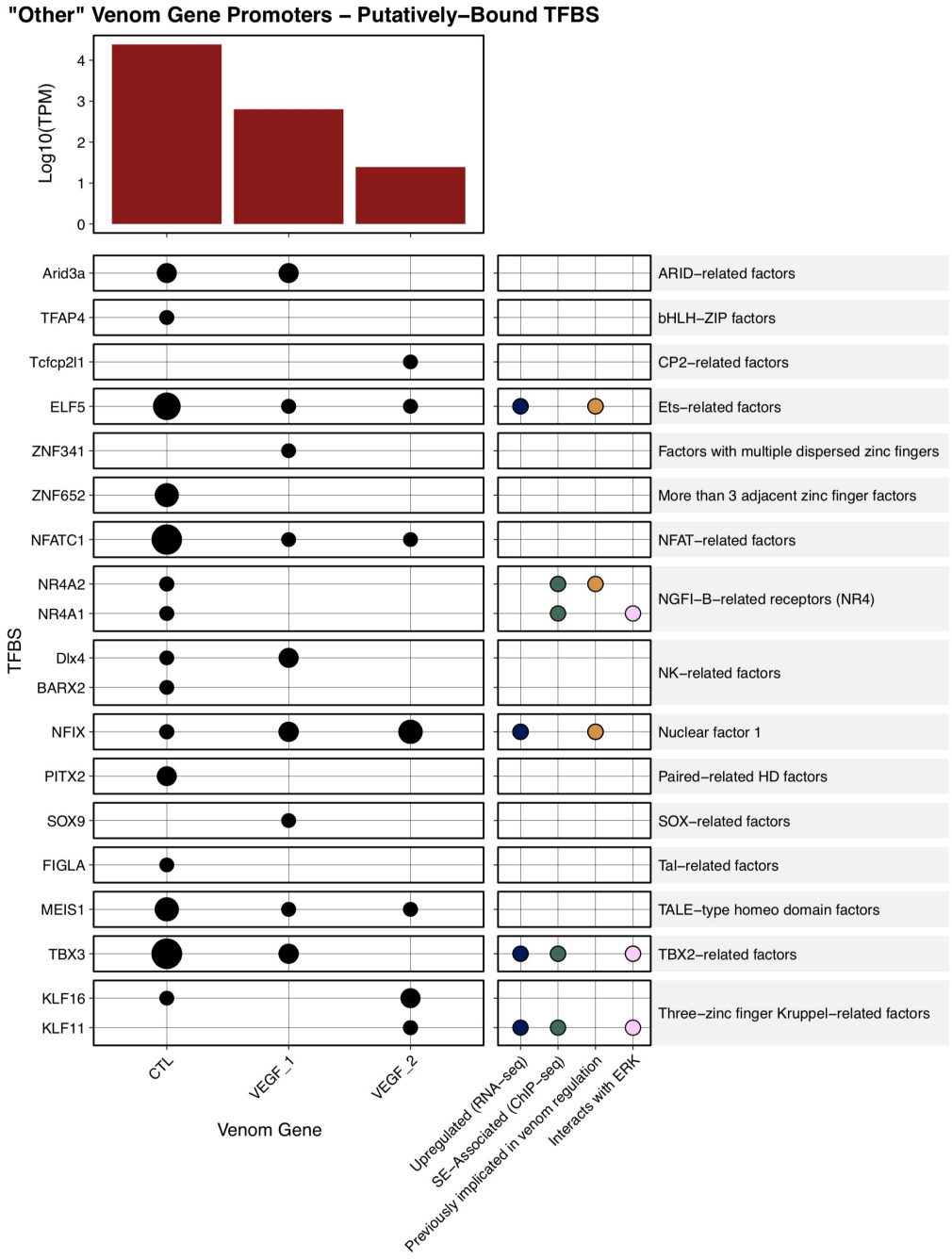


Figure S6. TFs with one or more putatively bound TFBS in the promoters "other" venom genes. A dot in the left panel indicates presence of one or more bound TFBS, with dot size scaled by number of TFBS (larger = more bound TFBS positions). Gene expression is shown at the top, and functional annotations and TF family are shown to the right.

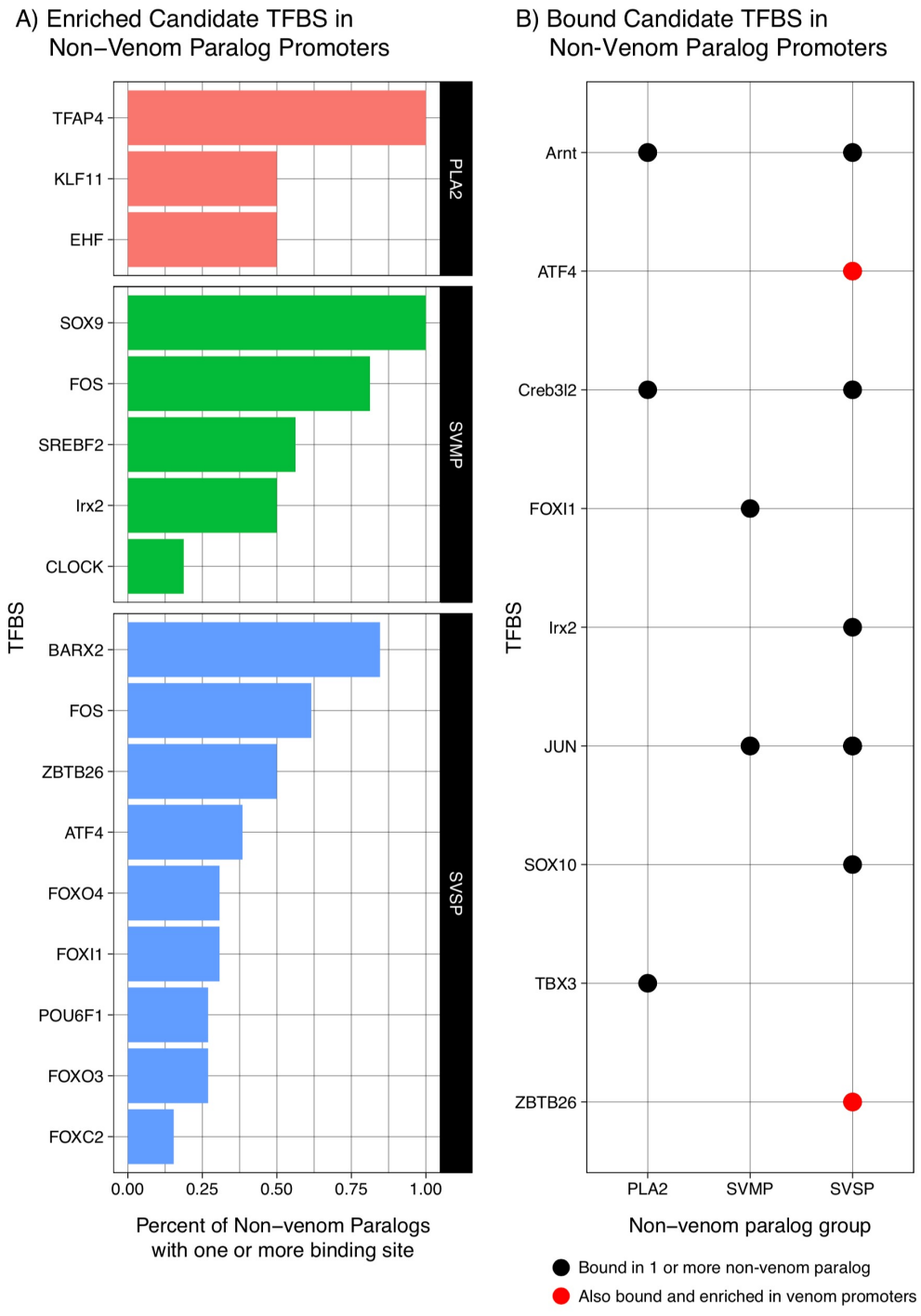


Figure S7. Candidate TFBS in promoters of non-venom paralogs (NVPs). A) TFBS with significant enrichment ($p < 0.05$) in the promoter regions of NVPs of PLA2, SVMP, and SVSP venom genes. All TFBS are enriched, and the x-axis shows the percent of NVPs in each group with one or more of each TFBS. B) For each NVP group, presence of one or more bound TFBS for an enriched TF is shown with a dot, and red dots indicate presence of a bound TFBS that is also enriched and bound in the related venom gene family of a given NVP group.

Figure S8. Putative enhancer regions (vPERs) for Cysteine-rich secretory proteins (CRISPs). vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

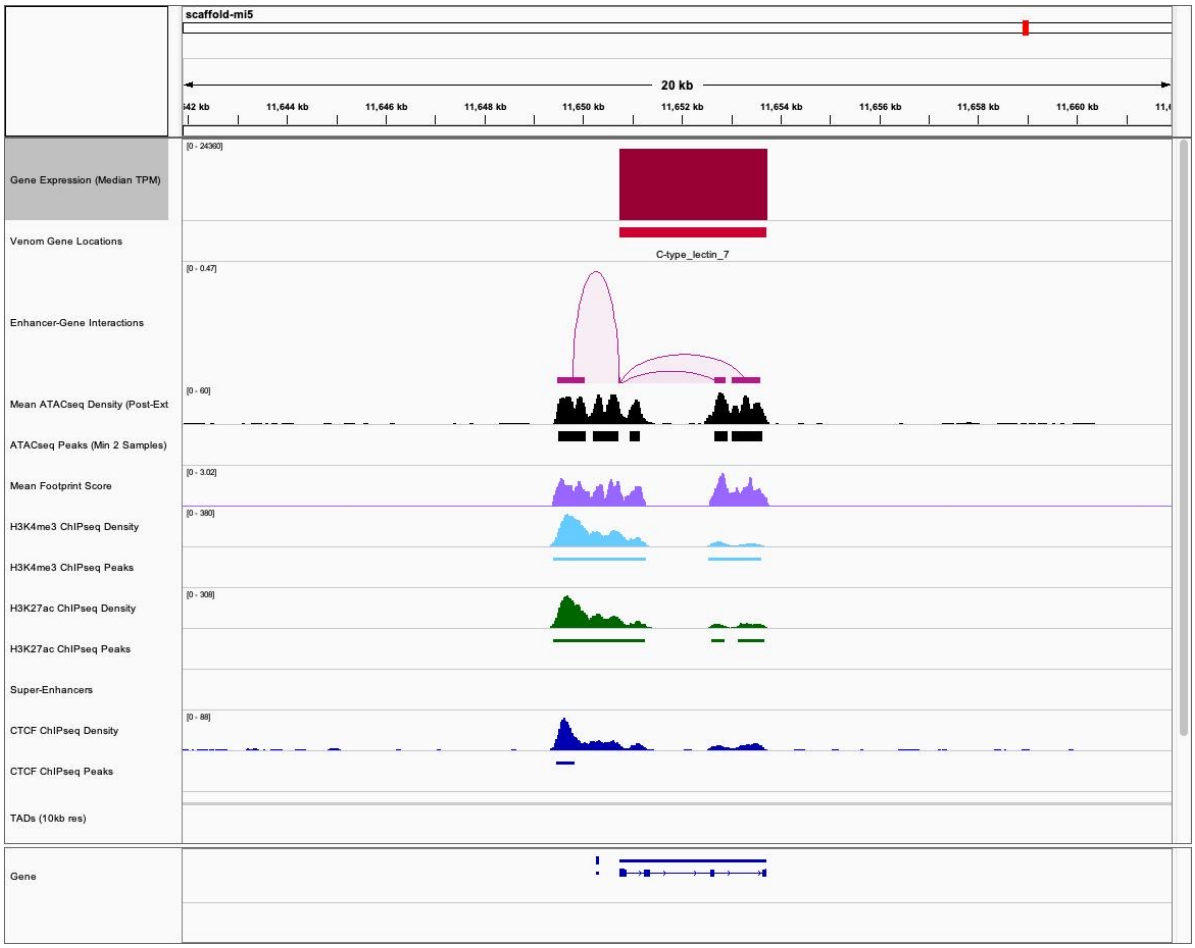


Figure S9. Putative enhancer regions (vPERs) for C-Type lectins (CTL). vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

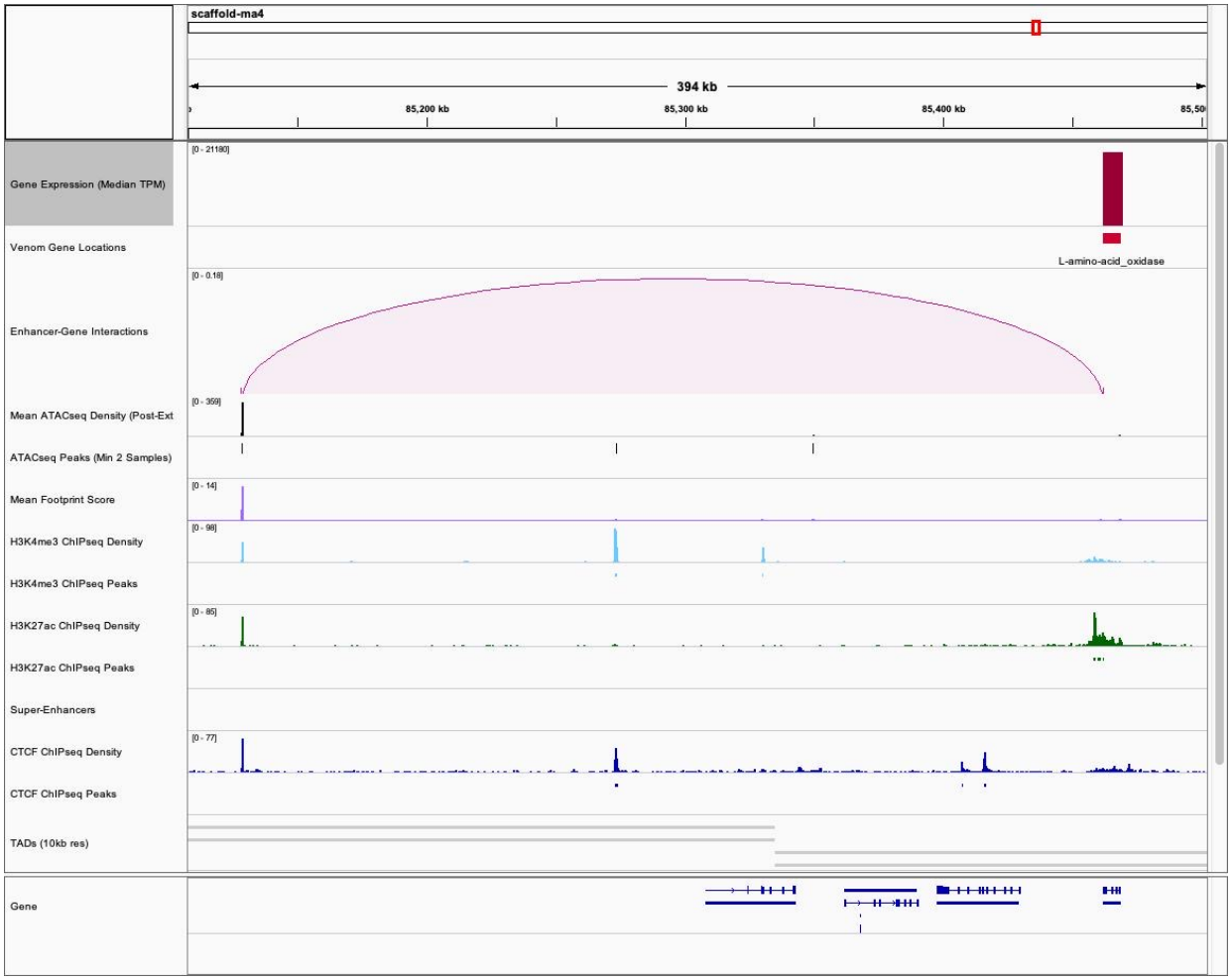


Figure S10. Putative enhancer regions (vPERs) for L-Amino Acid Oxidase 3 (LAAO3). vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

Figure S11. Putative enhancer regions (vPERs) for Vascular Endothelial Growth Factor A (VEGFA). vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

Figure S12. Putative enhancer regions (vPERs) for Vespryn. vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

Figure S13. Putative enhancer regions (vPERs) for glutaminyl cyclase 1 (v_QC1). vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

Figure S14. Putative enhancer regions (vPERs) for glutaminyl cyclase 2 (v_QC2). vPER inferences are shown in the "Enhancer-Gene Predictions" track; arcs begin at the promoter and end at the inferred vPER region (marked with a thicker purple bar). For ATAC-seq and ChIP-seq data, peak regions are marked with a bar underneath the read density plots.

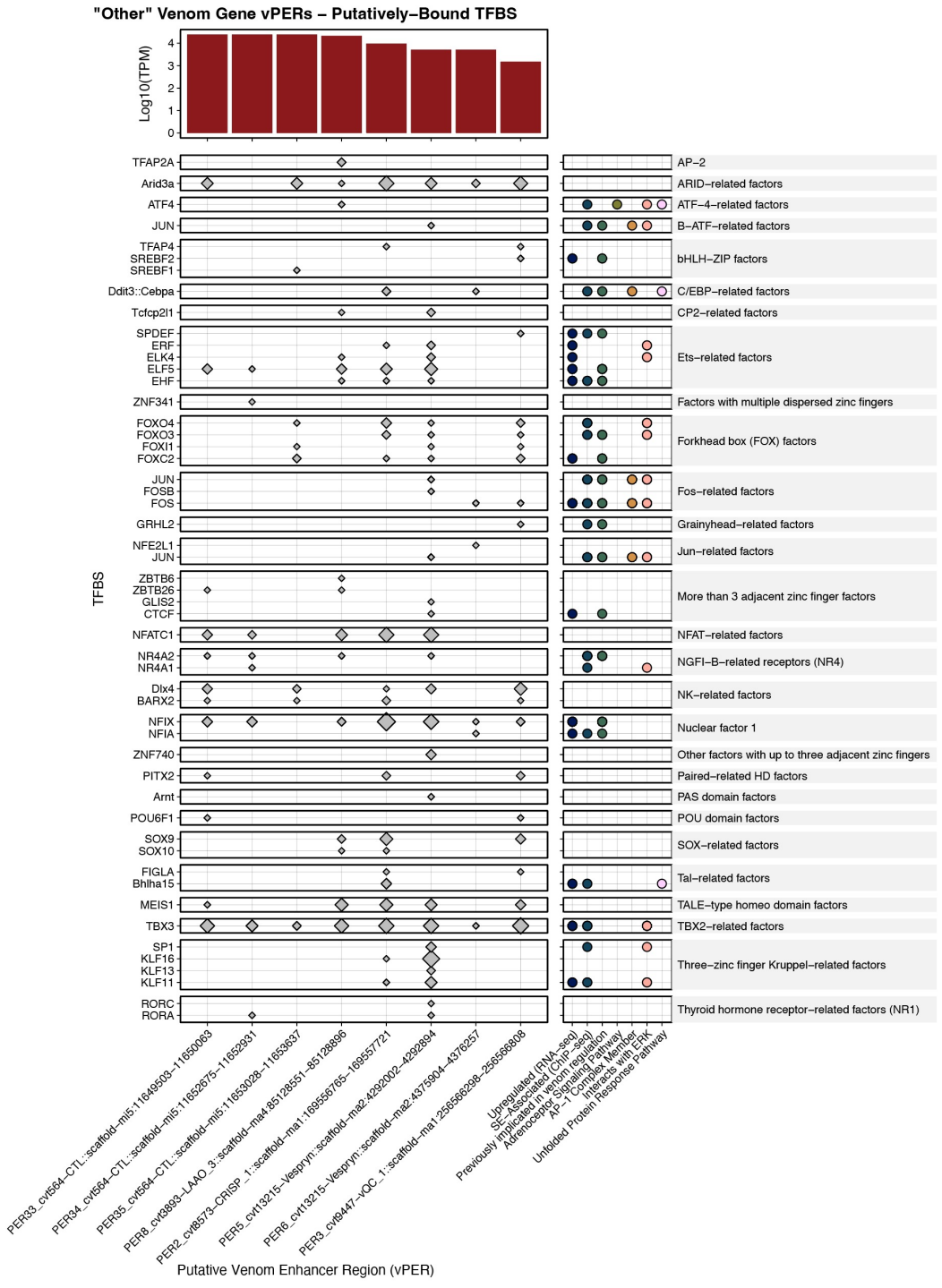


Figure S15. TFs with one or more putatively bound TFBS in the putative enhancers of "other" venom genes. A dot in the left panel indicates presence of one or more bound TFBS, with dot size scaled by number of TFBS (larger = more bound TFBS positions). Gene expression is shown at the top, and functional annotations and TF family are shown to the right.

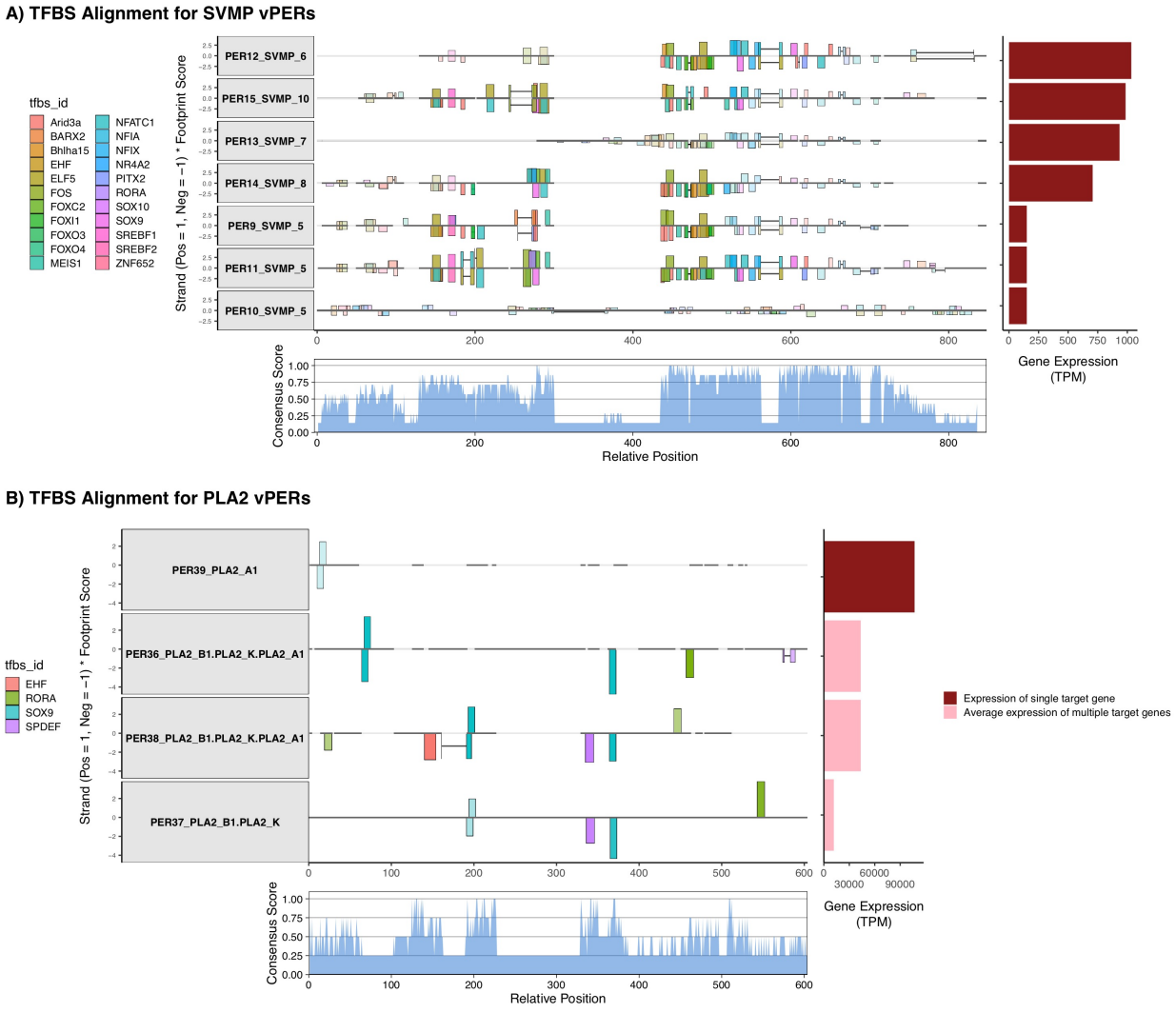


Figure S16. Putative enhancer alignments of major venom gene clusters with TFBS inferences. Colored vertical bars on each promoter indicate presence of a TFBS for an enriched TF (bar size scaled by ATAC-seq footprint score, and the TFBS orientation indicated by its position above or below the line). Alignment gaps shown by faded regions of center lines, and colored lines connecting TFBS bars indicate TFBS that span gaps. Gene expression is shown to the right, and conservation of aligned sequences are shown below. If a putative enhancer targets multiple genes (i.e., PER36 in B), the averaged expression of all target genes is shown (colored pink). SVSP enhancers are not shown due to the large quantity of putative enhancers, of which the majority did not align cleanly.

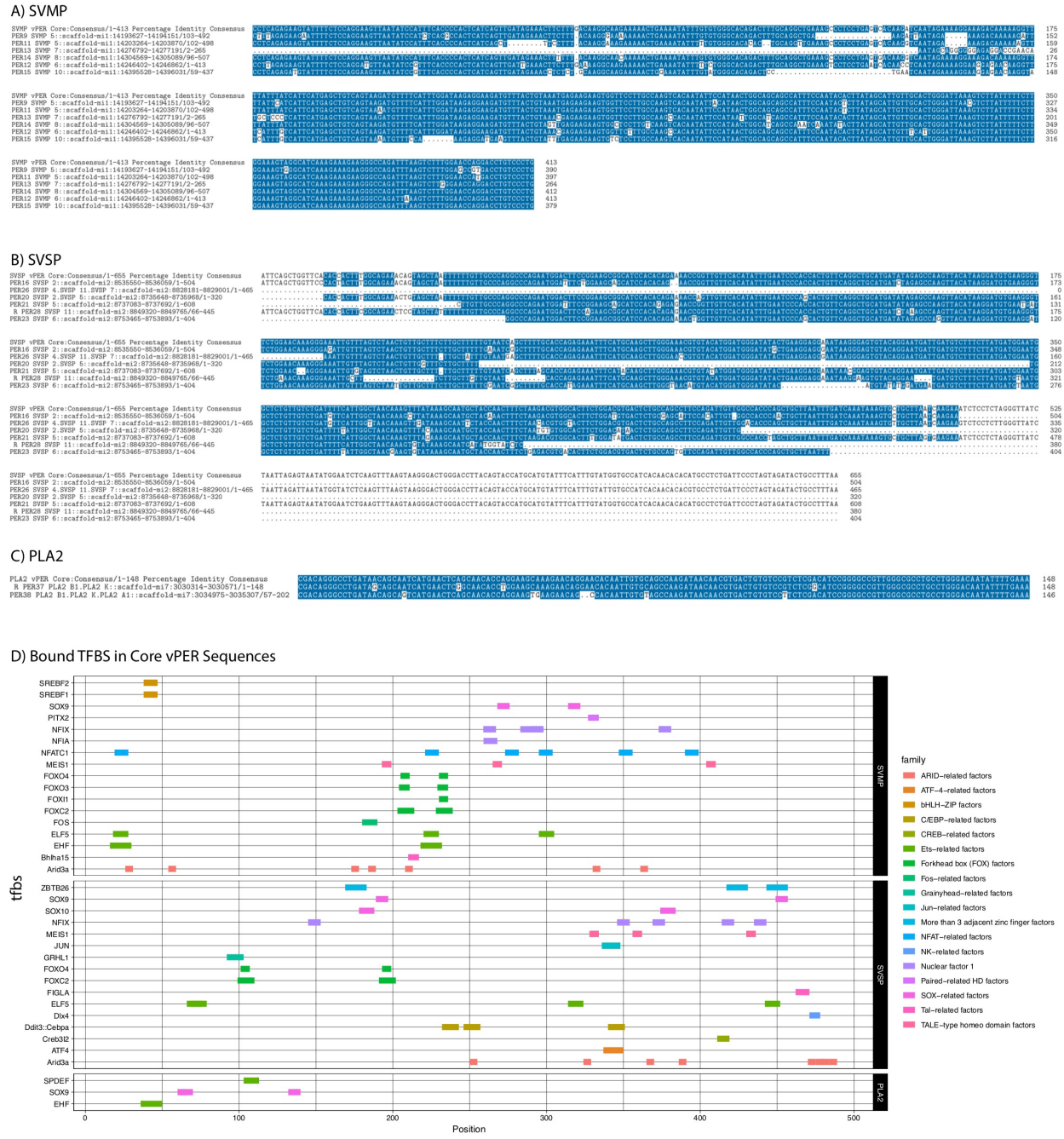
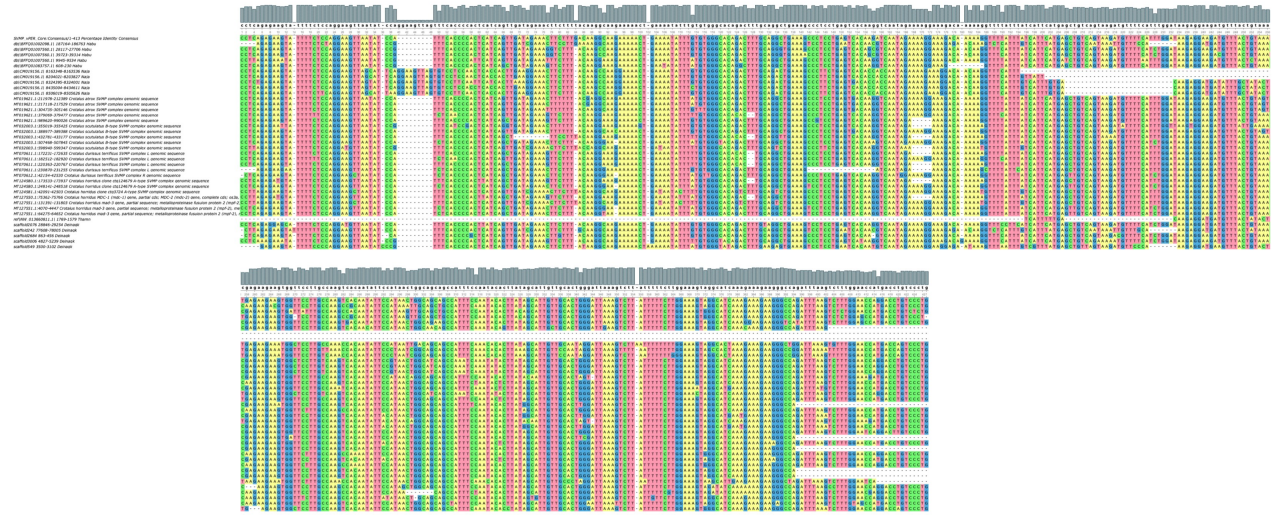


Figure S17. Curated "core" enhancer sequences for the the three major venom gene families. A-C) Sequence alignments of a subset of putative enhancer regions that showed the highest similarity and were thus used to generate a consensus sequence to represent the "core" enhancer of a given venom gene family. The top sequence in each alignment is the consensus "core" enhancer sequence. Positions highlighted in blue are identical between >50% of sequences. D) Positions of enriched and bound TFBS in the "core" enhancer sequences. TFBS positions are shown here if they were present and bound in at least 50% of the input sequences used to generate the consensus "core" sequence.

A) SVMP Core vPER - Blast Hits to Other Snake Species



B) PLA2 Core vPER - Blast Hits to Other Snake Species

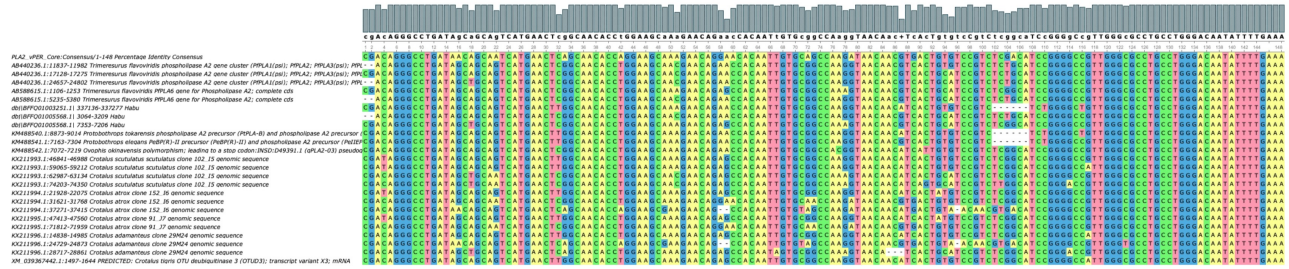


Figure S18. Conserved putative enhancer region sequences across snake species. Nucleotide alignments show BLASTn hits for the "core" enhancer sequence of A) SVMPs and B) PLA2s. The top 5 best blast hits (based on e-value and bit-score) are shown for each species. Sequence conservation and a consensus sequence is shown above the alignments.

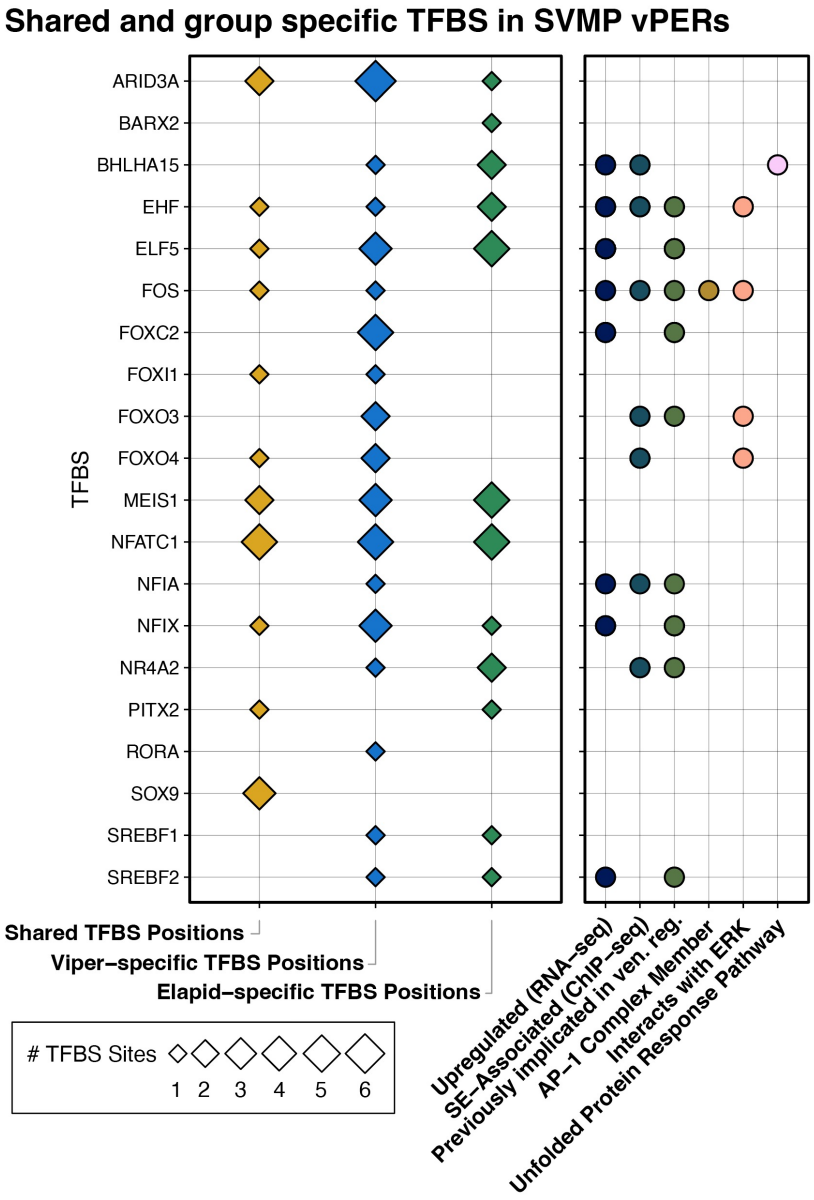


Figure S19. Shared and lineage specific TFBS positions in SVMP vPERs. TFBS identified in SVMP vPER BLAST hits were classified as viper-specific if present at a given position in the majority of viperid sequences, while not present at that position in the majority of elapid sequences (and vice versa for elapid-specific sequences). Shared TFBS positions are present in the majority of both elapid and viperid sequences. The size of the diamonds in the left panel corresponds to the number of TFBS positions for a given transcription factor that fall into each categorization. Dots in the right panel indicate relevant functional annontations for each transcription factor.

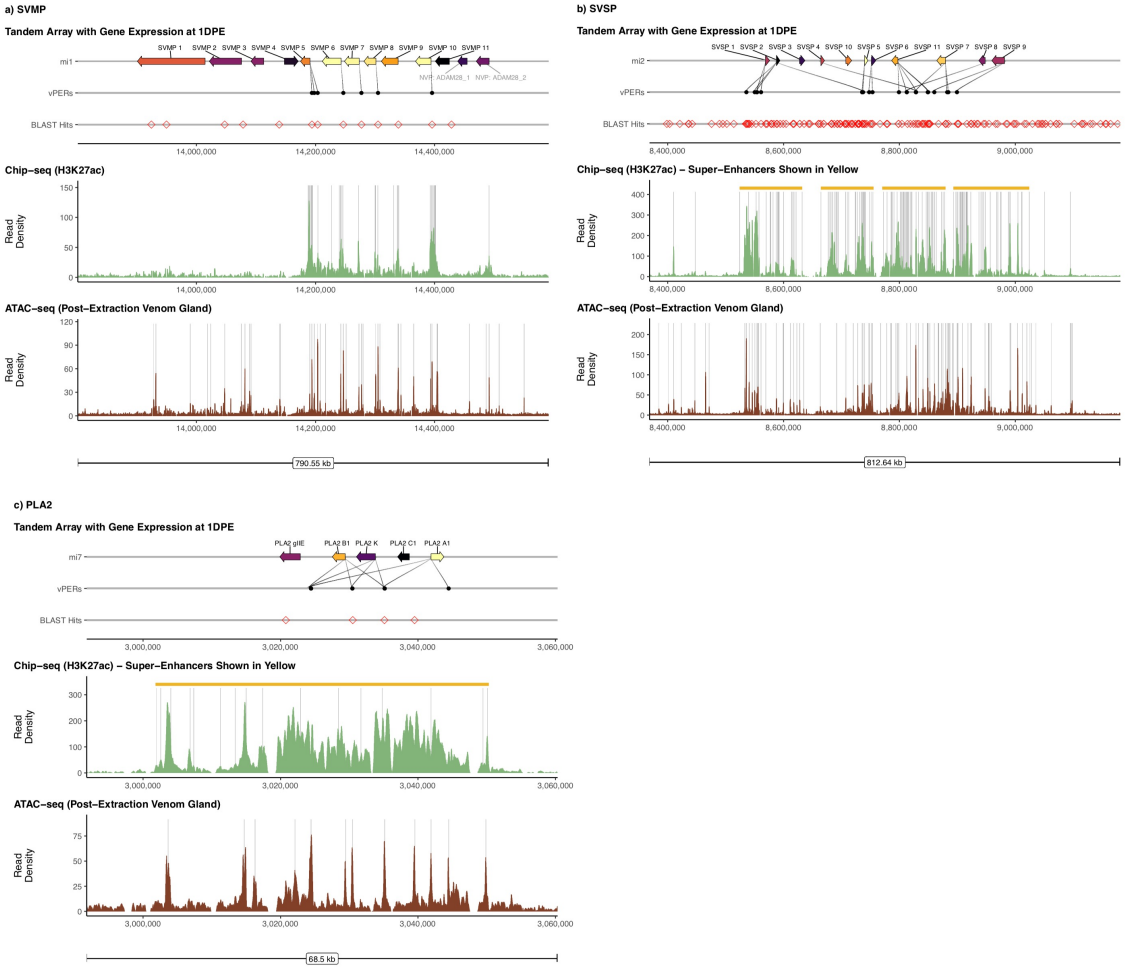


Figure S20. Results of BLASTn searches of core vPERs against the Prairie Rattlesnake genome. For A) SVMPs, B) SVSPs, and C) PLA2s, the core vPER sequence was searched back against the genome using BLASTn. BLAST hits with an e-value < 0.000001 (red diamonds) are shown in the major venom array regions. Local H3K27ac ChIP-seq (green) and ATAC-seq (dark brown) read density is shown below, with peak regions shown with vertical grey bars.

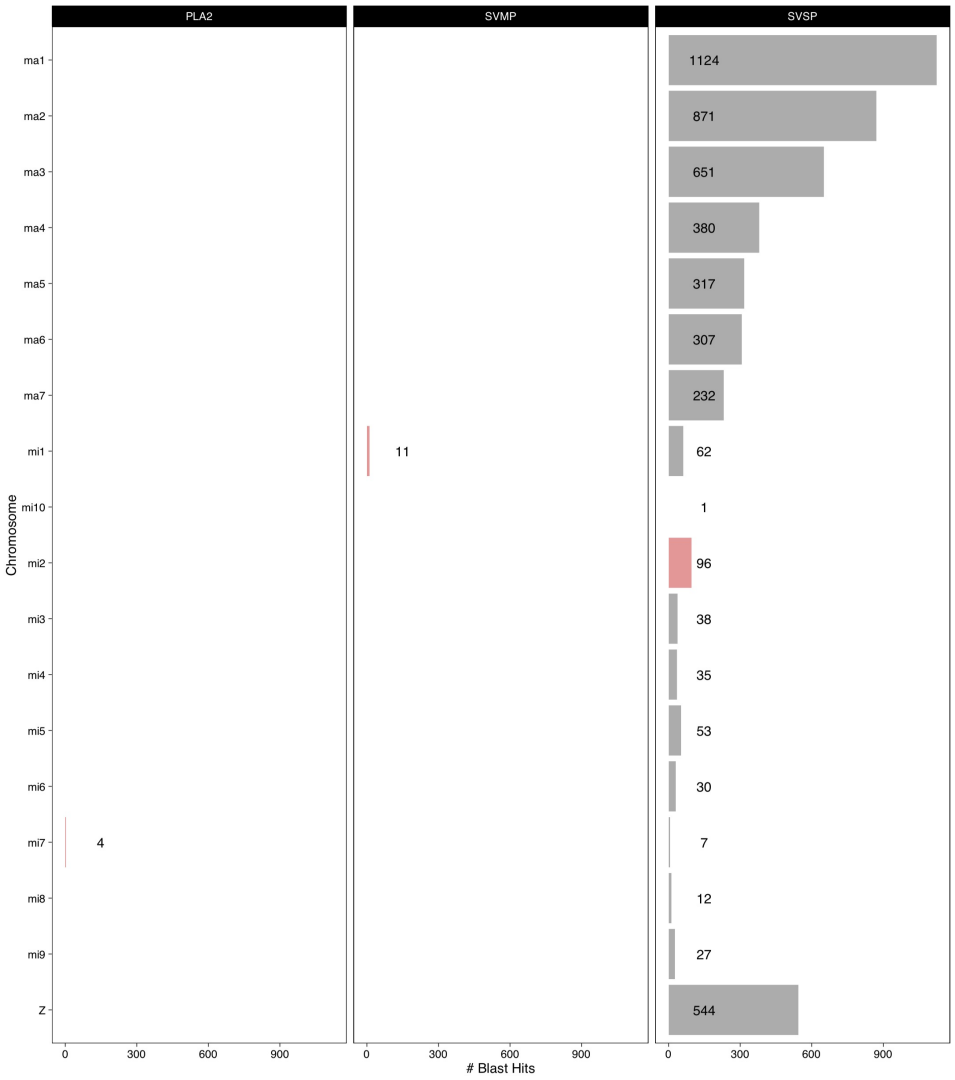


Figure S21. Bar plot showing the number of significant BLAST hits (e < 0.000001) to the Prairie Rattlesnake genome found when using the core vPER sequence for each family as the query sequence. Bars in red indicate the chromosome on which the query vPER sequence venom cluster reside.

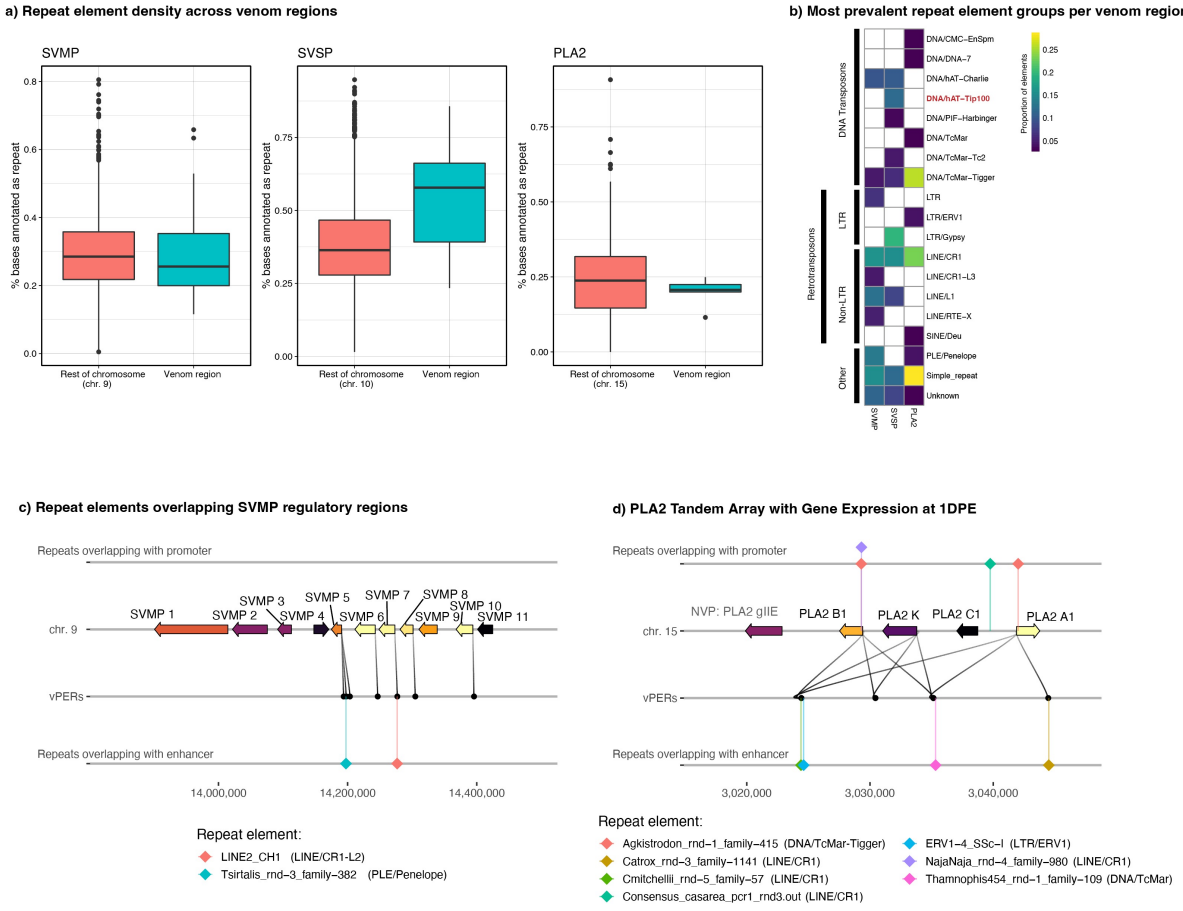


Figure S22. Characterization of repeat elements across venom gene regions. A) The percent composition of repeat elements in 10kb windows (the percent of bases per window annotated as repetitive) compared between windows that overlap a given venom gene region (blue) and those across the remainder of the corresponding chromosome. B) The top 10 most abundant repeat element types per venom gene region, shown as a proportion of all elements in that region. Brighter colors indicate a higher proportion of a given element within a particular venom gene region. Note that the DNA/hAT-Tip100 class of repeat element implicated in SVSP regulatory regions (shown in red) is not among the most abundant elements classes in SVMP and PLA2 regions. C-D) Repeat elements found to overlap significantly with C) SVMP and D) PLA2 regulatory regions (promoters and vPERs) based on Giggle analyses (one-tailed Fisher's exact test; $p < 0.05$). Genes and associated vPERs are shown for each venom cluster, with gene arrow color indicating gene expression in the venom gland at 1DPE (brighter colors indicate higher expression). Repeat elements found to overlap promoters, if present, are indicated on the top line, and those found to overlap vPER regions are shown on the bottom line.

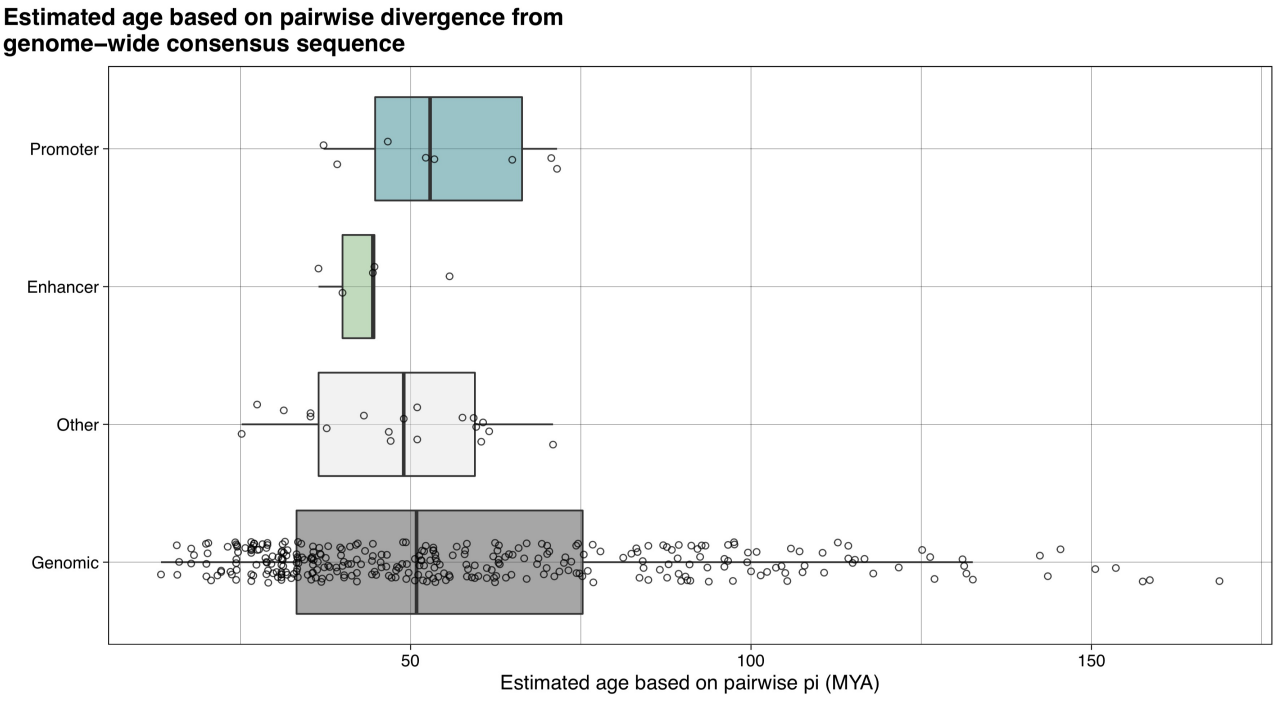


Figure S23. Estimated age of Cv1-hAT-Tip100 elements in the Prairie Rattlesnake genome. Estimated age of individual element copies based on pairwise divergence between each copy and the genome-wide consensus sequence, using the mutation rate of 2.24 x 109 following (Pasquesi et al., 2018).

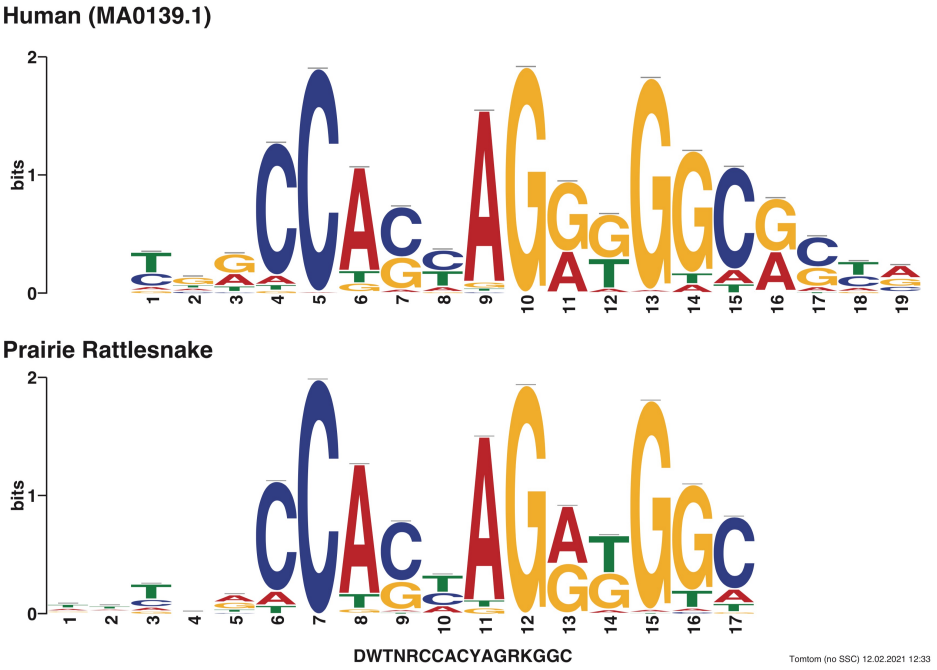


Figure S24. Comparison of the CTCF binding motif in humans (MA0139.1) and the CTCF binding motif for Prairie Rattlesnake inferred using CTCF ChIP-seq.

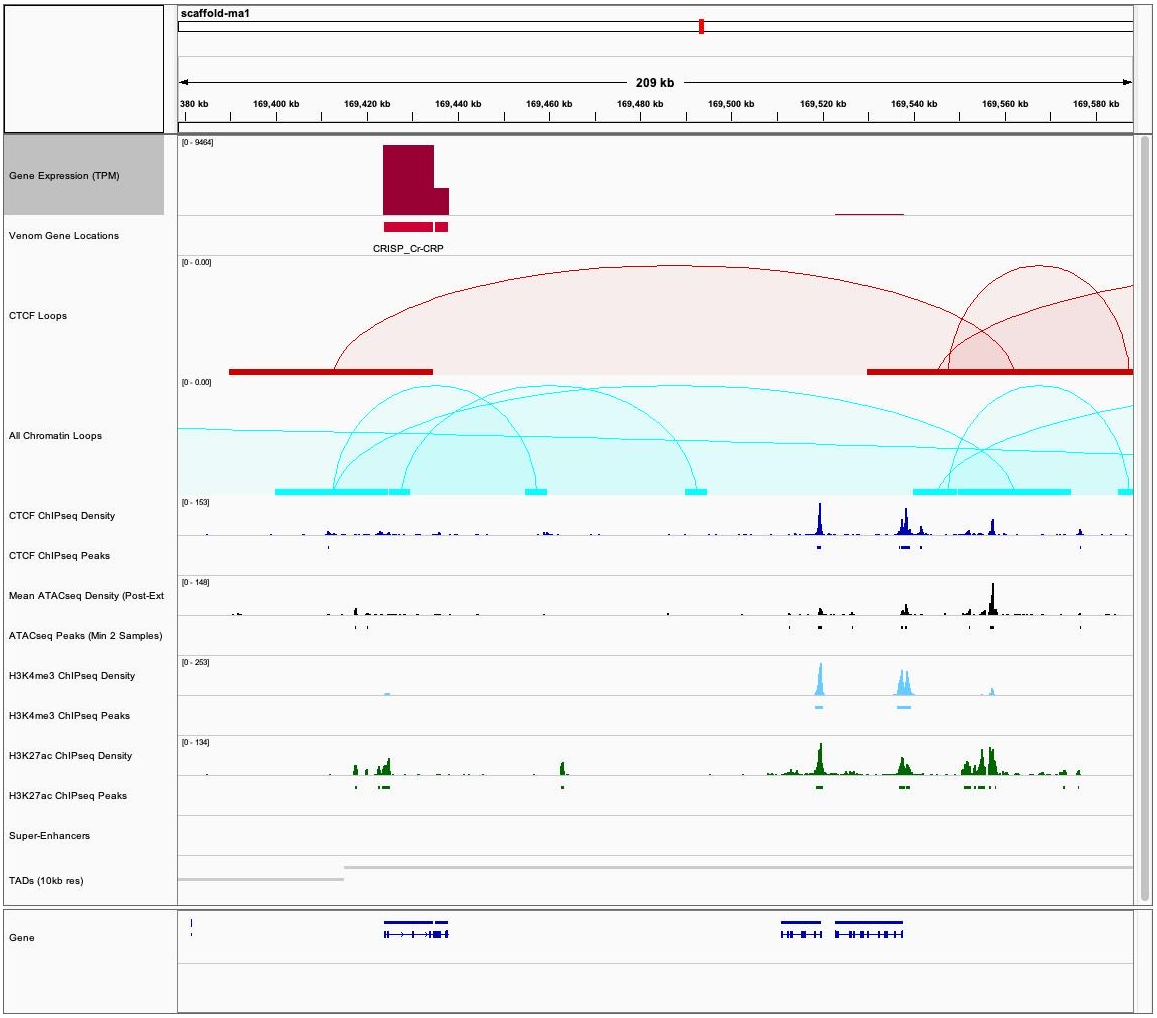


Figure S25. Inferred chromatin loops near the CRISP venom genes. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

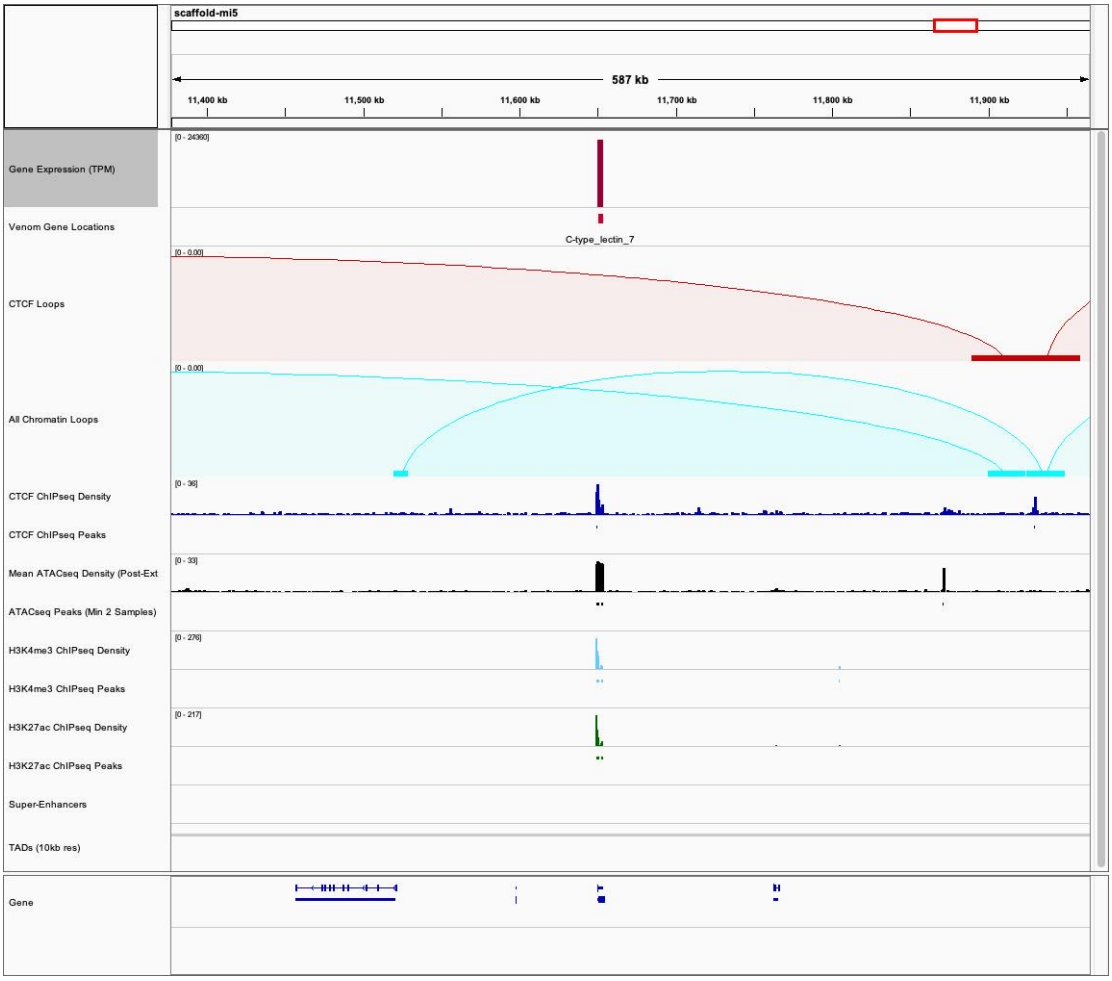


Figure S26. Inferred chromatin loops near the C-Type lectin venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

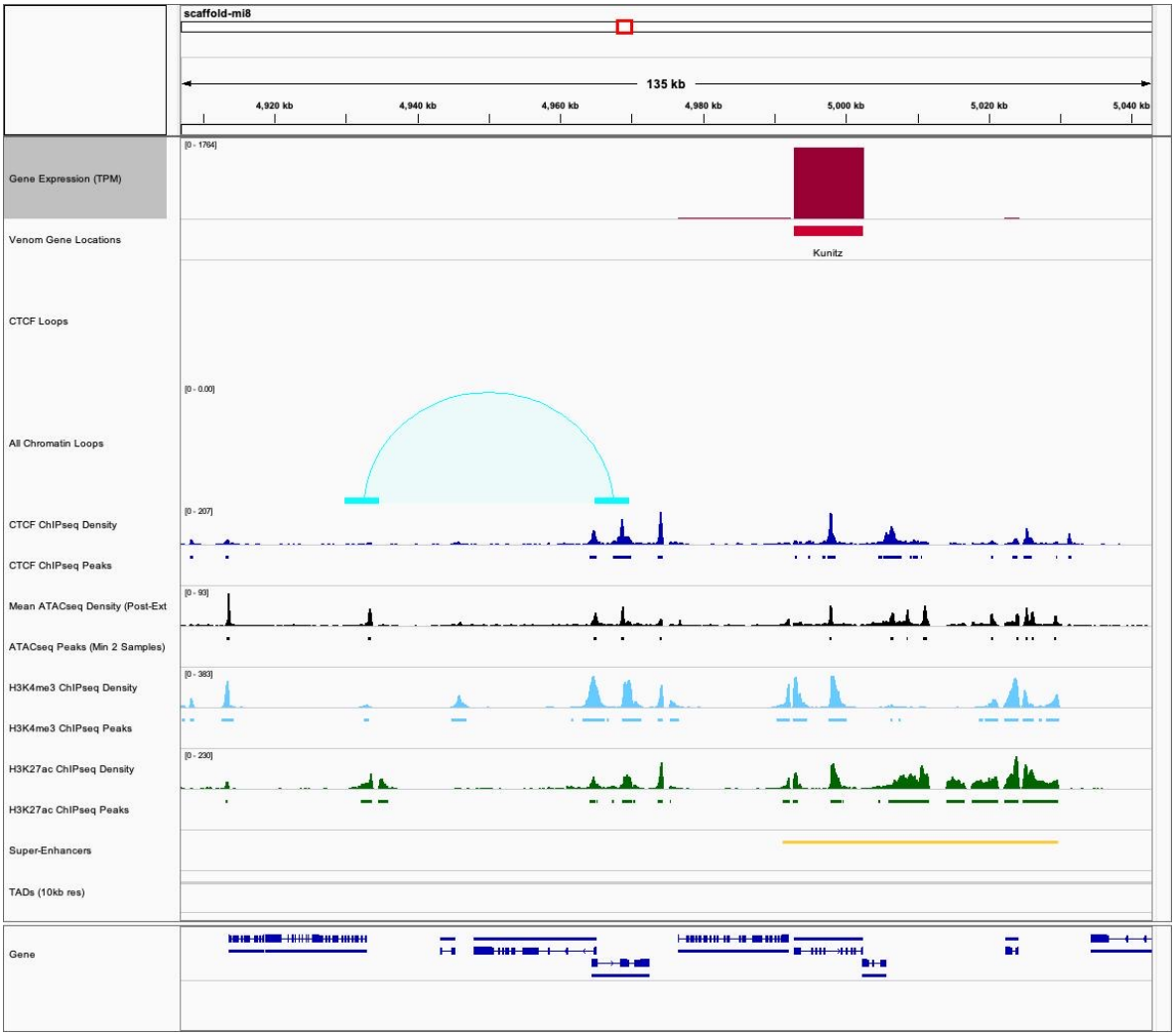


Figure S27. Inferred chromatin loops near the Kunitz venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

Figure S28. Inferred chromatin loops near the L-Amino Acid Oxidase 3 venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

Figure S29. Inferred chromatin loops near the RNA exonuclease 4 venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

Figure S30. Inferred chromatin loops near the VEGFA venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

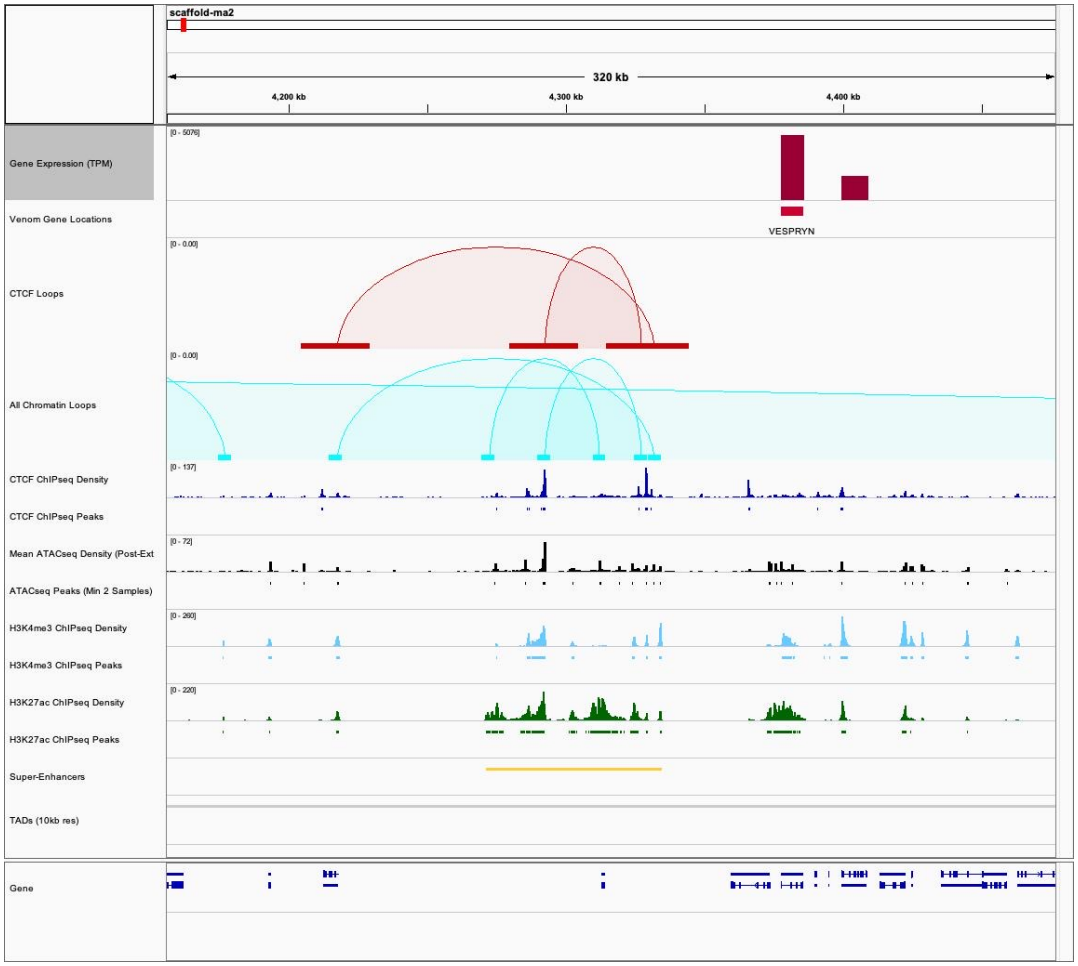


Figure S31. Inferred chromatin loops near the Vespryn venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

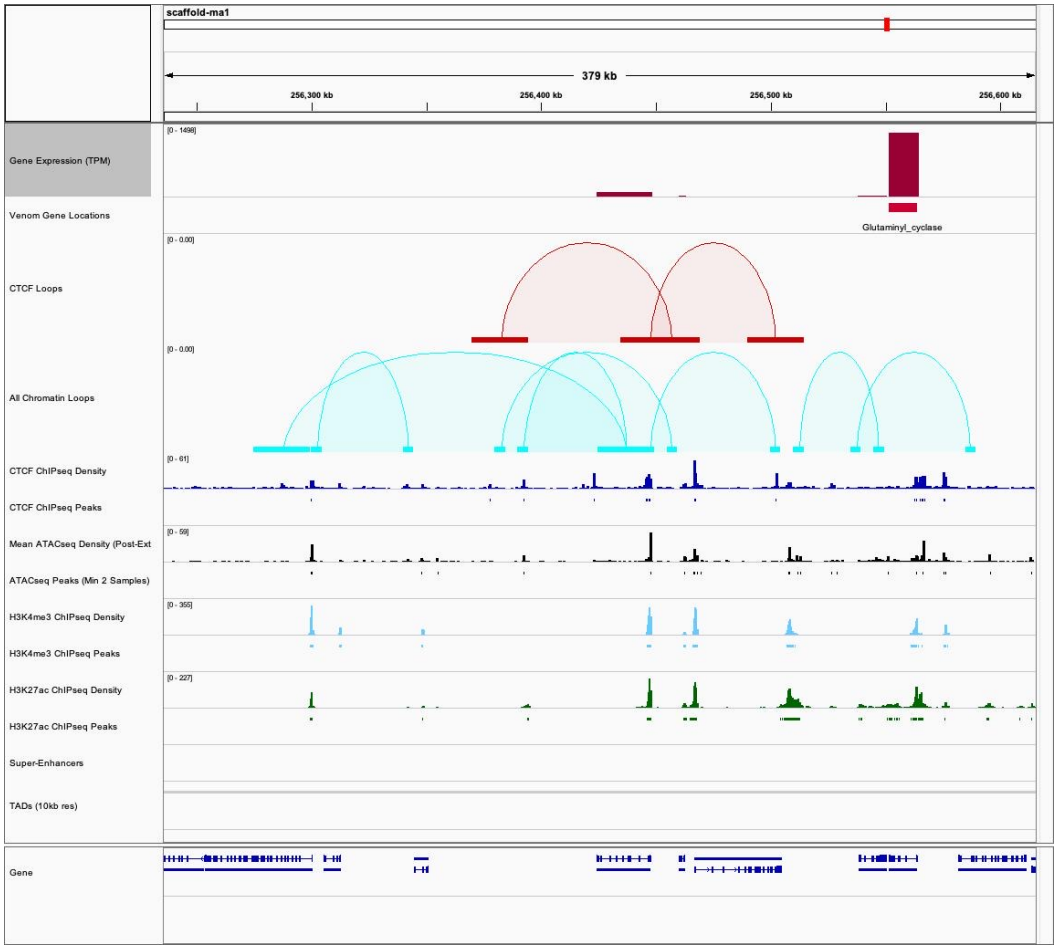


Figure S32. Inferred chromatin loops near the glutaminyl cyclase 1 (v_QC1) venom gene. Chromatin loops are shown in blue, and if present, chromatin loops with CTCF ChIP-seq peaks containing CTCF TFBS motifs near both end of the loop (CTCF Loops) are shown in red. Topologically associated domains (TADs) are shown in grey at the bottom.

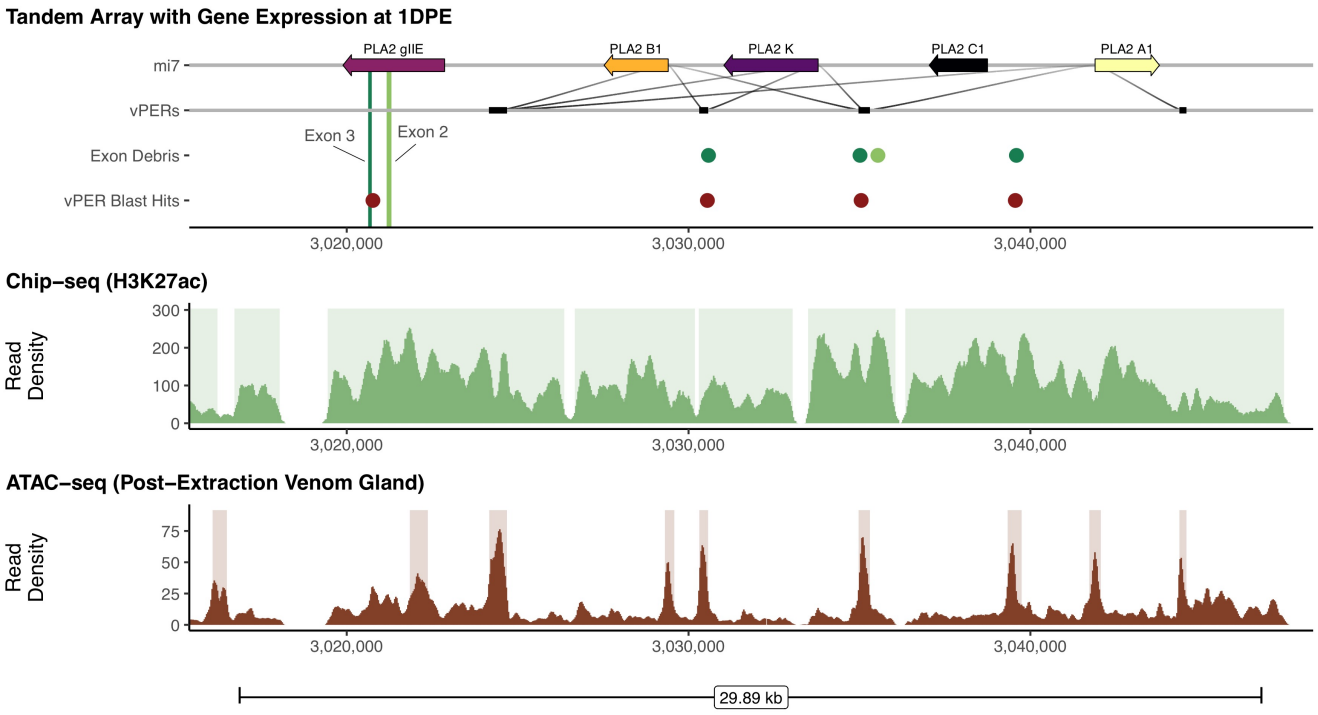


Figure S33. PLA2 enhancers may be derived from incomplete duplication of their non-venom paralog, PLA2 gIIE. The second and third exon of PLA2gIIE are marked with green bars, and the green dots in the PLA2gIIE Exon Debris row indicate exonic debris corresponding to these exon regions that resulted from partial duplication of this gene. Below, significant Blast results to the core vPER sequence for PLA2s correspond to the third exon of PLA2gIIE and multiple vPER regions. Active enhancer (H3K27ac) ChIP-seq and ATAC-seq are shown below, with peaks shown as shaded rectangles.