1. Methods

(a) Variant Calling

Merged reads were mapped to the consensus transcriptome using BWA-MEM and reads with more than two mismatches — via gaps or nucleotide differences — were removed [1]. Picard version 2.12.1 (broadinstitute.github.io/picard/) was used to sort and index the aligned reads and GATK version 3.6 (software.broadinstitute.org/gatk/) was then used for local realignment of regions with indels. GATK was also used to remove reads shorter than 120 nucleotides using the OverclippedRead and ReadLength filters. GATK's HaplotypeCaller was used to call variants and GenotypeGVCFs was used to perform joint genotyping. We then filtered SNPs using the VariantFiltration function of GATK with the following criteria: QualByDepth < 2.0, FisherStrand > 60.0, MappingQuality < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0. Variants were phased using WhatsHap version 0.15 [2]. We further filtered SNPs by manually reassigning SNPs as missing data when transcript coverage was 0 for greater than 5% of the transcript's total length in order to avoid partial transcripts and poorly mapped reads.

(b) Sequence Evolution

To summarize the variant data, we used SnpEff v.4.3 to determine whether each SNP was synonymous or nonsynonymous based on the consensus transcriptome [3]. All statistical analyses were performed in R version 3.5.1 [4]. We calculated the average number of SNPs per kilobase and tested for significant differences betweeen toxins and nontoxins using a linear regression. We performed a X^2-test to determine if the number of synonymous and nonsynonymous mutations was significantly associated with whether the gene was a toxin or nontoxin. Additionally, nucleotide diversity (π) was calculated using vcftools version 0.1.15 [5] and significant differences in mean π between toxins and nontoxins was determined by linear regression.

We tested for evidence of selection in toxin and nontoxin loci by calculating Tajima's D and Weir and Cockerham's F_{ST} across all individuals and populations. Tajima's D and F_{ST} were calculated for each locus using vcftools [5]. Then, to determine if – on average – toxins had significantly higher coding sequence evolution than nontoxins, we performed linear regressions between toxin and nontoxin estimates of Tajima's D and F_{ST}. To account for differences in sample sizes between toxins and nontoxins, we randomly sub-sampled nontoxins to toxin sample size and performed 1,000 bootstrap replicates of the regression analyses. The proportion of the bootstraps (*b*) which detected significance ($p < 0.05$) is reported alongside the results from the full dataset.

To look for specific toxins under selection, we generated a null distribution from the nontoxin transcripts and looked for toxin outliers outside the 95th percentile. Additionally, to test if low-expression toxins were under stronger positive or balancing selection pressures, we used a linear regression to correlate average toxin expression [6] to estimates of Tajima's D and F_{ST}. We tested positive and negative values of Tajima's D separately to account for potential differences in expression related to alternate selection pressures.

Lastly, we tested for differentiation between lineages using F_{ST}. Pairwise F_{ST} was calculated in vcftools for each locus between the four phylogeographic lineages (*i.e* Colorado, Sonora, North Mojave, South Mojave). We calculated the mean toxin and nontoxin F_{ST} estimates to examine overall divergence of the lineages and to compare toxins to nontoxins. We performed linear regression to assess significant differences in mean estimates of toxins and nontoxins. To determine if any toxin loci were significantly differentiated between any two populations, we generated null distributions from the nontoxin estimates for each pairwise lineage comparison and looked for toxin outliers outside the 95th percentile. To determine if the average F_{ST} between populations was correlated with geographic distance, we performed Mantel tests in the R package vegan [7] based on the average GPS coordinate for each population. Finally, to maximize our ability to detect evidence of sequence differentiation, we combined lineages into larger clades and re-performed F_{ST} analyses. Larger clades were based on the phylogeny recovered by [6] or on geography. Based on the phylogeny, the West clade consisted of the North and South Mojave lineages, while the Colorado and Sonoran lineages were combined into an "East" clade. Based on geography, however, the West clade included the Colorado lineage and the East clade was composed of only the Sonoran lineage.

References

1. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv1303.3997*
2. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schönhuth A, Marschall T. 2016 WhatsHap: fast and accurate read-based phasing. *bioRxiv* (doi:10.1101/085050)
3. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92.

(doi:10.4161/fly.19695)
4. R Core Team. 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
5. Danecek P *et al.* 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)
6. Hofmann EP, Rautsaw RM, Strickland JL, Holding ML, Hogan MP, Mason AJ, Rokyta DR, Parkinson CL. 2018 Comparative venom-gland transcriptomics and venom proteomics of four Sidewinder Rattlesnake (*Crotalus cerastes*) lineages reveal little differential expression despite individual variation. *Sci. Rep.* **8**, 15534. (doi:10.1038/s41598-018-33943-5)
7. Oksanen J *et al.* 2017 vegan: Community Ecology Package.