Diverse Gene Regulatory Mechanisms Alter Rattlesnake Venom Gene Expression at Fine Evolutionary Scales

Siddharth S. Gopalan[1], Blair W. Perry[1,2], Yannick Z. Francioli[1], Drew R. Schield[3], Hannah D. Guss[1], Justin M. Bernstein[1], Kaas Ballard[1], Cara F. Smith[4], Anthony J. Saviola[4], Richard H. Adams[5], Stephen P. Mackessy[6], and Todd A. Castoe[1,*]

[1]Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA
[2]School of Biological Sciences, Washington State University, Pullman, WA 99164, USA
[3]Department of Biology, University of Virginia, Charlottesville, VA 22903, USA
[4]Department of Biochemistry and Molecular Genetics, University of Colorado Denver, Aurora, CO 80045, USA
[5]Department of Entomology and Plant Pathology, University of Arkansas Agricultural Experimental Station, University of Arkansas, Fayetteville, AR 72701, USA
[6]Department of Biological Sciences, University of Northern Colorado, Greeley, CO 80639, USA

*Corresponding author: E-mail: todd.castoe@uta.edu.

Accepted: May 08, 2024

Abstract

Understanding and predicting the relationships between genotype and phenotype is often challenging, largely due to the complex nature of eukaryotic gene regulation. A step towards this goal is to map how phenotypic diversity evolves through genomic changes that modify gene regulatory interactions. Using the Prairie Rattlesnake (*Crotalus viridis*) and related species, we integrate mRNA-seq, proteomic, ATAC-seq and whole-genome resequencing data to understand how specific evolutionary modifications to gene regulatory network components produce differences in venom gene expression. Through comparisons within and between species, we find a remarkably high degree of gene expression and regulatory network variation across even a shallow level of evolutionary divergence. We use these data to test hypotheses about the roles of specific trans-factors and cis-regulatory elements, how these roles may vary across venom genes and gene families, and how variation in regulatory systems drive diversity in venom phenotypes. Our results illustrate that differences in chromatin and genotype at regulatory elements play major roles in modulating expression. However, we also find that enhancer deletions, differences in transcription factor expression, and variation in activity of the insulator protein CTCF also likely impact venom phenotypes. Our findings provide insight into the diversity and gene-specificity of gene regulatory features and highlight the value of comparative studies to link gene regulatory network variation to phenotypic variation.

Key words: ATAC-seq, chromatin, cis-regulatory element, CTCF, enhancer, gene regulatory networks.

Significance

The breath of factors involved in the regulation of eukaryotic genes makes it challenging to quantify their individual contributions to gene expression differences, and to identify genomic mechanisms that give rise to phenotypic variation. Here, we address this challenge by leveraging naturally existing regulatory and phenotypic variation in snake venom systems across a closely related group of rattlesnakes. Across venom genes and gene families, we find that variation in chromatin and genotype at regulatory elements play dominant roles in modulating expression. Our results provide new perspectives on the extent of standing variation that may impact gene regulatory function even at shallow evolutionary divergences in a highly adaptive trait, highlighting the diversity and specificity of the genomic mechanisms that may underlie such variation.

Introduction

Understanding how phenotypes evolve through genomic changes that modify gene regulatory interactions is central to understanding the basis of organismal diversity, and for linking variation in genotype to phenotype (Crombach and Hogeweg 2008; Romero et al. 2012; Wittkopp and Kalay 2012). However, the complexity of eukaryotic gene regulation poses many challenges for inferring how genomic variation manifests in phenotypic variation. Differences in gene expression can be driven by synergistic contributions of a variety of factors, including differences in transcription factor (TF) expression or activation (Spitz and Furlong 2012), differences in chromatin state that modulates access to cis-regulatory elements (CREs) (Buenrostro et al. 2015), variation in genotype at cis-elements that impacts TF binding (Rockman and Wray 2002; Wittkopp and Kalay 2012), and the activity of noncoding RNAs (Zheng et al. 2023). Studying how gene regulatory networks (GRNs) evolves to modulate expression of phenotypes across populations and species has the potential to provide new insights into the regulatory roles these factors play and thus provide a framework for linking regulatory network variation with trait variation. There are, however, few examples that provide baseline expectations for the relative contributions of chromatin accessibility changes, trans-factor differences, or sequence variation at CREs to gene expression differences at fine scales, such as between populations or among closely related species (e.g. Edsall et al. 2019; Barr et al. 2023). Accordingly, our understanding of which components of GRNs play predominant roles in generating gene expression differences at such fine scales, and how this regulatory architecture varies across genes, remains incomplete.

Snake venom provides a powerful system to map the relationships between genotypic, regulatory, and phenotypic variation due to the number of distinct venom gene families that contribute proteins to venom (Mackessy 2010; Tasoulis and Isbister 2017; Schield et al. 2019a; Casewell et al. 2020; Zancolli and Casewell 2020; Mackessy 2021). The diversity of venom composition across populations and species also provides comparative power to study evolutionary change at shallow scales of evolutionary divergence (Rokyta et al. 2015; Amazonas et al. 2018; Hofmann et al. 2018; Casewell et al. 2020; Colis-Torres et al. 2021). Additionally, snake venom systems are attractive models because of their direct relationships between venom gene expression, venom protein production, and venom phenotype (Casewell et al. 2012, 2013; Rokyta et al. 2015; Holding et al. 2016; Zancolli and Casewell 2020). Among snakes, the Prairie Rattlesnake (*Crotalus viridis*) has emerged as a model for studying among-population venom variation (Smith et al. 2023), and for understanding the glandular physiology and gene regulatory mechanisms associated with venom expression (Schield et al. 2019a; Perry et al. 2020, 2022; Westfall et al. 2023). Recent studies have identified candidate enhancers, promoters, TFs and TF binding sites (TFBSs) involved in venom gene regulation within this species (Perry et al. 2022) and have used single-cell approaches to confirm the roles of distinct TFs in regulating different venom loci (Westfall et al. 2023). These studies provide key foundations for exploring how differences in venom gene regulatory components may underlie the extensive variation in venom expression in *C. viridis* and related species. Notably, *C. viridis* venom phenotypes differ significantly in the primary components of their venom profile between southern and northern populations, with venom dominated by myotoxins in northern populations, versus snake venom metalloproteinases (SVMPs) in southern populations (Smith et al. 2023). Compared to *C. viridis*, closely related species (including *C. oreganus concolor, C. o. lutosus,* and *C. cerberus*) display remarkably different venom composition, including variation in expression levels of distinct venom families, as well as variation in certain paralogs within gene families (Mackessy 2010). Accordingly, this evolutionary variation provides a rich system to investigate the fundamental functional genomic underpinnings of venom phenotypic variation.

Here, we integrate multilevel functional genomic datasets and whole-genome resequencing data from Prairie Rattlesnakes (*C. viridis*) and three closely related species (*C. oreganus concolor, C. o. lutosus,* and *C. cerberus*) to survey the gene regulatory mechanisms underlying venom variation within this clade. Our sampling design is optimized to maximize phenotypic variation in venom composition across a continuum of genomic divergence in a relatively shallow phylogenetic transect of populations and species (<5 MY divergence), enhancing our ability to link changes in gene regulatory features to variation in phenotype. We use these data to explore variation in phenotype and regulatory features, such as trans-factor expression differences, chromatin and nucleotide differences at CREs, and evidence for differences in TF occupancy at CREs that exists within and between species.

We address the overarching hypothesis that that venom phenotypic variation is driven by underlying gene regulatory variation, including variable expression of relevant transcription factors, as well as chromatin state, TF occupancy, and nucleotide variation at CREs. We also hypothesize that diversity in a subset of gene regulatory network features might play consistent and dominant roles in driving expression variation, and that these patterns are consistent across all genes or paralogs within gene families. To test these hypotheses, we integrate mRNA-seq and proteomics to measure venom expression and composition, ATAC-seq data to compare chromatin accessibility and evidence of TF occupancy, and genome resequencing data to understand the contributions of CRE nucleotide differences among snake lineages and across venom genes and gene families. Our initial results

indicated that the predictive importance of regulatory features is highly gene-specific. Based on this finding, we explore several gene-specific examples in detail, which individually highlight the diversity of distinct regulatory mechanisms (or combinations of mechanisms) that appear to impact gene expression differences.

Results

Variation in Venom mRNA and Protein Expression

To quantify venom expression differences in an evolutionary context, we measured mRNA expression from both left and right venom glands of 12 individuals from four species and subspecies (supplementary table S1, Supplementary Material online). Venom genes exhibiting low expression across all samples were manually identified and subsequently excluded from all analyses (supplementary fig. S1, Supplementary Material online). We found no evidence of substantial differences in mRNA expression between left and right glands from the same individual, particularly for venom genes (supplementary fig. S2, Supplementary Material online). Therefore, we combined left and right gland expression data per individual to provide estimates of gene expression for most downstream analyses, unless otherwise noted. Our mRNA-seq data demonstrated substantial diversity in the gene expression of many venom genes across individuals, particularly myotoxin a/crotamine (hereafter myotoxin) and SVMPs (Fig. 1a). This variation is significantly greater than that observed in nonvenom paralogs (nonvenom metalloproteinases, phospholipase A2s, serine proteases, and beta-defensins; supplementary fig. S3, Supplementary Material online). Both within and across species, venom gene expression variation was the highest in myotoxin, a gene with a high degree of copy number variation (Gopalan et al. 2022) and several SVMP paralogs, the latter of which represent 9 out of the 20 most variably expressed venom genes across all samples, and 8 out of 20 within *C. viridis* (supplementary fig. S4, Supplementary Material online). This is consistent with prior evidence that proteomic variation in SVMP and myotoxin are major axes of venom variation across the range of *C. viridis* (Smith et al. 2023), which we find also applies to cross-species comparisons (e.g. *C. o. lutosus* expresses SVMPs relatively highly and myotoxin lowly, while *C. o. concolor* expresses the opposite profile). Other venom gene families with highly variable expression include phospholipases A_2 (PLA2s) and snake venom serine proteases (SVSPs; supplementary fig. S4, Supplementary Material online).

Venom proteomic profiles were broadly consistent with venom toxin abundance inferred from mRNA-seq data (Fig. 1b), and a principal component analysis (PCA) of venom proteome composition across individuals separated species primarily by PC1 (73.12% variance explained), and populations of *C. viridis* by PC2 (12.02% variance explained; supplementary fig. S5, Supplementary Material online). To estimate the relationship between venom gland-derived venom gene mRNA expression and venom protein abundance, we followed the method of Rokyta et al. (2015) to scale and transform count-based gene expression (VST-normalized counts) and protein abundances (estimated from chromatographic peak intensity) using the centered-log ratio transform for each venom gene and its matched protein per individual ($R^2 = 0.35$) (Fig. 1c).

Venom-associated TF Expression Correlates With Venom Variation

As an initial step to understand how differences in gene regulatory components explain venom gene expression, we focused on differences in trans-regulatory factor (TF) expression. We find that the top ranked TFs by expression are also often implicated in venom regulation (Perry et al. 2022; Westfall et al. 2023). TF expression varies considerably both within and among species, especially when compared to a background set of TFs not implicated in venom regulation (Fig. 2a). We also used DESeq2 (Love et al. 2014) to assess differential expression within and across species. However, we did not find evidence for differentially expressed TFs within different *C. viridis* populations. Across species, we did find evidence for the differential expression of 15 TFs: DLX3, EOMES, GABPA, GATA4, GATA6, HNF4A, MYF6, MYOG, NR1H4, PAX1, PBX1, SOX13, TBX19, TFAPC2, and VAX1, which, along with known TFs of importance (Perry et al. 2022) we include for downstream analyses of TF binding analysis.

This suggests that venom expression variation may be partly driven by differences in the expression of trans-regulatory factors, especially across species. To investigate this further, we tested for evidence of distinct co-expression modules between populations and species which may correlate with species identity by analyzing global venom gland mRNA data (including all genes) using WGCNA (Langfelder and Horvath 2008), through the estimation of module-gene significance values (Fig. 2b). Here, we analyzed left and right venom gland samples separately as biologically relevant replicates to increase power to detect co-expression modules. These modules were dominated by TFs, and modules with high scores in *C. viridis* include many TFs previously implicated in regulating *C. viridis* venom composition (Perry et al. 2022; supplementary table S2, Supplementary Material online). In addition to venom-associated TFs, the top *C. viridis* co-expression module also included venom genes (including CRISPs, SVSPs, SVMPs, and CTLs), chromatin regulators, and other TFs related to ERK and UPR signaling—key pathways hypothesized to coordinate venom expression (Perry et al. 2022; Westfall et al. 2023). We find each species is associated with distinct co-expression modules, indicating evolutionary lability in trans-acting factor expression (supplementary fig. S6, Supplementary Material online). Differences between genes

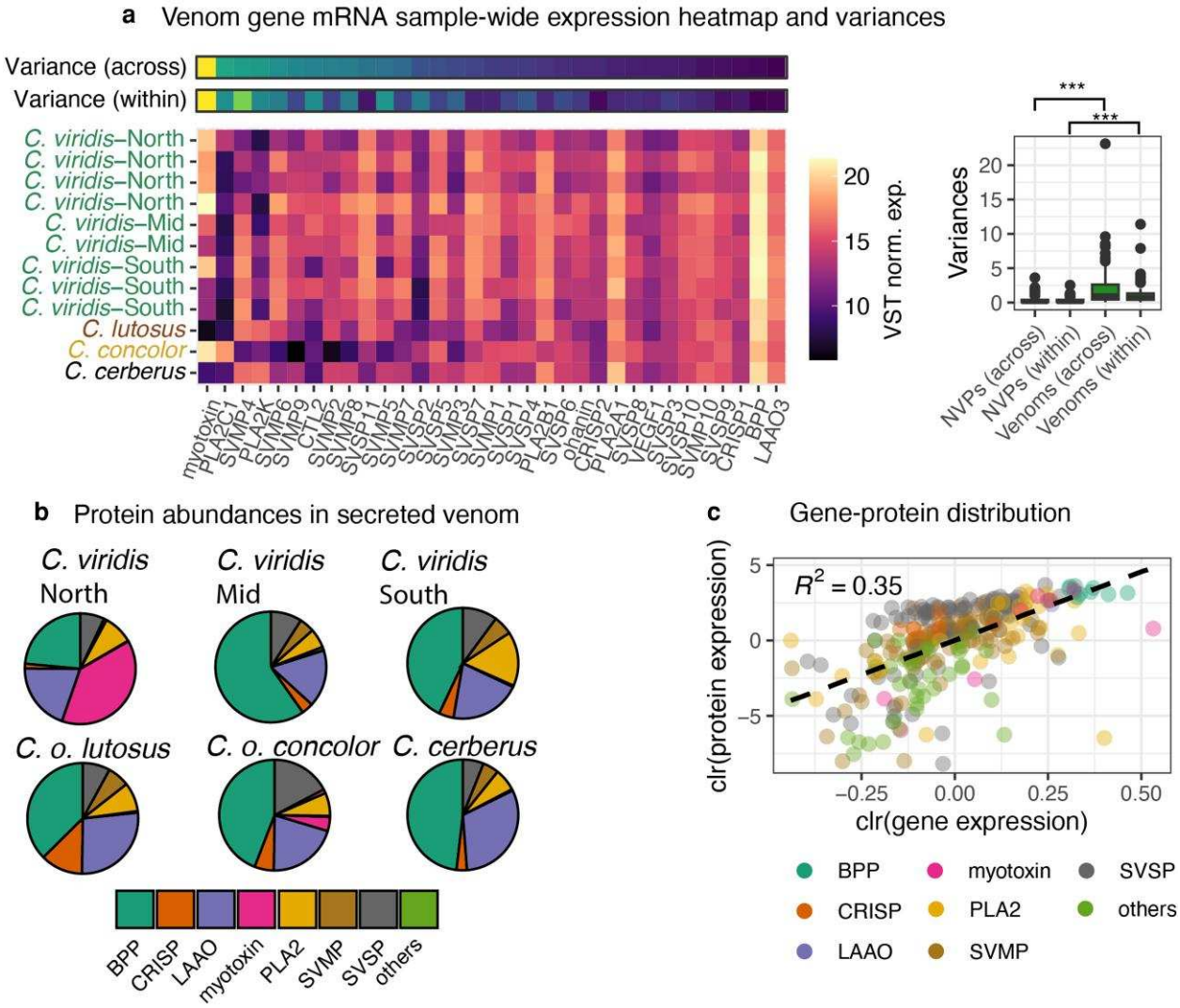

Fig. 1.—Toxin genes and their derived proteins display high expression variation. a) Venom gene expression for all individuals displayed as a heatmap. Variance, across all samples (across) and within *C. viridis* (within), in gene expression is shown as two rows above the heatmap, with brighter colors indicating higher variance and darker colors lower. Note that variances have been square root transformed to aid visualization; unscaled variances can be found in supplementary fig. S4, Supplementary Material online. To the right, the boxplot shows expression variance for venom genes and select nonvenom paralogs (a disintegrin and metalloproteinases (ADAMs), phospholipase A2s, beta-defensins and serine proteases; collectively NVPs; full list found in supplementary fig. S3, Supplementary Material online), both across all samples and within *C. viridis*. The asterisks represent *P*-values of a 2-sample t-test comparing groups (*** $P < 0.01$). Only significant comparisons are shown. b) Averaged venom protein abundances for each sampling group are displayed as pie charts. c) Linear correlation between protein and gene abundances. Gene and protein abundances were transformed using centered-log ratio (clr) transformation.

comprising these species-specific modules include venom-associated TFs (Perry et al. 2022), as well as TFs without prior known links to venom regulation (Fig. 2b).

To test for evidence that the expression of TFs was predictive of venom expression, we calculated gene–gene expression correlation coefficients between venom-associated TFs and venom gene expression across all samples and find several TFs whose expression is highly predictive of the expression of specific venom genes (Fig. 2c). For example, expression of myotoxin is strongly correlated with the expression of XBP1 ($\rho = 0.80$; *P*-value = 0.001) and ATF4 ($\rho = 0.79$; *P*-value = 0.002), the latter of which has been previously predicted to have a binding site in the myotoxin promoter (Gopalan et al. 2022). Additionally, FOS and DDIT3 are significantly (*P*-value < 0.05) positively correlated with the expression of four distinct venom genes: BPP, SVSP7, SVSP10, and SVSP11 ($\rho = 0.75$ to 0.82).

Broad Evidence that CRE Chromatin State, SNPs and TF Binding Underlie Venom Variation

In a prior study, we integrated ChIP-seq, ATAC-seq, and Hi-C data to infer CREs associated with venom loci in

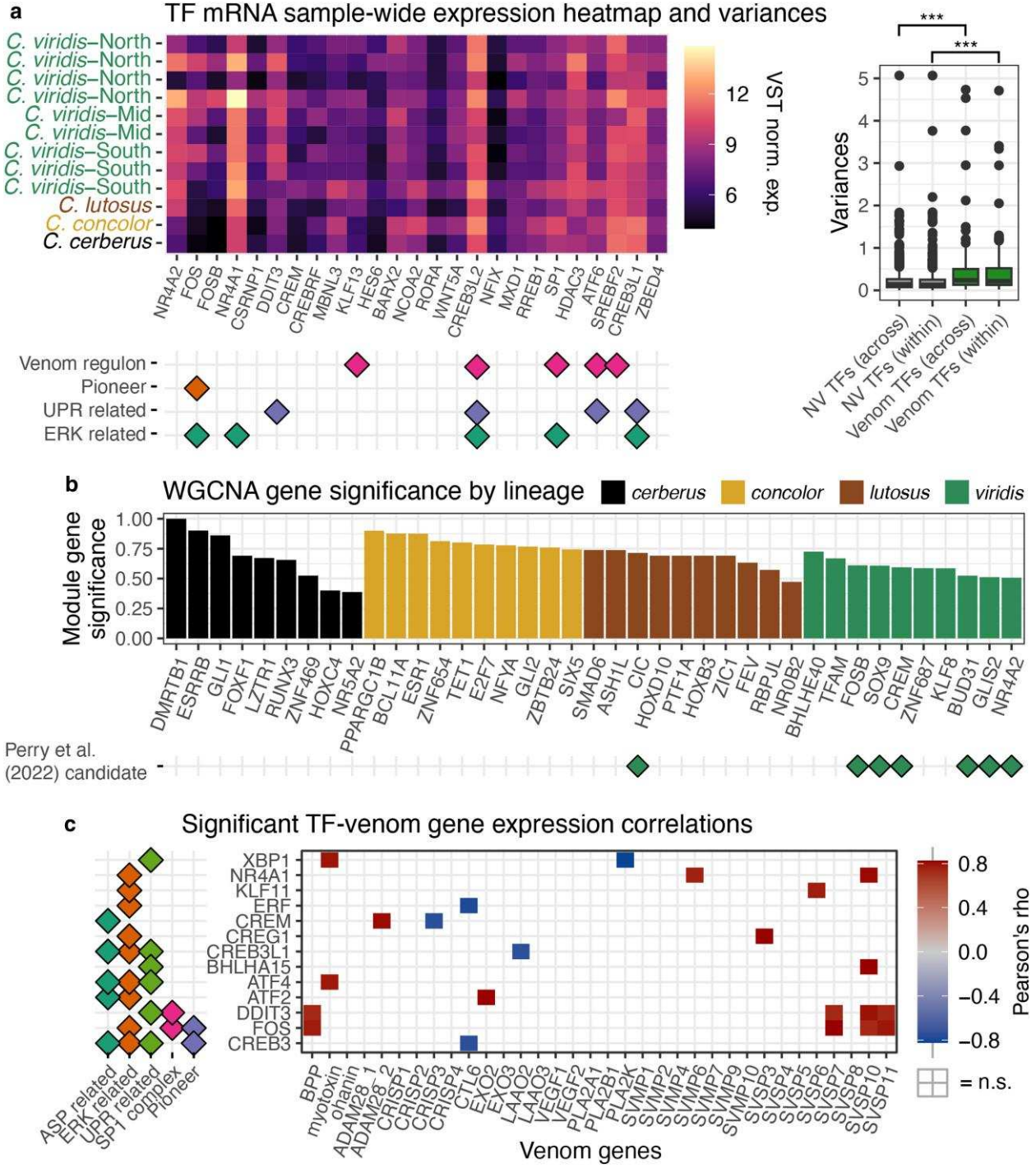

Fig. 2.—TF expression varies across lineages, suggesting role of trans-factor expression in venom variation. a) The top 25 TFs sorted by expression variance across all samples. To the right, the boxplot shows VST expression variances for all venom-associated TFs (from Perry et al. 2022; N = 161) and TFs not associated with venom, both within *C. viridis* and across all samples. The asterisks represent *P*-values of a 2-sample t-test comparing groups (***$P < 0.01$). b) WGCNA gene significance for TFs within the co-expression module that is most significant for each lineage variable. The functional annotations below 2a were taken from Perry et al. (2022) for pioneer TFs, UPR and ERK related TFs, and Westfall et al. (2023) for venom regulons. c) The matrix of Pearson's correlation coefficients between expression of candidate TFs and venom genes are displayed only for significant correlations. Pearson's rho scalebar on the right represents positive negative correlations. An uncolored box represents not significant (n.s.) correlations. Functional annotations come from Perry et al. (2022) and Westfall et al. (2023).

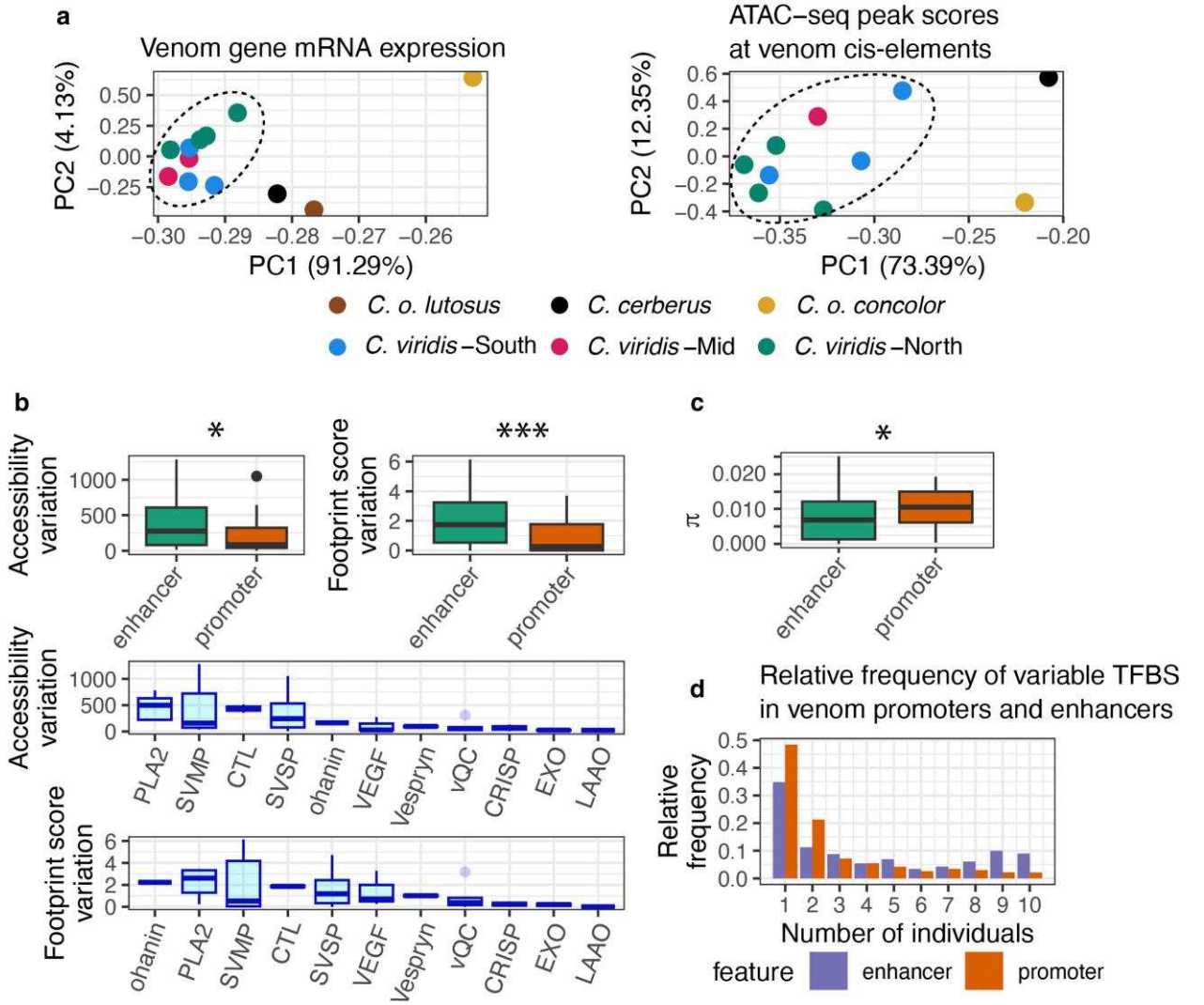

Fig. 3.—Abundant chromatin accessibility, TF binding and standing nucleotide variation exist in venom CREs. a) PCAs of venom gene mRNA expression and ATAC-seq peak scores at venom cis-elements (promoters and enhancers of venom genes) demonstrate variance partitioning across populations and species, corresponding to the two main PC axes. The dashed line encircles *C. viridis* samples. b) Chromatin accessibility and peak accessibility variation for enhancers and promoters. The variance sorted accessibility and footprint scores across venom gene families are shown below. c) Nucleotide diversity (π) for venom enhancers and promoters. d) Frequency of variable TFBSs within enhancers and promoters across samples. This is interpreted in a similar manner to a site frequency spectrum, where each bar represents the fraction of TFBSs that are shared by that many individuals. Asterisks above boxplots indicate statistical significance for parametric 2-sample t-tests: $*P < 0.05$; $***P < 0.001$.

C. viridis (Perry et al. 2022), which we use here as a base set of known CREs for downstream analyses. To investigate the roles of cis-regulatory feature variation, we first characterized differences in ATAC-seq derived chromatin accessibility, ATAC-seq derived TF footprint score (likelihood of TF occupancy) within venom gene CREs, and genotype derived nucleotide variation at venom gene CREs. The similarity in the PCAs of chromatin accessibility at venom gene CREs and of venom gene expression suggests that these two metrics broadly covary according to population ancestry (Fig. 3a). To further dissect the relevance of specific types of CRE variation, we quantified variation in CRE accessibility and footprint scores at promoters and enhancers and find that enhancers consistently showed greater variation in both accessibility and TF binding compared to promoters (Fig. 3b). We also find that the CREs of venom gene families that show the greatest accessibility and footprint score variation are those that displayed the highest and most variable expression, including PLA_2s, SVMPs, SVSPs, and CTL2, pointing to a high-level correspondence between mRNA-seq and ATAC-seq data (Fig. 3b; supplementary fig. S4, Supplementary Material online). Despite this, we do not find statistical evidence that variation in accessibility or in footprint scores are linearly correlated with variation in

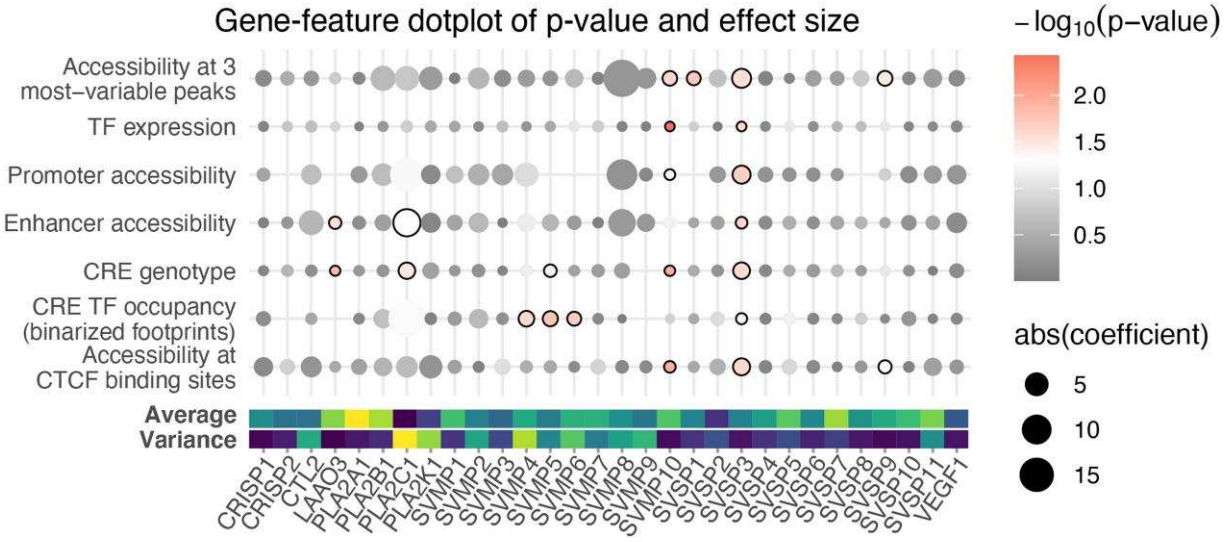

Fig. 4.—Linking toxin gene expression and regulatory variation. Results of the linear modeling on a gene-by-gene basis. Absolute values of regression coefficients and log-transformed *P*-values from multiple linear regression are shown as point size and point color respectively. Absolute values were used to assess only the effect size of the inputs. The color scale shifts from gray to red at the point of significance ($P < 0.05$). Points where the correlation is significant are also outlined in black. Sample-wide average gene expression and gene expression variance are shown as colored bars below, where brighter colors indicate higher values.

gene expression, suggesting multifactor, and nonlinear interactions play a larger role.

Because nucleotide variation at CREs can impact TF binding and thus gene regulation, we also assessed nucleotide diversity (π) from sample-wide single-nucleotide polymorphisms (SNPs) detected at venom gene CREs (Fig. 3c). This subset of SNP calls were of high quality, with an average cross-sample depth of 54.2, and average VCF quality score of 81 (probability of base call error $\approx 1/10^8$ on average; supplementary table S3, Supplementary Material online). We find that promoter sequences tended to be more variable than enhancers, despite enhancers having greater variation in chromatin accessibility and TF occupancy (Fig. 3b). Of 41 predicted venom enhancers and 50 venom promoters, only 7 enhancers showed no genetic variation ($\pi = 0$) across all individuals, 3 of which were SVSP enhancers (supplementary fig. S7, Supplementary Material online). To assess the functional implications in cases where we detected genetic variation, we predicted variants in CREs that modified the presence or the absence of TFBS and used this to assess the frequencies of these variants across individuals (Fig. 3d). We find that while nucleotide variation affecting TFBSs is common (CRE variants affect the presence and absence of 9395 TFBSs at venom gene CREs), 47% of variable TFBSs at promoters and enhancers are unique to a single sampled individual, highlighting the extensive variation in CRE sequences that exists across populations and species that is likely relevant to variation in venom expression.

Distinct Types of Regulatory Feature Variation Explain Expression of Distinct Venom Genes

Considering the diversity and high dimensionality of regulatory features that may affect gene expression, we first used phylogenetic PCA to reduce dimensionality of regulatory feature variation, then applied multiple linear regression using these principal components as predictor variables to identify what types of regulatory variation are associated with gene expression variation at a broad scale. These features include accessibility, both genotype and TF occupancy at previously identified CREs (Perry et al. 2022), expression of venom-regulating TFs, and accessibility at other potential cis-elements such as variably accessible peaks and binding sites of the insulator protein CTCF across venom gene clusters. Quantifying accessibility at CTCF loci is important in understanding potential variation in the structure of topologically associated domains, which can cause expression differences between physically adjacent venom genes (Perry et al. 2022). As a prerequisite for modeling, we ensured that candidate genes had a well-understood genomic context (i.e. sequencing of the adjacent region in the reference, enhancer predictions and CTCF predictions). This precluded a focus on myotoxin or BPP, which have genomic contexts that have yet to be well resolved. We explored relationships between regulatory variation and venom expression on a gene-by-gene basis (Fig. 4). We find that the most predictive regulatory characteristics are highly gene-specific, although CRE genotype, TF occupancy at previously identified venom gene CREs, as well as de novo identified (previously unannotated) ATAC-seq peaks

predict expression for most venom genes. TF occupancy of venom-regulating TFs for example correlates with the expression of three physically adjacent SVMPs paralogs (SVMP4, SVMP5, and SVMP6). Some genes, such as LAAO3 and SVSP9, correlate with only a few specific regulatory feature types, while other genes (e.g. SVMP10 and SVSP3) respond to a suite of features. Nonsignificant model results do not seem to be related to low gene expression in most cases, though high feature coefficients appear to be the result of high expression variance in at least the case of PLA2C1. Overall, our linear models highlighted a subset of venom loci for which gene expression was strongly associated with variation in regulatory factors, and in some cases, with gene-level specificity. The results of the linear modeling provided a set of potential genes of interest with respect to understanding the effects of specific regulatory inputs, and so we further investigated several specific venom loci in gene-specific vignettes.

SVMP6 Expression Responds to Variable Trans-factor Binding at its Enhancer

Considering that SVMP gene and proteomic expression is highly variable within and between species (Fig. 1a, and Smith et al. 2023), we compared chromatin accessibility across samples at the SVMP gene cluster and find that variance in accessibility tends to be much higher at enhancers than promoters (Figs. 3b and 5a; supplementary fig. S8, Supplementary Material online). To investigate these relationships further, we focused on the SVMP paralog SVMP6, which showed highly variable gene expression (Fig. 1a, supplementary fig. S4, Supplementary Material online), high levels of nucleotide diversity at its enhancer (supplementary fig. S7, Supplementary Material online), and significant correlations between SVMP6 expression and TF occupancy at CREs based on linear modeling (Fig. 5b). None of these patterns are confounded by excessive structural variation at the enhancer (supplementary fig. S9, Supplementary Material online). We first assessed TFBS occupancy differences (based on ATAC-seq footprint scores) between samples at the SVMP6 enhancer by quantifying the total number of binding events for each TF and find evidence for variable TFBS occupancy across samples of *C. viridis*, and very low predicted levels of TFBS occupancy in *C. o. concolor* that corresponds with very low SVMP6 expression in this species (Fig. 5c). The high degree of variation in TFBS occupancy suggest that there may be differences in cell populations with respect to TF binding, or that TFs may cooperatively bind to activate the enhancer. ATAC-seq footprint scores suggest that TFs such as GATA6 and GATA4 are bound only in *C. viridis*, while others such as PITX2, EHF and DDIT3 vary both in frequency of binding and presence across samples. This indicates that variation in TF binding at the SVMP6 enhancer is indeed associated with variation in gene expression across samples within and among species.

To investigate how evidence for variable TF binding at this enhancer may be related to the nucleotide variation at this locus, we focused on enhancer variants at known TFBSs that were also associated with differences in estimated TF occupancy across samples. This highlighted two variants, one SNP and one indel, which together impact TFBSs of as many as six TFs in the *C. viridis* samples and are absent in *C. o. concolor* and *C. cerberus* (Fig. 5d). These differentially bound TFs include two pioneer transcription factors (GATA4 and FOS) that can initiate regulatory events by opening chromatin (Cirillo et al. 2002; Fleming et al. 2013). This example highlights the roles of TF occupancy differences, which can be driven by allelic variants at enhancers, as a mechanism leading to differential gene expression within and between species.

SVSP9 Expression Responds to Accessibility at Enhancers, Silencers, and Insulation by CTCF

SVSPs represent a major component of rattlesnake venoms and show high degrees of gene expression variation and ATAC-seq variation across our samples compared to other venom genes (Figs. 1a and 3b, and 6a). For one SVSP paralog, SVSP9, our linear modeling suggests its expression is significantly correlated with accessibility at a known CTCF-binding site, and accessibility at additional nonannotated loci (loci with highly variable accessibility not previously identified as a CREs; Fig. 6b). To further investigate these loci, we examined ATAC-seq density across samples at the entire SVSP locus (Fig. 6a), and at the three ATAC-seq peaks within this gene cluster that showed significant ($P < 0.05$) correlations between their accessibility and SVSP9 gene expression (Fig. 6c). For both regions not previously annotated, we find moderately strong individual correlations ($R^2 > 0.5$) between their accessibility and SVSP9 expression, but with opposing effects, suggesting one may represent a putative enhancer while the second may represent a putative silencer (Figs. 6c and d). We also find evidence that accessibility at a previously predicted binding site for the insulator protein CTCF (Perry et al. 2022), located between the promoter of SVSP9 and its putative enhancers, is negatively correlated with SVSP9 expression (Figs. 6a c, and d). These findings provide evidence for how gene expression may vary across populations and species through the modulation of chromatin accessibility at CREs through both positive (enhancer) and negative (silencer and CTCF) gene regulatory interactions. Notably, these findings also highlight the potential role of the insulator protein CTCF, through its regulation of chromatin loops and enhancer–promoter interaction, in generating inter-population and inter-species gene expression diversity (Fig. 6e).

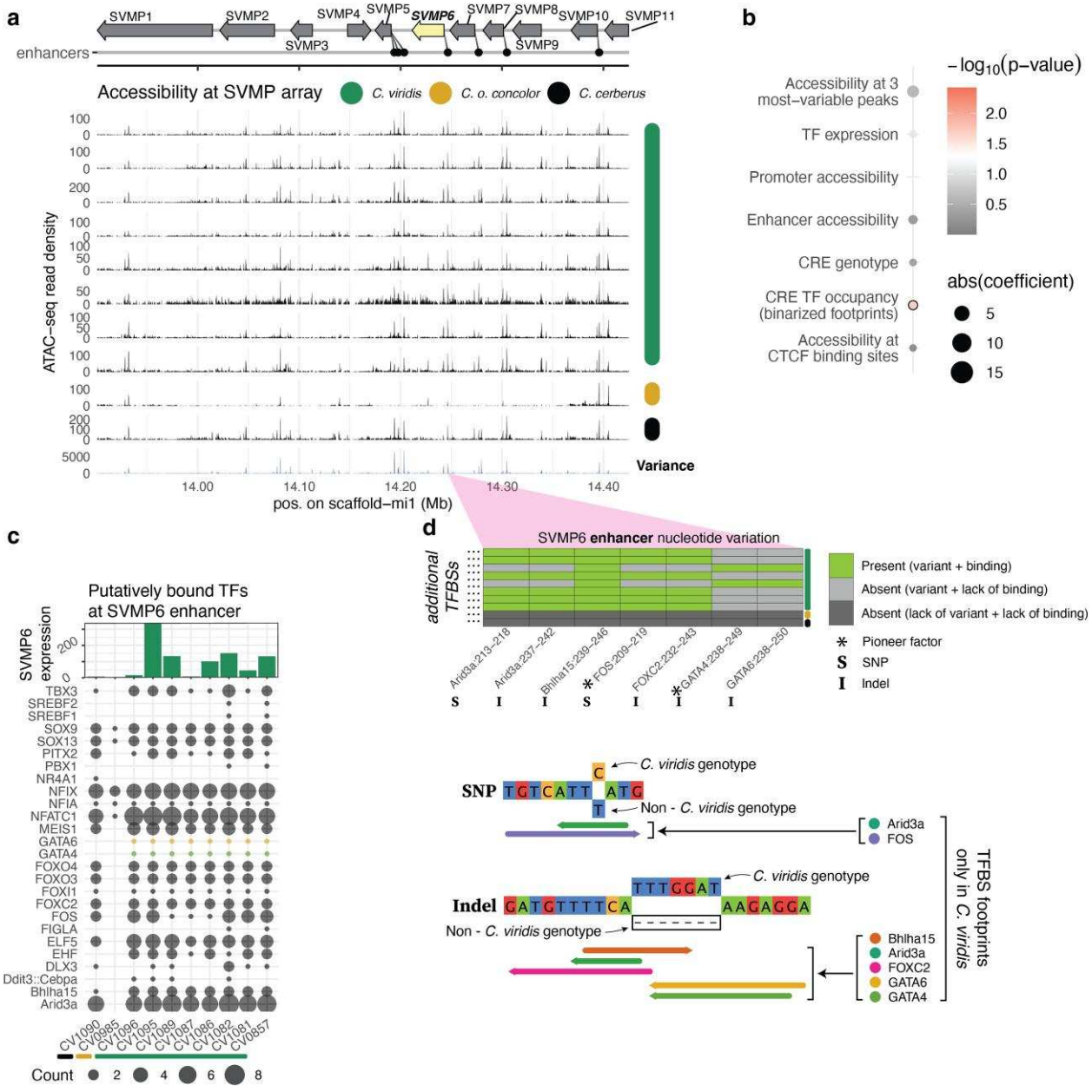

Fig. 5.—Nucleotide variation causes TF occupancy differences at enhancer, driving SVMP6 expression variation. a) The SVMP gene array and enhancers redrawn from Perry et al. (2022). Venom gland ATAC-seq for *C. viridis* and non-*C. viridis* individuals are shown as read pileup tracks. Variance in ATAC-seq density is shown as the bottom-most track. b) Results of multiple linear modeling for SVMP6 redrawn from Fig. 4. The significant feature is circled in black. c) TF binding frequency at the SVMP6 enhancer is shown, with SVMP6 expression displayed as a histogram at the top. GATA4 and GATA6 are highlighted with different colors. The expression is shown as DESeq2-normalized counts in thousands. d) The SNP and indel variants that modify the TF occupancy at TFBS sequences is shown for TFBS sequences in the SVMP6 enhancer. Below this, TFBS motifs which are affected by the SNP and indel variants are drawn onto the sequence, as well as the genotypes of *C. viridis* and non-*C. viridis* individuals. The direction of each motif is indicated by the arrow, and individual colors represent separate TFBSs.

Variation in Myotoxin Expression is Predicted by TF Binding and Expression

Though the genomic context of myotoxin remains poorly resolved, which has prevented identification of distal regulatory loci, it is notable for being the most variably expressed venom gene across our sampling (Fig. 1a). Our transcriptomic data identified strong correlations between expression of myotoxin and two TFs, ATF4, and XBP1 (Fig. 2c). The promoter sequence is known and is completely conserved across sampled individuals (supplementary fig. S7,

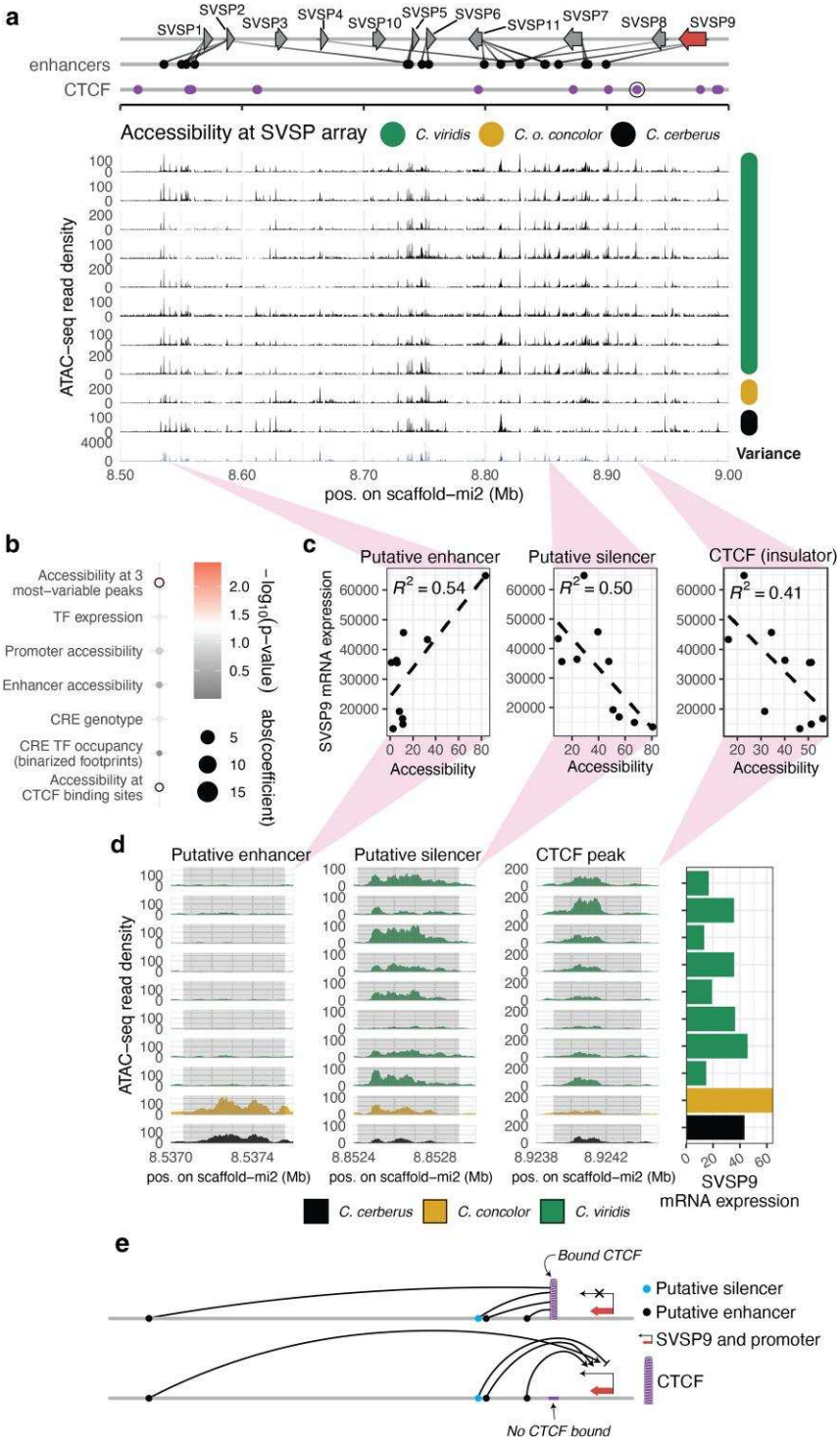

Fig. 6.—*De-novo* and CTCF-bound loci explain SVSP9 expression. a) The SVSP gene array and its predicted enhancers (redrawn from Perry et al. 2022), with CTCF binding inferred from ChIP-seq (Perry et al. 2022) shown below. The locus with accessibility that was significantly correlated with SVSP9 expression is circled. Venom gland ATAC-seq for *C. viridis* and non-*C. viridis* individuals are shown as read pileup tracks. Variance in read density is shown as the bottom-most track. b) Results of linear modeling for SVSP9, redrawn from Fig. 4b. The significant features are outlined with a black circle. c) Linear regressions between chromatin accessibility at the three loci of interest (a putative enhancer, silencer and a CTCF-bound locus) and SVSP9 gene expression. All linear models are significant at $P < 0.05$. d) Accessibility landscapes at the loci of interest are shown with a SVSP9 gene expression histogram shown at the far right. The light gray rectangles show the location of ATAC-seq peaks called by MACS2. Peaks have been centered and length standardized. e) A proposed model for how SVSP9 gene regulation responds to various input loci, and how this may be inhibited by CTCF binding.

Supplementary Material online), which, based on ATAC-seq derived TF footprint scores, does not contribute strongly to TF binding differences (supplementary fig. S10, Supplementary Material online). The promoter is predicted to be bound by ATF4 in all samples with accessible chromatin, and promoter accessibility corresponds with gene expression (supplementary fig. S10, Supplementary Material online). While no evidence of XBP1 binding was detected in the promoter, it is possible that it may bind an enhancer that has yet to be identified, form a complex with other TFs and thus not leave detectable chromatin footprints, or play a role in higher-level regulation of ATF4 or other myotoxin-regulating factors.

SVSP2 Expression Knocked out by Individual-specific Enhancer Deletions

In contrast to the CREs of other venom gene families, SVSP enhancers generally have very little or no nucleotide diversity (supplementary fig. S7, Supplementary Material online). Although our linear modeling provided no clear evidence of strongly associated genomic features, we find that other non-modeled features (e.g. structural variants) may be relevant (supplementary figs. S11 and S12, Supplementary Material online). The SVSP2 locus stood out as it was among the most variably expressed venom genes in *C. viridis* (supplementary fig. S4, Supplementary Material online), yet its two adjacent enhancers (PER17 and PER 17) showed no SNP variation across samples (supplementary fig. S7, Supplementary Material online). To test for potential effects of larger structural variation, we analyzed genome resequencing read density for *C. viridis* individuals versus the reference genome and find evidence for a several kilobase deletion affecting these enhancers in two *C. viridis* individuals from southern latitude populations (supplementary fig. S12, Supplementary Material online), which corresponds with low expression of this gene in these individuals (supplementary fig. S11, Supplementary Material online). These results highlight a case where gene expression variation may occur through the action of larger effect structural variation that exists among populations within species.

Discussion

While the rapid evolution of GRNs and the subsequent changes in gene expression are likely major drivers of adaptation and functional divergence (Wittkopp 2007; Emerson and Li 2010; Thompson et al. 2015), identifying the relative contributions of distinct regulatory components to gene and gene family expression variation, and ultimately phenotypic variation, remains challenging (Romero et al. 2012). Snake venom systems provide a uniquely powerful system, with extensive variation in venom gene expression in multiple gene families across closely related populations and species, to identify how variation in gene regulatory components contributes to gene expression variation. The ability to simultaneously measure matched protein, mRNA, and regulatory variation from the same individual during venom production affords the opportunity to more clearly link relationships between phenotype, gene expression, and regulatory variation in a comparative experimental framework. We leveraged this system here to highlight remarkable fine-scale evolutionary variation underlying phenotypic variation in a keystone adaptive trait (venom), and to further link specific mechanisms of regulatory variation to phenotypic variation.

We find that chromatin accessibility at CREs, CRE genotype variation, and predicted TF binding all influence gene expression, but to varying degrees across specific genes and gene families. Much of this is driven by high levels of nucleotide and accessibility variation at venom gene CREs, both between and even within species. We also find evidence that the specific types of gene regulatory components that contribute to venom expression variation are not only diverse but are also remarkably gene and gene family specific. In addition to canonical expectations that chromatin, TF-CRE interactions, and CRE genotype underly phenotypic variation, we also find evidence that trans-regulatory factor (i.e. TF) variation and variation in the action of the insulator protein CTCF may also play major roles in generating within and between species expression variation. Broadly, these findings establish expectations that even at shallow levels of divergence, a diversity of regulatory mechanisms may shape phenotypic variation, and that distinct genomic mechanisms may often dominate the modulation of gene expression for particular genes and gene families.

Roles of Nucleotide, Chromatin Accessibility and TF Variation

Considering the fine scale of evolutionary divergence surveyed here, we observed notably high degrees of nucleotide diversity at venom gene CREs that provide substantial "raw material" for generating variation in TF binding and chromatin accessibility that may impact venom gene expression. Indeed, we find evidence that venom gene expression is frequently related to CRE chromatin accessibility as well as CRE genotype at these venom loci, likely because both factors are key determinants of TF occupancy (Wittkopp and Kalay 2012). We showcase the regulation of the venom metalloproteinase SVMP6 to demonstrate how mutations influencing expression can confer species-specific TF binding, a pattern supported by linear modeling of regulatory network effects.

TF binding and regulatory activity can also vary depending on the expression of the transcription factors themselves. Our results suggest different suites of co-expressed TFs, many of which have been previously implicated in venom regulation, follow population- and species-specific trends, implying that distinct venom-regulating TF expression also contributes to

venom gene expression variation. Based on gene expression correlations, some TFs, such as the pioneer factor FOS (Fleming et al. 2013) and DDIT3 appear to co-regulate venom genes. Both TFs are components of the AP-1 TF complex, a major regulatory complex stimulated by venom depletion (Luna et al. 2009) and are known to physically interact (Oughtred et al. 2021). Additional correlation-based evidence comes from the myotoxin gene, which correlates with the expression of ATF4 and XBP1. These findings are notable because they highlight the correspondence between inferences from transcriptomic correlations and independent inferences of TF-CRE interactions from ATAC-seq data, both of which are consistent with prior inferences for the roles of the unfolded protein response (UPR) pathway, of which ATF4 and XPB1 are both members, in regulating venom genes. Being a gene of interest based on expression variation, myotoxin differs from most other venom gene families, such as SVMPs, SVSPs, and PLA2s, in that the genome assembly and annotation of this region remains poorly resolved (Schield et al. 2019a; Gopalan et al. 2022). From what we do understand, myotoxin paralog number appears to vary substantially even within *C. viridis*, yet paralogs appear to be identical in protein-coding sequence (Gopalan et al. 2022). Thus, unlike other multigene venom families, myotoxin expression may be primarily modulated by dosage (e.g. gene copy number variation), although our data also suggest that regulation of trans-acting factors (and potentially chromatin variation) may also play key roles.

Taken together, our results suggest that most phenotypic differences in venom between species are likely driven by changes in TF expression in the venom gland, whereas expression is tuned at finer scales by functional nucleotide variation and variable chromatin access at CREs. Indeed, recent findings have supported the hypothesis that the larger effect-size changes of the trans-regulatory environment may tend to evolutionarily persist when restricted to only some tissues (Barr et al. 2023). This would suggest that a fraction of observed venom compositional variation between lineages may result from divergence in trans-regulatory factor expression variation in the venom gland.

A Role for Variation in CTCF-mediated Insulation in Expression Variation

The protein CTCF, originally identified as a transcriptional repressor, is known to play broad roles as an "insulator" through its roles in defining chromatin boundaries and directing of chromatin looping structures that can modulate enhancer–promoter interactions (Lobanenkov et al. 1990; Ong and Corces 2014; Ren et al. 2017). Prior studies on snake venom regulation have identified the roles of CTCF in directing gene regulatory interactions across multiple venom gene clusters (Schield et al. 2019a; Liao et al. 2021; Perry et al. 2022). Based on modeling and additional analyses, we find evidence for the effects of binding of the insulator protein CTCF on gene expression variation. Our results suggest that CTCF-mediated insulation may be used to direct gene expression changes across recent evolutionary scales. The example demonstrating this leverages the complex regulatory architecture of the viperid SVSP cluster, which is a result of chromatin loops, often guided by CTCF, forming topologically associated domains isolating paralogs from their neighbors (Perry et al. 2022). Our sampling encompassing fine-scaled evolutionary variation has allowed us to identify additional features and associations related to the regulatory nature of SVSP9. We find that accessibility at a known CTCF-bound locus between the promoter and enhancers of SVSP9 produces a negative correlation with gene expression, consistent with the expected effects of CTCF as an insulator that can negatively mediate enhancer–promoter interactions through its action in mediating chromatin loops. While we do identify two new putative regulatory loci and a regulatory role of a CTCF locus for SVSP9, given the often multienhancer nature of viperid venom genes (Perry et al. 2022), this does not exclude the presence of other distal regulatory loci beyond our search space which could more accurately explain the regulatory nature of SVSP9.

Roles of Functional and Structural Variation at CREs

Our findings also provide new insight into the potentially distinct roles and mechanisms of functional diversity in promoters and enhancers in the context of evolutionary modulation of gene expression, with enhancer variation being dominated by chromatin variation while promoter variation is dominated by genotype variation. In snake venom genes, enhancers tend to be less genetically variable than promoters, yet show higher variation in chromatin accessibility and TF binding. Whether this is a generalizable trend, or specific to venom genes, remains unresolved. A recent study has suggested that snake venom genes may have elevated allelic diversity due to pervasive balancing selection (Schield et al. 2022), which may also drive elevated diversity at the proximal promoter loci of these genes but be reduced as more distant enhancer loci.

Prior studies on snake venom gene clusters have linked venom composition and gene expression variation to larger-scale genomic mechanisms such as structural diversity, which drives venom compositional variation between species (Casewell et al. 2011; Dowell et al. 2016; Giorgianni et al. 2020; Margres et al. 2021). Additional studies have also quantified chromatin accessibility (Margres et al. 2021; Perry et al. 2022) and DNA methylation (Margres et al. 2021), linking these to variation in expression across venom genes within single individual snakes (Margres et al. 2021; Perry et al. 2022). The work presented here extends the findings of prior studies through the integration of functional genomic data across multiple individuals and species that enables the contextualization of the evolutionary roles of chromatin state as well as

genomic variation in generating venom gene expression. This now allows for a far more comprehensive understanding of precise mechanisms by which modifications to chromatin access and nucleotides at CREs act as regulatory inputs to tune a highly selected phenotype within and across species.

Several prior studies have attempted to define expectations for the roles of general cis- and trans-effects in driving divergent inter-species diversity through independently measuring cis-element activity, chromatin accessibility, and gene expression across species (Berthelot et al. 2018; Pizzollo et al. 2018; Edsall et al. 2019; Barr et al. 2023). Developing an integrated quantitative understanding of gene expression variation in the context of multiple forms of regulatory variation has, however, remained a challenge. The distinct nature of our experimental design here, using shallow-divergence comparative studies, holds great potential as an alternative and productive way forward for detecting molecular variation and linking these diverse sources of variation to their relevance in directing gene expression, particularly in model systems in which mutagenesis is not feasible.

One critical axis of gene regulatory variation that was not directly explored in this study is the role of noncoding RNAs. Prior studies have implicated miRNAs as key underlying factors that explain divergent venom expression patterns (Durban et al. 2013; Zheng et al. 2023), and long noncoding RNAs that may also be involved in snake venom diversity and regulation (Gopalan et al. 2022; Zheng et al. 2023) have been identified. Though one study alternatively found that posttranscriptional mechanisms play a negligible role in venom regulation (Rokyta et al. 2015), our comparison of mRNA versus protein abundance highlights multiple venom genes that show lower than expected protein abundance compared to mRNA abundance (including myotoxin, some PLA_2s, and SVMPs), consistent with miRNAs playing a posttranscriptional regulatory role in venom composition variation. Future work to integrate the roles of noncoding RNAs more directly in modulating venom gene expression phenotypes would provide a more comprehensive, and likely more complex, understanding of the factors that ultimately modulate venom expression phenotypes and venom composition.

Conclusion

Recent studies have used hybrid or cybrid experimental designs to provide valuable insight into the relative roles of cis- versus trans-gene regulatory components in modulating gene expression phenotypes. In contrast, this study represents one of a few (Wittkopp et al. 2008; Jones et al. 2012) that has interrogated naturally existing variation in GRNs at fine evolutionary scales. Consequently, it provides valuable baseline expectations for the extent and functional impacts of naturally occurring gene regulatory variants. Our findings highlight a surprisingly high degree of naturally occurring gene regulatory variation and the extensive diversity of underlying mechanisms that appear to play dominant roles in different genes and gene families. This relatively small-scale study suggests that more powerful larger-scale comparative functional genomics studies hold exciting promise as hypothesis-generating and testing platforms for gene regulatory function, and for inferring how regulatory variation may manifest in phenotypic variation.

Materials and Methods

Tissue Sampling

All animal collection, housing, and sampling was conducted according to an approved and registered IACUC protocol (2303D-SM-S-26; S.P. Mackessy) at the University of Northern Colorado, and animals were collected under approved state permits (Arizona, Colorado, Utah, New Mexico, and Texas). To initiate venom production, venom was manually extracted from both venom glands one day prior to sacrifice. Animals were anesthetized using isoflurane and humanely sacrificed by severing the spinal cord. Left and right venom gland, right accessory venom gland, skin, pancreas, skeletal muscle, heart, and liver tissues were immediately dissected out and snap frozen in liquid nitrogen. For this study, only venom, blood, left and right venom gland tissues were used.

mRNA-seq and Venom Protein Data Generation and Analysis

Total RNA was extracted from snap-frozen tissues using TRIzol reagent (Invitrogen Life Technologies, No. 15596026). For this study, all RNA extractions were performed in a single batch. A single left venom gland sample was excluded from the study due to poor data quality, leaving a total of 23 venom gland tissues. Library preparation and sequencing were performed by Novogene (Sacramento, California). Briefly, mRNA was selected from total RNA using poly-T oligo-attached magnetic beads, followed by fragmentation, reverse transcription, adapter ligation, and amplification by PCR. The library was quality checked for size distribution using a Bioanalyzer (Agilent 5400). mRNA libraries were then sequenced on an Illumina NovaSeq 6000 using 150 bp paired-end reads. Raw reads were quality trimmed using Trimmomatic v0.39 with the settings LEADING:20 TRAILING:20 MINLEN:32 AVGQUAL:30 (Bolger et al. 2014), and resulting paired reads were mapped to the annotated *Crotalus viridis* reference genome (NCBI GCA_003400415.2, Schield et al. 2019a) using STAR v2.7.9a (Dobin et al. 2013). Reads mapped to genic features in the reference annotation were counted by exon and summarized by gene using featureCounts v1.6.3 (Liao et al. 2014) to provide estimates of gene expression. Differential gene expression between *C. viridis* and non-*C. viridis*

individuals, and individuals within *C. viridis* populations was performed using DEseq2 v1.30.1 (Love et al. 2014) in R (R Core Team 2022). TFs found to be differentially expressed across species were considered "of significance" and were appended to a previously generated set of TFs from a prior study (Perry et al. 2022) for the purposes of TFBS scanning (see below). DESeq2 was then used to produce library-size normalized count matrices (using the "counts' command) and variance stabilizing transformed count matrices (using the "vst" command), the latter of which was used to produce heatmaps in R.

WGCNA (Langfelder and Horvath 2008) was used to perform module co-expression analyses and to estimate module-trait significance values. WGCNA was run twice with standard settings. It was run initially with all left and right venom gland samples (N = 23) to estimate module-trait significance values for species identity (i.e. *C. viridis*, *C. o. lutosus*, *C. o. concolor* and *C. cerberus*). To generate gene–gene correlation matrices from gene expression, Pearson's rho was calculated in R using the "rcorr" function from the "Hmisc" package (cran.r-project.org/web/packages/Hmisc) and the coefficient matrix was filtered for *P*-value < 0.05 and FDR < 0.1 to produce a significance-filtered TF-venom gene correlation matrix.

Venom Proteomics

Lyophilized venoms were resuspended in 8 M urea/0.1 M Tris (pH 8.5), reduced with 5 mM TCEP (tris (2-carboxyethyl) phosphine) for 20 min, and alkylated with 50 mM 2-chloroacetamide for 15 min in the dark all at room temperature. Samples were diluted 4 times with 100 mM Tris–HCl (pH 8.5) and trypsin digested at an enzyme/substrate ratio of 1:20 overnight at 37°C. Digestion was stopped with formic acid (FA), and proteolytic peptides were purified with Pierce C18 Spin Tips (ThermoFisher Scientific). Samples were dried in a speed vacuum and resuspended in 0.1% FA.

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) was performed using an Easy nLC 1000 instrument coupled with a Q Exactive HF Mass Spectrometer (both from ThermoFisher Scientific). Digested peptides were loaded on a C_{18} column (100 μM inner diameter × 20 cm) packed in-house with 2.7 μm Cortecs C18 resin, and separated at a flow rate of 0.4 μl/min with solution A (0.1% FA) and solution B (0.1% FA in acetonitrile) under the following conditions: isocratic at 4% B for 3 min, followed by 4% to 32% B for 102 min, 32% to 55% B for 5 min, 55% to 95% B for 1 min and isocratic at 95% B for 9 min. Mass spectrometry was performed in data-dependent acquisition mode. Full MS scans were obtained from *m/z* 300 to 1800 at a resolution of 660,000, an automatic gain control (AGC) target of 1×10^6, and a maximum injection time (IT) of 50 ms. The top 15 most abundant precursors with an intensity threshold of 9.1×10^3 were selected for MS/MS acquisition at a 15,000 resolution, 1×10^5 AGC, and a maximal IT of 110 ms. The isolation window was set to 2.0 m/*z* and ions were fragmented at a normalized collision energy of 30. Dynamic exclusion was set to 20 s.

Fragmentation spectra were interpreted against a database containing translated sequences derived from a public transcriptome (Schield et al. 2019a) using the MSFragger-based FragPipe computational platform (Kong et al. 2017; Yu et al. 2020). Our reference proteome database contains highly specific protein sequences that increase the likelihood of unique peptide mapping, in contrast to some publicly available databases where the high degree of homology between proteins in the database may cause multiple mapping of peptides to proteins. Contaminants and reverse decoys were added to the database automatically. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine was selected as a variable modification. The precursor-ion mass tolerance and fragment-ion mass tolerance were set at 10 and 12 ppm, respectively. Up to 2 missed tryptic cleavages were allowed and the protein-level false discovery rate (FDR) was set to <1%.

ATAC-seq Data Generation, Processing, and Analysis

ATAC-seq data were generated for right venom gland tissue samples by Active Motif (Carlsbad, California), derived from snap-frozen glands of the same animals used for mRNA-seq. Raw ATAC-seq reads were mapped to the *C. viridis* reference genome using the "mem" algorithm from bwa v0.7.17 with default settings (Li 2013). Procedures for ATAC-seq data processing were largely based on an existing set of methods laid out in Perry et al. (2022). Briefly, PCR duplicates were removed using Picard Tools v2.22.6 (broadinstitute.github.io/picard), and samtools v1.9 (Li et al. 2009) was used to remove all nonunique alignments and improperly paired reads. The "randsample" command from MACS2 v2.2.7.1 (Zhang et al. 2008) was used to randomly down-sample reads to the number of tags present in the sample with the fewest tags. ATAC-seq peaks were called using MACS2 with a q-value cutoff of 0.001. To assess ATAC-seq data quality, we calculated the fraction of reads in peaks (FRiP) for each sample using featureCounts (Liao et al. 2014). Two ATAC-seq samples (a mid-latitude *C. viridis* (CV1081) and the *C. o. lutosus* (CV0987) individual) were excluded from subsequent ATAC-seq analyses due to low FRiP scores (supplementary table S1, Supplementary Material online). The "merge" command from bedtools v2.29.2 (Quinlan and Hall 2010) was used to merge partially overlapping peak regions between two or more samples. This set of merged peak regions was used for downstream analyses. Bigwig files of raw read coverage in each sample were generated using the "bamCoverage" command in deepTools v3.1.3

(Ramírez et al. 2016) with a bin size of 32 bp. The "multiBigwigSummary" command with options "–BED" and "–outRawCounts' was then used to output a length-normalized average ATAC-seq signal matrix for the merged peak set. edgeR v3.32.1 in R was used to calculate TMM normalization factors for all samples, and these factors were then used to generate normalized bigwig files again using the "bamCoverage" command in deepTools. These processed, normalized bigWig files were used to produce ATAC-seq read depth tracks in R using the ggcoverage (Song and Wang 2023) package in R.

Generation and Analysis of Genome resequencing Data

High coverage, re-sequenced genomes for the 12 samples (supplementary table S1, Supplementary Material online) used for prior analyses were also generated. DNA was extracted from snap-frozen blood using a phenol–chloroform–isoamyl (Invitrogen Life Technologies, No. 15593031) extraction protocol. Libraries were prepared from the DNA elution using Illumina Nextera Flex kits which were then sequenced on an Illumina NovaSeq 6000 using 150 bp paired-end reads, targeting an average coverage of 50X. Reads were filtered using Trimmomatic with the same settings specified above. These reads were then mapped to the reference genome using "bwa" at a mean unique read mapping rate of 97.93% and a mean coverage of 50.2X (supplementary table S1, Supplementary Material online).

Methods for variant calling and filtering were performed following methods previously described (Schield et al. 2022). Briefly, individual genomic variants were called using the "HaplotypeCaller" command from GATK v4.0.8.1 (McKenna et al. 2010) following best practices recommendations, and the resulting individual genomic variant call format (gVCF) files were then combined with the "CombineGVCF" command. The cohort gVCF was hard filtered based on GATK's parameter threshold recommendations with the "VariantFiltration" and "SelectVariants' commands, which resulted in 17,051,557 variants.

Variants from the gVCF were projected onto the reference venom CRE sequences using the "consensus' command from bcftools v1.16 (Danecek and McCarthy 2017) to produce individual variant sequences for each venom CRE. These variants were also checked for coverage depth, and base call error (supplementary table S3, Supplementary Material online). The reference CRE sequences used here were obtained from a prior study that investigated venom regulatory architecture in *C. viridis* by integrating multiple functional genomics approaches (including ChIP-seq, ATAC-seq, and chromatin contact data) to assign venom genes to genomic regions (Perry et al. 2022). Nucleotide diversity (π) from these sequence files was calculated using a custom R script, in which consensus sequences were first aligned using muscle (Edgar 2004). To identify variants that were found only in *C. viridis,* the gVCF was filtered using bcftools "filter" command to retain variants where all *C. viridis* samples contain the reference allele or the alternate allele, but non-*C. viridis* samples contain the opposite, respectively.

TFBS Scanning and Footprinting Analyses

The JASPAR 2022 non-redundant vertebrate motif database (Castro-Mondragon et al. 2022) was subset to retain the 161 TFs of interest with respect to venom gene regulation from a prior study (Perry et al. 2022) as well as TFs differentially expressed between *C. viridis* and non-*C. viridis* from venom mRNA-seq data, described above. The individual variant CRE sequence files described above were concatenated and scanned for TFBSs with this custom JASPAR motif set with the "scan" option in Ciiider v0.9 (Gearing et al. 2019), using the default motif similarity threshold of 0.15. Differential binding was assessed using ATAC-seq footprinting analysis, which was performed using TOBIAS v.0.12.4 (Bentsen et al. 2020) following the methods described in a prior study (Perry et al. 2022). Briefly, insertion site bias was corrected using the "ATACorrect" command, footprint scores were calculated using "ScoreBigwig" and "BINDetect" was used to calculate sample-specific footprint score binding thresholds. The sample-wide set of scanned TFBS regions was used as input to deepTools "multiBigwigSummary", using the same options described above, to produce a sequence length-normalized matrix of ATAC-seq scores at all TFBSs which were then binarized using the binding threshold to contrast bound and unbound TFBSs per individual. VCF variants were then intersected with the bound TFBSs with a custom R script to assess differentially bound TFBSs which contain variants at the motif.

Exploring Evidence of Evolutionary Correlates With Venom Expression Variation

Input feature tables for linear modeling were constructed per venom gene by assembling datasets as follows. For each individual, tables contained DeSEQ2-normalized gene expression for the gene of interest, accessibility scores at peaks falling within promoters and/or enhancers for that gene, accessibility scores at peaks containing loci bound by CTCF (Perry et al. 2022) which fall within a window defined as a $\pm$ 1 kb extension around the furthest separated features of a venom gene array (i.e. known CREs or coding regions), accessibility at the top three non-CRE, non-CTCF associated peaks with the highest variation in ATAC-seq scores within the same venom array windows defined above, binarized footprints for venom-regulating TFs binding TFBSs in CREs for the gene ('0' = TF is not bound at TFBS in that sample, "1' = TF is bound at TFBS in that sample), DeSEQ2-normalized expression for all venom-regulating TFs, and numerically recoded genotypes ('0' = homozygous reference, "1' =

heterozygous, "2' = homozygous alternate) for variants that occur within CREs of that venom gene.

A guide species tree for phylogenetic PCA was obtained from a prior study (Schield et al. 2019b), and middle, southern, and northern latitude *C. viridis* populations were collapsed. All original feature values were first transformed into phylogenetically independent contrasts using the "pic" function from the ape package (Paradis et al. 2004) to account for shared covariance among the species (Felsenstein 1985). We explored for evidence of evolutionary correlates with venom expression variation using principal component regression based on phylogenetic independent contrasts computed for each set of features. Specifically, we used this approach to evaluate whether variation across regulatory features (and classes of features) were correlated with venom expression according to the classes of input predictor features described above. Where the number of measured variables exceeded the number of samples, PCA of the phylogenetically corrected features was performed to obtain the first principal axis (PC1) for that feature class; these components were subsequently used as input predictor variables for multiple regression using the phylogenetic independent contrasts of normalized venom expression as the response variable. PCAs were conducted for the phylogenetic contrasts of each feature using the "prcomp" base function in R, and linear models were fit using the "lm" function.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

Support for this work was provided by a National Science Foundation (IOS-2307044) to T.A.C., S.P.M., A.J.S., and R.H.A., and (DEB-2208959) to J.M.B.

Data Availability

ATAC-seq, RNA-seq, and whole-genome resequencing data generated in this study has been submitted to the NCBI under BioProject accession PRJNA1061517. Processed ATAC-seq data have also been submitted to NCBI's Gene Expression Omnibus, under accession GSE254420. Code and additional data for reproducing main text figures and key analyses are available at github.com/SidG13/CrotalusVenomFxnGenomics. Any additional information required to reanalyze the data reported here is available upon request to the Author for Correspondence.

Literature Cited

Amazonas DR, Portes-Junior JA, Nishiyama-Jr MY, Nicolau CA, Chalkidis HM, Mourão RHV, Grazziotin FG, Rokyta DR, Gibbs HL, Valente RH, et al. Molecular mechanisms underlying intraspecific variation in snake venom. J Proteomics. 2018:181:60–72. https://doi.org/10.1016/j.jprot.2018.03.032.

Barr KA, Rhodes KL, Gilad Y. The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. Genome Biol. 2023:24(1):207. https://doi.org/10.1186/s13059-023-03019-3.

Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun. 2020:11(1):4267. https://doi.org/10.1038/s41467-020-18035-1.

Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol. 2018:2(1):152–163. https://doi.org/10.1038/s41559-017-0377-2.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014:30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015:523(7561):486–490. https://doi.org/10.1038/nature14590.

Casewell NR, Huttley GA, Wüster W. Dynamic evolution of venom proteins in squamate reptiles. Nat Commun. 2012:3(1):1–10. https://doi.org/10.1038/ncomms2065.

Casewell NR, Jackson TNW, Laustsen AH, Sunagar K. Causes and consequences of snake venom variation. Trends Pharmacol Sci. 2020:41(8):570–581. https://doi.org/10.1016/j.tips.2020.05.006.

Casewell NR, Wagstaff SC, Harrison RA, Renjifo C, Wüster W. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. Mol Biol Evol. 2011:28(9):2637–2649. https://doi.org/10.1093/molbev/msr091.

Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. Complex cocktails: the evolutionary novelty of venoms. Trends Ecol Evol. 2013:28(4):219–229. https://doi.org/10.1016/j.tree.2012.10.020.

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022:50(D1):D165–D173. https://doi.org/10.1093/nar/gkab1113.

Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol Cell. 2002:9(2):279–289. https://doi.org/10.1016/S1097-2765(02)00459-8.

Colis-Torres A, Neri-Castro E, Strickland JL, Olvera-Rodríguez A, Borja M, Calvete J, Jones J, Parkinson CL, Bañuelos J, López de León J, et al. Intraspecific venom variation of Mexican West Coast Rattlesnakes (*Crotalus basiliscus*) and its implications for antivenom production. Biochimie. 2021:192:111–124. https://doi.org/10.1016/j.biochi.2021.10.006.

Crombach A, Hogeweg P. Evolution of evolvability in gene regulatory networks. PLoS Comput Biol. 2008:4(7):e1000112. https://doi.org/10.1371/journal.pcbi.1000112.

Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017:33(13):2037–2039. https://doi.org/10.1093/bioinformatics/btx100.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013:29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dowell NL, Giorgianni MW, Kassner VA, Selegue JE, Sanchez EE, Carroll SB. The deep origin and recent loss of venom toxin genes

in rattlesnakes. Curr Biol. 2016:26(18):2434–2445. https://doi.org/10.1016/j.cub.2016.07.038.
Durban J, Pérez A, Sanz L, Gómez A, Bonilla F, Rodríguez S, Chacón D, Sasa M, Angulo Y, Gutiérrez JM, et al. Integrated "omics" profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, *Crotalus simus simus*. BMC Genomics. 2013:14(1):1–17. https://doi.org/10.1186/1471-2164-14-234.
Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004:32(5):1792–1797. https://doi.org/10.1093/nar/gkh340.
Edsall LE, Berrio A, Majoros WH, Swain-Lenz D, Morrow S, Shibata Y, Safi A, Wray GA, Crawford GE, Allen AS. Evaluating chromatin accessibility differences across multiple primate species using a joint modeling approach. Genome Biol Evol. 2019:11(10):3035–3053. https://doi.org/10.1093/gbe/evz218.
Emerson JJ, Li W-H. The genetic basis of evolutionary change in gene expression levels. Phil Trans R Soc Lond B Biol Sci. 2010:365(1552):2581–2590. https://doi.org/10.1098/rstb.2010.0005.
Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985:125(1):1–15. https://doi.org/10.1086/284325.
Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K. NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res. 2013:23(8):1195–1209. https://doi.org/10.1101/gr.148080.112.
Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, Gould JA, Forster SC, Hertzog PJ. CiiiDER: a tool for predicting and analysing transcription factor binding sites. PLoS One. 2019:14(9):e0215495. https://doi.org/10.1371/journal.pone.0215495.
Giorgianni MW, Dowell NL, Griffin S, Kassner VA, Selegue JE, Carroll SB. The origin and diversification of a novel protein family in venomous snakes. Proc Natl Acad Sci U S A. 2020:117(20):10911–10920. https://doi.org/10.1073/pnas.1920011117.
Gopalan SS, Perry BW, Schield DR, Smith CF, Mackessy SP, Castoe TA. Origins, genomic structure and copy number variation of snake venom myotoxins. Toxicon. 2022:216:92–106. https://doi.org/10.1016/j.toxicon.2022.06.014.
Hofmann EP, Rautsaw RM, Strickland JL, Holding ML, Hogan MP, Mason AJ, Rokyta DR, Parkinson CL. Comparative venom-gland transcriptomics and venom proteomics of four Sidewinder Rattlesnake (*Crotalus cerastes*) lineages reveal little differential expression despite individual variation. Sci Rep. 2018:8(1):1–15. https://doi.org/10.1038/s41598-018-33943-5.
Holding ML, Biardi JE, Gibbs HL. Coevolution of venom function and venom resistance in a rattlesnake predator and its squirrel prey. Proc Royal Soc B. 2016:283(1829):20152841. https://doi.org/10.1098/rspb.2015.2841.
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012:484(7392):55–61. https://doi.org/10.1038/nature10944.
Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat Methods. 2017:14(5):513–520. https://doi.org/10.1038/nmeth.4256.
Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008:9(1):559. https://doi.org/10.1186/1471-2105-9-559.
Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013. https://arxiv.org/abs/1303.3997.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009:25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
Liao X, Guo S, Yin X, Liao B, Li M, Su H, Li Q, Pei J, Gao J, Lei J, et al. Hierarchical chromatin features reveal the toxin production in *Bungarus multicinctus*. Chin Med. 2021:16(1):90. https://doi.org/10.1186/s13020-021-00502-6.
Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014:30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656.
Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, Goodwin GH. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5′-flanking sequence of the chicken c-myc gene. Oncogene. 1990:5:1743–1753.
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014:15(12):1–21. https://doi.org/10.1186/s13059-014-0550-8.
Luna MSA, Hortencio TMA, Ferreira ZS, Yamanouye N. Sympathetic outflow activates the venom gland of the snake *Bothrops jararaca* by regulating the activation of transcription factors and the synthesis of venom gland proteins. J Exp Biol. 2009:212(10):1535–1543. https://doi.org/10.1242/jeb.030197.
Mackessy SP. Evolutionary trends in venom composition in the Western Rattlesnakes (*Crotalus viridis* sensu lato): toxicity vs. tenderizers. Toxicon. 2010:55(8):1463–1474. https://doi.org/10.1016/j.toxicon.2010.02.028.
Mackessy SP. Handbook of venoms and toxins of reptiles. Florida, Boca Raton: CRC press; 2021.
Margres MJ, Rautsaw RM, Strickland JL, Mason AJ, Schramer TD, Hofmann EP, Stiers E, Ellsworth SA, Nystrom GS, Hogan MP, et al. The Tiger Rattlesnake genome reveals a complex genotype underlying a simple venom phenotype. Proc Natl Acad Sci U S A. 2021:118(4):e2014634118. https://doi.org/10.1073/pnas.2014634118.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010:20(9):1297–1303. https://doi.org/10.1101/gr.107524.110.
Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014:15(4):234–246. https://doi.org/10.1038/nrg3663.
Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021:30(1):187–200. https://doi.org/10.1002/pro.3978.
Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004:20(2):289–290. https://doi.org/10.1093/bioinformatics/btg412.
Perry BW, Gopalan SS, Pasquesi GIM, Schield DR, Westfall AK, Smith CF, Koludarov I, Chippindale PT, Pellegrino MW, Chuong EB, et al. Snake venom gene expression is coordinated by novel regulatory architecture and the integration of multiple co-opted vertebrate pathways. Genome Res. 2022:32(6):1058–1073. https://doi.org/10.1101/gr.276251.121.
Perry BW, Schield DR, Westfall AK, Mackessy SP, Castoe TA. Physiological demands and signaling associated with snake venom production and storage illustrated by transcriptional analyses of venom glands. Sci Rep. 2020:10(1):1–10. https://doi.org/10.1038/s41598-020-75048-y.
Pizzollo J, Nielsen WJ, Shibata Y, Safi A, Crawford GE, Wray GA, Babbitt CC. Comparative serum challenges show divergent patterns of gene expression and open chromatin in human and

chimpanzee. Genome Biol Evol. 2018:10(3):826–839. https://doi.org/10.1093/gbe/evy041.
Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010:26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.
R Core Team. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2022. Available from: http://www.r-project.org/.
Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016:44(W1):W160–W165. https://doi.org/10.1093/nar/gkw257.
Ren G, Jin W, Cui K, Rodrigez J, Hu G, Zhang Z, Larson DR, Zhao K. CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. Mol Cell. 2017:67(6):1049–1058.e6. https://doi.org/10.1016/j.molcel.2017.08.026.
Rockman MV, Wray GA. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol. 2002:19(11):1991–2004. https://doi.org/10.1093/oxfordjournals.molbev.a004023.
Rokyta DR, Margres MJ, Calvin K. Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. G3 (Bethesda). 2015:5(11):2375–2382. https://doi.org/10.1534/g3.115.020578.
Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012:13(7):505–516. https://doi.org/10.1038/nrg3229.
Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, Adams RH, Corbin AB, Smith CF, Ramesh B, et al. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. Genome Res. 2019a:29(4):590–601. https://doi.org/10.1101/gr.240952.118.
Schield DR, Perry BW, Adams RH, Card DC, Jezkova T, Pasquesi GIM, Nikolakis ZL, Row K, Meik JM, Smith CF, et al. Allopatric divergence and secondary contact with gene flow: a recurring theme in rattlesnake speciation. Biol J Linn Soc. 2019b:128(1):149–169. https://doi.org/10.1093/biolinnean/blz077.
Schield DR, Perry BW, Adams RH, Holding ML, Nikolakis ZL, Gopalan SS, Smith CF, Parker JM, Meik JM, DeGiorgio M, et al. The roles of balancing selection and recombination in the evolution of rattlesnake venom. Nat Ecol Evol. 2022:6(9):1367–1380. https://doi.org/10.1038/s41559-022-01829-5.
Smith CF, Nikolakis ZL, Ivey K, Perry BW, Schield DR, Balchan NR, Parker J, Hansen KC, Saviola AJ, Castoe TA, et al. Snakes on a plain: biotic and abiotic factors determine venom compositional variation in a wide-ranging generalist rattlesnake. BMC Biol. 2023:21(1):136. https://doi.org/10.1186/s12915-023-01626-x.
Song Y, Wang J. Ggcoverage: an R package to visualize and annotate genome coverage for various NGS data. BMC Bioinformatics. 2023:24(1):309. https://doi.org/10.1186/s12859-023-05438-2.
Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012:13(9):613–626. https://doi.org/10.1038/nrg3207.
Tasoulis T, Isbister GK. A review and database of snake venom proteomes. Toxins (Basel). 2017:9(9):290. https://doi.org/10.3390/toxins9090290.
Thompson D, Regev A, Roy S. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annu Rev Cell Dev Bi. 2015:31(1):399–428. https://doi.org/10.1146/annurev-cellbio-100913-012908.
Westfall AK, Gopalan SS, Perry BW, Adams RH, Saviola AJ, Mackessy SP, Castoe TA. Single-cell heterogeneity in snake venom expression is hardwired by co-option of regulators from progressively activated pathways. Genome Biol Evol. 2023:15(6):evad109. https://doi.org/10.1093/gbe/evad109.
Wittkopp PJ. Variable gene expression in eukaryotes: a network perspective. J Exp Biol. 2007:210(9):1567–1575. https://doi.org/10.1242/jeb.002592.
Wittkopp PJ, Haerum BK, Clark AG. Regulatory changes underlying expression differences within and between *Drosophila* species. Nat Genet. 2008:40(3):346–350. https://doi.org/10.1038/ng.77.
Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2012:13(1):59–69. https://doi.org/10.1038/nrg3095.
Yu F, Teo GC, Kong AT, Haynes SE, Avtonomov DM, Geiszler DJ, Nesvizhskii AI. Identification of modified peptides using localization-aware open search. Nat Commun. 2020:11(1):4065. https://doi.org/10.1038/s41467-020-17921-y.
Zancolli G, Casewell NR. Venom systems as models for studying the origin and regulation of evolutionary novelties. Mol Biol Evol. 2020:37(10):2777–2790. https://doi.org/10.1093/molbev/msaa133.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008:9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137.
Zheng H, Wang J, Fan H, Wang S, Ye R, Li L, Wang S, Li A, Lu Y. Comparative venom multiomics reveal the molecular mechanisms driving adaptation to diverse predator–prey ecosystems in closely related sea snakes. Mol Biol Evol. 2023:40(6):msad125. https://doi.org/10.1093/molbev/msad125.

Associate editor: Toni Gossmann